

*entropy*

# New Developments in Statistical Information Theory Based on Entropy and Divergence Measures

---

Edited by  
Leandro Pardo

Printed Edition of the Special Issue Published in *Entropy*

# **New Developments in Statistical Information Theory Based on Entropy and Divergence Measures**



# New Developments in Statistical Information Theory Based on Entropy and Divergence Measures

Special Issue Editor

**Leandro Pardo**

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade



*Special Issue Editor*

Leandro Pardo

Universidad Complutense de Madrid

Spain

*Editorial Office*

MDPI

St. Alban-Anlage 66

4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Entropy* (ISSN 1099-4300) from 2017 to 2019 (available at: [https://www.mdpi.com/journal/entropy/special-issues/Divergence\\_Measures](https://www.mdpi.com/journal/entropy/special-issues/Divergence_Measures))

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, Article Number, Page Range.

**ISBN 978-3-03897-936-4 (Pbk)**

**ISBN 978-3-03897-937-1 (PDF)**

© 2019 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license, which allows users to download, copy and build upon published articles, as long as the author and publisher are properly credited, which ensures maximum dissemination and a wider impact of our publications.

The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons license CC BY-NC-ND.

# Contents

<b>About the Special Issue Editor</b>	vii
<b>Leandro Pardo</b>	
Reprinted from: <i>Entropy</i> 2019, 391, 21, doi:10.3390/e21040391 . . . . .	1
<b>Abhik Ghosh and Ayanendranath Basu</b>	
A Generalized Relative $(\alpha, \beta)$ -Entropy: Geometric Properties and Applications to Robust Statistical Inference	
Reprinted from: <i>Entropy</i> 2018, 347, 20, doi:10.3390/e20050347 . . . . .	8
<b>Yuefeng Wu and Giles Hooker</b>	
Asymptotic Properties for Methods Combining the Minimum Hellinger Distance Estimate and the Bayesian Nonparametric Density Estimate	
Reprinted from: <i>Entropy</i> 2018, 20, 955, doi:10.3390/e20120955 . . . . .	40
<b>Elena Castilla, Nirian Martín, Leandro Pardo and Kostantinos Zografos</b>	
Composite Likelihood Methods Based on Minimum Density Power Divergence Estimator	
Reprinted from: <i>Entropy</i> 2018, 20, 18, doi:10.3390/e20010018 . . . . .	60
<b>Michał Broniatowski, Jana Jurečková, M. Ashok Kumar and Emilie Miranda</b>	
Composite Tests under Corrupted Data	
Reprinted from: <i>Entropy</i> 2019, 21, 63, doi:10.3390/e21010063 . . . . .	80
<b>Osamah Abdullah</b>	
Convex Optimization via Symmetrical Hölder Divergence for a WLAN Indoor Positioning System	
Reprinted from: <i>Entropy</i> 2018, 20, 639, doi:10.3390/e20090639 . . . . .	103
<b>Michał Broniatowski, Jana Jurečková and Jan Kalina</b>	
Likelihood Ratio Testing under Measurement Errors	
Reprinted from: <i>Entropy</i> 2018, 20, 966, doi:10.3390/e20120966 . . . . .	117
<b>M. Virtudes Alba-Fernández, M. Dolores Jiménez-Gamero and F. Javier Ariza-López</b>	
Minimum Penalized $\phi$ -Divergence Estimation under Model Misspecification	
Reprinted from: <i>Entropy</i> 2018, 20, 329, doi:10.3390/e20050329 . . . . .	126
<b>Marianthi Markatou and Yang Chen</b>	
Non-Quadratic Distances in Model Assessment	
Reprinted from: <i>Entropy</i> 2018, 20, 464, doi:10.3390/e20060464 . . . . .	141
<b>Maria Kateri</b>	
$\phi$ -Divergence in Contingency Table Analysis	
Reprinted from: <i>Entropy</i> 2018, 20, 324, doi:10.3390/e20050324 . . . . .	161
<b>Takayuki Kawashima and Hironori Fujisawa</b>	
Robust and Sparse Regression via $\gamma$ -Divergence	
Reprinted from: <i>Entropy</i> 2017, 19, 608, doi:10.3390/e19110608 . . . . .	173
<b>Chunming Zhang and Zhengjun Zhang</b>	
Robust-BD Estimation and Inference for General Partially Linear Models	
Reprinted from: <i>Entropy</i> 2017, 19, 625, doi:10.3390/e19110625 . . . . .	194

<b>Aida Toma and Cristinca Fulga</b>	
Robust Estimation for the Single Index Model Using Pseudodistances	
Reprinted from: <i>Entropy</i> <b>2018</b> , <i>20</i> , 374, doi:10.3390/e20050374 . . . . .	<b>223</b>
<b>Lei Li, Anand N. Vidyashankar, Guoqing Diao and Ejaz Ahmed</b>	
Robust Inference after Random Projections via Hellinger Distance for Location-Scale Family	
Reprinted from: <i>Entropy</i> <b>2019</b> , <i>21</i> , 348, doi:10.3390/e21040348 . . . . .	<b>243</b>
<b>Xiao Guo and Chunming Zhang</b>	
Robustness Property of Robust-BD Wald-Type Test for Varying-Dimensional General Linear Models	
Reprinted from: <i>Entropy</i> <b>2018</b> , <i>20</i> , 168, doi:10.3390/e20030168 . . . . .	<b>283</b>
<b>Kei Hirose and Hiroki Masuda</b>	
Robust Relative Error Estimation	
Reprinted from: <i>Entropy</i> <b>2018</b> , <i>20</i> , 632, doi:10.3390/e20090632 . . . . .	<b>311</b>

## About the Special Issue Editor

**Leandro Pardo** has been a Full Professor at the Department of Statistics and Operational Research, Faculty of Mathematics, Complutense University of Madrid, Spain, since 1993. He holds a Ph.D. in Mathematics and has been leading research projects since 1987. He is the author of over 250 research paper in refereed statistical journals, such as the *Journal of Multivariate Analysis*, *Biometrics*, *Bernoulli*, *Journal of Statistical Planning and Inference*, *Entropy*, *IEEE Transactions on Information Theory*, *Biometrical Journal*, *Statistics and Computing*, *Advances in Data Analysis and Classification*, *TEST*, *Psychometrika*, *Sankhya* (The Indian Journal of Statistics), *Annals of the Institute of Statistical Mathematics*, *Statistics & Probability Letters*, *Journal of Nonparametric Statistics*, *Australian and New Zealand Journal of Statistics*, *Statistica Neerlandica*, *Statistica Sinica*, among others. He has published more than 10 academic books and a research book: “*Statistical Inference Based on Divergence measures*” (Chapman & Hall/CRC). He was the Editor in Chief for *TEST* between 2005 and 2008 and an Associate Editor for the *Journal of Multivariate Analysis*, *Journal of Statistical Planning and Inference*, *TEST*, *Communications in Statistics (Theory and Methods)*, *Communications in Statistics (Simulation and Computation)* and *Revista Matemática Complutense*. He has been a scientific visitor to some Universities. In 2004, Professor Pardo was elected as the “Distinguished Eugene Lukacs Professor” at the Booling Green University (Booling, Green, Ohio) [https://en.wikipedia.org/wiki/Lukacs\\_Distinguished\\_Professor](https://en.wikipedia.org/wiki/Lukacs_Distinguished_Professor). He has been the Chair of the Department of Statistics and Operational Research, Faculty of Mathematics, and was the President of the Spanish Society of Statistics and Operations Research (SEIO) between 2013 and 2016.



## Editorial

# New Developments in Statistical Information Theory Based on Entropy and Divergence Measures

**Leandro Pardo**

Department of Statistics and Operation Research, Faculty of Mathematics, Universidad Complutense de Madrid, 28040 Madrid, Spain; lpardo@mat.ucm.es

Received: 28 March 2019; Accepted: 9 April 2019; Published: 11 April 2019



In the last decades the interest in statistical methods based on information measures and particularly in pseudodistances or divergences has grown substantially. Minimization of a suitable pseudodistance or divergence measure gives estimators (minimum pseudodistance estimators or minimum divergence estimators) that have nice robustness properties in relation to the classical maximum likelihood estimators with a not significant loss of efficiency. For more details we refer the monographs of Basu et al. [1] and Pardo [2]. Parametric test statistics based on the minimum divergence estimators have also given interesting results in relation to the robustness in comparison with the classical likelihood ratio test, Wald test statistic and Rao's score statistic. Worthy of special mention are the Wald-type test statistics obtained as an extension of the classical Wald test statistic. These test statistics are based on minimum divergence estimators instead of the maximum likelihood estimators and have been considered in many different statistical problems: Censoring, see Ghosh et al. [3], equality of means in normal and lognormal models, see Basu et al. [4,5], logistic regression models, see Basu et al. [6], polytomous logistic regression models, see Castilla et al. [7], composite likelihood methods, see Martín et al. [8], etc.

This Special Issue focuses on original and new research based on minimum divergence estimators, divergence statistics as well as parametric tests based on pseudodistances or divergences, from a theoretical and applied point of view, in different statistical problems with special emphasis on efficiency and robustness. It comprises 15 selected papers that address novel issues, as well as specific topics illustrating the importance of the divergence measures or pseudodistances in statistics. In the following, the manuscripts are presented in alphabetical order.

The paper, "A Generalized Relative  $(\alpha, \beta)$ -Entropy Geometric properties and Applications to Robust Statistical Inference", by A. Ghosh and A. Basu [9], proposes an alternative information theoretic formulation of the logarithmic super divergence (LSD), Magie et al. [10], as a two parametric generalization of the relative  $\alpha$ -entropy, which they refer as the general  $(\alpha, \beta)$ -entropy. The paper explores its relation with various other entropies and divergences, which also generates a two-parameter extension of Renyi entropy measure as a by-product. The paper is primarily focused on the geometric properties of the relative  $(\alpha, \beta)$ -entropy or the LSD measures: Continuity and convexity in both the arguments along with an extended Pythagorean relation under a power-transformation of the domain space. They also derived a set of sufficient conditions under which the forward and the reverse projections of the relative  $(\alpha, \beta)$ -entropy exist and are unique. Finally, they briefly discuss the potential applications of the relative  $(\alpha, \beta)$ -entropy or the LSD measures in statistical inference, in particular, for robust parameter estimation and hypothesis testing. The results in the reverse projection of the relative  $(\alpha, \beta)$ -entropy establish, for the first time, the existence and uniqueness of the minimum LSD estimators. Numerical illustrations are also provided for the problem of estimating the binomial parameter.

In the work "Asymptotic Properties for methods Combining the Minimum Hellinger Distance Estimate and the Bayesian Nonparametric Density Estimate", Wu, Y. and Hooker, G. [11], pointed out

that in frequentist inference, minimizing the Hellinger distance (Beran et al. [12]) between a kernel density estimate and a parametric family produces estimators that are both robust to outliers and statistically efficient when the parametric family contains the data-generating distribution. In this paper the previous results are extended to the use of nonparametric Bayesian density estimators within disparity methods. They proposed two estimators: One replaces the kernel density estimator with the expected posterior density using a random histogram prior; the other transforms the posterior over densities into a posterior over parameters through minimizing the Hellinger distance for each density. They show that it is possible to adapt the mathematical machinery of efficient influence functions from semiparametric models to demonstrate that both estimators introduced in this paper are efficient in the sense of achieving the Cramér-Rao lower bound. They further demonstrate a Bernstein-von-Mises result for the second estimator, indicating that its posterior is asymptotically Gaussian. In addition, the robustness properties of classical minimum Hellinger distance estimators continue to hold.

In “Composite Likelihood Methods Based on Minimum Density Power Divergence Estimator”, E. Castilla, N. Martin, L. Pardo and K. Zografos [13] pointed out that the classical likelihood function requires exact specification of the probability density function, but in most applications, the true distribution is unknown. In some cases, where the data distribution is available in an analytic form, the likelihood function is still mathematically intractable due to the complexity of the probability density function. There are many alternatives to the classical likelihood function; in this paper, they focus on the composite likelihood. Composite likelihood is an inference function derived by multiplying a collection of component likelihoods; the particular collection used is a conditional determined by the context. Therefore, the composite likelihood reduces the computational complexity, so that it is possible to deal with large datasets and very complex models even when the use of standard likelihood methods is not feasible. Asymptotic normality of the composite maximum likelihood estimator (CMLE) still holds with the Godambe information matrix to replace the expected information in the expression of the asymptotic variance-covariance matrix. This allows the construction of composite likelihood ratio test statistics, Wald-type test statistics, as well as score-type statistics. A review of composite likelihood methods is given in Varin [14]. They mentioned at this point that CMLE, as well as the respective test statistics are seriously affected by the presence of outliers in the set of available data. The main purpose of this paper is to introduce a new robust family of estimators, namely, composite minimum density power divergence estimators (CMDPDE), as well as a new family of Wald-type test statistics based on the CMDPDE in order to get broad classes of robust estimators and test statistics. A simulation study is presented, in order to study the robustness of the CMDPDE, as well as the performance of the Wald-type test statistics based on CMDPDE.

The paper “Composite Tests under Corrupted Data”, by M. Broniatowski, J. Jurecková, A. Kumar Moses and E. Miranda [15] investigate test procedures under corrupted data. They assume that the observations  $Z_i$  are mismeasured, due to the presence of measurement errors. Thus, instead of observing  $Z_i$  for  $i = 1, \dots, n$ , we observe  $X_i = Z_i + \sqrt{\delta}V_i$ , with an unknown parameter  $\delta$  and an unobservable random variable  $V_i$ . It is assumed that the random variables  $Z_i$  are independent and identically distributed, as are the  $X_i$  and the  $V_i$ . The test procedure aims at deciding between two simple hypotheses pertaining to the density of the variable  $Z_i$ , namely  $f_0$  and  $g_0$ . In this setting, the density of the  $V_i$  is supposed to be known. The procedure which they propose aggregates likelihood ratios for a collection of values of  $\delta$ . A new definition of least-favorable hypotheses for the aggregate family of tests is presented, and a relation with the Kullback-Leibler divergence between the sets  $f_\delta(\delta)$  and  $g_\delta(\delta)$  is presented. Finite-sample lower bounds for the power of these tests are presented, both through analytical inequalities and through simulation under the least-favorable hypotheses. Since no optimality holds for the aggregation of likelihood ratio tests, a similar procedure is proposed, replacing the individual likelihood ratio by some divergence based test statistics. It is shown and discussed that the resulting aggregated test may perform better than the aggregate likelihood ratio procedure.

The article “Convex Optimization via Symmetrical Hölder Divergence for a WLAN Indoor Positioning System”, by O. Abdullah [16], uses the Hölder divergence, which generalizes the idea of divergence in information geometry by smooth the non-metric of statistical distances in a way that are not required to follow the law of indiscernibles. The inequality of log-ratio gap pseudo-divergence is built to measure the statistical distance of two classes based on Hölder’s ordinary divergence. By experiment, the WiFi signal suffers from multimodal distribution; nevertheless, the Hölder divergence is considered the proper divergence to measure the dissimilarities between probability densities since the Hölder divergence is a projective divergence that does not need the distribution be normalized and allows the closed form expressions when the expansion family is an affine natural space like multinomial distributions. Hölder divergences encompass both the skew Bhattacharyya divergences and Cauchy-Schwarz divergence, Nielsen et al. [17], and can be symmetrized, and the symmetrized Hölder divergence outperformed the symmetrized Cauchy-Schwarz divergence over the dataset of Gaussians. Both Cauchy-Schwarz divergences are part of a projective divergence distance family with a closed-form expression that does not need to be normalized when considering closed-form expressions with an affine and conic parameter space, such as multivariate or multinomial distributions.

In the paper “Likelihood Ratio Testing under Measurement Errors”, M. Broniatowski, J. Jurecková and J. Kalina [18] consider the likelihood ratio test of a simple null hypothesis (with density  $f_0$ ) against a simple alternative hypothesis (with density  $g_0$ ) in the situation that observations  $X_i$  are mismeasured due to the presence of measurement errors. Thus instead of  $X_i$  for  $i = 1, \dots, n$ , we observe  $Z_i = X_i + \sqrt{\delta}V_i$  with unobservable parameter  $\delta$  and unobservable random variable  $V_i$ . When we ignore the presence of measurement errors and perform the original test, the probability of type I error becomes different from the nominal value, but the test is still the most powerful among all tests on the modified level. Further, they derive the minimax test of some families of misspecified hypotheses and alternatives.

The paper “Minimum Penalized  $\phi$ -Divergence Estimation under Model Misspecification”, by M. V. Alba-Fernández, M. D. Jiménez-Gamero and F. J. Ariza-López [19], focuses on the consequences of assuming a wrong model for multinomial data when using minimum penalized  $\phi$ -divergence, also known as minimum penalized disparity estimators, to estimate the model parameters. These estimators are shown to converge to a well-defined limit. An application of the results obtained shows that a parametric bootstrap consistently estimates the null distribution of a certain class of test statistics for model misspecification detection. An illustrative application to the accuracy assessment of the thematic quality in a global land cover map is included.

In “Non-Quadratic Distances in Model Assessment”, M. Markatou and Y. Chen [20] consider that as a natural way to measure model adequacy is by using statistical distances as loss functions. A related fundamental question is how to construct loss functions that are scientifically and statistically meaningful. In this paper, they investigate non-quadratic distances and their role in assessing the adequacy of a model and/or ability to perform model selection. They first present the definition of a statistical distance and its associated properties. Three popular distances, total variation, the mixture index of fit and the Kullback-Leibler distance, are studied in detail, with the aim of understanding their properties and potential interpretations that can offer insight into their performance as measures of model misspecification. A small simulation study exemplifies the performance of these measures and their application to different scientific fields is briefly discussed.

In “ $\phi$ -Divergence in Contingency Table Analysis”, M. Kateri [21] presents a review about the role of  $\phi$ -divergence measures, see Pardo [2], in modelling association in two-way contingency tables, and illustrated it for the special case of uniform association in ordinal contingency tables. This is targeted at pointing out the potential of this modelling approach and the generated families of models. Throughout this paper a multinomial sampling scheme is assumed. For the models considered here, the other two classical sampling schemes for contingency tables (independent Poisson and product multinomial) are inferentially equivalent. Furthermore, for ease of presentation, we restricted here

to two-way tables. The proposed models extend straightforwardly to multi-way tables. For two or higher-dimensional tables, the subset of models that are linear in their parameters (i.e., multiplicative Row-Column (RC) and RC(M)-type terms are excluded) belong to the family of homogeneous linear predictor models, Goodman [22] and can thus be fitted using the R-package *mph*.

In “Robust and Sparse Regression via  $\gamma$ -Divergence”, T. Kawashima and H. Fujisawa [23] study robust and sparse regression based on the  $\gamma$ -divergence. They showed desirable robust properties under both homogeneous and heterogeneous contamination. In particular, they presented the Pythagorean relation for the regression case, although it was not shown in Kanamori and Fujisawa, [24]. In most of the robust and sparse regression methods, it is difficult to obtain the efficient estimation algorithm, because the objective function is non-convex and non-differentiable. Nonetheless, they succeeded to propose the efficient estimation algorithm, which has a monotone decreasing property of the objective function by using the Majorization–Minimization algorithm (MM-algorithm). The numerical experiments and real data analyses suggested that their method was superior to comparative robust and sparse linear regression methods in terms of both accuracy and computational costs. However, in numerical experiments, a few results of performance measure “true negative rate (TNR)” were a little less than the best results. Therefore, if more sparsity of coefficients is needed, other sparse penalties, e.g., the Smoothly Clipped Absolute Deviations (SCAD), see Fan et al. [25] and the Minimax Concave Penalty (MCP), see Zhang [26], can also be useful.

The manuscript “Robust-Bregman Divergence (BD) Estimation and Inference for General Partially Linear Models”, by C. Zhang and Z. Zhang [27], proposes a class of “robust-Bregman divergence (BD)” estimators of both the parametric and nonparametric components in the general partially linear model (GPLM), which allows the distribution of the response variable to be partially specified, without being fully known. Using the local-polynomial function estimation method, they proposed a computationally-efficient procedure for obtaining “robust-BD” estimators and established the consistency and asymptotic normality of the “robust-BD” estimator of the parametric component  $\beta_0$ . For inference procedures of  $\beta_0$  in the GPLM, they show that the Wald-type test statistic,  $W_n$ , constructed from the “robust-BD” estimators is asymptotically distribution free under the null, whereas the likelihood ratio-type test statistic,  $\Lambda_n$ , is not. This provides an insight into the distinction from the asymptotic equivalence (Fan and Huang, [28]) between  $W_n$  and  $\Lambda_n$  in the partially linear model constructed from profile least-squares estimators using the non-robust quadratic loss. Numerical examples illustrate the computational effectiveness of the proposed “robust-BD” estimators and robust Wald-type test in the appearance of outlying observations.

In “Robust Estimation for the Single Index Model Using Pseudodistances”, A. Toma and C. Fulga [29] consider minimum pseudodistance estimators for the parameters of the single index model (model to reduce the number of parameters in portfolios), see Sharpe [30], and using them they construct new robust optimal portfolios. When outliers or atypical observations are present in the data set, the new portfolio optimization method based on robust minimum pseudodistance estimates yields better results than the classical single index method based on maximum likelihood estimates, in the sense that it leads to larger returns for smaller risks. In literature, there exist various methods for robust estimation in regression models. In the present paper, they proposed the method based on the minimum pseudodistance approach, which suppose to solve a simple optimization problem. In addition, from a theoretical point of view, these estimators have attractive properties, such as being redescending robust, consistent, equivariant and asymptotically normally distributed. The comparison with other known robust estimators of the regression parameters, such as the least median of squares estimators, the S-estimators or the minimum density power divergence estimators, shows that the minimum pseudodistance estimators represent an attractive alternative that may be considered in other applications too. They study properties of the estimators, such as, consistency, asymptotic normality, robustness and equivariance and illustrate the benefits of the proposed portfolio optimization method through examples for real financial data.

The paper “Robust Inference after Random Projections via Hellinger Distance for Location-scale Family”, by L. Li, A. N. Vidyashankar, G. Diao and E. Ahmed [31], proposes Hellinger distance based methods to obtain robust estimates for mean and variance in a location-scale model that takes into account (i) storage issues, (ii) potential model misspecifications, and (iii) presence of aberrant outliers. These issues—which are more likely to occur when dealing with massive amounts of data—if not appropriately accounted in the methodological development, can lead to inaccurate inference and misleading conclusions. On the other hand, incorporating them in the existing methodology may not be feasible due to a computational burden. Our extensive simulations show the usefulness of the methodology and hence can be applied in a variety of scientific settings. Several theoretical and practical questions concerning robustness in a big data setting arise.

The paper “Robustness Property of Robust-BD Wald-Type Test for Varying-Dimensional General Linear Models” by X. Guo and C. Zhang [32], aims to demonstrate the robustness property of the robust-BD Wald-type test in Zhang et al. [33]. Nevertheless, it is a nontrivial task to address this issue. Although the local stability for the Wald-type tests have been established for the M-estimators, see Heritier and Ronchetti, [34], generalized method of moment estimators, Ronchetti and Trojan, [35], minimum density power divergence estimator, Basu et al. [36] and general M-estimators under random censoring, Ghosh et al. [3], their results for finite-dimensional settings are not directly applicable to our situations with a diverging number of parameters. Under certain regularity conditions, we provide rigorous theoretical derivations for robust testing based on the Wald-type test statistics. The essential results are approximations of the asymptotic level and power under contaminated distributions of the data in a small neighborhood of the null and alternative hypotheses, respectively.

The manuscript “Robust Relative Error Estimation” by K. Hirose and H. Masuda [37], presents a relative error estimation procedure that is robust against outliers. The proposed procedure is based on the  $\gamma$ -likelihood function, which is constructed by  $\gamma$ -cross entropy, Fujisawa and Eguchi, [38]. They showed that the proposed method has the redescending property, a desirable property in robust statistics literature. The asymptotic normality of the corresponding estimator together with a simple consistent estimator of the asymptotic covariance matrix are derived, which allows the construction of approximate confidence sets. Besides the theoretical results, they have constructed an efficient algorithm, in which we minimize a convex loss function at each iteration. The proposed algorithm monotonically decreases the objective function at each iteration.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

- Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2011.
- Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2006.
- Ghosh, A.; Basu, A.; Pardo, L. Robust Wald-type tests under random censoring. *ArXiv* **2017**, arXiv:1708.09695.
- Basu, A.; Mandal, A.; Martín, N.; Pardo, L. A Robust Wald-Type Test for Testing the Equality of Two Means from Log-Normal Samples. *Methodol. Comput. Appl. Probab.* **2019**, *21*, 85–107. [[CrossRef](#)]
- Basu, A.; Mandal, A.; Martin, N.; Pardo, L. Robust tests for the equality of two normal means based on the density power divergence. *Metrika* **2015**, *78*, 611–634. [[CrossRef](#)]
- Basu, A.; Ghosh, A.; Mandal, A.; Martín, N.; Pardo, L. A Wald-type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator. *Electron. J. Stat.* **2017**, *11*, 2741–2772. [[CrossRef](#)]
- Castilla, E.; Ghosh, A.; Martín, N.; Pardo, L. New robust statistical procedures for polytomous logistic regression models. *Biometrics* **2019**, in press, doi:10.1111/biom.12890. [[CrossRef](#)]
- Martín, N.; Pardo, L.; Zografos, K. On divergence tests for composite hypotheses under composite likelihood. *Stat. Pap.* **2019**, in press. [[CrossRef](#)]

9. Ghosh, A.; Basu, A. A Generalized Relative  $(\alpha, \beta)$ -Entropy: Geometric Properties and Applications to Robust Statistical Inference. *Entropy* **2018**, *20*, 347. [[CrossRef](#)]
10. Maji, A.; Ghosh, A.; Basu, A. The Logarithmic Super Divergence and Asymptotic Inference Properties. *ASTA Adv. Stat. Anal.* **2016**, *100*, 99–131. [[CrossRef](#)]
11. Wu, Y.; Hooker, G. Asymptotic Properties for Methods Combining the Minimum Hellinger Distance Estimate and the Bayesian Nonparametric Density Estimate. *Entropy* **2018**, *20*, 955. [[CrossRef](#)]
12. Beran, R. Minimum Hellinger Distance Estimates for Parametric Models. *Ann. Stat.* **1977**, *5*, 445–463. [[CrossRef](#)]
13. Castilla, E.; Martín, N.; Pardo, L.; Zografos, K. Composite Likelihood Methods Based on Minimum Density Power Divergence Estimator. *Entropy* **2018**, *20*, 18. [[CrossRef](#)]
14. Varin, C.; Reid, N.; Firth, D. An overview of composite likelihood methods. *Stat. Sin.* **2011**, *21*, 4–42.
15. Broniatowski, M.; Jurečková, J.; Moses, A.K.; Miranda, E. Composite Tests under Corrupted Data. *Entropy* **2019**, *21*, 63. [[CrossRef](#)]
16. Abdulla, O. Convex Optimization via Symmetrical Hölder Divergence for a WLAN Indoor Positioning System. *Entropy* **2018**, *20*, 639. [[CrossRef](#)]
17. Nielsen, F.; Sun, K.; Marchand-Maillet, S. k-Means Clustering with Hölder Divergences. In Proceedings of the International Conference on Geometric Science of Information, Paris, France, 7–9 November 2017.
18. Broniatowski, M.; Jurečková, J.; Kalina, J. Likelihood Ratio Testing under Measurement Errors. *Entropy* **2018**, *20*, 966. [[CrossRef](#)]
19. Alba-Fernández, M.V.; Jiménez-Gamero, M.D.; Ariza-López, F.J. Minimum Penalized  $\phi$ -Divergence Estimation under Model Misspecification. *Entropy* **2018**, *20*, 329. [[CrossRef](#)]
20. Markatou, M.; Chen, Y. Non-Quadratic Distances in Model Assessment. *Entropy* **2018**, *20*, 464. [[CrossRef](#)]
21. Kateri, M.  $\phi$ -Divergence in Contingency Table Analysis. *Entropy* **2018**, *20*, 324. [[CrossRef](#)]
22. Goodman, L.A. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Am. Stat. Assoc.* **1981**, *76*, 320–334.
23. Kawashima, T.; Fujisawa, H. Robust and Sparse Regression via  $\gamma$ -Divergence. *Entropy* **2017**, *19*, 608. [[CrossRef](#)]
24. Kanamori, T.; Fujisawa, H. Robust estimation under heavy contamination using unnormalized models. *Biometrika* **2015**, *102*, 559–572. [[CrossRef](#)]
25. Fan, J.; Li, R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [[CrossRef](#)]
26. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [[CrossRef](#)]
27. Zhang, C.; Zhang, Z. Robust-BD Estimation and Inference for General Partially Linear Models. *Entropy* **2017**, *19*, 625. [[CrossRef](#)]
28. Fan, J.; Huang, T. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **2005**, *11*, 1031–1057. [[CrossRef](#)]
29. Toma, A.; Fulga, C. Robust Estimation for the Single Index Model Using Pseudodistances. *Entropy* **2018**, *20*, 374. [[CrossRef](#)]
30. Sharpe, W.F. A simplified model to portfolio analysis. *Manag. Sci.* **1963**, *9*, 277–293. [[CrossRef](#)]
31. Li, L.; Vidyashankar, A.N.; Diao, G.; Ahmed, E. Robust Inference after Random Projections via Hellinger Distance for Location-scale Family. *Entropy* **2019**, *21*, 348. [[CrossRef](#)]
32. Guo, X.; Zhang, C. Robustness Property of Robust-BD Wald-Type Test for Varying-Dimensional General Linear Models. *Entropy* **2018**, *20*, 168. [[CrossRef](#)]
33. Zhang, C.M.; Guo, X.; Cheng, C.; Zhang, Z.J. Robust-BD estimation and inference for varying-dimensional general linear models. *Stat. Sin.* **2012**, *24*, 653–673. [[CrossRef](#)]
34. Heritier, S.; Ronchetti, E. Robust bounded-influence tests in general parametric models. *J. Am. Stat. Assoc.* **1994**, *89*, 897–904. [[CrossRef](#)]
35. Ronchetti, E.; Trojani, F. Robust inference with GMM estimators. *J. Econom.* **2001**, *101*, 37–69. [[CrossRef](#)]
36. Basu, A.; Ghosh, A.; Martin, N.; Pardo, L. Robust Wald-type tests for non-homogeneous observations based on minimum density power divergence estimator. *Metrika* **2018**, *81*, 493–522. [[CrossRef](#)]

37. Hirose, K.; Masuda, H. Robust Relative Error Estimation. *Entropy* **2018**, *20*, 632. [[CrossRef](#)]
38. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# A Generalized Relative $(\alpha, \beta)$ -Entropy: Geometric Properties and Applications to Robust Statistical Inference

Abhik Ghosh and Ayanendranath Basu \*

Indian Statistical Institute, Kolkata 700108, India; abhianik@gmail.com

\* Correspondence: ayanbasu@isical.ac.in; Tel.: +91-33-2575-2806

Received: 30 March 2018; Accepted: 1 May 2018; Published: 6 May 2018

**Abstract:** Entropy and relative entropy measures play a crucial role in mathematical information theory. The relative entropies are also widely used in statistics under the name of divergence measures which link these two fields of science through the minimum divergence principle. Divergence measures are popular among statisticians as many of the corresponding minimum divergence methods lead to robust inference in the presence of outliers in the observed data; examples include the  $\phi$ -divergence, the density power divergence, the logarithmic density power divergence and the recently developed family of logarithmic super divergence (LSD). In this paper, we will present an alternative information theoretic formulation of the LSD measures as a two-parameter generalization of the relative  $\alpha$ -entropy, which we refer to as the general  $(\alpha, \beta)$ -entropy. We explore its relation with various other entropies and divergences, which also generates a two-parameter extension of Renyi entropy measure as a by-product. This paper is primarily focused on the geometric properties of the relative  $(\alpha, \beta)$ -entropy or the LSD measures; we prove their continuity and convexity in both the arguments along with an extended Pythagorean relation under a power-transformation of the domain space. We also derive a set of sufficient conditions under which the forward and the reverse projections of the relative  $(\alpha, \beta)$ -entropy exist and are unique. Finally, we briefly discuss the potential applications of the relative  $(\alpha, \beta)$ -entropy or the LSD measures in statistical inference, in particular, for robust parameter estimation and hypothesis testing. Our results on the reverse projection of the relative  $(\alpha, \beta)$ -entropy establish, for the first time, the existence and uniqueness of the minimum LSD estimators. Numerical illustrations are also provided for the problem of estimating the binomial parameter.

**Keywords:** relative entropy; logarithmic super divergence; robustness; minimum divergence inference; generalized renyi entropy

## 1. Introduction

Decision making under uncertainty is the backbone of modern information science. The works of C. E. Shannon and the development of his famous entropy measure [1–3] represent the early mathematical foundations of information theory. The Shannon entropy and the corresponding relative entropy, commonly known as the Kullback-Leibler divergence (KLD), has helped to link information theory simultaneously with probability [4–8] and statistics [9–13]. If  $P$  and  $Q$  are two probability measures on a measurable space  $(\Omega, \mathcal{A})$  and have absolutely continuous densities  $p$  and  $q$ , respectively, with respect to a common dominating  $\sigma$ -finite measure  $\mu$ , then the Shannon entropy of  $P$  is defined as

$$\mathcal{E}(P) = - \int p \log(p) d\mu, \quad (1)$$

and the KLD measure between  $P$  and  $Q$  is given by

$$\mathcal{RE}(P, Q) = \int p \log \left( \frac{p}{q} \right) d\mu. \quad (2)$$

In statistics, the minimization of the KLD measure produces the most likely approximation as given by the maximum likelihood principle; the latter, in turn, has a direct equivalence to the (Shannon) entropy maximization criterion in information theory. For example, if  $\Omega$  is finite and  $\mu$  is the counting measure, it is easy to see that  $\mathcal{RE}(P, U) = \log |\Omega| - \mathcal{E}(P)$ , where  $U$  is the uniform measure on  $\Omega$ . Minimization of this relative entropy, or equivalently maximization of the Shannon entropy, with respect to  $P$  within a suitable convex set  $\mathbb{E}$ , generates the most probable distribution for an independent identically distributed finite source having true marginal probability in  $\mathbb{E}$  with non-informative (uniform) prior probability of guessing [14,15]. In general, with a finite source,  $\mathcal{RE}(P, Q)$  denotes the penalty in expected compressed length if the compressor assumes a mismatched probability  $Q$  [16,17]. The corresponding general minimizer of  $\mathcal{RE}(P, Q)$  given  $Q$ , namely its forward projection, and other geometric properties of  $\mathcal{RE}(P, Q)$  are well studied in the literature; see [18–29] among others.

Although the maximum entropy or the minimum divergence criterion based on the classical Shannon entropy  $\mathcal{E}(P)$  and the KLD measure  $\mathcal{RE}(P, Q)$  is still widely used in major (probabilistic) decision making problems in information science and statistics [30–43], there also exist many different useful generalizations of these quantities to address eminent issues in quantum statistical physics, complex codings, statistical robustness and many other topics of interest. For example, if we consider the standardized cumulant of compression length in place of the expected compression length in Shannon's theory, the optimum distribution turns out to be the maximizer of a generalization of the Shannon entropy [44,45] which is given by

$$\mathcal{E}_\alpha(P) = \frac{1}{1-\alpha} \log \left( \int p^\alpha d\mu \right), \quad \alpha > 0, \alpha \neq 1 \quad (3)$$

provided  $p \in L_\alpha(\mu)$ , the complete vector space of functions for which the  $\alpha$ -th power of their absolute values are  $\mu$ -integrable. This general entropy functional is popular by the name Renyi entropy of order  $\alpha$  [46] and covers many important entropy measures like Hartley entropy at  $\alpha \rightarrow 0$  (for finite source), Shannon entropy at  $\alpha \rightarrow 1$ , collision entropy at  $\alpha = 2$  and the min-entropy at  $\alpha \rightarrow \infty$ . The corresponding Renyi divergence measure is given by

$$\mathcal{D}_\alpha(P, Q) = \frac{1}{\alpha-1} \log \left( \int p^\alpha q^{1-\alpha} d\mu \right), \quad \alpha > 0, \alpha \neq 1, \quad (4)$$

whenever  $p, q \in L_\alpha(\mu)$  and coincides with the classical KLD measure at  $\alpha \rightarrow 1$ . The Renyi entropy and the Renyi divergence are widely used in recent complex physical and statistical problems; see, for example, [47–56]. Other non-logarithmic extensions of Shannon entropy include the classical  $f$ -entropies [57], the Tsallis entropy [58] as well as the more recent generalized  $(\alpha, \beta, \gamma)$ -entropy [59,60] among many others; the corresponding divergences and the minimum divergence criteria are widely used in critical information theoretic and statistical problems; see [57,59–70] for details.

We have noted that there is a direct information theoretic connection of KLD to the Shannon entropy under mismatched guessing by minimizing the expected compressed length. However, such a connection does not exist between the Renyi entropy  $\mathcal{E}_\alpha(P)$  and the Renyi divergence  $\mathcal{D}_\alpha(P, Q)$  as recently noted by [17,71]. Herein, it has been shown that, for a finite source with marginal distribution  $P$  and a (prior) mismatched compressor distribution  $Q$ , the penalty in the normalized cumulant of compression length is not  $\mathcal{D}_\alpha(P, Q)$ ; rather it is given by  $\mathcal{D}_{1/\alpha}(P_\alpha, Q_\alpha)$  where  $P_\alpha$  and  $Q_\alpha$  are defined by

$$\frac{dP_\alpha}{d\mu} = p_\alpha = \frac{p^\alpha}{\int p^\alpha d\mu}, \quad \frac{dQ_\alpha}{d\mu} = q_\alpha = \frac{q^\alpha}{\int q^\alpha d\mu}. \quad (5)$$

The new quantity  $\mathcal{D}_{1/\alpha}(P_\alpha, Q_\alpha)$  also gives a measure of discrimination (i.e., is a divergence) between the probability distributions  $P$  and  $Q$  and coincides with the KLD at  $\alpha \rightarrow 1$ . This functional is referred to as the relative  $\alpha$ -entropy in the terminology of [72] and has the simpler form

$$\begin{aligned}\mathcal{RE}_\alpha(P, Q) &:= \mathcal{D}_{1/\alpha}(P_\alpha, Q_\alpha) \\ &= \frac{\alpha}{1-\alpha} \log \int pq^{\alpha-1} d\mu - \frac{1}{1-\alpha} \log \int p^\alpha d\mu + \log \int q^\alpha d\mu, \quad \alpha > 0, \alpha \neq 1.\end{aligned}\tag{6}$$

The geometric properties of this relative  $\alpha$ -entropy along with its forward and reverse projections have been studied recently [16,73]; see Section 2.1 for some details. This quantity had, however, already been proposed earlier as a statistical divergence, although for  $\alpha \geq 1$  only, by [74] while developing a robust estimation procedure following the generalized method-of-moments approach of [75]. Later authors referred to the divergence proposed in [74] as the logarithmic density power divergence (LDPD) measure. The advantages of the minimum LDPD estimator in terms of robustness against outliers in data have been studied by, among other, [66,74]. Fujisawa [76], Fujisawa and Eguchi [77] have also used the same divergence measure with  $\gamma = (\alpha - 1) \geq 0$  in different statistical problems and have referred to it as the  $\gamma$ -divergence. Note that, the formulation in (6) extends the definition of the divergence over the  $0 < \alpha < 1$  region as well.

Motivated by the substantial advantages of the minimum LDPD inference in terms of statistical robustness against outlying observations, Maji et al. [78,79] have recently developed a two-parameter generalization of the LDPD family, namely the logarithmic super divergence (LSD) family, given by

$$\begin{aligned}\mathcal{LSD}_{\tau,\gamma}(P, Q) &= \frac{1}{B} \log \int p^{1+\tau} d\mu - \frac{1+\tau}{AB} \log \int p^A q^B d\mu + \frac{1}{A} \log \int q^{1+\tau} d\mu, \\ &\text{with } A = 1 + \gamma(1 - \tau), B = 1 + \tau - A, \quad \tau \geq 0, \gamma \in \mathbb{R}.\end{aligned}\tag{7}$$

This rich superfamily of divergences contain many important divergence measures including the LDPD at  $\gamma = 0$  and the Kullback-Leibler divergence at  $\tau = \gamma \rightarrow 0$ ; this family also contains a transformation of Renyi divergence at  $\tau = 0$  which has been referred to as the logarithmic power-divergence family by [80]. As shown in [78,79], the statistical inference based on some of the new members of this LSD family, outside the existing ones including the LDPD, provide much better trade-off between the robustness and efficiency of the corresponding minimum divergence estimators.

The statistical benefits of the LSD family over the LDPD family raise a natural question: is it possible to translate this robustness advantage of the LSD family of divergences to the information theoretic context, through the development of a corresponding generalization of the relative  $\alpha$ -entropy in (6)? In this paper, we partly answer this question by defining an independent information theoretic generalization of the relative  $\alpha$ -entropy measure coinciding with the LSD measure. We will refer to this new generalized relative entropy measure as the “*Relative  $(\alpha, \beta)$ -entropy*” and study its properties for different values of  $\alpha > 0$  and  $\beta \in \mathbb{R}$ . In particular, this new formulation will extend the scope of the LSD measure for  $-1 < \tau < 0$  as well and generate several interesting new divergence and entropy measures. We also study the geometric properties of all members of the relative  $(\alpha, \beta)$ -entropy family, or equivalently the LSD measures, including their continuity in both the arguments and a Pythagorean-type relation. The related forward projection problem, i.e., the minimization of the relative  $(\alpha, \beta)$ -entropy in its first argument, is also studied extensively.

In summary, the main objective of the present paper is to study the geometric properties of the LSD measure through the new information theoretic or entropic formulation (or the relative  $(\alpha, \beta)$ -entropy). Our results indeed generalize the properties of the relative  $\alpha$ -entropy from [16,73]. The specific and significant contributions of the paper can be summarized as follows.

1. We present a two parameter extension of the relative  $\alpha$ -entropy measure in (6) motivated by the logarithmic  $S$ -divergence measures. These divergence measures are known to generate more robust statistical inference compared to the LDPD measures related to the relative  $\alpha$ -entropy.

2. In the new formulation of the relative  $(\alpha, \beta)$ -entropy, the LSD measures are linked with several important information theoretic divergences and entropy measures like the ones named after Renyi. A new divergence family is discovered corresponding to  $\alpha \rightarrow 0$  case (properly standardized) for the finite measure cases.
3. As a by-product of our new formulation, we get a new two-parameter generalization of the Renyi entropy measure, which we refer to as the Generalized Renyi entropy (GRE). This opens up a new area of research to examine the detailed properties of GRE and its use in complex problems in statistical physics and information theory. In this paper, we show that this new GRE satisfies the basic entropic characteristics, i.e., it is zero when the argument probability is degenerate and is maximum when the probability is uniform.
4. Here we provide a detailed geometric analysis of the robust LSD measure, or equivalently the relative  $(\alpha, \beta)$ -entropy in our new formulation. In particular, we show their continuity or lower semi-continuity with respect to the first argument depending on the values of the tuning parameters  $\alpha$  and  $\beta$ . Also, its lower semi-continuity with respect to the second argument is proved.
5. We also study the convexity of the LSD measures (or the relative  $(\alpha, \beta)$ -entropies) with respect to its argument densities. The relative  $\alpha$ -entropy (i.e., the relative  $(\alpha, \beta)$ -entropy at  $\beta = 1$ ) is known to be quasi-convex [16] only in its first argument. Here, we will show that, for general  $\alpha > 0$  and  $\beta \neq 1$ , the relative  $(\alpha, \beta)$ -entropies are not quasi-convex on the space of densities, but they are always quasi-convex with respect to both the arguments on a suitably (power) transformed space of densities. Such convexity results in the second argument were unavailable in the literature even for the relative  $\alpha$ -entropy, which we will introduce in this paper through a transformation of space.
6. Like the relative  $\alpha$ -entropy, but unlike the relative entropy in (2), our new relative  $(\alpha, \beta)$ -entropy also does not satisfy the data processing inequalities. However, we prove an extended Pythagorean relation for the relative  $(\alpha, \beta)$ -entropy which makes it reasonable to treat them as “squared distances” and talk about their projections.
7. The forward projection of a relative entropy or a suitable divergence, i.e., their minimization with respect to the first argument, is very important for both statistical physics and information theory. This is indeed equivalent to the maximum entropy principle and is also related to the Gibbs conditioning principle. In this paper, we will examine the conditions under which such a forward projection of the relative  $(\alpha, \beta)$ -entropy (or, LSD) exists and is unique.
8. Finally, for completeness, we briefly present the application of the LSD measure or the relative  $(\alpha, \beta)$ -entropy measure in robust statistical inference in the spirit of [78,79] but now with extended range of tuning parameters. It uses the reverse projection principle; a result on the existence of the minimum LSD functional is first presented with the new formulation of this paper. Numerical illustrations are provided for the binomial model, where we additionally study their properties for the extended tuning parameter range  $\alpha \in (0, 1)$  as well as for some new divergence families (related to  $\alpha = 0$ ). Brief indications of the potential use of these divergences in testing of statistical hypotheses are also provided.

Although we are primarily discussing the logarithmic entropies like the Renyi entropy and its generalizations in this paper, it is important to point out that non-logarithmic entropies including the f-entropy and the Tsallis entropy are also very useful in several applications with real systems. Recently, several complex physical and social systems have been observed to follow the theory developed from such non-logarithmic, non-additive entropies instead of the classical additive Shannon entropy. In particular, the Tsallis entropy has led to the development of the nonextensive statistical mechanics [61,64] to solve several critical issues in modern physics. Important areas of application include, but certainly are not limited to, the motion of cold atoms in dissipative optical lattices [81,82], the magnetic field fluctuations in the solar wind and related q-triplet [83], the distribution of velocity

in driven dissipative dusty plasma [84], spin glass relaxation [85], the interaction of trapped ion with a classical buffer gas [86], different high energy collisional experiments [87–89], derivation of the black hole entropy [90], along with water engineering [63], text mining [65] and many others. Therefore, it is also important to investigate the possible generalizations and manipulations of such non-logarithmic entropies both from mathematical and application point of view. However, as our primary interest here is in logarithmic entropies, we have, to keep the focus clear, otherwise avoided the description and development of non-logarithmic entropies in this paper.

Although there are many applications of extended and general non-additive entropy and divergence measures, there are also some criticisms of these non-additive measures that should be kept in mind. It is of course possible to employ such quantities simply as new descriptors of the complexity of systems, but at the same time, it is known that the minimization of a generalized divergence (or maximization of the corresponding entropy) under constraints in order to determine an optimal probability assignment leads to inconsistencies for information measures other than the Kullback-Leibler divergence. See, for instance [91–96], among others. So, one needs to be very careful in discriminating the application of the newly introduced entropies and divergence measures for the purposes of inference under given information, from the ones where it is used as a measure of complexity. In this respect, we would like to emphasize that, the main advantage of our two-parameter extended family of LSD or relative  $(\alpha, \beta)$ -entropy measures in parametric statistical inference is in their strong robustness property against possible contamination (generally manifested through outliers) in the sample data. The classical additive Shannon entropy and Kullback-Leibler divergence produce non-robust inference even under a small proportion of data contamination, but the extremely high robustness of the LSD has been investigated in detail, with both theoretical and empirical justifications, by [78,79]; in this respect, we will present some numerical illustrations in Section 5.2. Another important issue could be to decide whether to stop at the two-parameter level for information measures or to extend it to three-parameters, four-parameters, etc. It is not an easy question to answer. However, we have seen that many members of the two-parameter family of LSD measures generate highly robust inference along with a desirable trade-off between efficiency under pure data and robustness under contaminated data. Therefore a two-parameter system appears to work well in practice. Since it is a known principle that one “should not multiply entities beyond necessity”, we will, for the sake of parsimony, restrict ourselves to the second level of generalization for robust statistical inference, at least until there is further convincing evidence that the next higher level of generalization can produce a significant improvement.

## 2. The Relative $(\alpha, \beta)$ -Entropy Measure

### 2.1. Definition: An Extension of the Relative $\alpha$ -Entropy

In order to motivate the development of our generalized relative  $(\alpha, \beta)$ -entropy measure, let us first briefly describe an alternative formulation of the relative  $\alpha$ -entropy following [16]. Consider the mathematical set-up of Section 1 with  $\alpha > 0$  and assume that the space  $L_\alpha(\mu)$  is equipped with the norm

$$\|f\|_\alpha = \begin{cases} (\int |f|^\alpha d\mu)^{1/\alpha} & \text{if } \alpha \geq 1, f \in L_\alpha(\mu), \\ \int |f|^\alpha d\mu & \text{if } 0 < \alpha < 1, f \in L_\alpha(\mu), \end{cases} \quad (8)$$

and the corresponding metric  $d_\alpha(g, f) = \|g - f\|_\alpha$  for  $g, f \in L_\alpha(\mu)$ . Then, the relative  $\alpha$ -entropy between two distributions  $P$  and  $Q$  is obtained as a function of the Cressie-Read power divergence measure [97], defined below in (11), between the escort measures  $P_\alpha$  and  $Q_\alpha$  defined in (5). Note that the disparity family or the  $\phi$ -divergence family [18,98–103] between  $P$  and  $Q$  is defined as

$$D_\phi(P, Q) = \int q\phi\left(\frac{p}{q}\right)d\mu, \quad (9)$$

for a continuous convex function  $\phi$  on  $[0, \infty)$  satisfying  $\phi(0) = 0$  and with the usual convention  $0\phi(0/0) = 0$ . We consider the  $\phi$ -function given by

$$\phi(u) = \phi_\lambda(u) = \text{sign}(\lambda(\lambda + 1)) \left( u^{\lambda+1} - 1 \right), \quad \lambda \in \mathbb{R}, u \geq 0, \quad (10)$$

with the convention that, for any  $u > 0$ ,  $0\phi_\lambda(u/0) = 0$  if  $\lambda < 0$  and  $0\phi_\lambda(u/0) = \infty$  if  $\lambda > 0$ . The corresponding  $\phi$ -divergence has the form

$$D_\lambda(P, Q) = D_{\phi_\lambda}(P, Q) = \text{sign}(\lambda(\lambda + 1)) \int q \left[ \left( \frac{p}{q} \right)^{\lambda+1} - 1 \right] d\mu, \quad (11)$$

which is just a positive multiple of the Cressie-Read power divergence with the multiplicative constant being  $|\lambda(1 + \lambda)|$ ; when this constant is present, the case  $\lambda = 0$  leads to the KLD measure in a limiting sense. Note that, our  $\phi$ -function in (10) differs slightly from the one used by [16] in that we use  $\text{sign}(\lambda(\lambda + 1))$  in place of  $\text{sign}(\lambda)$  there; this is to make the divergence in (11) non-negative for all  $\lambda \in \mathbb{R}$  ([16] considered only  $\lambda > -1$ ) which will be needed to define our generalized relative entropy. Then, given an  $\alpha > 0$ , [16,17] set  $\lambda = \alpha^{-1} - 1 (> -1)$  and show that the relative  $\alpha$ -entropy of  $P$  with respect to  $Q$  can be obtained as

$$\mathcal{RE}_\alpha(P, Q) = \mathcal{RE}_\alpha^\mu(P, Q) = \frac{1}{\lambda} \log [\text{sign}(\lambda) D_\lambda(P_\alpha, Q_\alpha) + 1]. \quad (12)$$

It is straightforward to see that the above formulation (12) coincides with the definition given in (6). We often suppress the superscript  $\mu$  whenever the underlying measure is clear from the context; in most applications in information theory and statistics it is either counting measure or the Lebesgue measure depending on whether the distribution is discrete or continuous.

We can now change the tuning parameters in the formulation given by (12) suitably as to arrive at the more general form of the LSD family in (7). For this purpose, let us fix  $\alpha > 0$ ,  $\beta \in \mathbb{R}$  and assume that  $p, q \in L_\alpha(\mu)$  are the  $\mu$ -densities of  $P$  and  $Q$ , respectively. Instead of considering the re-parametrization  $\lambda = \alpha^{-1} - 1$  as above, we now consider the two-parameter re-parametrization  $\lambda = \beta\alpha^{-1} - 1 \in \mathbb{R}$ . Note that, the feasible range of  $\lambda$ , in order to make  $\alpha > 0$ , now clearly depends on  $\beta$  through  $\alpha = \frac{\beta}{1+\lambda} > 0$ ; whenever  $\beta > 0$  we have  $-1 < \lambda < \infty$  and if  $\beta < 0$  we need  $-\infty < \lambda < -1$ . We have already taken care of this dependence through the modified  $\phi$  function defined in (10) which ensures that  $D_\lambda(\cdot, \cdot)$  is non-negative for all  $\lambda \in \mathbb{R}$ . So we can again use the relation as in (12), after suitable standardization due to the additional parameter  $\beta$ , to define a new generalized relative entropy measure as given in the following definition.

**Definition 1** (Relative  $(\alpha, \beta)$ -entropy). *Given any  $\alpha > 0$  and  $\beta \in \mathbb{R}$ , put  $\lambda = \frac{\beta}{\alpha} - 1$  (i.e.,  $\alpha = \frac{\beta}{1+\lambda}$ ). Then, the relative  $(\alpha, \beta)$ -entropy of  $P$  with respect to  $Q$  is defined as*

$$\mathcal{RE}_{\alpha,\beta}(P, Q) = \mathcal{RE}_{\alpha,\beta}^\mu(P, Q) = \frac{1}{\beta\lambda} \log [\text{sign}(\beta\lambda) D_\lambda(P_\alpha, Q_\alpha) + 1]. \quad (13)$$

The cases  $\beta = 0$  and  $\lambda = 0$  (i.e.,  $\beta = \alpha$ ) are defined in limiting sense; see Equations (15) and (16) below.

A straightforward simplification gives a simpler form of this new relative  $(\alpha, \beta)$ -entropy which coincides with the LSD measure as follows.

$$\begin{aligned} \mathcal{RE}_{\alpha,\beta}(P, Q) &= \frac{1}{\alpha - \beta} \log \int p^\alpha d\mu - \frac{\alpha}{\beta(\alpha - \beta)} \log \int p^\beta q^{\alpha-\beta} d\mu + \frac{1}{\beta} \log \int q^\alpha d\mu, \\ &= \mathcal{LSD}_{\alpha-1, \frac{\beta-1}{2-\alpha}}(P, Q). \end{aligned} \quad (14)$$

Note that, it coincides with the relative  $\alpha$ -entropy  $\mathcal{RE}_\alpha(P, Q)$  at the choice  $\beta = 1$ . For the limiting cases, it leads to the forms

$$\mathcal{RE}_{\alpha,0}(P, Q) = \frac{\int \log(q/p)q^\alpha d\mu}{\int q^\alpha d\mu} + \frac{1}{\alpha} \log \left( \frac{\int p^\alpha d\mu}{\int q^\alpha d\mu} \right), \quad (15)$$

$$\mathcal{RE}_{\alpha,\infty}(P, Q) = \frac{\int \log(p/q)p^\alpha d\mu}{\int p^\alpha d\mu} + \frac{1}{\alpha} \log \left( \frac{\int q^\alpha d\mu}{\int p^\alpha d\mu} \right). \quad (16)$$

By the divergence property of  $D_\lambda(\cdot, \cdot)$ , all the relative  $(\alpha, \beta)$ -entropies are non-negative and valid statistical divergences. Note that, in view of (14), the formulation (13) extends the scope of LSD measure, defined in (7), for  $\tau \in (-1, 0)$ .

**Proposition 1.** For any  $\alpha > 0$  and  $\beta \in \mathbb{R}$ ,  $\mathcal{RE}_{\alpha,\beta}(P, Q) \geq 0$  for all probability measures  $P$  and  $Q$ , whenever it is defined. Further,  $\mathcal{RE}_{\alpha,\beta}(P, Q) = 0$  if and only in  $P = Q[\mu]$ .

Also, it is important to identify the cases where the relative  $(\alpha, \beta)$ -entropy is not finitely defined, which can be obtained from the definition and convention related to  $D_\lambda$  divergence; these are summarized in the following proposition.

**Proposition 2.** For any  $\alpha > 0$ ,  $\beta \in \mathbb{R}$  and distributions  $P, Q$  having  $\mu$ -densities in  $L_\alpha(\mu)$ , the relative  $(\alpha, \beta)$ -entropy  $\mathcal{RE}_{\alpha,\beta}(P, Q)$  is a finite positive number except for the following three cases:

1.  $P$  is not absolutely continuous with respect to  $Q$  and  $\alpha < \beta$ , in which case  $\mathcal{RE}_{\alpha,\beta}(P, Q) = +\infty$ .
2.  $P$  is mutually singular to  $Q$  and  $\alpha > \beta$ , in which case also  $\mathcal{RE}_{\alpha,\beta}(P, Q) = +\infty$ .
3.  $0 < \beta < \alpha$  and  $D_\lambda(P_\alpha, Q_\alpha) \geq 1$ , in which case also  $\mathcal{RE}_{\alpha,\beta}(P, Q)$  is undefined.

The above two propositions completely characterize the values and existence of our new relative  $(\alpha, \beta)$ -entropy measure. In the next subsection, we will now explore its relation with other existing entropies and divergence measures; along the way we will get some new ones as by-products of our generalized relative entropy formulation.

## 2.2. Relations with Different Existing or New Entropies and Divergences

The relative  $(\alpha, \beta)$ -entropy measures form a large family containing several existing relative entropies and divergences. Its relation with some popular ones are summarized in the following proposition; the proof is straightforward from definitions and hence omitted.

**Proposition 3.** For  $\alpha > 0$ ,  $\beta \in \mathbb{R}$  and distributions  $P, Q$ , the following results hold (whenever the relevant integrals and divergences are defined finitely, even in limiting sense).

1.  $\mathcal{RE}_{1,1}(P, Q) = \mathcal{RE}(P, Q)$ , the KLD measure.
2.  $\mathcal{RE}_{\alpha,1}(P, Q) = \mathcal{RE}_\alpha(P, Q)$ , the relative  $\alpha$ -entropy.
3.  $\mathcal{RE}_{1,\beta}(P, Q) = \frac{1}{\beta} \mathcal{D}_\beta(P, Q)$ , a scaled Renyi divergence, which also coincides with the logarithmic power divergence measure of [80].
4.  $\mathcal{RE}_{\alpha,\beta}(P, Q) = \frac{1}{\beta} \mathcal{D}_{\beta/\alpha}(P_\alpha, Q_\alpha)$ , where  $P_\alpha$  and  $Q_\alpha$  are as defined in (5).

**Remark 1.** Note that, items 3 and 4 in Proposition 3 indicate a possible extension of the Renyi divergence measure over negative values of the tuning parameter  $\beta$  as follows:

$$\mathcal{D}_\beta^*(P, Q) = \frac{1}{\beta} \mathcal{D}_\beta(P, Q), \quad \beta \in \mathbb{R} \setminus \{0\}, \quad \mathcal{D}_0^*(P, Q) = \int q \log \left( \frac{q}{p} \right) d\mu.$$

Note that this modified Renyi divergence also coincides with the KLD measure at  $\beta = 1$ . Statistical applications of this divergence family have been studied by [80].

However, not all the members of the family of relative  $(\alpha, \beta)$ -entropies are distinct or symmetric. For example,  $\mathcal{RE}_{\alpha,0}(P, Q) = \mathcal{RE}_{\alpha,\alpha}(Q, P)$  for any  $\alpha > 0$ . The following proposition characterizes all such identities.

**Proposition 4.** For  $\alpha > 0$ ,  $\beta \in \mathbb{R}$  and distributions  $P$ ,  $Q$ , the relative  $(\alpha, \beta)$ -entropy  $\mathcal{RE}_{\alpha,\beta}(P, Q)$  is symmetric if and only if  $\beta = \frac{\alpha}{2}$ . In general, we have  $\mathcal{RE}_{\alpha,\frac{\alpha}{2}-\gamma}(P, Q) = \mathcal{RE}_{\alpha,\frac{\alpha}{2}+\gamma}(Q, P)$  for any  $\alpha > 0$ ,  $\gamma \in \mathbb{R}$ .

Recall that the KLD measure is linked to the Shannon entropy and the relative  $\alpha$ -entropy is linked with the Renyi entropy when the prior mismatched probability is uniform over the finite space. To derive such a relation for our general relative  $(\alpha, \beta)$ -entropy, let us assume  $\mu(\Omega) < \infty$  and let  $U$  denote the uniform probability measure on  $\Omega$ . Then, we get

$$\mathcal{RE}_{\alpha,\beta}(P, U) = \frac{1}{\beta} [\log \mu(\Omega) - \mathcal{E}_{\alpha,\beta}(P)], \quad \beta \neq 0 \quad (17)$$

where the functional  $\mathcal{E}_{\alpha,\beta}(P)$  is given in Definition 2 below and coincides with the Renyi entropy at  $\beta = 1$ . Thus, it can be used to define a two-parameter generalization of the Renyi entropy as follows.

**Definition 2** (Generalized Renyi Entropy). For any probability measure  $P$  over a measurable space  $\Omega$ , we define the generalized Renyi entropy (GRE) of order  $(\alpha, \beta)$  as

$$\mathcal{E}_{\alpha,\beta}(P) = \frac{1}{\beta - \alpha} \log \left[ \left( \int p^\alpha d\mu \right)^\beta \right], \quad \alpha > 0, \beta \in \mathbb{R}, \beta \neq 0, \alpha; \quad (18)$$

$$\mathcal{E}_{\alpha,\alpha}(P) = -\frac{\int \log(p) p^\alpha d\mu}{\int p^\alpha d\mu} + \frac{1}{\alpha} \log \left( \int p^\alpha d\mu \right), \quad \alpha > 0. \quad (19)$$

Note that, at  $\beta = 1$ , we have  $\mathcal{E}_{\alpha,1}(P) = \mathcal{E}_\alpha(P)$ , the usual Renyi entropy measure of order  $\alpha$ .

The GRE is a new entropy to the best of our knowledge, and does not belong to the general class of entropy functionals as given in [104] which covers many existing entropies (including most, if not all, classical entropies). The following property of the functional  $\mathcal{E}_{\alpha,\beta}(P)$  is easy to verify and justifies its use as a new entropy functional. To keep the focus of the present paper clear on the relative  $(\alpha, \beta)$ -entropy, further properties of the GRE will be explored in our future work.

**Theorem 1** (Entropic characteristics of GRE). For any probability measure  $P$  over a finite measure space  $\Omega$ , we have  $0 \leq \mathcal{E}_{\alpha,\beta}(P) \leq \log \mu(\Omega)$  for all  $\alpha > 0$  and  $\beta \in \mathbb{R} \setminus \{0\}$ . The two extremes are attained as follows.

1.  $\mathcal{E}_{\alpha,\beta}(P) = 0$  if  $P$  is degenerate at a point in  $\Omega$  (no uncertainty).
2.  $\mathcal{E}_{\alpha,\beta}(P) = \log \mu(\Omega)$  if  $P$  is uniform over  $\Omega$  (maximum uncertainty).

**Example 1** (Normal Distribution). Consider distributions  $P_i$  from the most common class of multivariate ( $s$ -dimensional) normal distributions having mean  $\mu_i \in \mathbb{R}^s$  and variance matrix  $\Sigma_i$  for  $i = 1, 2$ . It is known that the Shannon and the Renyi entropies of  $P_1$  are, respectively, given by

$$\begin{aligned} \mathcal{E}(P_1) &= \frac{s}{2} + \frac{s}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_1|, \\ \mathcal{E}_\alpha(P_1) &= \frac{s \log \alpha}{2\alpha - 1} + \frac{s}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_1|, \quad \alpha > 0, \alpha \neq 1. \end{aligned}$$

With the new entropy measure, GRE, the entropy of the normal distribution  $P_1$  can be seen to have the form

$$\begin{aligned}\mathcal{E}_{\alpha,\beta}(P_1) &= \frac{s}{2} \frac{(\alpha \log \beta - \beta \log \alpha)}{(\beta - \alpha)} + \frac{s}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_1|, \quad \alpha > 0, \beta \in \mathbb{R} \setminus \{0, \alpha\}, \\ \mathcal{E}_{\alpha,\alpha}(P_1) &= \frac{s}{2}(1 - \log \alpha) + \frac{s}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_1|, \quad \alpha > 0.\end{aligned}$$

Interestingly, the GRE of a normal distribution is effectively the same as its Shannon entropy or Renyi entropy up to an additive constant. However, similar characteristic does not hold between the relative entropy (KLD) and relative  $(\alpha, \beta)$ -entropy. The KLD measure between two normal distributions  $P_1$  and  $P_2$  is given by

$$\mathcal{RE}(P_1, P_2) = \frac{1}{2} \text{Trace}(\Sigma_2^{-1} \Sigma_1) + \frac{1}{2} (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \log \left( \frac{|\Sigma_2|}{|\Sigma_1|} \right) - \frac{s}{2},$$

whereas the general relative  $(\alpha, \beta)$ -entropy, with  $\alpha > 0$  and  $\beta \in \mathbb{R} \setminus \{0, \alpha\}$ , has the form

$$\begin{aligned}\mathcal{RE}_{\alpha,\beta}(P_1, P_2) &= \frac{\alpha}{2} (\mu_2 - \mu_1)^T [\beta \Sigma_2 + (\alpha - \beta) \Sigma_1]^{-1} (\mu_2 - \mu_1) \\ &\quad + \frac{1}{2\beta(\beta - \alpha)} \log \left( \frac{|\Sigma_2|^{\beta} |\Sigma_1|^{\alpha - \beta}}{|\beta \Sigma_2 + (\alpha - \beta) \Sigma_1|^{\alpha}} \right) - \frac{s\alpha \log \alpha}{2\beta(\alpha - \beta)}.\end{aligned}$$

Note that the relative  $(\alpha, \beta)$ -entropy gives a more general divergence measure which utilizes different weights for the variance (or precision) matrix of the two normal distributions.

**Example 2** (Exponential Distribution). Consider the exponential distribution  $P$  having density  $p_\theta(x) = \theta e^{-\theta x} I(x \geq 0)$  with  $\theta > 0$ . This distribution is very useful in lifetime modeling and reliability engineering; it is also the maximum entropy distribution of a non-negative random variable with fixed mean. The Shannon and the Renyi entropies of  $P$  are, respectively, given by

$$\mathcal{E}(P) = 1 - \log \theta, \quad \text{and} \quad \mathcal{E}_\alpha(P) = \frac{\log \alpha}{\alpha - 1} - \log \theta, \quad \alpha > 0, \alpha \neq 1.$$

A simple calculation leads to the following form of the our new GRE measure of the exponential distribution  $P$ .

$$\begin{aligned}\mathcal{E}_{\alpha,\beta}(P) &= \frac{(\alpha \log \beta - \beta \log \alpha)}{(\beta - \alpha)} - \log \theta, \quad \alpha > 0, \beta \in \mathbb{R} \setminus \{0, \alpha\}, \\ \mathcal{E}_{\alpha,\alpha}(P) &= (1 - \log \alpha) - \log \theta, \quad \alpha > 0.\end{aligned}$$

Once again, the new GRE is effectively the same as the Shannon entropy or the Renyi entropy, up to an additive constant, for the exponential distribution as well.

Further, if  $P_1$  and  $P_2$  are two exponential distributions with parameters  $\theta_1$  and  $\theta_2$ , respectively, the relative entropy (KLD) and the relative  $(\alpha, \beta)$ -entropy between them are given by

$$\begin{aligned}\mathcal{RE}(P_1, P_2) &= \frac{\theta_2}{\theta_1} + \log \theta_1 - \log \theta_2 - 1, \\ \mathcal{RE}_{\alpha,\beta}(P_1, P_2) &= \frac{\alpha}{\beta(\alpha - \beta)} \log [\beta \theta_1 + (\alpha - \beta) \theta_2] - \frac{1}{\alpha - \beta} \log \theta_1 - \frac{1}{\beta} \log \theta_2 - \frac{\alpha \log \alpha}{\beta(\alpha - \beta)},\end{aligned}$$

for  $\alpha > 0$  and  $\beta \in \mathbb{R} \setminus \{0, \alpha\}$ . Clearly, the contributions of both the distribution is weighted differently by  $\beta$  and  $(\alpha - \beta)$  in their relative  $(\alpha, \beta)$ -entropy measure.

Before concluding this section, we study the nature of our relative  $(\alpha, \beta)$ -entropy as  $\alpha \rightarrow 0$ . For this purpose, we restrict ourselves to the case of finite measure spaces with  $\mu(\Omega) < \infty$ . It is again straightforward to note that  $\lim_{\alpha \rightarrow 0} \mathcal{RE}_{\alpha,\beta}(P, Q) = 0$  for any  $\beta \in \mathbb{R}$  and any distributions  $P$  and  $Q$  on  $\Omega$ .

However, if we take the limit after scaling the relative entropy measure by  $\alpha$  we get a non-degenerate divergence measure as follows.

$$\mathcal{RE}_\beta^*(P, Q) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha} \mathcal{RE}_{\alpha,\beta}(P, Q) = \frac{1}{\beta^2} \left[ \log \int \left( \frac{p}{q} \right)^\beta d\mu - \frac{\beta}{\mu(\Omega)} \int \log \left( \frac{p}{q} \right) d\mu - \log \mu(\Omega) \right],$$

for  $\beta \in \mathbb{R} \setminus \{0\}$ , and

$$\mathcal{RE}_0^*(P, Q) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha} \mathcal{RE}_{\alpha,0}(P, Q) = \frac{1}{2\mu(\Omega)} \left[ \int \{\log(p/q)\}^2 d\mu - \frac{1}{\mu(\Omega)} \left\{ \int \log(p/q) d\mu \right\}^2 \right].$$

These interesting relative entropy measures again define a subfamily of valid statistical divergences, from its construction. The particular member at  $\beta = 1$  is linked to the LDPD (or the  $\gamma$ -divergence) with tuning parameter  $-1$  and can be thought of as a logarithmic extension of the famous Itakura–Saito divergence [105] given by

$$D_{IS}(P, Q) = \int \left( \frac{p}{q} \right) d\mu - \int \log \left( \frac{p}{q} \right) d\mu - \mu(\Omega). \quad (20)$$

This Itakura–Saito-divergence has been successfully applied to non-negative matrix factorization in different applications [106] which can be extended by using the new divergence family  $\mathcal{RE}_\beta^*(P, Q)$  in future works.

### 3. Geometry of the Relative $(\alpha, \beta)$ -Entropy

#### 3.1. Continuity

We start the exploration of the geometric properties of the relative  $(\alpha, \beta)$ -entropy with its continuity over the functional space  $L_\alpha(\mu)$ . In the following, we interchangeably use the notation  $\mathcal{RE}_{\alpha,\beta}(p, q)$  and  $D_\lambda(p, q)$  to denote  $\mathcal{RE}_{\alpha,\beta}(P, Q)$  and  $D_\lambda(P, Q)$ , respectively. Our results generalize the corresponding properties of the relative  $\alpha$ -entropy from [16,73] to our relative  $(\alpha, \beta)$ -entropy or equivalent LSD measure.

**Proposition 5.** For a given  $q \in L_\alpha(\mu)$ , consider the function  $p \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$  from  $p \in L_\alpha(\mu)$  to  $[0, \infty]$ . This function is lower semi-continuous in  $L_\alpha(\mu)$  for any  $\alpha > 0$ ,  $\beta \in \mathbb{R}$ . Additionally, it is continuous in  $L_\alpha(\mu)$  when  $\alpha > \beta > 0$  and the relative entropy is finitely defined.

**Proof.** First let us consider any  $\alpha > 0$  and take  $p_n \rightarrow p$  in  $L_\alpha(\mu)$ . Then,  $\|p_n\|_\alpha \rightarrow \|p\|_\alpha$ . Also,  $|p_n^\alpha - p^\alpha| \leq |p_n|^\alpha + |p|^\alpha$  and hence a general version of the dominated convergence theorem yields  $p_n^\alpha \rightarrow p^\alpha$  in  $L_1(\mu)$ . Thus, we get

$$p_{n,\alpha} := \frac{p_n^\alpha}{\int p_n^\alpha d\mu} \rightarrow p_\alpha \quad \text{in } L_1(\mu). \quad (21)$$

Further, following ([107], Lemma 1), we know that the function  $h \rightarrow \int \phi_\lambda(h) d\nu$  is lower semi-continuous in  $L_1(\nu)$  for any  $\lambda \in \mathbb{R}$  and any probability measure  $\nu$  on  $(\Omega, \mathcal{A})$ . Taking  $\nu = Q_\alpha$ , we get from (21) that  $p_{n,\alpha}/q_\alpha \rightarrow p_\alpha/q_\alpha$  in  $L_1(\nu)$ . Therefore, the above lower semi-continuity result along with (9) implies that

$$\liminf_{n \rightarrow \infty} D_\lambda(p_{n,\alpha}, q_\alpha) \geq D_\lambda(p_\alpha, q_\alpha) \geq 0, \quad \lambda \in \mathbb{R}. \quad (22)$$

Now, note that the function  $\psi(u) = \frac{1}{\rho} \log(\text{sign}(\rho)u + 1)$  is continuous and increasing on  $[0, \infty)$  for  $\rho > 0$  and on  $[0, 1)$  for  $\rho < 0$ . Thus, combining (22) with the definition of the relative  $(\alpha, \beta)$ -entropy in (13), we get that

$$\liminf_{n \rightarrow \infty} \mathcal{RE}_{\alpha, \beta}(p_n, q) \geq \mathcal{RE}_{\alpha, \beta}(p, q), \quad (23)$$

i.e., the function  $p \mapsto \mathcal{RE}_{\alpha, \beta}(p, q)$  is lower semi-continuous.

Finally, consider the case  $\alpha > \beta > 0$ . Note that the dual space of  $L_{\alpha/\beta}(\mu)$  is  $L_{\frac{\alpha}{\alpha-\beta}}(\mu)$  since  $\alpha > \beta > 0$ . Also, for  $q \in L_\alpha(\mu)$ , we have  $\left(\frac{q}{\|q\|_\alpha}\right)^{\alpha-\beta} \in L_{\frac{\alpha}{\alpha-\beta}}(\mu)$ , the dual space of the Banach space  $L_{\alpha/\beta}(\mu)$ . Therefore, the function  $T : L_{\alpha/\beta}(\mu) \mapsto \mathbb{R}$  defined by

$$T(h) = \int h \left( \frac{q}{\|q\|_\alpha} \right)^{\alpha-\beta} d\mu, \quad h \in L_{\alpha/\beta}(\mu),$$

is a bounded linear functional and hence continuous. Now, take  $p_n \rightarrow p$  in  $L_\alpha(\mu)$  so that  $\|p_n\|_\alpha \rightarrow \|p\|_\alpha$  as  $n \rightarrow \infty$ . Therefore,  $\left(\frac{p_n}{\|p_n\|_\alpha}\right) \rightarrow \left(\frac{p}{\|p\|_\alpha}\right)$  in  $L_\alpha(\mu)$  implying  $\left(\frac{p_n}{\|p_n\|_\alpha}\right)^\beta \rightarrow \left(\frac{p}{\|p\|_\alpha}\right)^\beta$  in  $L_{\alpha/\beta}(\mu)$ . Hence, by the continuity of  $T$  on  $L_{\alpha/\beta}(\mu)$ , we get

$$T \left( \left( \frac{p_n}{\|p_n\|_\alpha} \right)^\beta \right) \rightarrow T \left( \left( \frac{p}{\|p\|_\alpha} \right)^\beta \right), \quad \text{as } n \rightarrow \infty.$$

However, from (14), we get

$$\mathcal{RE}_{\alpha, \beta}(p_n, q) = \frac{\alpha}{\beta(\beta - \alpha)} \log T \left( \left( \frac{p_n}{\|p_n\|_\alpha} \right)^\beta \right) \rightarrow \frac{\alpha}{\beta(\beta - \alpha)} \log T \left( \left( \frac{p}{\|p\|_\alpha} \right)^\beta \right) = \mathcal{RE}_{\alpha, \beta}(p, q). \quad (24)$$

This proves the continuity of  $\mathcal{RE}_{\alpha, \beta}(p, q)$  in its first argument when  $\alpha > \beta > 0$ .  $\square$

**Remark 2.** Whenever  $\Omega$  is finite (discrete) equipped with the counting measure  $\mu$ , all integrals in the definition of  $\mathcal{RE}_{\alpha, \beta}(P, Q)$  become finite sums and any limit can be taken inside these finite sums. Thus, whenever defined finitely, the function  $p \mapsto \mathcal{RE}_{\alpha, \beta}(p, q)$  is always continuous in this case.

**Remark 3.** For a general infinite space  $\Omega$ , the function  $p \mapsto \mathcal{RE}_{\alpha, \beta}(p, q)$  is not necessarily continuous for the cases  $\alpha < \beta$ . This can be seen by using the same counterexample as given in Remark 3 of [16]. However, it is yet to be verified if this function can be continuous for  $\beta < 0$  cases.

**Proposition 6.** For a given  $p \in L_\alpha(\mu)$ , consider the function  $q \mapsto \mathcal{RE}_{\alpha, \beta}(p, q)$  from  $q \in L_\alpha(\mu)$  to  $[0, \infty]$ . This function is lower semi-continuous in  $L_\alpha(\mu)$  for any  $\alpha > 0$  and  $\beta \in \mathbb{R}$ .

**Proof.** Fix an  $\alpha > 0$  and  $\beta \in \mathbb{R}$ , which in turn fixes a  $\lambda \in \mathbb{R}$ . Note that, the relative  $(\alpha, \beta)$ -entropy measure can be re-expressed from (13) as

$$\mathcal{RE}_{\alpha, \beta}(p, q) = \frac{1}{\beta\lambda} \log \left[ \text{sign}(\beta\lambda) D_{-(\lambda+1)}(q_\alpha, p_\alpha) + 1 \right]. \quad (25)$$

Now, consider a sequence  $q_n \rightarrow q$  in  $L_\alpha(\mu)$  and proceed as in the proof of Proposition 5 using ([107], Lemma 1) to obtain

$$\liminf_{n \rightarrow \infty} D_{-(\lambda+1)}(q_n, p_\alpha) \geq D_{-(\lambda+1)}(q_\alpha, p_\alpha) \geq 0, \quad \lambda \in \mathbb{R}. \quad (26)$$

Now, whenever  $D_{-(\lambda+1)}(q_\alpha, p_\alpha) = 1$  with  $\beta\lambda < 0$  or  $D_{-(\lambda+1)}(q_\alpha, p_\alpha) = \infty$  with  $\beta\lambda > 0$ , we get from (25) and (26) that

$$\liminf_{n \rightarrow \infty} \mathcal{RE}_{\alpha,\beta}(p, q_n) = \mathcal{RE}_{\alpha,\beta}(p, q) = +\infty. \quad (27)$$

In all other cases, we consider the function  $\psi(u) = \frac{1}{\rho} \log(\text{sign}(\rho)u + 1)$  as in the proof of Proposition 5. This function is continuous and increasing whenever the corresponding relative entropy is finitely defined for all tuning parameter values; on  $[0, \infty)$  for  $\rho > 0$  and on  $[0, 1)$  for  $\rho < 0$ . Hence, again combining (26) with (25) through the function  $\psi$ , we conclude that

$$\liminf_{n \rightarrow \infty} \mathcal{RE}_{\alpha,\beta}(p, q_n) \geq \mathcal{RE}_{\alpha,\beta}(p, q). \quad (28)$$

Therefore, the function  $q \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$  is also lower semi-continuous.  $\square$

**Remark 4.** As in Remark 2, whenever  $\Omega$  is finite (discrete) and is equipped with the counting measure  $\mu$ , the function  $q \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$  is continuous in  $L_\alpha(\mu)$  for any fixed  $p \in L_\alpha(\mu)$ ,  $\alpha > 0$  and  $\beta \in \mathbb{R}$ .

### 3.2. Convexity

It has been shown in [16] that the relative  $\alpha$ -entropy (i.e.,  $\mathcal{RE}_{\alpha,1}(p, q)$ ) is neither convex nor bi-convex, but it is quasi-convex in  $p$ . For general  $\beta \neq 1$ , however, the relative  $(\alpha, \beta)$ -entropy  $\mathcal{RE}_{\alpha,\beta}(p, q)$  is not even quasi-convex in  $p \in L_\alpha(\mu)$ ; rather it is quasi-convex on the  $\beta$ -power transformed space of densities,  $L_\alpha(\mu)^\beta = \{p^\beta : p \in L_\alpha(\mu)\}$ , as described in the following theorem. Note that, for  $\alpha, \beta > 0$ ,  $L_\alpha(\mu)^\beta = L_{\alpha/\beta}(\mu)$ . Here we define the lower level set  $B_{\alpha,\beta}(q, r) = \{p : \mathcal{RE}_{\alpha,\beta}(p, q) \leq r\}$  and its power-transformed set  $B_{\alpha,\beta}(q, r)^\beta = \{p^\beta : p \in B_{\alpha,\beta}(q, r)\}$ , for any  $q \in L_\alpha(\mu)$  and  $r > 0$ .

**Theorem 2.** For any given  $\alpha > 0$ ,  $\beta \in \mathbb{R}$  and  $q \in L_\alpha(\mu)$ , the sets  $B_{\alpha,\beta}(q, r)^\beta$  are convex for all  $r > 0$ . Therefore, the function  $p^\beta \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$  is quasi-convex on  $L_\alpha(\mu)^\beta$ .

**Proof.** Note that, at  $\beta = 1$ , our theorem coincides with Proposition 5 of [16]; so we will prove the result for the case  $\beta \neq 1$ . Fix  $\alpha, r > 0$ , a real  $\beta \notin \{1, \alpha\}$ ,  $q \in L_\alpha(\mu)$ , and  $p_0, p_1 \in B_{\alpha,\beta}(q, r)$ . Then  $p_0^\beta, p_1^\beta \in B_{\alpha,\beta}(q, r)^\beta$ . For  $\tau \in [0, 1]$ , we consider  $p_\tau^\beta = \tau p_1^\beta + \bar{\tau} p_0^\beta$  with  $\bar{\tau} = 1 - \tau$ . We need to show that  $p_\tau^\beta \in B_{\alpha,\beta}(q, r)^\beta$ , i.e.,  $\mathcal{RE}_{\alpha,\beta}(p_\tau, q) \leq r$ .

Now, from (14), we have

$$\mathcal{RE}_{\alpha,\beta}(p, q) = \frac{1}{\beta\lambda} \log \int \left( \frac{p}{\|p\|_\alpha} \right)^\beta \left( \frac{q}{\|q\|_\alpha} \right)^{\alpha-\beta} d\mu = \frac{1}{\beta\lambda} \log \int \left( \frac{p_\alpha}{q_\alpha} \right)^{\beta/\alpha} dQ_\alpha. \quad (29)$$

Since  $p_0^\beta, p_1^\beta \in B_{\alpha,\beta}(q, r)^\beta$ , we have

$$\text{sign}(\beta\lambda) \int \left( \frac{p_\tau}{\|p_\tau\|_\alpha} \right)^\beta \left( \frac{q}{\|q\|_\alpha} \right)^{\alpha-\beta} d\mu \leq \text{sign}(\beta\lambda) e^{r\beta\lambda}, \quad \text{for } \tau = 0, 1. \quad (30)$$

For any  $\tau \in (0, 1)$ , we get

$$\begin{aligned} \text{sign}(\beta\lambda) \int \left( \frac{p_\tau}{\|p_\tau\|_\alpha} \right)^\beta \left( \frac{q}{\|q\|_\alpha} \right)^{\alpha-\beta} d\mu &= \text{sign}(\beta\lambda) \int \left( \frac{\tau p_1^\beta + \bar{\tau} p_0^\beta}{\|p_\tau\|_\alpha^\beta} \right) \left( \frac{q}{\|q\|_\alpha} \right)^{\alpha-\beta} d\mu, && [\text{by definition of } p_\tau] \\ &\leq \text{sign}(\beta\lambda) e^{r\beta\lambda} \frac{\tau \|p_1\|_\alpha^\beta + \bar{\tau} \|p_0\|_\alpha^\beta}{\|p_\tau\|_\alpha^\beta}, && [\text{by (30)}]. \end{aligned} \quad (31)$$

Now, using the extended Minkowski's inequalities from Lemma 1, given below, along with (31) and noting that  $\beta\lambda = \beta(\beta - \alpha)/\alpha$ , we get that

$$\text{sign}(\beta\lambda) \int \left( \frac{p_\tau}{\|p_\tau\|_\alpha} \right)^\beta \left( \frac{q}{\|q\|_\alpha} \right)^{\alpha-\beta} d\mu \leq \text{sign}(\beta\lambda) e^{r\beta\lambda}.$$

Therefore, by (29) and the fact that  $\frac{1}{\rho} \log(\text{sign}(\rho)u)$  is increasing in  $u$ , we finally get  $\mathcal{RE}_{\alpha,\beta}(p_\tau, q) \leq r$ . This proves the result for  $\alpha \neq \beta$ .

The case  $\beta = \alpha$  can be proved in a similar manner and is left as an exercise to the readers.  $\square$

**Lemma 1** (Extended Minkowski's inequality). *Fix  $\alpha > 0$ , a real  $\beta \notin \{1, \alpha\}$ ,  $p_0, p_1 \in L_\alpha(\mu)$ , and  $\tau \in [0, 1]$ . Define  $p_\tau^\beta = \tau p_1^\beta + \bar{\tau} p_0^\beta$  with  $\bar{\tau} = 1 - \tau$ . Then we have the following inequalities:*

$$\|p_\tau\|_\alpha^\beta \geq \tau \|p_1\|_\alpha^\beta + \bar{\tau} \|p_0\|_\alpha^\beta, \quad \text{if } \beta(\beta - \alpha) > 0, \quad (32)$$

$$\|p_\tau\|_\alpha^\beta \leq \tau \|p_1\|_\alpha^\beta + \bar{\tau} \|p_0\|_\alpha^\beta, \quad \text{if } \beta(\beta - \alpha) < 0. \quad (33)$$

**Proof.** It follows by using the Jensen's inequality and the convexity of the function  $x^{\beta/\alpha}$ .  $\square$

Next, note in view of Proposition 4 that, for any  $p, q \in L_\alpha(\mu)$ ,  $\mathcal{RE}_{\alpha,\beta}(p, q) = \mathcal{RE}_{\alpha,\alpha-\beta}(q, p)$ . Using this result along with the above theorem, we also get the quasi-convexity of the relative  $(\alpha, \beta)$ -entropy  $\mathcal{RE}_{\alpha,\beta}(p, q)$  in  $q$  over a different power transformed space of densities. This leads to the following theorem.

**Theorem 3.** *For any given  $\alpha > 0$ ,  $\beta \in \mathbb{R}$  and  $p \in L_\alpha(\mu)$ , the function  $q^{\alpha-\beta} \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$  is quasi-convex on  $L_\alpha(\mu)^{\alpha-\beta}$ . In particular, for the choice  $\beta = \alpha - 1$ , the function  $q \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$  is quasi-convex on  $L_\alpha(\mu)$ .*

**Remark 5.** *Note that, at  $\alpha = \beta = 1$ , the  $\mathcal{RE}_{1,1}(p, q)$  coincides with the KLD measure (or relative entropy) which is quasi-convex in both the arguments  $p$  and  $q$  on  $L_\alpha(\mu)$ .*

### 3.3. Extended Pythagorean Relation

Motivated by the quasi-convexity of  $\mathcal{RE}_{\alpha,\beta}(p, q)$  on  $L_\alpha(\mu)^\beta$ , we now present a Pythagorean-type result for the general relative  $(\alpha, \beta)$ -entropy over the power-transformed space. It generalizes the corresponding result for relative  $\alpha$ -entropy [16]; the proof is similar to that in [16] with necessary modifications due to the transformation of the domain space.

**Theorem 4** (Pythagorean Property). *Fix an  $\alpha > 0$ ,  $\beta \in \mathbb{R}$  with  $\beta \neq \alpha$  and  $p_0, p_1, q \in L_\alpha(\mu)$ . Define  $p_\tau \in L_\alpha(\mu)$  by  $p_\tau^\beta = \tau p_1^\beta + \bar{\tau} p_0^\beta$  for  $\tau \in [0, 1]$  and  $\bar{\tau} = 1 - \tau$ .*

- (i) Suppose  $\mathcal{RE}_{\alpha,\beta}(p_0, q)$  and  $\mathcal{RE}_{\alpha,\beta}(p_1, q)$  are finite. Then,  $\mathcal{RE}_{\alpha,\beta}(p_\tau, q) \geq \mathcal{RE}_{\alpha,\beta}(p_0, q)$  for all  $\tau \in [0, 1]$ , i.e., the back-transformation of line segment joining  $p_1^\beta$  and  $p_0^\beta$  on  $L_\alpha(\mu)^\beta$  to  $L_\alpha(\mu)$  does not intersect  $B_{\alpha,\beta}(q, \mathcal{RE}_{\alpha,\beta}(p_0, q))$ , if and only if

$$\mathcal{RE}_{\alpha,\beta}(p_1, q) \geq \mathcal{RE}_{\alpha,\beta}(p_1, p_0) + \mathcal{RE}_{\alpha,\beta}(p_0, q). \quad (34)$$

- (ii) Suppose  $\mathcal{RE}_{\alpha,\beta}(p_\tau, q)$  is finite for some fixed  $\tau \in (0, 1)$ . Then, the back-transformation of line segment joining  $p_1^\beta$  and  $p_0^\beta$  on  $L_\alpha(\mu)^\beta$  to  $L_\alpha(\mu)$  does not intersect  $B_{\alpha,\beta}(q, \mathcal{RE}_{\alpha,\beta}(p_\tau, q))$  if and only if

$$\mathcal{RE}_{\alpha,\beta}(p_1, q) = \mathcal{RE}_{\alpha,\beta}(p_1, p_\tau) + \mathcal{RE}_{\alpha,\beta}(p_\tau, q), \quad (35)$$

$$\text{and } \mathcal{RE}_{\alpha,\beta}(p_0, q) = \mathcal{RE}_{\alpha,\beta}(p_0, p_\tau) + \mathcal{RE}_{\alpha,\beta}(p_\tau, q). \quad (36)$$

**Proof of Part (i).** Let  $P_{\tau,\alpha}$  to be the probability measure having  $\mu$ -density  $p_{\tau,\alpha} = \frac{p_\tau^\alpha}{\int p_\tau^\alpha d\mu}$  for  $\tau \in [0, 1]$ . Also note that, with  $\lambda = \beta/\alpha - 1$ , we have

$$D_\lambda(P_\alpha, Q_\alpha) = \text{sign}(\beta\lambda) \left[ \int \left( \frac{p}{\|p\|_\alpha} \right)^\beta (q_\alpha)^{-\lambda} d\mu - 1 \right], \quad \text{for } p, q \in L_\alpha(\mu). \quad (37)$$

Thus, (34) is equivalent to the statement

$$\text{sign}(\beta\lambda) \|p_0\|_\alpha^\beta \int p_1^\beta (q_\alpha)^{-\lambda} d\mu \geq \text{sign}(\beta\lambda) \int p_1^\beta (p_{0,\alpha})^{-\lambda} d\mu \cdot \int p_0^\beta (q_\alpha)^{-\lambda} d\mu. \quad (38)$$

and we have

$$D_\lambda(P_{\tau,\alpha}, Q_\alpha) = \text{sign}(\beta\lambda) \left[ \int \left( \frac{p_\tau}{\|p_\tau\|_\alpha} \right)^\beta (q_\alpha)^{-\lambda} d\mu - 1 \right] = \text{sign}(\beta\lambda) \frac{s(\tau)}{t(\tau)}, \quad (39)$$

where  $s(\tau) = \int p_\tau^\beta (q_\alpha)^{-\lambda} d\mu$  and  $t(\tau) = \|p_\tau\|_\alpha^\beta$ . Now consider the two implications separately.

*Only if statement:* Now, let us assume that  $\mathcal{RE}_{\alpha,\beta}(p_\tau, q) \geq \mathcal{RE}_{\alpha,\beta}(p_0, q)$  for all  $\tau \in (0, 1)$ . Then, we get  $\frac{1}{\tau} [D_\lambda(P_{\tau,\alpha}, Q_\alpha) - D_\lambda(P_{0,\alpha}, Q_\alpha)] \geq 0$  for all  $\tau \in (0, 1)$ . Letting  $\tau \downarrow 0$ , we get that

$$\frac{\partial}{\partial \tau} D_\lambda(P_{\tau,\alpha}, Q_\alpha) \Big|_{\tau=0} \geq 0. \quad (40)$$

In order to find the derivative of  $D_\lambda(P_{\tau,\alpha}, Q_\alpha)$ , we first note that

$$\frac{s(\tau) - s(0)}{\tau} = \frac{1}{\tau} \left[ \int p_\tau^\beta (q_\alpha)^{-\lambda} d\mu - \int p_0^\beta (q_\alpha)^{-\lambda} d\mu \right] = \int (p_1^\beta - p_0^\beta) (q_\alpha)^{-\lambda} d\mu,$$

and hence

$$s'(0) = \lim_{\tau \downarrow 0} \frac{s(\tau) - s(0)}{\tau} = \int (p_1^\beta - p_0^\beta) (q_\alpha)^{-\lambda} d\mu. \quad (41)$$

Further, using a simple modification of the techniques in the proof of ([16], Theorem 9), it is easy to verify that the derivative of  $t(\tau)$  with respect to  $\tau$  exists and is given by

$$t'(\tau) = \left( \int p_\tau^\alpha d\mu \right)^{\frac{(\beta-\alpha)}{\alpha}} \int p_0^{\alpha-\beta} (p_1^\beta - p_0^\beta) d\mu.$$

Hence we get

$$t'(0) = \left( \int p_0^\alpha d\mu \right)^{\frac{(\beta-\alpha)}{\alpha}} \int p_0^{\alpha-\beta} (p_1^\beta - p_0^\beta) d\mu = \int p_1^\beta (p_{0,\alpha})^{-\lambda} d\mu - \|p_0\|_\alpha^\beta. \quad (42)$$

Therefore, the derivative of  $D_\lambda(P_{\tau,\alpha}, Q_\alpha) = \text{sign}(\beta\lambda)s(\tau)/t(\tau)$  exists and is given by  $\text{sign}(\beta\lambda) [t(0)s'(0) - t'(0)s(0)]/t(0)^2$ . Therefore, using (40), we get that

$$\text{sign}(\beta\lambda)t(0)s'(0) \geq \text{sign}(\beta\lambda)t'(0)s(0), \quad (43)$$

which implies (38) after substituting the values from (41) and (42).

*If statement:* Now, let us assume that (34)—or equivalently (38)—holds true. Further, as in the derivation of (38), we can start from the trivial statement

$$\mathcal{RE}_{\alpha,\beta}(p_0, q) = \mathcal{RE}_{\alpha,\beta}(p_0, p_0) + \mathcal{RE}_{\alpha,\beta}(p_0, q),$$

to deduce

$$\text{sign}(\beta\lambda) \|\| p_0 \| \|_\alpha^\beta \int p_0^\beta (q_\alpha)^{-\lambda} d\mu = \text{sign}(\beta\lambda) \int p_0^\beta (p_{0,\alpha})^{-\lambda} d\mu \cdot \int p_0^\beta (q_\alpha)^{-\lambda} d\mu. \quad (44)$$

Now, multiply (38) by  $\tau$  and (44) by  $\bar{\tau}$ , and add to get

$$\text{sign}(\beta\lambda) \|\| p_0 \| \|_\alpha^\beta \int p_\tau^\beta (q_\alpha)^{-\lambda} d\mu \geq \text{sign}(\beta\lambda) \int p_\tau^\beta (p_{0,\alpha})^{-\lambda} d\mu \cdot \int p_0^\beta (q_\alpha)^{-\lambda} d\mu.$$

In view of (37), this implies that

$$\mathcal{RE}_{\alpha,\beta}(p_\tau, q) \geq \mathcal{RE}_{\alpha,\beta}(p_\tau, p_0) + \mathcal{RE}_{\alpha,\beta}(p_0, q) \geq \mathcal{RE}_{\alpha,\beta}(p_0, q).$$

This proves the if statement of Part (i) completing the proof.  $\square$

**Proof of Part (ii).** Note that the if statement follows directly from Part (i).

To prove the only if statement, we first show that  $\mathcal{RE}_{\alpha,\beta}(p_1, q)$  and  $\mathcal{RE}_{\alpha,\beta}(p_0, q)$  are finite since  $\mathcal{RE}_{\alpha,\beta}(p_\tau, q)$  is finite. For this purpose, we note that  $p_1^\beta \leq \tau^{-1} p_\tau^\beta$  by the definition of  $p_\tau$  and hence  $(p_1/q)^\beta \leq \tau^{-1} (p_\tau/q)^\beta$ . Therefore, we get

$$\left( \frac{p_{1,\alpha}}{q_\alpha} \right)^{\beta/\alpha} = \left( \frac{p_1}{q} \right)^\beta \left( \frac{\|q\|}{\|p_1\|} \right)^\beta \leq \frac{1}{\tau} \left( \frac{p_\tau}{q} \right)^\beta \left( \frac{\|q\|}{\|p_1\|} \right)^\beta = \frac{1}{\tau} \left( \frac{p_{\tau,\alpha}}{q_\alpha} \right)^\beta \left( \frac{\|p_\tau\|}{\|p_1\|} \right)^\beta. \quad (45)$$

Integration with respect to  $Q_\alpha$  and using (29), we get  $\mathcal{RE}_{\alpha,\beta}(p_1, q) \leq \mathcal{RE}_{\alpha,\beta}(p_\tau, q) + c < \infty$ , where  $c$  is a constant. Similarly one can also show that  $\mathcal{RE}_{\alpha,\beta}(p_0, q) < \infty$ .

Therefore, we can apply Part (i) to conclude that

$$\mathcal{RE}_{\alpha,\beta}(p_1, q) \geq \mathcal{RE}_{\alpha,\beta}(p_1, p_\tau) + \mathcal{RE}_{\alpha,\beta}(p_\tau, q), \text{ and } \mathcal{RE}_{\alpha,\beta}(p_0, q) \geq \mathcal{RE}_{\alpha,\beta}(p_0, p_\tau) + \mathcal{RE}_{\alpha,\beta}(p_\tau, q). \quad (46)$$

These relations imply that

$$\text{sign}(\beta\lambda) \|\| p_\tau \| \|_\alpha^\beta \int p_1^\beta (q_\alpha)^{-\lambda} d\mu \geq \text{sign}(\beta\lambda) \int p_1^\beta (p_{\tau,\alpha})^{-\lambda} d\mu \cdot \int p_\tau^\beta (q_\alpha)^{-\lambda} d\mu, \quad (47)$$

$$\text{and } \text{sign}(\beta\lambda) \|\| p_\tau \| \|_\alpha^\beta \int p_0^\beta (q_\alpha)^{-\lambda} d\mu \geq \text{sign}(\beta\lambda) \int p_0^\beta (p_{\tau,\alpha})^{-\lambda} d\mu \cdot \int p_\tau^\beta (q_\alpha)^{-\lambda} d\mu. \quad (48)$$

The proof of the above results proceed in a manner analogous to the proof of (38). Now, if either of the inequalities in (46) is strict, the corresponding inequality in (47) or (48) will also be strict. Then, multiplying (47) and (48) by  $\tau$  and  $\bar{\tau}$ , respectively, and adding them we get (44) with a strict inequality (in place of an equality), which is a contradiction. Hence, both inequalities in (46) must be equalities implying (35) and (36). This completes the proof.  $\square$

Note that, at  $\beta = 1$ , the above theorem coincides with Theorem 9 of [16]. However, for general  $\alpha, \beta$  as well, the above extended Pythagorean relation for the relative  $(\alpha, \beta)$ -entropy suggests that it behaves "like" a squared distance (although with a non-linear space transformation). So, one can meaningfully define its projection on to a suitable set which we will explore in the following sections.

#### 4. The Forward Projection of Relative $(\alpha, \beta)$ -Entropy

The forward projection, i.e., minimization with respect to the first argument given a fixed second argument, leads to the important maximum entropy principle of information theory; it also relates to the Gibbs conditioning principle from statistical physics [16]. Let us now formally define and study the forward projection of the relative  $(\alpha, \beta)$ -entropy. Let  $\mathbb{S}^*$  denote the set of probability measure on  $(\Omega, \mathcal{A})$  and let the set of corresponding  $\mu$ -densities be denoted by  $\mathbb{S} = \{p = dP/d\mu : P \in \mathbb{S}^*\}$ .

**Definition 3** (Forward  $(\alpha, \beta)$ -Projection). Fix  $Q \in \mathbb{S}^*$  having  $\mu$ -density  $q \in L_\alpha(\mu)$ . Let  $\mathbb{E} \subset \mathbb{S}$  with  $\mathcal{RE}_{\alpha,\beta}(p, q) < \infty$  for some  $p \in \mathbb{E}$ . Then,  $p^* \in \mathbb{E}$  is called the forward projection of the relative  $(\alpha, \beta)$ -entropy or simply the forward  $(\alpha, \beta)$ -projection (or forward LSD projection) of  $q$  on  $\mathbb{E}$  if it satisfies the relation

$$\mathcal{RE}_{\alpha,\beta}(p^*, q) = \inf_{p \in \mathbb{E}} \mathcal{RE}_{\alpha,\beta}(p, q). \quad (49)$$

Note that we must assume that,  $\mathbb{E} \subset L_\alpha(\mu)$  so that the above relative  $(\alpha, \beta)$ -entropy is finitely defined for  $p \in \mathbb{E}$ .

We first prove the uniqueness of the forward  $(\alpha, \beta)$ -projection from the Pythagorean property, whenever it exists. The following theorem describe the connection of the forward  $(\alpha, \beta)$ -projection with Pythagorean relation; the proof is same as that of ([16], Theorem 10) using Theorem 4 and hence omitted for brevity.

**Theorem 5.** Consider the set  $\mathbb{E} \subset \mathbb{S}$  such that  $\mathbb{E}^\beta$  is convex and fix  $q \in L_\alpha(\mu)$ . Then,  $p^* \in \mathbb{E} \cap B_{\alpha,\beta}(q, \infty)$  is a forward  $(\alpha, \beta)$ -projection of  $q$  on  $\mathbb{E}$  if and only if every  $p \in \mathbb{E} \cap B_{\alpha,\beta}(q, \infty)$  satisfies

$$\mathcal{RE}_{\alpha,\beta}(p, q) \geq \mathcal{RE}_{\alpha,\beta}(p, p^*) + \mathcal{RE}_{\alpha,\beta}(p^*, q). \quad (50)$$

Further, if  $(p^*)^\beta$  is an algebraic inner point of  $\mathbb{E}^\beta$ , i.e., for every  $p \in \mathbb{E}$  there exists  $p' \in \mathbb{E}$  and  $\tau \in (0, 1)$  such that  $(p^*)^\beta = \tau p^\beta + (1 - \tau)(p')^\beta$ , then every  $p \in \mathbb{E}$  satisfies  $\mathcal{RE}_{\alpha,\beta}(p, q) < \infty$  and

$$\mathcal{RE}_{\alpha,\beta}(p, q) = \mathcal{RE}_{\alpha,\beta}(p, p^*) + \mathcal{RE}_{\alpha,\beta}(p^*, q), \text{ and } \mathcal{RE}_{\alpha,\beta}(p', q) = \mathcal{RE}_{\alpha,\beta}(p', p^*) + \mathcal{RE}_{\alpha,\beta}(p^*, q).$$

**Corollary 1** (Uniqueness of Forward  $(\alpha, \beta)$ -Projection). Consider the set  $\mathbb{E} \subset \mathbb{S}$  such that  $\mathbb{E}^\beta$  is convex and fix  $q \in L_\alpha(\mu)$ . If a forward  $(\alpha, \beta)$ -projection of  $q$  on  $\mathbb{E}$  exists, it must be unique a.s.[ $\mu$ ].

**Proof.** Suppose  $p_1^*$  and  $p_2^*$  are two forward  $(\alpha, \beta)$ -projection of  $q$  on  $\mathbb{E}$ . Then, by definition,  $\mathcal{RE}_{\alpha,\beta}(p_1^*, q) = \mathcal{RE}_{\alpha,\beta}(p_2^*, q) < \infty$ . Applying Theorem 5 with  $p^* = p_1^*$  and  $p = p_2^*$ , we get

$$\mathcal{RE}_{\alpha,\beta}(p_2^*, q) \geq \mathcal{RE}_{\alpha,\beta}(p_2^*, p_1^*) + \mathcal{RE}_{\alpha,\beta}(p_1^*, q).$$

Hence  $\mathcal{RE}_{\alpha,\beta}(p_2^*, p_1^*) \leq 0$  or  $\mathcal{RE}_{\alpha,\beta}(p_2^*, p_1^*) = 0$  by non-negativity of relative entropy, which further implies that  $p_1^* = p_2^*$  a.s.[ $\mu$ ] by Proposition 1.  $\square$

Next we will show the existence of the forward  $(\alpha, \beta)$ -projection under suitable conditions. We need to use an extended Apollonius Theorem for the  $\phi$ -divergence measure  $D_\lambda$  used in the definition (13) of the relative  $(\alpha, \beta)$ -entropy. Such a result is proved in [16] for the special case  $\alpha(1 + \lambda) = 1$ ; the following lemma extends it for the general case  $\alpha(1 + \lambda) = \beta \in \mathbb{R}$ .

**Lemma 2.** Fix  $p_0, p_1, q \in L_\alpha(\mu)$ ,  $\tau \in [0, 1]$  and  $\alpha(1 + \lambda) = \beta \in \mathbb{R}$  with  $\alpha > 0$  and define  $r$  satisfying

$$r^\beta = \frac{\frac{\tau}{||p_1||_\alpha^\beta} p_1^\beta + \frac{1-\tau}{||p_0||_\alpha^\beta} p_0^\beta}{\frac{\tau}{||p_1||_\alpha^\beta} + \frac{1-\tau}{||p_0||_\alpha^\beta}}. \quad (51)$$

Let  $p_{j,\alpha} = p_j^\alpha / \int p_j^\alpha d\mu$  for  $j = 0, 1$ , and similarly  $q_\alpha$  and  $r_\alpha$ . Then, if  $\beta(\beta - \alpha) > 0$  we have

$$\tau D_\lambda(p_{1,\alpha}, q_\alpha) + (1 - \tau) D_\lambda(p_{0,\alpha}, q_\alpha) \geq \tau D_\lambda(p_{1,\alpha}, r_\alpha) + (1 - \tau) D_\lambda(p_{0,\alpha}, r_\alpha) + D_\lambda(r_\alpha, q_\alpha), \quad (52)$$

but the inequality gets reversed if  $\beta(\beta - \alpha) < 0$ .

**Proof.** By (37), we get

$$\begin{aligned}
& \tau D_\lambda(p_{1,\alpha}, q_\alpha) + (1 - \tau) D_\lambda(p_{0,\alpha}, q_\alpha) - \tau D_\lambda(p_{1,\alpha}, r_\alpha) - (1 - \tau) D_\lambda(p_{0,\alpha}, r_\alpha) \\
&= \text{sign}(\beta\lambda)\tau \int \left( \frac{p_1}{\|p_1\|_\alpha} \right)^\beta \left[ (q_\alpha)^{-\lambda} - (r_\alpha)^{-\lambda} \right] d\mu + \text{sign}(\beta\lambda)(1 - \tau) \int \left( \frac{p_0}{\|p_0\|_\alpha} \right)^\beta \left[ (q_\alpha)^{-\lambda} - (r_\alpha)^{-\lambda} \right] d\mu \\
&= \text{sign}(\beta\lambda) \|r\|_\alpha^\beta \left[ \frac{\tau}{\|p_1\|_\alpha^\beta} + \frac{1 - \tau}{\|p_0\|_\alpha^\beta} \right] \int \left( \frac{r}{\|r\|_\alpha} \right)^\beta \left[ (q_\alpha)^{-\lambda} - (r_\alpha)^{-\lambda} \right] d\mu \\
&= \text{sign}(\beta\lambda) \|r\|_\alpha^\beta \left[ \frac{\tau}{\|p_1\|_\alpha^\beta} + \frac{1 - \tau}{\|p_0\|_\alpha^\beta} \right] D_\lambda(R_\alpha, Q_\alpha).
\end{aligned}$$

Then the Lemma follows by an application of the extended Minkowski's inequalities (32) and (33) from Lemma 1.  $\square$

We now present the sufficient conditions for the existence of the forward  $(\alpha, \beta)$ -projection in the following theorem.

**Theorem 6** (Existence of Forward  $(\alpha, \beta)$ -Projection). *Fix  $\alpha > 0$  and  $\beta \in \mathbb{R}$  with  $\beta \neq \alpha$  and  $q \in L_\alpha(\mu)$ . Given any set  $\mathbb{E} \subset \mathbb{S}$  for which  $\mathbb{E}^\beta$  is convex and closed and  $\mathcal{RE}_{\alpha,\beta}(p, q) < \infty$  for some  $p \in \mathbb{E}$ , a forward  $(\alpha, \beta)$ -projection of  $q$  on  $\mathbb{E}$  always exists (and it is unique by Corollary 1).*

**Proof.** We prove it separately for the cases  $\beta\lambda > 0$  and  $\beta\lambda < 0$ , extending the arguments from [16]. The case  $\beta\lambda = 0$  can be obtained from these two cases by standard limiting arguments and hence omitted for brevity.

*The Case  $\beta\lambda > 0$ :*

Consider a sequence  $\{p_n\} \subset \mathbb{E}$  such that  $D_\lambda(p_{n,\alpha}, q_\alpha) < \infty$  for each  $n$  and  $D_\lambda(p_{n,\alpha}, q_\alpha) \rightarrow \inf_{p \in \mathbb{E}} D_\lambda(p_\alpha, q_\alpha)$  as  $n \rightarrow \infty$ . Then, by Lemma 2 applied to  $p_m$  and  $p_n$  with  $\tau = 1/2$ , we get

$$\frac{1}{2}D_\lambda(p_{m,\alpha}, q_\alpha) + \frac{1}{2}D_\lambda(p_{n,\alpha}, q_\alpha) \geq \frac{1}{2}D_\lambda(p_{m,\alpha}, r_{m,n,\alpha}) + \frac{1}{2}D_\lambda(p_{n,\alpha}, r_{m,n,\alpha}) + D_\lambda(r_{m,n,\alpha}, q_\alpha), \quad (53)$$

where  $r_{m,n}$  is defined by

$$r_{m,n}^\beta = \frac{\frac{\tau}{\|p_m\|_\alpha^\beta} p_m^\beta + \frac{1-\tau}{\|p_n\|_\alpha^\beta} p_n^\beta}{\frac{\tau}{\|p_m\|_\alpha^\beta} + \frac{1-\tau}{\|p_n\|_\alpha^\beta}}. \quad (54)$$

Note that, since  $\mathbb{E}^\beta$  is convex,  $r_{m,n} \in \mathbb{E}^\beta$  and so  $r_{m,n} \in \mathbb{E}$ . Also, using the non-negativity of divergence, (53) leads to

$$0 \leq \frac{1}{2}D_\lambda(p_{m,\alpha}, r_{m,n,\alpha}) + \frac{1}{2}D_\lambda(p_{n,\alpha}, r_{m,n,\alpha}) \leq \frac{1}{2}D_\lambda(p_{m,\alpha}, q_\alpha) + \frac{1}{2}D_\lambda(p_{n,\alpha}, q_\alpha) - D_\lambda(r_{m,n,\alpha}, q_\alpha). \quad (55)$$

Taking limit as  $m, n \rightarrow \infty$ , one can see that  $\left[ \frac{1}{2}D_\lambda(p_{m,\alpha}, q_\alpha) + \frac{1}{2}D_\lambda(p_{n,\alpha}, q_\alpha) - D_\lambda(r_{m,n,\alpha}, q_\alpha) \right] \rightarrow 0$  and hence  $[D_\lambda(p_{m,\alpha}, r_{m,n,\alpha}) + D_\lambda(p_{n,\alpha}, r_{m,n,\alpha})] \rightarrow 0$ . Thus,  $D_\lambda(p_{m,\alpha}, r_{m,n,\alpha}) \rightarrow 0$  as  $m, n \rightarrow \infty$  by non-negativity. This along with a generalization of Pinker's inequality for  $\phi$ -divergence ([100], Theorem 1) gives

$$\lim_{m,n \rightarrow \infty} \|p_{m,\alpha} - r_{m,n,\alpha}\|_T = 0, \quad (56)$$

whenever  $\lambda(1 + \lambda) > 0$  (which is true since  $\beta\lambda > 0$ ); here  $\|\cdot\|_T$  denotes the total variation norm. Now, by triangle inequality

$$\|p_{m,\alpha} - p_{n,\alpha}\|_T \leq \|p_{m,\alpha} - r_{m,n,\alpha}\|_T + \|p_{n,\alpha} - r_{m,n,\alpha}\|_T \rightarrow 0, \quad \text{as } m, n \rightarrow \infty.$$

Thus,  $\{p_{n,\alpha}\}$  is Cauchy in  $L_1(\mu)$  and hence converges to some  $g \in L_1(\mu)$ , i.e.,

$$\lim_{n \rightarrow \infty} \int |p_{n,\alpha} - g| d\mu = 0, \quad (57)$$

and  $g$  is a probability density with respect to  $\mu$  since each  $p_n$  is so. Also, (57) implies that  $p_{n,\alpha} \rightarrow g$  in  $[\mu]$ -measure and hence  $p_{n,\alpha}^{1/\alpha} \rightarrow g^{1/\alpha}$  in  $L_\alpha(\mu)$  by an application of generalized dominated convergence theorem.

Next, as in the proof of ([16], Theorem 8), we can show that  $\|p_n\|_\alpha$  is bounded and hence  $\|p_n\|_\alpha \rightarrow c$  for some  $c > 0$ , possibly working with a subsequence if needed. Thus we have  $p_n = \|p_n\|_\alpha p_{n,\alpha}^{1/\alpha} \rightarrow cg^{1/\alpha}$  in  $L_\alpha(\mu)$ . However, since  $\mathbb{E}^\beta$  is closed, we have  $\mathbb{E}$  is closed and hence  $cg^{1/\alpha} = p^*$  for some  $p^* \in \mathbb{E}$ . Further, since  $\int g d\mu = 1$ , we must have  $c = \|p^*\|_\alpha$  and hence  $g = p_\alpha^*$ . Since  $p_n \rightarrow p^*$  and  $p^* \in \mathbb{E}$ , Proposition 5 implies that

$$\mathcal{RE}_{\alpha,\beta}(p^*, q) \leq \liminf_{n \rightarrow \infty} \mathcal{RE}_{\alpha,\beta}(p_n, q) = \inf_{p \in \mathbb{E}} \mathcal{RE}_{\alpha,\beta}(p, q) \leq \mathcal{RE}_{\alpha,\beta}(p^*, q),$$

where the second equality follows by continuity of the function  $f(u) = (\beta\lambda)^{-1} \log(\text{sign}(\beta\lambda)u + 1)$ , definitions of  $p_n$  sequence and (13). Hence, we must have  $\mathcal{RE}_{\alpha,\beta}(p^*, q) = \inf_{p \in \mathbb{E}} \mathcal{RE}_{\alpha,\beta}(p, q)$ , i.e.,  $p^*$  is a forward  $(\alpha, \beta)$ -projection of  $q$  on  $\mathbb{E}$ .

*The Case  $\beta\lambda < 0$ :*

Note that, in this case, we must have  $0 < \beta < \alpha$ , since  $\alpha > 0$ . Then, using (29), we can see that

$$\begin{aligned} \inf_{p \in \mathbb{E}} \mathcal{RE}_{\alpha,\beta}(p, q) &= \frac{1}{\beta\lambda} \log \left[ \sup_{p \in \mathbb{E}} \int \left( \frac{p}{\|p\|_\alpha} \right)^\beta \left( \frac{q}{\|q\|_\alpha} \right)^{\alpha-\beta} d\mu \right] \\ &= \frac{1}{\beta\lambda} \log \left[ \sup_{h \in \tilde{\mathbb{E}}} \int h g d\mu \right], \end{aligned} \quad (58)$$

where  $g = \left( \frac{q}{\|q\|_\alpha} \right)^{\alpha-\beta} \in L_{\frac{\alpha}{\alpha-\beta}}(\mu)$  and

$$\tilde{\mathbb{E}} = \left\{ s \left( \frac{p}{\|p\|_\alpha} \right)^\beta : p \in \mathbb{E}, s \in [0, 1] \right\} \subset L_{\alpha/\beta}(\mu).$$

Now, since  $\mathbb{E}^\beta$  and hence  $\mathbb{E}$  is closed, one can show that  $\tilde{\mathbb{E}}$  is also closed; see, e.g., the proof of ([16], Theorem 8). Next, we will show that  $\tilde{\mathbb{E}}$  is also convex. For take  $s_1 \left( \frac{p_1}{\|p_1\|_\alpha} \right)^\beta \in \tilde{\mathbb{E}}$  and  $s_0 \left( \frac{p_0}{\|p_0\|_\alpha} \right)^\beta \in \tilde{\mathbb{E}}$  for some  $s_0, s_1 \in [0, 1]$  and  $p_0, p_1 \in \mathbb{E}$ , and take any  $\tau \in [0, 1]$ . Note that

$$\tau s_1 \left( \frac{p_1}{\|p_1\|_\alpha} \right)^\beta + (1 - \tau)s_0 \left( \frac{p_0}{\|p_0\|_\alpha} \right)^\beta = s_\tau \left( \frac{p_\tau}{\|p_\tau\|_\alpha} \right)^\beta,$$

where

$$p_\tau^\beta = \frac{\tau s_1 \left( \frac{p_1}{\|p_1\|_\alpha} \right)^\beta + (1 - \tau)s_0 \left( \frac{p_0}{\|p_0\|_\alpha} \right)^\beta}{\frac{\tau s_1}{\|p_1\|_\alpha^\beta} + \frac{(1 - \tau)s_0}{\|p_0\|_\alpha^\beta}}, \quad \text{and} \quad s_\tau = \left[ \frac{\tau s_1}{\|p_1\|_\alpha^\beta} + \frac{(1 - \tau)s_0}{\|p_0\|_\alpha^\beta} \right] \|p_\tau\|_\alpha^\beta.$$

However, by convexity of  $\mathbb{E}^\beta$ ,  $p_\tau \in \mathbb{E}$  and also  $0 \leq s_\tau \leq 1$  by the extended Minkowski inequality (33). Therefore,  $s_\tau \left( \frac{p_\tau}{\|p_\tau\|_\alpha} \right)^\beta \in \widetilde{\mathbb{E}}$  and hence  $\widetilde{\mathbb{E}}$  is convex.

Finally, since  $0 < \beta < \alpha$ ,  $L_{\alpha/\beta}(\mu)$  is a reflexive Banach space and hence the closed and convex  $\widetilde{\mathbb{E}} \subset L_{\alpha/\beta}(\mu)$  is also closed in the weak topology. So, the unit ball is compact in the weak topology by the Banach-Alaoglu theorem and hence its closed subset  $\widetilde{\mathbb{E}}$  is also weakly compact. However, since  $g$  belongs to the dual space of  $L_{\alpha/\beta}(\mu)$ , the linear functional  $h \mapsto \int hg d\mu$  is continuous in weak topology and also increasing in  $s$ . Hence its supremum over  $\widetilde{\mathbb{E}}$  is attained at  $s = 1$  and some  $p^* \in \mathbb{E}$ , which is the required forward  $(\alpha, \beta)$ -projection.  $\square$

Before concluding this section, we will present one example of the forward  $(\alpha, \beta)$ -projection onto a transformed-linear family of distributions.

**Example 3** (An example of the forward  $(\alpha, \beta)$ -projection). Fix  $\alpha > 0$ ,  $\beta \in \mathbb{R} \setminus \{0, \alpha\}$  and  $q \in L_\alpha(\mu)$  related to the measure  $Q$ . Consider measurable functions  $f_i : \Omega \mapsto \mathbb{R}$  for  $i \in I$ , an index set, and the family of distributions

$$\mathbb{L}_\beta^* = \left\{ P \in \mathbb{S}^* : \int f_\gamma dP_\beta = 0 \right\} \subset \mathbb{S}^*.$$

Let us denote the corresponding  $\mu$ -density set by  $\mathbb{L}_\beta = \left\{ p = \frac{dP}{d\mu} : P \in \mathbb{L}_\beta^* \right\}$ . We assume that,  $\mathbb{L}_\beta^*$  is non-empty, every  $P \in \mathbb{L}_\beta^*$  is absolute continuous with respect to  $\mu$  and  $\mathbb{L}_\beta \subset L_\alpha(\mu)$ .

Then,  $p^*$  is the forward  $(\alpha, \beta)$ -projection of  $q$  on  $\mathbb{L}_\beta$  if and only if there exists a function  $g$  in the  $L_1(Q_\beta)$ -closure of the linear space spanned by  $\{f_i : i \in I\}$  and a subset  $N \subset \Omega$  such that, for every  $P \in \mathbb{L}_\beta^*$

$$\begin{cases} P(N) = 0 & \text{if } \alpha < \beta, \\ c \int_N q^{\alpha-\beta} dP_\beta \leq \int_{\Omega \setminus N} g dP_\beta & \text{if } \alpha > \beta, \end{cases}$$

with  $c = \frac{\int (p^*)^\alpha d\mu}{\int (p^*)^\beta q^{\alpha-\beta} d\mu}$  and  $p^*$  satisfies

$$\begin{aligned} p^*(x)^{\alpha-\beta} &= cq(x)^{\alpha-\beta} + g(x), & \text{if } x \notin N, \\ p^*(x) &= 0, & \text{if } x \in N. \end{aligned}$$

The proof follows by extending the arguments of the proof of ([16], Theorem 11) and hence it is left as an exercise to the readers.

**Remark 6.** Note that, at the special case  $\beta = 1$ ,  $\mathbb{L}_1^*$  is a linear family of distributions and the above example coincides with ([16], Theorem 11) on the forward projection of relative  $\alpha$ -entropy on  $\mathbb{L}_1^*$ . However, it is still an open question to derive the forward  $(\alpha, \beta)$ -projection on  $\mathbb{L}_1^*$ .

## 5. Statistical Applications: The Minimum Relative Entropy Inference

### 5.1. The Reverse Projection and Parametric Estimation

As in the case of the forward projection of a relative entropy measure, we can also define the reverse projection by minimizing it with respect to the second argument over a convex set  $\mathbb{E}$  keeping the first argument fixed. More formally, we use the following definition.

**Definition 4** (Reverse  $(\alpha, \beta)$ -Projection). Fix  $p \in L_\alpha(\mu)$  and let  $\mathbb{E} \subset \mathbb{S}$  with  $\mathcal{RE}_{\alpha,\beta}(p, q) < \infty$  for some  $q \in \mathbb{E}$ . Then,  $q^* \in \mathbb{E}$  is called the reverse projection of the relative  $(\alpha, \beta)$ -entropy or simply the reverse  $(\alpha, \beta)$ -projection (or reverse LSD projection) of  $p$  on  $\mathbb{E}$  if it satisfies the relation

$$\mathcal{RE}_{\alpha,\beta}(p, q^*) = \inf_{q \in \mathbb{E}} \mathcal{RE}_{\alpha,\beta}(p, q). \quad (59)$$

We can get sufficient conditions for the existence and uniqueness of the reverse  $(\alpha, \beta)$ -projection directly from Theorem 6 and the fact that  $\mathcal{RE}_{\alpha, \beta}(p, q) = \mathcal{RE}_{\alpha, \alpha-\beta}(q, p)$ ; this is presented in the following theorem.

**Theorem 7** (Existence and Uniqueness of Reverse  $(\alpha, \beta)$ -Projection). *Fix  $\alpha > 0$  and  $\beta \in \mathbb{R}$  with  $\beta \neq \alpha$  and  $p \in L_\alpha(\mu)$ . Given any set  $\mathbb{E} \subset \mathbb{S}$  for which  $\mathbb{E}^{\alpha-\beta}$  is convex and closed and  $\mathcal{RE}_{\alpha, \beta}(p, q) < \infty$  for some  $q \in \mathbb{E}$ , a reverse  $(\alpha, \beta)$ -projection of  $p$  on  $\mathbb{E}$  exists and is unique.*

The reverse projection is mostly used in statistical inference where we fix the first argument of a relative entropy measure (or divergence measure) at the empirical data distribution and minimize the relative entropy with respect to the model family of distributions in its second argument. The resulting estimator, commonly known as the minimum distance or minimum divergence estimator, yields the reverse projection of the observed data distribution on the family of model distributions with respect to the relative entropy or divergence under consideration. This approach was initially studied by [9–13] to obtain the popular maximum likelihood estimator as the reverse projection with respect to the relative entropy in (2). More recently, this approach has become widely popular, but with more general relative entropies or divergence measures, to obtain robust estimators against possible contamination in the observed data. Let us describe it more rigorously in the following for our relative  $(\alpha, \beta)$ -entropy.

Suppose we have independent and identically distributed data  $X_1, \dots, X_n$  from a true distribution  $G$  having density  $g$  with respect to some common dominating measure  $\mu$ . We model  $g$  by a parametric model family of  $\mu$ -densities  $\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ , where it is assumed that both  $g$  and  $f_\theta$  have the same support independent of  $\theta$ . Our objective is to infer about the unknown parameter  $\theta$ . In minimum divergence inference, an estimator of  $\theta$  is obtained by minimizing the divergence measure between (an estimate of)  $g$  and  $f_\theta$  with respect to  $\theta \in \Theta$ . Maji et al. [78] have considered the LSD (or equivalently the relative  $(\alpha, \beta)$ -entropy) as the divergence under consideration and defined the corresponding minimum divergence functional at  $G$ , say  $T_{\alpha, \beta}(G)$ , through the relation

$$\mathcal{RE}_{\alpha, \beta}\left(g, f_{T_{\alpha, \beta}(G)}\right) = \min_{\theta \in \Theta} \mathcal{RE}_{\alpha, \beta}(g, f_\theta), \quad (60)$$

whenever the minimum exists. We will refer to  $T_{\alpha, \beta}(G)$  as the minimum relative  $(\alpha, \beta)$ -entropy (MRE) functional, or the minimum LSD functional in the language of [78,79]. Note that, if  $g \in \mathcal{F}$ , i.e.,  $g = f_{\theta_0}$  for some  $\theta_0 \in \Theta$ , then we must have  $T_{\alpha, \beta}(G) = \theta_0$ . If  $g \notin \mathcal{F}$ , we call  $T_{\alpha, \beta}(G)$  as the “best fitting parameter” value, since  $f_{T_{\alpha, \beta}(G)}$  is the closest model element to  $g$  in the LSD sense. In fact, for  $g \notin \mathcal{F}$ ,  $T_{\alpha, \beta}(G)$  is nothing but the reverse  $(\alpha, \beta)$ -projection of the true density  $g$  on the model family  $\mathcal{F}$ , which exists and is unique under the sufficient conditions of Theorem 7. Therefore, under identifiability of the model family  $\mathcal{F}$  we get the existence and uniqueness of the MRE functional, which is presented in the following corollary. Although this estimator was first introduced by [78] in terms of the LSD, the results concerning the existence of the estimate were not provided.

**Corollary 2** (Existence and Uniqueness of the MRE Functional). *Consider the above parametric estimation problem with  $g \in L_\alpha(\mu)$  and  $\mathcal{F} \subset L_\alpha(\mu)$ . Fix  $\alpha > 0$  and  $\beta \in \mathbb{R}$  with  $\beta \neq \alpha$  and assume that the model family  $\mathcal{F}$  is identifiable in  $\theta$ .*

1. Suppose  $g = f_{\theta_0}$  for some  $\theta_0 \in \Theta$ . Then the unique MRE functional is given by  $T_{\alpha, \beta}(G) = \theta_0$ .
2. Suppose  $g \notin \mathcal{F}$ . If  $\mathcal{F}^{\alpha-\beta}$  is convex and closed and  $\mathcal{RE}_{\alpha, \beta}(g, f_\theta) < \infty$  for some  $\theta \in \Theta$ , the MRE functional  $T_{\alpha, \beta}(G)$  exists and is unique.

Further, under standard differentiability assumptions, we can obtain the estimating equation of the MRE functional  $T_{\alpha,\beta}(G)$  as given by

$$\left[ \int f_\theta^\alpha u_\theta d\mu \right] \left[ \int f_\theta^{\alpha-\beta} g^\beta d\mu \right] = \left[ \int f_\theta^{\alpha-\beta} g^\beta u_\theta d\mu \right] \left[ \int f_\theta^\alpha d\mu \right], \quad (61)$$

where  $u_\theta(x) = \frac{\partial}{\partial \theta} \ln f_\theta(x)$ . It is important to note that, at  $\beta = \alpha = 1$ , the MRE functional  $T_{1,1}(G)$  coincides with the maximum likelihood functional since  $\mathcal{RE}_{1,1} = \mathcal{RE}$ , the KLD measure. Based on the estimating Equation (61), Maji et al. [78] extensively studied the theoretical robustness properties of the MRE functional against gross-error contamination in data through the higher order influence function analysis. The classical first order influence function was seen to be inadequate for this purpose; it becomes independent of  $\beta$  at the model but the real-life performance of the MRE functional critically depends on both  $\alpha$  and  $\beta$  [78,79] as we will also see in Section 5.2.

In practice, however, the true data generating density is not known and so we need to use some empirical estimate in place of  $g$  and the resulting value of the MRE functional is called the minimum relative  $(\alpha, \beta)$ -entropy estimator (MREE) or the minimum LSD estimator in the terminology of [78,79]. Note that, when the data are discrete and  $\mu$  is the counting measure, one can use a simple estimate of  $g$  given by the relative frequencies  $r_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i = x)$ , where  $I(A)$  is the indicator function of the event  $A$ ; the corresponding MREE is then obtained by solving (61) with  $g(x)$  replaced by  $r_n(x)$  and integrals replaced by sums over the discrete support. Asymptotic properties of this MREE under discrete models are well-studied by [78,79] for the tuning parameters  $\alpha \geq 1$  and  $\beta \in \mathbb{R}$ ; the same line of argument can be used to extend them also for the cases  $\alpha \in (0, 1)$  in a straightforward manner.

However, in case of continuous data, there is no such simple estimator available to use in place of  $g$  unless  $\beta = 1$ . When  $\beta = 1$ , the estimating Equation (61) depends on  $g$  through the terms  $\int f_\theta^{\alpha-1} g d\mu = \int f_\theta^{\alpha-1} dG$  and  $\int f_\theta^{\alpha-1} u_\theta g d\mu = \int f_\theta^{\alpha-1} u_\theta dG$ ; so we can simply use the empirical distribution function  $G_n$  in place of  $G$  and solve the resulting equation to obtain the corresponding MREE. However, for  $\beta \neq 1$ , we must use a non-parametric kernel estimator  $g_n$  of  $g$  in (61) to obtain the MREE under continuous models; this leads to complications including bandwidth selection while deriving the asymptotics of the resulting MREE. One possible approach to avoid such complications is to use the smoothed model technique, which has been applied in [108] for the case of minimum  $\phi$ -divergence estimators. Another alternative approach has been discussed in [109,110]. However, the detailed analyses of the MREE under the continuous model, in either of the above approaches, are yet to be studied so far.

## 5.2. Numerical Illustration: Binomial Model

Let us now present numerical illustrations under the common binomial model to study the finite sample performance of the MREEs. Along with the known properties of the MREE at  $\alpha \geq 1$  (i.e., the minimum LSD estimators with  $\tau \geq 0$  from [78,79]), here we will additionally explore their properties in case of  $\alpha \in (0, 1)$  and for the new divergences  $\mathcal{RE}_\beta^*(P, Q)$  related to  $\alpha = 0$ .

Suppose  $X_1, \dots, X_n$  are random observations from a true density  $g$  having support  $\chi = \{0, 1, 2, \dots, m\}$  for some positive integer  $m$ . We model  $g$  by the Binomial( $m, \theta$ ) densities  $f_\theta(x) = \binom{n}{x} \theta^x (1 - \theta)^{m-x}$  for  $x \in \chi$  and  $\theta \in [0, 1]$ . Here an estimate  $\hat{g}$  of  $g$  is given by the relative frequency  $\hat{g}(x) = r_n(x)$ . For any  $\alpha > 0$  and  $\beta \in \mathbb{R}$ , the relative  $(\alpha, \beta)$ -entropy between  $\hat{g}$  and  $f_\theta$  is given by

$$\begin{aligned} \mathcal{RE}_{\alpha,\beta}(\hat{g}, f_\theta) &= \frac{1}{\beta} \log \left[ \sum_{x=0}^m \binom{n}{x}^\alpha \left( \frac{\theta}{1-\theta} \right)^{\alpha x} (1-\theta)^{m\alpha} \right] + \frac{1}{\alpha-\beta} \log \left[ \sum_{x=0}^m r_n(x)^\alpha \right] \\ &\quad - \frac{\alpha}{\beta(\alpha-\beta)} \log \left[ \sum_{x=0}^m \binom{n}{x}^{\alpha-\beta} \left( \frac{\theta}{1-\theta} \right)^{(\alpha-\beta)x} (1-\theta)^{m(\alpha-\beta)} r_n(x)^\beta \right], \end{aligned}$$

which can be minimized with respect to  $\theta \in [0, 1]$  to obtain the corresponding MREE of  $\theta$ . Note that, it is also the solution of the estimating Equation (61) with  $g(x)$  replaced by the relative frequency  $r_n(x)$ . However, in this example,  $u_\theta(x) = \frac{x-m\theta}{\theta(1-\theta)}$  and hence the MREE estimating equation simplifies to

$$\frac{\sum_{x=0}^m \binom{n}{x}^\alpha (x - m\theta) \left(\frac{\theta}{1-\theta}\right)^{\alpha x}}{\sum_{x=0}^m \binom{n}{x}^\alpha \left(\frac{\theta}{1-\theta}\right)^{\alpha x}} = \frac{\sum_{x=0}^m (x - m\theta) \binom{n}{x}^{\alpha-\beta} \left(\frac{\theta}{1-\theta}\right)^{(\alpha-\beta)x} r_n(x)^\beta}{\sum_{x=0}^m \binom{n}{x}^{\alpha-\beta} \left(\frac{\theta}{1-\theta}\right)^{(\alpha-\beta)x} r_n(x)^\beta}. \quad (62)$$

We can numerically solve the above estimating equation over  $\theta \in [0, 1]$ , or equivalently over the transformed parameter  $p := \frac{\theta}{1-\theta} \in [0, \infty]$ , to obtain the corresponding MREE (i.e., the minimum LSD estimator).

We simulate random sample of size  $n$  from a binomial population with true parameter  $\theta_0 = 0.1$  with  $m = 10$  and numerically compute the MREE. Repeating this exercise 1000 times, we can obtain an empirical estimate of the bias and the mean squared error (MSE) of the MREE of  $10\theta$  (since  $\theta$  is very small in magnitude). Tables 1 and 2 present these values for sample sizes  $n = 20, 50, 100$  and different values of tuning parameters  $\alpha > 0$  and  $\beta > 0$ ; their existences are guaranteed by Corollary 2. Note that the choice  $\alpha = 1 = \beta$  gives the maximum likelihood estimator whereas  $\beta = 1$  only yields the minimum LDPD estimator with parameter  $\alpha$ . Next, in order to study the robustness, we contaminate 10% of each sample by random observations from a distant binomial distribution with parameters  $\theta = 0.9$  and  $m = 10$  and repeat the above simulation exercise; the resulting bias and MSE for the contaminated samples are given in Tables 3 and 4. Our observations from these tables can be summarized as follows.

- Under pure data with no contamination, the maximum likelihood estimator (the MREE at  $\alpha = 1 = \beta$ ) has the least bias and MSE as expected, which further decrease as sample size increases.
- As we move away from  $\alpha = 1$  and  $\beta = 1$  in either direction, the MSEs of the corresponding MREEs under pure data increase slightly; but as long as the tuning parameters remain within a reasonable window of the  $(1, 1)$  point and neither component is very close to zero, this loss in efficiency is not very significant.
- When  $\alpha$  or  $\beta$  approaches zero, the MREEs become somewhat unstable generating comparatively larger MSE values. This is probably due to the presence of inliers under the discrete binomial model. Note that, the relative  $(\alpha, \beta)$ -entropy measures with  $\beta \leq 0$  are not finitely defined for the binomial model if there is just only one empty cell present in the data.
- Under contamination, the bias and MSE of the maximum likelihood estimator increase significantly but many MREEs remains stable. In particular, the MREEs with  $\beta \geq \alpha$  and the MREEs with  $\beta$  close to zero are non-robust against data contamination. Many of the remaining members of the MREE family provide significantly improved robust estimators.
- In the entire simulation, the combination  $(\alpha = 1, \beta = 0.7)$  appears to provide the most stable results. In Table 4, the best results are available along a tubular region which moves from the top left-hand to the bottom right-hand of the table subject to the conditions that  $\alpha > \beta$  and none of them are very close to zero.
- Based on our numerical experiments, the optimum range of values of  $\alpha, \beta$  providing the most robust minimum relative  $(\alpha, \beta)$ -estimators are  $\alpha = 0.9, 1, 0.5 \leq \beta \leq 0.7$  and  $1 < \alpha \leq 1.5, 0.5 \leq \beta < 1$ . Note that this range includes the estimators based on the logarithmic power divergence measure as well as the new LSD measures with  $\alpha < 1$ .
- Many of the MREEs, which belong to the optimum range mentioned in the last item and are close to the combination  $\alpha = 1 = \beta$ , generally also provide the best trade-off between efficiency under pure data and robustness under contaminated data.

In summary, many MREEs provide highly robust estimators under data contamination along with only a very small loss in efficiency under pure data. These numerical findings about the finite sample behavior of the MREEs under the binomial model and the corresponding optimum range of tuning parameters, for the subclass with  $\alpha \geq 1$ , are consistent with the findings of [78,79] who used a Poisson model. Additionally, our illustrations shed lights on the properties of the MREEs at  $\alpha < 1$  as well and show that some MREEs in this range, e.g., at  $\alpha = 0.9$  and  $\beta = 0.5$ , also yield optimum estimators in terms of the dual goal of high robustness and high efficiency.

**Table 1.** Bias of the MREE for different  $\alpha$ ,  $\beta$  and sample sizes  $n$  under pure data.

$\beta$	$\alpha$									
	0.3	0.5	0.7	0.9	1	1.1	1.3	1.5	1.7	2
$n = 20$										
0.1	-0.210	-0.416	-0.397	-0.311	-0.277	-0.227	-0.130	0.021	0.024	0.122
0.3	2.218	-0.273	-0.229	-0.160	-0.141	-0.115	-0.096	-0.068	-0.036	0.034
0.5	-0.127	0.001	-0.125	-0.088	-0.082	-0.069	-0.058	-0.042	-0.032	-0.019
0.7	-0.093	-0.110	-0.010	-0.046	-0.044	-0.029	-0.023	-0.031	-0.023	-0.020
0.9	-0.066	-0.056	-0.028	-0.001	-0.015	-0.002	0.008	0.000	-0.006	-0.013
1	-0.041	-0.045	-0.017	0.005	-0.002	0.011	0.014	0.012	0.008	-0.003
1.3	-0.035	-0.013	0.023	0.036	0.030	0.039	0.088	0.039	0.035	0.021
1.5	-0.003	0.012	0.048	0.053	0.047	0.058	0.053	0.170	0.048	0.035
1.7	0.012	0.028	0.058	0.067	0.061	0.070	0.070	0.058	0.269	0.045
2	0.008	0.049	0.078	0.084	0.078	0.086	0.087	0.078	0.069	0.444
$n = 50$										
0.1	-0.085	-0.301	-0.254	-0.183	-0.156	-0.106	-0.002	0.114	0.292	0.245
0.3	1.829	-0.176	-0.150	-0.078	-0.066	-0.042	-0.045	-0.014	0.005	0.030
0.5	-0.056	0.099	-0.054	-0.037	-0.033	-0.026	-0.019	-0.009	-0.007	-0.005
0.7	-0.009	-0.059	0.035	-0.012	-0.013	-0.005	-0.002	-0.009	-0.002	0.006
0.9	-0.031	-0.031	-0.009	0.012	0.002	0.013	0.021	0.015	0.008	0.004
1	0.014	-0.023	0.000	0.011	0.009	0.019	0.022	0.020	0.018	0.004
1.3	0.002	-0.004	0.022	0.034	0.027	0.030	0.084	0.034	0.035	0.028
1.5	0.009	0.023	0.038	0.044	0.037	0.042	0.034	0.174	0.040	0.032
1.7	0.028	0.029	0.049	0.054	0.047	0.050	0.047	0.036	0.277	0.039
2	0.040	0.051	0.065	0.068	0.059	0.063	0.060	0.051	0.041	0.464
$n = 100$										
0.1	-0.028	-0.216	-0.175	-0.113	-0.103	-0.063	0.036	0.169	0.452	0.349
0.3	1.874	-0.135	-0.125	-0.052	-0.044	-0.022	-0.038	-0.023	0.009	0.024
0.5	-0.002	0.146	-0.034	-0.026	-0.025	-0.021	-0.019	-0.001	-0.008	-0.009
0.7	0.000	-0.042	0.045	-0.009	-0.013	-0.009	0.000	-0.009	-0.008	-0.001
0.9	0.007	-0.025	-0.015	0.001	-0.004	0.005	0.009	0.013	-0.001	-0.003
1	0.014	-0.010	-0.007	-0.001	-0.001	0.005	0.009	0.014	0.010	0.009
1.3	0.036	0.010	0.006	0.015	0.010	0.010	0.065	0.012	0.019	0.014
1.5	0.041	0.023	0.018	0.022	0.017	0.018	0.006	0.158	0.016	0.015
1.7	0.052	0.027	0.028	0.032	0.024	0.025	0.016	0.009	0.267	0.019
2	0.056	0.043	0.042	0.043	0.033	0.034	0.023	0.020	0.013	0.454

**Table 2.** MSE of the MREE for different  $\alpha$ ,  $\beta$  and sample sizes  $n$  under pure data.

$\beta$	$\alpha$									
	0.3	0.5	0.7	0.9	1	1.1	1.3	1.5	1.7	2
$n = 20$										
0.1	0.347	0.251	0.222	0.145	0.122	0.106	0.098	0.242	0.206	0.240
0.3	7.506	0.147	0.100	0.069	0.063	0.059	0.059	0.062	0.098	0.169
0.5	0.238	0.076	0.067	0.051	0.049	0.047	0.050	0.055	0.064	0.101
0.7	0.177	0.091	0.056	0.045	0.044	0.043	0.045	0.055	0.056	0.071
0.9	0.163	0.085	0.061	0.045	0.042	0.043	0.047	0.053	0.058	0.064
1	0.171	0.085	0.064	0.045	0.042	0.045	0.048	0.053	0.058	0.063
1.3	0.148	0.082	0.065	0.052	0.046	0.046	0.061	0.055	0.058	0.065
1.5	0.146	0.085	0.069	0.056	0.050	0.050	0.051	0.087	0.061	0.065
1.7	0.150	0.085	0.070	0.060	0.053	0.055	0.055	0.056	0.134	0.066
2	0.132	0.091	0.076	0.065	0.059	0.060	0.060	0.060	0.061	0.265
$n = 50$										
0.1	0.334	0.170	0.118	0.066	0.044	0.037	0.067	0.195	0.401	0.275
0.3	5.050	0.093	0.051	0.026	0.021	0.020	0.024	0.027	0.035	0.050
0.5	0.196	0.059	0.030	0.018	0.017	0.018	0.021	0.026	0.030	0.037
0.7	0.191	0.053	0.031	0.018	0.016	0.017	0.023	0.025	0.028	0.035
0.9	0.131	0.050	0.029	0.019	0.016	0.018	0.022	0.025	0.028	0.029
1	0.154	0.044	0.031	0.018	0.017	0.020	0.022	0.024	0.027	0.031
1.3	0.112	0.046	0.029	0.023	0.018	0.018	0.033	0.028	0.029	0.031
1.5	0.108	0.049	0.033	0.024	0.020	0.022	0.022	0.059	0.031	0.031
1.7	0.119	0.049	0.036	0.026	0.022	0.023	0.025	0.025	0.108	0.033
2	0.108	0.053	0.040	0.030	0.025	0.026	0.028	0.029	0.028	0.249
$n = 100$										
0.1	0.295	0.139	0.085	0.038	0.022	0.022	0.068	0.201	0.583	0.403
0.3	4.770	0.075	0.039	0.016	0.011	0.011	0.017	0.019	0.023	0.035
0.5	0.189	0.061	0.022	0.011	0.009	0.012	0.016	0.017	0.022	0.023
0.7	0.141	0.038	0.024	0.010	0.009	0.010	0.014	0.017	0.018	0.021
0.9	0.123	0.035	0.021	0.011	0.009	0.011	0.012	0.015	0.019	0.021
1	0.122	0.036	0.019	0.010	0.009	0.011	0.013	0.016	0.017	0.020
1.3	0.114	0.035	0.019	0.012	0.009	0.010	0.021	0.016	0.017	0.019
1.5	0.105	0.037	0.019	0.012	0.010	0.011	0.012	0.045	0.017	0.020
1.7	0.097	0.034	0.021	0.014	0.011	0.012	0.014	0.014	0.092	0.020
2	0.088	0.039	0.023	0.016	0.012	0.013	0.013	0.016	0.016	0.227

**Table 3.** Bias of the MREE for different  $\alpha$ ,  $\beta$  and sample sizes  $n$  under contaminated data.

$\beta$	$\alpha$									
	0.3	0.5	0.7	0.9	1	1.1	1.3	1.5	1.7	2
$n = 20$										
0.1	-0.104	-0.382	-0.340	-0.243	-0.131	-0.071	0.090	0.188	0.295	0.379
0.3	3.287	-0.157	-0.187	-0.135	-0.113	-0.091	-0.045	0.013	0.107	0.237
0.5	2.691	1.483	-0.024	-0.067	-0.069	-0.043	-0.031	-0.010	-0.003	0.051
0.7	3.004	2.546	1.168	0.036	-0.017	-0.008	0.003	0.006	0.005	0.010
0.9	3.133	2.889	2.319	0.917	0.222	0.058	0.019	0.023	0.017	0.022
1	3.183	2.986	2.558	1.619	0.805	0.214	0.039	0.030	0.031	0.019
1.3	3.239	3.121	2.902	2.550	2.262	1.872	0.613	0.077	0.049	0.040
1.5	3.255	3.170	3.012	2.775	2.606	2.396	1.676	0.571	0.069	0.051
1.7	3.271	3.194	3.071	2.903	2.790	2.661	2.256	1.489	0.578	0.057
2	3.289	3.216	3.122	3.012	2.942	2.865	2.649	2.305	1.690	0.682

**Table 3.** Cont.

$\beta$	$\alpha$									
	0.3	0.5	0.7	0.9	1	1.1	1.3	1.5	1.7	2
$n = 50$										
0.1	0.384	-0.170	-0.189	-0.132	-0.054	0.024	0.104	0.171	0.261	0.382
0.3	3.549	0.000	-0.122	-0.086	-0.077	-0.053	-0.023	0.029	0.054	0.118
0.5	2.875	1.771	0.040	-0.048	-0.048	-0.029	-0.013	-0.015	-0.017	0.003
0.7	3.091	2.698	1.294	0.048	-0.010	-0.014	-0.001	0.004	0.001	-0.005
0.9	3.205	2.945	2.379	0.939	0.226	0.045	0.009	0.013	0.012	0.013
1	3.240	3.011	2.612	1.609	0.793	0.196	0.018	0.014	0.021	0.012
1.3	3.316	3.171	2.925	2.548	2.239	1.819	0.554	0.034	0.020	0.020
1.5	3.346	3.223	3.034	2.780	2.596	2.363	1.589	0.502	0.035	0.022
1.7	3.362	3.254	3.100	2.916	2.791	2.643	2.199	1.383	0.518	0.025
2	3.373	3.281	3.162	3.035	2.955	2.865	2.622	2.236	1.575	0.650
$n = 100$										
0.1	0.610	-0.138	-0.105	-0.031	0.002	0.040	0.117	0.184	0.270	0.381
0.3	3.906	0.136	-0.071	-0.050	-0.052	-0.028	-0.028	-0.008	0.023	0.066
0.5	2.927	1.934	0.101	-0.034	-0.027	-0.016	0.006	0.000	-0.003	-0.008
0.7	3.122	2.761	1.348	0.066	0.004	-0.007	0.007	0.011	0.012	0.000
0.9	3.241	2.955	2.406	0.958	0.238	0.047	0.004	0.014	0.022	0.017
1	3.289	3.045	2.651	1.622	0.798	0.202	0.010	0.011	0.016	0.023
1.3	3.362	3.204	2.944	2.567	2.245	1.812	0.533	0.028	0.015	0.022
1.5	3.384	3.269	3.058	2.802	2.610	2.369	1.567	0.485	0.027	0.018
1.7	3.405	3.305	3.133	2.940	2.811	2.658	2.196	1.357	0.504	0.018
2	3.421	3.327	3.204	3.065	2.980	2.886	2.633	2.234	1.541	0.637

**Table 4.** MSE of the MREE for different  $\alpha, \beta$  and sample sizes  $n$  under contaminated data.

$\beta$	$\alpha$									
	0.3	0.5	0.7	0.9	1	1.1	1.3	1.5	1.7	2
$n = 20$										
0.1	0.403	0.248	0.465	0.576	1.025	1.093	1.613	1.565	1.626	1.591
0.3	12.595	0.142	0.103	0.075	0.192	0.188	0.362	0.590	1.016	1.537
0.5	7.443	2.268	0.088	0.062	0.058	0.059	0.065	0.189	0.241	0.527
0.7	9.209	6.645	1.410	0.069	0.056	0.058	0.063	0.068	0.119	0.208
0.9	9.982	8.493	5.512	0.882	0.119	0.068	0.065	0.069	0.075	0.090
1	10.292	9.072	6.672	2.692	0.693	0.117	0.068	0.070	0.076	0.087
1.3	10.664	9.916	8.574	6.641	5.240	3.610	0.430	0.079	0.079	0.089
1.5	10.778	10.229	9.238	7.850	6.940	5.883	2.917	0.389	0.079	0.087
1.7	10.884	10.379	9.599	8.582	7.942	7.234	5.235	2.326	0.403	0.087
2	11.004	10.515	9.915	9.233	8.814	8.369	7.177	5.472	2.998	0.547
$n = 50$										
0.1	1.552	0.815	0.741	0.703	0.966	1.190	1.129	1.224	1.165	1.210
0.3	14.969	0.105	0.047	0.030	0.078	0.075	0.280	0.559	0.566	0.881
0.5	8.345	3.190	0.049	0.025	0.021	0.022	0.025	0.029	0.035	0.184
0.7	9.634	7.335	1.694	0.031	0.020	0.022	0.027	0.029	0.033	0.039
0.9	10.353	8.723	5.712	0.898	0.077	0.027	0.028	0.030	0.032	0.039
1	10.578	9.126	6.871	2.619	0.645	0.067	0.027	0.030	0.033	0.039
1.3	11.069	10.129	8.608	6.548	5.064	3.359	0.329	0.033	0.034	0.038
1.5	11.263	10.457	9.268	7.787	6.801	5.648	2.576	0.279	0.032	0.038
1.7	11.371	10.655	9.676	8.567	7.854	7.051	4.908	1.968	0.298	0.037
2	11.449	10.833	10.060	9.275	8.793	8.276	6.947	5.079	2.560	0.461

**Table 4.** Cont.

$\beta$	$\alpha$									
	0.3	0.5	0.7	0.9	1	1.1	1.3	1.5	1.7	2
$n = 100$										
0.1	2.102	0.399	0.808	0.945	0.924	0.929	0.891	1.012	1.233	1.120
0.3	17.185	0.141	0.033	0.018	0.013	0.014	0.018	0.142	0.258	0.453
0.5	8.624	3.768	0.056	0.015	0.011	0.015	0.017	0.018	0.022	0.028
0.7	9.809	7.646	1.828	0.024	0.011	0.013	0.018	0.019	0.020	0.023
0.9	10.559	8.764	5.812	0.927	0.070	0.018	0.017	0.020	0.021	0.023
1	10.870	9.312	7.058	2.648	0.645	0.057	0.017	0.019	0.021	0.023
1.3	11.342	10.306	8.691	6.619	5.068	3.312	0.297	0.020	0.020	0.023
1.5	11.494	10.727	9.379	7.880	6.845	5.646	2.484	0.251	0.021	0.021
1.7	11.632	10.960	9.848	8.675	7.932	7.101	4.866	1.873	0.272	0.022
2	11.739	11.102	10.297	9.422	8.910	8.363	6.973	5.040	2.420	0.430

### 5.3. Application to Testing Statistical Hypothesis

We end the paper with a very brief indication on the potential of the relative  $(\alpha, \beta)$ -entropy or the LSD measure in statistical hypothesis testing problems. The minimum possible value of the relative entropy or divergence measure between the data and the null distribution indicates the amount of departure from null and hence can be used to develop a statistical testing procedure.

Consider the parametric estimation set-up as in Section 5.1 with  $g \in \mathcal{F}$  and fix a parameter value  $\theta_0 \in \Theta$ . Suppose we want to test the simple null hypothesis in the one sample case given by

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0.$$

Maji et al. [78] have developed the LSD-based test statistics for the above testing problem as given by

$$T_{n,\alpha,\beta}^{(1)} = 2n\mathcal{RE}_{\alpha,\beta}(f_{\hat{\theta}_{\alpha,\beta}}, f_{\theta_0}), \quad (63)$$

where  $\hat{\theta}_{\alpha,\beta}$  is the MREE with parameters  $\alpha$  and  $\beta$ . [78,79] have also developed the LSD-based test for a simple two-sample problem where two independent samples of sizes  $n_1$  and  $n_2$  are given from true densities  $f_{\theta_1}, f_{\theta_2} \in \mathcal{F}$ , respectively and we want to test for the homogeneity of the two samples through the hypothesis

$$H_0 : \theta_1 = \theta_2 \quad \text{against} \quad H_1 : \theta_1 \neq \theta_2.$$

The proposed test statistics for this two-sample problem has the form

$$T_{n,\alpha,\beta}^{(2)} = \frac{2n_1 n_2}{n_1 + n_2} \mathcal{RE}_{\alpha,\beta}(f_{(1)\hat{\theta}_{\alpha,\beta}}, f_{(2)\hat{\theta}_{\alpha,\beta}}), \quad (64)$$

where  $(1)\hat{\theta}_{\alpha,\beta}$  and  $(2)\hat{\theta}_{\alpha,\beta}$  are the MREEs of  $\theta_1$  and  $\theta_2$ , respectively, obtained from the two samples separately. Note that, at  $\alpha = \beta = 1$ , both the test statistics in (63) and (64) become asymptotically equivalent to the corresponding likelihood ratio tests under the respective null hypothesis. Maji et al. [78,79] have studied the asymptotic properties of these two tests, which have asymptotic null distributions as linear combinations of chi-square distributions. They have also numerically illustrated the benefits of these LSD or relative  $(\alpha, \beta)$ -entropy-based tests, although with tuning parameters  $\alpha \geq 1$  only, to achieve robust inference against possible contamination in the sample data.

The same approach can also be used to develop robust tests for more complex hypothesis testing problems based on the relative  $(\alpha, \beta)$ -entropy or the LSD measures, now with parameters  $\alpha > 0$ , and also using the new divergences  $\mathcal{RE}_\beta^*(\cdot, \cdot)$ . For example, consider the above one sample set-up and a subset  $\Theta_0 \subset \Theta$  and let we are interested in testing the composite hypothesis

$$H_0 : \theta \in \Theta_0 \text{ against } H_1 : \theta \notin \Theta_0.$$

with similar motivation from (63) and (64), we can construct relative entropy or LSD-based test statistics for testing the above composite hypothesis as given by

$$\widetilde{T_{n,\alpha,\beta}}^{(1)} = 2n\mathcal{RE}_{\alpha,\beta}(f_{\widehat{\theta}_{\alpha,\beta}}, f_{\widehat{\theta}_{\alpha,\beta}}), \quad (65)$$

where  $\widetilde{\theta}_{\alpha,\beta}$  is the restricted MREE with parameters  $\alpha$  and  $\beta$  obtained by minimizing the relative entropy over  $\theta \in \Theta_0$  and  $\widehat{\theta}_{\alpha,\beta}$  is the corresponding unrestricted MREE obtained by minimizing over  $\theta \in \Theta$ . It will surely be of significant interest to study the asymptotic and robustness properties of this relative entropy-based test for the above composite hypothesis under one sample or even more general hypotheses with two or more samples. However, considering the length of the present paper, which is primarily focused on the geometric properties of entropies and relative entropies, we have deferred the detailed analyses of such MREE-based hypothesis testing procedures in a future report.

## 6. Conclusions

We have explored the geometric properties of the LSD measures through a new information theoretic formulation when we develop this divergence measure as a natural extension of the relative  $\alpha$ -entropy; we refer to it as the two-parameter relative  $(\alpha, \beta)$ -entropy. It is shown to be always lower semicontinuous in both the arguments, but is continuous in its first argument only if  $\alpha > \beta > 0$ . We also proved that the relative  $(\alpha, \beta)$ -entropy is quasi-convex in both its arguments after a suitable (different) transformation of the domain space and derive an extended Pythagorean relation under these transformations. Along with the study of its forward and reverse projections, statistical applications are also discussed.

It is worthwhile to note that the information theoretic divergences can also be used to define new measures of robustness and efficiency of a parameter estimate; one can then obtain the optimum robust estimator, along Hampel's infinitesimal principle, to achieve the best trade-off between these divergence-based summary measures [111–113]. In particular, the LDPD measure, a prominent member of our LSD or relative  $(\alpha, \beta)$ -entropy family, has been used by [113] who have illustrated important theoretical properties including different types of equivariance of the resulting optimum estimators besides their strong robustness properties. A similar approach can also be used with our general relative  $(\alpha, \beta)$ -entropies to develop estimators with enhanced optimality properties, establishing a better robustness-efficiency trade-off.

The present work opens up several interesting problems to be solved in future research as already noted throughout the paper. In particular, we recall that the relative  $\alpha$ -entropy has an interpretation from the problem of guessing under source uncertainty [17,71]. As an extension of relative  $\alpha$ -entropy, a similar information theoretic interpretation of the relative  $(\alpha, \beta)$ -entropy (i.e., the LSD) is expected and its proper interpretation will be a useful development. Additionally, we have obtained a new extension of the Renyi entropy as a by-product and detailed study of this new entropy measure and its potential applications may lead to a new aspect of the mathematical information theory. Also, statistical applications of these measures need to be studied thoroughly specially for the continuous models, where the complications of a kernel density estimator is unavoidable, and for testing complex composite hypotheses from one or more samples. We hope to pursue some of these interesting extensions in future.

**Author Contributions:** Conceptualization, A.B. and A.G.; Methodology, A.B. and A.G.; Coding and Numerical Work, A.G.; Validation, A.G.; Formal Analysis, A.G. and A.B.; Investigation, A.G. and A.B.

**Funding:** The research of the first author is funded by the INSPIRE Faculty Research Grant from the Department of Science and Technology, Government of India.

**Acknowledgments:** The authors wish to thank four anonymous referees whose comments have led to a significantly improved version of the paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

KLD	Kullback-Leibler Divergence
LDPD	Logarithmic Density Power Divergence
LSD	Logarithmic Super Divergence
GRE	Generalized Renyi Entropy
MRE	Minimum Relative $(\alpha, \beta)$ -entropy
MREE	Minimum Relative $(\alpha, \beta)$ -entropy Estimator
MSE	Mean Squared Error

## References

- Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
- Shannon, C.E. Communication in the presence of noise. *Proc. IRE* **1949**, *37*, 10–21.
- Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, USA, 1949.
- Khinchin, A.I. The entropy concept in probability theory. *Uspekhi Matematicheskikh Nauk* **1953**, *8*, 3–20.
- Khinchin, A.I. On the fundamental theorems of information theory. *Uspekhi Matematicheskikh Nauk* **1956**, *11*, 17–75.
- Khinchin, A.I. *Mathematical Foundations of Information Theory*; Dover Publications: New York, NY, USA, 1957.
- Kolmogorov, A.N. *Foundations of the Theory of Probability*; Chelsea Publishing Co.: New York, NY, USA, 1950.
- Kolmogorov, A.N. On the Shannon theory of information transmission in the case of continuous signals. *IRE Trans. Inf. Theory* **1956**, *IT-2*, 102–108.
- Kullback, S. An application of information theory to multivariate analysis. *Ann. Math. Stat.* **1952**, *23*, 88–102.
- Kullback, S. A note on information theory. *J. Appl. Phys.* **1953**, *24*, 106–107.
- Kullback, S. Certain inequalities in information theory and the Cramer-Rao inequality. *Ann. Math. Stat.* **1954**, *25*, 745–751.
- Kullback, S. An application of information theory to multivariate analysis II. *Ann. Math. Stat.* **1956**, *27*, 122–145.
- Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
- Rosenkrantz, R.D. *E T Jaynes: Papers on Probability, Statistics and Statistical Physics*; Springer Science and Business Media: New York, NY, USA, 1983.
- Van Campenhout, J.M.; Cover, T.M. Maximum entropy and conditional probability. *IEEE Trans. Inf. Theory* **1981**, *27*, 483–489.
- Kumar, M.A.; Sundaresan, R. Minimization Problems Based on Relative  $\alpha$ -Entropy I: Forward Projection. *IEEE Trans. Inf. Theory* **2015**, *61*, 5063–5080.
- Sundaresan, R. Guessing under source uncertainty. *Proc. IEEE Trans. Inf. Theory* **2007**, *53*, 269–287.
- Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **1975**, *3*, 146–158.
- Csiszár, I. Sanov property, generalized I -projection, and a conditional limit theorem. *Ann. Probab.* **1984**, *12*, 768–793.
- Csiszár, I.; Shields, P. *Information Theory and Statistics: A Tutorial*; NOW Publishers: Hanover, NH, USA, 2004.
- Csiszár, I.; Tusnády, G. Information geometry and alternating minimization procedures. *Stat. Decis.* **1984**, *1*, 205–237.
- Amari, S.I.; Karakida, R.; Oizumi, M. Information Geometry Connecting Wasserstein Distance and Kullback-Leibler Divergence via the Entropy-Relaxed Transportation Problem. *arXiv* **2017**, arXiv:1709.10219.
- Costa, S.I.; Santos, S.A.; Strapasson, J.E. Fisher information distance: A geometrical reading. *Discret. Appl. Math.* **2015**, *197*, 59–69.

24. Nielsen, F.; Sun, K. Guaranteed bounds on the Kullback-Leibler divergence of univariate mixtures. *IEEE Signal Process. Lett.* **2016**, *23*, 1543–1546.
25. Amari, S.I.; Cichocki, A. Information geometry of divergence functions. *Bull. Pol. Acad. Sci. Tech. Sci.* **2010**, *58*, 183–195.
26. Contreras-Reyes, J.E.; Arellano-Valle, R.B. Kullback-Leibler divergence measure for multivariate skew-normal distributions. *Entropy* **2012**, *14*, 1606–1626.
27. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya Centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466.
28. Pinski, F.J.; Simpson, G.; Stuart, A.M.; Weber, H. Kullback–Leibler approximation for probability measures on infinite dimensional spaces. *SIAM J. Math. Anal.* **2015**, *47*, 4091–4122.
29. Attouch, H.; Bolte, J.; Redont, P.; Souleyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Math. Oper. Res.* **2010**, *35*, 438–457.
30. Eliazar, I.; Sokolov, I.M. Maximization of statistical heterogeneity: From Shannon’s entropy to Gini’s index. *Phys. A Stat. Mech. Appl.* **2010**, *389*, 3023–3038.
31. Monthus, C. Non-equilibrium steady states: Maximization of the Shannon entropy associated with the distribution of dynamical trajectories in the presence of constraints. *J. Stat. Mech. Theory Exp.* **2011**, *2011*, P03008.
32. Bafrouei, H.H.; Ohadi, A. Application of wavelet energy and Shannon entropy for feature extraction in gearbox fault detection under varying speed conditions. *Neurocomputing* **2014**, *133*, 437–445.
33. Batty, M. Space, Scale, and Scaling in Entropy Maximizing. *Geogr. Anal.* **2010**, *42*, 395–421.
34. Oikonomou, T.; Bagci, G.B. Entropy Maximization with Linear Constraints: The Uniqueness of the Shannon Entropy. *arXiv* **2018**, arXiv:1803.02556.
35. Hoang, D.T.; Song, J.; Periwal, V.; Jo, J. Maximizing weighted Shannon entropy for network inference with little data. *arXiv* **2017**, arXiv:1705.06384.
36. Sriraman, T.; Chakrabarti, B.; Trombettoni, A.; Muruganandam, P. Characteristic features of the Shannon information entropy of dipolar Bose-Einstein condensates. *J. Chem. Phys.* **2017**, *147*, 044304.
37. Sun, M.; Li, Y.; Gemmeke, J.F.; Zhang, X. Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence. *IEEE Trans. Audio Speech Lang. Process.* **2015**, *23*, 1233–1242.
38. Garcia-Fernandez, A.F.; Vo, B.N. Derivation of the PHD and CPHD Filters Based on Direct Kullback-Leibler Divergence Minimization. *IEEE Trans. Signal Process.* **2015**, *63*, 5812–5820.
39. Giantomassi, A.; Ferracuti, F.; Iarlori, S.; Ippoliti, G.; Longhi, S. Electric motor fault detection and diagnosis by kernel density estimation and Kullback-Leibler divergence based on stator current measurements. *IEEE Trans. Ind. Electron.* **2015**, *62*, 1770–1780.
40. Harmouche, J.; Delpha, C.; Diallo, D.; Le Bihan, Y. Statistical approach for nondestructive incipient crack detection and characterization using Kullback-Leibler divergence. *IEEE Trans. Reliab.* **2016**, *65*, 1360–1368.
41. Hua, X.; Cheng, Y.; Wang, H.; Qin, Y.; Li, Y.; Zhang, W. Matrix CFAR detectors based on symmetrized Kullback-Leibler and total Kullback-Leibler divergences. *Digit. Signal Process.* **2017**, *69*, 106–116.
42. Ferracuti, F.; Giantomassi, A.; Iarlori, S.; Ippoliti, G.; Longhi, S. Electric motor defects diagnosis based on kernel density estimation and Kullback-Leibler divergence in quality control scenario. *Eng. Appl. Artif. Intell.* **2015**, *44*, 25–32.
43. Matthews, A.G.D.G.; Hensman, J.; Turner, R.; Ghahramani, Z. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *J. Mach. Learn. Res.* **2016**, *51*, 231–239.
44. Arikan, E. An inequality on guessing and its application to sequential decoding. *IEEE Trans. Inf. Theory* **1996**, *42*, 99–105.
45. Campbell, L.L. A coding theorem and Renyi’s entropy. *Inf. Control* **1965**, *8*, 423–429.
46. Renyi, A. On measures of entropy and information. In *Proceedings of 4th Berkeley Symposium on Mathematical Statistics and Probability I*; University of California: Berkeley, CA, USA, 1961; pp. 547–561.
47. Wei, B.B. Relations between heat exchange and Rényi divergences. *Phys. Rev. E* **2018**, *97*, 042107.
48. Kumar, M.A.; Sason, I. On projections of the Rényi divergence on generalized convex sets. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016.

49. Sadeghpour, M.; Baratpour, S.; Habibirad, A. Exponentiality test based on Renyi distance between equilibrium distributions. *Commun. Stat.-Simul. Comput.* **2017**, doi:10.1080/03610918.2017.1366514.
50. Markel, D.; El Naqa, I.I. PD-0351: Development of a novel segmentation framework using the Jensen Renyi divergence for adaptive radiotherapy. *Radiother. Oncol.* **2014**, *111*, S134.
51. Bai, S.; Lepoint, T.; Roux-Langlois, A.; Sakzad, A.; Stehlé, D.; Steinfeld, R. Improved security proofs in lattice-based cryptography: Using the Rényi divergence rather than the statistical distance. *J. Cryptol.* **2018**, *31*, 610–640.
52. Dong, X. The gravity dual of Rényi entropy. *Nat. Commun.* **2016**, *7*, 12472.
53. Kusuki, Y.; Takayanagi, T. Renyi entropy for local quenches in 2D CFT from numerical conformal blocks. *J. High Energy Phys.* **2018**, *2018*, 115.
54. Kumbhakar, M.; Ghoshal, K. One-Dimensional velocity distribution in open channels using Renyi entropy. *Stoch. Environ. Res. Risk Assess.* **2017**, *31*, 949–959.
55. Xing, H.J.; Wang, X.Z. Selective ensemble of SVDDs with Renyi entropy based diversity measure. *Pattern Recog.* **2017**, *61*, 185–196.
56. Nie, F.; Zhang, P.; Li, J.; Tu, T. An Image Segmentation Method Based on Renyi Relative Entropy and Gaussian Distribution. *Recent Patents Comput. Sci.* **2017**, *10*, 122–130.
57. Ben Bassat, M. f-entropies, probability of error, and feature selection. *Inf. Control* **1978**, *39*, 277–292.
58. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487.
59. Kumar, S.; Ram, G.; Gupta, V. Axioms for  $(\alpha, \beta, \gamma)$ -entropy of a generalized probability scheme. *J. Appl. Math. Stat. Inf.* **2013**, *9*, 95–106.
60. Kumar, S.; Ram, G. A generalization of the Havrda-Charvat and Tsallis entropy and its axiomatic characterization. *Abstr. Appl. Anal.* **2014**, *2014*, 505184.
61. Tsallis, C.; Brigatti, E. Nonextensive statistical mechanics: A brief introduction. *Contin. Mech. Thermodyn.* **2004**, *16*, 223–235.
62. Rajesh, G.; Sunoj, S.M. Some properties of cumulative Tsallis entropy of order  $\alpha$ . *Stat. Pap.* **2016**, doi:10.1007/s00362-016-0855-7.
63. Singh, V.P. *Introduction to Tsallis Entropy Theory in Water Engineering*; CRC Press: Boca Raton, FL, USA, 2016.
64. Pavlos, G.P.; Karakatsanis, L.P.; Iliopoulos, A.C.; Pavlos, E.G.; Tsonis, A.A. Nonextensive Statistical Mechanics: Overview of Theory and Applications in Seismogenesis, Climate, and Space Plasma. In *Advances in Nonlinear Geosciences*; Tsonis, A., Ed.; Springer: Cham, Switzerland, 2018; pp. 465–495.
65. Jamaati, M.; Mehri, A. Text mining by Tsallis entropy. *Phys. A Stat. Mech. Appl.* **2018**, *490*, 1368–1376.
66. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2011.
67. Leise, F.; Vajda, I. On divergence and information in statistics and information theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412.
68. Pardo, L. *Statistical Inference Based on Divergences*; CRC/Chapman-Hall: London, UK, 2006.
69. Vajda, I. *Theory of Statistical Inference and Information*; Kluwer: Boston, MA, USA, 1989.
70. Stummer, W.; Vajda, I. On divergences of finite measures and their applicability in statistics and information theory. *Statistics* **2010**, *44*, 169–187.
71. Sundaresan, R. A measure of discrimination and its geometric properties. In Proceedings of the IEEE International Symposium on Information Theory, Lausanne, Switzerland, 30 June–5 July 2002.
72. Lutwak, E.; Yang, D.; Zhang, G. Cramér-Rao and moment-entropy inequalities for Renyi entropy and generalized Fisher information. *IEEE Trans. Inf. Theory* **2005**, *51*, 473–478.
73. Kumar, M.A.; Sundaresan, R. Minimization Problems Based on Relative  $\alpha$ -Entropy II: Reverse Projection. *IEEE Trans. Infor. Theory* **2015**, *61*, 5081–5095.
74. Jones, M.C.; Hjort, N.L.; Harris, I.R.; Basu, A. A comparison of related density-based minimum divergence estimators. *Biometrika* **2001**, *88*, 865–873.
75. Windham, M. Robustifying model fitting. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 599–609.
76. Fujisawa, H. Normalized estimating equation for robust parameter estimation. *Elect. J. Stat.* **2013**, *7*, 1587–1606.
77. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081.

78. Maji, A.; Ghosh, A.; Basu, A. The Logarithmic Super Divergence and its use in Statistical Inference. *arXiv* **2014**, arXiv:1407.3961.
79. Maji, A.; Ghosh, A.; Basu, A. The Logarithmic Super Divergence and Asymptotic Inference Properties. *ASyA Adv. Stat. Anal.* **2016**, *100*, 99–131.
80. Maji, A.; Chakraborty, S.; Basu, A. Statistical Inference Based on the Logarithmic Power Divergence. *Rashi* **2017**, *2*, 39–51.
81. Lutz, E. Anomalous diffusion and Tsallis statistics in an optical lattice. *Phys. Rev. A* **2003**, *67*, 051402.
82. Douglas, P.; Bergamini, S.; Renzoni, F. Tunable Tsallis Distributions in Dissipative Optical Lattices. *Phys. Rev. Lett.* **2006**, *96*, 110601.
83. Burlaga, L.F.; Viñas, A.F. Triangle for the entropic index  $q$  of non-extensive statistical mechanics observed by Voyager 1 in the distant heliosphere. *Phys. A Stat. Mech. Appl.* **2005**, *356*, 375.
84. Liu, B.; Goree, J. Superdiffusion and Non-Gaussian Statistics in a Driven-Dissipative 2D Dusty Plasma. *Phys. Rev. Lett.* **2008**, *100*, 055003.
85. Pickup, R.; Cywinski, R.; Pappas, C.; Farago, B.; Fouquet, P. Generalized Spin-Glass Relaxation. *Phys. Rev. Lett.* **2009**, *102*, 097202.
86. Devoe, R. Power-Law Distributions for a Trapped Ion Interacting with a Classical Buffer Gas. *Phys. Rev. Lett.* **2009**, *102*, 063001.
87. Khachatryan, V.; Sirunyan, A.; Tumasyan, A.; Adam, W.; Bergauer, T.; Dragicevic, M.; Erö, J.; Fabjan, C.; Friedl, M.; Fröhwirth, R.; et al. Transverse-Momentum and Pseudorapidity Distributions of Charged Hadrons in pp Collisions at  $\sqrt{s} = 7$  TeV. *Phys. Rev. Lett.* **2010**, *105*, 022002.
88. Chatrchyan, S.; Khachatryan, V.; Sirunyan, A.M.; Tumasyan, A.; Adam, W.; Bergauer, T.; Dragicevic, M.; Erö, J.; Fabjan, C.; Friedl, M.; et al. Charged particle transverse momentum spectra in pp collisions at  $\sqrt{s} = 0.9$  and 7 TeV. *J. High Energy Phys.* **2011**, *2011*, 86.
89. Adare, A.; Afanasiev, S.; Aidala, C.; Ajitanand, N.; Akiba, Y.; Al-Bataineh, H.; Alexander, J.; Aoki, K.; Aphecetche, L.; Armendariz, R.; et al. Measurement of neutral mesons in  $p + p$  collisions at  $\sqrt{s} = 200$  GeV and scaling properties of hadron production. *Phys. Rev. D* **2011**, *83*, 052004.
90. Majhi, A. Non-extensive statistical mechanics and black hole entropy from quantum geometry. *Phys. Lett. B* **2017**, *775*, 32–36.
91. Shore, J.E.; Johnson, R.W. Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–37.
92. Caticha, A.; Giffin, A. Updating Probabilities. *AIP Conf. Proc.* **2006**, *872*, 31–42.
93. Presse, S.; Ghosh, K.; Lee, J.; Dill, K.A. Nonadditive Entropies Yield Probability Distributions with Biases not Warranted by the Data. *Phys. Rev. Lett.* **2013**, *111*, 180604.
94. Presse, S. Nonadditive entropy maximization is inconsistent with Bayesian updating. *Phys. Rev. E* **2014**, *90*, 052149.
95. Presse, S.; Ghosh, K.; Lee, J.; Dill, K.A. Reply to C. Tsallis' “Conceptual Inadequacy of the Shore and Johnson Axioms for Wide Classes of Complex Systems”. *Entropy* **2015**, *17*, 5043–5046.
96. Vanslette, K. Entropic Updating of Probabilities and Density Matrices. *Entropy* **2017**, *19*, 664.
97. Cressie, N.; Read, T.R.C. Multinomial goodness-of-fit tests. *J. R. Stat. Soc. B* **1984**, *46*, 440–464.
98. Csiszár, I. Eine informations theoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hung. Acad. Sci.* **1963**, *3*, 85–107. (In German)
99. Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. *Stud. Scientiarum Math. Hung.* **1967**, *2*, 299–318.
100. Csiszár, I. On topological properties of  $f$ -divergences. *Stud. Scientiarum Math. Hung.* **1967**, *2*, 329–339.
101. Csiszár, I. A class of measures of informativity of observation channels. *Priodica Math. Hung.* **1972**, *2*, 191–213.
102. Csiszár, I. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Stat.* **1991**, *19*, 2032–2066.
103. Lindsay, B.G. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Stat.* **1994**, *22*, 1081–1114.
104. Esteban, M.D.; Morales, D. A summary of entropy statistics. *Kybernetika* **1995**, *31*, 337–346.
105. Itakura, F.; Saito, S. Analysis synthesis telephony based on the maximum likelihood method. In Proceedings of the 6th International Congress on Acoustics, Tokyo, Japan, 21–28 August 1968.

106. Fevotte, C.; Bertin, N.; Durrieu, J.L. Nonnegative Matrix Factorization with the Itakura–Saito Divergence: With application to music analysis. *Neural Comput.* **2009**, *21*, 793–830.
107. Teboulle, M.; Vajda, I. Convergence of best  $\phi$ -entropy estimates. *IEEE Trans. Inf. Theory* **1993**, *39*, 297–301.
108. Basu, A.; Lindsay, B.G. Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Ann. Inst. Stat. Math.* **1994**, *46*, 683–705.
109. Broniatowski, M.; Keziou, A. Parametric estimation and tests through divergences and the duality technique. *J. Multivar. Anal.* **2009**, *100*, 16–36.
110. Broniatowski, M.; Vajda, I. Several applications of divergence criteria in continuous families. *Kybernetika* **2012**, *48*, 600–636.
111. Toma, A. Optimal robust M-estimators using divergences. *Stat. Probab. Lett.* **2009**, *79*, 1–5.
112. Marazzi, A.; Yohai, V. Optimal robust estimates using the Hellinger distance. *Adv. Data Anal. Classif.* **2010**, *4*, 169–179.
113. Toma, A.; Leoni-Aubin, S. Optimal robust M-estimators using Renyi pseudodistances. *J. Multivar. Anal.* **2010**, *115*, 359–373.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Asymptotic Properties for Methods Combining the Minimum Hellinger Distance Estimate and the Bayesian Nonparametric Density Estimate

Yuefeng Wu <sup>1,\*†</sup> and Giles Hooker <sup>2</sup>

<sup>1</sup> Department of Mathematics and Computer Science, University of Missouri Saint Louis, St. Louis, MO 63121, USA

<sup>2</sup> Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA; gih27@cornell.edu

\* Correspondence: wuyue@umsl.edu; Tel.: +1-314-516-6348

† Current address: ESH 320, One University Blvd. Saint Louis, MO 63121, USA.

Received: 18 October 2018; Accepted: 7 December 2018; Published: 11 December 2018

**Abstract:** In frequentist inference, minimizing the Hellinger distance between a kernel density estimate and a parametric family produces estimators that are both robust to outliers and statistically efficient when the parametric family contains the data-generating distribution. This paper seeks to extend these results to the use of nonparametric Bayesian density estimators within disparity methods. We propose two estimators: one replaces the kernel density estimator with the expected posterior density using a random histogram prior; the other transforms the posterior over densities into a posterior over parameters through minimizing the Hellinger distance for each density. We show that it is possible to adapt the mathematical machinery of efficient influence functions from semiparametric models to demonstrate that both our estimators are efficient in the sense of achieving the Cramér–Rao lower bound. We further demonstrate a Bernstein–von–Mises result for our second estimator, indicating that its posterior is asymptotically Gaussian. In addition, the robustness properties of classical minimum Hellinger distance estimators continue to hold.

**Keywords:** robustness; efficiency; Bayesian nonparametric; Bayesian semi-parametric; asymptotic property; minimum disparity methods; Hellinger distance; Bernstein von Mises theorem

---

## 1. Introduction

This paper develops Bayesian analogs of minimum Hellinger distance methods. In particular, we aim to produce methods that enable a Bayesian analysis to be both robust to unusual values in the data and to retain their asymptotic precision when a proposed parametric model is correct.

All statistical models include assumptions which may or may not be true of the mechanisms producing a given data set. Robustness is a desired property in which a statistical procedure is relatively insensitive to deviations from these assumptions. For frequentist inference, concerns are largely associated with distributional robustness: the shape of the true underlying distribution deviates slightly from the assumed model. Usually, this deviation represents the situation where there are some outliers in the observed data set; see [1] for example. For Bayesian procedures, the deviations may come from the model, prior distribution, or utility function, or some combination thereof. Much of the literature on Bayesian robustness has been concerned with the prior distribution or utility function. By contrast, the focus of this paper is robustness with respect to outliers in a Bayesian context, a relatively understudied form of robustness for Bayesian models. For example, we know that Bayesian models with heavy tailed data distributions are robust with respect to outliers for the case of one single location parameter estimated by many observations. However, as a consequence of the Crámer–Rao lower

bound and the efficiency of the MLE, modifying likelihoods to account for outliers will usually result in a loss of precision in parameter estimates when they are not necessary. The methods we propose, and the study of their robustness properties, will provide an alternative means of making any i.i.d. data distribution robust to outliers that do not lose efficiency when no outliers are present. We speculate that they can be extended beyond i.i.d. data as in [2], but we do not pursue this here.

Suppose we are given the task of estimating  $\theta_0 \in \Theta$  from independent and identically distributed univariate random variables  $X_1, \dots, X_n$ , where we assume each  $X_i$  has density  $f_{\theta_0} \in \mathcal{F} = \{f_\theta : \theta \in \Theta\}$ . Within the frequentist literature, minimum Hellinger distance estimates proceed by first estimating a kernel density  $\hat{g}_n(x)$  and then choosing  $\theta$  to minimize the Hellinger distance  $h(f_\theta, g_n) = [\int \{f_\theta^{1/2}(x) - \hat{g}_n^{1/2}(x)\}^2 dx]^{1/2}$ . The minimum Hellinger distance estimator was shown in [3] to have the remarkable properties of being both robust to outliers and statistically efficient, in the sense of asymptotically attaining the information bound, when the data are generated from  $f_{\theta_0}$ . These methods have been generalized to a class of minimum disparity estimators, based on alternative measures of the difference between a kernel density estimate and a parametric model, which have been studied since then, e.g., [4–8]. While some adaptive M-estimators can be shown to retain both robustness and efficiency, e.g., [9], minimum disparity methods are the only generic methods we are aware of that retain both properties and can also be readily employed within a Bayesian context. In this paper, we only consider Hellinger distance in order to simplify the mathematical exposition; the extension to more general disparity methods can be made following similar developments to those in [5,7].

Recent methodology proposed in [2] suggested the use of disparity-based methods within Bayesian inference via the construction of a “disparity likelihood” by replacing the likelihood function when calculating the Bayesian posterior distribution; they demonstrated that the resulting expected *a posteriori* estimators retain the frequentist properties studied above. These methods first obtain kernel density estimates from data and then calculate the disparity between the estimated density function and the corresponding density functions in the parametric family.

In this paper, we propose the use of Bayesian non-parametric methods instead of the classical kernel methods in applying the minimum Hellinger distance method. One method we proposed is just to replace the kernel density estimate used in classical minimum Hellinger distance estimate by the Bayesian nonparametric expected *a posteriori* density, which we denote by MHB (minimum Hellinger distance method using a Bayesian nonparametric density estimate). The second method combines the minimum Hellinger distance estimate with the Bayesian nonparametric posterior to give a posterior distribution of the parameter of interest. This latter method is our main focus. We show that it is more robust than usual Bayesian methods and demonstrate that it retains asymptotic efficiency, hence the precision of the estimate is maintained. So far as we are aware, this is the first Bayesian method that can be applied generically and retain both robustness and (asymptotic) efficiency. We denote it by BHM (Bayesian inference using a minimum Hellinger distance).

To study the properties of the proposed new methods, we treat both MHB and BMH as special cases of semi-parametric models. The general form of a semi-parametric model has a natural parametrization  $(\theta, \eta) \mapsto P_{\theta, \eta}$ , where  $\theta \in \Theta$  is a Euclidean parameter and  $\eta \in H$  belongs to an infinite-dimensional set. For such models,  $\theta$  is the parameter of primary interest, while  $\eta$  is a nuisance parameter. Asymptotic properties of some of Bayesian semi-parametric models have been discussed in [10]. Our disparity based methods involve parameters in Euclidean space and Hilbert space, with the former being of most interest. However, unlike many semi-parametric models in which  $P_{\theta, \eta} \in \mathcal{P}$  is specified jointly by  $\theta$  and  $\eta$ , in our case, the finite dimensional parameter and the nonparametric density functions are parallel specifications of the data distribution. Therefore, standard methods to study asymptotic properties of semi-parametric models will not apply to the study of disparity-based methods. Nevertheless, considering the problem of estimating  $\psi(P)$  of some function  $\psi : \mathcal{P} \mapsto \mathbb{R}^d$ , where  $\mathcal{P}$  is the space of the probability models  $P$ , semi-parametric models and disparity-based methods can be unified into one framework.

The MHB and BMH methods are introduced in detail in Section 2, where we also discuss some related concepts and results, such as tangent sets, information, consistency, and the specific nonparametric prior that we employ. In Section 3, both MHB and BMH are shown to be efficient, in the sense that asymptotically the variance of the estimate achieves the lower bound of the Cramér–Rao theorem. For MHB, we show that asymptotic normality of the estimate holds, where the asymptotic variance is the inverse of the Fisher information. For BMH, we show that the Bernstein von Mises (BvM) theorem holds. The robustness property and further discussion of these two methods are given in Sections 4 and 5, respectively. A broader discussion is given in Section 6.

## 2. Minimum Hellinger Distance Estimates

Assume that random variables  $X_1, \dots, X_n$  are independent and identically distributed (iid) with density belonging to a specified parametric family  $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ , where all the  $f_\theta$  in the family have the same support, denoted by  $\text{supp}(f)$ . For simplicity, we use  $\mathbb{X}_n$  to denote the random variables  $X_1, \dots, X_n$ . More flexibly, we model  $\mathbb{X}_n \sim g^n$ , where  $g$  is a probability density function with respect to the Lebesgue measure on  $\text{supp}(f)$ . Let  $\mathcal{G}$  denote the collection of all such probability density functions. If the parametric family contains the data-generating distribution, then  $g = f_\theta$  for some  $\theta$ . Formally, we can denote the probability model of the observations in the form of a semi-parametric model  $(\theta, g) \mapsto P_{\theta, g}$ . We aim at estimating  $\theta$  and consider  $g$  as a nuisance parameter, which is typical of semi-parametric models.

Let  $\pi$  denote a prior on  $\mathcal{G}$ , and for any measurable subset  $B \subset \mathcal{G}$ , the posterior probability of  $g \in B$  given  $\mathbb{X}_n$  is

$$\pi(B | \mathbb{X}_n) = \frac{\int_B \prod_{i=1}^n g(X_i) \pi(dg)}{\int_{\mathcal{G}} \prod_{i=1}^n g(X_i) \pi(dg)}.$$

Let  $g_n^* = \int g \pi(dg | \mathbb{X}_n)$  denote the Bayesian nonparametric expected *a posteriori* estimate. Our first proposed method can be described formally as follows:

MHB: Minimum Hellinger distance estimator with Bayesian nonparametric density estimation:

$$\hat{\theta}_1 = \operatorname{argmin}_{\theta \in \Theta} h(f_\theta, g_n^*). \quad (1)$$

This estimator replaces the kernel density estimate in the classical minimum Hellinger distance method introduced in [3] by the posterior expectation of the density function.

For this method, we will view  $\hat{\theta}_1$  as the value at  $g_n^*$  of a functional  $T : \mathcal{G} \mapsto \Theta$ , which is defined via

$$\|f_{T(g)}^{1/2} - g^{1/2}\| = \min_{t \in \Theta} \|f_t^{1/2} - g^{1/2}\| \quad (2)$$

where  $\|\cdot\|$  denotes the  $L_2$  metric. We can also write  $\hat{\theta}_1$  as  $T(g_n^*)$ .

In a more general form, what we estimate is the value  $\psi(P)$  of some functional  $\psi : \mathcal{P} \mapsto \mathbb{R}^d$ , where the  $P$  stands for the common distribution from which data are generated, and  $\mathcal{P}$  is the set of all possible values of  $P$ , which also denotes the corresponding probability model. In the setting of minimum Hellinger distance estimation, the model  $\mathcal{P}$  is set as  $\mathcal{F} \times \mathcal{G}$ ,  $P$  can be specified as  $P_{\theta, g}$ , and  $\psi(P) = \psi(P_{\theta, g}) = \theta$ . For the methods we proposed in this paper, we will focus on the functional  $T : \mathcal{G} \mapsto \Theta$ , for a given  $\mathcal{F}$ , as defined above. Note that the constraint associated with the family  $\mathcal{F}$  is implicitly applied by  $T$ .

Using functional  $T$ , we can also propose a Bayesian method, which assigns nonparametric prior on the density space and gives inference on the unknown parameter  $\theta$  of a parametric family as follows:

BMH: Bayesian inference with minimum Hellinger distance estimation:

$$\pi(\theta | \mathbb{X}_n) = \pi(T(g) | \mathbb{X}_n). \quad (3)$$

A nonparametric prior  $\pi$  on the space  $\mathcal{G}$  and the observation  $\mathbb{X}_n$  leads to the posterior distribution  $\pi(g | \mathbb{X}_n)$ , which can then be converted to the posterior distribution of the parameter  $\theta \in \Theta$  through the functional  $T : \mathcal{G} \mapsto \Theta$ .

In the following subsections, we discuss properties associated with the functional  $T$  as well as the consistency of MHB and BHM, and we provide a detailed example of the random histogram prior that we will employ and its properties that will be used for the discussion of efficiency in Section 2.1.

### 2.1. Tangent Space and Information

In this subsection, we obtain the efficient influence function of the functional  $T$  on the linear span of the tangent set on  $g_0$  and show that the local asymptotic normality (LAN) expansion related to the norm of the efficient influence function attains the Caramér–Rao bound. These results play important roles in showing that BvM holds for the BMH method in the next section.

Estimating the parameter by  $T(g)$  under the assumption  $g \in \mathcal{G}$  uses less information than estimating this parameter for  $g \in \mathcal{G}^* \subset \mathcal{G}$ . Hence, the lower bound of the variance of  $T(g)$  for  $g \in \mathcal{G}$  should be at least the supremum of the lower bounds of all parametric sub-models  $\mathcal{G}^* = \{G_\lambda : \lambda \in \Lambda\} \subset \mathcal{G}$ .

To use mathematical tools such as functional analysis to study the properties of the proposed methods, we introduce some notations and concepts below. Without loss of generality, we consider one-dimensional sub-models  $\mathcal{G}^*$ , which pass through the “true” distribution, denoted by  $G_0$  with density function  $g_0$ . We say a sub-model indexed by  $t$ ,  $\{g_t : 0 < t < \epsilon\} \subset \mathcal{G}$ , is differentiable in quadratic mean at  $t = 0$  if we have that, for some measurable function  $q : \text{supp}(g_0) \mapsto \mathbb{R}$ ,

$$\int \left[ \frac{dG_t^{1/2} - dG_0^{1/2}}{t} - \frac{1}{2} q dG_0^{1/2} \right]^2 \rightarrow 0 \quad (4)$$

where  $G_t$  is the cumulative distribution function associated with  $g_t$ . Functions  $q(x)$ s are known as the score functions associated with each sub-model. The collection of these score functions, which is called a tangent set of the model  $\mathcal{G}$  at  $g_0$  and denoted by  $\dot{\mathcal{G}}_{g_0}$ , is induced by the collection of all sub-models that are differentiable at  $g_0$ .

We say that  $T$  is differentiable at  $g_0$  relative to a given tangent set  $\dot{\mathcal{G}}_{g_0}$ , if there exists a continuous linear map  $\dot{T}_{g_0} : L_2(G_0) \mapsto \mathbb{R}$  such that for every  $q \in \dot{\mathcal{G}}_{g_0}$  and a sub-model  $t \mapsto g_t$  with score function  $q$ , there is

$$\frac{T(g_t) - T(g_0)}{t} \rightarrow \dot{T}_{g_0} q \quad (5)$$

where  $L^2(G_0) = \{q : \text{supp}(g_0) \mapsto \mathbb{R}, \int q^2(x) g_0(x) dx < \infty\}$ . By the Riesz representation theorem for Hilbert spaces, the map  $\dot{T}_{g_0}$  can always be written in the form of an inner product with a fixed vector-valued, measurable function  $\tilde{T}_{g_0} : \text{supp}(g_0) \mapsto \mathbb{R}$ ,

$$\dot{T}_{g_0} q = \langle \tilde{T}_{g_0}, q \rangle_{G_0} = \int \tilde{T}_{g_0} q dG_0.$$

Let  $\tilde{T}_{g_0}$  denote the unique function in  $\overline{\text{lin}}\dot{\mathcal{G}}_{g_0}$ , the closure of the linear span of the tangent set. The function  $\tilde{T}_{g_0}$  is the efficient influence function and can be found as the projection of any other “influence function” onto the closed linear span of the tangent set.

For a sub-model  $t \mapsto g_t$  whose score function is  $q$ , the Fisher information about  $t$  at 0 is  $G_0 q^2 = \int q^2 dG_0$ . In this paper, we use the notation  $Fg$  to denote  $\int g dF$  for a general function  $g$  and distribution  $F$ . Therefore, the “optimal asymptotic variance” for estimating the functional  $t \mapsto T(g_t)$ , evaluated at  $t = 0$ , is greater than or equal to the Caramér–Rao bound

$$\frac{(dT(g_t)/dt)^2}{G_0 q^2} = \frac{\langle \tilde{T}_{g_0}, q \rangle_{G_0}^2}{\langle q, q \rangle_{G_0}}.$$

The supremum of the right-hand side (RHS) of the above expression over all elements of the tangent set is a lower bound for estimating  $T(g)$  given model  $\mathcal{G}$ , if the true model is  $g_0$ . The supremum can be expressed in the norm of the efficient influence function  $\tilde{T}_{g_0}$  by Lemma 25.19 in [11]. The lemma and its proof is quite neat, and we reproduce it here for the completeness of the argument.

**Lemma 1.** Suppose that the functional  $T : \mathcal{G} \mapsto \mathbb{R}$  is differentiable at  $g_0$  relative to the tangent set  $\tilde{\mathcal{G}}_{g_0}$ . Then

$$\sup_{q \in \text{lin}\tilde{\mathcal{G}}_{g_0}} \frac{\langle \tilde{T}_{g_0}, q \rangle_{G_0}^2}{\langle q, q \rangle_{G_0}} = G_0 \tilde{T}_{g_0}^2.$$

**Proof.** This is a consequence of the Cauchy–Schwarz inequality  $(G_0 \tilde{T}_{g_0} q)^2 \leq G_0 \tilde{T}_{g_0}^2 G_0 q^2$  and the fact that, by definition, the efficient influence function,  $\tilde{T}_{g_0}$ , is contained in the closure of  $\text{lin}\tilde{\mathcal{G}}_{G_0}$ .  $\square$

Now we show that functional  $T$  is differentiable under some mild conditions and construct its efficient influence function in the following theorem.

**Theorem 1.** For the functional  $T$  defined in Equation (2), and for  $t \in \Theta \subset \mathbb{R}$ , let  $s_t(x)$  denote  $f_\theta^{1/2}(x)$  for  $\theta = t$ . We assume that there exist  $\dot{s}_t(x)$  and  $\ddot{s}_t(x)$  both in  $L_2$ , such that for  $\alpha$  in a neighborhood of zero,

$$s_{t+\alpha}(x) = s_t(x) + \alpha \dot{s}_t(x) + \alpha u_\alpha(x) \quad (6)$$

$$\dot{s}_{t+\alpha}(x) = \dot{s}_t(x) + \alpha \ddot{s}_t(x) + \alpha v_\alpha(x), \quad (7)$$

where  $u_\alpha$  and  $v_\alpha$  converge to zero as  $\alpha \rightarrow 0$ . Assuming  $T(g_0) \in \text{int}(\Theta)$ , the efficient influence function of  $T$  is

$$\tilde{T}_{g_0} = \left( - \left[ \int \ddot{s}_{T(g_0)}(x) g_0^{1/2}(x) dx \right]^{-1} + a_t \right) \frac{\dot{s}_{T(g_0)}(x)}{2 g_0^{1/2}(t)} \quad (8)$$

where  $a_t$  converges to 0 as  $t \rightarrow 0$ . In particular, for  $g_0 = f_\theta$ ,

$$\tilde{T}_{f_\theta} = \left( - \left[ \int \ddot{s}_\theta(x) s_\theta(x) dx \right]^{-1} + a_t \right) \frac{\dot{s}_\theta(x)}{2 s_\theta(x)}. \quad (9)$$

**Proof.** Let the  $t$ -indexed sub-model be

$$g_t := (1 + tq(x))g_0(x)$$

where  $q(x)$  satisfies  $\int q(x)g_0(x)dx = 0$  and  $q \in L_2(g_0)$ . By direct calculation, we see that  $q$  is the score function associated with such a sub-model at  $t = 0$  in the sense of Equation (4) and thus the collection of  $q$  is the maximal tangent set.

By the definition of  $T$ ,  $T(g_0)$  maximizes  $\int s_t(x)g_0^{1/2}(x)dx$ . From Equation (6), we have that

$$\lim_{\alpha \rightarrow 0} \alpha^{-1} \int [s_{t+\alpha}(x) - s_t(x)]g_0^{1/2}(x)dx = \int \dot{s}_t(x)g_0^{1/2}(x)dx. \quad (10)$$

Since  $T(g_0) \in \text{int}(\Theta)$ , we have that

$$\int \dot{s}_{T(g_0)}(x)g_0^{1/2}(x)dx = 0. \quad (11)$$

Similarly,  $\int \dot{s}_{T(g_t)}(x)g_t^{1/2}(x)dx = 0$ . Using Equation (7) to substitute  $\dot{s}_{T(g_t)}$ , we have that

$$0 = \int [\dot{s}_{T(g_0)}(x) + \ddot{s}_{T(g_0)}(x)(T(g_t) - T(g_0)) + v_t(x)(T(g_t) - T(g_0))]g_t^{1/2}(x)dx$$

where  $v_t(x)$  converge in  $L_2$  to zero as  $t \rightarrow 0$  since  $T(g_t) \rightarrow T(g_0)$ . Thus,

$$\begin{aligned} & \lim_{t \rightarrow 0} \frac{1}{t} [T(g_t) - T(g_0)] \\ &= - \lim_{t \rightarrow 0} \frac{1}{t} \left[ \int (\dot{s}_{T(g_0)}(x) + v_t(x)) g_t^{1/2}(x) dx \right]^{-1} \int \dot{s}_{T(g_0)}(x) g_t^{1/2}(x) dx \\ &= \lim_{t \rightarrow 0} \frac{1}{t} \left( - \left[ \int (\dot{s}_{T(g_0)}(x)) g_0^{1/2}(x) dx \right]^{-1} + a_t \right) \int \dot{s}_{T(g_0)}(x) (g_t^{1/2}(x) - g_0^{1/2}(x)) dx \\ &= \left( - \left[ \int (\dot{s}_{T(g_0)}(x)) g_0^{1/2}(x) dx \right]^{-1} + a_t \right) \int \frac{\dot{s}_{T(g_0)}(x)}{2g_0^{1/2}(x)} q(x) g_0(x) dx. \end{aligned}$$

Since by the definition of  $\tilde{T}$ , which requires  $\int \tilde{T}_{g_0} g_0(x) dx = 0$ , we have that

$$\begin{aligned} \tilde{T}_{g_0} &= \left( - \left[ \int \dot{s}_{T(g_0)}(x) g_0^{1/2}(x) dx \right]^{-1} + a_t \right) \left( \frac{\dot{s}_{T(g_0)}(x)}{2g_0^{1/2}(x)} - \int \frac{\dot{s}_{T(g_0)}(x)}{2} g_0^{1/2}(x) dx \right) \\ &= \left( - \left[ \int \dot{s}_{T(g_0)}(x) g_0^{1/2}(x) dx \right]^{-1} + a_t \right) \frac{\dot{s}_{T(g_0)}(x)}{2g_0^{1/2}(x)}. \end{aligned}$$

By the same argument we can show that, when  $g_0 = f_\theta$ , Equation (9) holds.  $\square$

Some relatively accessible conditions under which Equations (6) and (7) hold are given by Lemmas 1 and 2 in [3]. We do not repeat them here.

Now we can expand  $T$  at  $g_0$  as

$$T(g) - T(g_0) = \langle \frac{g - g_0}{g_0}, \tilde{T}_{g_0} \rangle_{G_0} + \tilde{r}(g, g_0) \quad (12)$$

where  $\tilde{T}$  is given in Theorem 1 and  $\tilde{r} = 0$ .

## 2.2. Consistency of MHB and BMH

Since  $T(g)$  may have more than one value, the notation  $T(g)$  is used to denote any arbitrary one of the possible values. In [3], the existence, continuity in Hellinger distance, and uniqueness of functional  $T$  are ensured under the following condition:

**A1** (i)  $\Theta$  is compact, (ii)  $\theta_1 \neq \theta_2$  implies  $f_{\theta_1} \neq f_{\theta_2}$  on a set of positive Lebesgue measures, and (iii), for almost every  $x$ ,  $f_\theta(x)$  is continuous in  $\theta$ .

When a Bayesian nonparametric density estimator is used, we assume the posterior consistency:

**A2** For any given  $\epsilon > 0$ ,  $\pi\{g : h(g, f_{\theta_0}) > \epsilon \mid \mathbb{X}_n\} \rightarrow 0$  in probability.

Under Conditions A1 and A2, consistency holds for MHB and BMH.

**Theorem 2.** Suppose that Conditions A1 and A2 hold, then

1.  $\|g_n^{*1/2} - f_{\theta_0}^{1/2}\|^2 \rightarrow 0$  in probability,  $T(g_n^*) \rightarrow T(f_{\theta_0})$  in probability, and hence  $\hat{\theta}_1 \rightarrow \theta_0$  in probability;
2. For any given  $\epsilon > 0$ ,  $\pi(|\theta - \theta_0| > \epsilon \mid \mathbb{X}_n) \rightarrow 0$  in probability.

**Proof.** Part 1: To show that  $\|g_n^{*1/2} - f_{\theta_0}^{1/2}\|^2 \rightarrow 0$  in probability, which is equivalent to showing that  $\int (\int g\pi(dg | \mathbb{X}_n)^{1/2} - f_{\theta_0}^{1/2})^2 dx \rightarrow 0$  in probability, it is sufficient to show that  $\int |\int g\pi(dg | \mathbb{X}_n) - f_{\theta_0}| dx \rightarrow 0$  in probability, since  $h^2(f, g) \leq \|f - g\|_1$ . We have that

$$\begin{aligned} \int \left| \int g\pi(dg | \mathbb{X}_n) - f_{\theta_0} \right| dx &= \int \left| \int (g - f_{\theta_0})\pi(dg | \mathbb{X}_n) \right| dx \\ &\leq \int \int |g - f_{\theta_0}| \pi(dg | \mathbb{X}_n) dx \\ &= \int \int |g - f_{\theta_0}| dx \pi(dg | \mathbb{X}_n) \\ &\leq \int \sqrt{2}h(g, f_{\theta_0}) \pi(dg | \mathbb{X}_n). \end{aligned}$$

Note that the change of order of integration is due to Fubini's theorem and the last inequality is due to  $\|f - g\|_1 \leq \sqrt{2}h(f, g)$ . Split the integral on the right-hand side of the above expression into two parts:

$$\int_{\mathcal{A}} \sqrt{2}h(g, f_{\theta_0}) \pi(dg | \mathbb{X}_n) + \int_{\mathcal{A}^c} \sqrt{2}h(g, f_{\theta_0}) \pi(dg | \mathbb{X}_n)$$

where  $\mathcal{A} = \{g : h(g, f_{\theta_0}) \leq \epsilon\}$  for any given  $\epsilon > 0$ . The first term is bounded by  $\epsilon$  by construction. By Condition A1, the posterior measure of  $\mathcal{A}^c$  goes to 0 in probability as  $n \rightarrow \infty$ . Since Hellinger distance is bounded by 2, so does the second term above. This completes the proof for  $\|g_n^{*1/2} - f_{\theta_0}^{1/2}\|^2 \rightarrow 0$  in probability.

To show  $T(g_n^*) \rightarrow T(f_{\theta_0})$  and  $\hat{\theta}_1 \rightarrow \theta_0$  in probability, we need that the functional  $T$  is continuous and unique at  $f_{\theta_0}$ , which is proved by Theorem 1 in [3] under Condition A1.

Part 2: By Condition A1 and Theorem 1 in [3], the functional  $T$  is continuous and unique at  $f_{\theta_0}$ . Hence, for any given  $\epsilon > 0$ , there exist  $\delta > 0$  such that  $|T(g) - T(f_{\theta_0})| < \epsilon$  when  $h(g, f_{\theta_0}) < \delta$ . By Condition A2, we have that  $\pi(h(g, f_{\theta_0}) < \delta) \rightarrow 1$ , which implies that  $\pi(|\theta - \theta_0| < \epsilon) \rightarrow 1$  in probability.  $\square$

It should be noted that, if we change the  $\epsilon$  in Condition A2 to  $\epsilon_n$ , a sequence converging to 0, then we can apply the results for the concentration rate of the Bayesian nonparametric density estimation here. However, such an approach cannot lead to the general "efficiency" claim, no matter in the form of rate of concentration or asymptotic normality. There are two reasons for this. First, the rate of concentration for Bayesian nonparametric posterior is about  $n^{-2/5}$  for a rather general situation and  $(\log n)^a \times n^{-1/2}$ , where  $a > 0$ , for some special cases (see [12–14]). This concentration rate is not sufficient in many situations to directly imply that the concentration of the corresponding parametric estimates achieves the lower bound of the variance given in the Cramér–Rao theorem. Second, the Hellinger distances between pairs of densities as functions of parameters vary among different parametric families. Therefore, obtaining the rate of concentration in parameters from the rate of convergence in the densities cannot be generally applied to different distribution families.

It should also be noted that, although  $\Theta$  is required to be compact in Condition A1, Theorem 2 is useful for a  $\Theta$  that is not compact, as long as the parametric family  $f_\theta : \theta \in \Theta$  can be re-parameterized where the space of new parameters can be embedded within a compact set. An example of re-parameterizing a general location-scale family with parameters  $\mu \in \mathbb{R}$  and  $\sigma \in \mathbb{R}^+$  to a family with parameters  $t_1 = \tan^{-1}(\mu)$  and  $t_2 = \tan^{-1}(\sigma)$ , where  $\Theta_{(t_1, t_2)} = (-\pi/2, \pi/2) \times (0, \pi/2)$  and  $\Theta \subset \bar{\Theta} = [-\pi/2, \pi/2] \times [0, \pi/2]$ , is discussed in [3], and the conclusions of Theorem 1 in [3] is still valid for a location-scale family. Therefore, Theorem 2 remains valid for the same type of the families, whose parameter space may not be compact and for the same reasons; the compactness requirement stated in the theorem is mainly for mathematical simplicity.

### 2.3. Prior on Density Functions

We introduce a random histogram as an example for priors used in Bayesian nonparametric density estimation. It can be seen as a simplified version of a Dirichlet process mixture (DPM) prior, which is commonly used in practice. Both DPM and random histogram are mixture densities. While DPM uses a Dirichlet process to model the weights within an infinite mixture of kernels, the random histogram prior only has a finite number of components. Another difference is that, although we specify the form of the kernel function for DPM, the kernel function could be any density function in general, while the random histogram uses only the uniform density as its mixing kernel. Nevertheless, the limit on the finite number of the mixing components is not that important in practice, since the Dirichlet process will always be truncated in computation. In the next section, we will verify that the random histogram satisfies the conditions that are needed for our proposed methods to be efficient. On the other hand, although we believe that DPM should also lead to efficiency, the authors are unaware of the theoretical results or tools required to prove it. This is mostly due to the flexibility of DPM, which in turn significantly increases the mathematical complexity of the analysis.

For any  $k \in \mathbb{N}$ , denote the set of all regular  $k$  bin histograms on  $[0, 1]$  by  $\mathcal{H}_k = \{f \in L^2([0, 1]) : m(x) = \sum_{j=1}^k f_j \mathbf{1}_{I_j}(x), f_j \in \mathbb{R}, j = 1, \dots, k\}$ , where  $I_j = [(j-1)/k, j/k)$ . Denote the unit simplex in  $\mathbb{R}^k$  by  $\mathcal{S}_k = \{\omega \in [0, 1]^k : \sum_{j=1}^k \omega_j = 1\}$ . The subset of  $\mathcal{H}_k$ ,  $\mathcal{H}_k^1 = \{f \in L^2(\mathbb{R}), f(x) = f_{\omega, k} = k \cdot \sum_{j=1}^k \omega_j \mathbf{1}_{I_j}(x), (\omega_1, \dots, \omega_k) \in \mathcal{S}_k\}$ , denotes the collection of densities on  $[0, 1]$  in the form of a histogram.

The set  $\mathcal{H}_k$  is a closed subset of  $L_2[0, 1]$ . For any function  $f \in L_2[0, 1]$ , denote its projection in the  $L_2$  sense on  $\mathcal{H}_k$  by  $f_{[k]}$ , where  $f_{[k]} = k \sum_{j=1}^k \mathbf{1}_{I_j} \int_{I_j} f$ .

We assign priors on  $\mathcal{H}_k^1$  via  $k$  and  $(\omega_1, \dots, \omega_k)$  for each  $k$ . A degenerate case is to let  $k = K_n = o(n)$ . Otherwise, let  $p_k$  be a distribution on positive integers, where

$$k \sim p_k, e^{-b_1 k \log(k)} \leq p_k(k) \leq e^{-b_2 k \log(k)} \quad (13)$$

for all  $k$  large enough and some  $0 < b_1 < b_2 < \infty$ . For example, Condition (13) is satisfied by the Poisson distribution, which is commonly used in Bayesian nonparametric models.

Conditionally on  $k$ , we consider a Dirichlet prior on  $\omega = \{\omega_1, \dots, \omega_k\}$ :

$$\omega \sim \mathcal{D}(\alpha_{1,k}, \dots, \alpha_{k,k}), \quad c_1 k^{-a} \leq \alpha_{j,k} \leq c_2 \quad (14)$$

for some fixed constants  $a, c_1, c_2 > 0$  and any  $1 \leq j \leq k$ . For posterior consistency, we need the following condition:

$$\sup_{k \in \mathcal{K}_n} \sum_{j=1}^k \alpha_{j,k} = o(\sqrt{n}) \quad (15)$$

where  $\mathcal{K}_n \subset \{1, 2, \dots, \lfloor n/(\log n)^2 \rfloor\}$ .

The consistency result of this prior is given by Proposition 1 in the supplement to [15]. For  $n \geq 2, k \geq 1, M > 0$ , let

$$A_{n,k}(M) = \{g \in \mathcal{H}_k^1, h(g, g_{0,[k]}) < M \epsilon_{n,k}\} \quad (16)$$

where  $\epsilon_{n,k}^2 = k \log n / n$  denote a neighborhood of  $g_{0,[k]}$ , and we have that

- (a) there exist  $c, M > 0$  such that

$$P_0 \left[ \exists k \leq \frac{n}{\log n}; \pi[g \notin A_{n,k}(M) \mid \mathbb{X}_n, k] > e^{-ck \log n} \right] = o(1). \quad (17)$$

- (b) Suppose  $g_0 \in \mathcal{C}^\beta$  with  $0 < \beta \leq 1$ , if  $k_n(\beta) = (n \log n)^{1/(2\beta+1)}$  and  $\epsilon_n(\beta) = k_n(\beta)^{-\beta}$ , then, for  $k_1$  and a sufficiently large  $M$ ,

$$\pi[h(g_0, g) \leq M\epsilon_n(\beta); k \leq k_1 k_n(\beta) | \mathbb{X}_n] = 1 + o_p(1), \quad (18)$$

where  $\mathcal{C}^\beta$  denotes the class of  $\beta$ -Hölder functions on  $[0, 1]$ .

This means that the posterior of the density function concentrates around the projection  $g_{0[k]}$  of  $g_0$  and around  $g_0$  itself in terms of the Hellinger distance. We can easily conclude that  $\pi(\mathcal{K}_n | \mathbb{X}_n) = 1 + o(1)$  from Equation (18) for  $g_0 \in \mathcal{C}^\beta$ .

It should be noted that, although the priors we defined above are on the densities on  $[0, 1]$ , this is for mathematical simplicity, which could easily be extended to the space of probability densities on any given compact set. Further, transformations of  $\mathbb{X}_n$ , similar to those discussed at the end of Section 2.2, can extend the analysis to the real line (refer to [3,16] for more example and details).

### 3. Efficiency

We say that both MHB and BMH methods are efficient if the lower bound of the variance of the estimate, in the sense of Cramér and Rao's theorem, is achieved.

#### 3.1. Asymptotic Normality of MHB

Consider the maximal tangent set at  $g_0$ , which is defined as  $\mathcal{H}_T = \{q \in L^2(g_0), \int q g_0 = 0\}$ . Denote the inner product on  $\mathcal{H}_T$  by  $\langle q_1, q_2 \rangle_L = \int q_1 q_2 g_0$ , which induces the L-norm as

$$\|q\|_L^2 = \int_0^1 (q - G_0 g)^2 g_0. \quad (19)$$

Note that the inner product  $\langle \cdot, \cdot \rangle_L$  is equivalent to the inner product introduced in Section 2.1, and the induced L-norm corresponds to the local asymptotic normality (LAN) expansion. Refer to [17] and Theorem 25.14 in [11] for more details.

With functional  $T$  and priors on  $g$  defined in the previous section, Theorem 3 shows that the MHB method is efficient when the parametric family contains the true model.

**Theorem 3.** Let two priors  $\pi_1$  and  $\pi_2$  be defined by Equations (13)–(14) and let a prior on  $k$  be either a Dirac mass at  $k = K_n = n^{1/2}(\log n)^{-2}$  for  $\pi_1$  or  $k \sim \pi_k$  given by Equation (13) for  $\pi_2$ . Then the limit distribution of  $n^{1/2}[T(g_n^*) - T(g_0)]$  under  $g_0$  as  $n \rightarrow \infty$  is  $\text{Norm}(0, \|\tilde{T}_{g_0}\|_L^2)$ , where  $\|\tilde{T}_{g_0}\|_L^2 = I(\theta_0)^{-1}$  when  $g_0 = f_{\theta_0}$ .

**Proof.** To prove this result, we verify Lemma 25.23 in [11], which is equivalent to showing that

$$\sqrt{n}(T(g_n^*) - T(g_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{T}_{g_0}(X_i) + o_p(1).$$

By the consistency result provided for priors  $\pi_1$  and  $\pi_2$  in the previous section, we consider only  $g_n^* \in A_{n,k}$  for an  $n$  that is sufficiently large. Then by Equation (12) we have that

$$\sqrt{n}(T(g_n^*) - T(g_0)) = \sqrt{n} \left\langle \frac{g_n^* - g_0}{g_0}, \tilde{T}_{g_0} \right\rangle_L + o_p(1).$$

Therefore, showing

$$\sqrt{n} \int_0^1 (g_n^*(x) - g_0(x)) \tilde{T}_{g_0}(x) dx = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{T}_{g_0}(X_i) + o_p(1)$$

will complete the proof. Due to  $\int_0^1 g_0(x) \tilde{T}_{g_0}(x) dx = 0$ , we now need to show that  $\int_0^1 g_n^*(x) \tilde{T}_{g_0}(x) dx = (1/n) \sum_{i=1}^n \tilde{T}_{g_0}(X_i) + o_p(1)$ . By the law of large numbers, we have that  $\frac{1}{n} \sum_{i=1}^n \tilde{T}_{g_0}(X_i) - G_0 \tilde{T}_{g_0} = o_p(1)$ , and  $\int_0^1 g_n^*(x) \tilde{T}_{g_0}(x) dx - G_0 \tilde{T}_{g_0} = o_p(1)$  due to the posterior consistency demonstrated above. Therefore, we have that

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \tilde{T}_{g_0}(X_i) - \int_0^1 g_n^*(x) \tilde{T}_{g_0}(x) dx \\ &= \frac{1}{n} \sum_{i=1}^n \tilde{T}_{g_0}(X_i) - \int_0^1 g_n^*(x) \tilde{T}_{g_0}(x) dx + \int_0^1 g_n^*(x) \tilde{T}_{g_0}(x) dx - G_0 \tilde{T}_{g_0} \\ &= o_p(1). \end{aligned}$$

□

### 3.2. The Bernstein von Mises Theorem for BMH

Theorem 2.1 in [15] yielded a general result and approach to show that the BvM Theorem holds for smooth functionals in some semi-parametric models. The theorem shows that, under the continuity and consistency condition, the moment generating function (MGF) of the parameter endowed with a posterior distribution can be calculated approximately through the local asymptotic normal (LAN) expansion, and its convergence to an MGF of some normal random variable can then be shown under some assumptions on the statistical model.

We will show that the BvM theorem holds for the BMH Method via Theorem 4. The result also shows that the approach given in [15] can be applied not only to simple examples but also to relatively complicated frameworks. To prove it, we introduced Lemma 2, which is modified from Proposition 1 in [15], the proof of which was not given explicitly in the original paper.

For mathematical simplicity, we assume that the true density  $f_{\theta_0}$  belongs to the set  $\mathcal{F}$ , which is restricted to the space of all densities that are bounded away from 0 and  $\infty$  on  $[0, 1]$ . As noted above, the compactness of the domain can be relaxed by considering transformations of the parameters and random variables.

To state the lemma, we need several more notations. Assume that the functional  $T$  satisfies Equation (12) with bounded efficient influence function  $\tilde{T}_{g_0} \neq 0$ . We denote  $\tilde{T}_{g_0}$  by  $\tilde{T}$ , where  $\tilde{T}_{[k]}$  denotes the projection of  $\tilde{T}$  on  $\mathcal{H}_k$ . For  $k \geq 1$ , let

$$\begin{aligned} \hat{T}_k &= T(g_{0[k]}) + \frac{\mathbb{G}_n \tilde{T}_{[k]}}{\sqrt{n}}, \quad V_k = \|\tilde{T}_{[k]}\|_L^2 \\ \hat{T} &= T(g_0) + \frac{\mathbb{G}_n \tilde{T}}{\sqrt{n}}, \quad V = \|\tilde{T}\|_L^2 \end{aligned} \quad (20)$$

and denote

$$\mathbb{G}_n(g) = W_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(x_i) - G_0(g)]. \quad (21)$$

**Lemma 2.** Let  $g_0$  belong to  $\mathcal{G}$ , let the prior  $\pi$  be defined as in Section 2.3, and let Conditions (13, 14, 15) be satisfied. Consider estimating a functional  $T(g)$ , differentiable with respect to the tangent set  $\mathcal{H}_T := \{q \in L^2(g_0), \int_{[0,1]} q g_0 = 0\} \subset \mathcal{H} = L^2(g_0)$ , with efficient influence function  $\tilde{T}_{g_0}$  bounded on  $[0, 1]$ , and with  $\tilde{r}$  defined in Equation (12), for  $\mathcal{H}_n$  as introduced in Equation (15). If

$$\max_{k \in \mathcal{H}_n} \left| \|\tilde{T}_{[k]}\|_L^2 - \|\tilde{T}\|_L^2 \right| = o_p(1) \quad (22)$$

$$\max_{k \in \mathcal{H}_n} \mathbb{G}_n(\tilde{T}_{[k]} - \tilde{T}) = o_p(1) \quad (23)$$

$$\sup_{k \in \mathcal{K}_n} \sup_{g \in A_{n,k}(M)} \sqrt{n} \tilde{r}(g, g_0) = o_p(1) \quad (24)$$

for any  $M > 0$  and  $A_{n,k}(M)$  defined as in (16), as  $n \rightarrow \infty$ , and

$$\max_{k \in \mathcal{K}_n} \sqrt{n} \left| \int (\tilde{T} - \tilde{T}_{[k]})(g - g_0) \right| = o(1), \quad (25)$$

then the BvM theorem for the functional  $T$  holds.

**Proof.** To show that BvM holds is to show that the posterior distribution converges to a normal distribution. If we have that

$$\begin{aligned} \pi[\sqrt{n}(T - \hat{T}_k) \leq z \mid \mathbb{X}_n] &= \sum_{k \in \mathcal{K}_n} \pi[k \mid \mathbb{X}_n] \pi[\sqrt{n}(T - \hat{T}) \leq z + \sqrt{n}(\hat{T} - \hat{T}_k) \mid \mathbb{X}_n, k] + o_p(1) \\ &= \sum_{k \in \mathcal{K}_n} \pi[k \mid \mathbb{X}_n] \Phi\left(\frac{z + \sqrt{n}(\hat{T} - \hat{T}_k)}{\sqrt{V_k}}\right) + o_p(1), \end{aligned} \quad (26)$$

then the proof will be completed by showing that the RHS of Equation (26) reduces from the mixture of normal to the target law  $N(0, V)$ .

By Condition (22), we have that  $V_k$  goes to  $V$  uniformly for  $k \in \mathcal{K}_n$ . Due to the definition of  $\tilde{T}$  and the Lemma 4 result (iii) in the supplement of [15], we have that

$$\begin{aligned} \sqrt{n}(\hat{T} - \hat{T}_k) &= \sqrt{n}\left(T(g_0) - T(g_{[k]})\right) + \mathbb{G}_n(\tilde{T} - \tilde{T}_{[k]}) \\ &= \sqrt{n} \int \tilde{T}(g_{0[k]} - g_0) + \mathbb{G}_n(\tilde{T}_{[k]} - \tilde{T}) + o_p(1) \\ &= \sqrt{n} \int (\tilde{T} - \tilde{T}_{[k]})(g_{0[k]} - g_0) + \mathbb{G}_n(\tilde{T}_{[k]} - \tilde{T}) + o_p(1). \end{aligned}$$

By Conditions (25) and (23), the last line converges to 0 uniformly for  $k \in \mathcal{K}_n$ .

Therefore, showing that for any given  $k$ , Equation (26) holds will complete the proof. We prove this by showing that the MGF (Laplace transformation) of the posterior distribution of the parameter of interest converges to the MGF of some normal distribution, which implies that the posterior converges to the normal distribution weakly by Lemmas 1 and 2 in supplement to [15] or Theorem 2.2 in [18].

First, consider the deterministic  $k = K_n$  case. We calculate the MGF as

$$E[e^{t\sqrt{n}(T(g) - \hat{T}(g_{0[k]}))} \mid \mathbb{X}_n, A_n] = \frac{\int_{A_n} e^{t\sqrt{n}(T(g) - \hat{T}(g_{0[k]})) + l_n(g) - l_n(g_{0[k]})} d\pi(g)}{\int_{A_n} e^{l_n(g) - l_n(g_{0[k]})} d\pi(g)} \quad (27)$$

where  $l_n(g)$  is the log-likelihood for given  $g$  and  $\mathbb{X}_n$ . Based on the LAN expansion of the log-likelihood and the smoothness of the functional, the exponent in the numerator on the RHS of the equation can be transformed with respect to  $\tilde{T}_{(k)} = \tilde{T}_{[k]} - \int \tilde{T}_{[k]} g_{0[k]}$

$$\begin{aligned} t\sqrt{n}(T(g) - \hat{T}_k) + l_n(g) - l_n(g_{0[k]}) \\ &= t\sqrt{n} \left( T(g) - T(g_{0[k]}) - \frac{\mathbb{G}_n \tilde{T}_{[k]}}{\sqrt{n}} \right) + l_n(g) - l_n(g_{0[k]}) \\ &= t\sqrt{n} \left( \left\langle \log \frac{g}{g_{0[k]}} - \int \log \frac{g}{g_{0[k]}} g_{0[k]}, \tilde{T}_{[k]} \right\rangle_L + \mathcal{B}(g, g_{0[k]}) + \tilde{r}(g, g_{0[k]}) - \frac{\mathbb{G}_n \tilde{T}_{[k]}}{\sqrt{n}} \right) \\ &\quad - \frac{1}{2} \left\| \sqrt{n} \log \frac{g}{g_{0[k]}} \right\|_L^2 + W_n \left( \sqrt{n} \log \frac{g}{g_{0[k]}} \right) + R_{n,k}(g, g_{0[k]}) \end{aligned}$$

where  $\mathcal{B}(g, g_0) = \int_0^1 [\log(g/g_0) - (g - g_0)/g_0](x) \tilde{T}_{g_0}(x) g_0(x) dx$ . Note that  $\mathbb{G}_n = W_n$  and add a term of  $(t^2/2) \|\tilde{T}_{(k)}\|_L^2$ . Re-arranging the RHS expression above, we have

$$\begin{aligned} & t\sqrt{n}(T(g) - \hat{T}_k) + l_n(g) - l_n(g_{0,[k]}) \\ &= -\frac{n}{2} \left\| \log \frac{g}{g_{0,[k]}} - \frac{t}{\sqrt{n}} \tilde{T}_{(k)} \right\|_{L,k}^2 + \sqrt{n} W_n \left( \log \frac{g}{g_{0,[k]}} - \frac{t}{\sqrt{n}} \tilde{T}_{(k)} \right) \\ &\quad + \frac{t^2}{2} \|\tilde{T}_{(k)}\|_{L,k}^2 + t\sqrt{n}\mathcal{B}_{n,k} + R_{n,k}(g, g_{0,[k]}) + \tilde{r}(g, g_{0,[k]}) \\ &= -\frac{n}{2} \left\| \log \frac{ge^{-\frac{t}{\sqrt{n}}\tilde{T}_{(k)}}}{g_{0,[k]}} \right\|_{L,k}^2 + \sqrt{n} W_n \left( \log \frac{ge^{-\frac{t}{\sqrt{n}}\tilde{T}_{(k)}}}{g_{0,[k]}} \right) + \frac{t^2}{2} \|\tilde{T}_{(k)}\|_{L,k}^2 \\ &\quad + t\sqrt{n}\mathcal{B}_{n,k} + R_{n,k}(g, g_{0,[k]}) + \tilde{r}(g, g_{0,[k]}). \end{aligned}$$

This is because the cross term in calculating the first term in the second line above is equal to the inner product term in the equation above it.

Let  $g_{t,k} = ge^{-\frac{t}{\sqrt{n}}\tilde{T}_{(k)}} / Ge^{-\frac{t}{\sqrt{n}}\tilde{T}_{(k)}}$ , the RHS of the above equation can be written as

$$\frac{t^2}{2} \|\tilde{T}_{(k)}\|_{L,k}^2 + l_n(g_{t,k}) - l_n(g_{0,[k]}) + o(1). \quad (28)$$

Substituting the corresponding terms on the RHS of Equation (27) by (28), we have that

$$E[e^{t\sqrt{n}(T(g) - \hat{T}(g_{0,[k]}))} | \mathbb{X}_n, A_n] = e^{(t^2/2)\|\tilde{T}_{(k)}\|_{L,k}^2 + o(1)} \times \frac{\int_{A_{n,k}} e^{l_n(g_{t,k}) - l_n(g_{0,[k]})} d\pi_k(g)}{\int_{A_{n,k}} e^{l_n(g) - l_n(g_{0,[k]})} d\pi_k(g)}. \quad (29)$$

Notice that the integration in the denominator of the second term is an expectation based on a Dirichlet distribution on  $\omega$  as described in Equation (14) and that  $g_{t,k} = k \sum_{j=1}^k \zeta_j \mathbf{1}_{I_j}$ , where

$$\zeta_j = \frac{\omega_j \gamma_j^{-1}}{\sum_{j=1}^k \omega_j \gamma_j^{-1}} \quad (30)$$

with  $\gamma_j = e^{t\tilde{T}_j/\sqrt{n}}$  and  $\tilde{T}_j := k \int_{I_j} \tilde{T}_{(k)}$ . Let  $S_{\gamma^{-1}(\omega)} = \sum_{j=1}^k \omega_j \gamma_j^{-1}$ , by (30). We then have  $S_{\gamma}^{-1}(\zeta) = S_{\gamma^{-1}}(\omega)$ . Now using these notations,

$$\begin{aligned} \frac{\int_{A_{n,k}} e^{l_n(g_{t,k}) - l_n(g_{0,[k]})} d\pi_k(g)}{\int_{A_{n,k}} e^{l_n(g) - l_n(g_{0,[k]})} d\pi_k(g)} &= \frac{\int_{A_{n,k}} e^{l_n(g_{t,k}) - l_n(g_{0,[k]})} \prod_{j=1}^k \omega_j^{\alpha_{j,k}-1} / B(\alpha) d\omega}{\int_{A_{n,k}} e^{l_n(g) - l_n(g_{0,[k]})} \prod_{j=1}^k \omega_j^{\alpha_{j,k}-1} / B(\alpha) d\omega} \\ &= \frac{\int_{A_{n,k}} e^{l_n(k \sum_{j=1}^k \frac{\omega_j \gamma_j^{-1}}{\sum_{j=1}^k \omega_j \gamma_j^{-1}} \mathbf{1}_{I_j}) - l_n(g_{0,[k]})} \prod_{j=1}^k \omega_j^{\alpha_{j,k}-1} d\omega}{\int_{A_{n,k}} e^{l_n(k \sum_{j=1}^k \omega_j \mathbf{1}_{I_j}) - l_n(g_{0,[k]})} \prod_{j=1}^k \omega_j^{\alpha_{j,k}-1} d\omega} \\ &= \frac{\int_{A_{n,k}} e^{l_n(k \sum_{j=1}^k \zeta_j \mathbf{1}_{I_j}) - l_n(g_{0,[k]})} \Delta_{\zeta} \prod_{j=1}^k [\gamma_j \zeta_j S_{\gamma}^{-1}(\zeta)]^{\alpha_{j,k}-1} d\zeta}{\int_{A_{n,k}} e^{l_n(k \sum_{j=1}^k \omega_j \mathbf{1}_{I_j}) - l_n(g_{0,[k]})} \prod_{j=1}^k \omega_j^{\alpha_{j,k}-1} d\omega} \end{aligned} \quad (31)$$

where  $\Delta_{\zeta} = S_{\gamma}^{-k}(\zeta) \prod_{j=1}^k \gamma_j$  is the Jacobian of the change of variable,  $(\omega_1, \dots, \omega_{k-1}) \rightarrow (\zeta_1, \dots, \zeta_{k-1})$ , which is given in Lemma 5 in supplement of [15], and  $B(\alpha) = \prod_{i=1}^k \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^k \alpha_i)$  is the constant for normalizing Dirichlet distribution.

Notice that, over the set  $A_{n,k}$ ,

$$\begin{aligned} \prod_{j=1}^k [\gamma_j S_\gamma^{-1}(\zeta)]^{\alpha_{j,k}-1} \Delta_\zeta &= S_\gamma(\zeta)^{-\sum_{j=1}^k \alpha_{j,k}} \gamma_j^{\sum_{j=1}^k \alpha_{j,k}} \\ &= S_\gamma(\zeta)^{-\sum_{j=1}^k \alpha_{j,k}} e^{t \sum_{j=1}^k \alpha_{j,k} T_j / \sqrt{n}} \\ &= e^{t \sum_{j=1}^k \alpha_{j,k} T_j / \sqrt{n}} \left( 1 - \frac{t}{\sqrt{n}} \int_0^1 \bar{T}_{(k)}(g - g_0) + O(n^{-1}) \right)^{\sum_{j=1}^k \alpha_{j,k}}, \end{aligned} \quad (32)$$

since

$$S_{\gamma^{-1}}(\omega) = \int_0^1 e^{-t\bar{T}_{(k)}(x)/\sqrt{n}} g_{[k]}(x) dx = 1 - \frac{t}{\sqrt{n}} \int_0^1 \bar{T}_{(k)}(g_{[k]} - g_0) + O(n^{-1})$$

by Taylor's expansion. Expression (32) converges to 1 under Condition (15), so Expression (31) converges to

$$\frac{\int_{A_{n,k}} e^{l_n(k \sum_{j=1}^k \zeta_j \mathbb{1}_{I_j}) - l_n(g_{0[k]})} \prod_{j=1}^k \zeta_j^{\alpha_{j,k}-1} / B(\alpha_k) d\zeta}{\int_{A_{n,k}} e^{l_n(k \sum_{j=1}^k \omega_j \mathbb{1}_{I_j}) - l_n(g_{0[k]})} \prod_{j=1}^k \omega_j^{\alpha_{j,k}-1} / B(\alpha_k) d\omega} \quad (33)$$

since, when  $\|\omega - \omega_0\|_1 \leq M \sqrt{k \log n} / \sqrt{n}$ ,

$$\|\zeta - \omega_0\|_1 \leq \|\omega - \omega_0\|_1 + \|\omega - \zeta\|_1 = \frac{M \sqrt{k \log n} + 2|t| \|\tilde{T}\|_\infty}{\sqrt{n}} \leq (M+1) \frac{\sqrt{k \log n}}{\sqrt{n}}$$

and, vice versa, when  $\|\zeta - \omega_0\|_1 \leq M \sqrt{k \log n} / \sqrt{n}$ ,

$$\|\omega - \omega_0\|_1 \leq \|\omega - \zeta\|_1 + \|\omega_0 - \zeta\|_1 = \frac{M \sqrt{k \log n} + 2|t| \|\tilde{T}\|_\infty}{\sqrt{n}} \leq (M+1) \frac{\sqrt{k \log n}}{\sqrt{n}}.$$

Choosing  $M$ , such that

$$\pi \left[ \|\omega - \omega_0\|_1 \leq (M+1) \sqrt{k \log n} \mid \mathbb{X}_n, k \right] = 1 + o_p(1), \quad (34)$$

Expression (33) equals  $1 + o_p(1)$ . Notice that  $\|\bar{T}_{(k)}\|_{L,k} = \|\bar{T}_{[k]}\|_L$ . We then have that

$$E^\pi \left[ e^{t\sqrt{n}(T(g) - \hat{T}_k)} \mid \mathbb{X}_n, A_{n,k} \right] = e^{t^2 \|\bar{T}_{[k]}\|_L^2} (1 + o_p(1)), \quad (35)$$

which completes the proof for a fixed  $k$  case.

For a random  $k$  case, the proof will follow the same steps as the corresponding part in the proof for Theorem 4.2 in [15]. For completeness, we briefly sketch the proof here. Since  $k$  is not fixed, we will calculate  $E^\pi[e^{t\sqrt{n}(T(f) - \hat{T}_k)} \mid \mathbb{X}_n]$  on  $B_n = \bigcup_{1 \leq k \leq n} A_{n,k} \cap \{f = f_{\omega,k}, k \in \mathcal{K}_n\}$ . Consider  $\mathcal{K}_n$  a subset of  $\{1, 2, \dots, n/\log^2 n\}$  such that  $\pi(\mathcal{K}_n \mid \mathbb{X}_n) = 1 + o_p(1)$  by the concentration property (a) of the random histogram, we have that  $\pi[B_n \mid \mathbb{X}_n] = 1 + o_p(1)$ . We rewrite the left-hand side (LHS) of Equation (35) as  $E^\pi[e^{t\sqrt{n}(T(f) - \hat{T}_k)} \mid \mathbb{X}_n, B_{n,k}]$ , which is also equal to  $e^{t^2 \|\bar{T}_{[k]}\|_L^2} (1 + o_p(1))$ . Notice that  $o(1)$  in this expression is uniform in  $k$ . This is because it holds in the proof for a deterministic case for any given  $k < n$ . Therefore,

$$\begin{aligned} E^\pi \left[ e^{t\sqrt{n}(T(f) - \hat{T})} \mid \mathbb{X}_n, B_n \right] &= \sum_{k \in \mathcal{K}_n} E^\pi \left[ e^{t\sqrt{n}(T(f) - \hat{T}_k) + \hat{T}_k - \hat{T}} \mid \mathbb{X}_n, A_{n,k}, k \right] \pi[k \mid \mathbb{X}_n] \\ &= (1 + o(1)) \sum_{k \in \mathcal{K}_n} e^{t^2 V_k / 2 + t\sqrt{n}(\hat{T}_k - \hat{T})} \pi[k \mid \mathbb{X}_n]. \end{aligned}$$

Using Equations (23) and (25) together with the continuous mapping theorem for the exponential function yields that the last display converges in probability to  $e^{t^2V/2}$  as  $n \rightarrow \infty$ , which completes the proof.  $\square$

The following theorem shows that Method 2 is efficient, the proof of which consists in verifying that the conditions in the above lemma are satisfied.

**Theorem 4.** Suppose  $g_0 \in \mathcal{C}^\beta$  with  $\beta > 0$ . Let the prior on  $k$  be either a Dirac mass at  $k = K_n = n^{1/2}(\log n)^{-2}$  or  $k \sim \pi_k$  given by (13), and let two priors  $\pi_1$  and  $\pi_2$  be defined by (14) and satisfy (13). Then, for all  $\beta > 1/2$ , the BvM holds for  $T(f)$  for both  $\pi_1$  and  $\pi_2$ .

**Proof.** For  $T(f)$  such that Equation (12) is satisfied, Condition (24) is satisfied obviously.

For Equation (23), the empirical process  $\mathbb{G}_n(\tilde{T}_{[k]} - \tilde{T})$  is controlled and will converge to 0 by applying Lemma 19.33 in [11].

Condition (25) is satisfied by Lemma 3 below.

Now we show that Equation (22) holds:

$$\begin{aligned}\|\tilde{T}_f\|_L^2 - \|\tilde{T}_{[k]}\|_L^2 &\leq \left| \int \dot{s}_{T(f)}(x)dx - \int \frac{\dot{s}_{T(f_{[k]})(x)}f(x)}{f_{[k]}(x)}dx \right| \\ &\lesssim \left| \int \dot{s}_{T(f)}(x)f_{[k]}(x) - \dot{s}_{T(f_{[k]})}(x)f(x) \right| \\ &= \left| \int \dot{s}_{T(f_{[k]})}(x)[f_{[k]}(x) - f(x)] \right| \\ &\lesssim \int |f_{[k]}(x) - f(x)|dx.\end{aligned}$$

The last equality is based on Conclusion (3) in Lemma 4 in [19], and the last inequality is due to the assumption that  $\tilde{T}$  is bounded. Then the last term is controlled by  $h(f, f_n)$ , which completes the proof.  $\square$

**Lemma 3.** Under the same conditions as in Theorem 4, Equation (25) holds.

**Proof.** Since  $\tilde{T} = \left( - \left[ \int \ddot{s}_{T(g_0)}(x)g_0^{\frac{1}{2}}(x)dx \right]^{-1} + a_t \right) \frac{\dot{s}_{T(g_0)}(x)}{2g_0^{\frac{1}{2}}(x)}$ , under the deterministic  $k$ -prior with  $k = K_n = n^{1/2}(\log n)^{-2}$  and  $\beta > 1/2$ ,

$$\left| \int (\tilde{T} - \tilde{T}_{[k]})(g_0 - g_{0[k]}) \right| \lesssim h^2(g_0, g_{0[k]}) = o(1/\sqrt{n}).$$

For the random  $k$ -prior, since we restrict  $g$  to be bounded from above and below, so the Hellinger and  $L^2$ -distances considered are comparable. For a given  $k \in \mathcal{K}_n$ , by definition, there exists  $g_k^* \in \mathcal{H}_k^1$  with  $h(g_0, g_k^*) \leq M\epsilon_n(\beta)$ , so

$$h^2(g_0, g_{0[k]}) \lesssim \int (g_0 - g_{0[k]})^2 \leq \int (g_0 - g_k^*)^2 \lesssim h^2(g_0, g_k^*) \lesssim \epsilon_n^2(\beta),$$

which completes the proof.  $\square$

#### 4. Robustness Properties

In frequentist analysis, robustness is usually measured by the influence function and breakdown point of estimators. These have been used to study robustness in minimum Hellinger distance estimators in [3] and in more general minimum disparity estimators in [2,7].

In Bayesian inference, robustness is labeled “outlier rejection” and is studied under the framework of the “theory of conflict resolution”. There is a large literature on this topic, e.g., [20–22]. While the results of [22] are only about symmetric distributions, [23] provides corresponding results covering a wider class of distributions with tails in the general exponential power family. These results provide a complete theory for the case of many observations and a single location parameter.

We examine the behavior of the methods MHB and BMH under a mixture model for gross errors. Let  $\delta_z$  denote the uniform density of the interval  $(z - \epsilon, z + \epsilon)$ , where  $\epsilon > 0$  is small, and let  $f_{\theta,\alpha,z} = (1 - \alpha)f_\theta + \alpha\delta_z$ , where  $\theta \in \Theta$  and  $\alpha \in [0, 1]$  and  $z$  is a real number. The density  $f_{\theta,\alpha,z}$  models a situation, where  $100(1 - \alpha)\%$  observations are distributed from  $f_\theta$ , and  $100\alpha\%$  of the observations are the gross errors located near  $z$ .

**Theorem 5.** For every  $\alpha \in (0, 1)$  and every  $\theta \in \Theta$ , denote the mixture model for gross errors by  $f_{\theta,\alpha,z}$ . We then have that  $\lim_{z \rightarrow \infty} \lim_{n \rightarrow \infty} T(g_n^*) = \theta$ , under the assumptions of Theorem 3 and that, for the BMH method,  $\pi(T(g) | \mathbb{X}_n) \rightarrow \phi(\theta, \|\tilde{T}_{f_{\theta,\alpha,z}}\|_L^2)$  in the distribution as  $n \rightarrow \infty$  and  $z \rightarrow \infty$ , where  $\phi$  denotes the probability function of the normal distribution, when conditions in Theorem 4 are satisfied.

**Proof.** By Theorem 7 in [3], for functional  $T$ , as we defined and under the conditions in this theorem, we have that

$$\lim_{z \rightarrow \infty} T(f_{\theta,\alpha,z}) = \theta.$$

We also have that, for MHB, under conditions of Theorem 3,  $\lim_{n \rightarrow \infty} T(g_n^*) \rightarrow T(f_{\theta,\alpha,z})$  in probability. Combining the two results,  $\lim_{z \rightarrow \infty} \lim_{n \rightarrow \infty} T(g_n^*) = \theta$ , when the data is generated from a contaminated distribution as  $f_{\theta,\alpha,z}$ . Similarly, by Theorem 4, we have that  $\pi(T(g) | \mathbb{X}_n) \rightarrow \phi(T(f_{\theta,\alpha,z}), \|\tilde{T}_{f_{\theta,\alpha,z}}\|_L^2)$  in distribution as  $n \rightarrow \infty$ , and which converges to  $\phi(\theta, \|\tilde{T}_{f_{\theta,\alpha,z}}\|_L^2)$ , as  $z \rightarrow \infty$ .  $\square$

## 5. Demonstration

We provide a demonstration of both BMH and MHB methods on two data sets: the classical Newcomb light speed data (see [24,25]), in which 2 out of 66 values are clearly negative outliers, and a bivariate simulation containing 10% contamination in two asymmetric locations.

We have implemented the BMH and MHB methods using two Bayesian nonparametric priors:

1. the random histogram prior studied in this paper based on a fixed  $k = 100$  with the range naturally extended to the range of the observed data (this is applied only to our first univariate example).
2. the popular Dirichlet Process (DP) kernel mixture of the form

$$\begin{aligned} y_i | \mu_i, \Sigma_i &\sim N(\mu_i, \Sigma_i) \\ (\mu_i, \Sigma_i) | G &\sim G \\ G | \alpha, G_0 &\sim DP(\alpha G_0) \end{aligned}$$

where the baseline distribution is the conjugate normal-inverted Wishart,

$$G_0 = N(\mu | m_1, (1/k_0)\Sigma) IW(\Sigma | \nu_1, \psi_1).$$

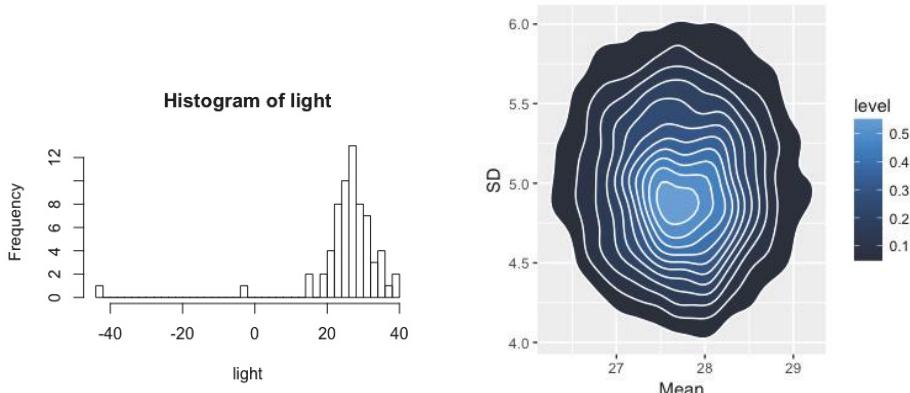
Note that, when  $y_i$  values are univariate observations, the inverse Wishart (IW) distribution reverts to an inverse Gamma distribution. To complete the model specification, independent hyperpriors are assumed

$$\begin{aligned}
\alpha \mid a_0, b_0 &\sim \text{Gamma}(a_0, b_0) \\
m_1 \mid m_2, s_2 &\sim N(m_2, s_2) \\
k_0 \mid \tau_1, \tau_2 &\sim \text{Gamma}(\tau_1/2, \tau_2/2) \\
\psi_1 \mid \nu_2, \psi_2 &\sim IW(\nu_2, \psi_2).
\end{aligned}$$

We obtain posteriors for both using BUGS. We have elected to use BUGS here as opposed to the package DPpackage within R despite the latter's rather efficient MCMC algorithms because our BMH method requires direct access to samples from the posterior distribution as opposed to the expected *a posteriori* estimate. The R package distrEx is then used to construct the sampled density functions and calculated the Hellinger distance between the sampled densities from the nonparametric model and the assumed normal distribution. The R package optimx is also used to find the minima of the Hellinger distances. The time cost of our methods are dominated by the optimization step rather than by the obtaining of samples from the posterior density.

We first apply BMH and MHB on the Simon Newcomb's measurements to measure the speed of light. The data contains 66 observations. For this example, we specify the parameters and hyper-parameters of the DPM as  $\alpha = 1$ ,  $m_2 = 0$ ,  $s_2 = 1000$ ,  $\tau_1 = 1$ ,  $\tau_2 = 100$ , and  $\nu_2 = 2$ ,  $\psi_2 = 1$ . We plot the data and a bivariate contour of the BMH posterior for both the mean and variance of the assumed normal in Figure 1, where, despite outliers, the BvM result is readily apparent.

Table 1 summarizes these estimates. We report the estimated mean and variance with and without the obvious outliers as well as the same quantities estimated using both MHB and BMH methods with the last of these being the expected *a posteriori* estimates. Quantities in parentheses given the "natural" standard error for each quantity: likelihood estimates correspond to standard normal theory—dividing the estimated standard error by  $\sqrt{n}$ , and BMH standard errors are obtained from the posterior distribution. For MHB, we used a bootstrap and note that, while the computational cost involved in estimating MHB is significantly lower than BMH when obtaining a point estimate, the standard errors require and MCMC chain for each bootstrap, significantly raising the cost of obtaining these estimates. We observe that both prior specifications result in parameter estimates that are identical to two decimal places and very close to those obtained after removing outliers.



**Figure 1.** Left: Histogram of the light speed data; Right: bivariate contour plots of the posterior for the mean and variance of these data from the BMH method.

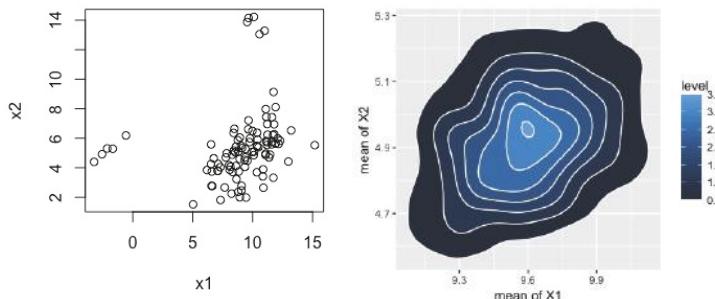
**Table 1.** Estimation results for Newcomb’s light speed data. Direct Estimate refers to the standard mean and variance estimates, and Without Outliers indicates the same estimates with outliers removed. The first row for each parameter gives the estimate under a Dirichlet process prior and the second using a random histogram. Standard errors for each estimate are given in parentheses: these are from the normal theory for the first two columns via a bootstrap for MHB and from posterior samples for BMH.

	Direct Estimate	Without Outliers	MHB	BMH
$\hat{\mu}$	26.21 (1.32)	27.75 (0.64)	27.72 (0.64) 27.72 (0.64)	27.73 (0.63) 27.73 (0.63)
$\hat{\sigma}$	10.75 (3.40)	5.08 (0.46)	5.07 (0.46) 5.07 (0.46)	5.00 (0.47) 5.00 (0.47)

To examine the practical implementation of methods that go beyond our theoretical results, we applied these methods to a simulated two-dimensional data set of 100 data points generated from a standard normal with two contamination distributions. Specifically, our data distribution comes from

$$\frac{9}{10}N\left(\begin{pmatrix} 10 \\ 5 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 5 \end{pmatrix}\right) + \frac{1}{20}N\left(\begin{pmatrix} -2 \\ 5 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}\right) + \frac{1}{20}N\left(\begin{pmatrix} 10 \\ 14 \end{pmatrix}, \begin{pmatrix} 0.4 & -0.1 \\ -0.1 & 0.4 \end{pmatrix}\right)$$

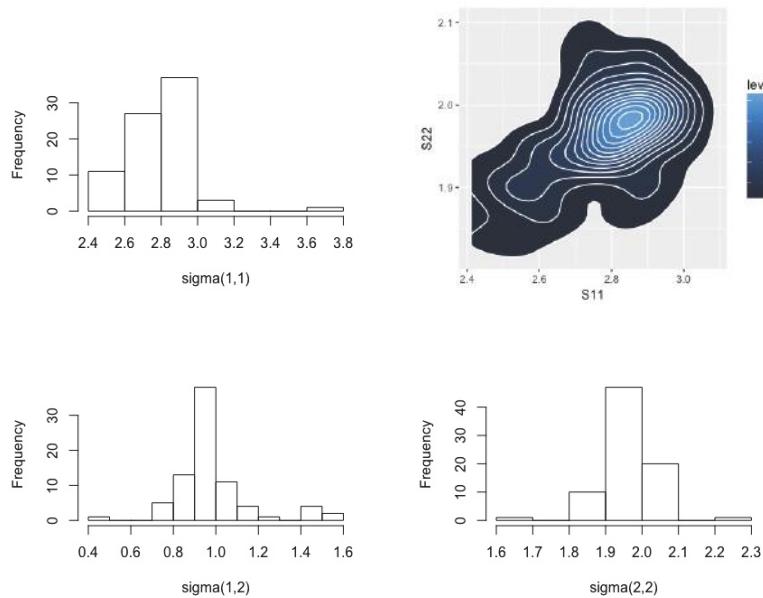
where exactly five points were generated from each of the second-two Gaussians. Our DP prior used the same hyper-parameters as above with the exception that  $\Psi_1$  was obtained from the empirical variance of the (contaminated) data, and  $(m_2, S_2)$  were extended to their 2-dimensional form as  $((0, 0)^T, diag(1000, 1000))$ . Figure 2 plots these data along with the posterior for the two means. Figure 3 provides posterior distributions for the components of the variance matrix. Table 2 presents estimation results for the full data and those with the contaminating distributions removed as well as from the BMH method. Here we again observe that BMH yields results that are very close to those obtained using the uncontaminated data. There is some more irregularity in our estimates, particularly in Figure 3, which we speculate is due to poor optimization. There is considerable scope to improve the numerics of minimum Hellinger distance methods more generally, but this is beyond the scope of this paper.



**Figure 2.** Left: simulated two-dimensional normal example with two contamination components; Right: BMH posterior for the mean vector  $(\mu_1, \mu_2)$ .

**Table 2.** Estimation results for a contaminated bivariate normal. We provide generating estimates, the natural maximum likelihood estimates with and without outliers and the BMH estimates. Reported BMH estimates are expected *a posteriori* estimates with posterior standard errors given in parentheses.

	$\mu_{01}$	$\mu_{02}$	$\Sigma_{11}$	$\Sigma_{12}$	$\Sigma_{22}$
True	10	5	3	1	2
Contaminated data	9.07	5.36	9.76	1.67	5.80
Data with outliers removed	9.62 (0.13)	4.91 (0.11)	3.45 (0.13)	1.49 (0.13)	2.29 (0.11)
Estimated by BMH	9.59 (0.27)	4.93 (0.19)	2.79 (0.18)	0.98 (0.18)	1.97 (0.076)



**Figure 3.** Posterior distributions for the elements of  $\Sigma$  in the simulated bivariate normal example.

## 6. Discussion

This paper investigates the use of minimum Hellinger distance methods that replace kernel density estimates with Bayesian nonparametric models. We show that simply substituting the expected *a posteriori* estimator will reproduce the efficiency and robustness properties of the classical disparity methods first derived in [3]. Further, inducing a posterior distribution on  $\theta$  through the posterior for  $g$  results in a Bernstein von Mises theorem and a distributional robustness result.

There are multiple potential extensions of this work. While we have focused on the specific pairing of Hellinger distance and random histogram priors, both of these can be generalized. A more general class of disparities was examined in [7], and we believe the extension of our methods to this class are straightforward. More general Bayesian nonparametric priors are discussed in [14], where the Dirichlet process prior has been particularly popular. Extensions to each of these priors will require separate analysis (e.g. [26]). Extensions of disparities to regression models were examined in [27] using a conditional density estimate, where equivalent Bayesian nonparametrics are not as well developed. Other modeling domains such as time series may require multivariate density estimates, resulting in further challenges.

Our results are a counterpoint to the Bayesian extensions of Hellinger distance methods in [2] where the kernel density was retained for  $g_n$  but a prior was given for  $\theta$  and the disparity treated as a log likelihood. Combining both these approaches represents a fully Bayesian implementation of disparity methods and is an important direction of future research.

**Author Contributions:** Conceptualization: Y.W. and G.H.; methodology: Y.W. and G.H.; formal analysis: Y.W.; investigation: Y.W. and G.H.; writing—original draft preparation: Y.W.; writing—review and editing: Y.W. and G.H.

**Funding:** This research based on work supported by NASA under award No(s) NNX15AK38A and by the National Science Foundation grants NSF DEB-0813743 and DMS-1712554.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

BMH	Bayesian Minimum Hellinger Method
MHB	Minimum Hellinger Method with Bayesian Density estimation
BvM	Bernstein von Mises
MGF	Moment Generating Function
DP	Dirichlet Process
DPM	Dirichlet Process Mixture
LAN	local asymptotic normality
BUGS	Bayesian inference Using Gibbs Sampling
MCMC	Markov Chain Monte Carlo
IW	inverse Wishart
RHS	Right hand side
LHS	Left hand side

## References

- Huber, P.J. *Robust Statistics*; Wiley: Hoboken, NJ, USA, 2004.
- Hooker, G.; Vidyashankar, A.N. Bayesian model robustness via disparities. *Test* **2014**, *23*, 556–584. [[CrossRef](#)]
- Beran, R. Minimum Hellinger Distance Estimates for Parametric Models. *Ann. Stat.* **1977**, *5*, 445–463. [[CrossRef](#)]
- Basu, A.; Lindsay, B.G. Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Ann. Inst. Statist. Math.* **1994**, *46*, 683–705. [[CrossRef](#)]
- Basu, A.; Sarkar, S.; Vidyashankar, A.N. Minimum Negative Exponential Disparity Estimation in Parametric Models. *J. Stat. Plan. Inference* **1997**, *58*, 349–370. [[CrossRef](#)]
- Pak, R.J.; Basu, A. Minimum Disparity Estimation in Linear Regression Models: Distribution and Efficiency. *Ann. Inst. Stat. Math.* **1998**, *50*, 503–521. [[CrossRef](#)]
- Park, C.; Basu, A. Minimum Disparity Estimation: Asymptotic Normality and Breakdown Point Results. *Bull. Inform. Cybern.* **2004**, *38*, 19–33.
- Lindsay, B.G. Efficiency versus Robustness: The case for minimum Hellinger distance and related methods. *Ann. Stat.* **1994**, *22*, 1081–1114. [[CrossRef](#)]
- Gervini, D.; Yohai, V.J. A class of robust and fully efficient regression estimators. *Ann. Stat.* **2002**, *30*, 583–616. [[CrossRef](#)]
- Wu, Y.; Ghosal, S. Posterior consistency for some semi-parametric problems. *Sankhyā Ser. A* **2008**, *70*, 267–313.
- Van der Vaart, A. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 2000.
- Ghosal, S.; Ghosh, J.K.; van der Vaart, A. Convergence rates of posterior distributions. *Ann. Stat.* **2000**, *28*, 500–531. [[CrossRef](#)]
- Ghosal, S.; van der Vaart, A. Convergence rates of posterior distributions for noniid observations. *Ann. Stat.* **2007**, *35*, 192–223. [[CrossRef](#)]
- Ghosh, J.K.; Ramamoorthi, R.V. *Bayesian Nonparametrics*; Springer: New York, NY, USA, 2003.
- Castillo, I.; Rousseau, J.A. Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Stat.* **2015**, *43*, 2353–2383. [[CrossRef](#)]
- Amewou-Atisso, M.; Ghosal, S.; Ghosh, J.; Ramamoorthi, R. Posterior consistency for semi-parametric regression problems. *Bernoulli* **2003**, *9*, 291–312. [[CrossRef](#)]
- Rivoirard, V.; Rousseau, J. Bernstein-von Mises theorem for linear functionals of the density. *Ann. Stat.* **2012**, *40*, 1489–1523. [[CrossRef](#)]
- Bagui, S.C.; Mehra, K.L. Convergence of Binomial, Poisson, Negative-Binomial, and Gamma to normal distribution: Moment generating functions technique. *Am. J. Math. Stat.* **2016**, *6*, 115–121.
- Castillo, I.; Nickl, R. Nonparametric Bernstein-von Mises Theorems in Gaussian White Noise. *Ann. Stat.* **2013**, *41*, 1999–2028. [[CrossRef](#)]
- De Finetti, B. The Bayesian approach to the rejection of outliers. In Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20 June–30 July 1960; University of California Press: Berkeley, CA, USA, 1961; Volume 1, pp. 199–210.

21. O'Hagan, A. On outlier rejection phenomena in Bayes inference. *J. R. Stat. Soc. B* **1979**, *41*, 358–367. [[CrossRef](#)]
22. O'Hagan, A. Outliers and credence for location parameter inference. *J. Am. Stat. Assoc.* **1990**, *85*, 172–176. [[CrossRef](#)]
23. Desgagné, A.; Angers, J.-F. Conflicting information and location parameter inference. *Metron* **2007**, *67*, 67–97.
24. Stigler, S.M. Do Robust Estimators Work with Real Data? *Ann. Stat.* **1977**, *5*, 1055–1098. [[CrossRef](#)]
25. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapman and Hall: Boca Raton, FL, USA, 2011.
26. Wu, Y.; Ghosal, S. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat.* **2008**, *3*, 298–331. [[CrossRef](#)]
27. Hooker, G. Consistency, Efficiency and Robustness of Conditional Disparity Methods. *Bernoulli* **2016**, *22*, 857–900. [[CrossRef](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Composite Likelihood Methods Based on Minimum Density Power Divergence Estimator

Elena Castilla <sup>1,\*</sup>, Nirian Martín <sup>2</sup>, Leandro Pardo <sup>1</sup> and Konstantinos Zografos <sup>3</sup>

<sup>1</sup> Department of Statistics and O.R. I, Complutense University of Madrid, 28040 Madrid, Spain; lpardo@mat.ucm.es

<sup>2</sup> Department of Statistics and O.R. II, Complutense University of Madrid, 28003 Madrid, Spain; nirian@estad.ucm.es

<sup>3</sup> Department of Mathematics, University of Ioannina, 45110 Ioannina, Greece; kzografi@uoi.gr

\* Correspondence: elecasti@mat.ucm.es; Tel.: +34-913944535

Received: 6 November 2017; Accepted: 28 December 2017; Published: 31 December 2017

**Abstract:** In this paper, a robust version of the Wald test statistic for composite likelihood is considered by using the composite minimum density power divergence estimator instead of the composite maximum likelihood estimator. This new family of test statistics will be called Wald-type test statistics. The problem of testing a simple and a composite null hypothesis is considered, and the robustness is studied on the basis of a simulation study. The composite minimum density power divergence estimator is also introduced, and its asymptotic properties are studied.

**Keywords:** composite likelihood; maximum composite likelihood estimator; Wald test statistic; composite minimum density power divergence estimator; Wald-type test statistics

## 1. Introduction

It is well known that the likelihood function is one of the most important tools in classical inference, and the resultant estimator, the maximum likelihood estimator (MLE), has nice efficiency properties, although it has not so good robustness properties.

Tests based on MLE (likelihood ratio test, Wald test, Rao's test, etc.) have, usually, good efficiency properties, but in the presence of outliers, the behavior is not so good. To solve these situations, many robust estimators have been introduced in the statistical literature, some of them based on distance measures or divergence measures. In particular, density power divergence measures introduced in [1] have given good robust estimators: minimum density power divergences estimators (MDPDE) and, based on them, some robust test statistics have been considered for testing simple and composite null hypotheses. Some of these tests are based on divergence measures (see [2,3]), and some others are used to extend the classical Wald test; see [4–6] and the references therein.

The classical likelihood function requires exact specification of the probability density function, but in most applications, the true distribution is unknown. In some cases, where the data distribution is available in an analytic form, the likelihood function is still mathematically intractable due to the complexity of the probability density function. There are many alternatives to the classical likelihood function; in this paper, we focus on the composite likelihood. Composite likelihood is an inference function derived by multiplying a collection of component likelihoods; the particular collection used is a conditional determined by the context. Therefore, the composite likelihood reduces the computational complexity so that it is possible to deal with large datasets and very complex models even when the use of standard likelihood methods is not feasible. Asymptotic normality of the composite maximum likelihood estimator (CMLE) still holds with the Godambe information matrix to replace the expected information in the expression of the asymptotic variance-covariance matrix. This allows the construction of composite likelihood ratio test statistics, Wald-type test statistics, as well

as score-type statistics. A review of composite likelihood methods is given in [7]. We have to mention at this point that CMLE, as well as the respective test statistics are seriously affected by the presence of outliers in the set of available data.

The main purpose of the paper is to introduce a new robust family of estimators, namely, composite minimum density power divergence estimators (CMDPDE), as well as a new family of Wald-type test statistics based on the CMDPDE in order to get broad classes of robust estimators and test statistics.

In Section 2, we introduce the CMDPDE, and we provide the associated estimating system of equations. The asymptotic distribution of the CMDPDE is obtained in Section 2.1. Section 2.2 is devoted to the definition of a family of Wald-type test statistics, based on CMDPDE, for testing simple and composite null hypotheses. The asymptotic distribution of these Wald-type test statistics is obtained, as well as some asymptotic approximations to the power function. A numerical example, presented previously in [8], is studied in Section 3. A simulation study based on this example is also presented (Section 3), in order to study the robustness of the CMDPDE, as well as the performance of the Wald-type test statistics based on CMDPDE. Proofs of the results are presented in the Appendix A.

## 2. Composite Minimum Density Power Divergence Estimator

We adopt here the notation by [9], regarding the composite likelihood function and the respective CMLE. In this regard, let  $\{f(\cdot; \theta), \theta \in \Theta \subseteq \mathbb{R}^p, p \geq 1\}$  be a parametric identifiable family of distributions for an observation  $y$ , a realization of a random  $m$ -vector  $Y$ . In this setting, the composite density based on  $K$  different marginal or conditional distributions has the form:

$$\mathcal{CL}(\theta, y) = \prod_{k=1}^K (f_{A_k}(y_j, j \in A_k; \theta))^{w_k}$$

and the corresponding composite log-density has the form:

$$c\ell(\theta, y) = \sum_{k=1}^K w_k \ell_{A_k}(\theta, y),$$

with:

$$\ell_{A_k}(\theta, y) = \log f_{A_k}(y_j, j \in A_k; \theta),$$

where  $\{A_k\}_{k=1}^K$  is a family of random variables associated either with marginal or conditional distributions involving some  $y_j$  and  $j \in \{1, \dots, m\}$  and  $w_k, k = 1, \dots, K$  are non-negative and known weights. If the weights are all equal, then they can be ignored. In this case, all the statistical procedures produce equivalent results.

Let  $y_1, \dots, y_n$  also be independent and identically distributed replications of  $y$ . We denote by:

$$c\ell(\theta, y_1, \dots, y_n) = \sum_{i=1}^n c\ell(\theta, y_i)$$

the composite log-likelihood function for the whole sample. In complete accordance with the classical MLE, the CMLE,  $\hat{\theta}_c$ , is defined by:

$$\hat{\theta}_c = \arg \max_{\theta \in \Theta} \sum_{i=1}^n c\ell(\theta, y_i) = \arg \max_{\theta \in \Theta} \sum_{i=1}^n \sum_{k=1}^K w_k \ell_{A_k}(\theta, y_i). \quad (1)$$

It can also be obtained by solving the equations.

$$u(\theta, y_1, \dots, y_n) = \mathbf{0}_p, \quad (2)$$

where:

$$u(\theta, y_1, \dots, y_n) = \frac{\partial c\ell(\theta, y_1, \dots, y_n)}{\partial \theta} = \sum_{i=1}^n \sum_{k=1}^K w_k \frac{\partial \ell_{A_k}(\theta, y_i)}{\partial \theta}.$$

We are going to see how it is possible to get the CMLE,  $\hat{\theta}_c$ , on the basis of the Kullback–Leibler divergence measure. We shall denote by  $g(y)$  the density generating the data with the respective distribution function denoted by  $G$ . The Kullback–Leibler divergence between the density function  $g(y)$  and the composite density function  $\mathcal{CL}(\theta, y)$  is given by:

$$\begin{aligned} d_{KL}(g(.), \mathcal{CL}(\theta,.)) &= \int_{\mathbb{R}^m} g(y) \log \frac{g(y)}{\mathcal{CL}(\theta,y)} dy \\ &= \int_{\mathbb{R}^m} g(y) \log g(y) dy - \int_{\mathbb{R}^m} g(y) \log \mathcal{CL}(\theta,y) dy. \end{aligned}$$

The term:

$$\int_{\mathbb{R}^m} g(y) \log g(y) dy$$

can be removed because it does not depend on  $\theta$ ; hence, we can define the following estimator of  $\theta$ , based on the Kullback–Leibler divergence:

$$\hat{\theta}_{KL} = \arg \min_{\theta} d_{KL}(g(.), \mathcal{CL}(\theta,.))$$

or equivalently:

$$\begin{aligned} \hat{\theta}_{KL} &= \arg \min_{\theta} \left( - \int_{\mathbb{R}^m} g(y) \log \mathcal{CL}(\theta,y) dy \right) \\ &= \arg \min_{\theta} \left( - \int_{\mathbb{R}^m} \log \mathcal{CL}(\theta,y) dG(y) \right). \end{aligned} \quad (3)$$

If we replace in (3) the distribution function  $G$  by the empirical distribution function  $G_n$ , we have:

$$\begin{aligned} \hat{\theta}_{KL} &= \arg \min_{\theta} \left( - \int_{\mathbb{R}^m} \log \mathcal{CL}(\theta,y) dG_n(y) \right) \\ &= \arg \min_{\theta} \left( - \frac{1}{n} \sum_{i=1}^n c\ell(\theta, y_i) \right) \end{aligned}$$

and this expression is equivalent to Expression (1). Therefore, the estimator  $\hat{\theta}_{KL}$  coincides with the CMLE. Based on the previous idea, we are going to introduce, in a natural way, the composite minimum density power divergence estimator (CMDPDE).

The CMLE,  $\hat{\theta}_c$ , obeys asymptotic normality (see [9]) and in particular:

$$\sqrt{n}(\hat{\theta}_c - \theta) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, (G_*(\theta))^{-1}\right),$$

where  $G_*(\theta)$  denotes the Godambe information matrix, defined by:

$$G_*(\theta) = H(\theta) (J(\theta))^{-1} H(\theta),$$

with  $H(\theta)$  being the sensitivity or Hessian matrix and  $J(\theta)$  being the variability matrix, defined, respectively, by:

$$\begin{aligned} H(\theta) &= E_{\theta}[-\frac{\partial}{\partial \theta} u(\theta, Y)^T], \\ J(\theta) &= \text{Var}_{\theta}[u(\theta, Y)] = E_{\theta}[u(\theta, Y)u(\theta, Y)^T], \end{aligned}$$

where the superscript  $T$  denotes the transpose of a vector or a matrix.

The matrix  $J(\theta)$  is nonnegative definite by definition. In the following, we shall assume that the matrix  $H(\theta)$  is of full rank. Since the component score functions can be correlated, we have  $H(\theta) \neq J(\theta)$ . If  $c\ell(\theta, y)$  is a true log-likelihood function, then  $H(\theta) = J(\theta) = I_F(\theta)$ ,  $I_F(\theta)$  being the Fisher information matrix of the model. Using multivariate version of the Cauchy–Schwarz inequality, we have that the matrix  $G_*(\theta) - I_F(\theta)$  is non-negative definite, i.e., the full likelihood function is more efficient than any other composite likelihood function (cf. [10], Lemma 4A).

We are now going to proceed to the definition of the CMDPDE, which is based on the density power divergence measure, defined as follows. For two densities  $p$  and  $q$  associated with two  $m$ -dimensional random variables, respectively, the density power divergence (DPD) between  $p$  and  $q$  was defined in [1] by:

$$d_\beta(p, q) = \int_{\mathbb{R}^m} \left\{ q(y)^{1+\beta} - \left(1 + \frac{1}{\beta}\right) q(y)^\beta p(y) + \frac{1}{\beta} p(y)^{1+\beta} \right\} dy, \quad (4)$$

for  $\beta > 0$ , while for  $\beta = 0$ , it is defined by:

$$\lim_{\beta \rightarrow 0} d_\beta(p, q) = d_{KL}(p, q).$$

For  $\beta = 1$ , Expression (4) reduces to the  $L_2$  distance:

$$L_2(p, q) = \int_{\mathbb{R}^m} (q(y) - p(y))^2 dy.$$

It is also interesting to note that (4) is a special case of the so-called Bregman divergence  $\int [T(p(y)) - T(q(y)) - \{p(y) - q(y)T'(q(y))\}] dy$ . If we consider  $T(l) = l^{1+\beta}$ , we get  $\beta$  times  $d_\beta(p, q)$ . The parameter  $\beta$  controls the trade-off between robustness and asymptotic efficiency of the parameter estimates (see the Simulation Section), which are the minimizers of this family of divergences. For more details about this family of divergence measures, we refer to [11].

In this paper, we are going to consider DPD measures between the density function  $g(y)$  and the composite density function  $\mathcal{CL}(\theta, y)$ , i.e.,

$$d_\beta(g(.), \mathcal{CL}(\theta, .)) = \int_{\mathbb{R}^m} \left\{ \mathcal{CL}(\theta, y)^{1+\beta} - \left(1 + \frac{1}{\beta}\right) \mathcal{CL}(\theta, y)^\beta g(y) + \frac{1}{\beta} g(y)^{1+\beta} \right\} dy \quad (5)$$

for  $\beta > 0$ , while for  $\beta = 0$ , we have,

$$\lim_{\beta \rightarrow 0} d_\beta(g(.), \mathcal{CL}(\theta, .)) = d_{KL}(g(.), \mathcal{CL}(\theta, .)).$$

The CMDPDE,  $\hat{\theta}_c^\beta$ , is defined by:

$$\hat{\theta}_c^\beta = \arg \min_{\theta \in \Theta} d_\beta(g(.), \mathcal{CL}(\theta, .)).$$

The term:

$$\int_{\mathbb{R}^m} g(y)^{1+\beta} dy$$

does not depend on  $\theta$ , and consequently, the minimization of (5) with respect to  $\theta$  is equivalent to minimizing:

$$\int_{\mathbb{R}^m} \left( \mathcal{CL}(\theta, y)^{1+\beta} - \left(1 + \frac{1}{\beta}\right) \mathcal{CL}(\theta, y)^\beta g(y) \right) dy$$

or:

$$\int_{\mathbb{R}^m} \mathcal{CL}(\theta, y)^{1+\beta} dy - \left(1 + \frac{1}{\beta}\right) \int_{\mathbb{R}^m} \mathcal{CL}(\theta, y)^\beta dG(y).$$

Now, we replace the distribution function  $G$  by the empirical distribution function  $G_n$ , and we get:

$$\int_{\mathbb{R}^m} \mathcal{CL}(\theta, y)^{1+\beta} dy - \left(1 + \frac{1}{\beta}\right) \frac{1}{n} \sum_{i=1}^n \mathcal{CL}(\theta, y_i)^\beta. \quad (6)$$

As a consequence, for a fixed value of  $\beta$ , the CMDPDE of  $\theta$  can be obtained by minimizing the expression given in (6); or equivalently, by maximizing the expression:

$$\frac{1}{n\beta} \sum_{i=1}^n \mathcal{CL}(\theta, y_i)^\beta - \frac{1}{1+\beta} \int_{\mathbb{R}^m} \mathcal{CL}(\theta, y)^{1+\beta} dy. \quad (7)$$

Under the differentiability of the model, the maximization of the function in Equation (7) leads to an estimating system of equations of the form:

$$\frac{1}{n} \sum_{i=1}^n \mathcal{CL}(\theta, y_i)^\beta \frac{\partial c\ell(\theta, y_i)}{\partial \theta} - \int_{\mathbb{R}^m} \frac{\partial c\ell(\theta, y)}{\partial \theta} \mathcal{CL}(\theta, y)^{1+\beta} dy = 0. \quad (8)$$

The system of Equations (8) can be written as:

$$\frac{1}{n} \sum_{i=1}^n \mathcal{CL}(\theta, y_i)^\beta u(\theta, y_i) - \int_{\mathbb{R}^m} u(\theta, y) \mathcal{CL}(\theta, y)^{1+\beta} dy = 0. \quad (9)$$

and the CMDPDE  $\hat{\theta}_c^\beta$  of  $\theta$  is obtained by the solution of (9). For  $\beta = 0$  in (9), we have:

$$\frac{1}{n} \sum_{i=1}^n u(\theta, y_i) - \int_{\mathbb{R}^m} u(\theta, y) \mathcal{CL}(\theta, y) dy.$$

but:

$$\int_{\mathbb{R}^m} u(\theta, y) \mathcal{CL}(\theta, y) dy = \frac{\partial}{\partial \theta} \mathcal{CL}(\theta, y) dy = 0$$

and we recover the estimating equation for the CMLE,  $\hat{\theta}_c$ , presented in (2).

## 2.1. Asymptotic Distribution of the Composite Minimum Density Power Divergence Estimator

Equation (9) can be written as follows:

$$\frac{1}{n} \sum_{i=1}^n \Psi_\beta(y_i, \theta) = 0$$

with:

$$\Psi_\beta(y_i, \theta) = \mathcal{CL}(\theta, y_i)^\beta u(\theta, y_i) - \int_{\mathbb{R}^m} u(\theta, y) \mathcal{CL}(\theta, y)^{1+\beta} dy.$$

Therefore, the CMDPDE,  $\hat{\theta}_c^\beta$ , is an M-estimator. In this case, it is well known (cf. [12]) that the asymptotic distribution of  $\hat{\theta}_c^\beta$  is given by:

$$\sqrt{n}(\hat{\theta}_c^\beta - \theta) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, (H_\beta(\theta))^{-1} J_\beta(\theta) (H_\beta(\theta))^{-1}\right),$$

being:

$$H_\beta(\theta) = E_\theta \left[ -\frac{\partial \Psi_\beta(Y, \theta)}{\partial \theta^T} \right]$$

and:

$$J_\beta(\theta) = E_\theta \left[ \Psi_\beta(Y, \theta) \Psi_\beta(Y, \theta)^T \right].$$

We are going to establish the expressions of  $H_\beta(\theta)$  and  $J_\beta(\theta)$ . In relation to  $H_\beta(\theta)$ , we have:

$$\begin{aligned}\frac{\partial \Psi_\beta(y, \theta)}{\partial \theta^T} &= \beta \mathcal{C}\mathcal{L}(\theta, y)^{\beta-1} \mathcal{C}\mathcal{L}(\theta, y) u(\theta, y)^T u(\theta, y) + \mathcal{C}\mathcal{L}(\theta, y)^\beta \frac{\partial u(\theta, y)^T}{\partial \theta} \\ &\quad - \int_{\mathbb{R}^m} \frac{\partial u(\theta, y)^T}{\partial \theta} \mathcal{C}\mathcal{L}(\theta, y)^{1+\beta} dy - (1+\beta) \int_{\mathbb{R}^m} \mathcal{C}\mathcal{L}(\theta, y)^\beta \mathcal{C}\mathcal{L}(\theta, y) u(\theta, y)^T u(\theta, y) dy\end{aligned}$$

and:

$$H_\beta(\theta) = E_\theta \left[ -\frac{\partial \Psi_\beta(Y, \theta)}{\partial \theta^T} \right] = \int_{\mathbb{R}^m} \mathcal{C}\mathcal{L}(\theta, y)^{\beta+1} u(\theta, y)^T u(\theta, y) dy. \quad (10)$$

In relation to  $J_\beta(\theta)$ , we have,

$$\begin{aligned}\Psi_\beta(Y, \theta) \Psi_\beta(Y, \theta)^T &= \left( \mathcal{C}\mathcal{L}(\theta, y)^\beta u(\theta, y) - \int_{\mathbb{R}^m} u(\theta, y) \mathcal{C}\mathcal{L}(\theta, y)^{1+\beta} dy \right) \\ &\quad \left( \mathcal{C}\mathcal{L}(\theta, y)^\beta u(\theta, y)^T - \int_{\mathbb{R}^m} u(\theta, y)^T \mathcal{C}\mathcal{L}(\theta, y)^{1+\beta} dy \right) \\ &= \mathcal{C}\mathcal{L}(\theta, y)^{2\beta} u(\theta, y) u(\theta, y)^T - \mathcal{C}\mathcal{L}(\theta, y)^\beta u(\theta, y) \int_{\mathbb{R}^m} u(\theta, y)^T \mathcal{C}\mathcal{L}(\theta, y)^{1+\beta} dy \\ &\quad - \mathcal{C}\mathcal{L}(\theta, y)^\beta u(\theta, y)^T \int_{\mathbb{R}^m} u(\theta, y) \mathcal{C}\mathcal{L}(\theta, y)^{1+\beta} dy \\ &\quad + \left( \int_{\mathbb{R}^m} u(\theta, y) \mathcal{C}\mathcal{L}(\theta, y)^{1+\beta} dy \right) \left( \int_{\mathbb{R}^m} u(\theta, y)^T \mathcal{C}\mathcal{L}(\theta, y)^{1+\beta} dy \right).\end{aligned}$$

Then,

$$J_\beta(\theta) = E_\theta \left[ \Psi_\beta(Y, \theta) \Psi_\beta(Y, \theta)^T \right] = \int_{\mathbb{R}^m} \mathcal{C}\mathcal{L}(\theta, y)^{2\beta+1} u(\theta, y) u(\theta, y)^T dy \quad (11)$$

$$- \int_{\mathbb{R}^m} \mathcal{C}\mathcal{L}(\theta, y)^{\beta+1} u(\theta, y) dy \int_{\mathbb{R}^m} u(\theta, y)^T \mathcal{C}\mathcal{L}(\theta, y)^{1+\beta} dy. \quad (12)$$

Based on the previous results, we have the following theorem.

**Theorem 1.** Under suitable regularity conditions, we have:

$$\sqrt{n}(\hat{\theta}_c^\beta - \theta) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N} \left( \mathbf{0}, (H_\beta(\theta))^{-1} J_\beta(\theta) (H_\beta(\theta))^{-1} \right),$$

where the matrices  $H_\beta(\theta)$  and  $J_\beta(\theta)$  were defined in (10) and (11), respectively.

**Remark 1.** If we apply the previous theorem for  $\beta = 0$ , then we get the CMLE, and the asymptotic variance covariance matrix coincides with the Godambe information matrix because:

$$H_\beta(\theta) = H(\theta) \text{ and } J_\beta(\theta) = J(\theta),$$

for  $\beta = 0$ .

## 2.2. Wald-Type Tests Statistics Based on the Composite Minimum Power Divergence Estimator

Wald-type test statistics based on MDPDE have been considered with excellent results in relation to the robustness in different statistical problems; see for instance [4–6].

Motivated by those works, we focus in this section on the definition and the study of Wald-type test statistics, which are defined by means of CMDPDE estimators instead of MDPDE estimators. In this context, if we are interested in testing:

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta \neq \theta_0, \quad (13)$$

we can consider the family of Wald-type test statistics:

$$W_{n,\beta}^0 = n(\hat{\theta}_c^\beta - \theta_0)^T \left( (H_\beta(\theta_0))^{-1} J_\beta(\theta_0) (H_\beta(\theta_0))^{-1} \right)^{-1} (\hat{\theta}_c^\beta - \theta_0). \quad (14)$$

For  $\beta = 0$ , we get the classical Wald-type test statistic considered in the composite likelihood methods (see for instance [7]).

In the following theorem, we present the asymptotic null distribution of the family of the Wald-type test statistics  $W_{n,\beta}^0$ :

**Theorem 2.** *The asymptotic distribution of the Wald-type test statistics given in (14) is a chi-square distribution with  $p$  degrees of freedom.*

The proof of this Theorem 2 is given in Appendix A.1.

**Theorem 3.** *Let  $\theta^*$  be the true value of the parameter  $\theta$ , with  $\theta^* \neq \theta_0$ . Then, it holds:*

$$\sqrt{n} \left( l(\hat{\theta}_c^\beta) - l(\theta^*) \right) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N(\mathbf{0}, \sigma_{W_\beta^0}^2(\theta^*)),$$

being:

$$l(\theta) = (\theta - \theta_0)^T \left( (H_\beta(\theta_0))^{-1} J_\beta(\theta_0) (H_\beta(\theta_0))^{-1} \right)^{-1} (\theta - \theta_0)$$

and:

$$\sigma_{W_\beta^0}^2(\theta^*) = 4(\theta^* - \theta_0)^T \left( (H_\beta(\theta_0))^{-1} J_\beta(\theta_0) (H_\beta(\theta_0))^{-1} \right)^{-1} (\theta^* - \theta_0). \quad (15)$$

The proof of the Theorem is outlined in Appendix A.2.

**Remark 2.** *Based on the previous result, we can approximate the power,  $\beta_{W_n^0}$ , of the Wald-type test statistics in  $\theta^*$  by:*

$$\begin{aligned} \beta_{W_{n,\beta}^0}(\theta^*) &= \Pr \left( W_{n,\beta}^0 > \chi_{p,\alpha}^2 / \theta = \theta^* \right) \\ &= \Pr \left( l(\hat{\theta}_c^\beta) - l(\theta^*) > \frac{\chi_{p,\alpha}^2}{n} - l(\theta^*) \middle| \theta = \theta^* \right) \\ &= \Pr \left( \sqrt{n} \left( l(\hat{\theta}_c^\beta) - l(\theta^*) \right) > \sqrt{n} \left( \frac{\chi_{p,\alpha}^2}{n} - l(\theta^*) \right) \middle| \theta = \theta^* \right) \\ &= \Pr \left( \sqrt{n} \frac{\left( l(\hat{\theta}_c^\beta) - l(\theta^*) \right)}{\sigma_{W_{n,\beta}^0}(\theta^*)} > \frac{\sqrt{n}}{\sigma_{W_{n,\beta}^0}(\theta^*)} \left( \frac{\chi_{p,\alpha}^2}{n} - l(\theta^*) \right) \middle| \theta = \theta^* \right) \\ &= 1 - \Phi_n \left( \frac{\sqrt{n}}{\sigma_{W_{n,\beta}^0}(\theta^*)} \left( \frac{\chi_{p,\alpha}^2}{n} - l(\theta^*) \right) \right), \end{aligned}$$

where  $\Phi_n$  is a sequence of distribution functions tending uniformly to the standard normal distribution function  $\Phi(x)$ .

It is clear that:

$$\lim_{n \rightarrow \infty} \beta_{W_{n,\beta}^0}(\theta^*) = 1$$

for all  $\alpha \in (0, 1)$ . Therefore, the Wald-type test statistics are consistent in the sense of Fraser.

In many practical hypothesis testing problems, the restricted parameter space  $\Theta_0 \subset \Theta$  is defined by a set of  $r$  restrictions of the form:

$$g(\theta) = \mathbf{0}_r \quad (16)$$

on  $\Theta$ , where  $g : \mathbb{R}^p \rightarrow \mathbb{R}^r$  is a vector-valued function such that the  $p \times r$  matrix:

$$\mathbf{G}(\theta) = \frac{\partial g(\theta)}{\partial \theta}^T \quad (17)$$

exists and is continuous in  $\theta$  and  $\text{rank}(\mathbf{G}(\theta)) = r$ ; where  $\mathbf{0}_r$  denotes the null vector of dimension  $r$ .

Now, we are going to consider composite null hypotheses,  $\Theta_0 \subset \Theta$ , in the way considered in (16), and our interest is in testing:

$$H_0 : \theta \in \Theta_0 \text{ against } H_1 : \theta \notin \Theta_0 \quad (18)$$

on the basis of a random sample of size  $n$ ,  $X_1, \dots, X_n$ .

**Definition 1.** The family of Wald-type test statistics for testing (18) is given by:

$$W_{n,\beta} = ng\left(\hat{\theta}_c^\beta\right)^T \left[ \mathbf{G}(\hat{\theta}_c^\beta)^T \left( \mathbf{H}_\beta(\hat{\theta}_c^\beta) \right)^{-1} \mathbf{J}_\beta(\hat{\theta}_c^\beta) \left( \mathbf{H}_\beta(\hat{\theta}_c^\beta) \right)^{-1} \mathbf{G}(\hat{\theta}_c^\beta) \right]^{-1} g\left(\hat{\theta}_c^\beta\right), \quad (19)$$

where the matrices  $\mathbf{G}(\theta)$ ,  $\mathbf{H}_\beta(\theta)$  and  $\mathbf{J}_\beta(\theta)$  were defined in (17), (10) and (11), respectively, and the function  $g$  in (16).

If we consider  $\beta = 0$ , then  $\hat{\theta}_c^\beta$  coincides with the CMLE,  $\hat{\theta}_c$ , of  $\theta$  and  $(\mathbf{H}_\beta(\hat{\theta}_c))^{-1} \mathbf{J}_\beta(\hat{\theta}_c) (\mathbf{H}_\beta(\hat{\theta}_c))^{-1}$  with the inverse of the Fisher information matrix, and then, we get the classical Wald test statistic considered in the composite likelihood methods.

In the next theorem, we present the asymptotic distribution of  $W_{n,\beta}$ .

**Theorem 4.** The asymptotic distribution of the Wald-type test statistics, given in (19), is a chi-square distribution with  $r$  degrees of freedom.

The proof of this Theorem is presented in Appendix A.3.

Consider the null hypothesis  $H_0 : \theta \in \Theta_0 \subset \Theta$ . By Theorem 4, the null hypothesis should be rejected if  $W_{n,\beta} \geq \chi_{r,\alpha}^2$ . The following theorem can be used to approximate the power function. Assume that  $\theta^* \notin \Theta_0$  is the true value of the parameter, so that  $\hat{\theta}_c^\beta \xrightarrow[n \rightarrow \infty]{a.s.} \theta^*$ .

**Theorem 5.** Let  $\theta^*$  be the true value of the parameter, with  $\theta^* \neq \theta_0$ . Then, it holds:

$$\sqrt{n} \left( l^*(\hat{\theta}_c^\beta) - l^*(\theta^*) \right) \xrightarrow[n \rightarrow \infty]{L} N(0, \sigma_{W_\beta}^2(\theta^*))$$

being:

$$l^*(\theta) = ng(\theta)^T \left[ \mathbf{G}(\theta_0)^T \left( \mathbf{H}_\beta(\theta_0) \right)^{-1} \mathbf{J}_\beta(\theta_0) \left( \mathbf{H}_\beta(\theta_0) \right)^{-1} \mathbf{G}(\theta_0) \right]^{-1} g(\theta)$$

and:

$$\sigma_{W_\beta}^2(\theta^*) = \left( \frac{\partial l^*(\theta)}{\partial \theta} \right)_{\theta=\theta^*}^T \left( \mathbf{H}_\beta(\theta_0) \right)^{-1} \mathbf{J}_\beta(\theta_0) \left( \mathbf{H}_\beta(\theta_0) \right)^{-1} \left( \frac{\partial l^*(\theta)}{\partial \theta} \right)_{\theta=\theta^*}. \quad (20)$$

### 3. Numerical Example

In this section, we shall consider an example, studied previously by [8], in order to study the robustness of CMLE. The aim of this section is to clarify the different issues that were discussed in the previous sections.

Consider the random vector  $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)^T$ , which follows a four-dimensional normal distribution with mean vector  $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)^T$  and variance-covariance matrix:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & 2\rho & 2\rho \\ \rho & 1 & 2\rho & 2\rho \\ 2\rho & 2\rho & 1 & \rho \\ 2\rho & 2\rho & \rho & 1 \end{pmatrix}, \quad (21)$$

i.e., we suppose that the correlation between  $Y_1$  and  $Y_2$  is the same as the correlation between  $Y_3$  and  $Y_4$ . Taking into account that  $\boldsymbol{\Sigma}$  should be semi-positive definite, the following condition is imposed:  $-\frac{1}{5} \leq \rho \leq \frac{1}{3}$ . In order to avoid several problems regarding the consistency of the CMLE of the parameter  $\rho$  (cf. [8]), we shall consider the composite likelihood function:

$$\mathcal{CL}(\boldsymbol{\theta}, \mathbf{y}) = f_{A_1}(\boldsymbol{\theta}, \mathbf{y})f_{A_2}(\boldsymbol{\theta}, \mathbf{y}),$$

where:

$$\begin{aligned} f_{A_1}(\boldsymbol{\theta}, \mathbf{y}) &= f_{12}(\mu_1, \mu_2, \rho, y_1, y_2), \\ f_{A_2}(\boldsymbol{\theta}, \mathbf{y}) &= f_{34}(\mu_3, \mu_4, \rho, y_3, y_4), \end{aligned}$$

where  $f_{12}$  and  $f_{34}$  are the densities of the marginals of  $\mathbf{Y}$ , i.e., bivariate normal distributions with mean vectors  $(\mu_1, \mu_2)^T$  and  $(\mu_3, \mu_4)^T$ , respectively, and common variance-covariance matrix:

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

with densities given by:

$$f_{h,h+1}(\mu_h, \mu_{h+1}, \rho, y_h, y_{h+1}) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}Q(y_h, y_{h+1})\right\}, \quad h \in \{1, 3\},$$

being:

$$Q(y_h, y_{h+1}) = (y_h - \mu_h)^2 - 2\rho(y_h - \mu_h)(y_{h+1} - \mu_{h+1}) + (y_{h+1} - \mu_{h+1})^2, \quad h \in \{1, 3\}.$$

By  $\boldsymbol{\theta}$ , we are denoting the parameter vector of our model, i.e.,  $\boldsymbol{\theta} = (\mu_1, \mu_2, \mu_3, \mu_4, \rho)^T$ . The system of equations that it is necessary to solve in order to obtain the CMDPDE:

$$\hat{\boldsymbol{\theta}}_c^\beta = \left( \hat{\mu}_{1,c}^\beta, \hat{\mu}_{2,c}^\beta, \hat{\mu}_{3,c}^\beta, \hat{\mu}_{4,c}^\beta, \hat{\rho}_c^\beta \right)^T,$$

is given (see Appendix A.4) by:

$$\frac{1}{n} \sum_{i=1}^n f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^{\beta-1} f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{1i} - \mu_1) + 2\rho(y_{2i} - \mu_2)] \right\} = 0, \quad (22)$$

$$\frac{1}{n} \sum_{i=1}^n f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^{\beta-1} f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{2i} - \mu_2) + 2\rho(y_{1i} - \mu_1)] \right\} = 0, \quad (23)$$

$$\frac{1}{n} \sum_{i=1}^n f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^{\beta-1} f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{3i} - \mu_3) + 2\rho(y_{4i} - \mu_4)] \right\} = 0, \quad (24)$$

$$\frac{1}{n} \sum_{i=1}^n f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^\beta f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{4i} - \mu_4) + 2\rho(y_{3i} - \mu_3)] \right\} = 0 \quad (25)$$

and:

$$\frac{1}{n\beta} \sum_{i=1}^n \frac{\partial \mathcal{CL}(\boldsymbol{\theta}, \mathbf{y}_i)^\beta}{\partial \rho} - \frac{\beta(2\pi)^{-2\beta}}{(\beta+1)^3} \frac{2\rho}{(1-\rho^2)^{\beta+1}} = 0, \quad (26)$$

being:

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\theta}, \mathbf{y}_i)^\beta}{\partial \rho} &= \frac{\rho}{1-\rho^2} \beta f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^\beta f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta \\ &\quad \left\{ 2 + \frac{1}{\rho} \{ (y_{1i} - \mu_1)(y_{2i} - \mu_2) + (y_{3i} - \mu_3)(y_{4i} - \mu_4) \} \right. \\ &\quad - \frac{1}{1-\rho^2} \left( (y_{1i} - \mu_1)^2 - 2\rho (y_{1i} - \mu_1)(y_{2i} - \mu_2) + (y_{2i} - \mu_2)^2 \right) \\ &\quad \left. - \frac{1}{1-\rho^2} \left( (y_{3i} - \mu_3)^2 - 2\rho (y_{3i} - \mu_3)(y_{4i} - \mu_4) + (y_{4i} - \mu_4)^2 \right) \right\}. \end{aligned}$$

After some heavy algebraic manipulations specified in Appendix A.5, the sensitivity and variability matrices are given by:

$$\mathbf{H}_\beta(\boldsymbol{\theta}) = \frac{C_\beta}{(\beta+1)(1-\rho^2)} \begin{pmatrix} 1 & -\rho & 0 & 0 & 0 \\ -\rho & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -\rho & 0 \\ 0 & 0 & -\rho & 1 & 0 \\ 0 & 0 & 0 & 0 & 2 \frac{(\rho^2+1)+2\rho^2\beta^2}{(1-\rho^2)(1+\beta)} \end{pmatrix} \quad (27)$$

and:

$$\mathbf{J}_\beta(\boldsymbol{\theta}) = \mathbf{H}_{2\beta}(\boldsymbol{\theta}) - \boldsymbol{\xi}_\beta(\boldsymbol{\theta}) \boldsymbol{\xi}_\beta(\boldsymbol{\theta})^T, \quad (28)$$

where  $C_\beta = \frac{1}{(\beta+1)^2} \left( \frac{1}{(2\pi)^2(1-\rho^2)} \right)^\beta$  and  $\boldsymbol{\xi}_\beta(\boldsymbol{\theta}) = (0, 0, 0, 0, \frac{2\rho\beta C_\beta}{(\beta+1)(1-\rho^2)})^T$ .

### Simulation Study

A simulation study, developed by using the R statistical programming environment, is presented in order to study the behavior of the CMDPDE, as well as the behavior of the Wald-type test statistics based on them. The theoretical model studied in the previous example is considered. The parameters in the model are:

$$\boldsymbol{\theta} = (\mu_1, \mu_2, \mu_3, \mu_4, \rho)^T$$

and we are interested in studying the behavior of the CMDPDE:

$$\hat{\boldsymbol{\theta}}_c^\beta = (\hat{\mu}_{1,c}^\beta, \hat{\mu}_{2,c}^\beta, \hat{\mu}_{3,c}^\beta, \hat{\mu}_{4,c}^\beta, \hat{\rho}_c^\beta)^T$$

as well as the behavior of the Wald-type test statistics for testing:

$$H_0 : \rho = \rho_0 \quad \text{against} \quad H_1 : \rho \neq \rho_0. \quad (29)$$

Through  $R = 10,000$  replications of the simulation experiment, we compare, for different values of  $\beta$ , the corresponding CMDPDE through the root of the mean square errors (RMSE), when the true value of the parameters is  $\boldsymbol{\theta} = (0, 0, 0, 0, \rho)^T$  and  $\rho \in \{-0.1, 0, 0.15\}$ . We pay special attention to the problem of the existence of some outliers in the sample, generating 5% of the samples with  $\boldsymbol{\theta} = (1, 3, -2, -1, \bar{\rho})^T$  and  $\bar{\rho} \in \{-0.15, 0.1, 0.2\}$ , respectively. Notice that, although the case  $\rho = 0$  has been considered; this case is less important taking into account the method of the theoretical model under consideration, and having the case of independent observations, the composite likelihood theory is useless. Results are presented in Tables 1 and 2. Two points deserve our attention. The first one is that, as expected, RMSEs for contaminated data are always greater than RMSEs for pure data and that the RMSEs decrease when the sample size  $n$  increases. The second is that, while in pure data, RMSEs are greater for big values of  $\beta$ , when working with contaminated data, the CMDPDE with medium-low

values of  $\beta$  ( $\beta \in \{0.1, 0.2, 0.3\}$ ) present the best behavior in terms of efficiency. These statements are also true for larger levels of contamination, noting that, when larger percentages are considered, larger values of  $\beta$  are also considerable in terms of efficiency (see Tables 3–5 for contamination equal to 10%, 15% and 20%, respectively). Considering the mean absolute error (MAE) for the evaluation of the accuracy, we obtain similar results (Table 6).

**Table 1.** RMSEs for pure data.

	<i>n</i> = 100			<i>n</i> = 200			<i>n</i> = 300		
	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$
$\beta = 0$	0.0958	0.0950	0.0948	0.0683	0.0668	0.0666	0.0553	0.0552	0.0551
$\beta = 0.1$	0.0972	0.0961	0.0966	0.0693	0.0676	0.0677	0.0560	0.0559	0.0561
$\beta = 0.2$	0.1009	0.0991	0.1007	0.0718	0.0697	0.0704	0.0581	0.0575	0.0585
$\beta = 0.3$	0.1061	0.1034	0.1062	0.0754	0.0727	0.0742	0.0612	0.0599	0.0619
$\beta = 0.4$	0.1123	0.1087	0.1127	0.0797	0.0762	0.0787	0.0649	0.0628	0.0659
$\beta = 0.5$	0.1195	0.1147	0.1200	0.0845	0.0803	0.0837	0.0691	0.0661	0.0702
$\beta = 0.6$	0.1274	0.1215	0.1280	0.0898	0.0848	0.0892	0.0737	0.0697	0.0748
$\beta = 0.7$	0.1361	0.1291	0.1369	0.0955	0.0897	0.0952	0.0786	0.0736	0.0797
$\beta = 0.8$	0.1456	0.1374	0.1467	0.1015	0.0905	0.1016	0.0839	0.0778	0.0849

**Table 2.** RMSEs for contaminated data (5%).

	<i>n</i> = 100			<i>n</i> = 200			<i>n</i> = 300		
	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$
$\beta = 0$	0.1371	0.1336	0.1287	0.1210	0.1167	0.1113	0.1144	0.1098	0.1047
$\beta = 0.1$	0.1105	0.1104	0.1081	0.0875	0.0874	0.0843	0.0778	0.0786	0.0748
$\beta = 0.2$	0.1061	0.1053	0.1047	0.0783	0.0777	0.0759	0.0660	0.0669	0.0643
$\beta = 0.3$	0.1091	0.1072	0.1083	0.0783	0.0766	0.0761	0.0646	0.0645	0.0635
$\beta = 0.4$	0.1147	0.1118	0.1146	0.0814	0.0788	0.0798	0.0668	0.0657	0.0665
$\beta = 0.5$	0.1215	0.1176	0.1220	0.0858	0.0823	0.0848	0.0703	0.0683	0.0709
$\beta = 0.6$	0.1292	0.1242	0.1302	0.0907	0.0864	0.0905	0.0744	0.0716	0.0758
$\beta = 0.7$	0.1375	0.1315	0.1391	0.0961	0.0911	0.0966	0.0790	0.0753	0.0810
$\beta = 0.8$	0.1465	0.1396	0.1486	0.1018	0.0962	0.1031	0.0838	0.0794	0.0863

**Table 3.** RMSEs for contaminated data (10%).

	<i>n</i> = 100			<i>n</i> = 200			<i>n</i> = 300		
	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$
$\beta = 0$	0.2107	0.2052	0.2000	0.2003	0.1944	0.1884	0.1968	0.1911	0.1844
$\beta = 0.1$	0.1500	0.1472	0.1436	0.1324	0.1305	0.1264	0.1259	0.1250	0.1204
$\beta = 0.2$	0.1238	0.1229	0.1192	0.0991	0.0987	0.0951	0.0881	0.0898	0.0858
$\beta = 0.3$	0.1173	0.1170	0.1139	0.0882	0.0871	0.0846	0.0735	0.0754	0.0726
$\beta = 0.4$	0.1189	0.1187	0.1170	0.0872	0.0849	0.0845	0.0705	0.0714	0.0706
$\beta = 0.5$	0.1237	0.1234	0.1234	0.0901	0.0868	0.0884	0.0721	0.0718	0.0734
$\beta = 0.6$	0.1301	0.1296	0.1311	0.0944	0.0903	0.0938	0.0753	0.0742	0.0779
$\beta = 0.7$	0.1375	0.1367	0.1396	0.0995	0.0947	0.1000	0.0793	0.0776	0.0831
$\beta = 0.8$	0.1467	0.1446	0.1488	0.1050	0.0996	0.1064	0.0837	0.0814	0.0884

**Table 4.** RMSEs for contaminated data (15%).

	<i>n</i> = 100			<i>n</i> = 200			<i>n</i> = 300		
	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$
$\beta = 0$	0.2912	0.2854	0.2788	0.2835	0.2770	0.2713	0.2814	0.2757	0.2687
$\beta = 0.1$	0.2036	0.1994	0.1951	0.1909	0.1874	0.1828	0.1871	0.185	0.1785
$\beta = 0.2$	0.1530	0.1497	0.1453	0.1325	0.1306	0.1252	0.1252	0.1256	0.1181
$\beta = 0.3$	0.1329	0.1295	0.1257	0.1049	0.1031	0.0976	0.0932	0.0945	0.0872
$\beta = 0.4$	0.1287	0.1249	0.1229	0.0957	0.0931	0.0893	0.0805	0.0815	0.0763
$\beta = 0.5$	0.1312	0.1272	0.1272	0.0949	0.0915	0.0902	0.0774	0.0777	0.0755
$\beta = 0.6$	0.1367	0.1323	0.1343	0.0977	0.0936	0.0947	0.0784	0.0781	0.0788
$\beta = 0.7$	0.1436	0.1389	0.1425	0.1019	0.0974	0.1005	0.0811	0.0804	0.0836
$\beta = 0.8$	0.1514	0.1465	0.1514	0.1070	0.1020	0.1069	0.0847	0.0837	0.0888

**Table 5.** RMSEs for contaminated data (20%).

	<i>n</i> = 100			<i>n</i> = 200			<i>n</i> = 300		
	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$
$\beta = 0$	0.3725	0.3680	0.3612	0.3684	0.3618	0.3554	0.3661	0.3610	0.3534
$\beta = 0.1$	0.2691	0.2657	0.2591	0.2625	0.2566	0.2506	0.2577	0.2547	0.2473
$\beta = 0.2$	0.1949	0.1921	0.1831	0.1819	0.1766	0.1683	0.1742	0.1723	0.1624
$\beta = 0.3$	0.1562	0.1537	0.1441	0.1345	0.1299	0.1204	0.1235	0.1222	0.1109
$\beta = 0.4$	0.1419	0.1391	0.1316	0.1126	0.1082	0.1003	0.0987	0.0971	0.0876
$\beta = 0.5$	0.1397	0.1366	0.1323	0.1050	0.1005	0.0962	0.0890	0.0867	0.0812
$\beta = 0.6$	0.1430	0.1395	0.1383	0.1042	0.0996	0.0990	0.0866	0.0837	0.0828
$\beta = 0.7$	0.1488	0.1450	0.1463	0.1066	0.1018	0.1043	0.0877	0.0843	0.0873
$\beta = 0.8$	0.1560	0.1518	0.1552	0.1106	0.1056	0.1105	0.0905	0.0866	0.0927

**Table 6.** MAEs for pure and contaminated data (5%, 10%, 15% and 20%),  $n = 100$ .

	Pure data		5%		10%		15%		20%	
	$\rho = -0.1$	$\rho = 0.15$								
$\beta = 0$	0.076	0.076	0.190	0.179	0.371	0.342	0.626	0.574	0.954	0.877
$\beta = 0.1$	0.077	0.077	0.167	0.163	0.289	0.277	0.464	0.437	0.697	0.652
$\beta = 0.2$	0.081	0.080	0.165	0.163	0.263	0.257	0.388	0.372	0.551	0.520
$\beta = 0.3$	0.085	0.085	0.172	0.170	0.264	0.260	0.370	0.359	0.495	0.473
$\beta = 0.4$	0.090	0.090	0.181	0.180	0.275	0.272	0.377	0.370	0.489	0.474
$\beta = 0.5$	0.095	0.095	0.192	0.192	0.290	0.289	0.394	0.391	0.504	0.496
$\beta = 0.6$	0.101	0.102	0.204	0.204	0.308	0.308	0.416	0.416	0.528	0.527
$\beta = 0.7$	0.108	0.109	0.218	0.218	0.328	0.329	0.441	0.444	0.558	0.561
$\beta = 0.8$	0.115	0.116	0.232	0.233	0.349	0.351	0.468	0.474	0.590	0.599

For a nominal size  $\alpha = 0.05$ , with the model under the null hypothesis given in (29), the estimated significance levels for different Wald-type test statistics are given by:

$$\hat{\alpha}_n^{(\beta)}(\rho_0) = \widehat{\Pr}(W_n^\beta > \chi_{1,0.05}^2 | H_0) = \frac{\sum_{i=1}^R I(W_{n,i}^\beta > \chi_{1,0.05}^2 | \rho_0)}{R},$$

with  $I(S)$  being the indicator function (with a value of one if  $S$  is true and zero otherwise). Empirical levels with the same previous parameter values are presented in Table 7 (pure data) and Table 8 (5% of outliers). While medium-high values of  $\beta$  are not recommended at all, CMLE is generally the best choice when working with pure data. However, the lack of robustness of the CMLE test is impressive, as can be seen in Table 8. The effect of contamination in medium-low values of  $\beta$  is much lighter, while for medium-high values of  $\beta$ , it can return to being deceptively beneficial.

**Table 7.** Levels for pure data.

n = 100				n = 200				n = 300			
$\rho_0 = -0.1$	$\rho_0 = 0$	$\rho_0 = 0.15$	$\rho_0 = -0.1$	$\rho_0 = 0$	$\rho_0 = 0.15$	$\rho_0 = -0.1$	$\rho_0 = 0$	$\rho_0 = 0.15$	$\rho_0 = -0.1$	$\rho_0 = 0$	$\rho_0 = 0.15$
$\beta = 0$	0.067	0.059	0.070	0.068	0.046	0.062	0.072	0.045	0.075		
$\beta = 0.1$	0.067	0.060	0.072	0.062	0.046	0.070	0.085	0.045	0.079		
$\beta = 0.2$	0.072	0.061	0.084	0.069	0.051	0.084	0.097	0.049	0.102		
$\beta = 0.3$	0.081	0.062	0.093	0.084	0.053	0.100	0.112	0.051	0.121		
$\beta = 0.4$	0.094	0.069	0.099	0.103	0.055	0.111	0.127	0.055	0.142		
$\beta = 0.5$	0.105	0.071	0.111	0.118	0.056	0.122	0.149	0.051	0.155		
$\beta = 0.6$	0.122	0.083	0.129	0.131	0.062	0.136	0.167	0.051	0.165		
$\beta = 0.7$	0.135	0.088	0.141	0.139	0.063	0.146	0.181	0.055	0.177		
$\beta = 0.8$	0.153	0.099	0.158	0.151	0.071	0.156	0.198	0.056	0.179		

**Table 8.** Levels for contaminated data (5%).

n = 100				n = 200				n = 300			
$\rho_0 = -0.1$	$\rho_0 = 0$	$\rho_0 = 0.15$	$\rho_0 = -0.1$	$\rho_0 = 0$	$\rho_0 = 0.15$	$\rho_0 = -0.1$	$\rho_0 = 0$	$\rho_0 = 0.15$	$\rho_0 = -0.1$	$\rho_0 = 0$	$\rho_0 = 0.15$
$\beta = 0$	0.357	0.223	0.081	0.638	0.429	0.155	0.788	0.623	0.240		
$\beta = 0.1$	0.121	0.113	0.056	0.207	0.191	0.077	0.287	0.284	0.100		
$\beta = 0.2$	0.065	0.074	0.048	0.066	0.099	0.049	0.086	0.129	0.059		
$\beta = 0.3$	0.057	0.067	0.071	0.057	0.066	0.059	0.065	0.077	0.073		
$\beta = 0.4$	0.075	0.066	0.087	0.067	0.058	0.081	0.079	0.060	0.095		
$\beta = 0.5$	0.090	0.062	0.107	0.080	0.061	0.110	0.105	0.051	0.128		
$\beta = 0.6$	0.096	0.063	0.126	0.095	0.063	0.131	0.117	0.049	0.151		
$\beta = 0.7$	0.109	0.073	0.137	0.101	0.061	0.141	0.127	0.047	0.159		
$\beta = 0.8$	0.125	0.083	0.147	0.109	0.061	0.149	0.141	0.049	0.171		

For finite sample sizes and nominal size  $\alpha = 0.05$ , the simulated powers are obtained under  $H_1$  in (29), when  $\rho \in \{-0.1, 0, 0.1\}$ ,  $\tilde{\rho} = 0.2$  and  $\rho_0 = 0.15$  (Tables 9 and 10). The (simulated) power for different composite Wald-type test statistics is obtained by:

$$\beta_n^{(\beta)}(\rho_0, \rho) = \Pr(W_n^\beta > \chi_{1,0.05}^2 | H_1) \text{ and } \hat{\beta}_n^{(\lambda)}(\rho_0, \rho) = \frac{\sum_{i=1}^R I(W_{n,i}^\beta > \chi_{1,0.05}^2 | \rho_0, \rho)}{R}.$$

As expected, when we get closer to the null hypothesis and when decreasing the sample sizes, the power decreases. With pure data, the best behavior is obtained with low values of  $\beta$ , and with this level of contamination (5%), the best results are obtained for medium values of  $\beta$ .

**Table 9.** Powers for pure data,  $\rho_0 = 0.15$ .

n = 100				n = 200				n = 300			
$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$
$\beta = 0$	0.945	0.603	0.141	1	0.871	0.180	1	0.962	0.265		
$\beta = 0.1$	0.954	0.588	0.157	1	0.863	0.207	1	0.96	0.299		
$\beta = 0.2$	0.952	0.557	0.158	1	0.825	0.213	1	0.944	0.315		
$\beta = 0.3$	0.941	0.510	0.153	0.999	0.783	0.213	1	0.913	0.313		
$\beta = 0.4$	0.925	0.465	0.154	0.999	0.734	0.210	1	0.885	0.301		
$\beta = 0.5$	0.904	0.424	0.159	0.996	0.677	0.202	1	0.845	0.289		
$\beta = 0.6$	0.873	0.395	0.153	0.990	0.618	0.197	0.999	0.789	0.277		
$\beta = 0.7$	0.830	0.361	0.153	0.985	0.555	0.183	0.999	0.733	0.261		
$\beta = 0.8$	0.789	0.322	0.161	0.974	0.499	0.179	0.997	0.678	0.246		

**Table 10.** Powers for contaminated data (5%),  $\rho_0 = 0.15$ .

	$n = 100$			$n = 200$			$n = 300$		
	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$	$\rho = -0.1$	$\rho = 0$	$\rho = 0.15$
$\beta = 0$	0.424	0.090	0.029	0.746	0.141	0.030	0.919	0.246	0.037
$\beta = 0.1$	0.716	0.222	0.041	0.954	0.397	0.029	0.994	0.569	0.037
$\beta = 0.2$	0.838	0.333	0.071	0.989	0.555	0.075	0.999	0.744	0.096
$\beta = 0.3$	0.881	0.383	0.105	0.993	0.633	0.121	0.999	0.803	0.161
$\beta = 0.4$	0.879	0.393	0.129	0.993	0.642	0.150	0.999	0.809	0.213
$\beta = 0.5$	0.865	0.381	0.135	0.992	0.621	0.168	0.999	0.797	0.241
$\beta = 0.6$	0.836	0.357	0.149	0.984	0.583	0.174	0.998	0.769	0.252
$\beta = 0.7$	0.808	0.332	0.146	0.980	0.531	0.173	0.997	0.713	0.256
$\beta = 0.8$	0.773	0.309	0.152	0.961	0.487	0.173	0.995	0.657	0.243

#### 4. Conclusions

The likelihood function is the basis of the maximum likelihood method in estimation theory, and it also plays a key role in the development of log-likelihood ratio tests. However, it is not so tractable in many cases, in practice. Maximum likelihood estimators are based on the likelihood function, and they can be easily obtained; however, there are cases where they do not exist or they cannot be obtained. In such a case, composite likelihood methods constitute an appealing methodology in the area of estimation and testing of hypotheses. On the other hand, the distance or divergence based on methods of estimation and testing have increasingly become fundamental tools in the field of mathematical statistics. The work in [13] is the first, to the best of our knowledge, to link the notion of composite likelihood with divergence based on methods for testing statistical hypotheses.

In this paper, MDPDE are introduced, and they are exploited to develop Wald-type test statistics for testing simple or composite null hypotheses, in a composite likelihood framework. The validity of the proposed procedures is investigated by means of simulations. The simulation results point out the robustness of the proposed information theoretic procedures in estimation and testing, in the composite likelihood context. There are several areas where the notions of divergence and composite likelihood are crucial, including spatial statistics and time series analysis. These are areas of interest, and they will be explored elsewhere.

**Acknowledgments:** We would like to thank the referees for their helpful comments and suggestions. Their comments have improved the paper. This research is supported by Grant MTM2015-67057-P, from Ministerio de Economía y Competitividad (Spain).

**Author Contributions:** All authors conceived and designed the study, conducted the numerical simulation and wrote the paper. All authors read and approved the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

#### Abbreviations

The following abbreviations are used in this manuscript:

MLE	Maximum likelihood estimator
CMLE	Composite maximum likelihood estimator
DPD	Density power divergence
MDPDE	Minimum density power divergence estimator
CMDPDE	Composite minimum density power divergence estimator
RMSE	Root of mean square error
MAE	Mean absolute error

## Appendix A. Proof of the Results

### Appendix A.1. Proof of Theorem 2

The result follows in a straightforward manner because of the asymptotic normality of  $\hat{\theta}_c^\beta$ ,

$$\sqrt{n}(\hat{\theta}_c^\beta - \theta_0) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, (H_\beta(\theta_0))^{-1} J_\beta(\theta_0) (H_\beta(\theta_0))^{-1}\right).$$

### Appendix A.2. Proof of Theorem 3

A first order Taylor expansion of  $l(\theta)$  at  $\hat{\theta}_c^\beta$  around  $\theta^*$  gives:

$$l(\hat{\theta}_c^\beta) - l(\theta^*) = \left( \frac{\partial l(\theta)}{\partial \theta} \right)_{\theta=\theta^*} (\hat{\theta}_c^\beta - \theta^*) + o_p\left(\|\hat{\theta}_c^\beta - \theta^*\|\right).$$

Now, the result follows because the asymptotic distribution of  $(l(\hat{\theta}_c^\beta) - l(\theta^*))$  coincides with the asymptotic distribution of  $\sqrt{n} \left( \frac{\partial l(\theta)}{\partial \theta} \right)_{\theta=\theta^*} (\hat{\theta}_c^\beta - \theta^*)$ .

### Appendix A.3. Proof of Theorem 4

We have:

$$\begin{aligned} g(\hat{\theta}_c^\beta) &= g(\theta_0) + G(\theta_0)^T (\hat{\theta}_c^\beta - \theta_0) + o_p\left(\|\hat{\theta}_c^\beta - \theta_0\|\right) \\ &= G(\theta_0)^T (\hat{\theta}_c^\beta - \theta_0) + o_p\left(\|\hat{\theta}_c^\beta - \theta_0\|\right), \end{aligned}$$

because  $g(\theta_0) = \mathbf{0}_r$ .

Therefore:

$$\sqrt{n}g(\hat{\theta}_c^\beta) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(\mathbf{0}, G_\beta(\theta_0)^T (H_\beta(\theta_0))^{-1} J_\beta(\theta_0) (H_\beta(\theta_0))^{-1} G_\beta(\theta_0))$$

because:

$$\sqrt{n}(\hat{\theta}_c^\beta - \theta_0) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N(\mathbf{0}, (H_\beta(\theta_0))^{-1} J_\beta(\theta_0) (H_\beta(\theta_0))^{-1}).$$

Now:

$$W_{n,\beta} = n g(\hat{\theta}_\beta)^T \left[ G(\theta_0)^T (H_\beta(\theta_0))^{-1} J_\beta(\theta_0) (H_\beta(\theta_0))^{-1} G(\theta_0) \right]^{-1} g(\hat{\theta}_\beta) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi_r^2.$$

### Appendix A.4. CMDPE for the Numerical Example

The estimator  $\hat{\theta}_c^\beta$  is obtained by maximizing Expression (6) with respect to  $\theta$ . Firstly, we are going to get:

$$\begin{aligned} \int_{\mathbb{R}^4} \frac{\partial \mathcal{CL}(\theta, y)^{1+\beta}}{\partial \theta} dy &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}^4} \mathcal{CL}(\theta, y)^{1+\beta} dy \\ &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}^4} f_{12}(\mu_1, \mu_2, \rho, y_1, y_2)^{\beta+1} f_{34}(\mu_3, \mu_4, \rho, y_3, y_4)^{\beta+1} dy_1 dy_2 dy_3 dy_4 \\ &= \frac{\partial}{\partial \theta} \left( \int_{\mathbb{R}^2} f_{12}(\mu_1, \mu_2, \rho, y_1, y_2)^{\beta+1} dy_1 dy_2 \int_{\mathbb{R}^2} f_{34}(\mu_3, \mu_4, \rho, y_3, y_4)^{\beta+1} dy_3 dy_4 \right). \end{aligned}$$

Based on [14] (p. 32):

$$\int_{\mathbb{R}^2} f_{12}(\mu_1, \mu_2, \rho, y_1, y_2)^{\beta+1} dy_1 dy_2 = \int_{\mathbb{R}^2} f_{34}(\mu_3, \mu_4, \rho, y_3, y_4)^{\beta+1} dy_3 dy_4 = \frac{(1-\rho^2)^{-\frac{\beta}{2}}}{\beta+1} (2\pi)^{-\beta}.$$

Then:

$$\int_{\mathbb{R}^4} \frac{\partial \mathcal{C}\mathcal{L}(\theta, y)^{1+\beta}}{\partial \theta} d\mathbf{y} = \frac{\partial}{\partial \theta} \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}(\theta, y)^{1+\beta} d\mathbf{y} = \frac{\partial}{\partial \theta} \frac{(1-\rho^2)^{-\beta}}{(\beta+1)^2} (2\pi)^{-2\beta}$$

and:

$$\frac{\partial}{\partial \mu_i} \frac{(1-\rho^2)^{-\beta}}{(\beta+1)^2} (2\pi)^{-2\beta} = 0, \quad i = 1, 2, 3, 4,$$

while:

$$\frac{\partial}{\partial \rho} \frac{(1-\rho^2)^{-\beta}}{(\beta+1)^2} (2\pi)^{-2\beta} = \frac{\beta(2\pi)^{-2\beta}}{(\beta+1)^2} \frac{2\rho}{(1-\rho^2)^{\beta+1}}.$$

Now, we are going to get:

$$\frac{1}{n\beta} \sum_{i=1}^n \frac{\partial \mathcal{C}\mathcal{L}(\theta, y_i)^\beta}{\partial \theta}$$

in order to obtain the CMDPDE,  $\hat{\theta}_c^\beta$ , by maximizing (6) with respect to  $\theta$ .

We have,

$$\mathcal{C}\mathcal{L}(\theta, y)^\beta = f_{12}(\mu_1, \mu_2, \rho, y_1, y_2)^\beta f_{34}(\mu_3, \mu_4, \rho, y_3, y_4)^\beta.$$

Therefore,

$$\frac{\partial \mathcal{C}\mathcal{L}(\theta, y_i)^\beta}{\partial \mu_1} = \beta f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^{\beta-1} \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{1i} - \mu_1) + 2\rho(y_{2i} - \mu_2)] \right\} f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta$$

and the expression:

$$\frac{1}{n\beta} \sum_{i=1}^n \frac{\partial \mathcal{C}\mathcal{L}(\theta, y_i)^\beta}{\partial \mu_1} = 0$$

leads to the estimator of  $\mu_1$ , given by:

$$\frac{1}{n} \sum_{i=1}^n f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^{\beta-1} f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{1i} - \mu_1) + 2\rho(y_{2i} - \mu_2)] \right\} = 0. \quad (\text{A1})$$

In a similar way:

$$\frac{\partial \mathcal{C}\mathcal{L}(\theta, y_i)^\beta}{\partial \mu_2} = \beta f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^{\beta-1} \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{2i} - \mu_2) + 2\rho(y_{1i} - \mu_1)] \right\} f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta,$$

$$\frac{\partial \mathcal{C}\mathcal{L}(\theta, y_i)^\beta}{\partial \mu_3} = \beta f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^\beta \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{3i} - \mu_3) + 2\rho(y_{4i} - \mu_4)] \right\} f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^{\beta-1}$$

and:

$$\frac{\partial \mathcal{C}\mathcal{L}(\theta, y_i)^\beta}{\partial \mu_4} = \beta f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^\beta \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{4i} - \mu_4) + 2\rho(y_{3i} - \mu_3)] \right\} f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^{\beta-1}.$$

Therefore, the equations:

$$\frac{1}{n\beta} \sum_{i=1}^n \frac{\partial \mathcal{C}\mathcal{L}(\theta, y_i)^\beta}{\partial \mu_2} = 0, \quad \frac{1}{n\beta} \sum_{i=1}^n \frac{\partial \mathcal{C}\mathcal{L}(\theta, y_i)^\beta}{\partial \mu_3} = 0 \quad \text{and} \quad \frac{1}{n\beta} \sum_{i=1}^n \frac{\partial \mathcal{C}\mathcal{L}(\theta, y_i)^\beta}{\partial \mu_4} = 0$$

lead to the estimators of  $\mu_2, \mu_3$  and  $\mu_4$ , which should be read as follows:

$$\frac{1}{n} \sum_{i=1}^n f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^{\beta-1} f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{2i} - \mu_2) + 2\rho(y_{1i} - \mu_1)] \right\} = 0, \quad (\text{A2})$$

$$\frac{1}{n} \sum_{i=1}^n f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^{\beta-1} f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{3i} - \mu_3) + 2\rho(y_{4i} - \mu_4)] \right\} = 0 \quad (\text{A3})$$

and:

$$\frac{1}{n} \sum_{i=1}^n f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^\beta f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta \left\{ -\frac{1}{2(1-\rho^2)} [-2(y_{4i} - \mu_4) + 2\rho(y_{3i} - \mu_3)] \right\} = 0. \quad (\text{A4})$$

Now, it is necessary to get:

$$\begin{aligned} \frac{\partial \mathcal{L}(\theta, \mathbf{y}_i)^\beta}{\partial \rho} &= \frac{\partial f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^\beta f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta}{\partial \rho} \\ &= \beta f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^{\beta-1} f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta \frac{\partial f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})}{\partial \rho} \\ &\quad + \beta f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^\beta f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^{\beta-1} \frac{\partial f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})}{\partial \rho}. \end{aligned}$$

However,  $\frac{\partial f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})}{\partial \rho}$  is given by:

$$\begin{aligned} &\frac{1}{2\pi} \frac{(-1)}{(1-\rho^2)} \frac{(-2\rho)}{2(1-\rho^2)^{\frac{1}{2}}} \exp \left\{ \frac{(-1)}{2(1-\rho^2)} [(y_{1i} - \mu_1)^2 - 2\rho(y_{1i} - \mu_1)(y_{2i} - \mu_2) + (y_{2i} - \mu_2)^2] \right\} \\ &+ \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} \exp \left\{ \frac{(-1)}{2(1-\rho^2)} [(y_{1i} - \mu_1)^2 - 2\rho(y_{1i} - \mu_1)(y_{2i} - \mu_2) + (y_{2i} - \mu_2)^2] \right\} \\ &\left[ \frac{-\rho}{(1-\rho^2)^2} ((y_{1i} - \mu_1)^2 - 2\rho(y_{1i} - \mu_1)(y_{2i} - \mu_2) + (y_{2i} - \mu_2)^2) + \frac{1}{(1-\rho^2)} (y_{1i} - \mu_1)(y_{2i} - \mu_2) \right] \\ &= \frac{\rho}{1-\rho^2} f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i}) + f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i}) \\ &\left[ \frac{-\rho}{(1-\rho^2)^2} ((y_{1i} - \mu_1)^2 - 2\rho(y_{1i} - \mu_1)(y_{2i} - \mu_2) + (y_{2i} - \mu_2)^2) + \frac{1}{(1-\rho^2)} (y_{1i} - \mu_1)(y_{2i} - \mu_2) \right] \\ &= f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i}) \frac{\rho}{1-\rho^2} \left[ 1 - \frac{1}{1-\rho^2} ((y_{1i} - \mu_1)^2 - 2\rho(y_{1i} - \mu_1)(y_{2i} - \mu_2) + (y_{2i} - \mu_2)^2) \right. \\ &\quad \left. + \frac{1}{\rho} (y_{1i} - \mu_1)(y_{2i} - \mu_2) \right]. \end{aligned}$$

In a similar way,  $\frac{\partial f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})}{\partial \rho}$  is given by:

$$\begin{aligned} &f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i}) \frac{\rho}{1-\rho^2} \left[ 1 - \frac{1}{1-\rho^2} ((y_{3i} - \mu_3)^2 - 2\rho(y_{3i} - \mu_3)(y_{4i} - \mu_4) + (y_{4i} - \mu_4)^2) \right. \\ &\quad \left. + \frac{1}{\rho} (y_{3i} - \mu_3)(y_{4i} - \mu_4) \right]. \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\partial \mathcal{L}(\theta, \mathbf{y}_i)^\beta}{\partial \rho} &= \frac{\rho}{1-\rho^2} \beta f_{12}(\mu_1, \mu_2, \rho, y_{1i}, y_{2i})^\beta f_{34}(\mu_3, \mu_4, \rho, y_{3i}, y_{4i})^\beta \\ &\quad \left\{ 2 + \frac{1}{\rho} \{ (y_{1i} - \mu_1)(y_{2i} - \mu_2) + (y_{3i} - \mu_3)(y_{4i} - \mu_4) \} \right. \\ &\quad - \frac{1}{1-\rho^2} ((y_{1i} - \mu_1)^2 - 2\rho(y_{1i} - \mu_1)(y_{2i} - \mu_2) + (y_{2i} - \mu_2)^2) \\ &\quad \left. - \frac{1}{1-\rho^2} ((y_{3i} - \mu_3)^2 - 2\rho(y_{3i} - \mu_3)(y_{4i} - \mu_4) + (y_{4i} - \mu_4)^2) \right\}. \quad (\text{A5}) \end{aligned}$$

Therefore, the equation in relation to  $\rho$  is given by:

$$\frac{1}{n\beta} \sum_{i=1}^n \frac{\partial \mathcal{L}(\theta, \mathbf{y}_i)^\beta}{\partial \rho} - \frac{1}{\beta+1} \int_{\mathbb{R}^m} \frac{\partial \mathcal{L}(\theta, \mathbf{y}_i)^\beta}{\partial \rho} d\mathbf{y} = 0$$

being:

$$\int_{\mathbb{R}^m} \frac{\partial \mathcal{CL}(\theta, y_i)^{\beta+1}}{\partial \theta} dy = \frac{\beta(2\pi)^{-2\beta}}{(\beta+1)^2} \frac{2\rho}{(1-\rho^2)^{\beta+1}} \quad (\text{A6})$$

and:

$$\frac{\partial \mathcal{CL}(\theta, y_i)^\beta}{\partial \rho}$$

was given in (A5).

Finally,

$$\hat{\theta}_c^\beta = \left( \hat{\mu}_{1,c}^\beta, \hat{\mu}_{2,c}^\beta, \hat{\mu}_{3,c}^\beta, \hat{\mu}_{4,c}^\beta, \hat{\rho}_c^\beta \right)^T$$

will be obtained as the solution of the system of equations given by (A1)–(A6).

#### Appendix A.5. Computation of Sensitivity and Variability Matrices in the Numerical Example

We want to compute:

$$\begin{aligned} H_\beta(\theta) &= \int_{\mathbb{R}^m} \mathcal{CL}(\theta, y)^{\beta+1} u(\theta, y)^T u(\theta, y) dy \\ J_\beta(\theta) &= \int_{\mathbb{R}^m} \mathcal{CL}(\theta, y)^{2\beta+1} u(\theta, y)^T u(\theta, y) dy \\ &\quad - \int_{\mathbb{R}^m} \mathcal{CL}(\theta, y)^{\beta+1} u(\theta, y) dy \int_{\mathbb{R}^m} (u(\theta, y))^T \mathcal{CL}(\theta, y)^{\beta+1} dy. \end{aligned}$$

First of all, we can see that:

$$\begin{aligned} \mathcal{CL}(\theta, y)^{\beta+1} &= (f_{A_1}(\theta, y) f_{A_2}(\theta, y))^{\beta+1} \\ &= \left( \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} Q(y_1, y_2)\right\} \cdot \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} Q(y_3, y_4)\right\} \right)^{\beta+1} \\ &= \left( \frac{1}{(2\pi)^2(1-\rho^2)} \right)^{\beta+1} \exp\left\{-\frac{\beta+1}{2(1-\rho^2)} [Q(y_1, y_2) + Q(y_3, y_4)]\right\} \\ &= \frac{1}{(\beta+1)^2} \left( \frac{1}{(2\pi)^2(1-\rho^2)} \right)^\beta \frac{(\beta+1)^2}{(2\pi)^2(1-\rho^2)} \exp\left\{-\frac{\beta+1}{2(1-\rho^2)} [Q(y_1, y_2) + Q(y_3, y_4)]\right\} \\ &= C_\beta \cdot \mathcal{CL}_\beta^*, \end{aligned}$$

where  $C_\beta = \frac{1}{(\beta+1)^2} \left( \frac{1}{(2\pi)^2(1-\rho^2)} \right)^\beta$  and  $\mathcal{CL}_\beta^* = \mathcal{CL}_\beta(\theta, y)^* \sim \mathcal{N}(\mu, \Sigma^*)$ , with  $\Sigma^* = \frac{1}{\beta+1} \Sigma$ .

While  $u(\theta, y) = \frac{\partial \log \mathcal{CL}(\theta, y)}{\partial \theta}$ , we will denote as  $u(\theta, y)^*$  to  $u(\theta, y)^* = \frac{\partial \log \mathcal{CL}_\beta^*}{\partial \theta}$ . Then:

$$\begin{aligned} u(\theta, y) &= \frac{\partial \log \mathcal{CL}(\theta, y)}{\partial \theta} = \frac{1}{\beta+1} \frac{\partial \log \mathcal{CL}(\theta, y)^{\beta+1}}{\partial \theta} = \frac{1}{\beta+1} \frac{\partial \log (C_\beta \cdot \mathcal{CL}_\beta^*)}{\partial \theta} \\ &= \frac{1}{\beta+1} \left( \frac{\partial \log C_\beta}{\partial \theta} + \frac{\partial \log \mathcal{CL}_\beta^*}{\partial \theta} \right) = \frac{1}{\beta+1} \left( \frac{\partial \log C_\beta}{\partial \theta} + u(\theta, y)^* \right). \end{aligned} \quad (\text{A7})$$

Further,

$$\begin{aligned} \int_{\mathbb{R}^m} \mathcal{CL}(\theta, y)^{\beta+1} u(\theta, y) dy &= \int_{\mathbb{R}^m} \mathcal{CL}(\theta, y)^{\beta+1} \frac{\partial \log \mathcal{CL}(\theta, y)}{\partial \theta} dy = \int_{\mathbb{R}^m} \mathcal{CL}(\theta, y)^\beta \frac{\partial \mathcal{CL}(\theta, y)}{\partial \theta} dy \\ &= \int_{\mathbb{R}^m} \frac{1}{\beta+1} \frac{\partial \mathcal{CL}(\theta, y)^{\beta+1}}{\partial \theta} dy = \frac{1}{\beta+1} \frac{\partial}{\partial \theta} \int_{\mathbb{R}^m} \mathcal{CL}(\theta, y)^{\beta+1} dy \\ &= \frac{1}{\beta+1} \frac{\partial C_\beta}{\partial \theta} = (0, 0, 0, 0, \frac{2\rho\beta C_\beta}{(\beta+1)(1-\rho^2)})^T = \xi_\beta(\theta). \end{aligned} \quad (\text{A8})$$

Now:

$$\begin{aligned}
 & \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}^{\beta+1}u(\theta, y)^T u(\theta, y) dy \\
 &= \int_{\mathbb{R}^4} (\mathcal{C}_\beta \cdot \mathcal{C}\mathcal{L}_\beta^*) \frac{1}{(\beta+1)^2} \left( \frac{\partial \log C_\beta}{\partial \theta} + u(\theta, y)^* \right)^T \left( \frac{\partial \log C_\beta}{\partial \theta} + u(\theta, y)^* \right) dy \\
 &= \frac{C_\beta}{(\beta+1)^2} \int_{\mathbb{R}^4} \left[ \left( \frac{\partial \log C_\beta}{\partial \theta} \right)^T \left( \frac{\partial \log C_\beta}{\partial \theta} \right) \mathcal{C}\mathcal{L}_\beta^* \right. \\
 &\quad \left. + \mathcal{C}\mathcal{L}_\beta^* (u(\theta, y)^*)^T \frac{\partial \log C_\beta}{\partial \theta} + \mathcal{C}\mathcal{L}_\beta^* \left( \frac{\partial \log C_\beta}{\partial \theta} \right)^T u(\theta, y)^* + \mathcal{C}\mathcal{L}_\beta^* (u(\theta, y)^*)^T u(\theta, y)^* \right] dy \\
 &= \frac{C_\beta}{(\beta+1)^2} \left[ \left( \frac{\partial \log C_\beta}{\partial \theta} \right)^T \left( \frac{\partial \log C_\beta}{\partial \theta} \right) \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}_\beta^* dy + \left( \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}_\beta^* u(\theta, y)^* dy \right)^T \left( \frac{\partial \log C_\beta}{\partial \theta} \right) \right. \\
 &\quad \left. + \left( \frac{\partial \log C_\beta}{\partial \theta} \right)^T \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}_\beta^* u(\theta, y)^* dy + \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}_\beta^* (u(\theta, y)^*)^T u(\theta, y)^* dy \right] \\
 &= \frac{C_\beta}{(\beta+1)^2} \left[ K^T K + \left( \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}_\beta^* u(\theta, y)^* dy \right)^T K + K^T \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}_\beta^* u(\theta, y)^* dy + \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}_\beta^* (u(\theta, y)^*)^T u(\theta, y)^* dy \right],
 \end{aligned} \tag{A9}$$

where  $K = \frac{\partial \log C_\beta}{\partial \theta} = (0, 0, 0, 0, \frac{2\rho \cdot \beta}{1-\rho^2})$ . However:

$$\begin{aligned}
 \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}_\beta^* u(\theta, y)^* dy &= \int_{\mathbb{R}^4} \left( \frac{1}{C_\beta} \mathcal{C}\mathcal{L}(\theta, y)^{\beta+1} \right) \left[ (\beta+1)u(\theta, y) - \frac{\partial \log C_\beta}{\partial \theta} \right] dy \\
 &= \frac{\beta+1}{C_\beta} \left[ \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}(\theta, y)^{\beta+1} u(\theta, y) dy \right] - \frac{K}{C_\beta} \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}(\theta, y)^{\beta+1} dy \\
 &= \frac{1}{C_\beta} \frac{\partial C_\beta}{\partial \theta} - K = K - K = \mathbf{0},
 \end{aligned}$$

and thus, (A9) can be expressed as:

$$\int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}(\theta, y)^{\beta+1} u(\theta, y)^T u(\theta, y) dy = \frac{C_\beta}{(\beta+1)^2} \left[ K^T K + \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}_\beta^* (u(\theta, y)^*)^T u(\theta, y)^* dy \right].$$

On the other hand, it is not difficult to prove that:

$$\int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}_\beta^* (u(\theta, y)^*)^T u(\theta, y)^* dy = C \cdot \int_{\mathbb{R}^4} \mathcal{C}\mathcal{L}(\theta, y) u(\theta, y)^T u(\theta, y) dy = C \cdot H_0(\theta),$$

where  $C = diag(\beta+1, \beta+1, \beta+1, \beta+1, 1)$  and ([13]):

$$H_0(\theta) = \begin{pmatrix} \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} & 0 & 0 & 0 \\ \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} & 0 \\ 0 & 0 & \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & 0 \\ 0 & 0 & 0 & 0 & \frac{2(\rho^2+1)}{(1-\rho^2)^2} \end{pmatrix}. \tag{A10}$$

Therefore,

$$H_\beta(\theta) = \frac{C_\beta}{(\beta+1)^2} [C \cdot H_0(\theta) + K^T K],$$

that is:

$$\mathbf{H}_\beta(\boldsymbol{\theta}) = \frac{C_\beta}{(\beta+1)(1-\rho^2)} \begin{pmatrix} 1 & -\rho & 0 & 0 & 0 \\ -\rho & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -\rho & 0 \\ 0 & 0 & -\rho & 1 & 0 \\ 0 & 0 & 0 & 0 & 2\frac{(\rho^2+1)+2\rho^2\beta^2}{(1-\rho^2)(1+\beta)} \end{pmatrix}. \quad (\text{A11})$$

Note that, for  $\beta = 0$ , (A11) reduces to (A10).

On the other hand, the expression of the variability matrix  $\mathbf{J}_\beta(\boldsymbol{\theta})$  can be obtained from Expressions (27) and (A8) as:

$$\mathbf{J}_\beta(\boldsymbol{\theta}) = \mathbf{H}_{2\beta}(\boldsymbol{\theta}) - \boldsymbol{\xi}_\beta(\boldsymbol{\theta})\boldsymbol{\xi}_\beta(\boldsymbol{\theta})^T. \quad (\text{A12})$$

## References

- Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimizing a density power divergence. *Biometrika* **1998**, *85*, 549–559.
- Basu, A.; Mandal, A.; Martín, N.; Pardo, L. Testing statistical hypotheses based on the density power divergence. *Ann. Inst. Stat. Math.* **2013**, *65*, 319–348.
- Basu, A.; Mandal, A.; Martín, N.; Pardo, L. Robust tests for the equality of two normal means based on the density power divergence. *Metrika* **2015**, *78*, 611–634.
- Basu, A.; Mandal, A.; Martín, N.; Pardo, L. Generalized Wald-type tests based on minimum density power divergence estimators. *Statistics* **2016**, *50*, 1–26.
- Basu, A.; Ghosh, A.; Mandal, A.; Martín, N.; Pardo, L. A Wald-type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator. *Electron. J. Stat.* **2017**, *11*, 2741–2772.
- Ghosh, A.; Mandal, A.; Martín, N.; Pardo, L. Influence analysis of robust Wald-type tests. *J. Multivar. Anal.* **2016**, *147*, 102–126.
- Varin, C.; Reid, N.; Firth, D. An overview of composite likelihood methods. *Stat. Sin.* **2011**, *21*, 4–42.
- Xu, X.; Reid, N. On the robustness of maximum composite estimate. *J. Stat. Plan. Inference* **2011**, *141*, 3047–3054.
- Joe, H.; Reid, N.; Somg, P.X.; Firth, D.; Varin, C. Composite Likelihood Methods. Report on the Workshop on Composite Likelihood; 2012. Available online: <http://www.birs.ca/events/2012/5-day-workshops/12w5046> (accessed on 28 December 2017).
- Lindsay, G. Composite likelihood methods. *Contemp. Math.* **1998**, *80*, 221–239.
- Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapman & Hall/CRC: Boca Raton, FA, USA, 2011.
- Maronna, R.A.; Martin, R.D.; Yohai, V.J. *Time Series, in Robust Statistics: Theory and Methods*; John Wiley & Sons, Ltd.: Chichester, UK, 2006.
- Martín, N.; Pardo, L.; Zografos, K. On divergence tests for composite hypotheses under composite likelihood. In *Statistical Papers*; Springer: Berlin/Heidelberg, Germany, 2017.
- Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman & Hall/CRC: Boca Raton, FA, USA, 2006.



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Composite Tests under Corrupted Data

Michel Broniatowski <sup>1,\*</sup>, Jana Jurečková <sup>2,3</sup>, Ashok Kumar Moses <sup>4</sup> and Emilie Miranda <sup>1,5</sup>

<sup>1</sup> Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, 75005 Paris, France; emilie.miranda@upmc.fr

<sup>2</sup> Institute of Information Theory and Automation, The Czech Academy of Sciences, 18208 Prague, Czech Republic; jurecko@karlin.mff.cuni.cz

<sup>3</sup> Faculty of Mathematics and Physics, Charles University, 18207 Prague, Czech Republic

<sup>4</sup> Department of ECE, Indian Institute of Technology, Palakkad 560012, India; ashokm@iitpkd.ac.in

<sup>5</sup> Safran Aircraft Engines, 77550 Moissy-Cramayel, France

\* Correspondence: michel.broniatowski@sorbonne-universite.fr

Received: 11 November 2018; Accepted: 10 January 2019; Published: 14 January 2019



**Abstract:** This paper focuses on test procedures under corrupted data. We assume that the observations  $Z_i$  are mismeasured, due to the presence of measurement errors. Thus, instead of  $Z_i$  for  $i = 1, \dots, n$ , we observe  $X_i = Z_i + \sqrt{\delta}V_i$ , with an unknown parameter  $\delta$  and an unobservable random variable  $V_i$ . It is assumed that the random variables  $Z_i$  are i.i.d., as are the  $X_i$  and the  $V_i$ . The test procedure aims at deciding between two simple hypotheses pertaining to the density of the variable  $Z_i$ , namely  $f_0$  and  $g_0$ . In this setting, the density of the  $V_i$  is supposed to be known. The procedure which we propose aggregates likelihood ratios for a collection of values of  $\delta$ . A new definition of least-favorable hypotheses for the aggregate family of tests is presented, and a relation with the Kullback-Leibler divergence between the sets  $(f_\delta)_\delta$  and  $(g_\delta)_\delta$  is presented. Finite-sample lower bounds for the power of these tests are presented, both through analytical inequalities and through simulation under the least-favorable hypotheses. Since no optimality holds for the aggregation of likelihood ratio tests, a similar procedure is proposed, replacing the individual likelihood ratio by some divergence based test statistics. It is shown and discussed that the resulting aggregated test may perform better than the aggregate likelihood ratio procedure.

**Keywords:** composite hypotheses; corrupted data; least-favorable hypotheses; Neyman Pearson test; divergence based testing; Chernoff Stein lemma

## 1. Introduction

A situation which is commonly met in quality control is the following: Some characteristic  $Z$  of an item is supposed to be random, and a decision about its distribution has to be made based on a sample of such items, each with the same distribution  $F_0$  (with density  $f_0$ ) or  $G_0$  (with density  $g_0$ ). The measurement device adds a random noise  $V_\delta$  to each measurement, mutually independent and independent of the item, with a common distribution function  $H_\delta$  and density  $h_\delta$ , where  $\delta$  is an unknown scaling parameter. Therefore the density of the measurement  $X := Z + V_\delta$  is either  $f_\delta := f_0 * h_\delta$  or  $g_\delta := g_0 * h_\delta$ , where  $*$  denotes the convolution operation. We denote  $F_\delta$  (respectively  $G_\delta$ ) to be the distribution function with density  $f_\delta$  (respectively  $g_\delta$ ).

The problem of interest, studied in [1], is how the measurement errors can affect the conclusion of the likelihood ratio test with statistics

$$L_n := \frac{1}{n} \sum \log \frac{g_0}{f_0}(X_i).$$

For small  $\delta$ , the result of [2] enables us to estimate the true log-likelihood ratio (true Kullback-Leibler divergence) even when we only dispose of locally perturbed data by additive measurement errors. The distribution function  $H_0$  of the measurement errors is considered unknown, up to zero expectation and unit variance. When we use the likelihood ratio test, while ignoring the possible measurement errors, we can incur a loss in both errors of the first and second kind. However, it is shown, in [1], that for small  $\delta$  the original likelihood ratio test (LRT) is still the most powerful, only on a slightly changed significance level. The test problem leads to a composite of null and alternative classes  $\mathbf{H}_0$  or  $\mathbf{H}_1$  of distributions of random variables  $Z + V_\delta$  with  $V_\delta := \sqrt{\delta}V$ , where  $V$  has distribution  $H_1$ . If those families are bounded by alternating Choquet capacities of order 2, then the minimax test is based on the likelihood ratio of the pair of the least-favorable distributions of  $\mathbf{H}_0$  and  $\mathbf{H}_1$ , respectively (see Huber and Strassen [3]). Moreover, Eguchi and Copas [4] showed that the overall loss of power caused by a misspecified alternative equals the Kullback-Leibler divergence between the original and the corrupted alternatives. Surprisingly, the value of the overall loss is independent of the choice of null hypothesis. The arguments of [2] and of [5] enable us to approximate the loss of power locally, for a broad set of alternatives. The asymptotic behavior of the loss of power of the test based on sampled data is considered in [1], and is supplemented with numerical illustration.

#### *Statement of the Test Problem*

Our aim is to propose a class of statistics for testing the composite hypotheses  $\mathbf{H}_0$  and  $\mathbf{H}_1$ , extending the optimal Neyman-Pearson LRT between  $f_0$  and  $g_0$ . Unlike in [1], the scaling parameter  $\delta$  is not supposed to be small, but merely to belong to some interval bounded away from 0.

We assume that the distribution  $H$  of the random variable (r.v.)  $V$  is known; indeed, in the tuning of the offset of a measurement device, it is customary to perform a large number of observations on the noise under a controlled environment.

Therefore, this first step produces a good basis for the modelling of the distribution of the density  $h$ . Although the distribution of  $V$  is known, under operational conditions the distribution of the noise is modified: For a given  $\delta$  in  $[\delta_{\min}, \delta_{\max}]$  with  $\delta_{\min} > 0$ , denote by  $V_\delta$  a r.v. whose distribution is obtained through some transformation from the distribution of  $V$ , which quantifies the level of the random noise. A classical example is when  $V_\delta = \sqrt{\delta}V$ , but at times we have a weaker assumption, which amounts to some decomposability property with respect to  $\delta$ : For instance, in the Gaussian case, we assume that for all  $\delta, \eta$ , there exists some r.v.  $W_{\delta, \eta}$  such that  $V_{\delta+\eta} =_d V_\delta + W_{\delta, \eta}$ , where  $V_\delta$  and  $W_{\delta, \eta}$  are independent.

The test problem can be stated as follows: A batch of i.i.d. measurements  $X_i := Z_i + V_{\delta,i}$  is performed, where  $\delta > 0$  is unknown, and we consider the family of tests of  $\mathbf{H}_0(\delta):=[X \text{ has density } f_\delta]$  vs.  $\mathbf{H}_1(\delta):=[X \text{ has density } g_\delta]$ , with  $\delta \in \Delta = [\delta_{\min}, \delta_{\max}]$ . Only the  $X_i$  are observed. A class of combined tests of  $\mathbf{H}_0$  vs.  $\mathbf{H}_1$  is proposed, in the spirit of [6–9].

Under every fixed  $n$ , we assume that  $\delta$  is allowed to run over a finite set  $p_n$  of components of the vector  $\Delta_n := [\delta_{\min} = \delta_{0,n}, \dots, \delta_{p_n,n} = \delta_{\max}]$ . The present construction is essentially non-asymptotic, neither on  $n$  nor on  $\delta$ , in contrast with [1], where  $\delta$  was supposed to lie in a small neighborhood of 0. However, with increasing  $n$ , it would be useful to consider that the array  $(\delta_{j,n})_{j=1}^{p_n}$  is getting dense in  $\Delta = [\delta_{\min}, \delta_{\max}]$  and that

$$\lim_{n \rightarrow \infty} \frac{\log p_n}{n} = 0. \quad (1)$$

For the sake of notational brevity, we denote by  $\Delta$  the above grid  $\Delta_n$ , and all suprema or infima over  $\Delta$  are supposed to be over  $\Delta_n$ . For any event  $B$  and any  $\delta$  in  $\Delta$ ,  $F_\delta(B)$  (respectively  $G_\delta(B)$ ) designates the probability of  $B$  under the distribution  $F_\delta$  (respectively  $G_\delta$ ). Given a sequence of levels  $\alpha_n$ , we consider a sequence of test criteria  $T_n := T_n(X_1, \dots, X_n)$  of  $\mathbf{H}_0(\delta)$ , and the pertaining critical regions

$$T_n(X_1, \dots, X_n) > A_n, \quad (2)$$

such that

$$F_\delta(T_n(X_1, \dots, X_n) > A_n) \leq \alpha_n \quad \forall \delta \in \Delta,$$

leading to rejection of  $\mathbf{H}_0(\delta)$  for at least some  $\delta \in \Delta$ .

In an asymptotic context, it is natural to assume that  $\alpha_n$  converges to 0 as  $n$  increases, since an increase in the sample size allows for a smaller first kind risk. For example, in [8],  $\alpha_n$  takes the form  $\alpha_n := \exp\{-na_n\}$  for some sequence  $a_n \rightarrow \infty$ .

In the sequel, the Kullback-Leibler discrepancy between probability measures  $Q$  and  $P$ , with respective densities  $p$  and  $q$  (with respect to the Lebesgue measure on  $\mathbb{R}$ ), is denoted

$$K(Q, P) := \int \log \frac{q(x)}{p(x)} q(x) dx$$

whenever defined, and takes value  $+\infty$  otherwise.

The present paper handles some issues with respect to this context. In Section 2, we consider some test procedures based on the supremum of Likelihood Ratios (LR) for various values of  $\delta$ , and define  $T_n$ . The threshold for such a test is obtained for any level  $\alpha_n$ , and a lower bound for its power is provided. In Section 3, we develop an asymptotic approach to the Least Favorable Hypotheses (LFH) for these tests. We prove that asymptotically least-favorable hypotheses are obtained through minimization of the Kullback-Leibler divergence between the two composite classes  $\mathbf{H}0$  and  $\mathbf{H}1$  independently upon the level of the test.

We next consider, in Section 3.3, the performance of the test numerically; indeed, under the least-favorable pair of hypotheses we compare the power of the test (as obtained through simulation) with the theoretical lower bound, as obtained in Section 2. We show that the minimal power, as measured under the LFH, is indeed larger than the theoretical lower bound—this result shows that the simulation results overperform on theoretical bounds. These results are developed in a number of examples.

Since no argument plays in favor of any type of optimality for the test based on the supremum of likelihood ratios for composite testing, we consider substituting those ratios with other kinds of scores in the family of divergence-based concepts, extending the likelihood ratio in a natural way. Such an approach has a long history, stemming from the seminal book by Liese and Vajda [10]. Extensions of the Kullback-Leibler based criterions (such as the likelihood ratio) to power-type criterions have been proposed for many applications in Physics and in Statistics (see, e.g., [11]). We explore the properties of those new tests under the pair of hypotheses minimizing the Kullback-Leibler divergence between the two composite classes  $\mathbf{H}0$  and  $\mathbf{H}1$ . We show that, in some cases, we can build a test procedure whose properties overperform the above supremum of the LRTs, and we provide an explanation for this fact. This is the scope of Section 4.

## 2. An Extension of the Likelihood Ratio Test

For any  $\delta$  in  $\Delta$ , let

$$T_{n,\delta} := \frac{1}{n} \sum_{i=1}^n \log \frac{g_\delta}{f_\delta}(X_i), \quad (3)$$

and define

$$T_n := \sup_{\delta \in \Delta} T_{n,\delta}.$$

Consider, for fixed  $\delta$ , the Likelihood Ratio Test with statistics  $T_{n,\delta}$  which is uniformly most powerful (UMP) within all tests of  $\mathbf{H}0(\delta) := p_T = f_\delta$  vs.  $\mathbf{H}1(\delta) := p_T = g_\delta$ , where  $p_T$  designates the distribution of the generic r.v.  $X$ . The test procedure to be discussed aims at solving the question: Does there exist some  $\delta$ , for which  $\mathbf{H}0(\delta)$  would be rejected vs.  $\mathbf{H}1(\delta)$ , for some prescribed value of the first kind risk?

Whenever  $\mathbf{H}0(\delta)$  is rejected in favor of  $\mathbf{H}1(\delta)$ , for some  $\delta$ , we reject  $\mathbf{H}0:=f_0 = g_0$  in favor of  $\mathbf{H}1:=f_0 \neq g_0$ . A critical region for this test with level  $\alpha_n$  is defined by

$$T_n > A_n,$$

with

$$\begin{aligned} P_{\mathbf{H}0}(\mathbf{H}1) &= \sup_{\delta \in \Delta} F_\delta(T_n > A_n) \\ &= \sup_{\delta \in \Delta} F_\delta \left( \bigcup_{\delta'} T_{n,\delta'} > A_n \right) \leq \alpha_n. \end{aligned}$$

Since, for any sequence of events  $B_1, \dots, B_{p_n}$ ,

$$F_\delta \left( \bigcup_{k=1}^{p_n} B_k \right) \leq p_n \max_{1 \leq k \leq p_n} F_\delta(B_k),$$

it holds that

$$P_{\mathbf{H}0}(\mathbf{H}1) \leq p_n \max_{\delta \in \Delta} \max_{\delta' \in \Delta} F_\delta(T_{n,\delta'} > A_n). \quad (4)$$

An upper bound for  $P_{\mathbf{H}0}(\mathbf{H}1)$  can be obtained, making use of the Chernoff inequality for the right side of (4), providing an upper bound for the risk of first kind for a given  $A_n$ . The correspondence between  $A_n$  and this risk allows us to define the threshold  $A_n$  accordingly.

Turning to the power of this test, we define the risk of second kind by

$$\begin{aligned} P_{H_1}(\mathbf{H}0) &:= \sup_{\eta \in \Delta} G_\eta(T_n \leq A_n) \\ &= \sup_{\eta \in \Delta} G_\eta \left( \sup_{\delta \in \Delta} T_{n,\delta} \leq A_n \right) \\ &= \sup_{\eta \in \Delta} G_\eta \left( \bigcap_{\delta \in \Delta} T_{n,\delta} \leq A_n \right) \\ &\leq \sup_{\eta \in \Delta} G_\eta(T_{n,\eta} \leq A_n), \end{aligned} \quad (5)$$

a crude bound which, in turn, can be bounded from above through the Chernoff inequality, which yields a lower bound for the power of the test under any hypothesis  $g_\eta$  in  $\mathbf{H}1$ .

Let  $\alpha_n$  denote a sequence of levels, such that

$$\limsup_{n \rightarrow \infty} \alpha_n < 1.$$

We make use of the following hypothesis:

$$\sup_{\delta \in \Delta} \sup_{\delta' \in \Delta} \int \log \frac{f_{\delta'}}{g_{\delta'}} f_\delta < 0. \quad (6)$$

**Remark 1.** Since

$$\int \log \frac{f_{\delta'}}{g_{\delta'}} f_\delta = K(F_\delta, G_{\delta'}) - K(F_\delta, F_{\delta'}),$$

making use of the Chernoff-Stein Lemma (see Theorem A1 in the Appendix A), Hypothesis (6) means that any LRT with  $H0: p_T = f_\delta$  vs.  $H1: p_T = g_{\delta'}$  is asymptotically more powerful than any LRT with  $H0: p_T = f_\delta$  vs.  $H1: p_T = f_{\delta'}$ .

Both hypotheses (7) and (8), which are defined below, are used to provide the critical region and the power of the test.

For all  $\delta, \delta'$  define

$$Z_{\delta'} := \log \frac{g_{\delta'}}{f_{\delta'}}(X),$$

and let

$$\varphi_{\delta, \delta'}(t) := \log E_{F_\delta}(\exp(tZ_{\delta'})) = \log \int \left( \frac{g_{\delta'}(x)}{f_{\delta'}(x)} \right)^t f_\delta(x) dx.$$

With  $\mathcal{N}_{\delta, \delta'}$ , the set of all  $t$  such that  $\varphi_{\delta, \delta'}(t)$  is finite, we assume

$$\mathcal{N}_{\delta, \delta'} \text{ is a non void open neighborhood of } 0. \quad (7)$$

Define, further,

$$J_{\delta, \delta'}(x) := \sup_t tx - \varphi_{\delta, \delta'}(t),$$

and let

$$J(x) := \min_{(\delta, \delta') \in \Delta \times \Delta} J_{\delta, \delta'}(x).$$

For any  $\eta$ , let

$$W_\eta := -\log \frac{g_\eta}{f_\eta}(X),$$

and let

$$\psi_\eta(t) := \log E_{G_\eta}(\exp(tW_\eta)).$$

Let  $\mathcal{M}_\eta$  be the set of all  $t$  such that  $\psi_\eta(t)$  is finite. Assume

$$\mathcal{M}_\eta \text{ is a non void neighborhood of } 0. \quad (8)$$

Let

$$I_\eta(x) := \sup_t tx - \log E_{G_\eta}(\exp(tW_\eta)), \quad (9)$$

and

$$I(x) := \inf_\eta I_\eta(x).$$

We also assume an accessory condition on the support of  $Z_{\delta'}$  and  $W_\eta$ , respectively under  $F_\delta$  and under  $G_\eta$  (see (A2) and (A5) in the proof of Theorem A1). Suppose the regularity assumptions (7) and (8) are fulfilled for all  $\delta, \delta'$  and  $\eta$ . Assume, further, that  $p_n$  fulfills (1).

The following result holds:

**Proposition 2.** Whenever (6) holds, for any sequence of levels  $\alpha_n$  bounded away from 1, defining

$$A_n := J^{-1} \left( -\frac{1}{n} \log \frac{\alpha_n}{p_n} \right),$$

it holds, for large  $n$ , that

$$P_{\mathbf{H}0}(\mathbf{H}1) = \sup_{\delta \in \Delta} F_\delta(T_n > A_n) \leq \alpha_n$$

and

$$P_{\mathbf{H}1}(\mathbf{H}1) = \sup_{\delta \in \Delta} G_\delta(T_n > A_n) \geq 1 - \exp(-nI(A_n)).$$

### 3. Minimax Tests under Noisy Data, Least-Favorable Hypotheses

#### 3.1. An Asymptotic Definition for the Least-Favorable Hypotheses

We prove that the above procedure is asymptotically minimax for testing the composite hypothesis  $H_0$  against the composite alternative  $H_1$ . Indeed, we identify the least-favorable hypotheses, say  $F_{\delta_*} \in H_0$  and  $G_{\delta_*} \in H_1$ , which lead to minimal power and maximal first kind risk for these tests. This requires a discussion of the definition and existence of such a least-favourable pair of hypotheses in an asymptotic context; indeed, for a fixed sample size, the usual definition only leads to an explicit definition in very specific cases. Unlike in [1], the minimax tests will not be in the sense of Huber and Strassen. Indeed, on one hand, hypotheses  $H_0$  and  $H_1$  are not defined in topological neighbourhoods of  $F_0$  and  $G_0$ , but rather through a convolution under a parametric setting. On the other hand, the specific test of  $\{H_0(\delta), \delta \in \Delta\}$  against  $\{H_1(\delta), \delta \in \Delta\}$  does not require capacities dominating the corresponding probability measures.

Throughout the subsequent text, we shall assume that there exists  $\delta_*$  such that

$$\min_{\delta \in \Delta} K(F_\delta, G_\delta) = K(F_{\delta_*}, G_{\delta_*}). \quad (10)$$

We shall call the pair of distributions  $(F_{\underline{\delta}}, G_{\underline{\delta}})$  least-favorable for the sequence of tests  $1\{T_n > A_n\}$  if they satisfy

$$\begin{aligned} F_\delta(T_n \leq A_n) &\geq F_{\underline{\delta}}(T_n \leq A_n) \\ &\geq G_{\underline{\delta}}(T_n \leq A_n) \geq G_\delta(T_n \leq A_n) \end{aligned} \quad (11)$$

for all  $\delta \in \Delta$ . The condition of unbiasedness of the test is captured by the central inequality in (11).

Because, for finite  $n$ , such a pair can be constructed only in few cases, we should take a recourse of (11) to the asymptotics  $n \rightarrow \infty$ . We shall show that any pair of distributions  $(F_{\delta_*}, G_{\delta_*})$  achieving (10) are least-favorable. Indeed, it satisfies the inequality (11) asymptotically on the logarithmic scale.

Specifically, we say that  $(F_{\underline{\delta}}, G_{\underline{\delta}})$  is a least-favorable pair of distributions when, for any  $\delta \in \Delta$ ,

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{1}{n} \log F_{\underline{\delta}}(T_n \leq A_n) &\geq \lim_{n \rightarrow \infty} \frac{1}{n} \log G_{\underline{\delta}}(T_n \leq A_n) \\ &\geq \limsup_{n \rightarrow \infty} \frac{1}{n} \log G_\delta(T_n \leq A_n). \end{aligned} \quad (12)$$

Define the total variation distance

$$d_{TV}(F_\delta, G_\delta) := \sup_B |F_\delta(B) - G_\delta(B)|,$$

where the supremum is over all Borel sets  $B$  of  $\mathbb{R}$ . We will assume that, for all  $n$ ,

$$\alpha_n < 1 - \sup_{\delta \in \Delta} d_{TV}(F_\delta, G_\delta). \quad (13)$$

We state our main result, whose proof is deferred to the Appendix B.

**Theorem 3.** *For any level  $\alpha_n$  satisfying (13), the pair  $(F_{\delta_*}, G_{\delta_*})$  is a least-favorable pair of hypotheses for the family of tests  $1\{T_n \geq A_n\}$ , in the sense of (12).*

#### 3.2. Identifying the Least-Favorable Hypotheses

We now concentrate on (10).

For given  $\delta \in [\delta_{\min}, \delta_{\max}]$  with  $\delta_{\min} > 0$ , the distribution of the r.v.  $V_\delta$  is obtained through some transformation from the known distribution of  $V$ . The classical example is  $V_\delta = \sqrt{\delta}V$ , which is a scaling, and where  $\sqrt{\delta}$  is the signal to noise ratio. The following results state that the Kullback-Leibler discrepancy  $K(F_\delta, G_\delta)$  reaches its minimal value when the noise  $V_\delta$  is “maximal”, under some additivity property with respect to  $\delta$ . This result is not surprising: Adding noise deteriorates the ability to discriminate between the two distributions  $F_0$  and  $G_0$ —this effect is captured in  $K(F_\delta, G_\delta)$ , which takes its minimal value for the maximal  $\delta$ .

**Proposition 4.** Assume that, for all  $\delta, \eta$ , there exists some r.v.  $W_{\delta, \eta}$  such that  $V_{\delta+\eta} =_d V_\delta + W_{\delta, \eta}$  where  $V_\delta$  and  $W_{\delta, \eta}$  are independent. Then

$$\delta_* = \delta_{\max}.$$

This result holds as a consequence of Lemma A5 in the Appendix C.

In the Gaussian case, when  $h$  is the standard normal density, Proposition 4 holds, since  $h_{\delta+\eta} = h_\delta * h_{\eta-\delta}$  with  $h_\varepsilon(x) := (1/\sqrt{\varepsilon}) h(x/\sqrt{\varepsilon})$ . In order to model symmetric noise, we may consider a symmetrized Gamma density as follows: Set  $h_\delta(x) := (1/2) \gamma^+(1, \delta)(x) + (1/2) \gamma^-(1, \delta)(x)$ , where  $\gamma^+(1, \delta)$  designates the Gamma density with scale parameter 1 and shape parameter  $\delta$ , and  $\gamma^-(1, \delta)$  the Gamma density on  $\mathbb{R}^-$  with same parameter. Hence a r.v. with density  $h_\delta$  is symmetrically distributed and has variance  $2\delta$ . Clearly,  $h_{\delta+\eta}(x) = h_\delta * h_\eta(x)$ , which shows that Proposition 4 also holds in this case. Note that, except for values of  $\delta$  less than or equal to 1, the density  $h_\delta$  is bimodal, which does not play in favour of such densities for modelling the uncertainty, due to the noise. In contrast with the Gaussian case,  $h_\delta$  cannot be obtained from  $h_1$  by any scaling. The centred Cauchy distribution may help as a description of heavy tailed symmetric noise, and keeps uni-modality through convolution; it satisfies the requirements of Proposition 4 since  $f_\delta * f_\eta(x) = f_{\delta+\eta}(x)$  where  $f_\varepsilon(x) := \varepsilon/\pi(x^2 + \varepsilon^2)$ . In this case,  $\delta$  acts as a scaling, since  $f_\delta$  is the density of  $\delta X$  where  $X$  has density  $f_1$ .

In practice, the interesting case is when  $\delta$  is the variance of the noise and corresponds to a scaling of a generic density, as occurs for the Gaussian case or for the Cauchy case. In the examples, which will be used below, we also consider symmetric, exponentially distributed densities (Laplace densities) or symmetric Weibull densities with a given shape parameter. The Weibull distribution also fulfills the condition in Proposition 4, being infinitely divisible (see [12]).

### 3.3. Numerical Performances of the Minimax Test

As frequently observed, numerical results deduced from theoretical bounds are of poor interest due to the sub-optimality of the involved inequalities. They may be sharpened on specific cases. This motivates the need for simulation. We study two cases, which can be considered as benchmarks.

- A. In the first case,  $f_0$  is a normal density with expectation 0 and variance 1, whereas  $g_0$  is a normal density with expectation 0.3 and variance 1.
- B. The second case handles a situation where  $f_0$  and  $g_0$  belong to different models:  $f_0$  is a log-normal density with location parameter 1 and scale parameter 0.2, whereas  $g_0$  is a Weibull density on  $\mathbb{R}^+$  with shape parameter 5 and scale parameter 3. Those two densities differ strongly, in terms of asymptotic decay. They are, however, very close to one another in terms of their symmetrized Kullback-Leibler divergence (the so-called Jeffrey distance). Indeed, centering on the log-normal distribution  $f_0$ , the closest among all Weibull densities is at distance 0.10—the density  $g_0$  is at distance 0.12 from  $f_0$ .

Both cases are treated, considering four types of distribution for the noise:

- a. The noise  $h_\delta$  is a centered normal density with variance  $\delta^2$ ;
- b. the noise  $h_\delta$  is a centered Laplace density with parameter  $\lambda(\delta)$ ;
- c. the noise  $h_\delta$  is a symmetrized Weibull density with shape parameter 1.5 and variable scale parameter  $\beta(\delta)$ ; and

d. the noise  $h_\delta$  is Cauchy with density  $h_\delta(x) = \gamma(\delta)/\pi (\gamma(\delta)^2 + x^2)$ .

In order to compare the performances of the test under those four distributions, we have adopted the following rule: The parameter of the distribution of the noise is tuned such that, for each value  $\underline{\delta}$ , it holds that  $P(|V_{\underline{\delta}}| > \underline{\delta}) = \Phi(1) - \Phi(-1) \sim 0.65$ , where  $\Phi$  stands for the standard Gaussian cumulative function. Thus, distributions b to d are scaled with respect to the Gaussian noise with variance  $\delta^2$ .

In both cases A and B, the range of  $\delta$  is  $\Delta = (\delta_{\min} = 0.1, \delta_{\max})$ , and we have selected a number of possibilities for  $\delta_{\max}$ , ranging from 0.2 to 0.7.

In case A, we selected  $= \delta_{\max}^2 = 0.5$ , which has a signal-to-noise ratio equal to 0.7, a commonly chosen bound in quality control tests.

In case B, the variance of  $f_0$  is roughly 0.6 and the variance of  $g_0$  is roughly 0.4. The maximal value of  $\delta_{\max}^2$  is roughly 0.5. This is thus a maximal upper bound for a practical modeling.

We present some power functions, making use of the theoretical bounds together with the corresponding ones based on simulation runs. As seen, the performances in the theoretical approach is weak. We have focused on simulation, after some comparison with the theoretical bounds.

### 3.3.1. Case A: The Shift Problem

In this subsection, we evaluate the quality of the theoretical power bound, defined in the previous sections. Thus, we compare the theoretical formula to the empirical lower performances obtained through simulations under the least-favorable hypotheses.

#### Theoretical Power Bound

While supposedly valid for finite  $n$ , the theoretical power bound given by (A8) still assumes some sort of asymptotics, since a good approximation of the bound implies a fine discretization of  $\Delta$  to compute  $I(A_n) = \inf_{\eta \in \Delta_n} I_\eta(A_n)$ . Thus, by condition (1),  $n$  has to be large. Therefore, in the following, we will compute this lower bound for  $n$  sufficiently large (that is, at least 100 observations), which is also consistent with industrial applications.

#### Numerical Power Bound

In order to obtain a minimal bound for the power of the composite test, we compute the power of the test  $\mathbf{H}_0(\delta_*)$  against  $\mathbf{H}_1(\delta_*)$ , where  $\delta_*$  defines the LFH pair  $(F_{\delta_*}, G_{\delta_*})$ .

Following Proposition 4, the LFH for the test defined by  $T_n$  when the noise follows a Gaussian, a Cauchy, or a symmetrized Weibull distribution is achieved for  $(F_{\delta_{\max}}, G_{\delta_{\max}})$ .

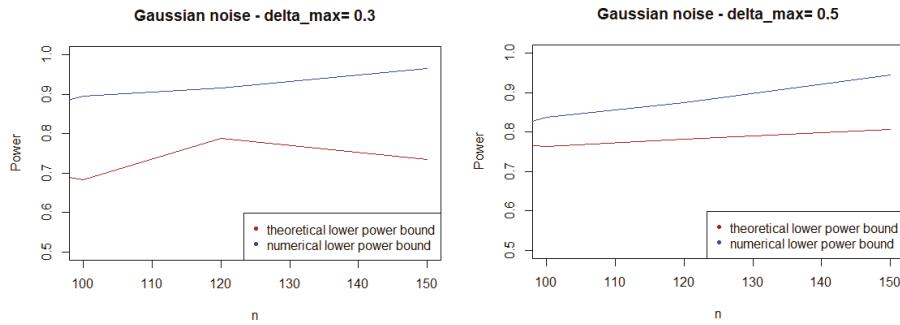
When the noise follows a Laplace distribution, the pair of LFH is the one that satisfies:

$$(F_{\delta_*}, G_{\delta_*}) = \arg \min_{(F_\delta, G_\delta), \delta \in \Delta_n} K(F_\delta, G_\delta). \quad (14)$$

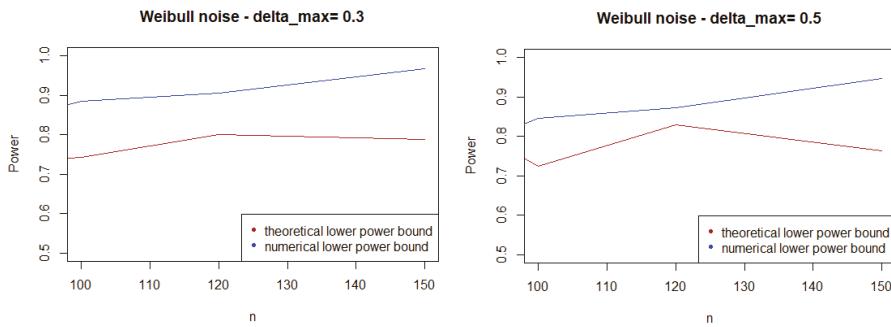
In both of the cases A and B, this condition is also satisfied for  $\delta^* = \delta_{\max}$ .

#### Comparison of the Two Power Curves

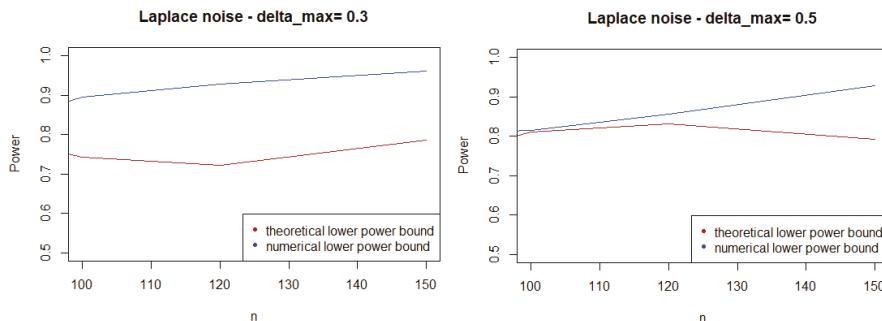
As expected, Figures 1–3 show that the theoretical lower bound is always below the empirical lower bound, when  $n$  is high enough to provide a good approximation of  $I(A_n)$ . This is also true when the noise follows a Cauchy distribution, but for a bigger sample size than in the figures above ( $n > 250$ ).



**Figure 1.** Theoretical and numerical power bound of the test of case A under Gaussian noise (with respect to  $n$ ), for the first kind risk  $\alpha = 0.05$ .



**Figure 2.** Theoretical and numerical power bound of the test of case A under symmetrized Weibull noise (with respect to  $n$ ), for the first kind risk  $\alpha = 0.05$ .



**Figure 3.** Theoretical and numerical power bound of the test of case A under a symmetrized Laplacian noise (with respect to  $n$ ), for the first kind risk  $\alpha = 0.05$ .

In most cases, the theoretical bound tends to largely underestimate the power of the test, when compared to its minimal performance given by simulations under the least-favorable hypotheses. The gap between the two also tends to increase as  $n$  grows. This result may be explained by the

large bound provided by (5), while the numerical performances are obtained with respect to the least-favorable hypotheses.

From a computational perspective, the computational cost of the theoretical bound is far higher than its numeric counterpart.

### 3.3.2. Case B: The Tail Thickness Problem

The calculation of the moment-generating function, appearing in the formula of  $I_\eta(x)$  in (9), is numerically unstable, which renders the computation of the theoretical bound impossible. Thus, in the following sections, the performances of the test will be evaluated numerically, through Monte Carlo replications.

## 4. Some Alternative Statistics for Testing

### 4.1. A Family of Composite Tests Based on Divergence Distances

This section provides a similar treatment as above, dealing now with some extensions of the LRT test to the same composite setting. The class of tests is related to the divergence-based approach to testing, and it includes the cases considered so far. For reasons developed in Section 3.3, we argue through simulation and do not develop the corresponding large deviation approach.

The statistics  $T_n$  can be generalized in a natural way, by defining a family of tests depending on some parameter  $\gamma$ . For  $\gamma \neq 0, 1$ , let

$$\phi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)}$$

be a function defined on  $(0, \infty)$  with values in  $(0, \infty)$ , setting

$$\phi_0(x) := -\log x + x - 1$$

and

$$\phi_1(x) := x \log x - x + 1.$$

For  $\gamma \leq 2$ , this class of functions is instrumental in order to define the so-called power divergences between probability measures, a class of pseudo-distances widely used in statistical inference (see, for example, [13]).

Associated to this class, consider the function

$$\begin{aligned} \varphi_\gamma(x) &:= -\frac{d}{dx}\phi_\gamma(x) \\ &= \frac{1 - x^{\gamma-1}}{\gamma - 1} \text{ for } \gamma \neq 0, 1. \end{aligned}$$

We also consider

$$\begin{aligned} \varphi_1(x) &:= -\log x \\ \varphi_0(x) &:= \frac{1}{x} - 1, \end{aligned}$$

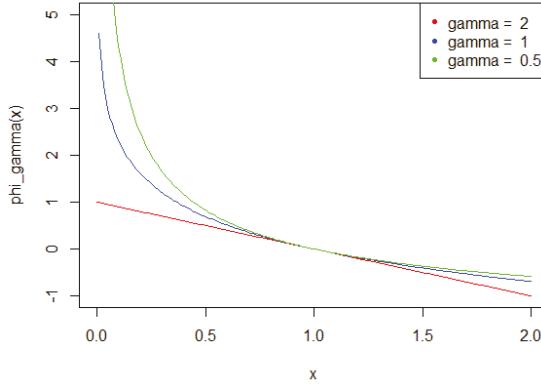
from which the statistics

$$T_{n,\delta}^\gamma := \frac{1}{n} \sum_{i=1}^n \varphi_\gamma(X_i)$$

and

$$T_n^\gamma := \sup_{\delta} T_{n,\delta}^\gamma$$

are well defined, for all  $\gamma \leq 2$ . Figure 4 illustrates the functions  $\varphi_\gamma$ , according to  $\gamma$ .



**Figure 4.**  $\varphi_\gamma$  for  $\gamma = 0.5, 1$ , and  $2$ .

Fix a risk of first kind  $\alpha$ , and the corresponding power of the LRT pertaining to  $H0(\delta_*)$  vs.  $H1(\delta_*)$  by

$$1 - \beta := G_{\delta_*} \left( T_{n,\delta_*}^1 > s_\alpha \right),$$

with

$$s_\alpha := \inf \left\{ s : F_{\delta_*} \left( T_{n,\delta_*}^1 > s \right) \leq \alpha \right\}.$$

Define, accordingly, the power of the test, based on  $T_n^\gamma$  under the same hypotheses, by

$$s_\alpha^\gamma := \inf \left\{ s : F_{\delta_*} \left( T_n^\gamma > s \right) \leq \alpha \right\}$$

and

$$1 - \beta' := G_{\delta_*} \left( T_n^\gamma > s_\alpha^\gamma \right).$$

First,  $\delta_*$  defines the pair of hypotheses  $(F_{\delta_*}, G_{\delta_*})$ , such that the LRT with statistics  $T_{n,\delta_*}^1$  has maximal power among all tests  $H0(\delta_*)$  vs.  $H1(\delta_*)$ . Furthermore, by Theorem A1, it has minimal power on the logarithmic scale among all tests  $H0(\delta)$  vs.  $H1(\delta)$ .

On the other hand,  $(F_{\delta_*}, G_{\delta_*})$  is the LF pair for the test with statistics  $T_n^1$  among all pairs  $(F_\delta, G_\delta)$ .

These two facts allow for the definition of the loss of power, making use of  $T_n^1$  instead of  $T_{n,\delta_*}^1$  for testing  $H0(\delta_*)$  vs.  $H1(\delta_*)$ . This amounts to considering the price of aggregating the local tests  $T_{n,\delta}^1$ , a necessity since the true value of  $\delta$  is unknown. A natural indicator for this loss consists in the difference

$$\Delta_n^1 := G_{\delta_*} \left( T_{n,\delta_*}^1 > s_\alpha \right) - G_{\delta_*} \left( T_n^1 > s_\alpha^1 \right) \geq 0.$$

Consider, now, aggregated test statistics  $T_n^\gamma$ . We do not have at hand a similar result, as in Proposition 2. We, thus, consider the behavior of the test  $H0(\delta_*)$  vs.  $H1(\delta_*)$ , although  $(F_{\delta_*}, G_{\delta_*})$  may not be a LFH for the test statistics  $T_n^\gamma$ . The heuristics, which we propose, makes use of the corresponding loss of power with respect to the LRT, through

$$\Delta_n^\gamma := G_{\delta_*} \left( T_{n,\delta_*}^1 > s_\alpha \right) - G_{\delta_*} \left( T_n^\gamma > s_\alpha^\gamma \right).$$

We will see that it may happen that  $\Delta_n^\gamma$  improves over  $\Delta_n^1$ . We define the optimal value of  $\gamma$ , say  $\gamma^*$ , such that

$$\Delta_n^{\gamma^*} \leq \Delta_n^\gamma,$$

for all  $\gamma$ .

In the various figures hereafter, NP corresponds to the LRT defined between the LFH's  $(F_{\delta_*}, G_{\delta_*})$ , KL to the test with statistics  $T_n^1$  (hence, as presented Section 2), HELL corresponds to  $T_n^{1/2}$ , which is associated to the Hellinger power divergence, and  $G = 2$  corresponds to  $\gamma = 2$ .

#### 4.2. A Practical Choice for Composite Tests Based on Simulation

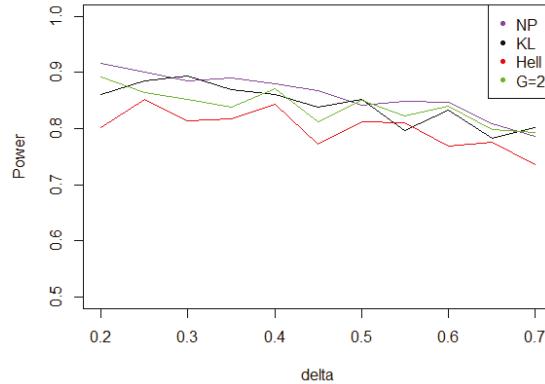
We consider the same cases A and B, as described in Section 3.3.

As stated in the previous section, the performances of the different test statistics are compared, considering the test of  $H_0(\delta_*)$  against  $H_1(\delta_*)$  where  $\delta^*$  is defined, as explained in Section 3.3 as the LFH for the test  $T_n^1$ . In both cases A and B, this corresponds to  $\delta^* = \delta_{\max}$ .

##### 4.2.1. Case A: The Shift Problem

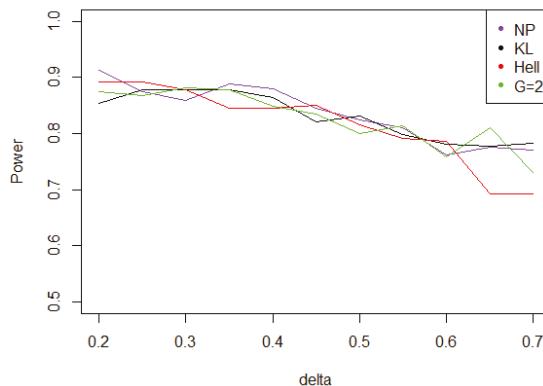
Overall, the aggregated tests perform well, when the problem consists in identifying a shift in a distribution. Indeed, for the three values of  $\gamma$  (0.5, 1, and 2), the power remains above 0.7 for any kind of noise and any value of  $\delta_*$ . Moreover, the power curves associated to  $T_n^\gamma$  mainly overlap with the optimal test  $T_{n,\delta_*}^1$ .

- Under Gaussian noise, the power remains mostly stable over the values of  $\delta_*$ , as shown by Figure 5. The tests with statistics  $T_n^1$  and  $T_n^2$  are equivalently powerful for large values of  $\delta_*$ , while the first one achieves higher power when  $\delta_*$  is small.



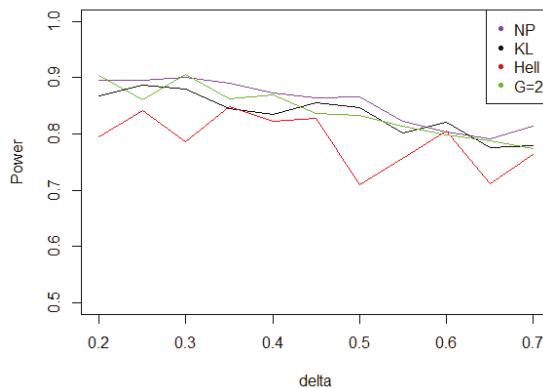
**Figure 5.** Power of the test of case A under Gaussian noise (with respect to  $\delta_{\max}$ ), for the first kind risk  $\alpha = 0.05$  and sample size  $n = 100$ .

- When the noise follows a Laplace distribution, the three power curves overlap the NP power curve, and the different test statistics can be indifferently used. Under such a noise, the alternative hypotheses are extremely well distinguished by the class of tests considered, and this remains true as  $\delta_*$  increases (cf. Figure 6).



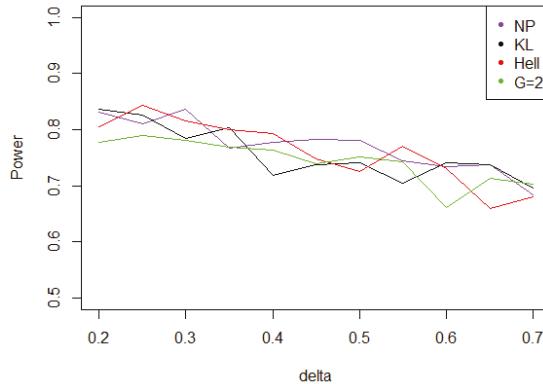
**Figure 6.** Power of the test of case A under Laplacian noise (with respect to  $\delta_{\max}$ ), for the first kind risk  $\alpha = 0.05$  and sample size  $n = 100$ .

- c. Under the Weibull hypothesis,  $T_n^1$  and  $T_n^2$  perform similarly well, and almost always as well as  $T_{n,\delta_*}^1$ , while the power curve associated to  $T_n^{1/2}$  remains below. Figure 7 illustrates that, as  $\delta_{\max}$  increases, the power does not decrease much.



**Figure 7.** Power of the test of case A under symmetrized Weibull noise (with respect to  $\delta_{\max}$ ), for the first kind risk  $\alpha = 0.05$  and sample size  $n = 100$ .

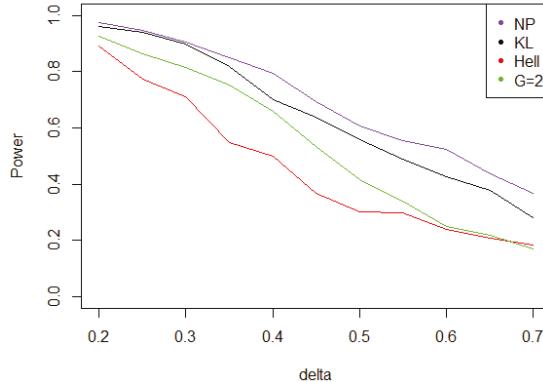
- d. Under a Cauchy assumption, the alternate hypotheses are less distinguishable than under any other parametric hypothesis on the noise, since the maximal power is about 0.84, while it exceeds 0.9 in cases a, b, and c (cf. Figures 5–8). The capacity of the tests to discriminate between  $H_0(\delta_{\max})$  and  $H_1(\delta_{\max})$  is almost independent of the value of  $\delta_{\max}$ , and the power curves are mainly flat.



**Figure 8.** Power of the test of case A under noise following a Cauchy distribution (with respect to  $\delta_{\max}$ ), for the first kind risk  $\alpha = 0.05$  and sample size  $n = 100$ .

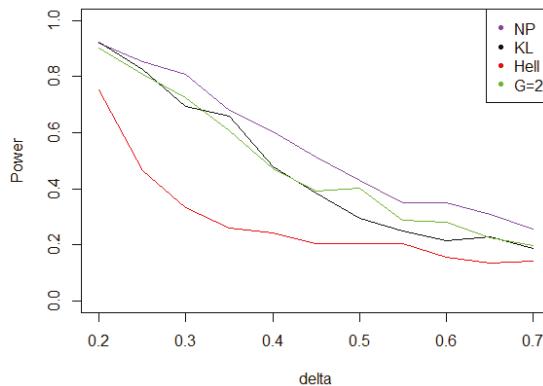
#### 4.2.2. Case B: The Tail Thickness Problem

- a. With the noise defined by case A (Gaussian noise), for  $\text{KL} (\gamma = 1)$ ,  $\delta_* = \delta_{\max}$  due to Proposition 4 and statistics  $T_n^1$  provides the best power uniformly upon  $\delta_{\max}$ . Figure 9 shows a net decrease of the power as  $\delta_{\max}$  increases (recall that the power is evaluated under the least favorable alternative  $G_{\delta_{\max}}$ ).



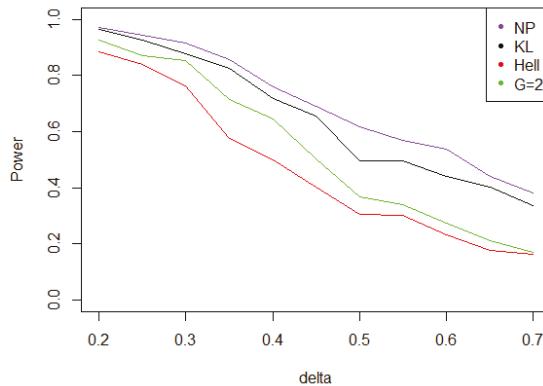
**Figure 9.** Power of the test of case B under Gaussian noise (with respect to  $\delta_{\max}$ ), for the first kind risk  $\alpha = 0.05$  and sample size  $n = 100$ . The NP curve corresponds to the optimal Neyman Pearson test under  $\delta_{\max}$ . The KL, Hellinger, and  $G = 2$  curves stand respectively for  $\gamma = 1$ ,  $\gamma = 0.5$ , and  $\gamma = 2$  cases.

- b. When the noise follows a Laplace distribution, the situation is quite peculiar. For any value of  $\delta$  in  $\Delta$ , the modes  $M_{G_{\delta_{\max}}}$  and  $M_{F_{\delta_{\max}}}$  of the distributions of  $(f_\delta/g_\delta)(X)$  under  $G_{\delta_{\max}}$  and under  $F_{\delta_{\max}}$  are quite separated; both larger than 1. Also, for  $\delta$  all the values of  $|\phi_\gamma(M_{G_{\delta_{\max}}}) - \phi_\gamma(M_{F_{\delta_{\max}}})|$  are quite large for large values of  $\gamma$ . We may infer that the distributions of  $\phi_\gamma((f_\delta/g_\delta)(X))$  under  $G_{\delta_{\max}}$  and under  $F_{\delta_{\max}}$  are quite distinct for all  $\delta$ , which in turn implies that the same fact holds for the distributions of  $T_n^\gamma$  for large  $\gamma$ . Indeed, simulations presented in Figure 10 show that the maximal power of the test tends to be achieved when  $\gamma = 2$ .



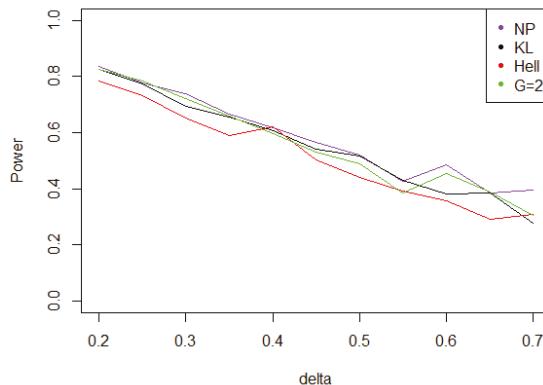
**Figure 10.** Power of the test of case B under Laplacian noise (with respect to  $\delta_{\max}$ ), for the first kind risk  $\alpha = 0.05$  and sample size  $n = 100$ .

- c. When the noise follows a symmetric Weibull distribution, the power function when  $\gamma = 1$  is very close to the power of the LRT between  $F_{\delta_{\max}}$  and  $G_{\delta_{\max}}$  (cf. Figure 11). Indeed, uniformly on  $\delta$ , and on  $x$ , the ratio  $(f_\delta / g_\delta)(x)$  is close to 1. Therefore, the distribution of  $T_n^1$  is close to that of  $T_{n,\delta_{\max}}^1$ , which plays in favor of the KL composite test.



**Figure 11.** Power of the test of case B under symmetrized Weibull noise (with respect to  $\delta_{\max}$ ), for the first kind risk  $\alpha = 0.05$  and sample size  $n = 100$ .

- d. Under a Cauchy distribution, similarly to case A, Figure 12 shows that  $T_n^\gamma$  achieves the maximal power for  $\gamma = 1$  and 2, closely followed by  $\gamma = 0.5$ .



**Figure 12.** Power of the test of case B under a noise following a Cauchy distribution (with respect to  $\delta_{\max}$ ), for the first kind risk  $\alpha = 0.05$  and sample size  $n = 100$ .

## 5. Conclusions

We have considered a composite testing problem, where simple hypotheses in either **H0** and **H1** were paired, due to corruption in the data. The test statistics were defined through aggregation of simple likelihood ratio tests. The critical region for this test and a lower bound of its power was produced. We have shown that this test is minimax, evidencing the least-favorable hypotheses. We have considered the minimal power of the test under such a least favorable hypothesis, both theoretically and by simulation, and for a number of cases (including corruption by Gaussian, Laplacian, symmetrized Weibull, and Cauchy noise). Whatever this noise, the actual minimal power, as measured through simulation, was quite higher than obtained through analytic developments. Least-favorable hypotheses were defined in an asymptotic sense, and were proved to be the pair of simple hypotheses in **H0** and **H1** which are closest, in terms of the Kullback-Leibler divergence; this holds as a consequence of the Chernoff-Stein Lemma. We, next, considered aggregation of tests where the likelihood ratio was substituted by a divergence-based statistics. This choice extended the former one, and may produce aggregate tests with higher power than obtained through aggregation of the LRTs, as exemplified and analysed. Open questions are related to possible extensions of the Chernoff-Stein Lemma for divergence-based statistics.

**Author Contributions:** Conceptualization, M.B. and J.J.; Methodology, M.B., J.J., A.K.M. and E.M.; Software, E.M.; Validation, M.B., A.K.M. and E.M.; Formal Analysis, M.B. and J.J.; Writing Original Draft Preparation, M.B., J.J., E.M. and A.K.M.; Supervision, M.B.

**Funding:** The research of Jana Jurečková was supported by the Grant 18-01137S of the Czech Science Foundation. Michel Broniatowski and M. Ashok Kumar would like to thank the Science and Engineering Research Board of the government of India for the financial support for their collaboration through the VAJRA scheme.

**Acknowledgments:** The authors are thankful to Jan Kalina for discussion; they also thank two anonymous referees for comments which helped to improve on a former version of this paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Proof of Proposition 2

### Appendix A.1. The Critical Region of the Test

Define

$$Z_{\delta'} := \log \frac{g_{\delta'}}{f_{\delta'}}(X),$$

which satisfies

$$\begin{aligned} E_{F_\delta}(Z_{\delta'}) &= \int \log \frac{g_{\delta'}}{f_{\delta'}}(x) f_\delta(x) dx \\ &= \int \log \frac{g_{\delta'}}{f_\delta}(x) f_\delta(x) dx + \int \log \frac{f_\delta}{f_{\delta'}}(x) f_\delta(x) dx \\ &= K(F_\delta, G_{\delta'}) - K(F_\delta, G_{\delta'}). \end{aligned}$$

Note that, for all  $\delta$ ,

$$K(F_\delta, G_{\delta'}) - K(F_\delta, G_{\delta'}) = \int \log \frac{g_{\delta'}}{f_{\delta'}} f_\delta$$

is negative for  $\delta'$  close to  $\delta$ , assuming that

$$\delta' \mapsto \int \log \frac{g_{\delta'}}{f_{\delta'}} f_\delta$$

is a continuous mapping. Assume, therefore, that (6) holds, which means that the classes of distributions  $(G_\delta)$  and  $(F_\delta)$  are somehow well separated. This implies that  $E_{F_\delta}(Z_{\delta'}) < 0$ , for all  $\delta$  and  $\delta'$ .

In order to obtain an upper bound for  $F_\delta(T_{n,\delta'}(\mathbf{X}_n) > A_n)$ , for all  $\delta, \delta'$  in  $\Delta$ , through the Chernoff Inequality, consider

$$\varphi_{\delta,\delta'}(t) := \log E_{F_\delta}(\exp(tZ_{\delta'})) = \log \int \left( \frac{g_{\delta'}(x)}{f_{\delta'}(x)} \right)^t g_\delta(x) dx.$$

Let

$$t^+(\mathcal{N}_{\delta,\delta'}) := \sup \{t \in \mathcal{N}_{\delta,\delta'} : \varphi_{\delta,\delta'}(t) < \infty\}.$$

The function  $(\delta, \delta', x) \mapsto J_{\delta,\delta'}(x)$  is continuous on its domain, and since  $t \mapsto \varphi_{\delta,\delta'}(t)$  is a strictly convex function which tends to infinity as  $t$  tends to  $t^+(\mathcal{N}_{\delta,\delta'})$ , it holds that

$$\lim_{x \rightarrow \infty} J_{\delta,\delta'}(x) = +\infty$$

for all  $\delta, \delta'$  in  $\Delta_n$ .

We now consider an upper bound for the risk of first kind on a logarithmic scale.

We consider

$$A_n > E_{F_\delta}(Z_{\delta'}), \quad (\text{A1})$$

for all  $\delta, \delta'$ . Then, by the Chernoff inequality

$$\frac{1}{n} \log F_\delta(T_{n,\delta'}(\mathbf{X}_n) > A_n) \leq -J_{\delta,\delta'}(A_n).$$

Since  $A_n$  should satisfy

$$\exp(-nJ_{\delta,\delta'}(A_n)) \leq \alpha_n,$$

with  $\alpha_n$  bounded away from 1,  $A_n$  surely satisfies (A1) for large  $n$ .

The mapping  $m_{\delta,\delta'}(t) := (d/dt) \varphi_{\delta,\delta'}(t)$  is a homeomorphism from  $\mathcal{N}_{\delta,\delta'}$  onto the closure of the convex hull of the support of the distribution of  $Z_{\delta'}$  under  $F_\delta$  (see, e.g., [14]). Denote

$$\text{ess sup}_\delta Z_{\delta'} := \sup \{x : \text{for all } \epsilon > 0, F_\delta(Z_{\delta'} \in (x - \epsilon, x) > 0)\}.$$

We assume that

$$\text{ess sup}_\delta Z_{\delta'} = +\infty, \quad (\text{A2})$$

which is convenient for our task, and quite common in practical industrial modelling. This assumption may be weakened, at notational cost mostly. It follows that

$$\lim_{t \rightarrow t^+ (\mathcal{N}_{\delta, \delta'})} m_{\delta, \delta'}(t) = +\infty.$$

It holds that

$$J_{\delta, \delta'}(E_{F_\delta}(Z_{\delta'})) = 0,$$

and, as seen previously

$$\lim_{x \rightarrow \infty} J_{\delta, \delta'}(x) = +\infty.$$

On the other hand,

$$m_{\delta, \delta'}(0) = E_{F_\delta}(Z_{\delta'}) = K(F_\delta, F_{\delta'}) - K(F_\delta, G_{\delta'}) < 0.$$

Let

$$\begin{aligned} \mathcal{I} &:= \left( \sup_{\delta, \delta'} E_{F_\delta}(Z_{\delta'}), \infty \right) \\ &= \left( \sup_{\delta, \delta'} K(F_\delta, F_{\delta'}) - K(F_\delta, G_{\delta'}), \infty \right). \end{aligned}$$

By (A2), the interval  $\mathcal{I}$  is not void.

We now define  $A_n$  such that (4) holds, namely

$$P_{H0}(\mathbf{H1}) \leq p_n \max_{\delta} \max_{\delta'} F_\delta(T_{n, \delta'} > A_n) \leq \alpha_n$$

holds for any  $\alpha_n$  in  $(0, 1)$ . Note that

$$A_n \geq \max_{\delta, \delta'} E_{F_\delta}(Z_{\delta'}) = \max_{(\delta, \delta') \in \Delta \times \Delta} K(F_\delta, F_{\delta'}) - K(F_\delta, G_{\delta'}), \quad (\text{A3})$$

for all  $n$  large enough, since  $\alpha_n$  is bounded away from 1.

The function

$$J(x) := \min_{(\delta, \delta') \in \Delta \times \Delta} J_{\delta, \delta'}(x)$$

is continuous and increasing, as it is the infimum of a finite collection of continuous increasing functions defined on  $\mathcal{I}$ .

Since

$$P_{H0}(\mathbf{H1}) \leq p_n \exp(-nJ(A_n)),$$

given  $\alpha_n$ , define

$$A_n := J^{-1}\left(-\frac{1}{n} \log \frac{\alpha_n}{p_n}\right). \quad (\text{A4})$$

This is well defined for  $\alpha_n \in (0, 1)$ , as  $\sup_{(\delta, \delta') \in \Delta \times \Delta} E_{F_\delta}(Z_{\delta'}) < 0$  and  $-(1/n) \log(\alpha_n/p_n) > 0$ .

### Appendix A.2. The Power Function

We now evaluate a lower bound for the power of this test, making use of the Chernoff inequality to get an upper bound for the second risk.

Starting from (5),

$$P_{H1}(\mathbf{H0}) \leq \sup_{\eta \in \Delta} G_\eta(T_{n, \eta} \leq A_n),$$

and define

$$W_\eta := -\log \frac{g_\eta}{f_\eta}(x).$$

It holds that

$$E_{G_\eta}(W_\eta) = \int \log \frac{f_\eta(x)}{g_\eta(x)} g_\eta(x) dx = -K(G_\eta, F_\eta),$$

and

$$m_\eta(t) := (d/dt) \log E_{G_\eta}(\exp t W_\eta),$$

which is an increasing homeomorphism from  $\mathcal{M}_\eta$  onto the closure of the convex hull of the support of  $W_\eta$  under  $G_\eta$ . For any  $\eta$ , the mapping

$$x \mapsto I_\eta(x)$$

is a strictly increasing function of  $\mathcal{K}_\eta := (E_{G_\eta}(W_\eta), \infty)$  onto  $(0, +\infty)$ , where the same notation as above holds for  $\text{ess sup}_\eta W_\eta$  (here under  $G_\eta$ ), and where we assumed

$$\text{ess sup}_\eta W_\eta = \infty \quad (\text{A5})$$

for all  $\eta$ .

Assume that  $A_n$  satisfies

$$A_n \in \mathcal{K} := \bigcap_{\eta \in \Delta} \mathcal{K}_\eta \quad (\text{A6})$$

namely

$$A_n \geq \sup_{\eta \in \Delta} E_{G_\eta}(W_\eta) = -\inf_{\eta \in \Delta} K(G_\eta, F_\eta). \quad (\text{A7})$$

Making use of the Chernoff inequality, we get

$$P_{H_1}(\mathbf{H0}) \leq \exp \left( -n \inf_{\eta \in \Delta} I_\eta(A_n) \right).$$

Each function  $x \mapsto I_\eta(x)$  is increasing on  $(E_{G_\eta}(W_\eta), \infty)$ . Therefore the function

$$x \mapsto I(x) := \inf_{\eta \in \Delta} I_\eta(x)$$

is continuous and increasing, as it is the infimum of a finite number of continuous increasing functions on the same interval  $\mathcal{K}$ , which is not void due to (A5).

We have proven that, whenever (A7) holds, a lower bound for the test of **H0** vs. **H1** is given by

$$\begin{aligned} P_{H_1}(\mathbf{H1}) &\geq 1 - \exp(-nI(A_n)) \\ &= 1 - \exp \left( -nI \left( J^{-1} \left( -\frac{1}{n} \log \frac{\alpha_n}{p_n} \right) \right) \right). \end{aligned} \quad (\text{A8})$$

We now collect the above discussion, in order to complete the proof.

### Appendix A.3. A Synthetic Result

The function  $J$  is one-to-one from  $I$  onto  $K := (J \left( \sup_{(\delta, \delta') \in \Delta \times \Delta} E_\delta(Z_{\delta'}) \right), \infty)$ . Since  $F_\delta, J_{\delta, \delta'}(E_\delta(Z_{\delta'})) = 0$ , it follows that  $J \left( \sup_{(\delta, \delta') \in \Delta \times \Delta} E_\delta(Z_{\delta'}) \right) \geq 0$ . Since  $E_{F_\delta}(Z_{\delta'}) = K(F_\delta, F_{\delta'}) - K(F_\delta, G_{\delta'}) < 0$ , whenever  $\alpha_n$  in  $(0, 1)$  there exists a unique  $A_n \in (-\inf_{(\delta, \delta') \in \Delta \times \Delta} (K(F_\delta, G_{\delta'}) - K(F_\delta, F_{\delta'})), \infty)$  which defines the critical region with level  $\alpha_n$ .

For the lower bound on the power of the test, we have assumed  $A_n \in \mathcal{K} = \left( \sup_{\eta \in \Delta} E_\eta (W_\eta), \infty \right) = \left( -\inf_{\eta \in \Delta} K(G_\eta, F_\eta), \infty \right)$ .

In order to collect our results in a unified setting, it is useful to state some connection between  $\inf_{(\delta, \delta') \in \Delta \times \Delta} [K(F_\delta, G_{\delta'}) - K(F_\delta, F_{\delta'})]$  and  $\inf_{\eta \in \Delta} K(G_\eta, F_\eta)$ . See (A3) and (A7).

Since  $K(G_\delta, F_\delta)$  is positive, it follows from (6) that

$$\sup_{(\delta, \delta') \in \Delta \times \Delta} \int \log \frac{f_{\delta'}}{g_{\delta'}} f_\delta < \sup_{\delta \in \Delta} K(G_\delta, F_\delta), \quad (\text{A9})$$

which implies the following fact:

Let  $\alpha_n$  be bounded away from 1. Then (A3) is fulfilled for large  $n$ , and therefore there exists  $A_n$  such that

$$\sup_{\delta \in \Delta} F_\delta (T_n > A_n) \leq \alpha_n.$$

Furthermore, by (A9), Condition (A7) holds, which yields the lower bound for the power of this test, as stated in (A8).

## Appendix B. Proof of Theorem 3

We will repeatedly make use of the following result (Theorem 3 in [15]), which is an extension of the Chernoff-Stein Lemma (see [16]).

**Theorem A1.** [Krafft and Plachky] Let  $x_n$ , such that

$$F_\delta (T_{n,\delta} > x_n) \leq \alpha_n$$

with  $\limsup_{n \rightarrow \infty} \alpha_n < 1$ . Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log G_\delta (T_{n,\delta} \leq x_n) = -K (F_\delta, G_\delta).$$

**Remark A2.** The above result indicates that the power of the Neyman-Pearson test only depends on its level on the second order on the logarithmic scale.

Define  $A_{n,\delta_*}$  such that

$$F_{\delta_*} (T_n \leq A_n) = F_{\delta_*} (T_{n,\delta_*} \leq A_{n,\delta_*}).$$

This exists and is uniquely defined, due to the regularity of the distribution of  $T_{n,\delta_*}$  under  $F_{\delta_*}$ . Since  $\mathbf{1}[T_{n,\delta_*} > A_n]$  is the likelihood ratio test of  $\mathbf{H}_0(\delta_*)$  against  $\mathbf{H}_1(\delta_*)$  of the size  $\alpha_n$ , it follows, by unbiasedness of the LRT, that

$$F_{\delta_*} (T_n \leq A_n) = F_{\delta_*} (T_{n,\delta_*} \leq A_{n,\delta_*}) \geq G_{\delta_*} (T_{n,\delta_*} \leq A_{n,\delta_*}).$$

We shall later verify the validity of the conditions of Theorem A1; namely, that

$$\limsup_{n \rightarrow \infty} F_{\delta_*} (T_{n,\delta_*} \leq A_{n,\delta_*}) < 1. \quad (\text{A10})$$

Assuming (A10) we get, by Theorem A1,

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log F_{\delta_*} (T_n \leq A_n) \geq \lim_{n \rightarrow \infty} \frac{1}{n} \log G_{\delta_*} (T_{n,\delta_*} \leq A_{n,\delta_*}) = -K (F_{\delta_*}, G_{\delta_*}).$$

We shall now prove that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log G_{\delta_*}(T_{n,\delta_*} \leq A_{n,\delta_*}) = \lim_{n \rightarrow \infty} \frac{1}{n} \log G_{\delta_*}(T_n \leq A_n).$$

Let  $B_{n,\delta_*}$ , such that

$$G_{\delta_*}(T_{n,\delta_*} \leq B_{n,\delta_*}) = G_{\delta_*}(T_n \leq A_n).$$

By regularity of the distribution of  $T_{n,\delta_*}$  under  $G_{\delta_*}$ , such a  $B_{n,\delta_*}$  is defined in a unique way. We will prove that the condition in Theorem A1 holds, namely

$$\limsup_{n \rightarrow \infty} F_{\delta_*}(T_{n,\delta_*} \leq B_{n,\delta_*}) < 1. \quad (\text{A11})$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log G_{\delta_*}(T_{n,\delta_*} \leq A_{n,\delta_*}) = \lim_{n \rightarrow \infty} \frac{1}{n} \log G_{\delta_*}(T_n \leq A_n) = -K(F_{\delta_*}, G_{\delta_*}).$$

Incidentally, we have obtained that  $\lim_{n \rightarrow \infty} \frac{1}{n} \log G_{\delta_*}(T_n \leq A_n)$  exists. Therefore we have proven that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log F_{\delta_*}(T_n \leq A_n) \geq \lim_{n \rightarrow \infty} \frac{1}{n} \log G_{\delta_*}(T_n \leq A_n),$$

which is a form of unbiasedness. For  $\delta \neq \delta_*$ , let  $B_{n,\delta}$  be defined by

$$G_\delta(T_{n,\delta} \leq B_{n,\delta}) = G_\delta(T_n \leq A_n).$$

As above,  $B_{n,\delta}$  is well-defined. Assuming

$$\limsup_{n \rightarrow \infty} F_\delta(T_{n,\delta} \leq B_{n,\delta}) < 1, \quad (\text{A12})$$

it follows, from Theorem A1, that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log G_\delta(T_n \leq A_n) = \lim_{n \rightarrow \infty} \frac{1}{n} \log G_\delta(T_{n,\delta} \leq B_{n,\delta}) = -K(F_\delta, G_\delta).$$

Since  $K(F_{\delta_*}, G_{\delta_*}) \leq K(F_\delta, G_\delta)$ , we have proven

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log F_{\delta_*}(T_n \leq A_n) \geq \lim_{n \rightarrow \infty} \frac{1}{n} \log G_{\delta_*}(T_n \leq A_n) \geq \lim_{n \rightarrow \infty} \frac{1}{n} \log G_\delta(T_n \leq A_n).$$

It remains to verify the conditions (A10)–(A12). We will only verify (A12), as the two other conditions differ only by notation. We have

$$\begin{aligned} G_\delta(T_{n,\delta} > B_{n,\delta}) &= G_\delta(T_n > A_n) \leq F_\delta(T_n > A_n) + d_{TV}(F_\delta, G_\delta) \\ &\leq \alpha_n + d_{TV}(F_\delta, G_\delta) < 1, \end{aligned}$$

by hypothesis (13). By the law of large numbers, under  $G_\delta$

$$\lim_{n \rightarrow \infty} T_{n,\delta} = K(G_\delta, F_\delta) [G_\delta - \text{a.s.}].$$

Therefore, for large  $n$ ,

$$\liminf_{n \rightarrow \infty} B_{n,\delta} \geq K(G_\delta, F_\delta) [G_\delta - \text{a.s.}].$$

Since, under  $F_\delta$ ,

$$\lim_{n \rightarrow \infty} T_{n,\delta} = -K(F_\delta, G_\delta) [F_\delta - \text{a.s.}],$$

this implies that

$$\lim_{n \rightarrow \infty} F_\delta(T_{n,\delta} > B_{n,\delta}) < 1.$$

## Appendix C. Proof of Proposition 4

We now prove the three lemmas that we used.

**Lemma A3.** Let  $P$ ,  $Q$ , and  $R$  denote three distributions with respective continuous and bounded densities  $p$ ,  $q$ , and  $r$ . Then

$$K(P * R, Q * R) \leq K(P, Q). \quad (\text{A13})$$

**Proof.** Let  $\mathcal{P} := (A_1, \dots, A_K)$  be a partition of  $\mathbb{R}$  and  $p := (p_1, \dots, p_K)$  denote the probabilities of  $A_1, \dots, A_K$  under  $P$ . Set the same definition for  $q_1, \dots, q_K$  and for  $r_1, \dots, r_K$ . Recall that the log-sum inequality writes

$$\left( \sum a_i \right) \log \frac{\sum b_i}{\sum c_i} \leq \sum a_i \log \frac{b_i}{c_i}$$

for positive vectors  $(a_i)_i$ ,  $(b_i)_i$  and  $(c_i)_i$ . By the above inequality, for any  $i \in \{1, \dots, K\}$ , denoting  $(p * r)$  to be the convolution of  $p$  and  $r$ ,

$$(p * r)_j \log \frac{(p * r)_j}{(q * r)_j} \leq \sum_{i=1}^K p_j r_{i-j} \log \frac{p_j r_{i-j}}{q_j r_{i-j}}.$$

Summing over  $j \in \{1, \dots, K\}$  yields

$$\sum_{j=1}^K (p * r)_j \log \frac{(p * r)_j}{(q * r)_j} \leq \sum_{j=1}^K p_j \log \frac{p_j}{q_j},$$

which is equivalent to

$$K_{\mathcal{P}}(P * R, Q * R) \leq K_{\mathcal{P}}(P, Q),$$

where  $K_{\mathcal{P}}$  designates the Kullback-Leibler divergence defined on  $\mathcal{P}$ . Refine the partition and go to the limit (Riemann Integrals), to obtain (A13)  $\square$

We now set a classical general result which states that, when  $R_\delta$  denotes a family of distributions with some decomposability property, then the Kullback-Leibler divergence between  $P * R_\delta$  and  $Q * R_\delta$  is a decreasing function of  $\delta$ .

**Lemma A4.** Let  $P$  and  $Q$  satisfy the hypotheses of Lemma A3 and let  $(R_\delta)_{\delta > 0}$  denote a family of p.m.'s on  $\mathbb{R}$ , and denote accordingly  $V_\delta$  to be a r.v. with distribution  $R_\delta$ . Assume that, for all  $\delta$  and  $\eta$ , there exists a r.v.  $W_{\delta, \eta}$ , independent upon  $V_\delta$ , such that

$$V_{\delta+\eta} =_d V_\delta + W_{\delta, \eta}.$$

Then the function  $\delta \mapsto K(P * R_\delta, Q * R_\delta)$  is non-increasing.

**Proof.** Using Lemma A3, it holds that, for positive  $\eta$ ,

$$\begin{aligned} K(P * R_{\delta+\eta}, Q * R_{\delta+\eta}) &= K((P * R_\delta) * W_{\delta, \eta}, (Q * R_\delta) * W_{\delta, \eta}) \\ &\leq K(P * R_\delta, Q * R_\delta), \end{aligned}$$

which proves the claim.  $\square$

**Lemma A5.** Let  $P$ ,  $Q$ , and  $R$  be three probability distributions with respective continuous and bounded densities  $p$ ,  $q$ , and  $r$ . Assume that

$$K(P, Q) \leq K(Q, P),$$

where all involved quantities are assumed to be finite. Then

$$K(P * R, Q * R) \leq K(Q * R, P * R).$$

**Proof.** We proceed as in Lemma A3, using partitions and denoting by  $p_1, \dots, p_K$  the induced probability of  $P$  on  $\mathcal{P}$ . Then,

$$\begin{aligned} K_{\mathcal{P}}(P * R, Q * R) - K_{\mathcal{P}}(Q * R, P * R) &= \sum_i \sum_j (p_j r_{i-j} + q_j r_{i-j}) \log \frac{\sum_j p_j r_{i-j}}{\sum_j q_j r_{i-j}} \\ &\leq \sum_j \sum_i (p_j r_{i-j} + q_j r_{i-j}) \log \frac{p_j}{q_j} \\ &= \sum_j (p_j + q_j) \log \frac{p_j}{q_j} \\ &= K_{\mathcal{P}}(P, Q) - K_{\mathcal{P}}(Q, P) \leq 0, \end{aligned}$$

where we used the log-sum inequality and the fact that  $K(P, Q) \leq K(Q, P)$  implies  $K_{\mathcal{P}}(P, Q) \leq K_{\mathcal{P}}(Q, P)$ , by the data-processing inequality.  $\square$

## References

1. Broniatowski, M.; Jurečková, J.; Kalina, J. Likelihood ratio testing under measurement errors. *Entropy* **2018**, *20*, 966. [[CrossRef](#)]
2. Guo, D. Relative entropy and score function: New information-estimation relationships through arbitrary additive perturbation. In Proceedings of the IEEE International Symposium on Information Theory (ISIT 2009), Seoul, Korea, 28 June–3 July 2009; pp. 814–818.
3. Huber, P.; Strassen, V. Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Stat.* **1973**, *2*, 251–273. [[CrossRef](#)]
4. Eguchi, S.; Copas, J. Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. *J. Multivar. Anal.* **2006**, *97*, 2034–2040. [[CrossRef](#)]
5. Narayanan, K.R.; Srinivasa, A.R. On the thermodynamic temperature of a general distribution. *arXiv* **2007**, arXiv:0711.1460.
6. Bahadur, R.R. Stochastic comparison of tests. *Ann. Math. Stat.* **1960**, *31*, 276–295. [[CrossRef](#)]
7. Bahadur, R.R. *Some Limit Theorems in Statistics*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1971.
8. Birgé, L. Vitesses maximales de décroissance des erreurs et tests optimaux associés. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **1981**, *55*, 261–273. [[CrossRef](#)]
9. Tusnády, G. On asymptotically optimal tests. *Ann. Stat.* **1987**, *5*, 385–393. [[CrossRef](#)]
10. Liese, F.; Vajda, I. *Convex Statistical Distances*; Teubner: Leipzig, Germany, 1987.
11. Tsallis, C. Possible generalization of BG statistics. *J. Stat. Phys.* **1987**, *52*, 479–485. [[CrossRef](#)]
12. Goldie, C. A class of infinitely divisible random variables. *Proc. Camb. Philos. Soc.* **1967**, *63*, 1141–1143. [[CrossRef](#)]
13. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; CRC Press: Boca Raton, FL, USA, 2011.
14. Barndorff-Nielsen, O. *Information and Exponential Families in Statistical Theory*; John Wiley & Sons: New York, NY, USA, 1978.
15. Kraft, O.; Plachky, D. Bounds for the power of likelihood ratio tests and their asymptotic properties. *Ann. Math. Stat.* **1970**, *41*, 1646–1654. [[CrossRef](#)]
16. Chernoff, H. Large-sample theory: Parametric case. *Ann. Math. Stat.* **1956**, *27*, 1–22. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Convex Optimization via Symmetrical Hölder Divergence for a WLAN Indoor Positioning System

Osamah Abdullah

Department of Electrical Power Engineering Techniques, Al-Ma'moun University College, Baghdad 00964, Iraq; osamah.abdullah@wmich.edu

Received: 3 July 2018; Accepted: 14 August 2018; Published: 25 August 2018



**Abstract:** Modern indoor positioning system services are important technologies that play vital roles in modern life, providing many services such as recruiting emergency healthcare providers and for security purposes. Several large companies, such as Microsoft, Apple, Nokia, and Google, have researched location-based services. Wireless indoor localization is key for pervasive computing applications and network optimization. Different approaches have been developed for this technique using WiFi signals. WiFi fingerprinting-based indoor localization has been widely used due to its simplicity, and algorithms that fingerprint WiFi signals at separate locations can achieve accuracy within a few meters. However, a major drawback of WiFi fingerprinting is the variance in received signal strength (RSS), as it fluctuates with time and changing environment. As the signal changes, so does the fingerprint database, which can change the distribution of the RSS (multimodal distribution). Thus, in this paper, we propose that symmetrical Hölder divergence, which is a statistical model of entropy that encapsulates both the skew Bhattacharyya divergence and Cauchy–Schwarz divergence that are closed-form formulas that can be used to measure the statistical dissimilarities between the same exponential family for the signals that have multivariate distributions. The Hölder divergence is asymmetric, so we used both left-sided and right-sided data so the centroid can be symmetrized to obtain the minimizer of the proposed algorithm. The experimental results showed that the symmetrized Hölder divergence consistently outperformed the traditional k nearest neighbor and probability neural network. In addition, with the proposed algorithm, the position error accuracy was about 1 m in buildings.

**Keywords:** information geometry; centroid; Bregman information; Hölder divergence; indoor localization

---

## 1. Introduction

The global positioning system (GPS) is the world's most utilized location system, but it cannot be used to accurately identify indoor locations due to the lack of line-of-sight between GPS receivers and satellites. Smartphones can provide location-based services in pervasive computing; they bring the power of GPS inside buildings. A previous study [1] showed that the global indoor positioning market is expected to grow from \$935.05 million in 2014 to approximately \$4.42 billion in 2019, corresponding to compound annual growth rate of 36.5%. Many technologies have been used instead of GPS, such as radiofrequency identification, Bluetooth, magnetic field variations, ultrasound, light-emitting diode light bulbs, ZigBee, and WiFi signals, to create high-accuracy indoor localization-based systems. These technologies are considered from a cost perspective.

With the widespread use of smart phones in the past decade, there has been an increasing demand to use indoor positioning systems (IPSs) to determine the position of objects and people inside buildings. In general, there are trade-offs between cost and an IPS technology. For example, ultrasonic technology has high accuracy but is also costly due to the large installation required. Since deployment of the

WiFi infrastructure, it has been widely used to estimate the position of an object. The received signal strength (RSS) is a metric value that can be obtained from existing WiFi access points (APs) by any device equipped with a WiFi network adapter. The WiFi infrastructure does not require installation costs or specific hardware [2,3]. Nevertheless, IPSs face many challenges in indoor environments due to the unique properties and transient phenomena such as multipath propagation and signal attenuation. Signal attenuation is caused by people, furniture, and walls, which can limit the ability to design an accurate positioning system [4,5].

IPPs can be classified into two main categories: fingerprint-based techniques and log-distance propagation model algorithms, the latter of can be divided into angulation and lateration methods. Lateration methods calculate the absolute or relative position of an object by measuring distances from multiple reference points using geometry information such as angle of arrival, time of arrival, and time difference of arrival from the signals of APs. However, lateration-based techniques suffer from inaccurate location estimation; for example, it was reported in Reference [6] that the average localization distance error is 24.73 ft with a width of 80 ft and a length of 200 ft in a typical office scenario. Such inaccurate estimations occur for two reasons: non-line-of-sight propagation and inaccurate calculation of one or more of the APs' axes. Thus, fingerprinting-based localization has become the more dominant technique in IPSs and has two major phases. First, the offline phase, in which the RSS value is recorded with their coordinates at predetermined reference points (RPs) to generate a radio map database [7–9].

The k nearest neighbor (kNN) is one simple way to estimate the location of an object by using the Euclidean distance to estimate the dissimilarity between the offline and online phases. The kNN algorithm has low accuracy and is easy to implement compared to other algorithms, such as Bayesian modeling and statistical learning, which have been used to estimate the location of an object. The localization distance error is one of the most fundamental metrics that determine the accuracy and reliability of the system. Variation in WiFi signals is an important issue [10,11]. There are several factors that affect WiFi signal propagation such as human bodies, radiofrequency (RF) equipment, and physical obstructions. These factors cause multiple issues, such as multipath wave propagation and signal attenuation, which can decrease the accuracy of the localization system [12].

The values stored in data maps represent the mean value of the RSS. Some approaches presume that the RSS distribution is Gaussian [13], whereas others presume non-Gaussian distributions [14]. Nevertheless, WiFi-based indoor localization systems have many advantages such as low cost and availability. Different hardware can significantly affect the accuracy of IPSs; for instance, it was reported in Reference [12] that RSS values collected using different smartphones at the same time and same location had different values. Furthermore, the orientation of the body can also contribute to the variance of the RSS signal; thus, the human body can be a significant signal attenuator.

In this paper, we use the Hölder divergence, which generalizes the idea of divergence in information geometry by smooth the non-metric of statistical distances in a way that are not required to follow the law of indiscernibles. The inequality of log-ratio gap pseudo-divergence is built to measure the statistical distance of two classes based on Hölder's ordinary divergence. By experiment, the WiFi signal suffers from multimodal distribution; nevertheless, the Hölder divergence is considered the proper divergence to measure the dissimilarities between probability densities since the Hölder divergence is a projective divergence that does not need the distribution be normalized and allows the closed form expressions when the expansion family is an affine natural space like multinomial distributions.

Hölder divergences encompass both the skew Bhattacharyya divergences and Cauchy–Schwarz divergence and can be symmetrized, and the symmetrized Hölder divergence outperformed the symmetrized Cauchy–Schwarz divergence over the dataset of Gaussians. Both Cauchy–Schwarz divergences are part of a projective divergence distance family with a closed-form expression that does not need to be normalized when considering closed-form expressions with an affine and conic parameter space, such as multivariate or multinomial distributions.

The fingerprinting-based localization has two phases, the off-line phase and the on-line phase. In the off-line phase, we propose a procedure with a high characterization distribution. The RSS values were taken from four different orientations ( $45^\circ$ ,  $135^\circ$ ,  $225^\circ$ , and  $315^\circ$ ) to prevent body-blocking effects, with a scan performed for 100 s in each direction to reduce the effects of signal variation.

The fingerprinting radio-maps were decomposed into many clusters using k-means-Bregman. The symmetrized k-means-Bregman showed unique results; the left-side centroid is the same Jensen–Shannon information radius as the right-side centroid that generalized the mean value of the cluster. Nevertheless, the right-side centroid was independent and always coincided with the center of the mass of the cluster point set. The symmetrized k-means-Bregman can be geometrically interpreted as a unique intersection of the linking between the two-sided centroid and the mixed-type bisector, and that generalized the two-sided centroid for a symmetrized k-means-Bregman.

## 2. Related Work

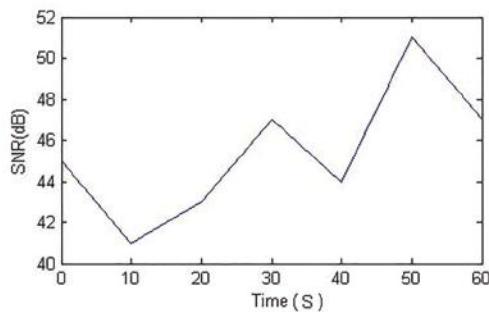
Most research on WiFi fingerprinting localization algorithms has focused on improvements in collecting fingerprinting data, which can decrease localization distance error and improve accuracy. Different algorithms have been proposed, some of which use the propagation properties of the signal, others that use ray tracing [15], and still others that use crowdsourcing-based inertial sensor data and indoor WiFi signal propagation models. Fingerprint-based location methods suffer from time variation between the offline and online phases. kNN is considered a pioneer algorithm that is used in localization-based algorithms. It uses the Euclidean distance to measure the similarity and dissimilarity between runtime and training data, after which the distance is sorted in increasing order. Some researchers use clustering techniques to reduce the impact of time variation by clustering the fingerprinting radio map into multi-partitions, after that the cluster that has lowest RSS-based distance will be chosen [15].

The cluster filtered kNN method was proposed in Reference [16] to partition the fingerprint radio map using hierarchical clustering; the proposed algorithm showed some improvement in the results. To improve the accuracy of the positioning system, Altintas and Serif [17] replaced the k-means algorithm with hierarchical clustering, which led to some improvement in the localization distance error. Likewise, it was proposed to incorporate kNN information into the fuzzy c-means clustering algorithm, so that a cluster could be chosen that matches an object's location to estimate its location; the proposed algorithm resulted in little improvement in localization distance error within 2 m [18]. In Reference [19], affinity propagation was proposed with the coarse positioning algorithm to cluster the off-line of the database; the coarse algorithm works within one or more clusters to estimate the location of the object.

A new idea was proposed in Reference [20] by using a probabilistic distribution measurement, using a Bayesian network as a probabilistic framework to estimate the object's location. The authors in Reference [21] proposed a modified probability neural network to estimate the location of the object, and this method outperformed the lateration technique. The authors in Reference [22] used a histogram of the RSS as a kernel method to estimate the object's location. In Reference [23], the Kullback–Leibler divergence (KLD) algorithm was proposed to estimate the probability density function (PDF) as a composite hypothesis test between the test point and fingerprinting radio map, whereas in Reference [24], to estimate the location of the object, the authors assumed that the RSS had a multivariate Gaussian and used the KLD algorithm to estimate the PDF impact of the test point on the fingerprinting radio map. In Reference [25], a low energy RSS-based Bluetooth technique was proposed to create a radio map for fingerprinting, after which probabilistic kernel regression based on the KLD was used to estimate the location of the object. The localization distance error was approximately 1 m in an office environment.

### 3. Overall Structure of the IPS

A typical WiFi fingerprint-based localization scenario was performed, in which a person held a smartphone device that had WiFi access, which was used to collect RSS measurements from different APs at various locations within the College of Engineering and Applied Sciences (CEAS) at Western Michigan University (WMU). As mentioned in Reference [26], an RSS distribution from multiple APs as a multimodal distribution commonly occurs. In our study, the signal-to-noise ratio was recorded for 35 min in a long corridor for a single AP. The mobile robot would stop every five minutes at each location and move 4 m further, and these steps were repeated for seven locations. We noticed values that differed by as much as 10 dBm, as shown in Figure 1.



**Figure 1.** Signal-to-noise ratio of received strength signal indicator variations over time.

There are many parameters that can affect the distribution of a signal such as diffraction, reflection, and pedestrian traffic [27]. We looked for a scenario that would lead to a better distribution of the AP signals. During the offline phase, a realistic scenario was performed that took signal variation into account. Because the human body can be an obstacle for signals, including the person holding the phone and the pedestrian in traffic, the fingerprint radio map was recorded from four different directions ( $45^\circ$ ,  $135^\circ$ ,  $225^\circ$ , and  $315^\circ$ ). At each RP, the RSS data were collected within the time sample, which was denoted as  $\{q_{i,j}^{(\circ)}(\tau), \tau = 1, \dots, t, t = 100\}$ , where  $(\circ)$  is the orientation direction and  $t$  represents the number of time samples. The covariance matrix and average of the RSS were calculated from four different directions, and 10 scans were used to create the radio map of the fingerprinting database, as represented by  $Q^{(\circ)}$  [28]:

$$Q^{(\circ)} = \begin{pmatrix} q_{1,1}^{(\circ)} & q_{1,2}^{(\circ)} & \cdots & q_{1,N}^{(\circ)} \\ q_{2,1}^{(\circ)} & q_{2,2}^{(\circ)} & \cdots & q_{2,N}^{(\circ)} \\ \vdots & \vdots & \ddots & \vdots \\ q_{L,1}^{(\circ)} & q_{L,2}^{(\circ)} & \cdots & q_{L,N}^{(\circ)} \end{pmatrix} \quad (1)$$

where  $q_{i,j}^{(\circ)} = \frac{1}{t} \sum_{\tau=1}^t q_{i,j}^{(\circ)}(\tau)$  and  $t = 10$ , which were arbitrarily chosen from 100 time samples. This can help us calculate the average value of RSS data over time for different APs,  $i = 1, 2, \dots, L$ ,  $j = 1, 2, \dots, N$ , where  $L$  is the number of APs and  $N$  represents the number of RPs. The variance vector of each RP can be defined as:

$$\Delta_j^{(\circ)} = [\Delta_{1,j}^{(\circ)}, \Delta_{2,j}^{(\circ)}, \Delta_{3,j}^{(\circ)}, \dots, \Delta_{L,j}^{(\circ)}] \quad (2)$$

where

$$\Delta_{i,j}^{(\circ)} = \frac{1}{t-1} \sum_{\tau=1}^t (q_{i,j}^{(\circ)}(\tau) - q_{i,j}^{(\circ)})^2 \quad (3)$$

where  $\Delta_{i,j}^{(\circ)}$  is the variance for AP  $i$  at RP  $j$  with orientation  $(\circ)$ ; thus, the database table of the radio map is  $(x_j, y_j, q_j^{(\circ)}, \Delta_j^{(\circ)})$  with  $q_j^{(\circ)}$  defined as:

$$q_j^{(\circ)} = [q_{1,j}^{(\circ)}, q_{2,j}^{(\circ)}, q_{3,j}^{(\circ)}, \dots, q_{L,j}^{(\circ)}] \quad (4)$$

During the online phase, the RSS measurement is denoted as:

$$p_r = [p_{1,r}, p_{2,r}, p_{3,r}, \dots, p_{L,r}] \quad (5)$$

#### 4. Bregman Divergence Algorithm Formulation

The heterogeneity of RSS data makes it difficult to design IPSs with high accuracy that are dependent on fingerprinting-based locations. Indeed, the  $L_p$ -norm and usual Euclidean distance do not always lead to IPSs with the highest accuracy, especially for systems with various histograms and other geometric features. It has been shown that using the information-theoretic relative entropy, known as the KLD, can lead to better results [29]. Bregman divergence has become a more attractive method for measuring similarity/dissimilarity between classes because it encapsulates the geometric Euclidean distance and information-theoretic relative entropy. The Bregman divergence  $D_F$  between two sets of data,  $p = (p_1, \dots, p_d)$  and  $q = (q_1, \dots, q_d)$ , and that associated with  $F$  (defined as a strictly convex function) can be defined as:

$$D_F(p, q) = F(p) - F(q) - \langle \nabla F(p), p - q \rangle \quad (6)$$

where  $\langle \dots, \dots \rangle$  denotes the dot product:

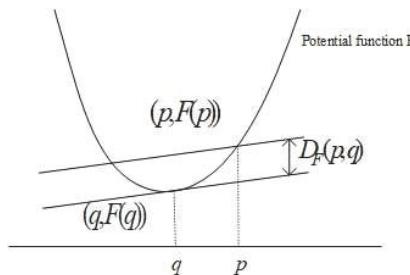
$$\langle p, q \rangle = \sum_{i=1}^d p^{(i)} q^{(i)} = p^T q \quad (7)$$

and  $\nabla F(p)$  denotes the gradient decent operator:

$$\nabla F(p) = \left[ \frac{\partial F}{\partial p_1}, \dots, \frac{\partial F}{\partial p_d} \right]^T \quad (8)$$

The Bregman distance unifies the KLD with the Euclidean distance by defining dissimilarity measurements as follows:

- The squared Euclidean distance is measured by substituting the convex function of the Bregman as  $F(p) = \sum_{i=1}^d p_i^2 = \langle p, p \rangle$ , as shown in Figure 2.
- The Bregman divergence will lead to the KLD if the strictly convex function used is



**Figure 2.** The Bregman divergence represents the vertical distance between the potential function and hyperplane at  $q$ .

$F(p) = \sum_{i=1}^d p_i \log p_i$ , which is defined as negative Shannon entropy. The KLD is defined as:

$$KL(p||q) = \sum_s p(S=s) \log\left(\frac{p(S=s)}{q(S=s)}\right) \quad (9)$$

In information-theoretic relative entropy, the Shannon entropy measures the uncertainty of a random variable by:

$$H(p) = p \log\frac{1}{p} \quad (10)$$

The KLD is equal to the cross-entropy of two discrete distributions minus the Shannon differential entropy [30]:

$$KL(p||q) = \sum_s H^x(p(s) || q(s)) - H(p(s)) \quad (11)$$

where  $H^x$  is the cross-entropy:

$$H^x(p(s) || q(s)) = \sum_s p(s) \log\frac{1}{q(s)} \quad (12)$$

Such a KLD has two major drawbacks. First, the output is undefined if  $q = 0$  and  $p \neq 0$ ; and second, the KLD is not bound by terms of metric distance. To avoid these drawbacks and avoid the  $\log(0)$  or to divide by 0, the authors in Reference [31] proposed a Jensen–Shannon divergence (JSD) dependent on the KLD as follows:

$$JSD(p||q) = \frac{1}{2} \left( KL\left(p, \frac{p+q}{2}\right) + KL\left(q, \frac{p+q}{2}\right) \right) \quad (13)$$

The JSD can be defined, is bound by an L1-metric, and is finite. In the same vein, the Bregman divergence ( $SD_F$ ) can be symmetrized as:

$$\begin{aligned} SD_F(p||q) &= \frac{1}{2} \left( D_F\left(p, \frac{p+q}{2}\right) + D_F\left(q, \frac{p+q}{2}\right) \right) \\ &= \frac{F(p) + F(q_j)}{2} - F\left(\frac{p+q_j}{2}\right) \end{aligned} \quad (14)$$

where  $p$  represents the test point dataset,  $q$  represents the fingerprint dataset, and  $j$  represents the number of APs that the smartphone has received. Because  $F$  is a strictly convex function, the  $SD(p||q)$  equals zero if and only if  $p = q$ ; the geometric interpretation for this is represented in Figure 3. For a positive definite matrix, the JBD is known as the Mahalanobis distance.

$$\begin{aligned} SD(p, q) &= \frac{F(p) + F(q)}{2} - F\left(\frac{p+q}{2}\right) \\ &= \frac{2\langle Qp, p \rangle + 2\langle Qq, q \rangle - 2\langle Q(p+q), p+q \rangle}{4} \\ &= \frac{1}{4} (\langle Qp, p \rangle + \langle Qq, q \rangle - 2\langle Qp, q \rangle) \\ &= \frac{1}{4} \langle Q(p - q), p - q \rangle \\ &= \frac{1}{4} \|p - q\|_Q^2 \end{aligned}$$

Due to RSS variation and the hardware variance problem, the fingerprinting database of the offline phase was clustered by using a clustering algorithms technique. The k-means algorithm was proposed by Lloyd in 1957 [32], who is considered a pioneer in clustering methods. In general, the k-means was used to solve the vector quantization problem. k-means is an iterative clustering algorithm that works by choosing random data points (seeds) to be the initial centroid (cluster center); the points of each cluster are associated with the closest cluster center. Each cluster center is updated and reiterated until the difference between any successive calculation goes below the “loss function” or convergence is met. The squared Euclidean distance is used to minimize the intra-cluster distance that leads to the

centroids. Lloyd [32] further proved that the iterative k-means algorithm monotonically converges to a local optima of the quadratic function loss (minimum variance loss). The cluster  $C_i$ 's center  $c_i$  is defined as follows:

$$c_i = \operatorname{argmin}_{p_j \in C_i} \sum \|p_j - c_i\| \quad (15)$$

$$= \operatorname{argmin}_{C_i} AVG_{L_2^2}(C_i, c) \quad (16)$$

$$c_i = \frac{1}{|C_i|} \sum_{p_j \in C_i} p_j \quad (17)$$

where  $c_i$  denotes the center of the cluster  $C_i$ , and  $|C_i|$  denotes the cardinality of  $C_i$ . In 2004, Reference [33] proposed a new clustering algorithm method, in which the k-means algorithm is modified by using the symmetric Bregman divergence. The minimum distance of the centroid of the point set has been defined as:

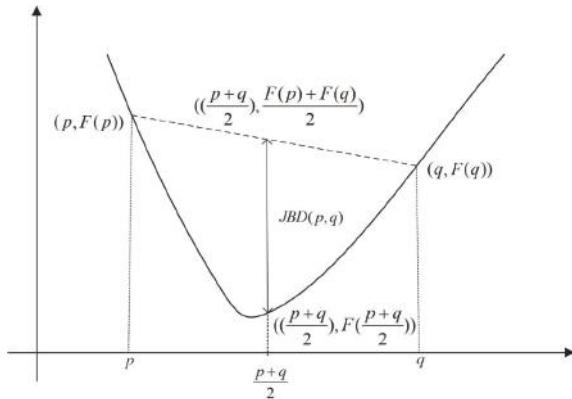
$$c = \operatorname{arg}_p \min = \frac{1}{n} \sum_i SD_F(p, p_i) \quad (18)$$

$$c_R^F = \operatorname{arg}_{c \in RP} \min \frac{1}{n} \sum_{i=1}^n SD_F(p_i || c) \quad (19)$$

$$c_L^F = \operatorname{arg}_{c \in RP} \min \frac{1}{n} \sum_{i=1}^n SD_F(c || p_i) \quad (20)$$

$$c^F = \operatorname{arg}_{c \in RP} \min \frac{1}{n} \sum_{i=1}^n \frac{SD_F(c || p_i) + SD_F(c || p_i)}{2} \quad (21)$$

where  $c_R^F$  and  $c_L^F$  represent the right- and left-sided centroid, the centroid  $c^F$  stands for the symmetrized Bregman divergence centroid, and  $n$  stands for the number of cells of the off-line database in each cluster.



**Figure 3.** Interpreting the Jensen-Bregman divergence.

## 5. Overall Structure of Proposed Positioning Algorithm

Designing an IPS by depending on fingerprinting-based locations is difficult because the environment suffers from inference and discrimination, which can lead to a heterogeneous RSS. As a result, depending on  $L_p$ -norm or square Euclidean distance algorithms do not always lead to systems with high accuracy. For example, it was proved in Reference [7] that the concave-convex procedure can obtain higher accuracy than algorithms that depend on the square Euclidean distance such as the kNN and probabilistic neural network (PNN). In this section, we introduce the symmetric

Hölder divergence. To measure the similarity between  $p$  and  $q$ , where  $rhs$  and  $lhs$  denote the right-hand side and left-hand side, respectively, one can use bi-parametric inequalities, i.e., one can use  $lhs(p,q) \leq rhs(p,q)$ , and a similarity can be measured by using the log-ratio gap:

$$D(p : q) = -\log\left(\frac{lhs(p,q)}{rhs(p,q)}\right) = \log\left(\frac{rhs(p,q)}{lhs(p,q)}\right) \geq 0 \quad (22)$$

The Hölder divergence between two values  $p(x)$  and  $q(x)$  is:

$$D^H(p : q) = -\log\left(\frac{\int p(x)^{\gamma/\alpha} q(x)^{\gamma/\beta} dx}{\left(\int p(x)^\gamma dx\right)^{1/\alpha} \left(\int q(x)^\gamma dx\right)^{1/\beta}}\right) \quad (23)$$

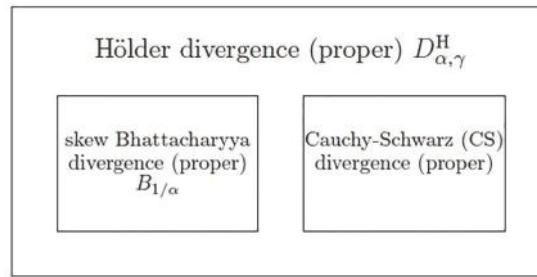
where  $\gamma$  represents the power of the absolute value Lebesgue integrable,  $\alpha, \beta$  represents the conjugate exponents, and  $p(x)$  and  $q(x)$  are positive measures as scalar values. Hölder divergence suffers from the law of the identity of indiscernible (self-distance is not equal to zero if  $p(x) = q(x)$ ), the triangle-inequality, and the symmetry. The Hölder divergence encapsulates both the one-parameter family of skew Bhattacharyya divergence and Cauchy–Schwarz divergence [34]. The Hölder divergence yields to the Cauchy–Schwarz divergence if we set  $\gamma, \alpha, \beta = 2$ :

$$D_{2,2}^H(p : q) = CS(p : q) := -\log\left(\frac{\int p(x)q(x)dx}{\left(\int p(x)^2 dx\right)^{1/2} \left(\int q(x)^2 dx\right)^{1/2}}\right) \quad (24)$$

The Hölder divergence will yield to the skew Bhattacharyya divergence if we set  $\gamma=1$ :

$$D_{\alpha,1}^H(p : q) = B_{1/\alpha}(p : q) := -\log\left(\int p(x)^{1/\alpha} q(x)^{1/\beta} dx\right) \quad (25)$$

The relationship between the divergence families is illustrated in Figure 4.



**Figure 4.** Hölder divergence encompasses the skew Bhattacharyya divergence and the Cauchy–Schwarz divergence.

Similarly, for conjugate exponents  $\beta$  and  $\alpha$ , the Hölder divergence satisfies:

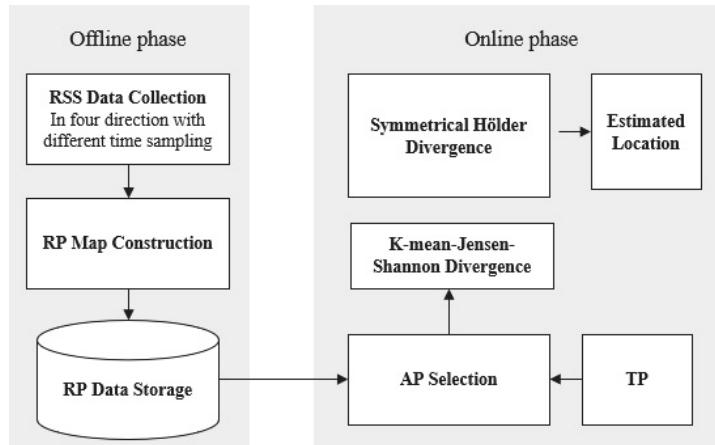
$$D_{\alpha,\gamma}^H(p : q) = D_{\beta,\gamma}^H(p : q) \quad (26)$$

The symmetrized Hölder divergence is:

$$D_{\alpha,1}^H(p : q) = \frac{1}{2}\left(D_{\alpha,\gamma}^H(p : q) + D_{\alpha,\gamma}^H(q : p)\right) \quad (27)$$

$$= \frac{1}{2}\left[F(\gamma p) + F(\gamma q) - F\left(\frac{\gamma}{\alpha}p + \frac{\gamma}{\beta}q\right) - F\left(\frac{\gamma}{\beta}p + \frac{\gamma}{\alpha}q\right)\right] \quad (28)$$

To improve the accuracy of the IPS, we proposed that sided and symmetrized Bregman centroids incorporate the symmetrized Hölder divergence. Furthermore, we introduce three different approaches to define the APs that will be used in the proposed algorithm, as shown in Figure 5.



**Figure 5.** The offline and online stages of location WiFi-based fingerprinting architecture.

- *Strongest APs (MaxMean) [35]*

Previous studies have proposed that the RSS be chosen based on the signal strength in the online phase, and that the same set of APs from the fingerprinting radio map be used in the calculations, with the assumption that the APs with the highest signal provide the highest coverage over time. However, the strongest AP scheme may not render a good criterion in our calculation.

- *Fisher Criterion:*

The Fisher criterion is a metric that is used to quantify the discrimination ability of APs across a fingerprinting radio map in four different orientations. The statistical properties of the RPs are used to determine the APs that will be used based on their performance. A score is pointed to each AP separately as [36]:

$$\xi_i = \frac{\sum_{j=1}^N \left( q_j^{i(o)} - \bar{q}_i \right)^2}{\sum_{j=1}^N \Delta_j^{i(o)}} \quad (29)$$

$$\bar{q}_i = \frac{1}{N} \sum_{j=1}^N q_j^i \quad (30)$$

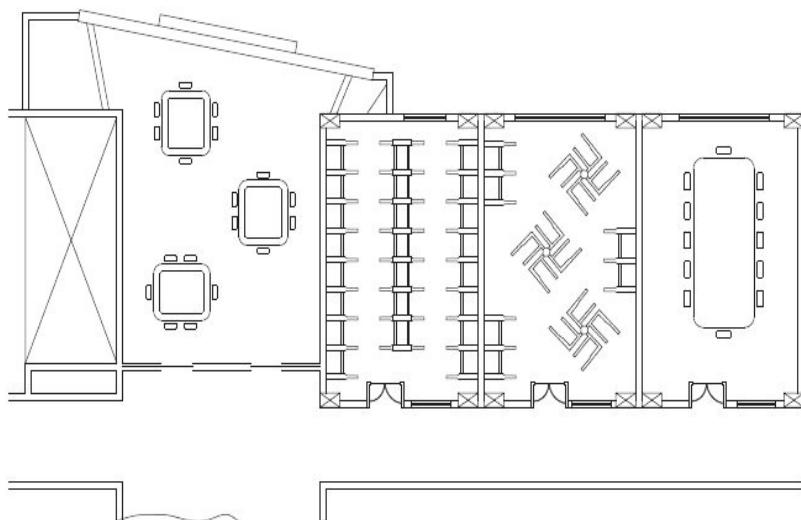
The Fisher criterion proposes that APs with higher variance are less reliable to use in IPS calculations; the APs will be sorted with respect to their score, and those with high scores will be much more likely to be selected. However, Fisher criterion discrimination is only used in offline fingerprinting based-localization. If one or more APs are not available in the online phase, the Fisher criterion is not suitable to use.

- *Random Selection*

Unlike the above schemes, in which APs are selected based on some criteria, in random selection, the APs are selected arbitrarily without considering AP performance. This scheme has less computational complexity, as the matrix of the APs needs to be generated at different runs and does not need the variance to be calculated, as with the Fisher criterion.

## 6. Simulation and Implementation Results

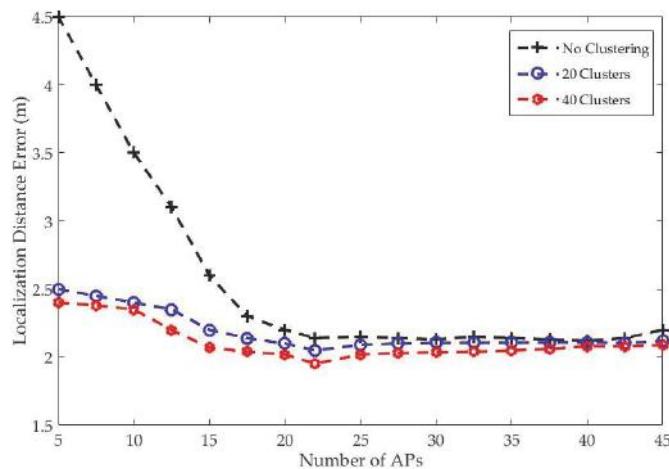
This section provides details on the proposed algorithms outlined in subsequent subsections. The RSS data were collected on the first floor of the CEAS at WMU with an area of interest map, as shown in Figure 6. A Samsung smartphone with operating system 4.4.2 (S5, Samsung Company, Suwon, Korea) was used to collect the RSS data. Furthermore, the proposed algorithms were implemented on an HP Laptop using Java software (HP, Beijing, China) with an Eclipse framework (Photon, IBM, NY, USA). Cisco Linksys E2500 Simultaneous Dual-Band Routers were used for the area of interest. The RSS value and MAC address of the WiFi APs were collected within a time frame of 1 s for 100 s over 84 RPs within an average grid of 1 m. At each RP, a total of 47 APs were detected throughout the area of interest.



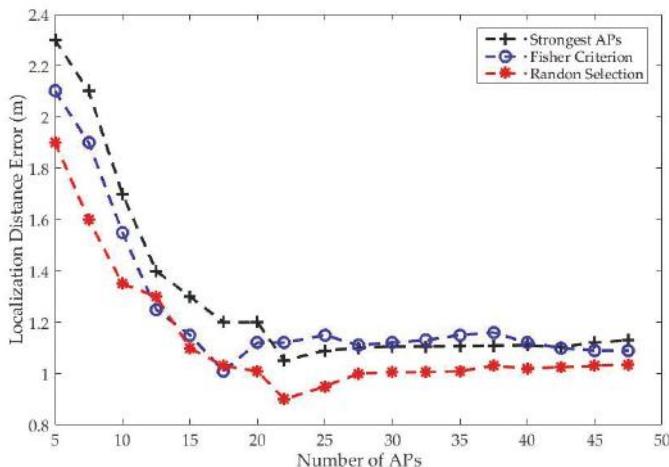
**Figure 6.** The layout used in the experimental work in the College of Engineering and Applied.

To evaluate the performance, online phase data were collected in varying environments on different days in 65 unknown locations with four repetitions as test points. The localization distance error was measured by calculating the Euclidean distance between the actual location of the testing point and the location that was estimated by the proposed algorithms. To reduce the RSS time variation, the k-means-Bregman divergence was used on the fingerprinting radio map to cluster the offline data. Figure 7 illustrates the effects of the clustering algorithms on localization distance error with the number of APs when five NNs are used. As shown in Figure 7, the localization distance error was decreased as the numbers of cluster increased, which reduced the area of interest that could improve object localization.

Figure 8 shows the localization distance error when a different AP selection scheme was used with the symmetrized Hölder divergence and k-mean-Bregman divergence, where the y-axis is the localization distance error and the x-axis is the number of APs. The Fisher criterion had the highest accuracy when the APs were less than 18, and the proposed random scheme achieved the next highest performance. The strongest AP scheme had a lower accuracy than the other schemes. In general, using more APs may not necessarily yield the lowest localization error. As shown in Figure 8, the best performance occurred when 22 APs were used; as the number of APs increased after that, the performance of the proposed systems decreased. Thus, we conclude that not only the number but also the selection scheme of APs can affect the IPS performance.



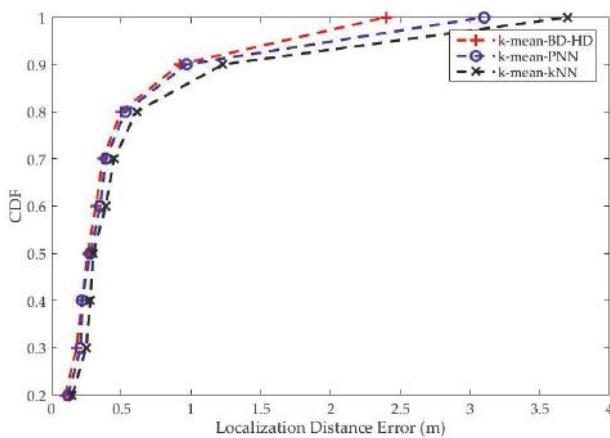
**Figure 7.** The implementation results of different number of clusters with respect to the average of the localization distance.



**Figure 8.** The implementation result of the average localization error under different AP selection schemes.

#### Comparison to Prior Work

The proposed fingerprint based-localization method is compared with prior fingerprinting approaches such as the kernel-based localization method, kNN. Figure 9 illustrates the corresponding cumulative probability distributions of the localization error for the three methods. In particular, the median error for the k-means-BD-HD was 0.92 m, 0.97 m for k-means-PNN, and 1.23 m for k-means-kNN.



**Figure 9.** Experiment results: The Cumulative distribution function (CDF) of localization error when using 50 nearest neighbors.

As noticed, the proposed k-means-BD-HD method provides a 90th percentile error of 0.92 m, while for k-means-PNN it was 0.97 m, and for k-means-kNN it was 1.23 m.

## 7. Conclusions

IPSSs incorporate the power of GPS and indoor mapping and have many potential applications that make them very important in modern life. For example, they can be used for healthcare services such as aiding people with impaired vision, and navigating unfamiliar buildings (e.g., malls, airports, subways). Several large companies, such as Apple, Google, and Microsoft, started a fund to initiate research on IPSSs. Cluster methods can be used to reduce the impact of time variation by clustering the fingerprinting radio map into multiple partitions and then choosing the cluster that has the lowest distance error. A radio map fingerprint was developed in CEAS to investigate different localization algorithms and compare different approaches such as kNN and PNN. We proposed a symmetrical Hölder divergence, which uses statistical entropy that encapsulates both skew Bhattacharyya divergence and Cauchy–Schwarz divergence, and assessed their performance with different AP selection schemes. The results were quite adequate for the indoor environment with an average error of less than 1 m. the symmetrical Hölder divergence that incorporated the k-means-Bregman divergence had the highest accuracy when 25 clusters were used with 22 APs.

We are currently in the process of investigating the user position inside smaller clusters/areas and position prediction error distributions and quantifying the localization variation of WiFi signals distributed in space.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

- Markets. *Indoor Localization Market by Positioning Systems, Map and Navigation, Location based Analysis, Monitoring and Emergency Services-Worldwide Market Forecasts and Analysis (2014–2019)*; Technical Report; Markets: Limassol, Cyprus, 2014.
- Torres-Sospedra, J.; Montoliu, R.; Trilles, S.; Belmonte, O.; Huerta, J. Comprehensive analysis of distance and similarity measures for Wi-Fi fingerprinting indoor positioning systems. *Expert Syst. Appl.* **2015**, *42*, 9263–9278. [CrossRef]

3. Jiang, P.; Zhang, Y.; Fu, W.; Liu, H.; Su, X. Indoor Mobile Localization Based on Wi-Fi Fingerprint’s Important Access Point. *Int. J. Distrib. Sens. Netw.* **2015**. [[CrossRef](#)]
4. Shchekotov, M. Indoor localization methods based on Wi-Fi lateration and signal strength data collection. In Proceedings of the 2015 17th Conference of Open Innovations Association (FRUCT), Yaroslavl, Russia, 20–24 April 2015.
5. Swangmuang, N.; Prashant, K. An Effective Location Fingerprint Model for Wireless Indoor Localization. *Pervasive Mob. Comput.* **2008**, *4*, 836–850. [[CrossRef](#)]
6. Wang, B.; Zhou, S.; Liu, W.; Mo, Y. Indoor Localization Based on Curve Fitting and Location Search Using Received Signal Strength. *IEEE Trans. Ind. Electron.* **2015**, *62*, 572–582. [[CrossRef](#)]
7. Abdullah, O.; Abdel-Qader, I.; Bazuin, B. A probability neural network-Jensen-Shannon divergence for a fingerprint based localization. In Proceedings of the 2016 Annual Conference on Information Science and Systems (CISS), Princeton, NJ, USA, 16–18 March 2016.
8. Abdullah, O.; Abdel-Qader, I. A PNNT-Jensen-Bregman Divergence symmetrization for a WLAN Indoor Positioning System. In Proceedings of the 2016 IEEE International Conference on Electro Information Technology (EIT), Grand Forks, ND, USA, 19–21 May 2016.
9. Abdullah, O.; Abdel-Qader, I.; Bazuin, B. Fingerprint-based technique for indoor positioning system via machine learning and convex optimization. In Proceedings of the 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 20–22 October 2016.
10. Abdullah, O.; Abdel-Qader, I.; Bazuin, B. K-means-Jensen-Shannon divergence for a WLAN indoor positioning system. In Proceedings of the 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 20–22 October 2016.
11. Abdullah, O.; Abdel-Qader, I.; Bazuin, B. Convex Optimization via Jensen-Bregman Divergence for WLAN Indoor Positioning System. *Int. J. Handheld Comput. Res.* **2017**, *8*, 29–41. [[CrossRef](#)]
12. Sharma, P.; Chakraborty, D.; Banerjee, N.; Banerjee, D.; Agarwal, S.D.; Mittal, S. KARMA: Improving WiFi-based indoor localization with dynamic causality calibration. In Proceedings of the 2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), Singapore, 30 June–3 July 2014.
13. Hähnle, B.; Dirk, B.; Fox, D. Gaussian processes for signal strength-based location estimation. In Proceedings of the Robotics: Science and Systems, Ann Arbor, MI, USA, 18–22 June 2006.
14. Chan, E.C.; Baci, G.; Mak, S. Using Wi-Fi Signal Strength to Localize in Wireless Sensor Networks. In Proceedings of the WRI International Conference on Communications and Mobile Computing, Yunnan, China, 6–8 January 2009.
15. Noh, Y.; Yamaguchi, H.; Lee, U.; Vij, P.; Joy, J.; Gerla, M. CLIPS: Infrastructure-free collaborative indoor positioning scheme for time-critical team operations. In Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom’13), San Diego, CA, USA, 18–22 March 2013; pp. 172–178.
16. Ma, J.; Li, X.; Tao, X.; Lu, J. Cluster filtered KNN: A WLAN based indoor positioning scheme. In Proceedings of the IEEE International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM’08), Newport Beach, CA, USA, 23–26 June 2008; pp. 1–8.
17. Altintas, B.; Serif, T. Improving RSS-based indoor positioning algorithm via K-Means clustering. In Proceedings of the 11th European Wireless Conference 2011—Sustainable Wireless Technologies (European Wireless), Vienna, Austria, 27–29 April 2011; pp. 1–5.
18. Sun, Y.; Xu, Y.; Ma, L.; Deng, Z. KNN-FCMhybrid algorithm for indoor location in WLAN. In Proceedings of the 2nd International Conference on Power Electronics and Intelligent Transportation System (PEITS’09), Shenzhen, China, 19–20 December 2009; Volume 2, pp. 251–254.
19. Tian, Z.; Tang, X.; Zhou, M.; Tan, Z. Fingerprint indoor positioning algorithm based on affinity propagation clustering. *EURASIP J. Wirel. Commun. Netw.* **2013**, *2013*, 272. [[CrossRef](#)]
20. Castro, P.; Chiu, P.; Kremeneck, T.; Muntz, R. A Probabilistic Location Service for Wireless Network Environments. In Proceedings of the International Conference on Ubiquitous Computing (Ubicomp’2001), Atlanta, GA, USA, 30 September–2 October 2001.

21. Chen, C.; Chen, Y.; Yin, L.; Hwang, R. A Modified Probability Neural Network Indoor Positioning Technique. In Proceedings of the 2012 International Conference on Information Security and Intelligent Control, Yunlin, Taiwan, 14–16 August 2012.
22. Roos, T.; Myllymäki, P.; Tirri, H.; Misikangas, P.; Sievänen, J. A Probabilistic Approach to WLAN User Location Estimation. *Int. J. Wirel. Inf. Netw.* **2002**, *9*, 155–164. [[CrossRef](#)]
23. Tsui, W.A.; Chuang, Y.; Chu, H. Unsupervised Learning for Solving RSS Hardware Variance Problem in WiFi Localization. *Mob. Netw. Appl.* **2009**, *14*, 677–691. [[CrossRef](#)]
24. Miliotis, D.; Kriara, L.; Papakonstantinou, A.; Tzagkarakis, G. Empirical Evaluation of Signal-Strength Fingerprint Positioning in Wireless LANs. In Proceedings of the 13th ACM International Conference on Modeling, Analysis, and Simulation of Wireless and Mobile Systems, Bodrum, Turkey, 17–21 October 2010.
25. Mirowski, P.; Steck, H.; Whiting, P.R.; Palaniappan, R.; MacDonald, M.; Ho, T.K. KL-Divergence Kernel Regression for Non-Gaussian Fingerprint Based Localization. In Proceedings of the 2011 International Conference on Indoor Positioning and Indoor Navigation, Guimaraes, Portugal, 21–23 September 2011.
26. Bahl, P.; Padmanabhan, V.N. RADAR: An in-building RF-based user location and tracking system. In Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE INFOCOM 2000, Tel Aviv, Israel, 26–30 March 2000.
27. Youssef, M.; Agrawala, A. The Horus WLAN Location Determination System. In Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services, Seattle, WA, USA, 6–8 June 2005.
28. Feng, C.; Au, W.S.A.; Valaee, S.; Tan, Z. Received-Signal-Strength-Based Indoor Positioning Using Compressive Sensing. *IEEE Trans. Mob. Comput.* **2012**, *11*, 1983–1993. [[CrossRef](#)]
29. Nielsen, F.; Nock, R. Skew Jensen-Bregman Voronoi Diagrams. In *Transaction on Computer Science XIV*; Springer: Berlin/Heidelberg, Germany, 2011.
30. Nielsen, F. A family of statistical symmetric divergences based on Jensen’s inequality. *arXiv*, 2010.
31. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [[CrossRef](#)]
32. Lloyd, S.P. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–136. [[CrossRef](#)]
33. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
34. Nielsen, F.; Sun, K.; Marchand-Maillet, S.  $k$ -Means Clustering with Hölder Divergences. In Proceedings of the International Conference on Geometric Science of Information, Paris, France, 7–9 November 2017.
35. Youssef, M.; Agrawala, A.; Udaya Shankar, A. WLAN location determination via clustering and probability distributions. In Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications, Fort Worth, TX, USA, 24–26 March 2003.
36. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*, 2nd ed.; Wiley-InterScience: Hoboken, NJ, USA, 2000.



© 2018 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Likelihood Ratio Testing under Measurement Errors

Michel Broniatowski <sup>1,\*</sup>, Jana Jurečková <sup>2,3</sup> and Jan Kalina <sup>4</sup>

<sup>1</sup> Faculté de Mathématiques, Laboratoire de Probabilité, Statistique et Modélisation, Université Pierre et Marie Curie (Sorbonne Université), 4 place Jussieu, 75252 Paris CEDEX 05, France

<sup>2</sup> Institute of Information Theory and Automation, The Czech Academy of Sciences, Pod Vodárenskou věží 4, 182 08 Prague 8, Czech Republic; jureckova@utia.cas.cz

<sup>3</sup> Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Prague 8, Czech Republic

<sup>4</sup> Institute of Computer Science, The Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic; kalina@cs.cas.cz

\* Correspondence: michel.broniatowski@sorbonne-universite.fr

Received: 13 November 2018; Accepted: 7 December 2018; Published: 13 December 2018



**Abstract:** We consider the likelihood ratio test of a simple null hypothesis (with density  $f_0$ ) against a simple alternative hypothesis (with density  $g_0$ ) in the situation that observations  $X_i$  are mismeasured due to the presence of measurement errors. Thus instead of  $X_i$  for  $i = 1, \dots, n$ , we observe  $Z_i = X_i + \sqrt{\delta}V_i$  with unobservable parameter  $\delta$  and unobservable random variable  $V_i$ . When we ignore the presence of measurement errors and perform the original test, the probability of type I error becomes different from the nominal value, but the test is still the most powerful among all tests on the modified level. Further, we derive the minimax test of some families of misspecified hypotheses and alternatives. The test exploits the concept of pseudo-capacities elaborated by Huber and Strassen (1973) and Buja (1986). A numerical experiment illustrates the principles and performance of the novel test.

**Keywords:** measurement errors; robust testing; two-sample test; misspecified hypothesis and alternative; 2-alternating capacities

---

## 1. Introduction

Measurement technologies are often affected by random errors; if the goal of the experiment is to compare two probability distributions using data, then the conclusion can be distorted if the data are affected by some measurement errors. If the data are mismeasured due to the presence of measurement errors, the statistical inference performed with them is biased and trends or associations in the data are deformed. This is common for a broad spectrum of applications e.g., in engineering, physics, biomedicine, molecular genetics, chemometrics, econometrics etc. Some observations can be even undetected, e.g., in measurements of magnetic or luminous flux in analytical chemistry when the flux intensity falls below some flux limit. Actually, we can hardly imagine real data free of measurement errors; the question is how severe the measurement errors are and what their influence on the data analysis is [1–3].

A variety of functional models have been proposed for handling measurement errors in statistical inference. Technicians, geologists, and other specialists are aware of this problem, and try to reduce the effect of measurement errors with various ad hoc procedures. However, this effect cannot be completely eliminated or substantially reduced unless we have some additional knowledge on the behavior of measurement errors.

There exists a rich literature on the statistical inference in the error-in-variables (EV) models as is evidenced by the monographs of Fuller [4], Carroll et al. [5], and Cheng and van Ness [6], and the references therein. The monographs [4] and [6] deal mostly with classical Gaussian set up while [5]

discusses numerous inference procedure under semi-parametric set up. Nonparametric methods in EV models are considered in [7,8] and in references therein, and in [9], among others. The regression quantile theory in the area of EV models was started by He and Liang [10]. Arias [11] used an instrumental variable estimator for quantile regression, considering biases arising from unmeasured ability and measurement errors. The papers dealing with practical aspects of measurement error models include [12–16], among others. Recent developments in treating the effect of measurement errors on econometric models was presented in [17] or [18]. The advantage of *rank and signed rank procedures* in the measurement errors models was discovered recently in [19–24]. The problem of interest in the present paper is to study how the measurement errors can affect the conclusion of the likelihood ratio test.

The distribution function of measurement errors is considered unknown, up to zero expectation and unit variance. When we use the the likelihood ratio test while ignoring the possible measurement errors, we can suffer a loss in both errors of the first and second kind. However, we show that under a small variance of measurement errors, the original likelihood ratio test is still most powerful, only on a slightly changed significance level.

On the other hand, we may consider the situation that  $\mathbf{H}_0$  or  $\mathbf{H}_1$  are classes of distributions of random variables  $Z + \sqrt{\delta}V$ . Hence, both hypothesis and alternative are composite as families  $\mathbf{H}_0$  and  $\mathbf{H}_1$ ; if they are bounded by alternating Choquet capacities of order 2, then we can look for a minimax test based on the ratio of the capacities, and/over on the ratio of the pair of the least favorable distributions of  $\mathbf{H}_0$  and  $\mathbf{H}_1$ , respectively (cf. Huber and Strassen [25]).

## 2. Likelihood Ratio Test under Measurement Errors

Our primary goal is to test the null hypothesis  $\mathcal{H}_0$  that independent observations  $\mathbf{X} = (X_1, \dots, X_n)^\top$  come from a population with a density  $f$  against the alternative  $\mathcal{H}_1$  that the true density is  $g$ , where  $f$  and  $g$  are fixed densities of our interest. For the identifiability, we shall assume that  $f$  and  $g$  are continuous and symmetric around 0. Although the alternative is the main concern of the experimenter, some measurement errors or just the nature may cause the situation that the true alternative should be considered as composite. Specifically,  $X_1, \dots, X_n$ , can be affected by additive measurement errors, what appears in numerous fields, as illustrated in Section 1.

Hence the alternative is  $\mathbf{H}_{1,\delta}$  under which the observations are  $Z_{i,\delta} = X_i + \sqrt{\delta}V_i$ , identically distributed with continuous density  $g_\delta$ . Here, both under the hypothesis and under the alternative,  $V_i$  are independent random variables, unobservable with unknown distribution, independent of  $X_i$ ;  $i = 1, \dots, n$ . The parameter  $\delta > 0$  is also unknown, only we assume that  $E V_i = 0$  and  $EV_i^2 = 1$ , for simplicity. The mismeasured, hence unobservable,  $X_i$  are assumed to have the density  $g$  under the alternative. Quite analogously, the mismeasured observations lead to a composite hypothesis  $\mathbf{H}_{0,\delta}$  under which the density of observations  $Z_{i,\delta} = X_i + \sqrt{\delta}V_i$  is  $f_\delta$  while the  $X_i$  are assumed to have density  $f$ .

If we knew  $f_\delta$  and  $g_\delta$ , we would use the Neyman-Pearson critical region

$$W = \left\{ \mathbf{z} : \sum_{i=1}^n \ln \left( \frac{g_\delta(z_i)}{f_\delta(z_i)} \right) \geq u \right\} \quad (1)$$

with  $u$  determined so that

$$P_{f_\delta} \left\{ \sum_{i=1}^n \ln \left( \frac{g_\delta(z_i)}{f_\delta(z_i)} \right) \geq u \right\} = \alpha,$$

with a significance level  $\alpha$ . Evidently

$$\int I \left[ \sum_{i=1}^n \ln \left( \frac{g_\delta(z_i)}{f_\delta(z_i)} \right) \geq u \right] \prod_{i=1}^n g_\delta(z_i) dz_i = \int I \left[ \sum_{i=1}^n \ln \left( \frac{g(x_i)}{f(x_i)} \right) \geq u \right] \prod_{i=1}^n g(x_i) dx_i$$

$$\int I \left[ \sum_{i=1}^n \ln \left( \frac{g_\delta(z_i)}{f_\delta(z_i)} \right) \geq u \right] \prod_{i=1}^n f_\delta(z_i) dz_i = \int I \left[ \sum_{i=1}^n \ln \left( \frac{g(x_i)}{f(x_i)} \right) \geq u \right] \prod_{i=1}^n f(x_i) dx_i.$$

Indeed, notice that

$$E_{g_\delta} \left\{ I \left[ \sum_{i=1}^n \ln \left( \frac{g_\delta(Z_i)}{f_\delta(Z_i)} \right) \geq u \right] \middle| V_1 = v_1, \dots, V_n = v_n \right\}$$

$$= E_g \left\{ I \left[ \sum_{i=1}^n \ln \left( \frac{g(X_i)}{f(X_i)} \right) \geq u \right] \middle| V_1 = v_1, \dots, V_n = v_n \right\} \quad \forall v_i \in \mathbb{R}, i = 1, \dots, n,$$

where the expectations are considered with respect to the conditional distribution; a similar equality holds for  $f_\delta$ .

Combining the integration transmission in the conditional distribution, we obtain

$$\int I \left[ \sum_{i=1}^n \ln \left( \frac{g_\delta(x_i + \sqrt{\delta}V_i)}{f_\delta(x_i + \sqrt{\delta}V_i)} \right) \geq u \right] \prod_{i=1}^n f(x_i) dx_i$$

$$\neq \int I \left[ \sum_{i=1}^n \ln \left( \frac{g(x_i)}{f(x_i)} \right) \geq u \right] \prod_{i=1}^n f(x_i) dx_i = \alpha, \quad (2)$$

hence the size of the critical region  $W$  when used for testing  $\mathcal{H}_0$  against  $\mathcal{H}_1$  differs from  $\alpha$ . Then we ask how the critical region  $W$  in (1) behaves when it is used as a test of  $\mathcal{H}_0$ . This problem we shall try to attack with an expansion of  $f_\delta, g_\delta$  in  $\delta$  close to zero.

## 2.1. Approximations of Densities

Put  $f = f_0$ ,  $g = g_0$  the densities of  $X$  under the hypotheses and alternative, respectively. For the identifiability, we shall assume that  $f_0$  and  $g_0$  are continuous and symmetric around 0. Denote  $f_\delta$  the density of  $Z_\delta = X + \sqrt{\delta}V$ . This means that  $X$  is affected by an additive measurement error  $\sqrt{\delta}V$ , where  $V$  is independent of  $X$  and  $EV = 0$ ,  $EV^2 = 1$ ,  $EV^4 < \infty$ . Notice that if densities of  $X$  and  $V$  are strongly unimodal, then that of  $Z$  is also strongly unimodal (see [26]). Under some additional conditions on  $f_0, g_0$ , we shall derive approximations of  $f_\delta$  and  $g_\delta$  for small  $\delta > 0$ . More precisely, we assume that both  $f_0$  and  $g_0$  have differentiable and integrable derivatives up to order 5. Then we have the following expansion of  $f_\delta$  and a parallel result for  $g_\delta$ :

**Theorem 1.** Assume that  $f_0$  and  $g_0$  are symmetric around 0, strongly unimodal with differentiable and integrable derivatives, up to the order 5. Then, as  $\delta \downarrow 0$ ,

$$f_\delta(z) = f_0(x + \sqrt{\delta}V) = f_0(x) + \frac{\delta}{2} \frac{d^2}{dz^2} f_0(x) + \frac{\delta^2}{4!} \frac{d^4}{dz^4} f_0(x) E(V^4) + o(\delta^2), \quad (3)$$

$$g_\delta(z) = g_0(x + \sqrt{\delta}V) = g_0(x) + \frac{\delta}{2} \frac{d^2}{dz^2} g_0(x) + \frac{\delta^2}{4!} \frac{d^4}{dz^4} g_0(x) E(V^4) + o(\delta^2)$$

**Proof.** Let  $\varphi(u, \delta) = E\{e^{iuZ}\}$  be the characteristic function of  $Z$ . Then

$$\begin{aligned}\varphi(u, \delta) &= E\{e^{iuX}\} E\{e^{iu\sqrt{\delta}V}\} = \varphi(0, 0)\varphi_V(u\sqrt{\delta}) \\ &= \varphi(u, 0) \left[ 1 + \frac{1}{2}\delta(iu)^2 + \frac{1}{4!}\delta^2(iu)^4 E(V^4) + o(\delta^2) \right] \\ &= \varphi(u, 0) \left[ 1 - \frac{\delta}{2}u^2 + \frac{1}{4!}\delta^2u^4 E(V^4) + o(\delta^2) \right],\end{aligned}$$

where  $\varphi_V$  denotes the characteristic function of  $V$ . Taking the inverse Fourier transform on both sides, we obtain (3), taking the above assumptions on  $V$  into account.  $\square$

Consider the problem of testing the hypothesis  $H_0$  that the observations are distributed according to density  $f_0$  against the alternative  $H_1$  that they are distributed according to density  $g_0$ . Parallelly, we consider the hypothesis  $H_{0,\delta}$  that observations are distributed according to  $g_\delta$  against the alternative  $H_{1,\delta}$  that the true density is  $g_\delta$ . Let  $\Phi(\mathbf{x})$  be the likelihood ratio test with critical region  $W = \left\{ \mathbf{x} : \sum_{i=1}^n \ln \left( \frac{g_0(x_i)}{f_0(x_i)} \right) > u \right\}$  and the significance level  $\alpha$ , and  $\Phi^* = \Phi^*(\mathbf{z})$  be the test with critical region  $W^* = \left\{ \mathbf{z} : \sum_{i=1}^n \ln \left( \frac{g_0(z_i)}{f_0(z_i)} \right) > u \right\}$  based on observations  $z_i = x_i + \sqrt{\delta}V_i$ ,  $i = 1, \dots, n$ . We know neither  $\delta$  nor  $V$ , hence the test  $\Phi^*$  is just an application of the critical region  $W$  for contaminated data  $Z_1, \dots, Z_n$ . Thus, due to our lack of information, we use the test  $\Phi$  even for testing  $H_{0,\delta}$  against  $H_{1,\delta}$ , and the performance of this test is of interest. This is described in the following theorem:

**Theorem 2** (Assume the conditions of Theorem 1). *Then, as  $\delta \downarrow 0$ , the test  $\Phi^*$  is the most powerful even for testing  $H_{0,\delta}$  against  $H_{1,\delta}$ , with a modified significance level satisfying*

$$\alpha_\delta \leq \alpha + \frac{\delta}{2} |f'_0(0)| + \frac{\delta^2}{24} EV^4 |f_0^{(3)}(0)| + \mathcal{O}(\delta).$$

**Proof.**

$$\begin{aligned}E_{f_0} \Phi^*(\mathbf{X}) &= \int I \left[ \ln \left( \frac{g_0(x + \sqrt{\delta}V)}{f_0(x + \sqrt{\delta}V)} \right) > u \right] f_0(x) dx \\ &= \int I \left[ \ln \left( \frac{g_0(x + \sqrt{\delta}V)}{f_0(x + \sqrt{\delta}V)} \right) > u \right] \frac{f_0(x)}{f_0(x + \sqrt{\delta}V)} f_0(x + \sqrt{\delta}V) dx \\ &= \int I \left[ \ln \left( \frac{g_0(x)}{f_0(x)} \right) > u \right] \frac{f_0(x - \sqrt{\delta}V)}{f_0(x)} f_0(x) dx.\end{aligned}$$

If  $f_0$  is symmetric, then the derivative  $f_0^{(k)}$  is symmetric for  $k$  even and skew-symmetric for  $k$  odd,  $k = 1, \dots, 4$ . Moreover, because  $|f'_0(x)|$  and  $|f_0^{(3)}(x)|$  are integrable, then  $\lim_{x \rightarrow \pm\infty} |f'_0(x)| = 0$  and  $\lim_{x \rightarrow \pm\infty} |f_0^{(3)}(x)| = 0$ . Hence, using the expansion (3), we obtain

$$\begin{aligned}E_{f_0} \Phi^*(\mathbf{X}) &= E_{f_0} \Phi(\mathbf{X}) + \int I \left[ \ln \left( \frac{g_0(x)}{f_0(x)} \right) > u \right] \left( \frac{\delta}{2} f''_0(x) + \frac{\delta^2}{24} EV^4 f^{(4)}(x) dx \right) + o(\delta^2) \\ &\leq E_{f_0} \Phi(\mathbf{X}) + \frac{\delta}{2} |f'_0(0)| + \frac{\delta^2}{24} EV^4 |f_0^{(3)}(0)| + o(\delta^2) = \alpha + \mathcal{O}(\delta) \text{ as } \delta \downarrow 0.\end{aligned}$$

$\square$

### 3. Robust Testing

If the observations are missmeasured or contaminated, we observe  $Z_\delta = Z + \sqrt{\delta}V$  with unknown  $\delta$  and unobservable  $V$  instead of  $Z$ . Hence, instead of simple  $f_0$  and  $g_0$ , we are led to composite hypothesis and alternative  $\mathcal{H}$  and  $\mathcal{K}$ . Following [25], we can try to find suitable 2-alternating capacities,

dominating  $\mathcal{H}$  and  $\mathcal{K}$  and to construct a pertaining minimax test. As before, we assume that  $Z$  and  $V$  are independent,  $EV = 0$ ,  $EV^2 = 1$ , and  $EV^4 < \infty$ . Moreover, we assume that  $f_0$  and  $g_0$  are symmetric, strongly unimodal and differentiable up to order 5, with derivatives integrable and increasing distribution functions  $F_0$  and  $G_0$ , respectively. The measurement errors  $V$  are assumed to satisfy

$$1 \leq EV^4 \leq K \quad (4)$$

with a fixed  $K$ ,  $0 < K < \infty$ . Hence the distribution of  $V$  is restricted to have the tails lighter than  $t$ -distribution with 4 degrees of freedom. We shall construct a pair of 2-alternating capacities around specific subfamilies of  $f_0$  and  $g_0$ .

Let us determine the capacity around  $g_0$ ; that for  $f_0$  is analogous. By Theorem 1 we have

$$g_\delta(z) = g_0(z) + \frac{\delta}{2} \frac{d^2}{dz^2} g_0(z) + \frac{\delta^2}{4!} \frac{d^4}{dz^4} g_0(z) E(V^4) + o(\delta^2), \quad \text{as } \delta \downarrow 0.$$

We shall concentrate on the following family  $\mathcal{K}^*$  of densities (similarly for  $f_0$ ):

$$\mathcal{K}^* = \left\{ g_{\delta,\kappa}^* : g_{\delta,\kappa}^*(z) = g_0(z) + \frac{\delta}{2} g_0''(z) + \kappa \frac{\delta^2}{24} g_0^{(4)}(z) \mid \delta \leq \Delta, 1 \leq \kappa \leq K \right\} \quad (5)$$

with fixed suitable  $\Delta, K > 0$ .

Indeed, under our assumptions, each  $g_{\delta,\kappa}^* \in \mathcal{K}^*$  is a positive and symmetric density satisfying

$$\sup_{\delta \leq \Delta, \kappa \leq K} \sup_{z \in \mathbb{R}} |g_{\delta,\kappa}^*(z) - g_0(z)| \leq CK\Delta^2 + o(\Delta^2)$$

for some  $C$ ,  $0 < C < \infty$ .

Let  $G_{\delta,\kappa}^*(B)$ ,  $B \in \mathcal{B}$ , be the probability distribution induced by density  $g_{\delta,\kappa}^* \in \mathcal{K}^*$ , with  $\mathcal{B}$  being the Borel  $\sigma$ -algebra. Then the set function

$$w(B) = \begin{cases} \sup \{G^*(B) : G^* \in \mathcal{K}^*\} & \text{if } B \neq \emptyset \\ 0 & \text{if } B = \emptyset \end{cases} \quad (6)$$

is a pseudo-capacity in the sense of Buja [27], i.e., satisfying

- (a)  $w(\emptyset) = 0$ ,  $w(\Omega) = 1$
- (b)  $w(A) \leq w(B)$   $\forall A \subset B$
- (c)  $w(A_n) \uparrow w(A)$   $\forall A_n \uparrow A$
- (d)  $w(A_n) \downarrow w(A)$   $\forall A_n \downarrow A \neq \emptyset$
- (e)  $w(A \cup B) + w(A \cap B) \leq w(A) + w(B)$ .

Analogously, consider a density  $f_0$ , symmetric around 0 and satisfying the assumptions of Theorem 1 as a simple hypothesis. Construct the family  $\mathcal{H}^*$  of densities and the corresponding family of distributions  $\{F_{\delta,\kappa}^*(\cdot), \delta \leq \Delta, \kappa \leq K\}$  similarly as above. Then the set function

$$v(B) = \begin{cases} \sup \{F^*(B) : F^* \in \mathcal{H}^*\} & \text{if } B \neq \emptyset \\ 0 & \text{if } B = \emptyset \end{cases} \quad (7)$$

is a pseudo-capacity in the sense of Buja [27].

Buja [27] showed that on any Polish space exists a (possibly different) topology which generates the same Borel algebra and on which every pseudo-capacity is a 2-alternating capacity in the sense of [25].

Let us now consider the problem of testing the hypothesis  $\mathcal{H} = \{F^* \in \mathcal{H}^* | F^*(\cdot) \leq v(\cdot)\}$  against the alternative  $\mathcal{K} = \{G^* \in \mathcal{K}^* | G^*(\cdot) \leq w(\cdot)\}$ , based on an independent random sample  $Z_1, \dots, Z_n$ . Assume that  $\mathcal{H}^*$  and  $\mathcal{K}^*$  satisfy (5). Then, following [27] and [25], we have the main theorem providing the minimax test of  $\mathcal{H}$  against  $\mathcal{K}$  with significance level  $\alpha \in (0, 1)$ :

**Theorem 3.** *The test*

$$\begin{aligned}\phi(z_1, \dots, z_n) &= \begin{cases} 1 & \text{if } \prod_{i=1}^n \pi(z_i) > \mathcal{C} \\ \gamma & \text{if } \prod_{i=1}^n \pi(z_i) = \mathcal{C} \\ 0 & \text{if } \prod_{i=1}^n \pi(z_i) < \mathcal{C} \end{cases} \\ &\text{where } \pi(\cdot) \text{ is a version of } \frac{dw}{dv}(\cdot) \text{ and } \mathcal{C} \text{ and } \gamma \text{ are chosen so that } E_v \phi(\mathbf{Z}) = \alpha, \text{ is a minimax test of } \mathcal{H} \text{ against } \mathcal{K} \text{ of level } \alpha.\end{aligned}$$

#### 4. Numerical Illustration

We assume to observe independent observations  $Z_{1,\delta}, \dots, Z_{n,\delta}$  for  $i = 1, \dots, n$ , where  $Z_{i,\delta} = X_i + \sqrt{\delta}V_i$  as described in Section 3, where  $X_1, \dots, X_n$  are independent identically distributed (with a distribution function  $F$ ) but unobserved. Let us further denote by  $\Phi$  the distribution function of  $N(0, 1)$  and by  $\Phi_\sigma^*$  the distribution function of  $N(0, \sigma^2)$ . The primary task here is to test  $\mathcal{H}_0 : F \equiv \Phi$  against

$$\mathcal{H}_1 : F(x) = (1 - \lambda)\Phi(x) + \lambda\Phi_\sigma^*(x), \quad x \in R,$$

with a fixed  $\sigma > 1$  and  $\lambda \in (0, 1)$ . We perform all the computations using the R software [28].

To describe our approach to computing the test, we will need the notation for the set of pseudo-distribution functions corresponding to the set of pseudo-densities  $\mathcal{H}^*$  denotes as

$$\tilde{\mathcal{H}}^* = \left\{ F_{\delta, \kappa}^* : F_{\delta, \kappa}^*(z) = \Phi(z) + \frac{\delta}{2}f'_0(z) + \kappa \frac{\delta^2}{24}f_0^{(3)}(z) \mid \delta \leq \Delta, 1 \leq \kappa \leq K \right\},$$

where  $\Phi$  denotes the distribution function of  $N(0, 1)$  distribution. Under the alternative, the set analogous to  $\mathcal{K}^*$  is defined as

$$\tilde{\mathcal{K}}^* = \left\{ G_{\delta, \kappa}^* : G_{\delta, \kappa}^*(z) = G_0(z) + \frac{\delta}{2}g'_0(z) + \kappa \frac{\delta^2}{24}g_0^{(3)}(z) \mid 0 \leq \delta \leq \Delta, 1 \leq \kappa \leq K \right\}.$$

Our task is to approximate

$$v((-\infty, z)) = \sup \{F_{\delta, \kappa}^*(z); F_{\delta, \kappa}^* \in \tilde{\mathcal{H}}^*\}, \quad z \in R, \quad (8)$$

and

$$w((-\infty, z)) = \sup \{G_{\delta, \kappa}^*(z); G_{\delta, \kappa}^* \in \tilde{\mathcal{K}}^*\}, \quad z \in R. \quad (9)$$

Here, the functions  $F_{\delta, \kappa}^*(z)$  and  $G_{\delta, \kappa}^*(z)$  are evaluated over a grid with step 0.05. Then, the maximization in (8) and (9) is performed for values of  $z$  over the grid and over four boundary values of  $(\delta, \kappa)^T$ , which are equal to  $(0, 0)^T$ ,  $(0, K)^T$ ,  $(\Delta, 0)^T$ , and  $(\Delta, K)^T$ . Additional computations with 10 randomly selected pairs of  $(\delta, \kappa)^T$  over  $\delta \in [0, \Delta]$  and  $\kappa \in [0, K]$  revealed that the optimum is attained in one of the boundary values. Further, the Radon-Nikodym derivatives of  $V$  and  $W$  are estimated by a finite difference approximation in order to compute the test statistic.

The test rejects  $\mathcal{H}_0$  if the test statistics  $\prod_{i=1}^n \pi(z_i)$  exceeds a critical value, which (as well as the  $p$ -value) can be approximated by a Monte Carlo simulation, i.e., by a repeated random generating random variables  $X_1, \dots, X_n$  under  $\mathcal{H}_0$ , and we generate them 10,000 times here.

We perform the following particular numerical study. We compute the critical value of the  $\alpha$ -test for  $n = 20$  (or  $n = 40$ ),  $\lambda = 0.25$ ,  $\sigma^2 = 3$ ,  $\Delta = 0.2$ ,  $K = 1.1$ , and  $\alpha = 0.05$ . Further, we are interested in evaluating the probability of rejecting this test for data generated from

$$F(x) = (1 - \tilde{\lambda})\Phi(x) + \tilde{\lambda}\Phi_{\tilde{\sigma}}^*(x), \quad x \in R, \quad (10)$$

with different values of  $\tilde{\lambda}$  and  $\tilde{\sigma}^2$ . Its values are shown in Table 1 (for  $n = 20$ ) and Table 2 (for  $n = 40$ ), which are approximated using (again) 10,000 randomly generated variables from (10). The boldface numbers are equal to the power of the test (under the simple  $H_1$ ). The proposed test seems meaningful, while its power is increased for  $n = 40$  compared to  $n = 20$ ; in addition, the power increases with an increasing  $\tilde{\lambda}$  if  $\tilde{\sigma}^2$  is retained; and the power also increases with an increasing  $\tilde{\sigma}^2$  if  $\tilde{\lambda}$  is retained.

**Table 1.** Probability of rejecting the test in the simulation with  $n = 20$ .

Value of $\tilde{\lambda}$	Value of $\tilde{\sigma}^2$			
	3	4	5	6
0.25	0.39	0.52	0.61	0.67
0.35	0.50	0.67	0.75	0.81
0.45	0.61	0.76	0.85	0.89

**Table 2.** Probability of rejecting the test in the simulation with  $n = 40$ .

Value of $\tilde{\lambda}$	Value of $\tilde{\sigma}^2$			
	3	4	5	6
0.25	0.55	0.73	0.82	0.87
0.35	0.72	0.86	0.93	0.96
0.45	0.82	0.94	0.97	0.99

## 5. Conclusions

The likelihood ratio test of  $f_0$  against  $g_0$  is considered in the situation that observations  $X_i$  are mismeasured due to the presence of measurement errors. Thus instead of  $X_i$  for  $i = 1, \dots, n$ , we observe  $Z_i = X_i + \sqrt{\delta}V_i$  with unobservable parameter  $\delta$  and unobservable random variable  $V_i$ . When we ignore the presence of measurement errors and perform the original test, the probability of type I error becomes different from the nominal value, but the test is still the most powerful among all tests on the modified level.

Under some assumptions on  $f_0$  and  $g_0$  and for  $\delta < \Delta$ ,  $EV^4 \leq K$ , we further construct a minimax likelihood ratio test of some families of distributions of the  $Z_i = X_i + \sqrt{\delta}V_i$ , based on the capacities of the Huber-Strassen type. The test treats the composite null and alternative hypotheses, which cover all possible measurement errors satisfying the assumptions. The advantage of the novel test is that it keeps the probability of type I error below the desired value ( $\alpha = 0.05$ ) across all possible measurement errors. The test is performed in a straightforward way, while the user must specify particular (not excessively large) values of  $\Delta$  and  $K$ . We do not consider this a limiting requirement, because parameters corresponding to the severity of measurement errors are commonly chosen in a similar way in numerous measurement error models [5,23] or robust optimization procedures [29]. The critical value of the test can be approximated by a simulation. The numerical experiment in Section 4 illustrates the principles and performance of the novel test.

**Author Contributions:** Methodology, M.B. and J.J.; Software, J.K.; Writing—Original Draft Preparation, M.B. and J.J.; Writing—Review & Editing, M.B., J.J. and J.K.; Funding Acquisition, M.B.

**Funding:** The research of Jana Jurečková was supported by the Grant 18-01137S of the Czech Science Foundation. The research of Jan Kalina was supported by the Grant 17-01251S of the Czech Science Foundation.

**Acknowledgments:** The authors would like to thank two anonymous referees for constructive advice.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Boyd, A.; Lankford, H.; Loeb, S.; Wyckoff, J. Measuring test measurement error: A general approach. *J. Educ. Behav. Stat.* **2013**, *38*, 629–663. [[CrossRef](#)]
- Brakenhoff, T.B.; Mitroiu, M.; Keogh, R.H.; Moons, K.G.M.; Groenwold, R.H.H.; van Smeden, M. Measurement error is often neglected in medical literature: A systematic review. *J. Clin. Epidemiol.* **2018**, *98*, 89–97. [[CrossRef](#)] [[PubMed](#)]
- Edwards, J.K.; Cole, S.R.; Westreich, D. All your data are always missing: Incorporating bias due to measurement error into the potential outcomes framework. *Int. J. Epidemiol.* **2015**, *44*, 1452–1459. [[CrossRef](#)] [[PubMed](#)]
- Fuller, W.A. *Measurement Error Models*; John Wiley & Sons: New York, NY, USA, 1987.
- Carroll, R.J.; Ruppert, D.; Stefanski, L.A.; Crainiceanu, C.M. *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2006
- Cheng, C.L.; van Ness, J.W. *Statistical Regression with Measurement Error*; Arnold: London, UK, 1999.
- Carroll, R.J.; Maca, J.D.; Ruppert, D. Nonparametric regression in the presence of measurement error. *Biometrika* **1999**, *86*, 541–554. [[CrossRef](#)]
- Carroll, R.J.; Delaigle, A.; Hall, P. Non-parametric regression estimation from data contaminated by a mixture of Berkson and classical errors. *J. R. Stat. Soc. B* **2007**, *69*, 859–878. [[CrossRef](#)]
- Fan, J.; Truong, Y.K. Nonparametric regression estimation involving errors-in-variables. *Ann. Stat.* **1993**, *21*, 23–37. [[CrossRef](#)]
- He, X.; Liang, H. Quantile regression estimate for a class of linear and partially linear errors-in-variables models. *Stat. Sin.* **2000**, *10*, 129–140.
- Arias, O.; Hallock, K.F.; Sosa-Escudero, W. Individual heterogeneity in the returns to schooling: Instrumental variables quantile regression using twins data. *Empir. Econ.* **2001**, *26*, 7–40. [[CrossRef](#)]
- Hyk, W.; Stojek, Z. Quantifying uncertainty of determination by standard additions and serial dilutions methods taking into account standard uncertainties in both axes. *Anal. Chem.* **2013**, *85*, 5933–5939. [[CrossRef](#)]
- Kelly, B.C. Some aspects of measurement error in linear regression of astronomical data. *Astrophys. J.* **2007**, *665*, 1489–1506. [[CrossRef](#)]
- Marques, T.A. Predicting and correcting bias caused by measurement error in line transect sampling using multiplicative error model. *Biometrics* **2004**, *60*, 757–763. [[CrossRef](#)] [[PubMed](#)]
- Rocke, D.M.; Lorenzato, S. A two-component model for measurement error in analytical chemistry. *Technometrics* **1995**, *37*, 176–184. [[CrossRef](#)]
- Akritis, M.G.; Bershady, M.A. Linear regression for astronomical data with measurement errors and intrinsic scatter. *Astrophys. J.* **1996**, *470*, 706–728. [[CrossRef](#)]
- Hausman, J. Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *J. Econ. Perspect.* **2001**, *15*, 57–67. [[CrossRef](#)]
- Hyslop, D.R.; Imbens, Q.W. Bias from classical and other forms of measurement error. *J. Bus. Econ. Stat.* **2001**, *19*, 475–481. [[CrossRef](#)]
- Jurečková, J.; Picek, J.; Saleh, A.K.M.E. Rank tests and regression rank scores tests in measurement error models. *Comput. Stat. Data Anal.* **2010**, *54*, 3108–3120. [[CrossRef](#)]
- Jurečková, J.; Koul, H.L.; Navrátil, R.; Picek, J. Behavior of R-estimators under Measurement Errors. *Bernoulli* **2016**, *22*, 1093–1112. [[CrossRef](#)]
- Navrátil, R.; Saleh, A.K.M.E. Rank tests of symmetry and R-estimation of location parameter under measurement errors. *Acta Univ. Palacki. Olomuc. Fac. Rerum Nat. Math.* **2011**, *50*, 95–102.
- Navrátil, R. Rank tests and R-estimates in location model with measurement errors. In *Proceedings of Workshop of the Jaroslav Hájek Center and Financial Mathematics in Practice I*; Masaryk University: Brno, Czech Republic, 2012; pp. 37–44.
- Saleh, A.K.M.E.; Picek, J.; Kalina, J. R-estimation of the parameters of a multiple regression model with measurement errors. *Metrika* **2012**, *75*, 311–328. [[CrossRef](#)]
- Sen, P.K.; Jurečková, J.; Picek, J. Rank tests for corrupted linear models. *J. Indian Stat. Assoc.* **2013**, *51*, 201–230.

25. Huber, P.; Strassen, V. Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Stat.* **1973**, *2*, 251–273. [[CrossRef](#)]
26. Ibragimov, I.A. On the composition of unimodal distributions. *Theor. Probab. Appl.* **1956**, *1*, 255–260. [[CrossRef](#)]
27. Buja, A. On the Huber-Strassen theorem. *Probab. Theory Relat. Fields* **1986**, *73*, 149–152. [[CrossRef](#)]
28. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017. Available online: <https://www.R-project.org/> (accessed on 15 September 2018).
29. Xanthopoulos, P.; Pardalos, P.M.; Trafalis, T.B. *Robust Data Mining*; Springer: New York, NY, USA, 2013.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Minimum Penalized $\phi$ -Divergence Estimation under Model Misspecification

M. Virtudes Alba-Fernández <sup>1,\*</sup>, M. Dolores Jiménez-Gamero <sup>2</sup> and F. Javier Ariza-López <sup>3</sup>

<sup>1</sup> Departamento de Estadística e Investigación Operativa, Universidad de Jaén, 23071, Jaén, Spain

<sup>2</sup> Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, 41012, Sevilla, Spain; dolores@us.es

<sup>3</sup> Departamento de Ingeniería Cartográfica, Geodésica y Fotogrametría, Universidad de Jaén, 23071, Jaén, Spain; fjariza@ujaen.es

\* Correspondence: mvalba@ujaen.es; Tel.: +34-953212142

Received: 8 March 2018; Accepted: 27 April 2018; Published: 30 April 2018



**Abstract:** This paper focuses on the consequences of assuming a wrong model for multinomial data when using minimum penalized  $\phi$ -divergence, also known as minimum penalized disparity estimators, to estimate the model parameters. These estimators are shown to converge to a well-defined limit. An application of the results obtained shows that a parametric bootstrap consistently estimates the null distribution of a certain class of test statistics for model misspecification detection. An illustrative application to the accuracy assessment of the thematic quality in a global land cover map is included.

**Keywords:** minimum penalized  $\phi$ -divergence estimator; consistency; asymptotic normality; goodness-of-fit; bootstrap distribution estimator; thematic quality assessment

---

## 1. Introduction

In many practical settings, individuals are classified into a finite number of unique nonoverlapping categories, and the experimenter collects the number of observations falling in each of such categories. In statistics, that sort data is called multinomial data. Examples arise in many scientific disciplines: in economics, when dealing with the number of different types of industries observed in a geographical area; in biology, when counting the number of individuals belonging to one of  $k$  species (see, for example, Pardo [1], pp. 94–95); in sports, when considering the number of injured players in soccer matches (see, for example, Pardo [1], p. 146); and many others.

When dealing with multinomial data, one often finds zero cell frequencies, even for large samples. Although many examples can be given, we will center on the following one, since two related data sets will be analyzed in Section 4. Zero cell frequencies are usually observed when the quality of the geographic information data is assessed, and specifically, when we pay attention to the thematic component of this quality. Roughly speaking, the thematic quality refers to the correctness of the qualitative aspect of an element (pixel, feature, etc.). To give an assessment of the thematic accuracy, a comparison is needed between the label considered as true of a feature and the label assigned to the same feature after a classification (among a number of labels previously stated). This way, each element/feature, which really belongs to a particular category, can be classified as belonging to the same category (correct assignment), or as belonging to another one (incorrect assignment). Given a sample of  $n$  elements belonging to a particular category, after collecting the number of elements correctly classified,  $X_1$ , and the number of incorrect classifications in a set of  $k - 1$  possible categories,  $X_i$ ,  $i = 2, \dots, k$ , we obtain a multinomial vector  $(X_1, X_2, \dots, X_k)^t$ , for which small or zero cell frequencies are often observed associated with the incorrect classifications,  $X_i$ ,  $i = 2, \dots, k$ .

Motivated by this example in the geographic information data context, as well as many others, along this paper, it will be assumed that the available information can be summarized by means of a random vector  $X = (X_1, \dots, X_k)^t$  having a  $k$ -cell multinomial distribution with parameters  $n$  and  $\pi = (\pi_1, \dots, \pi_k)^t \in \Delta_{0k} = \{(\pi_1, \dots, \pi_k)^t : \pi_i \geq 0, 1 \leq i \leq k, \sum_{i=1}^k \pi_i = 1\}$ ,  $X \sim \mathcal{M}_k(n; \pi)$  in short. Notice that, if  $\pi \in \Delta_{0k}$ , then some components of  $\pi$  may equal 0, implying that some cell frequencies can be equal to zero, even for large samples. In many instances, it is assumed that  $\pi$  belongs to a parametric family  $\pi \in \mathcal{P} = \{P(\theta) = (p_1(\theta), \dots, p_k(\theta))^t, \theta \in \Theta\} \subset \Delta_k = \{(\pi_1, \dots, \pi_k)^t : \pi_i > 0, 1 \leq i \leq k, \sum_{i=1}^k \pi_i = 1\}$ , where  $\Theta \subseteq \mathbb{R}^s$ ,  $k - s - 1 > 0$  and  $p_1(\cdot), \dots, p_k(\cdot)$  are known real functions.

When it is assumed that  $\pi \in \mathcal{P}$ ,  $\pi$  is usually estimated through  $P(\hat{\theta}) = (p_1(\hat{\theta}), \dots, p_k(\hat{\theta}))^t$  for some estimator  $\hat{\theta}$  of  $\theta$ . A common choice for  $\hat{\theta}$  is the maximum likelihood estimator (MLE), which is known to have good asymptotic properties. Basu and Sarkar [2] and Morales et al. [3] have shown that these properties are shared by a larger class of estimators: the minimum  $\phi$ -divergence estimators (M $\phi$ E). This class includes MLEs as a particular case. However, as illustrated in Mandal et al. [4], the finite sample performance of these estimators can be improved by modifying the weight that each  $\phi$ -divergence assigns to the empty cells. The resulting estimator is called the minimum penalized  $\phi$ -divergence estimator (MP $\phi$ E). Moreover, Mandal et al. [4] have shown that such estimators have the same asymptotic properties as the M $\phi$ E. Specifically, they are strongly consistent and, conveniently normalized, asymptotically normal. To derive these asymptotic properties, it is assumed that the probability model is correctly specified, that is to say, that we are sure about  $\pi \in \mathcal{P}$ .

If the parametric model is not correctly specified, Jiménez-Gamero et al. [5] have shown that, under certain assumptions, the M $\phi$ E still have a well defined limit, and, conveniently normalized, they are asymptotically normal. For the MLE, these results were known from those in [6]. Because, as argued before, the use of penalized  $\phi$ -divergences may lead to better performance of the resulting estimators, the aim of this piece of research is to investigate the asymptotic properties of the MP $\phi$ E under model misspecification. If the model considered is true, we obtain as a particular case the results in [4].

The usefulness of the results obtained is illustrated by applying them to the problem of testing goodness-of-fit to the parametric family  $\mathcal{P}$ ,

$$H_0 : \pi \in \mathcal{P},$$

against the alternative

$$H_1 : \pi \notin \mathcal{P},$$

using as a test statistic a penalized  $\phi_1$ -divergence between a nonparametric estimator of  $\pi$ , the relative frequencies, and a parametric estimator of  $\pi$ , obtained by assuming that the null hypothesis is true,  $P(\hat{\theta})$ ,  $\hat{\theta}$  being an MP $\phi_2$ E. Here,  $\phi_1$  and  $\phi_2$  may differ. The convenience of using this type of test statistics is justified in Mandal et al. [7]. Although these authors show that, under  $H_0$ , such test statistics are asymptotically distribution free, the asymptotic approximation to the null distribution of the test statistics in this class is rather poor. Some numerical examples illustrate this unsatisfactory behavior of the asymptotic approximation. By using the fact that the MP $\phi$ E always converges to a well-defined limit, whether the model in  $H_0$  is true or not, we prove that the bootstrap consistently estimates the null distribution of these test statistics. We then retake the previously cited numerical examples to exemplify the usefulness of the bootstrap approximation which, despite the demand for more computing time, is more accurate than that yielded by the asymptotic null distribution for small and moderate sample sizes.

The rest of the paper is organized as follows. Section 2 studies certain asymptotic properties of MP $\phi_2$ E; specifically, conditions are given for the strong consistency and asymptotic normality. Section 3 uses such results to prove that a parametric bootstrap provides a consistent estimator to the null distribution of test statistics based on penalized  $\phi$ -divergences for testing  $H_0$ . Section 4 displays an application of the results obtained in the context of a classification work in a cover land map.

Before ending this section we introduce some notation: all limits in this paper are taken when  $n \rightarrow \infty$ ;  $\xrightarrow{\mathcal{L}}$  denotes convergence in distribution;  $\xrightarrow{P}$  denotes convergence in probability;  $\xrightarrow{a.s.}$  denotes the almost sure convergence; let  $\{A_n\}$  be a sequence of random variables and let  $\epsilon \in \mathbb{R}$ , then  $A_n = O_p(n^{-\epsilon})$  means that  $n^\epsilon A_n$  is bounded in probability,  $A_n = o_p(n^{-\epsilon})$  means that  $n^\epsilon A_n \xrightarrow{P} 0$ , and  $A_n = o(n^{-\epsilon})$  means that  $n^\epsilon A_n \xrightarrow{a.s.} 0$ ;  $N_k(\mu, \Sigma)$  denotes the  $k$ -variate normal law with mean  $\mu$  and variance matrix  $\Sigma$ ; all vectors are column vectors; the superscript  $t$  denotes transpose; if  $x \in \mathbb{R}^k$ , with  $x^t = (x_1, \dots, x_k)$ , then  $Diag(x)$  is the  $k \times k$  diagonal matrix whose  $(i, i)$  entry is  $x_i$ ,  $1 \leq i \leq k$ , and

$$\Sigma_x = Diag(x) - xx^t;$$

$I_k$  denotes the  $k \times k$  identity matrix; to simplify notation, all 0s appearing in the paper represent vectors of the appropriate dimension.

## 2. Some Asymptotic Properties of MP $\phi$ E

Let  $X \sim \mathcal{M}_k(n; \pi)$ , with  $\pi \in \Delta_{0k}$ , and let  $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_k)^t$  be the vector of relative frequencies,

$$\hat{\pi}_i = \frac{X_i}{n}, \quad 1 \leq i \leq k. \quad (1)$$

Let  $\mathcal{P}$  be a parametric model satisfying Assumption 1 below.

**Assumption 1.**  $\mathcal{P} = \{P(\theta) = (p_1(\theta), \dots, p_k(\theta))^t, \theta \in \Theta\} \subset \Delta_k$ , where  $\Theta \subseteq \mathbb{R}^s$ ,  $k - s - 1 > 0$  and  $p_1(\cdot), \dots, p_k(\cdot) : \Theta \rightarrow \mathbb{R}$  are known twice continuously differentiable in  $\text{int}\Theta$  functions.

Let  $\phi : [0, \infty) \rightarrow \mathbb{R} \cup \{\infty\}$  be a continuous convex function. For arbitrary  $Q = (q_1, \dots, q_k)^t \in \Delta_{0k}$  and  $P = (p_1, \dots, p_k)^t \in \Delta_k$ , the  $\phi$ -divergence between  $Q$  and  $P$  is defined by (Csiszár [8])

$$D_\phi(Q, P) = \sum_{i=1}^k p_i \phi(q_i / p_i).$$

Note that

$$D_\phi(Q, P) = \sum_{i/q_i > 0} p_i \phi(q_i / p_i) + \phi(0) \sum_{i/q_i = 0} p_i.$$

The penalized  $\phi$ -divergence for the tuning parameter  $h$  between  $Q$  and  $P$  is defined from the above expression by replacing  $\phi(0)$  with  $h$  as follows (see Mandal et al. [4]):

$$D_{\phi,h}(Q, P) = \sum_{i/q_i > 0} p_i \phi(q_i / p_i) + h \sum_{i/q_i = 0} p_i.$$

If

$$\hat{\theta}_{\phi,h} = \arg \min_{\theta} D_{\phi,h}(\hat{\pi}, P(\theta)),$$

then  $\hat{\theta}_{\phi,h}$  is called the MP $\phi$ E of  $\theta$ .

In order to study some of the properties of  $\hat{\theta}_{\phi,h}$ , we will assume that  $\phi$  satisfies Assumption 2 below.

**Assumption 2.**  $\phi : [0, \infty) \rightarrow \mathbb{R}$  is a strictly convex function, twice continuously differentiable in  $(0, \infty)$ .

Assumption 2 is assumed when dealing with estimators based on minimum divergence, since it lets us take Taylor series expansions of  $D_\phi(\hat{\pi}, P(\theta))$ , which is useful to derive asymptotic properties of the M $\phi$ E. For example, Section 3 of Lindsay [9] assumes that the function  $\phi$  (he calls  $G$  what we call  $\phi$ ) is a thrice differentiable function (which is stronger than Assumption 2); Theorem 3 in Morales et al. [3]

requires, among other conditions,  $\phi$  to meet Assumption 2 to derive the consistency and asymptotic normality of M $\phi$ E's.

Assumption 2 is also assumed in Mandal et al. [4] (they call  $G$  what we call  $\phi$ ) to study the consistency and asymptotic normality of MP $\phi$ E's. Specifically, these authors show that, if  $\pi \in \mathcal{P}$  and  $\theta_0$  is the true parameter value, then, under suitable regularity conditions including Assumption 2, the MP $\phi$ E is consistent for  $\theta_0$ , and  $\sqrt{n}(\hat{\theta}_{\phi,h} - \theta_0)$  is asymptotically normal with a mean of 0 and a variance matrix equal to the inverse of the information matrix.

Next we will only assume that  $\pi \in \Delta_{0k}$ , that is, the assumption that  $\pi \in \mathcal{P}$  is dropped. In this context, we prove that the MP $\phi$ E is consistent for  $\theta_0$ , where now  $\theta_0$  is the parameter vector that minimizes  $D_{\phi,h}(\pi, P(\theta))$ , that is to say,  $\theta_0 = \arg \min_{\theta} D_{\phi,h}(\pi, P(\theta))$ . Note that  $\theta_0$  also depends on  $\phi$  and  $h$ , so to be rigorous we should denote it by  $\theta_{0,\phi,h}$ , but to simplify notation we will simply denote it as  $\theta_0$ . We also show that  $\sqrt{n}(\hat{\theta}_{\phi,h} - \theta_0)$  is asymptotically normal with a mean of 0. With this aim, we will also assume the following.

**Assumption 3.**  $D_{\phi,h}(\pi, P(\theta))$  has a unique minimum at  $\theta_0 \in \text{int}\Theta$ .

Assumption 3 is assumed in papers on estimators based on minimum divergence estimation. For example, it is Assumption A3(b) in [6], which states, that it is the fundamental identification condition for quasi-maximum likelihood estimators to have a well-defined limit; and it is contained in Assumptions 7 and 9 in [10], required for minimum chi-square estimators to have a well-defined limit; it also coincides with Assumption 30 in [9], imposed for the same reason.

Let  $\theta_0$  be as defined in Assumption 3. Then  $P(\theta_0)$  is the  $(\phi, h)$ -projection of  $\pi$  on  $\mathcal{P}$ . Section 3 in [11] shows that Assumption 3 holds for two-way tables when  $\mathcal{P}$  is the uniform association model, so the  $(\phi, h)$ -projection always exists for such model. Nevertheless, this projection may not exist, or may not be defined uniquely. See Example 2 in [12] for an instance where there is no unique minimum (because although  $\Theta$  is that example is convex, the family  $\{P(\theta), \theta \in \Theta\}$  is not convex, so the uniqueness of the projection is not guaranteed). Let  $\Delta_k(\phi, \mathcal{P}, h) = \{\pi \in \Delta_{0k} \text{ such that Assumption 3 holds}\}$ .

From now on, we will assume that the components of  $\pi$  are sorted so that  $\pi_1, \dots, \pi_m > 0$ , and  $\pi_{m+1} = \dots = \pi_k = 0$ , for some  $1 < m \leq k$ , where, if  $m = k$ , then it is understood that all components of  $\pi$  are positive. We will write  $\pi^+ = (\pi_1, \dots, \pi_m)^t$  and  $\hat{\pi}^+ = (\hat{\pi}_1, \dots, \hat{\pi}_m)^t$ . The next result shows the strong consistency and asymptotic normality of the MP $\phi$ E.

**Theorem 1.** Let  $\mathcal{P}$  be a parametric family satisfying Assumption 1. Let  $\phi$  be a real function satisfying Assumption 2. Let  $X \sim \mathcal{M}_k(n; \pi)$  with  $\pi \in \Delta_k(\phi, \mathcal{P}, h)$ . Then

- (a)  $\hat{\theta}_{\phi,h} \xrightarrow{a.s.} \theta_0$ .
  - (b)  $\sqrt{n} \begin{pmatrix} \hat{\pi}^+ - \pi^+ \\ \hat{\theta}_{\phi,h} - \theta_0 \end{pmatrix} \xrightarrow{\mathcal{L}} N_{m+s}(0, A\Sigma_{\pi^+} A^t)$ , where  $A^t = (I_m, G^t)$  and  $G$  is defined in Equation (7).
- In particular,

$$\sqrt{n}(\hat{\theta}_{\phi,h} - \theta_0) \xrightarrow{\mathcal{L}} N_s(0, G\Sigma_{\pi^+} G^t) \quad (2)$$

- (c)  $\sqrt{n} \begin{pmatrix} \hat{\pi}^+ - \pi^+ \\ P(\hat{\theta}_{\phi,h}) - P(\theta_0) \end{pmatrix} \xrightarrow{\mathcal{L}} N_{2m}(0, B\Sigma_{\pi^+} B^t)$ , where  $B^t = (I_m, G^t D_1(P(\theta_0)))$ , with  $D_1(P(\theta))$  defined in Equation (8).

**Remark 1.** Observe that, if  $m = k$ , then the penalization has no effect asymptotically; by contrast, if  $m < k$ , then the presence of the tuning parameter  $h$  influences the covariance matrix of the asymptotic law of  $\sqrt{n}(\hat{\theta}_{\phi,h} - \theta_0)$  and  $\sqrt{n}(P(\hat{\theta}_{\phi,h}) - P(\theta_0))$ .

**Remark 2.** If  $\pi \in \mathcal{P}$ , we obtain as a particular case the results in Mandal et al. [4]. Our conditions are weaker than those in [4]. The reason is that they allow an infinite number of categories, while we are assuming that such a number is finite,  $k$ . Therefore, when the number of categories is finite, the assumptions in [4] for the consistency and asymptotic normality of the MP $\phi$ E can be weakened.

As a consequence of Theorem 1, the following corollary gives the asymptotic behavior of  $D_{\phi_1, h_1}(\hat{\pi}, P(\hat{\theta}_{\phi_2, h_2}))$ , for arbitrary  $\phi_1, \phi_2$ , and  $h_1, h_2$ , that may or may not coincide. Part (a) of Corollary 1, which assumes that the model  $\mathcal{P}$  is correctly specified, has been previously proven in [7]. It is included here for the sake of completeness. Part (b), which describes the limit in law under alternatives is, to the best of our knowledge, new.

**Corollary 1.** Let  $\mathcal{P}$  be a parametric family satisfying Assumption 1. Let  $\phi_1$  and  $\phi_2$  be two real functions satisfying Assumption 2. Let  $X \sim \mathcal{M}_k(n; \pi)$  with  $\pi \in \Delta_k(\phi, \mathcal{P}, h)$ .

(a) For  $\pi \in \mathcal{P}$ ,

$$T = \frac{2n}{\phi_1''(1)} \{D_{\phi_1, h_1}(\hat{\pi}, P(\hat{\theta}_{\phi_2, h_2})) - \phi_1(1)\} \xrightarrow{\mathcal{L}} \chi_{k-s-1}^2.$$

(b) For  $\pi \in \Delta_k(\phi_2, \mathcal{P}, h_2) - \mathcal{P}$ , let  $\theta_0 = \arg \min_\theta D_{\phi_2, h_2}(\pi, P(\theta))$ . Then

$$W = \sqrt{n} \{D_{\phi_1, h_1}(\hat{\pi}, P(\hat{\theta}_{\phi_2, h_2})) - D_{\phi_1, h_1}(\pi, P(\theta_0))\} \xrightarrow{\mathcal{L}} N(0, \varrho^2)$$

where  $\varrho^2 = a^t B \Sigma_\pi B^t a$ , with  $B$ , as defined in Theorem 1 with  $\phi = \phi_2$  and  $h = h_2$ ,

$$a^t = \left( \phi_1' \left( \frac{\pi_1}{p_1(\theta_0)} \right), \dots, \phi_1' \left( \frac{\pi_m}{p_m(\theta_0)} \right), v_1, \dots, v_m, \underbrace{h_1, \dots, h_1}_{k-m \text{ times}} \right),$$

and  $v_i$ ,  $1 \leq i \leq m$ , are as defined in Equation (5) with  $\phi = \phi_1$  and  $h = h_1$ .

**Remark 3.** If  $\pi \in \mathcal{P}$ , the asymptotic behavior of the statistic  $T$  does not depend either on  $\phi_1, \phi_2$ , or on  $h_1, h_2$ . In fact, the asymptotic law of  $T$  is the same as if non-penalized divergences were used.

**Remark 4.** When  $\pi \in \Delta_k(\phi_2, \mathcal{P}, h_2) - \mathcal{P}$ , if  $m = k$ , then the asymptotic distribution of  $W$  does not depend on  $h_1, h_2$ ; by contrast, if  $m < k$ , then the asymptotic distribution of  $W$  does depend on  $h_1$  and  $h_2$ .

**Remark 5.** (Properties of the asymptotic test) As a consequence of Corollary 1(a), we have that for testing  $H_0$  vs.  $H_1$ , the test that rejects the null hypothesis when  $T \geq \chi_{k-s-1, 1-\alpha}^2$  is asymptotically correct, in the sense that  $P_0(T \geq \chi_{k-s-1, 1-\alpha}^2) \rightarrow \alpha$ , where  $\chi_{k-s-1, 1-\alpha}^2$  stands for the  $1 - \alpha$  percentile of the  $\chi_{k-s-1}^2$  distribution and  $P_0$  stands for the probability when the null hypothesis is true. From Corollary 1(b), it follows that such a test is consistent against fixed alternatives  $\pi \in \Delta_k(\phi_2, \mathcal{P}, h_2) - \mathcal{P}$ , in the sense that  $P(T \geq \chi_{k-s-1, 1-\alpha}^2) \rightarrow 1$ .

### 3. Application to Bootstrapping Goodness-Of-Fit Tests

As observed in Remark 5, the test that rejects  $H_0$  when  $T \geq \chi_{k-s-1, 1-\alpha}^2$  is asymptotically correct and consistent against fixed alternatives. Nevertheless, the  $\chi^2$  approximation to the null distribution of the test statistic is rather poor. Next we illustrate this fact with three examples. The last one is motivated by a real data set application in Section 4. All computations have been performed using programs written in the R language [13].

**Example 1.** Let  $X \sim \mathcal{M}_3(n; \pi)$ , with  $\pi \in \mathcal{P}$  so that

$$p_1(\theta) = \frac{1}{3} - \theta, \quad p_2(\theta) = \frac{2}{3} - \theta, \quad p_3(\theta) = 2\theta, \quad 0 < \theta < 1/3.$$

The problem of testing goodness-of-fit to this family is dealt with by considering as test statistic a penalized  $\phi_1$ -divergence and an MP $\phi_2$ E, with  $\phi_1$  and  $\phi_2$ , two members of the power-divergence family, defined as follows:

$$PD_\lambda(x) = \frac{1}{\lambda(\lambda+1)} \left( x^{(\lambda+1)} - x - \lambda(x-1) \right), \lambda \neq 0, -1,$$

$PD_0(x) = x \log(x) - x + 1$ , for  $\lambda = 0$ , and  $PD_{-1}(x) = -\log(x) + x - 1$ , for  $\lambda = -1$ . We thank an anonymous referee for pointing out that the power divergence family is also known as the  $\alpha$ -divergence family (see, for example, Section 4 of Amari [14]).

In order to evaluate the performance of the  $\chi^2$  approximation to the null distribution of  $T$ , we carried out an extensive simulation experiment. As a previous part of the simulation experiment, we evaluated the possible effect of the tuning parameter  $h_2$  on the accuracy of the MP $\phi_2$ E. For this goal, we generated 10,000 samples of size 200 from the parametric family with  $\theta = 0.3333$ , and calculated the MP $\phi_2$ E with  $h_2 = 0.5, 1, 2, 5, 10$  and  $\phi_2 = PD_{-2}$ , which correspond to the modified chi-square test statistic (see, for example, [1], p. 114). We calculated the root mean square deviation (RMSD) of the resulting estimations,

$$RMSD = \sqrt{\frac{\sum_{i=1}^{10,000} (\hat{\theta}_{-2,h_2} - \theta)^2}{10,000}},$$

obtaining 0.00156, 0.00128, 0.00128, 0.00128, and 0.00128, respectively. According to these results, there are rather small differences in the performance of the MP $\phi_2$ E for the values of  $h_2$  considered. Because of this, we fixed  $\phi_2 = PD_{-2}$  and  $h_2 = 0.5, 1, 2$ .

Next, to study the goodness of the asymptotic approximation, we generated 10,000 samples of size  $n = 100$  from the parametric family with  $\theta = 0.3333$ , and calculated the test statistic  $T$  with  $h_1 = h_2 = 0.5$  and  $\phi_1(x) = \phi_2(x) = PD_{-2}(x)$ , as well as the associated  $p$ -values corresponding to the asymptotic null distribution. We then computed the fraction of these  $p$ -values, which are less than or equal to the nominal values  $\alpha = 0.05, 0.10$  (top and below in tables). This experiment was repeated for  $n = 150, 200, h_1 = h_2 = 1, 2, \phi_1 = PD_1$  (which corresponds to the chi-square test statistic) and  $\phi_1 = PD_2$ . Table 1 shows the results obtained. We also considered the case  $h_1 \neq h_2$ , obtaining quite close outcomes. Table 2 displays the results obtained for  $n = 200$  and  $\phi_1 = \phi_2 = PD_{-2}$ . Looking at these tables, we conclude that the asymptotic null distribution does not provide an accurate estimation of the null distribution of  $T$  since the type I error probabilities are much greater than the nominal values, 0.05 and 0.10. Therefore, other approximations of the null distribution should be studied.

**Table 1.** Type I error probabilities obtained using asymptotic approximation for Example 1 with  $\theta = 0.3333, \phi_1 = PD_\lambda, \lambda \in \{-2, 1, 2\}, \phi_2 = PD_{-2}$ , and  $h_1 = h_2 \in \{0.5, 1, 2\}$ .

$\phi_1 = PD_{-2}$			$\phi_1 = PD_1$			$\phi_1 = PD_2$			
$h_1 = h_2$			$h_1 = h_2$			$h_1 = h_2$			
$n$	0.5	1	2	0.5	1	2	0.5	1	2
100	0.996	0.996	0.998	0.995	0.997	0.996	0.995	0.997	0.997
	0.996	0.996	0.998	0.995	0.997	0.996	0.995	0.997	0.997
150	0.995	0.995	0.996	0.994	0.995	0.996	0.994	0.994	0.995
	0.995	0.995	0.996	0.994	0.995	0.996	0.994	0.994	0.995
200	0.992	0.993	0.994	0.992	0.994	0.991	0.993	0.993	0.994
	0.992	0.994	0.994	0.992	0.994	0.991	0.993	0.993	0.994

**Table 2.** Type I error probabilities obtained using asymptotic approximation for Example 1 with  $n = 200, \theta = 0.3333, \phi_1 = \phi_2 = PD_{-2}, h_1 \neq h_2$ , and  $h_1, h_2 \in \{0.5, 1, 2\}$ .

$(h_1, h_2)$	(0.5, 1)	(1, 0.5)	(0.5, 2)	(2, 0.5)	(1, 2)	(2, 1)
	0.989	0.997	0.998	0.998	0.994	0.998
	0.999	0.997	0.998	0.998	0.994	0.999

**Example 2.** Let  $X \sim \mathcal{M}_3(n; \pi)$ , with  $\pi \in \mathcal{P}$  so that

$$p_1(\theta) = 0.5 - 2\theta, \quad p_2(\theta) = 0.5 + \theta, \quad p_3(\theta) = \theta, \quad 0 < \theta < 1/4.$$

We repeated the simulation schedule described in Example 1 for this law with  $\theta = 0.24$ . Tables 3 and 4 report the obtained results. In contrast to the results for Example 1, where the asymptotic approximation gives a rather liberal test, in this case the resulting test is very conservative. Therefore, we again conclude that the asymptotic null distribution does not provide an accurate estimation of the null distribution of  $T$ .

**Table 3.** Type I error probabilities obtained using asymptotic approximation for Example 2 with  $\theta = 0.24$ ,  $\phi_1 = PD_\lambda$ ,  $\lambda \in \{-2, 1, 2\}$ ,  $\phi_2 = PD_{-2}$ , and  $h_1 = h_2 \in \{0.5, 1, 2\}$ .

$n$	$\phi_1 = PD_{-2}$			$\phi_1 = PD_1$			$\phi_1 = PD_2$		
	$h_1 = h_2$			$h_1 = h_2$			$h_1 = h_2$		
	0.5	1	2	0.5	1	2	0.5	1	2
100	0.016	0.017	0.017	0.013	0.013	0.014	0.013	0.014	0.015
	0.034	0.036	0.036	0.031	0.030	0.031	0.030	0.033	0.033
150	0.018	0.019	0.017	0.014	0.014	0.014	0.013	0.015	0.016
	0.035	0.039	0.037	0.031	0.033	0.032	0.035	0.033	0.032
200	0.024	0.022	0.022	0.014	0.016	0.016	0.014	0.015	0.016
	0.043	0.042	0.040	0.032	0.034	0.032	0.032	0.035	0.033

**Table 4.** Type I error probabilities obtained using asymptotic approximation for Example 2 with  $n = 200$ ,  $\theta = 0.24$ ,  $\phi_1 = \phi_2 = PD_{-2}$ ,  $h_1 \neq h_2$ , and  $h_1, h_2 \in \{0.5, 1, 2\}$ .

$(h_1, h_2)$	(0.5, 1)	(1, 0.5)	(0.5, 2)	(2, 0.5)	(1, 2)	(2, 1)
	0.017	0.017	0.018	0.019	0.018	0.016
	0.035	0.033	0.035	0.040	0.036	0.034

**Example 3.** Let  $X \sim \mathcal{M}_4(n; \pi)$ , with  $\pi \in \mathcal{P}$  so that

$$p_1(\theta) = \theta^2, \quad p_2(\theta) = \theta(1 - \theta), \quad p_3(\theta) = \theta(1 - \theta), \quad p_4(\theta) = (1 - \theta)^2, \quad 0 < \theta < 1. \quad (3)$$

We repeated the simulation schedule described in Example 1 for this law with  $\theta = 0.8$ . Tables 5 and 6 report the results obtained. Looking at these tables, we see that the test based on asymptotic approximation is liberal, and conclude, as in the previous examples, that other approximations of the null distribution should be considered.

**Table 5.** Type I error probabilities obtained using asymptotic approximation for Example 3 with  $\theta = 0.8$ ,  $\phi_1 = PD_\lambda$ ,  $\lambda \in \{-2, 1, 2\}$ ,  $\phi_2 = PD_{-2}$ , and  $h_1 = h_2 \in \{0.5, 1, 2\}$ .

$n$	$\phi_1 = PD_{-2}$			$\phi_1 = PD_1$			$\phi_1 = PD_2$		
	$h_1 = h_2$			$h_1 = h_2$			$h_1 = h_2$		
	0.5	1	2	0.5	1	2	0.5	1	2
100	0.063	0.066	0.074	0.095	0.107	0.111	0.122	0.136	0.131
	0.122	0.120	0.125	0.157	0.165	0.161	0.181	0.190	0.182
150	0.063	0.064	0.066	0.083	0.082	0.084	0.099	0.105	0.100
	0.114	0.118	0.113	0.137	0.134	0.136	0.153	0.159	0.152
200	0.062	0.061	0.061	0.075	0.079	0.074	0.086	0.091	0.086
	0.111	0.111	0.115	0.129	0.137	0.123	0.145	0.148	0.144

**Table 6.** Type I error probabilities obtained using asymptotic approximation for Example 3 with  $n = 200$ ,  $\theta = 0.8$ ,  $\phi_1 = \phi_2 = PD_{-2}$ ,  $h_1 \neq h_2$ , and  $h_1, h_2 \in \{0.5, 1, 2\}$ .

$(h_1, h_2)$	(0.5, 1)	(1, 0.5)	(0.5, 2)	(2, 0.5)	(1, 2)	(2, 1)
	0.060	0.062	0.063	0.062	0.063	0.058
	0.108	0.114	0.113	0.112	0.113	0.109

The reason for the unsatisfactory results in the three examples is that the asymptotic approximation requires unaffordably large sample sizes when some cells have extremely small probabilities, which provoke the presence of zero cell frequencies. To appreciate this fact, notice that Example 1 requires  $n > 30,000$  to obtain expected cell frequencies greater than 10.

Motivated by these examples, the aim of this section is to study another way of approximating the null distribution of  $T$ , the bootstrap. The null bootstrap distribution of  $T$  is the conditional distribution of

$$T^* = \frac{2n}{\phi_1''(1)} \{D_{\phi_1, h_1}(\hat{\pi}^*, P(\hat{\theta}_{\phi_2, h_2}^*)) - \phi_1(1)\},$$

given  $(X_1, \dots, X_k)$ , where  $\hat{\pi}^*$  is defined as  $\hat{\pi}$  with  $(X_1, \dots, X_k)$  replaced by  $(X_1^*, \dots, X_k^*) \sim \mathcal{M}_k(n; P(\hat{\theta}_{\phi_2, h_2}))$ , and  $\hat{\theta}_{\phi_2, h_2}^* = \arg \min_{\theta} D_{\phi_2, h_2}(\hat{\pi}^*, P(\theta))$ .

Let  $P_*$  denote the bootstrap conditional probability law, given  $(X_1, \dots, X_k)$ . The next theorem gives the weak limit of  $T^*$ .

**Theorem 2.** Let  $\mathcal{P}$  be a parametric family satisfying Assumption 1. Let  $\phi_1$  and  $\phi_2$  be two real functions satisfying Assumption 2. Let  $X \sim \mathcal{M}_k(n; \pi)$  with  $\pi \in \Delta_k(\phi, \mathcal{P}, h)$ . Then

$$\sup_x |P_*(T^* \leq x) - P(Y \leq x)| \xrightarrow{P} 0$$

where  $Y \sim \chi_{k-s-1}^2$ .

Recall that, from Corollary 1(a), when  $H_0$  is true, the test statistic  $T$  converges in law to a  $\chi_{k-s-1}^2$  law. Thus, the result in Theorem 2 implies the consistency of the null bootstrap distribution of  $T$  as an estimator of the null distribution of  $T$ . It is important to remark that the result in Theorem 2 holds whether  $H_0$  is true or not, that is, the bootstrap properly estimates the null distribution, even if the available data does not obey the law in the null hypothesis. This is due to the fact that, under the assumed conditions, the MPfE always converges to a well-defined limit.

**Remark 6.** Properties of the Bootstrap Test. Similarly to Remark 5, as a consequence of Corollary 1(a) and Theorem 2, we have that, for testing  $H_0$  vs.  $H_1$ , the test that rejects the null hypothesis when  $T \geq T_{1-\alpha}^*$  is asymptotically correct, in the sense that  $P_0(T \geq T_{1-\alpha}^*) \rightarrow \alpha$ , where  $T_{1-\alpha}^*$  stands for the  $1 - \alpha$  percentile of the bootstrap distribution of  $T$ . From Corollary 1(b) and Theorem 2, it follows that such a test is consistent against fixed alternatives  $\pi \in \Delta_k(\phi_2, \mathcal{P}, h_2) - \mathcal{P}$ , in the sense that  $P(T \geq T_{1-\alpha}^*) \rightarrow 1$ .

In practice, the bootstrap  $p$ -value must be approximated by simulation as follows:

1. Calculate the observed value of the test statistic for the available data  $(X_1, \dots, X_k)$ ,  $T_{obs}$ .
2. Generate  $B$  bootstrap samples  $(X_1^{*b}, \dots, X_k^{*b}) \sim \mathcal{M}_k(n; P(\hat{\theta}_{\phi_2, h_2}))$ ,  $b = 1, \dots, B$ , and calculate the test statistic for each bootstrap sample obtaining  $T^{*b}$ ,  $b = 1, \dots, B$ .
3. Approximate the  $p$ -value by means of the expression

$$\hat{p}_{boot} = \frac{\text{card}\{b : T_b^{*b} \geq T_{obs}\}}{B}.$$

For the numerical experiments previously described, whose results are displayed in Tables 1–6, we also calculated the bootstrap  $p$ -values. This was done by generating  $B = 1000$  bootstrap samples to approximate each  $p$ -value, and calculating the fraction of these  $p$ -values, which are less than or equal to 0.05 and 0.10 (top and bottom in the tables). Tables 7–12 display the estimated type I error probabilities obtained by using the bootstrap approximation as well as those obtained with the asymptotic approximation (bootstrap, B, and asymptotic, A, in the tables) taken from Tables 1–6 in order to facilitate the comparison between them. Looking at Tables 7–12, we conclude that the bootstrap approximation is superior to the asymptotic one for small and moderate sample sizes, since in all cases the bootstrap type I error probabilities were closer to the nominal values than those obtained using the asymptotic null distribution. This superior performance of the bootstrap null distribution estimator has been noticed in other inferential problems, where  $\phi$ -divergences are used as test statistics (see, for example, [5,12,15,16]).

**Table 7.** Asymptotic and bootstrap type I error probabilities for Example 1 with  $\theta = 0.3333$ ,  $\phi_1 = PD_\lambda$ ,  $\lambda \in \{-2, 1, 2\}$ ,  $\phi_2 = PD_{-2}$ ,  $h_1 = h_2 \in \{0.5, 1, 2\}$ .

		$h_1 = h_2$		0.5		1		2	
$\phi_1$	$n$	B	A	B	A	B	A	B	A
$PD_{-2}$	100	0.051	0.996	0.048	0.996	0.048	0.998		
		0.110	0.996	0.103	0.996	0.109	0.998		
	150	0.055	0.995	0.050	0.995	0.056	0.996		
		0.106	0.995	0.101	0.995	0.109	0.996		
	200	0.053	0.992	0.053	0.993	0.056	0.994		
		0.103	0.992	0.106	0.994	0.108	0.994		
$PD_1$	100	0.057	0.995	0.056	0.997	0.055	0.996		
		0.110	0.995	0.110	0.997	0.107	0.996		
	150	0.054	0.994	0.052	0.995	0.055	0.996		
		0.110	0.994	0.104	0.995	0.114	0.996		
	200	0.055	0.992	0.051	0.994	0.052	0.991		
		0.106	0.992	0.103	0.994	0.106	0.991		
$PD_2$	100	0.055	0.995	0.056	0.997	0.054	0.997		
		0.110	0.995	0.109	0.997	0.107	0.997		
	150	0.054	0.994	0.055	0.994	0.056	0.995		
		0.107	0.994	0.106	0.994	0.110	0.995		
	200	0.054	0.993	0.053	0.993	0.055	0.994		
		0.107	0.993	0.105	0.993	0.108	0.994		

**Table 8.** Asymptotic and bootstrap type I error probabilities for Example 1 with  $n = 200$ ,  $\theta = 0.3333$ ,  $\phi_1 = \phi_2 = PD_{-2}$ ,  $h_1 \neq h_2$ , and  $h_1, h_2 \in \{0.5, 1, 2\}$ .

$(h_1, h_2)$	(0.5, 1)		(1, 0.5)		(0.5, 2)		(2, 0.5)		(1, 2)		(2, 1)	
	B	A	B	A	B	A	B	A	B	A	B	A
0.061	0.989	0.050	0.997	0.059	0.996	0.042	0.998	0.044	0.994	0.063	0.998	
0.107	0.999	0.113	0.997	0.106	0.996	0.095	0.998	0.105	0.994	0.115	0.999	

**Table 9.** Asymptotic and bootstrap type I error probabilities for Example 2 with  $\theta = 0.24$ ,  $\phi_1 = PD_\lambda$ ,  $\lambda \in \{-2, 1, 2\}$ ,  $\phi_2 = PD_{-2}$ , and  $h_1 = h_2 \in \{0.5, 1, 2\}$ .

		$h_1 = h_2$		0.5		1		2	
$\phi_1$	$n$	B	A	B	A	B	A	B	A
$PD_{-2}$	100	0.057	0.016	0.055	0.017	0.051	0.017		
		0.111	0.034	0.110	0.036	0.102	0.036		
	150	0.049	0.018	0.048	0.019	0.051	0.017		
		0.097	0.035	0.103	0.039	0.101	0.036		
	200	0.051	0.024	0.055	0.022	0.051	0.022		
		0.099	0.043	0.102	0.042	0.099	0.040		
$PD_1$	100	0.058	0.013	0.054	0.013	0.051	0.014		
		0.114	0.031	0.113	0.030	0.106	0.031		
	150	0.050	0.014	0.051	0.014	0.052	0.014		
		0.098	0.031	0.103	0.031	0.100	0.032		
	200	0.049	0.014	0.054	0.016	0.052	0.016		
		0.099	0.032	0.104	0.034	0.099	0.032		
$PD_2$	100	0.055	0.013	0.053	0.014	0.050	0.015		
		0.110	0.030	0.108	0.033	0.104	0.033		
	150	0.050	0.013	0.052	0.015	0.051	0.016		
		0.097	0.032	0.103	0.033	0.098	0.032		
	200	0.049	0.014	0.051	0.015	0.051	0.016		
		0.100	0.032	0.102	0.035	0.098	0.033		

**Table 10.** Asymptotic and bootstrap type I error probabilities for Example 2 with  $n = 200$ ,  $\theta = 0.24$ ,  $\phi_1 = \phi_2 = PD_{-2}$ ,  $h_1 \neq h_2$ , and  $h_1, h_2 \in \{0.5, 1, 2\}$ .

$(h_1, h_2)$	(0.5, 1)		(1, 0.5)		(0.5, 2)		(2, 0.5)		(1, 2)		(2, 1)	
	B	A	B	A	B	A	B	A	B	A	B	A
0.048	0.017	0.051	0.017	0.052	0.018	0.053	0.019	0.050	0.018	0.049	0.016	
0.101	0.035	0.099	0.033	0.100	0.035	0.105	0.040	0.103	0.036	0.101	0.034	

**Table 11.** Asymptotic and bootstrap type I error probabilities for Example 3 with  $\theta = 0.8$ ,  $\phi_1 = PD_\lambda$ ,  $\lambda \in \{-2, 1, 2\}$ ,  $\phi_2 = PD_{-2}$ , and  $h_1 = h_2 \in \{0.5, 1, 2\}$ .

		$h_1 = h_2$		0.5		1		2	
$\phi_1$	$n$	B	A	B	A	B	A	B	A
$PD_{-2}$	100	0.066	0.063	0.058	0.066	0.044	0.074		
		0.119	0.122	0.101	0.120	0.086	0.125		
	150	0.053	0.063	0.050	0.064	0.045	0.066		
		0.098	0.114	0.095	0.118	0.093	0.113		
	200	0.051	0.062	0.047	0.061	0.046	0.061		
		0.099	0.111	0.096	0.111	0.100	0.115		
$PD_1$	100	0.049	0.095	0.049	0.107	0.041	0.111		
		0.103	0.157	0.098	0.065	0.084	0.161		
	150	0.050	0.083	0.040	0.082	0.040	0.084		
		0.098	0.137	0.090	0.134	0.087	0.136		
	200	0.046	0.075	0.048	0.079	0.044	0.074		
		0.095	0.129	0.102	0.137	0.092	0.123		
$PD_2$	100	0.043	0.122	0.045	0.136	0.037	0.131		
		0.099	0.181	0.046	0.190	0.077	0.182		
	150	0.040	0.099	0.047	0.105	0.035	0.100		
		0.041	0.153	0.093	0.159	0.081	0.152		
	200	0.043	0.086	0.048	0.091	0.043	0.086		
		0.092	0.145	0.097	0.148	0.090	0.144		

**Table 12.** Asymptotic and bootstrap type I error probabilities for Example 3 with  $n = 200$ ,  $\theta = 0.8$ ,  $\phi_1 = \phi_2 = PD_{-2}$ ,  $h_1 \neq h_2$ , and  $h_1, h_2 \in \{0.5, 1, 2\}$ .

$(h_1, h_2)$	(0.5, 1)		(1, 0.5)		(0.5, 2)		(2, 0.5)		(1, 2)		(2, 1)	
	B	A	B	A	B	A	B	A	B	A	B	A
0.047	0.060	0.048	0.062	0.051	0.063	0.049	0.062	0.048	0.063	0.044	0.058	
0.095	0.108	0.099	0.114	0.099	0.113	0.097	0.112	0.099	0.113	0.092	0.109	

#### 4. Application to the Evaluation of the Thematic Classification in Global Land Cover Maps

This section displays the results of an application of our proposal to two real data sets related to the thematic quality assessment of a global land cover (GLC) map. The data comprise the results of two thematic classifications of the land cover category “Evergreen Broadleaf Trees” (EBL) and summarize the number of sample units correctly classified in this class, and the number of confusions with other land cover classes: “Deciduous Broadleaf Trees” (DBL), “Evergreen Needleleaf Trees” (ENL), and “Urban/Built Up” (U). The results of these two classifications were collected from two different global land cover maps: the Globcover map and the LC-CCI map (see Tsendbazar et al. [17] for additional details) and they are displayed in Table 13.

**Table 13.** Thematic classification of the Evergreen Broadleaf Trees (EBL) class.

	Globcover Map	LC-CCI Map
Classified Data		
EBL	165	172
DBL	13	5
ENL	7	5
U	0	0

Parametric specifications of the multinomial vector of probabilities are quite attractive since they describe in a concise way the classification pattern. Because of this, given the similarity between the two observed classifications in Table 13, we are interested in the search of a parametric model suitable to depict the thematic accuracy of this class in both GLC maps. For this purpose, we consider the parametric family in Equation (3) of Example 3. The presence of a zero cell frequency in each data set leads us to consider a penalized  $\phi$ -divergence as a test statistic for testing goodness-of-fit to such a parametric family.

Table 14 displays the observed values of the test statistic  $T$  and the associated bootstrap  $p$ -values for the goodness-of-fit test with respect to the parametric family in Equation (3) for the two observed classifications of the EBL class in Table 13. Looking at this table, it can be concluded that the null hypothesis cannot be rejected in both cases. Therefore, the parametric model in Equation (3) provides an adequate description of the thematic classification of the EBL class.

**Table 14.** Results of the goodness-of-fit test applied to the thematic classification of the EBL class.

Globcover Map			LC-CCI Map		
$\hat{\theta}_{-2,0.5} = 0.9490$			$\hat{\theta}_{-2,0.5} = 0.9721$		
$\phi_1$	$PD_{-2}$	$PD_1$	$PD_2$	$PD_{-2}$	$PD_1$
$T_{obs}$	2.3015	2.7618	3.0111	0.1432	0.1432
$\hat{p}_{boot}$	0.1700	0.2253	0.2926	0.9283	0.9200
$\hat{\theta}_{-2,1} = 0.9503$			$\hat{\theta}_{-2,1} = 0.9725$		
$T_{obs}$	2.7686	3.3752	3.6962	0.2821	0.2823
$\hat{p}_{boot}$	0.1801	0.2325	0.2671	0.8431	0.9162
$\hat{\theta}_{-2,2} = 0.9527$			$\hat{\theta}_{-2,2} = 0.9732$		
$T_{obs}$	3.6352	4.5400	5.0219	0.5492	0.5508
$\hat{p}_{boot}$	0.1300	0.2492	0.2584	0.7526	0.8144

## 5. Proofs

Notice that

$$\begin{aligned} D_{\phi,h}(\pi, P(\theta)) &= \sum_{i=1}^m p_i(\theta) \phi \left( \frac{\pi_i}{p_i(\theta)} \right) + h \sum_{i=m+1}^k p_i(\theta) \\ &= hI(m < k) + \sum_{i=1}^m p_i(\theta) \phi_h \left( \frac{\pi_i}{p_i(\theta)} \right) \end{aligned}$$

where  $I$  stands for the indicator function,  $\phi_h(x) = \phi(x) - h$ , if  $m < k$ , and  $\phi_h(x) = \phi(x)$ , if  $m = k$ . Let

$$D_{\phi,h}^+(\pi, P(\theta)) = \sum_{i=1}^m p_i(\theta) \phi_h \left( \frac{\pi_i}{p_i(\theta)} \right).$$

Clearly,

$$\arg \min_{\theta} D_{\phi,h}(\hat{\pi}, P(\theta)) = \arg \min_{\theta} D_{\phi,h}^+(\hat{\pi}, P(\theta)).$$

Note that, if Assumptions 1 and 2 hold, then Assumption 3 implies that

$$\frac{\partial}{\partial \theta} D_{\phi}^+(\pi, P(\theta_0)) = \sum_{i=1}^m \frac{\partial}{\partial \theta} p_i(\theta_0) v_i = 0 \quad (4)$$

where

$$v_i = \phi \left( \frac{\pi_i}{p_i(\theta_0)} \right) - \frac{\pi_i}{p_i(\theta_0)} \phi' \left( \frac{\pi_i}{p_i(\theta_0)} \right) - hI(m < k) \quad (5)$$

$1 \leq i \leq m$ , and  $\phi'(x) = \frac{\partial}{\partial x} \phi(x)$ . The  $s \times s$  matrix

$$\mathbb{D}_2 = \frac{\partial^2}{\partial \theta \partial \theta^t} D_{\phi}^+(\pi, P(\theta_0)) = \sum_{i=1}^m \frac{\partial^2}{\partial \theta \partial \theta^t} p_i(\theta_0) v_i + \sum_{i=1}^m \frac{\partial}{\partial \theta} p_i(\theta_0) \frac{\partial}{\partial \theta} p_i(\theta_0)^t w_i \quad (6)$$

is positive definite, where

$$w_i = \frac{\pi_i^2}{p_i^3(\theta_0)} \phi'' \left( \frac{\pi_i}{p_i(\theta_0)} \right),$$

$1 \leq i \leq m$ , and  $\phi''(x) = \frac{\partial^2}{\partial x^2} \phi(x)$ . Therefore, by the Implicit Function Theorem (see, for example, Dieudonne [18], p. 272), there is an open neighborhood  $U \subseteq (0, 1)^m$  of  $\pi^+$  and  $s$  unique functions,  $g_i : U \rightarrow \mathbb{R}$ ,  $1 \leq i \leq s$ , so that

- (i)  $\hat{\theta}_{\phi} = (g_1(\hat{\pi}^+), \dots, g_s(\hat{\pi}^+))^t$ ,  $\forall n \geq n_0$ , for some  $n_0 \in \mathbb{N}$ ;

- (ii)  $\theta_0 = (g_1(\pi^+), \dots, g_s(\pi^+))^t$ ;  
 (iii)  $g = (g_1, \dots, g_s)^t$  is continuously differentiable in  $U$  and the  $s \times m$  Jacobian matrix of  $g$  at  $(\pi_1, \dots, \pi_m)$  is given by

$$G = \mathbb{D}_2^{-1} D_1(P(\theta_0)) Diag(\omega) \quad (7)$$

where

$$D_1(P(\theta)) = \left( \frac{\partial}{\partial \theta} p_1(\theta), \dots, \frac{\partial}{\partial \theta} p_m(\theta) \right), \quad (8)$$

$$\omega = (\omega_1, \dots, \omega_m)^t,$$

$$\omega_i = \frac{\pi_i}{p_i^2(\theta_0)} \phi'' \left( \frac{\pi_i}{p_i(\theta_0)} \right),$$

and  $1 \leq i \leq m$ .

**Proof of Theorem 1.** Part (a) follows from (i) and (ii) above and the fact that  $\hat{\pi}^+ \rightarrow \pi^+$  a.s. From (i)–(ii), and taking into account that  $\sqrt{n}(\hat{\pi}^+ - \pi^+)$  is asymptotically normal, it follows that

$$\hat{\theta}_\phi = \theta_0 + G(\pi, P(\theta_0), \phi)(\hat{\pi} - \pi) + o_P(n^{-1/2}). \quad (9)$$

Parts (b) and (c) follow from Equation (9) and the asymptotic normality of  $\sqrt{n}(\hat{\pi}^+ - \pi^+)$ .  $\square$

**Proof of Corollary 1.** Part (a) was shown in Theorem 5.1 in [7]. To prove (b), we first demonstrate that

$$W = W_0 + r_n \quad (10)$$

where

$$W_0 = \sqrt{n} \left\{ \sum_{j=1}^m p_j(\hat{\theta}_{\phi_2, h_2}) \phi_1 \left( \frac{\hat{\pi}_j}{p_j(\hat{\theta}_{\phi_2, h_2})} \right) + h_1 \sum_{j=m+1}^k p_j(\hat{\theta}_{\phi_2, h_2}) - D_{\phi_1, h_1}(\pi, P(\theta_0)) \right\} + r_n,$$

and  $r_n = o_P(1)$ . Notice that

$$\begin{aligned} r_n &= \sqrt{n}\{h_1 - \phi_1(0)\} \sum_{j: \hat{\pi}_j=0, \pi_j>0} p_j(\hat{\theta}_{\phi_2, h_2}) \\ &= \sqrt{n}\{h_1 - \phi_1(0)\} \sum_{j=1}^m p_j(\hat{\theta}_{\phi_2, h_2}) I(\hat{\pi}_j = 0). \end{aligned}$$

Therefore,

$$0 \leq E|r_n| \leq \sqrt{n}|h_1 - \phi_1(0)| \sum_{j=1}^m P(\hat{\pi}_j = 0) = \sqrt{n}|h_1 - \phi_1(0)| \sum_{j=1}^m (1 - \pi_j)^n \rightarrow 0,$$

which implies  $r_n = o_P(1)$ . From Theorem 1 and Taylor expansion, it follows that  $W_0 \xrightarrow{\mathcal{L}} N(0, \sigma^2)$ ; hence, the result in part (b) is proven.  $\square$

**Proof of Theorem 2.** The proof of Theorem 2 is parallel to that of Theorem 2 in [5], so we omit it.  $\square$

**Author Contributions:** M.V. Alba-Fernández and M.D. Jiménez-Gamero conceived and designed the experiments; M.V. Alba-Fernández performed the experiments; M.V. Alba-Fernández and F.J. Ariza-López analyzed the data; F.J. Ariza-López contributed materials; M.V. Alba-Fernández and M.D. Jiménez-Gamero wrote the paper.

**Acknowledgments:** The authors thank the anonymous referees for their valuable time and careful comments, which improved the presentation of this paper. The research in this paper has been partially funded by grants: CTM2015-68276-R of the Spanish Ministry of Economy and Competitiveness (M.V. Alba-Fernández and F.J. Ariza-López) and MTM2017-89422-P of the Spanish Ministry of Economy, Industry and Competitiveness, ERDF support included (M.D. Jiménez-Gamero).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

MLE	maximum likelihood estimator
M $\phi$ E	minimum $\phi$ -divergence estimator
MP $\phi$ E	minimum penalized $\phi$ -divergence estimator
RMSD	root mean square deviation
B	bootstrap
A	asymptotic
GLC	global land cover
EBL	evergreen broadleaf trees
DBL	deciduous broadleaf trees
ENL	evergreen needleleaf trees
U	urban/built up

## References

- Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman & Hall: London, UK; CRC Press: Boca Raton, FL, USA, 2006.
- Basu, A.; Sarkar, S. On disparity based goodness-of-fit tests for multinomial models. *Stat. Probab. Lett.* **1994**, *19*, 307–312. [[CrossRef](#)]
- Morales, D.; Pardo, L.; Vajda, I. Asymptotic divergence of estimates of discrete distributions. *J. Stat. Plann. Inference* **1995**, *48*, 347–369. [[CrossRef](#)]
- Mandal, A.; Basu, A.; Pardo, L. Minimum disparity inference and the empty cell penalty: Asymptotic results. *Sankhya Ser. A* **2010**, *72*, 376–406. [[CrossRef](#)]
- Jiménez-Gamero, M.D.; Pino-Mejías, R.; Alba-Fernández, M.V.; Moreno-Rebolledo, J.L. Minimum  $\phi$ -divergence estimation in misspecified multinomial models. *Comput. Stat. Data Anal.* **2011**, *55*, 3365–3378. [[CrossRef](#)]
- White, H. Maximum likelihood estimation of misspecified models. *Econometrica* **1982**, *50*, 1–25. [[CrossRef](#)]
- Mandal, A.; Basu, A. Minimum disparity inference and the empty cell penalty: Asymptotic results. *Electron. J. Stat.* **2011**, *5*, 1846–1875. [[CrossRef](#)]
- Csiszár, I. Information type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **1967**, *2*, 299–318.
- Lindsay, B.G. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Stat.* **1994**, *22*, 1081–1114. [[CrossRef](#)]
- Vuong, Q.H.; Wang, W. Minimum  $\chi$ -square estimation and tests for model selection. *J. Econom.* **1993**, *56*, 141–168. [[CrossRef](#)]
- Alba-Fernández, M.V.; Jiménez-Gamero, M.D.; Lagos-Álvarez, B. Divergence statistics for testing uniform association in cross-classifications. *Inf. Sci.* **2010**, *180*, 4557–4571. [[CrossRef](#)]
- Jiménez-Gamero, M.D.; Pino-Mejías, R.; Rufián-Lizana, A. Minimum  $K_\phi$ -divergence estimators for multinomial models and applications. *Comput. Stat.* **2014**, *29*, 363–401. [[CrossRef](#)]
- R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017. Available online: <https://www.R-project.org/> (accessed on 29 April 2018).
- Amari, S. Integration of stochastic models by minimizing  $\alpha$ -divergence. *Neural Comput.* **2007**, *19*, 2780–2796. [[CrossRef](#)] [[PubMed](#)]
- Alba-Fernández, M.V.; Jiménez-Gamero, M.D. Bootstrapping divergence statistics for testing homogeneity in multinomial populations. *Math. Comput. Simul.* **2009**, *79*, 3375–3384. [[CrossRef](#)]
- Jiménez-Gamero, M.D.; Alba-Fernández, M.V.; Barranco-Chamorro, I.; Muñoz-García, J. Two classes of divergence statistics for testing uniform association. *Statistics* **2014**, *48*, 367–387. [[CrossRef](#)]

17. Tsendbazar, N.E.; de Bruina, S.; Mora, B.; Schoutenc, L.; Herolda, M. Comparative assessment of thematic accuracy of GLC maps for specific applications using existing reference data. *Int. J. Appl. Earth. Obs. Geoinf.* **2016**, *44*, 124–135. [[CrossRef](#)]
18. Dieudonne, J. *Foundations of Modern Analysis*; Academic Press: New York, NY, USA; London, UK, 1969.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Non-Quadratic Distances in Model Assessment

Marianthi Markatou \* and Yang Chen

Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, USA; ychen57@buffalo.edu

\* Correspondence: markatou@buffalo.edu

Received: 31 March 2018; Accepted: 13 June 2018; Published: 14 June 2018



**Abstract:** One natural way to measure model adequacy is by using statistical distances as loss functions. A related fundamental question is how to construct loss functions that are scientifically and statistically meaningful. In this paper, we investigate non-quadratic distances and their role in assessing the adequacy of a model and/or ability to perform model selection. We first present the definition of a statistical distance and its associated properties. Three popular distances, total variation, the mixture index of fit and the Kullback-Leibler distance, are studied in detail, with the aim of understanding their properties and potential interpretations that can offer insight into their performance as measures of model misspecification. A small simulation study exemplifies the performance of these measures and their application to different scientific fields is briefly discussed.

**Keywords:** model assessment; statistical distance; non-quadratic distance; total variation; mixture index of fit; Kullback-Leibler distance; divergence measure

---

## 1. Introduction

Model assessment, that is assessing the adequacy of a model and/or ability to perform model selection, is one of the fundamental components of statistical analyses. For example, in the model adequacy problem one usually begins with a fixed model and interest centers on measuring the model misspecification cost. A natural way to create a framework within which we can assess model misspecification is by using statistical distances as loss functions. These constructs measure the distance between the unknown distribution that generated the data and an estimate from the data model. By identifying statistical distances as loss functions, we can begin to understand the role distances play in model fitting and selection, as they become measures of the overall cost of model misspecification. This strategy will allow us to investigate the construction of a loss function as the maximum error in a list of model fit questions. Therefore, our fundamental question is the following. How can one design a loss function  $\rho$  that is scientifically and statistically meaningful? We would like to be able to attach a specific scientific meaning to the numerical values of the loss, so that a value of the distance equal to 4, for example, has an explicit interpretation in terms of our statistical goals. When we select between models, we would like to measure the quality of the approximation via the model's ability to provide answers to important scientific questions. This presupposes that the meaning of "best fitting model" should depend on the "statistical questions" being asked of the model.

Lindsay [1] discusses a distance-based framework for assessing model adequacy. A fundamental tenet of the framework for model adequacy put forward by Lindsay [1] is that it is possible and reasonable to carry out a model-based scientific inquiry without believing that the model is true, and without assuming that the truth is included in the model. All this of course, assuming that we have a way to measure the quality of the approximation to the "truth", is offered by the model. This point of view never assumes the correctness of the model. Of course, it is rather presumptuous to label any distribution as the truth as any basic modeling assumption generated by the sampling scheme that provided the data is never exactly true. An example of a basic modeling assumption

might be “ $X_1, X_2, \dots, X_n$  are independent, identically distributed from an unknown distribution  $\tau$ ”. This, as any other statistical assumption, is subject to question even in the most idealized of data collection frameworks. However, we believe that well designed experiments can generate data that is similar to data from idealized models, therefore we operate as if the basic assumption is true. This means that we assume that there is a true distribution  $\tau$  that generates the data, which is “knowable” if we can collect an infinite amount of data. Furthermore, we note that the basic modeling assumption will be the global framework for assessment of all more restrictive assumptions about the data generation mechanism. In a sense, it is the “nonparametric” extension of the more restrictive models that might be considered.

We let  $\mathcal{P}$  be the class of all distributions consistent with the basic assumptions. Hence  $\tau \in \mathcal{P}$ , and sets  $\mathcal{H} \in \mathcal{P}$  are called models. We assume that  $\tau \notin \mathcal{H}$ ; hence, there is a permanent model misspecification error. Statistical distances will then provide a measure for the model misspecification error.

One natural way to measure model adequacy is to define a loss function  $\rho(\tau, M)$  that describes the loss incurred when the model element  $M$  is used instead of the true distribution  $\tau$ . Such a loss function should, in principle, indicate, in an inferential sense, how far apart the two distributions  $\tau, M$  are. In the next section, we offer a formal definition of the concept of a statistical distance.

If the statistical questions of interest can be expressed as a list of functionals  $T(M)$  of the model  $M$  that we wish to be uniformly close to the same functionals  $T(\tau)$  of the true distribution, then we can turn the set of model fit questions into a distance via

$$\rho(\tau, M) = \sup_{T(\cdot)} |T(\tau) - T(M)|,$$

where the supremum is taken over the class of functionals of interest. Using the supremum of the individual errors is one way of assessing overall error, but using this measure has the nice feature that its value gives a bound on all individual errors. The statistical questions of interest may be global, such as: is the normal model correct in every aspect? Or we may be interested to have answers on a few key characteristics, such as the mean.

Lindsay et al. [2] introduced a class of statistical distances, called quadratic distances, and studied their use in the context of goodness-of-fit testing. Furthermore, Markatou et al. [3] discuss extensively the chi-squared distance, a special case of quadratic distance, and its role in robustness. In this paper, we study non-quadratic distances and their role in model assessment. The paper is organized as follows. Section 2 presents the definition of a statistical distance and its associated properties. Sections 3–5 discuss in detail three popular distances, total variation, the mixture index of fit and the Kullback-Leibler distance, with the aim of understanding their role in model assessment problems. The likelihood distance is also briefly discussed in Section 5. Section 6 illustrates computation and applications of total variation, mixture index of fit and Kullback-Leibler distances. Finally, Section 7 presents discussion and conclusions pertaining to the use of total variation and mixture index of fit distances.

## 2. Statistical Distances and Their Properties

If we adopt the usual convention that loss functions are nonnegative in their arguments, and zero if the correct model is used, and have larger value if the two distributions are not very similar, then the loss  $\rho(\tau, M)$  can also be viewed as a distance between  $\tau, M$ . In fact, we will always assume that for any two distributions  $F, G$

$$\rho(F, G) \geq 0 \text{ with } \rho(F, F) = 0.$$

If this holds, we will say that  $\rho$  is a statistical distance. Unlike the requirements for a metric, we do not require symmetry. In fact, there is no reason that the loss should be symmetric, as the roles of  $\tau, M$  are different. We also do not require  $\rho$  to be nonzero when the arguments differ. This zero

property will allow us to specify that two distributions are equivalent as far as our statistical purposes are concerned by giving them zero distance.

Furthermore, it is important to note that if  $\tau$  is in  $\mathcal{H}$  and  $\tau = M_{\theta_0}$ , and  $M \in \mathcal{H}$ , say  $M_\theta$ , then the distance  $\rho(\tau, M)$  induces a loss function on the parameter space via

$$L_\rho(\theta_0, \theta) \stackrel{\text{def}}{=} \rho(M_{\theta_0}, M_\theta).$$

Therefore, if  $\tau$  is in the model, the losses defined by  $\rho$  are parametric losses.

We begin within the discrete distribution framework. Let  $\mathcal{T} = \{0, 1, 2, \dots, T\}$ , where  $T$  is possibly infinite, be a discrete sample space. On this sample space we define a true probability density  $\tau(t)$ , as well as a family of densities  $\mathcal{M} = \{m_\theta(t) : \theta \in \Theta\}$ , where  $\Theta$  is the parameter space. Assume we have independent and identically distributed random variables  $X_1, X_2, \dots, X_n$  producing the realizations  $x_1, x_2, \dots, x_n$  from  $\tau(\cdot)$ . We record the data as  $d(t) = n(t)/n$ , where  $n(t)$  is the number of observations in the sample with value equal to  $t$ . We note here that we use the word “density” in a generic fashion that incorporates both, probability mass functions as well as probability density functions. A rather formal definition of the concept of statistical distance is as follows.

**Definition 1.** (Markatou et al. [3]) Let  $\tau, m$  be two probability density functions. Then  $\rho(\tau, m)$  is a statistical distance between the corresponding probability distributions if  $\rho(\tau, m) \geq 0$ , with equality if and only if  $\tau$  and  $m$  are the same for all statistical purposes.

We would require  $\rho(\tau, m)$  to indicate the worst mistake that we can make if we use  $m$  instead of  $\tau$ . The precise meaning of this statement is obvious in the case of total variation that we discuss in detail in Section 3 of the paper.

We would also like our statistical distances to be convex in their arguments.

**Definition 2.** Let  $\tau, m$  be a pair of probability density functions, with  $m$  being represented as  $m = \alpha m_1 + (1 - \alpha)m_2$ ,  $0 \leq \alpha \leq 1$ . We say that the statistical distance  $\rho(\tau, m)$  is convex in the right argument if

$$\rho(\tau, \alpha m_1 + (1 - \alpha)m_2) \leq \alpha \rho(\tau, m_1) + (1 - \alpha) \rho(\tau, m_2),$$

where  $m_1, m_2$  are two probability density functions.

**Definition 3.** Let  $\tau, m$  be a pair of probability density functions, and assume  $\tau = \gamma \tau_1 + (1 - \gamma)\tau_2$ ,  $0 \leq \gamma \leq 1$ . Then, we say that  $\rho(\tau, m)$  is convex in the left argument if

$$\rho(\gamma \tau_1 + (1 - \gamma)\tau_2, m) \leq \gamma \rho(\tau_1, m) + (1 - \gamma) \rho(\tau_2, m),$$

where  $\tau_1, \tau_2$  are two densities.

Lindsay et al. [2] define and study quadratic distances as measures of goodness of fit, a form of model assessment. In the next sections, we study non-quadratic distances and their role in the problem of model assessment. We begin with the total variation distance.

### 3. Total Variation

In this section, we study the properties of the total variation distance. We offer a loss function interpretation of this distance and discuss sensitivity issues associated with its use. We will begin with the case of discrete probability measures and then move to the case of continuous probability measures. The results presented here are novel and are useful in selecting the distances to be used in any given problem.

The total variation distance is defined as follows.

**Definition 4.** Let  $\tau, m$  be two probability distributions. We define the total variation distance between the probability mass functions  $\tau, m$  to be

$$V(\tau, m) = \frac{1}{2} \sum_t |\tau(t) - m(t)|.$$

This measure is also known as the  $L_1$ -distance (without the factor 1/2) or index of dissimilarity.

**Corollary 1.** The total variation distance takes values in the interval  $[0, 1]$ .

**Proof.** By definition  $V(\tau, m) \geq 0$  with equality if and only if  $\tau = m, \forall t$ . Moreover,  $|\tau(t) - m(t)| \leq |\tau(t)| + |m(t)|$ . But  $\tau, m$  are probability mass functions (or densities), therefore

$$|\tau(t) - m(t)| \leq \tau(t) + m(t)$$

and hence

$$\frac{1}{2} \sum_t |\tau(t) - m(t)| \leq \frac{1}{2} \left( \sum_t \tau(t) + \sum_t m(t) \right)$$

or, equivalently

$$\frac{1}{2} \sum_t |\tau(t) - m(t)| \leq \frac{1}{2}(1+1) = 1.$$

Therefore  $0 \leq V(\tau, m) \leq 1$ .  $\square$

**Proposition 1.** The total variation distance is a metric.

**Proof.** By definition, the total variation distance is non-negative. Moreover, it is symmetric because  $V(\tau, m) = V(m, \tau)$  and it satisfies the triangle inequality since

$$\begin{aligned} V(\tau, m) &= \frac{1}{2} \sum_t |\tau(t) - m(t)| \\ &= \frac{1}{2} \sum_t |\tau(t) - g(t) + g(t) - m(t)| \\ &\leq \frac{1}{2} \left( \sum_t |\tau(t) - g(t)| + \sum_t |g(t) - m(t)| \right) \\ &= V(\tau, g) + V(g, m). \end{aligned}$$

Thus, it is a metric.  $\square$

The following proposition states that the total variation distance is convex in both, left and right arguments.

**Proposition 2.** Let  $\tau, m$  be a pair of densities with  $\tau$  represented as  $\tau = \alpha\tau_1 + (1-\alpha)\tau_2, 0 \leq \alpha \leq 1$ . Then

$$V(\alpha\tau_1 + (1-\alpha)\tau_2, m) \leq \alpha V(\tau_1, m) + (1-\alpha)V(\tau_2, m).$$

Moreover, if  $m$  is represented as  $m = \gamma m_1 + (1-\gamma)m_2, 0 \leq \gamma \leq 1$ , then

$$V(\tau, \gamma m_1 + (1-\gamma)m_2) \leq \gamma V(\tau, m_1) + (1-\gamma)V(\tau, m_2).$$

**Proof.** It is a straightforward application of the definition of the total variation distance.  $\square$

The total variation measure has major implications for prediction probabilities. A statistically useful interpretation of the total variation distance is that it can be thought of as the worst error we can

commit in probability when we use the model  $m$  instead of the truth  $\tau$ . The maximum value of this error equals 1 and it occurs when  $\tau, m$  are mutually singular.

Denote by  $P_\tau$  the probability of a set under the measure  $\tau$  and by  $P_m$  the probability of a set under the measure  $m$ .

**Proposition 3.** *Let  $\tau, m$  be two probability mass functions. Then*

$$V(\tau, m) = \sup_{A \subset \mathcal{B}} |\mathbb{P}_\tau(A) - \mathbb{P}_m(A)|,$$

where  $A$  is a subset of the Borel set  $\mathcal{B}$ .

**Proof.** Define the sets  $B_1 = \{t : \tau(t) > m(t)\}$ ,  $B_2 = \{t : \tau(t) < m(t)\}$ ,  $B_3 = \{t : \tau(t) = m(t)\}$ . Notice that

$$\mathbb{P}_\tau(B_1) + \mathbb{P}_\tau(B_2) + \mathbb{P}_\tau(B_3) = \mathbb{P}_m(B_1) + \mathbb{P}_m(B_2) + \mathbb{P}_m(B_3) = 1.$$

Because on the set  $B_3$  the two probability mass functions are equal  $\mathbb{P}_\tau(B_3) = \mathbb{P}_m(B_3)$ , and hence

$$\mathbb{P}_\tau(B_1) - \mathbb{P}_m(B_1) = \mathbb{P}_m(B_2) - \mathbb{P}_\tau(B_2).$$

Note that, because of the nature of the sets  $B_1$  and  $B_2$ , both terms in the last expression are positive. Therefore

$$\begin{aligned} V(\tau, m) &= \frac{1}{2} \sum |\tau(t) - m(t)| \\ &= \frac{1}{2} \left( \sum_{t \in B_1} |\tau(t) - m(t)| + \sum_{t \in B_2} |\tau(t) - m(t)| + \sum_{t \in B_3} |\tau(t) - m(t)| \right) \\ &= \frac{1}{2} \{ (\mathbb{P}_\tau(B_1) - \mathbb{P}_m(B_1)) + (\mathbb{P}_m(B_2) - \mathbb{P}_\tau(B_2)) \} \\ &= \mathbb{P}_\tau(B_1) - \mathbb{P}_m(B_1). \end{aligned}$$

Furthermore

$$\sup_{A \subset \mathcal{B}} |\mathbb{P}_\tau(A) - \mathbb{P}_m(A)| = \max \left\{ \sup_{A \subset \mathcal{B}} (\mathbb{P}_\tau(A) - \mathbb{P}_m(A)), \sup_{A \subset \mathcal{B}} (\mathbb{P}_m(A) - \mathbb{P}_\tau(A)) \right\}.$$

But

$$\sup_{A \subset \mathcal{B}} (\mathbb{P}_\tau(A) - \mathbb{P}_m(A)) = \mathbb{P}_\tau(B_1) - \mathbb{P}_m(B_1)$$

and

$$\sup_{A \subset \mathcal{B}} (\mathbb{P}_m(A) - \mathbb{P}_\tau(A)) = \mathbb{P}_m(B_2) - \mathbb{P}_\tau(B_2) = \mathbb{P}_\tau(B_1) - \mathbb{P}_m(B_1).$$

Therefore

$$\sup_{A \subset \mathcal{B}} |\mathbb{P}_\tau(A) - \mathbb{P}_m(A)| = \mathbb{P}_\tau(B_1) - \mathbb{P}_m(B_1),$$

and hence

$$V(\tau, m) = \sup_{A \subset \mathcal{B}} |\mathbb{P}_\tau(A) - \mathbb{P}_m(A)|.$$

□

**Remark 1.** The model misspecification measure  $V(\tau, m)$  has a “minimax” expression

$$V(\tau, \mathcal{M}) = \inf_{m \in \mathcal{M}} \sup_A \{|\mathbb{P}_\tau(A) - \mathbb{P}_m(A)| : A \subset \mathcal{B}\}.$$

This indicates the sense in which the measure assesses the overall risk of using  $m$  instead of  $\tau$ , then chooses  $m$  that minimizes the aforementioned risk.

We now offer a testing interpretation of the total variation distance. We establish that the total variation distance can be obtained as a solution to a suitably defined optimization problem. It is obtained as that test function which maximizes the difference between the power and level of a suitably defined test problem.

**Definition 5.** A randomized test function for testing a statistical hypothesis  $H_0$  versus the alternative  $H_1$  is a (measurable) function  $\phi$  defined on  $\mathbb{R}^n$  and taking values in the interval  $[0, 1]$  with the following interpretation. If  $x$  is the observed value of  $X$  and  $\phi(x) = y$ , then a coin whose probability of falling heads is  $y$  is tossed and  $H_0$  is rejected when head appears. In the case where  $y$  is either 0 or 1,  $\forall x$ , the test is called non-randomized.

**Proposition 4.** Let  $H_0 : \tau(x) = f(x)$  versus  $H_1 : \tau(x) = g(x)$  and  $\phi(x)$  is a test function,  $f, g$  are probability mass functions. Then

$$V(f, g) = \max_{\phi} \{\mathbb{E}_{H_1}(\phi(X)) - \mathbb{E}_{H_0}(\phi(X))\}.$$

**Proof.** We have

$$\mathbb{E}_{H_1}(\phi(X)) - \mathbb{E}_{H_0}(\phi(X)) = \sum \phi(x)(g(x) - f(x)).$$

Then

$$\phi(x) = 1 \text{ if } x \in B_1 = \{x : g(x) > f(x)\},$$

So

$$\max_{\phi} \sum \phi(x)(g(x) - f(x)) = \mathbb{P}_g(B_1) - \mathbb{P}_f(B_1) = V(f, g).$$

□

An advantage of the total variation distance is that it is not sensitive to small changes in the density. That is, if  $\tau(t)$  is replaced by  $\tau(t) + e(t)$  where  $\sum_t e(t) = 0$  and  $\sum_t |e(t)|$  is small then

$$\begin{aligned} V(\tau + e, m) &= \frac{1}{2} \sum |\tau(t) + e(t) - m(t)| \\ &\leq \frac{1}{2} \sum |\tau(t) - m(t)| + \frac{1}{2} \sum |e(t)| \\ &= V(\tau, m) + \frac{1}{2} \sum |e(t)|. \end{aligned}$$

Therefore, when the changes in the density are small  $V(\tau + e, m) \approx V(\tau, m)$ . When describing a population, it is natural to describe it via the proportion of individuals in various subgroups. Having  $V(\tau, m)$  small would ensure uniform accuracy for all such descriptions. On the other hand, populations are also described in terms of a variety of other variables, such as means. Having the total variation measure small does not imply that means are close on the scale of standard deviation.

**Remark 2.** The total variation distance is not differentiable in the arguments. Using  $V(d, m_\theta)$  as an inference function, where  $d$  denotes the data estimate of  $\tau$  (i.e.,  $\hat{\tau}$ ), yields estimators of  $\theta$  that have the feature of not generating smooth, asymptotically normal estimators when the model is true [4]. This feature is related to the

*pathologies of the variation distance described by Donoho and Liu [5]. However, if parameter estimation is of interest, one can use alternative divergences that are free of these pathologies.*

We now study the total variation distance in continuous probability models.

**Definition 6.** *The total variation distance between two probability density functions  $\tau, m$  is defined as*

$$V(\tau, m) = \frac{1}{2} \int |\tau(x) - m(x)| dx.$$

The total variation distance has the same interpretation as in the discrete probability model case. That is

$$V(\tau, m) = \sup_{A \subset \mathcal{B}} |\mathbb{P}_\tau(A) - \mathbb{P}_m(A)|.$$

One of the important issues in the construction of distances in continuous spaces is the issue of invariance, because the behavior of distance measures under transformations of the data is of interest. Suppose we take a monotone transformation of the observed variable  $X$  and use the corresponding model distribution; how does this transformation affect the distance between  $X$  and the model?

Invariance seems to be desirable from an inferential point of view, but difficult to achieve without forcing one of the distributions to be continuous and appealing to the probability integral transform for a common scale. In multivariate continuous spaces, the problem of transformation invariance is even more difficult, as there is no longer a natural probability integral transformation to bring data and model on a common scale.

**Proposition 5.** *Let  $V(\tau_X, m_X)$  be the total variation distance between the densities  $\tau_X, m_X$  for a random variable  $X$ . If  $Y = a(X)$  is a one-to-one transformation of the random variable  $X$ , then*

$$V(\tau_X, m_X) = V(\tau_Y, m_Y).$$

**Proof.** Write

$$\begin{aligned} V(\tau_Y, m_Y) &= \frac{1}{2} \int |\tau_Y(y) - m_Y(y)| dy \\ &= \frac{1}{2} \int \left| \tau_X(b(y)) \cdot \left| \frac{d}{dy} b(y) \right| - m_X(b(y)) \cdot \left| \frac{d}{dy} b(y) \right| \right| dy \\ &= \frac{1}{2} \int |\tau_X(b(y)) - m_X(b(y))| \cdot \left| \frac{d}{dy} b(y) \right| dy, \end{aligned}$$

where  $b(y)$  is the inverse transformation. Next, we do a change of variable in the integral. Set  $x = b(y)$  from where we obtain  $y = a(x)$  and  $dy = a'(x)dx$ ; the prime denotes derivative with respect to the corresponding argument. Then

$$V(\tau_Y, m_Y) = \frac{1}{2} \int |\tau_X(x) - m_X(x)| \cdot |b'(a(x))| \cdot a'(x) dx.$$

But

$$\begin{aligned} b(a(x)) &= x \implies \frac{d}{dx} b(a(x)) = 1 \\ &\implies b'(a(x)) a'(x) = 1 \\ &\implies b'(a(x)) = \frac{1}{a'(x)}, \end{aligned}$$

hence

$$\begin{aligned} V(\tau_Y, m_Y) &= \frac{1}{2} \int |\tau_X(x) - m_X(x)| \cdot \frac{a'(x)}{|a'(x)|} dx \\ &= \frac{1}{2} \int |\tau_X(x) - m_X(x)| \cdot \text{sign}(a'(x)) dx. \end{aligned}$$

Now since  $a(\cdot)$  is a one-to-one transformation,  $a(x)$  is either increasing or decreasing on different segments of  $\mathbb{R}$ . Thus

$$V(\tau_Y, m_Y) = V(\tau_X, m_X),$$

where  $Y = a(X)$ .  $\square$

A fundamental problem with the total variation distance is that it cannot be used to compute the distance between a discrete distribution and a continuous distribution because the total variation distance between a continuous measure and a discrete measure is always the maximum possible, that is 1. This inability of the total variation distance to discriminate between discrete and continuous measures can be interpreted as asking “too many questions” at once, without any prioritization. This limits its use despite its invariant characteristics.

We now discuss the relationship between the total variation distance and Fisher information. Denote by  $m^{(n)}$  the joint density of  $n$  independent and identically distributed random variables. Then we have the following proposition.

**Proposition 6.** *The total variation distance is locally equivalent to the Fisher information number, that is*

$$\frac{1}{n} V(m_{\theta}^{(n)}, m_{\theta_0}^{(n)}) \rightarrow |\theta - \theta_0| \sqrt{\frac{I(\theta_0)}{2\pi}}, \text{ as } n \rightarrow \infty,$$

where  $m_{\theta}, m_{\theta_0}$  are two discrete probability models.

**Proof.** By definition

$$V(m_{\theta}^{(n)}, m_{\theta_0}^{(n)}) = \frac{1}{2} \sum |m_{\theta}^{(n)}(t) - m_{\theta_0}^{(n)}(t)|.$$

Now, expand  $m_{\theta}^{(n)}(t)$  using Taylor series in the neighborhood of  $\theta_0$  to obtain

$$m_{\theta}^{(n)}(t) \simeq m_{\theta_0}^{(n)}(t) + (\theta - \theta_0) (m_{\theta_0}^{(n)}(t))'$$

where the prime denotes derivative with respect to the parameter  $\theta$ . Further, write

$$(m_{\theta_0}^{(n)}(t))' = m_{\theta_0}^{(n)}(t) \left( \frac{d}{d\theta} \log m_{\theta}^{(n)}(t) \Big|_{\theta_0} \right)$$

to obtain

$$\begin{aligned} \frac{1}{n} V(m_{\theta}^{(n)}, m_{\theta_0}^{(n)}) &\simeq \frac{1}{2} |\theta - \theta_0| \mathbb{E} \left\{ \frac{1}{n} \left| \frac{d}{d\theta} \log m_{\theta}^{(n)}(t) \Big|_{\theta_0} \right| \right\} \\ &= \frac{1}{2} |\theta - \theta_0| \mathbb{E} \left\{ \left| \frac{1}{n} \sum_{i=1}^n u_{\theta_0}(t_i) \right| \right\}, \end{aligned}$$

where

$$u_{\theta_0}(t_i) = \frac{d}{d\theta} \log m_{\theta}(t_i) \Big|_{\theta_0}.$$

Therefore, assuming that  $\frac{1}{n} \sum u_{\theta_0}(t_i)$  converges to a normal random variable in absolute mean, then

$$\frac{1}{n} V(m_{\theta}^{(n)}, m_{\theta_0}^{(n)}) \rightarrow \frac{1}{2} |\theta - \theta_0| \sqrt{I(\theta_0)} \sqrt{\frac{2}{\pi}} = |\theta - \theta_0| \sqrt{\frac{I(\theta_0)}{2\pi}}, \text{ as } n \rightarrow \infty,$$

because  $E(u_{\theta_0}(t_i)) = 0$ ,  $\text{Var}(u_{\theta_0}(t_i)) = I(\theta_0)$  and  $E(|Z|) = \sqrt{\frac{2}{\pi}}$  when  $Z \sim N(0, 1)$ .  $\square$

The total variation is a non-quadratic distance. It is however related to a quadratic distance, the Hellinger distance, defined as  $H^2(\tau, m) = \frac{1}{2} \sum (\sqrt{\tau(t)} - \sqrt{m(t)})^2$  by the following inequality.

**Proposition 7.** Let  $\tau, m$  be two probability mass functions. Then

$$0 \leq H^2(\tau, m) \leq V(\tau, m) \leq [H^2(\tau, m)(2 - H^2(\tau, m))]^{\frac{1}{2}}.$$

**Proof.** Straightforward using the definitions of the distances involved and Cauchy-Swartz inequality. Holder's inequality provides  $1 - \sum \sqrt{\tau(t)m(t)} \geq 0$ .  $\square$

Note that  $2H^2(\tau, m) = \sum [\sqrt{\tau(t)} - \sqrt{m(t)}]^2$ ; the square root of this quantity, that is  $\{\sum [\sqrt{\tau(t)} - \sqrt{m(t)}]^2\}^{1/2}$ , is known as Matusita's distance [6,7]. Further, define the affinity between two probability densities by

$$\rho(\tau, m) = \sum_t \tau^{1/2}(t)m^{1/2}(t).$$

Then, it is easy to prove that

$$\sum_t [\sqrt{\tau(t)} - \sqrt{m(t)}]^2 = 2(1 - \rho(\tau, m)) \leq V(\tau, m) \leq 2 \{\sum_t [\sqrt{\tau(t)} - \sqrt{m(t)}]^2\}^{1/2}.$$

The above inequality indicates the relationship between total variation and Matusita's distance.

#### 4. Mixture Index of Fit

Rudas, Clogg, and Lindsay [8] proposed a new index of fit approach to evaluate the goodness of fit analysis of contingency tables based on the mixture model framework. The approach focuses attention on the discrepancy between the model and the data, and allows comparisons across studies. Suppose  $\mathcal{M}$  is the baseline model. The family of models which are proposed for evaluating goodness of fit is a two-point mixture model given by

$$\mathcal{M}_{\pi} = \{\tau : \tau(t) = (1 - \pi)m_{\theta}(t) + \pi e(t), m_{\theta}(t) \in \mathcal{M}, e(t) \text{ arbitrary}, \theta \in \Theta\}.$$

Here  $\pi$  denotes the mixing proportion, which is interpreted as the proportion of the population outside the model  $\mathcal{M}$ . In the robustness literature the mixing proportion corresponds to the contamination proportion, as explained below. In the contingency table framework  $m_{\theta}(t)$ ,  $e(t)$  describe the tables of probabilities for each latent class. The family of models  $\mathcal{M}_{\pi}$  defines a class of nested models as  $\pi$  varies from zero to one. Thus, if the model  $\mathcal{M}$  does not fit well the data, then by increasing  $\pi$ , the model  $\mathcal{M}_{\pi}$  will be an adequate fit for  $\pi$  sufficiently large.

We can motivate the index of fit by thinking of the population as being composed of two classes with proportions  $1 - \pi$  and  $\pi$  respectively. The first class is perfectly described by  $\mathcal{M}$ , whereas the second class contains the "outliers". The index of fit can then be interpreted as the fraction of the population intrinsically outside  $\mathcal{M}$ , that is, the proportion of outliers in the sample.

We note here that these ideas can be extended beyond the contingency table framework. In our setting, the probability distribution describing the true data generating mechanism may be written as  $\tau(t) = (1 - \pi)m_{\theta}(t) + \pi e(t)$ , where  $m_{\theta}(t) \in \mathcal{M}$  and  $e(t)$  is arbitrary. This representation of  $\tau(t)$  is arbitrary such that we can construct another representation  $\tau(t) = (1 - \pi - \delta)m_{\theta}(t) + (\pi + \delta)e(t)$ .

$\delta)e^*(t)$ . However, there always exists the smallest unique  $\pi$  such that there exists a representation of  $\tau(t)$  that puts the maximum proportion in one of the population classes. Next, we define formally the mixture index of fit.

**Definition 7.** (Rudas, Clogg, and Lindsay [8]) The mixture index of fit  $\pi^*$  is defined by

$$\pi^*(\tau) = \inf\{\pi : \tau(t) = (1 - \pi)m_\theta(t) + \pi e(t), m_\theta(t) \in \mathcal{M}, e(t) \text{ arbitrary}\}.$$

Notice that  $\pi^*(\tau)$  is a distance. This is because if we set  $\pi^*(\tau, m_\theta) = \inf\{\pi : \tau(t) = (1 - \pi)m_\theta(t) + \pi e(t), e(t) \text{ arbitrary}\}$  for a fixed  $m_\theta(t)$ , we have  $\pi^*(\tau, m_\theta) > 0$  and  $\pi^*(\tau, m_\theta) = 0$  if  $\tau = m_\theta$ .

**Definition 8.** Define the statistical distance  $\pi^*(\tau, \mathcal{M})$  as follows:

$$\pi^*(\tau, \mathcal{M}) = \inf_{m \in \mathcal{M}} \pi^*(\tau, m).$$

**Remark 3.** Note that, to be able to present Proposition 8 below, we have turned arbitrary discrete distributions into vectors. As an example, if the sample space  $\mathcal{T} = \{0, 1, 2\}$  and  $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \mathbb{P}(X = 2) = 1/3$ , we write this discrete distribution as the vector  $(1/3, 1/3, 1/3)^T$ . If, furthermore, we consider the vectors  $\vec{\delta}_0 = (1, 0, 0)^T$ ,  $\vec{\delta}_1 = (0, 1, 0)^T$ , and  $\vec{\delta}_2 = (0, 0, 1)^T$  as degenerate distributions assigning mass 1 at positions 0, 1, 2 then  $(1/3, 1/3, 1/3)^T = \frac{1}{3}\vec{\delta}_0 + \frac{1}{3}\vec{\delta}_1 + \frac{1}{3}\vec{\delta}_2$ . This representation of distributions is used in the proof of Proposition 8.

**Proposition 8.** The set of vectors  $\vec{\tau}$  satisfying the relationship  $\pi^*(\vec{\tau}, \vec{m}) \leq \pi_0$  is a simplex with extremal points  $(1 - \pi_0)\vec{m} + \pi_0\vec{\delta}_i$ , where  $\vec{\delta}_i$  is the vector with 1 at the  $(i + 1)$ th position and 0 everywhere else.

**Proof.** Given  $\vec{\tau}$  with  $\pi^* \leq \pi_0$ , there exists a representation of

$$\vec{\tau} = (1 - \pi_0)\vec{m} + \pi_0\vec{e}.$$

Write any arbitrary discrete distribution  $\vec{e}$  as follows:

$$\vec{e} = e_0\vec{\delta}_0 + \cdots + e_T\vec{\delta}_T,$$

where  $\sum_{i=0}^T e_i = 1$  and  $\delta_i$  takes the value 1 at the  $(i + 1)$ th position and the value 0 everywhere else. Then

$$(1 - \pi_0)\vec{m} + \pi_0\vec{e} = e_0[(1 - \pi_0)\vec{m} + \pi_0\vec{\delta}_0] + \cdots + e_T[(1 - \pi_0)\vec{m} + \pi_0\vec{\delta}_T],$$

which belongs to a simplex.  $\square$

**Proposition 9.** We have

$$\pi^*(\tau, m) = \sup_t \left\{ 1 - \frac{\tau(t)}{m(t)} \right\} = 1 - \inf_t \left\{ \frac{\tau(t)}{m(t)} \right\}.$$

**Proof.** Define

$$\lambda = 1 - \inf_t \left\{ \frac{\tau(t)}{m(t)} \right\} \text{ and } \bar{\lambda} = 1 - \lambda.$$

Then

$$\begin{aligned} \tau(t) - (1 - \lambda)m(t) &= \tau(t) - \inf_t \left\{ \frac{\tau(t)}{m(t)} \right\} m(t) \\ &= m(t) \left[ \frac{\tau(t)}{m(t)} - \inf_t \left\{ \frac{\tau(t)}{m(t)} \right\} \right] \geq 0, \end{aligned}$$

with equality at some  $t$ . Let now the error term be

$$e^*(t) = \frac{1}{\lambda} [\tau(t) - \bar{\lambda}m(t)].$$

Then  $\tau(t) = (1 - \lambda)m(t) + \lambda e^*(t)$  and  $\lambda$  cannot be made smaller without making  $e^*(t)$  negative at a point  $t_0$ . This concludes the proof.  $\square$

**Corollary 2.** We have

$$\pi^*(\tau, m) = 1$$

if there exists  $t_0$  such that  $\tau(t_0) = 0$  and  $m(t_0) > 0$ .

**Proof.** By Proposition 9  $\pi^* \leq 1$ , but it equals 1 at  $t_0$ .  $\square$

One of the advantages of the mixture index of fit is that it has an intuitive interpretation that does not depend upon the specific nature of the model being assessed. Liu and Lindsay [9] extended the results of Rudas et al. [8] to the Kullback-Leibler distance. Computational aspects of the mixture index of fit are discussed in Xi and Lindsay [4] as well as in Dayton [10] and Ispány and Verdes [11].

Finally, a new interpretation to the mixture index of fit was presented by Ispány and Verdes [11]. Let  $\mathcal{P}$  be the set of probability measures and  $\mathcal{H} \subset \mathcal{P}$ . If  $d$  is a distance measure on  $\mathcal{P}$  and  $N(\mathcal{H}, \pi) = \{Q : Q = (1 - \pi)M + \pi R, M \in \mathcal{H}, R \in \mathcal{P}\}$ , then  $\pi^* = \pi^*(\mathcal{P}, \mathcal{H})$  is the least non-negative solution of the equation  $d(\mathcal{P}, N(\mathcal{H}, \pi)) := \min_{Q \in N(\mathcal{H}, \pi)} d(P, Q) = 0$  in  $\pi$ .

Next, we offer some interpretations associated with the mixture index of fit. The statistical interpretations made with this measure are attractive, as any statement based on the model applies to at least  $1 - \pi^*$  of the population involved. However, while the “outlier” model seems interpretable and attractive, the distance itself is not very robust.

In other words, small changes in the probability mass function do not necessarily mean small changes in distance. This is because if  $m(t_0) = \varepsilon$ , then a change of  $\varepsilon$  in  $\tau(t_0)$  from  $\varepsilon$  to 0 causes  $\pi^*(\tau, m)$  to go to 1. Moreover, assume that our framework is that of continuous probability measures, and that our model is a normal density. If  $\tau(t)$  is a lighter tailed distribution than our normal model  $m(t)$ , then

$$\lim_{t \rightarrow \infty} \left\{ 1 - \frac{\tau(t)}{m(t)} \right\} = 1,$$

and therefore

$$\pi^*(\tau, m) = \sup_t \left\{ 1 - \frac{\tau(t)}{m(t)} \right\} = 1.$$

That is, *light tailed densities* are interpreted as 100% outliers. Therefore, the mixture index of fit measures error from the model in a “one-sided” way. This is in contrast to total variation, which measures the size of “holes” as well as the “outliers” by allowing the distributional errors to be neutral.

In what follows, we show that if we can find a mixture representation for the true distribution then this implies a small total variation distance between the true probability mass function and the assumed model  $m$ . Specifically, we have the following.

**Proposition 10.** Let  $\pi^*$  be the mixture index of fit. If  $\tau(t) = (1 - \pi)m(t) + \pi e(t)$ , then

$$V(\tau, m) \leq \pi^*.$$

**Proof.** Write

$$\begin{aligned} V(\tau, m) &= \frac{1}{2} \sum |(1 - \pi)m(t) + \pi e(t) - m(t)| \\ &= \frac{1}{2} \sum |\pi(e(t) - m(t))| \\ &= \frac{1}{2} \sum \pi^* |e(t) - m(t)|, \end{aligned}$$

with  $\pi = \pi^*$ . This is because there always exists the smallest unique  $\pi$  such that  $\tau(t)$  can be represented as a mixture model.

Thus, the above relationship can be written as

$$V(\tau, m) = \frac{1}{2} \pi^* \sum |e(t) - m(t)| = \pi^* V(e, m) \leq \pi^*.$$

□

There is a mixture representation that connects total variation with the mixture index of fit. This is presented below.

**Proposition 11.** Denote by

$$W(\tau, m) = \inf_{\pi} \{ \pi : (1 - \pi)\tau(t) + \pi e_1(t) = (1 - \pi)m(t) + \pi e_2(t) \}.$$

Then

$$W(\tau, m) = \frac{V(\tau, m)}{1 + V(\tau, m)}.$$

**Proof.** Fix  $\tau$ ; for any given  $m$  let  $(e_1, e_2, \tilde{\pi})$  be a solution to the equation

$$\tilde{\pi}\tau + (1 - \tilde{\pi})e_1 = \tilde{\pi}e_i + (1 - \tilde{\pi})e_{2i}, i = 1, 2, \dots, T. \quad (1)$$

Let  $q_{1i} = (1 - \tilde{\pi})e_{1i}$  and  $q_{2i} = (1 - \tilde{\pi})e_{2i}$  and note that since

$$\sum e_{1i} = \sum e_{2i} = 1$$

then

$$\sum q_{1i} = \sum q_{2i} = 1 - \tilde{\pi}.$$

Rewrite now Equation (1) as follows:

$$\begin{aligned} \tilde{\pi}\tau_i + q_{1i} &= \tilde{\pi}m_i + q_{2i} \\ \Rightarrow q_{2i} - q_{1i} &= \tilde{\pi}(\tau_i - m_i) \\ \Rightarrow q_{2i} - q_{1i} &= \tilde{\pi}(\tau_i - m_i)^+ - \tilde{\pi}(\tau_i - m_i)^-, \end{aligned}$$

where  $(x)^+ = \max(x, 0)$  and  $(x)^- = -\min(x, 0)$ . Thus, ignoring the constraints, every pair  $(e_{1i}, e_{2i})$  satisfying the equation above also satisfies

$$\begin{aligned} q_{1i} &= \tilde{\pi}(\tau_i - m_i)^- + \varepsilon_i, \\ q_{2i} &= \tilde{\pi}(\tau_i - m_i)^+ + \varepsilon_i, \end{aligned}$$

for some number  $\varepsilon_i$ . Moreover, such pair must have  $\varepsilon_i \geq 0$  in order the constraints  $q_{1i} \geq 0, q_{2i} \geq 0$  to be satisfied. Hence, varying  $\varepsilon_i$  over  $\varepsilon_i \geq 0$  gives a class of solutions. To determine  $\tilde{\pi}$ ,

$$\begin{aligned}\sum_i q_{1i} &= \sum_i (\tilde{\pi}(\tau_i - m_i)^- + \varepsilon_i) = 1 - \tilde{\pi}, \\ \sum_i q_{2i} &= \sum_i (\tilde{\pi}(\tau_i - m_i)^+ + \varepsilon_i) = 1 - \tilde{\pi},\end{aligned}$$

and adding these we obtain

$$\begin{aligned}2(1 - \tilde{\pi}) &= \tilde{\pi} \sum |\tau_i - m_i| + 2 \sum \varepsilon_i \\ \Rightarrow 2 &= \tilde{\pi}(2 + \sum |\tau_i - m_i|) + 2 \sum \varepsilon_i \\ \Rightarrow 2 - 2 \sum \varepsilon_i &= \tilde{\pi}(2 + \sum |\tau_i - m_i|) \\ \Rightarrow \tilde{\pi} &= \frac{2 - 2 \sum \varepsilon_i}{2 + \sum |\tau_i - m_i|},\end{aligned}$$

and the maximum value is obtained when  $\sum \varepsilon_i = 0 \Rightarrow \varepsilon_i = 0, \forall i$ . Therefore

$$\tilde{\pi} = \frac{2}{2 + \sum |\tau_i - m_i|} = \frac{1}{1 + \frac{1}{2} \sum |\tau_i - m_i|} = \frac{1}{1 + V(\tau, m)}$$

and so

$$W(\tau, m) = \frac{V(\tau, m)}{1 + V(\tau, m)}.$$

□

Therefore, for small  $V(\tau, m)$  the mixture index of fit and the total variation distance are nearly equal.

## 5. Kullback-Leibler Distance

The Kullback-Leibler distance [12] is extensively used in statistics and in particular in model selection. The celebrated AIC model selection criterion [13] is based on this distance. In this section, we present the Kullback-Leibler distance and some of its properties with particular emphasis on interpretations.

**Definition 9.** *The Kullback-Leibler distance between two densities  $\tau, m$  is defined as*

$$K^2(\tau, m) = \sum m(t) \log \left( \frac{m(t)}{\tau(t)} \right),$$

or

$$K^2(\tau, m) = \int m(t) \log \left( \frac{m(t)}{\tau(t)} \right) dt.$$

**Proposition 12.** *The Kullback-Leibler distance is nonnegative, that is*

$$K^2(\tau, m) \geq 0$$

with equality if and only if  $\tau(t) = m(t)$ .

**Proof.** Write

$$K^2(\tau, m) = \sum m(t) \left[ \log \left( \frac{m(t)}{\tau(t)} \right) + \frac{\tau(t)}{m(t)} - 1 \right] = \sum m(t) \left[ -\log \left( \frac{\tau(t)}{m(t)} \right) + \frac{\tau(t)}{m(t)} - 1 \right].$$

Set  $X = \frac{\tau(t)}{m(t)} \geq 0$ , then  $-\log X + X - 1$  is a convex, non-negative function that equals 0 at  $X = 1$ . Therefore  $K^2(\tau, m) \geq 0$ .  $\square$

**Definition 10.** We define the likelihood distance between two densities  $\tau, m$  as

$$\lambda^2(\tau, m) = \sum \tau(t) \log \left( \frac{\tau(t)}{m(t)} \right).$$

The intuition behind the above expression of the likelihood distance comes from the fact that the log-likelihood in the case of discrete random variables taking  $n_j$  discrete values,  $\sum_{j=1}^m n_j = n$ ,  $m$  is the number of groups, can be written, after appropriate algebraic manipulations, in the above form.

Alternatively, we can write the likelihood distance as

$$\lambda^2(\tau, m) = \sum m(t) \left[ \frac{\tau(t)}{m(t)} \log \left( \frac{\tau(t)}{m(t)} \right) - \frac{\tau(t)}{m(t)} + 1 \right],$$

and use this relationship to obtain insight into connections of the likelihood distance with the chi-squared measures studied by Markatou et al. [3].

Specifically, if we write the Pearson's chi-squared statistic as

$$P^2(\tau, m) = \sum m(t) \left[ \frac{\tau(t)}{m(t)} - 1 \right]^2,$$

then from the functional relationship  $r \log r - r + 1 \leq (r - 1)^2$  we obtain that  $\lambda^2(\tau, m) \leq P^2(\tau, m)$ . However, it is also clear from the right tails of the functions that there is no way to bound  $\lambda^2(\tau, m)$  below by a multiple of  $P^2(\tau, m)$ . Hence, these measures are not equivalent in the same way that Hellinger distance and symmetric chi-squared are (see Lemma 4, Markatou et al. [3]). In particular, knowing that  $\lambda^2(\tau, m)$  is small is no guarantee that all Pearson z-statistics are uniformly small.

On the other hand, one can show by the same mechanism that  $S^2 \leq 2k\lambda^2$ , where  $k < 32/9$  and  $S^2$  is the symmetric chi-squared distance given as

$$S^2(\tau, m) = \sum \frac{(\tau(t) - m(t))^2}{\frac{1}{2}\tau(t) + \frac{1}{2}m(t)}.$$

It is therefore true that small likelihood distance  $\lambda^2$  implies small z-statistics with blended variance estimators. However, the reverse is not true because the right tail in  $r$  for  $S^2$  is of magnitude  $r$ , as opposed to  $r \log r$  for the likelihood distance.

These comparisons provide some feeling for the statistical interpretation of the likelihood distance. Its meaning as a measure of model misspecification is unclear. Furthermore, our impression is that likelihood, like Pearson's chi-squared is too sensitive to outliers and gross errors in the data. Despite Kullback-Leibler's theoretical and computational advantages, a point of inconvenience in the context of model selection is the lack of symmetry. One can show that reversing the roles of the arguments in the Kullback-Leibler divergence can yield substantially different results. The sum of the Kullback-Leibler distance and the likelihood distance produces the symmetric Kullback-Leibler distance or J divergence. This measure is symmetric in the arguments, and when used as a model selection measure it is expected to be more sensitive than each of the individual components.

## 6. Computation and Applications of Total Variation, Mixture Index of Fit and Kullback-Leibler Distances

The distances discussed in this paper are used in a number of important applications. Euán et al. [14] use the total variation to detect changes in wave spectra, while Alvarez- Esteban et al. [15] cluster time series data on the basis of the total variation distance. The mixture index of fit has found a number of

applications in the area of social sciences. Rudas et al. [8] provided examples of the application of  $\pi^*$  to two-way contingency tables. Applications involving differential item functioning and latent class analysis were presented in Rudas and Zwick [16] and Dayton [17] respectively. Formann [18] applied it in regression models involving continuous variables. Finally, Revuelta [19] applied the  $\pi^*$  goodness-of-fit statistic to finite mixture item response models that were developed mainly in connection with Rasch models [20,21]. The Kullback-Leibler (KL) distance [12] is fundamental in information theory and its applications. In statistics, the celebrated Akaike information Criterion (AIC) [13,22], widely used in model selection, is based on the Kullback-Leibler distance. There are numerous additional applications of the KL distance in fields such as fluid mechanics, neuroscience, machine learning. In economics, Smith, Naik, and Tsai [23] use KL distance to simultaneously select the number of states and variables associated with Markov-switching regression models that are used in marketing and other business applications. KL distance is also used in diagnostic testing for ruling in or ruling out disease [24,25], as well as in a variety of other fields [26].

Table 1 presents the software, written in R, that can be used to compute the aforementioned distances. Additionally, Zhang and Dayton [27] present a SAS program to compute the two-point mixture index of fit for the two-class latent class analysis models with dichotomous variables. There are a number of different algorithms that can be used to compute the mixture index of fit for contingency tables. Rudas et al. [8] propose to use a standard EM algorithm, Xi and Lindsay [4] use sequential quadratic programming and discuss technical details and numerical issues related to applying nonlinear programming techniques to estimate  $\pi^*$ . Dayton [10] discusses explicitly the practical advantages associated with the use of nonlinear programming as well as the limitations, while Pan and Dayton [28] study a variety of additional issues associated with computing  $\pi^*$ . Additional algorithms associated with the computation of  $\pi^*$  can be found in Verdes [29] and Ispány and Verdes [11].

We now describe a simulation study that aims to illustrate the performance of the total variation, Kullback-Leibler, and mixture index of fit as model selection measures. Data are generated from either an asymmetric  $(1 - \varepsilon)N(0, 1) + \varepsilon N(\mu, \sigma^2)$  contamination model, or from a symmetric  $(1 - \varepsilon)N(0, 1) + \varepsilon N(0, \sigma^2)$  contamination model, where  $\varepsilon$  is the percentage of contamination. Specifically, we generate 500 Monte Carlo samples of sample sizes 200, 1000, and 5000 as follows. If the sample has size  $n$  and the percentage of contamination is  $\varepsilon$ , then  $n\varepsilon$  of the sample size is generated from model  $N(\mu, \sigma^2)$  or  $N(0, \sigma^2)$  and the remaining  $n(1 - \varepsilon)$  from a  $N(0, 1)$  model. We use  $\mu = 1, 5, 10$  and  $\sigma^2 = 1$  in the  $N(\mu, \sigma^2)$  model and  $\sigma^2 = 4, 9, 16$  in the  $N(0, \sigma^2)$  model. The total variation distance was computed between the simulated data and the  $N(0, 1)$  model. The Kullback-Leibler distance was calculated between the data generated from the aforementioned contamination models and a random sample of the same size  $n$  from  $N(0, 1)$ . When computing the mixture index of fit, we specified the component distribution as a normal distribution with initial mean 0 and variance 1. All simulations were carried out on a laptop computer with an Intel Core i7 processor and 64 bit Windows 7 operation system. The R packages used are presented in Table 1.

Tables 2 and 3 present means and standard deviations of the total variation and Kullback-Leibler distances as a function of the contamination model and the sample size. To compute the total variation distance we use the R function “TotalVarDist” of the R package “distrEx”. It smooths the empirical distribution of the provided data using a normal kernel and computes the distance between the smoothed empirical distribution and the provided continuous distribution (in our case this distribution is  $N(0, 1)$ ). We note here that the package “distrEx” provides an alternative option to compute the total variation which relies on discretizing the continuous distribution and then computes the distance between the discretized continuous distribution and the data. We think that smoothing the data to obtain an empirical estimator of the density and then calculating its distance from the continuous density is a more natural way to handle the difference in scale between the discrete data and the continuous model. Lindsay [1] and Markatou et al. [3] discuss this phenomenon and

call it discretization robustness. The Kullback-Leibler distance was computed using the function “KLD.matrix” of the R package “bioDist”.

**Table 1.** Computer packages for calculating total variation, mixture index of fit, and Kullback-Leibler distances.

Information	Total Variation	Kullback-Leibler	Mixture Index of Fit
R package	distrEx	bioDist	pistar
R function	TotalVarDist	KLD.matrix	pistar.uv
Dimension	Univariate	Univariate	Univariate
Website	<a href="https://cran.r-project.org/web/packages/distrEx/">https://cran.r-project.org/web/packages/distrEx/</a>	<a href="http://bioconductor.org/packages/release/bioc/html/bioDist.html">http://bioconductor.org/packages/release/bioc/html/bioDist.html</a>	<a href="https://rdrr.io/github/jmedzihorsky/pistar/man/">https://rdrr.io/github/jmedzihorsky/pistar/man/</a>

**Table 2.** Means and standard deviations (SD) of the total variation (TV) and Kullback-Leibler (KLD) distances. Data are generated from the model  $(1 - \varepsilon)N(0, 1) + \varepsilon N(\mu, 1)$  with  $\mu = 1, 5, 10$ . The sample size  $n$  is 200, 1000, 5000. The number of Monte Carlo replications is 500.

Contaminating Model	Percentage of Contamination ( $\varepsilon$ )	Summary	$n = 200$		$n = 1000$		$n = 5000$	
			TV	KLD	TV	KLD	TV	KLD
N(1, 1)	0.01	Mean	0.144	0.224	0.065	0.048	0.029	0.008
		SD	0.017	0.244	0.007	0.051	0.004	0.009
	0.05	Mean	0.146	0.255	0.069	0.065	0.034	0.017
		SD	0.017	0.267	0.009	0.059	0.004	0.015
	0.1	Mean	0.149	0.323	0.076	0.088	0.047	0.026
		SD	0.017	0.343	0.009	0.073	0.005	0.018
	0.2	Mean	0.162	0.482	0.097	0.147	0.081	0.059
		SD	0.020	0.462	0.011	0.123	0.006	0.030
	0.3	Mean	0.181	0.616	0.128	0.215	0.117	0.102
		SD	0.022	0.528	0.013	0.150	0.007	0.044
N(5, 1)	0.4	Mean	0.201	0.733	0.162	0.293	0.155	0.153
		SD	0.024	0.616	0.014	0.176	0.007	0.058
	0.5	Mean	0.232	0.937	0.198	0.392	0.192	0.207
		SD	0.026	0.735	0.014	0.203	0.007	0.067
	0.01	Mean	0.149	0.577	0.070	0.338	0.034	0.231
		SD	0.017	0.373	0.008	0.131	0.004	0.063
	0.05	Mean	0.167	1.416	0.092	1.041	0.060	0.838
		SD	0.020	0.499	0.009	0.248	0.004	0.138
	0.1	Mean	0.196	2.392	0.126	2.002	0.103	1.731
		SD	0.020	0.609	0.010	0.335	0.004	0.219
N(10, 1)	0.2	Mean	0.259	4.841	0.210	4.404	0.199	3.947
		SD	0.023	0.941	0.012	0.512	0.006	0.383
	0.3	Mean	0.336	7.924	0.302	7.305	0.297	6.652
		SD	0.028	1.182	0.014	0.730	0.007	0.569
	0.4	Mean	0.419	11.317	0.398	10.655	0.396	9.843
		SD	0.031	1.388	0.016	0.863	0.006	0.792
	0.5	Mean	0.506	15.045	0.495	14.443	0.494	13.573
		SD	0.035	1.768	0.016	1.027	0.007	0.999

**Table 3.** Means and standard deviations (SD) of the total variation (TV) and Kullback-Leibler (KLD) distances. Data are generated from the model  $(1 - \varepsilon)N(0, 1) + \varepsilon N(0, \sigma^2)$  with  $\sigma^2 = 4, 9, 16$ . The sample size  $n$  is 200, 1000, 5000. The number of Monte Carlo replications is 500.

Contaminating Model	Percentage of Contamination ( $\varepsilon$ )	Summary	$n = 200$		$n = 1000$		$n = 5000$	
			TV	KLD	TV	KLD	TV	KLD
$N(0, 4)$	0.01	Mean	0.145	0.263	0.066	0.068	0.030	0.021
		SD	0.017	0.250	0.008	0.058	0.003	0.014
	0.05	Mean	0.147	0.497	0.069	0.204	0.034	0.079
		SD	0.017	0.391	0.008	0.130	0.004	0.036
	0.1	Mean	0.154	0.778	0.076	0.368	0.044	0.181
		SD	0.018	0.527	0.008	0.168	0.004	0.062
	0.2	Mean	0.166	1.275	0.094	0.712	0.071	0.426
		SD	0.020	0.639	0.010	0.255	0.005	0.108
	0.3	Mean	0.182	1.797	0.118	1.067	0.101	0.671
		SD	0.021	0.738	0.012	0.324	0.006	0.158
$N(0, 9)$	0.4	Mean	0.201	2.320	0.144	1.407	0.133	0.924
		SD	0.021	0.875	0.012	0.403	0.006	0.198
	0.5	Mean	0.220	2.766	0.173	1.755	0.164	1.164
		SD	0.025	0.932	0.013	0.450	0.006	0.219
	0.01	Mean	0.146	0.369	0.067	0.122	0.031	0.046
		SD	0.018	0.348	0.007	0.089	0.003	0.022
	0.05	Mean	0.154	0.839	0.074	0.490	0.040	0.321
		SD	0.017	0.477	0.008	0.187	0.004	0.081
	0.1	Mean	0.164	1.414	0.087	0.945	0.058	0.661
		SD	0.018	0.602	0.009	0.256	0.005	0.120
$N(0, 16)$	0.2	Mean	0.189	2.529	0.120	1.748	0.101	1.300
		SD	0.021	0.801	0.011	0.366	0.005	0.188
	0.3	Mean	0.216	3.529	0.161	2.526	0.149	1.954
		SD	0.023	0.957	0.012	0.466	0.006	0.276
	0.4	Mean	0.252	4.608	0.205	3.444	0.196	2.660
		SD	0.026	1.071	0.014	0.549	0.006	0.339
	0.5	Mean	0.286	5.630	0.250	4.289	0.244	3.423
		SD	0.026	1.123	0.014	0.657	0.007	0.406

We observe from the results of Tables 2 and 3 that the total variation distance for small percentages of contamination is small and generally smaller than the Kullback-Leibler distance for both asymmetric and symmetric contamination models with a considerably smaller standard deviation. The above behavior of the total variation distance in comparison to the Kullback-Leibler manifests itself across all sample sizes used.

Table 4 presents the mixture index of fit computed using the R function “pistar.uv” from the R package “pistar” (<https://rdrr.io/github/jmedzihorsky/pistar/man/>; accessed on 5 June 2018). Since the fundamental assumption in the definition of the mixture index of fit is that the population on

which the index is applied is heterogeneous and expressed via the two-point model, we only used the asymmetric contamination model for various values of the contamination distribution.

**Table 4.** Means and standard deviations (SD) for the mixture index of fit. Data are generated from an asymmetric contamination model of the form  $(1 - \varepsilon)N(0, 1) + \varepsilon N(\mu, 1)$ ,  $\mu = 1, 5, 10$  with sample sizes,  $n$ , of 1000, 5000. The number of Monte Carlo replications is 500.

Percentage of Contamination $\varepsilon$	Summary	$N(1, 1)$		$N(5, 1)$		$N(10, 1)$	
		$n = 1000$	$n = 5000$	$n = 1000$	$n = 5000$	$n = 1000$	$n = 5000$
0.1	Mean	0.180	0.160	0.223	0.213	0.837	0.934
	SD	0.045	0.044	0.041	0.040	0.279	0.198
0.2	Mean	0.184	0.172	0.288	0.287	0.433	0.521
	SD	0.044	0.042	0.036	0.036	0.144	0.240
0.3	Mean	0.189	0.179	0.344	0.346	0.314	0.317
	SD	0.047	0.039	0.028	0.024	0.016	0.012
0.4	Mean	0.194	0.186	0.436	0.436	0.410	0.413
	SD	0.044	0.034	0.026	0.021	0.017	0.011
0.5	Mean	0.194	0.185	0.529	0.533	0.511	0.512
	SD	0.047	0.035	0.024	0.020	0.017	0.010

We observe that the mixture index of fit generally estimates well the mixing proportion  $\varepsilon$ . We observe (see Table 4) that when the second population is  $N(1, 1)$  the bias associated with estimating the mixing (or contamination) population can be as high as 30.6%. This is expected because the population  $N(1, 1)$  is very close to  $N(0, 1)$  creating essentially a unimodal sample. As the means of the two normal components get more separated, the mixture index of fit provides better estimates of the mixing quantity and the percentage of observations that need to be removed so that  $N(0, 1)$  provides a good fit to the remaining data points.

## 7. Discussion and Conclusions

Divergence measures are widely used in scientific work, and popular examples of these measures include the Kullback-Leibler divergence, Bregman Divergence [30], the power divergence family of Cressie and Read [31], the density power divergence family [32] and many others. Two relatively recent books that discuss various families of divergences are Pardo [33] and Basu et al. [34].

In this paper we discuss specific divergences that do not belong to the family of quadratic divergences, and examine their role in assessing model adequacy. The total variation distance might be preferable as it seems closest to a robust measure, in that if the two probability measures differ only on a set of small probability, such as a few outliers, then the distance must be small. This was clearly exemplified in Tables 2 and 3 of Section 6. Outliers influence chi-squared measures more. For example, the Pearson's chi-squared distance can be made dramatically larger by increasing the amount of data in a cell with small model probability  $m_\theta(t)$ . In fact, if there is data in a cell with model probability zero, the distance is infinite. Note that if data occur in a cell with probability, under the model, equal to zero, then it is possible that the model is not true. Still, even in this case, we might wish to use it on the premise that  $m_\theta$  provides a good approximation.

There is a pressing need for the further development of well-tested software for computing the mixture index of fit. This measure is intuitive and has found many applications in the social sciences. Reiczigel et al. [35] discuss bias-corrected point estimates of  $\pi^*$ , as well as a bootstrap test and new confidence limits, in the context of contingency tables. Well-developed and tested software will further popularize the dissemination and use of this method.

The mixture index of fit ideas were extended in the context of testing general model adequacy problems by Liu and Lindsay [9]. Recent work by Ghosh and Basu [36] presents a systematic procedure of generating new divergences. Ghosh and Basu [36], building upon the work of Liu and Lindsay [9], generate new divergences through suitable model adequacy tests using existing divergences. Additionally,

Dimova et al. [37] use the quadratic divergences introduced in Lindsay et al. [2] and construct a model selection criterion from which we can obtain AIC and BIC as special cases.

In this paper, we discuss non-quadratic distances that are used in many scientific fields where the problem of assessing the fitted models is of importance. In particular, our interest centered around the properties and potential interpretations of these distances, as we think this offers insight into their performance as measures of model misspecification. One important aspect for the dissemination and use of these distances is the existence of well-tested software that facilitates computation. This is an area where further development is required.

**Author Contributions:** M.M. developed the ideas and wrote the paper. Y.C. contributed to the proofs and simulations presented.

**Funding:** This research received no external funding.

**Acknowledgments:** The first author would like to acknowledge the many discussions and contributions towards the formation of these ideas by the late Bruce G. Lindsay, to whom this paper is dedicated.

**Conflicts of Interest:** The authors have no conflicts of interest.

## References

1. Lindsay, B.G. Statistical distances as loss functions in assessing model adequacy. In *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*; Taper, M.L., Lele, S.R., Eds.; The University of Chicago Press: Chicago, IL, USA, 2004; pp. 439–488.
2. Lindsay, B.G.; Markatou, M.; Ray, S.; Yang, K.; Chen, S.C. Quadratic distances on probabilities: A unified foundation. *Ann. Stat.* **2008**, *36*, 983–1006. [[CrossRef](#)]
3. Markatou, M.; Chen, Y.; Afendras, G.; Lindsay, B.G. Statistical distances and their role in robustness. In *New Advances in Statistics and Data Science*; Chen, D.G., Jin, Z., Li, G., Li, Y., Liu, A., Zhao, Y., Eds.; Springer: New York, NY, USA, 2017; pp. 3–26.
4. Xi, L.; Lindsay, B.G. A note on calculating the  $\pi^*$  index of fit for the analysis of contingency tables. *Sociol. Methods Res.* **1996**, *25*, 248–259. [[CrossRef](#)]
5. Donoho, D.L.; Liu, R.C. Pathologies of some minimum distance estimators. *Ann. Stat.* **1988**, *16*, 587–608. [[CrossRef](#)]
6. Matusita, K. On the theory of statistical decision functions. *Ann. Inst. Stat. Math.* **1951**, *3*, 17–35. [[CrossRef](#)]
7. Matusita, K. Decision rules, based on the distance, for problems of fit, two samples, and estimation. *Ann. Math. Stat.* **1955**, *26*, 631–640. [[CrossRef](#)]
8. Rudas, T.; Clogg, C.C.; Lindsay, B.G. A new index of fit based on mixture methods for the analysis of contingency tables. *J. Royal Stat. Soc. Series B* **1994**, *56*, 623–639.
9. Liu, J.; Lindsay, B.G. Building and using semiparametric tolerance regions for parametric multinomial models. *Ann. Stat.* **2009**, *37*, 3644–3659. [[CrossRef](#)]
10. Dayton, C.M. Applications and computational strategies for the two-point mixture index of fit. *Br. J. Math. Stat. Psychol.* **2003**, *56*, 1–13. [[CrossRef](#)] [[PubMed](#)]
11. Ispány, M.; Verdes, E. On the robustness of mixture index of fit. *J. Math. Sci.* **2014**, *200*, 432–440. [[CrossRef](#)]
12. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
13. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **1974**, *19*, 716–723. [[CrossRef](#)]
14. Euán, C.; Ortega, J.; Esteban, P.C.A. Detecting Changes in Wave Spectra Using the Total Variation Distance. In Proceedings of the 23rd International Offshore and Polar Engineering Conference. International Society of Offshore and Polar Engineers, Anchorage, AK, USA, 30 June–5 July 2013.
15. Alvarez-Esteban, P.C.; Euán, C.; Ortega, J. Time series clustering using the total variation distance with applications in oceanography. *Environmetrics* **2016**, *27*, 355–369. [[CrossRef](#)]
16. Rudas, T.; Zwick, R. Estimating the importance of differential item functioning. *J. Educ. Behav. Stat.* **1997**, *22*, 31–45. [[CrossRef](#)]
17. Dayton, M.C. *Latent Class Scaling Analysis*; Sage: Thousand Oaks, CA, USA, 1999.
18. Formann, A.K. Testing the Rasch model by means of the mixture fit index. *Br. J. Math. Stat. Psychol.* **2006**, *59*, 89–95. [[CrossRef](#)] [[PubMed](#)]

19. Revuelta, J. Estimating the  $\pi^*$  goodness of fit index for finite mixtures of item response models. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 93–113. [[CrossRef](#)] [[PubMed](#)]
20. Rost, J. Rasch models in latent classes: An integration of two approaches to item analysis. *Appl. Psychol. Meas.* **1990**, *14*, 271–282. [[CrossRef](#)]
21. Rost, J. A logistic mixture distribution model for polychotomous item responses. *Br. J. Math. Stat. Psychol.* **1991**, *44*, 75–92. [[CrossRef](#)]
22. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: New York, NY, USA, 2002.
23. Smith, A.; Naik, P.A.; Tsai, C.L. Markov-switching model selection using Kullback–Leibler divergence. *J. Econom.* **2006**, *134*, 553–577. [[CrossRef](#)]
24. Lee, W.C. Selecting diagnostic tests for ruling out or ruling in disease: the use of the Kullback–Leibler distance. *Int. J. Epidemiol.* **1999**, *28*, 521–523. [[CrossRef](#)] [[PubMed](#)]
25. Grimes, D.A.; Schulz, K.F. Refining clinical diagnosis with likelihood ratios. *Lancet* **2005**, *365*, 1500–1505. [[CrossRef](#)]
26. Cliff, O.M.; Prokopenko, M.; Fitch, R. Minimising the Kullback–Leibler Divergence for Model Selection in Distributed Nonlinear Systems. *Entropy* **2018**, *20*, 51. [[CrossRef](#)]
27. Zhang, D.; Dayton, C.M. JMASM30 PI-LCA: A SAS program computing the two-point mixture index of fit for two-class LCA Models with dichotomous variables (SAS). *J. Mod. Appl. Stat. Methods* **2010**, *9*, 314–331. [[CrossRef](#)]
28. Pan, X.; Dayton, C.M. Factors influencing the mixture index of model fit in contingency tables showing independence. *J. Mod. Appl. Stat. Methods* **2011**, *10*, 314–331. [[CrossRef](#)]
29. Verdes, E. Finding and characterization of local optima in the  $\pi^*$  problem for two-way contingency tables. *Stud. Sci. Math. Hung.* **2000**, *36*, 471–480.
30. Bregman, L.M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217. [[CrossRef](#)]
31. Cressie, N.; Read, T.R. Multinomial goodness-of-fit tests. *J. Royal Stat. Soc. Series B* **1984**, *46*, 440–464.
32. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559. [[CrossRef](#)]
33. Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2006.
34. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2011.
35. Reiczigel, J.; Ispány, M.; Tusnády, G.; Michaletzky, G.; Marozzi, M. Bias-corrected estimation of the Rudas–Clogg–Lindsay mixture index of fit. *Br. J. Math. Stat. Psychol.* **2017**. [[CrossRef](#)] [[PubMed](#)]
36. Ghosh, A.; Basu, A. A new family of divergences originating from model adequacy tests and application to robust statistical inference. *IEEE Trans. Inf. Theory* **2018**. [[CrossRef](#)]
37. Dimova, R.; Markatou, M.; Afendras, G. *Model Selection Based on the Relative Quadratic Risk*; Technical Report; Department of Biostatistics, University at Buffalo: Buffalo, NY, USA, 2018.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Review

# $\phi$ -Divergence in Contingency Table Analysis

Maria Kateri

Institute of Statistics, RWTH Aachen University, 52062 Aachen, Germany; maria.kateri@rwth-aachen.de

Received: 6 April 2018; Accepted: 20 April 2018; Published: 27 April 2018



**Abstract:** The  $\phi$ -divergence association models for two-way contingency tables is a family of models that includes the association and correlation models as special cases. We present this family of models, discussing its features and demonstrating the role of  $\phi$ -divergence in building this family. The most parsimonious member of this family, the model of  $\phi$ -scaled uniform local association, is considered in detail. It is implemented and representative examples are commented on.

**Keywords:** log-linear models; ordinal classification variables; association models; correlation models

---

## 1. Introduction

Contingency tables and their analysis are of special importance for various diverse fields, like medical sciences, psychology, education, demography and social sciences. In these fields, categorical variables and their cross-classification play a predominant role, since characteristics of interest are often categorical, nominal or most frequently ordinal. For example, diagnostic ratings, strength of opinions or preferences, educational attainments and socioeconomic characteristics are expressed in ordinal scales. The origins of contingency table analysis (CTA) lie back in 1900 with the well-known contributions by Pearson and Yule, while, for the history of CTA before 1900, we refer to the interesting paper by Stigler [1]. The interest lies mainly in identifying and describing structures of underlying association in terms of appropriate models or measures. Divergence measures have been employed in the CTA mainly for hypothesis testing (model fit) and estimation, leading to general families of test statistics and estimators. The family of  $\phi$ -divergence test statistics contain the classical likelihood ratio and Pearson test statistics as special cases while the maximum likelihood estimators (MLEs) belong to the family of the minimum  $\phi$ -divergence estimators ( $M\phi$ Es). Families of  $\phi$ -divergence based test statistics as well as  $M\phi$ Es for various standard models in CTA have a long history. However, their consideration and discussion is out of our scope. For a detailed overview and related references, we refer to the comprehensive book of Pardo [2]. For log-linear models, see also [3].

Here, we aim at highlighting a different structural role of  $\phi$ -divergence in contingency tables modelling, namely that of linking phenomenological different models, forming thus a family of models and providing a basis for their comparison, understanding and unified treatment. Through this approach, new insight is gained for the standard association and correlation models (see [4,5]) while further alternatives are considered. We restrict our discussion on two-dimensional contingency tables, but the models and approaches discussed are directly extendable to tables of higher dimension.

The organization of the paper is as follows. Preliminaries on log-linear models, divergence measures, association and correlation models for two-way tables are provided in Section 2. In the sequel, the general family of  $\phi$ -divergence based association models (AMs), which includes the classical association and correlation models as special cases, is reviewed in Section 3. The most parsimonious  $\phi$ -divergence based association model, that of  $\phi$ -scaled uniform local association, and its role in conditional testing of independence is considered and discussed in Section 4. For this family of models, the effect of the specific  $\phi$ -function used is illustrated by analysing representative examples in Section 5. Some final comments are provided in Section 6.

## 2. Preliminaries

Consider an  $I \times J$  contingency table  $\mathbf{n} = (n_{ij})$  with rows and columns classification variables  $X$  and  $Y$ , respectively, where  $n_{ij}$  is the observed frequency in cell  $(i, j)$ . The total sample size  $n = \sum_{i,j} n_{ij}$  is fixed and the random table  $\mathbf{N}$  is multinomial distributed  $\mathbf{N} \sim \mathcal{M}(n, \boldsymbol{\pi})$ , with probability table  $\boldsymbol{\pi} \in \Delta_{IJ}$ , where  $\Delta_{IJ} = \{\boldsymbol{\pi} = (\pi_{ij}) : \pi_{ij} > 0, \sum_{i,j} \pi_{ij} = 1\}$ . Let  $\mathbf{m} = E(\mathbf{N}) = n\boldsymbol{\pi}$  be the table of expected cell frequencies. Since the mapping  $(n, \boldsymbol{\pi}) \mapsto \mathbf{m}$  is one-to-one on  $\Delta_{IJ}$ ,  $\mathbf{m} \in \{\mathbf{m} = (m_{ij}) : m_{ij} > 0, \sum_{i,j} m_{ij} = n\}$  and models (hypotheses) for  $\boldsymbol{\pi}$  can equivalently be expressed in terms of  $\mathbf{m}$ . Furthermore, let  $\boldsymbol{\pi}_r = (\pi_{1+}, \dots, \pi_{I+})^T$  and  $\boldsymbol{\pi}_c = (\pi_{+1}, \dots, \pi_{+J})^T$  be the row and column marginal probabilities vectors, respectively, and  $\mathbf{p} = (p_{ij})$  the table of sample proportions with  $p_{ij} = n_{ij}/n$ .

The classical independence hypothesis for the classification variables  $X$  and  $Y$  ( $\boldsymbol{\pi} = \boldsymbol{\pi}_r \boldsymbol{\pi}_c^T = \boldsymbol{\pi}^I$ ) corresponds to the log-linear model of independence (I), defined in terms of expected cell frequencies as

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y, \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad (1)$$

where  $\lambda_i^X$  and  $\lambda_j^Y$  are the  $i$ -th row and  $j$ -th column main effects, respectively, while  $\lambda$  is the intercept. If the independence model is rejected, the interaction between  $X$  and  $Y$  is significant and the only alternative in the standard log-linear models set-up is the saturated model

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad (2)$$

which imposes no structure on  $\boldsymbol{\pi}$ . Identifiability constraints need to be imposed on the main effect and interaction parameters of these models, like  $\lambda_1^X = \lambda_1^Y = 0$  and  $\lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0$ , for all  $i, j$ .

An important generalized measure for measuring the divergence between two probability distributions is the  $\phi$ -divergence. Let  $\boldsymbol{\pi} = (\pi_{ij})$ ,  $\mathbf{q} = (q_{ij}) \in \Delta_{IJ}$  be two discrete finite bivariate probability distributions. Then, the  $\phi$ -divergence between  $\mathbf{q}$  and  $\boldsymbol{\pi}$  (or Csiszar's measure of information in  $\mathbf{q}$  about  $\boldsymbol{\pi}$ ), is given by

$$I_\phi^C(\mathbf{q}, \boldsymbol{\pi}) = \sum_{i,j} \pi_{ij} \phi(q_{ij}/\pi_{ij}), \quad (3)$$

where  $\phi$  is a real-valued strictly convex function on  $[0, \infty)$  with  $\phi(1) = \phi'(1) = 0$ ,  $0\phi(0/0) = 0$ ,  $0\phi(y/0) = \lim_{x \rightarrow \infty} \phi(x)/x$  (cf. [2]). Setting  $\phi(x) = x \log x$ , (3) is reduced to the Kullback–Leibler (KL) divergence

$$I^{KL}(\mathbf{q}, \boldsymbol{\pi}) = \sum_{i,j} q_{ij} \log(q_{ij}/\pi_{ij}), \quad (4)$$

while, for  $\phi(x) = (1-x)^2$ , Pearson's divergence is derived. If  $\phi(x) = \frac{x^{\lambda+1}-x}{\lambda(\lambda+1)}$ , (3) becomes the power divergence measure of Cressie and Read [6]

$$I_\lambda^{CR}(\mathbf{q}, \boldsymbol{\pi}) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^K q_i \left[ \left( \frac{q_i}{\pi_i} \right)^\lambda - 1 \right], \quad -\infty < \lambda < \infty, \quad \lambda \neq -1, 0. \quad (5)$$

For  $\lambda \rightarrow -1$  and  $\lambda \rightarrow -0$ , (5) converges to the  $I^{KL}(\boldsymbol{\pi}, \mathbf{q})$  and  $I^{KL}(\mathbf{q}, \boldsymbol{\pi})$ , respectively, while  $\lambda = 1$  corresponds to Pearson's divergence.

The goodness of fit (GOF) of model (1) is usually tested by the likelihood ratio test statistic  $G^2 = 2nI^{KL}(\mathbf{p}, \hat{\boldsymbol{\pi}})$  or Pearson's  $X^2 = 2nI_1^{CR}(\mathbf{p}, \hat{\boldsymbol{\pi}})$ , where  $\hat{\boldsymbol{\pi}}$  is the MLE of  $\boldsymbol{\pi}$  under (1). Both test statistics are under (1) asymptotically  $\chi^2_{(I-1)(J-1)}$  distributed. Alternatively,  $\phi$ -divergence test statistics can be used (see [3]).

## 2.1. Association Models

In case of ordinal classification variables, the association models (AMs) impose a special structure on the underlying association and thus provide non-saturated models of dependence. AMs are based on scores  $\mu = (\mu_1, \dots, \mu_I)$  and  $\nu = (\nu_1, \dots, \nu_J)$  assigned to the rows and columns of the ordinal classification variables, respectively. They are defined by the expression

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \zeta \mu_i \nu_j, \quad i = 1, \dots, I, j = 1, \dots, J, \quad (6)$$

where the row and column scores are standardized subject to weights  $w_1 = (w_{11}, \dots, w_{1I})^T$  and  $w_2 = (w_{21}, \dots, w_{2J})^T$ , respectively, i.e., it holds

$$\sum_i w_{1i} \mu_i = \sum_j w_{2j} \nu_j = 0 \quad \text{and} \quad \sum_i w_{1i} \mu_i^2 = \sum_j w_{2j} \nu_j^2 = 1. \quad (7)$$

Usually, the uniform ( $w_1 = \mathbf{1}_I$ ,  $w_2 = \mathbf{1}_J$ , where  $\mathbf{1}_k$  is a  $k \times 1$  vector of 1s) or the marginal weights ( $w_1 = \pi_r$ ,  $w_2 = \pi_c$ ) are used.

If  $\mu$  and  $\nu$  are both known and ordered, then (6) has just one parameter more than independence, parameter  $\zeta$ , and is known as the *Linear-by-Linear* (LL) AM. In case the vector  $\mu$  is unknown, (6) is the *Row effect* (R) AM, while the *Column effect* (C) AM is defined analogously. Finally, when the row and the column scores are all unknown parameters to be estimated, (6) is the *multiplicative Row-Column* (RC) AM. Scores that are unknown need not necessarily to be ordered. Note that models LL, R and C are log-linear while the RC is not. The degrees of freedom ( $df$ ) of these AMs equal  $df(LL) = (I-1)(J-1)-1$ ,  $df(R) = (I-1)(J-2)$ ,  $df(C) = (I-2)(J-1)$  and  $df(RC) = (I-2)(J-2)$ . The special LL model for which the row and column scores are equidistant for successive categories is known as the *Uniform* (U) AM.

In case the RC model is not of adequate fit, multiplicative row-column AMs of higher order can be considered. Such a model of  $M$ -th order is defined as

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \sum_{m=1}^M \zeta_m \mu_{im} \nu_{jm}, \quad i = 1, \dots, I, j = 1, \dots, J, \quad (8)$$

with  $1 \leq M \leq M^* = \min(I, J) - 1$ , and denoted by  $RC(M)$ . Model  $RC(M^*)$  is an equivalent expression of the saturated model (2). The sum  $\sum_{m=1}^M \zeta_m \mu_{im} \nu_{jm}$  in (8) corresponds to the generalized singular value decomposition of the matrix of interaction parameters of model (2),  $\Lambda = (\lambda_{ij}^{XY})_{I \times J'}$  and  $M$  is the rank of  $\Lambda$ . The  $\zeta_m$ s are the associated eigenvalues, satisfying thus  $\zeta_1 \geq \dots \geq \zeta_M > 0$ . Vectors  $\mu_m = (\mu_{1m}, \dots, \mu_{Im})$  and  $\nu_m = (\nu_{1m}, \dots, \nu_{Jm})$  are the corresponding row and column eigenvectors,  $m = 1, \dots, M$ , which are orthonormalized with respect to the weights  $w_1$  and  $w_2$ , i.e., the following constraints are satisfied:

$$\begin{aligned} \sum_i w_{1i} \mu_{im} &= \sum_j w_{2j} \nu_{jm} = 0, \quad m = 1, \dots, M, \\ \sum_i w_{1i} \mu_{im} \mu_{i\ell} &= \sum_j w_{2j} \nu_{jm} \nu_{j\ell} = \delta_{m\ell}, \quad m, \ell = 1, \dots, M, \end{aligned} \quad (9)$$

where  $\delta_{m\ell}$  is Kronecker's delta. It can easily be verified that  $df(RC(M)) = (I-M-1)(J-M-1)$ . AMs have been mainly developed by Goodman (see [4,5] and references therein) and are thus often referred to as Goodman's AMs. For a detailed presentation of the association models, their inference, properties, interpretation, the role of the weights used and associated literature, we refer to the book of Kateri [7] (Chapter 6).

For ease in understanding but also for interpretation purposes, it is convenient to think in terms of the local associations of the table and define thus AMs through the local odds ratios (LORs)

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}}, \quad i = 1, \dots, I-1, j = 1, \dots, J-1. \quad (10)$$

Recall that the  $(I-1) \times (J-1)$  table of LORs  $\theta = (\theta_{ij})$ , jointly with the marginal probabilities vectors  $\pi_r$  and  $\pi_c$ , specify uniquely the corresponding  $I \times J$  probability table  $\pi$ . Thus, given  $\pi_r$  and  $\pi_c$ , a model on  $\pi$  can equivalently be expressed in terms of  $\theta$ . Hence, model (8) is alternatively defined as

$$\log \theta_{ij} = \sum_{m=1}^M \zeta_m (\mu_{im} - \mu_{i+1,m})(v_{jm} - v_{j+1,m}), \quad (11)$$

for  $i = 1, \dots, I-1, j = 1, \dots, J-1$ . In this set-up, the differences of successive row and column scores are constant for the U model, equal to

$$\mu_i - \mu_{i+1} = \Delta_1 \text{ and } v_j - v_{j+1} = \Delta_2, \quad i = 1, \dots, I-1, j = 1, \dots, J-1, \quad (12)$$

since scores for successive categories are equidistant. Hence, the U model is equivalently defined as

$$\log \theta_{ij} = \zeta(\mu_i - \mu_{i+1})(v_j - v_{j+1}) = \zeta\Delta_1\Delta_2 = \log \theta = \theta^{(0)} \quad (13)$$

and is the model under which all local odds ratios are equal across the table, justifying its ‘uniform association’ characterization.

## 2.2. Correlation Models

A popular, mainly descriptive method for exploring the pattern of association in contingency tables is correspondence analysis (CA). The detailed discussion of CA is beyond our scope. For this, we refer to the book of Greenacre [8]. Correspondence analysis is a reparameterized version of the canonical correlation model of order  $M$

$$p_{ij} = p_i p_j \left( 1 + \sum_{m=1}^M \rho_m x_{im} y_{jm} \right), \quad i = 1, \dots, I, j = 1, \dots, J, \quad (14)$$

with  $1 \leq M \leq M^*$ . The row and column scores ( $\mathbf{x}_m = (x_{1m}, \dots, x_{Im})$  and  $\mathbf{y}_m = (y_{1m}, \dots, y_{Jm})$ ,  $m = 1, \dots, M$ ) satisfy constraints (9) subject to the marginal weights. Usually, it is assumed  $M = 2$  and the row and column scores (coordinates) are graphically displayed as points in two-dimensional plots. The similarities between (8), expressed in terms of  $\pi$  in its multiplicative form, and (14) are obvious. Thus, motivated by the inferential approaches for AMs, MLEs have been considered for model (14), leading to the row-column correlation model of order  $M$ , while, for  $M = 1$ , special correlation models of U, R or C type have also been discussed by Goodman [4,5].

## 3. $\phi$ -Divergence Based Association Models

AMs and correlation models were developed competitively and opposed to each other (cf. [5]), until Gilula et al. [9] linked them in an inspiring manner under an information theoretical approach. They proved that, under certain (common) conditions, both of them are the closest model to independence. Their difference lies on the divergence used for measuring their closeness to independence. AMs are the closest in terms of the KL divergence and correlation models in terms of the Pearson’s divergence. This result motivated subsequent research and led to the definition of general classes of dependence models by substituting the KL and Pearson divergences through generalized families of divergences.

Under the conditions of [9], namely for given marginal distributions ( $\pi_r$  and  $\pi_c$ ), given scores ( $\mu$  and  $\nu$ ) and given their correlation  $\rho = \text{corr}(\mu, \nu)$ , Kateri and Papaioannou [10] derived that the joint distribution  $\pi$  that is closest to independence in terms of the  $\phi$ -divergence is of the form

$$\pi_{ij} = \pi_{i+}\pi_{+j}F^{-1}(\alpha_i + \beta_j + \zeta\mu_i\nu_j), \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad (15)$$

where  $F^{-1}$  is the inverse function of  $F(x) = \phi'(x)$ . The scores  $\mu$  and  $\nu$  satisfy the constraints (7) with marginal weights. Under these constraints, it can easily be verified that  $\rho = \text{corr}(\mu, \nu) = \sum_{i,j} \mu_i\nu_j\pi_{ij}$ . Additionally, the identifiability constraints

$$\sum_i \pi_{i+}\alpha_i = \sum_j \pi_{+j}\beta_j = 0 \quad (16)$$

are imposed on parameters  $\alpha = (\alpha_1, \dots, \alpha_I)$  and  $\beta = (\beta_1, \dots, \beta_J)$ . The parameter  $\zeta$  is measuring association. It can be verified that (15), due to (7) with marginal weights and (16), leads to

$$\zeta(\pi, \mu, \nu) = \sum_{i,j} \pi_{i+}\pi_{+j}\mu_i\nu_j F\left(\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}\right) \quad (17)$$

and that  $\zeta = 0$  if and only if the independence model holds ( $\pi = \pi^I$ ). Furthermore, under model (15), the correlation  $\rho$  between the row and column scores is increasing in  $\zeta$  and the  $\phi$ -divergence measure  $I_\phi^C(\pi, \pi^I)$ , for given  $\phi$ -function, is increasing in  $|\zeta|$ .

Model (15), with known row and column scores, is the  $\phi$ -divergence based extension of the LL model and is denoted by  $\text{LL}_\phi$ . If the scores are additionally equidistant for successive categories, (15) becomes the  $\phi$ -divergence based U model,  $\text{U}_\phi$ , while the classes of models  $\text{R}_\phi$ ,  $\text{C}_\phi$  and  $\text{RC}_\phi$  are defined analogously. The standard LL, R, C or RC models correspond to  $\phi(x) = x \log x$ . The analogue correlation models, defined by (14) for  $M = 1$ , are derived for  $\phi(x) = (1 - x)^2$ , setting  $\mu = \mathbf{x}_1$ ,  $\nu = \mathbf{y}_1$  and  $\zeta = \rho_1$ . For the standard association and correlation models, (17) simplifies to

$$\zeta(\pi, \mu, \nu) = \sum_{i,j} \pi_{i+}\pi_{+j}\mu_i\nu_j \log(\pi_{ij}) \quad (18)$$

and

$$\zeta(\pi, \mu, \nu) = \sum_{i,j} \mu_i\nu_j\pi_{ij} = \text{corr}(\mu, \nu) = \rho, \quad (19)$$

respectively.

For the power divergence (for  $\phi(x) = \frac{x^{\lambda+1}-x}{\lambda(\lambda+1)}$ ,  $\lambda \neq -1, 0$ ), model (15) becomes

$$\pi_{ij} = \pi_{i+}\pi_{+j} \left[ \frac{1}{\lambda+1} + \lambda(\alpha_i + \beta_j + \zeta\mu_i\nu_j) \right]^{1/\lambda}, \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad (20)$$

considered by Rom and Sarkar [11]. Expression (20) defines parametric classes of AMs, controlled by the parameter  $\lambda$ , which are denoted by  $\text{LL}_\lambda$ ,  $\text{U}_\lambda$ ,  $\text{R}_\lambda$ ,  $\text{C}_\lambda$  or  $\text{RC}_\lambda$ , according to the assumption made for the row and column scores.

The  $\text{RC}(M)$  model,  $1 \leq M \leq M^*$ , is analogously generalized to  $\text{RC}_\phi(M)$ , the class of  $\phi$ -divergence AMs of order  $M$ , given by

$$\pi_{ij} = \pi_{i+}\pi_{+j}F^{-1}\left(\alpha_i + \beta_j + \sum_{m=1}^M \zeta_m\mu_{im}\nu_{jm}\right), \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad (21)$$

where the scores  $\mu_m$  and  $\nu_m$  satisfy restrictions (9) with marginal weights. The standard RC(M) model (8) is derived for  $\phi(x) = x \log x$ , while, for  $\phi(x) = (1-x)^2$ , model (21) becomes the correlation model (14) with  $\mu_m = \mathbf{x}_m$ ,  $\nu_m = \mathbf{y}_m$  and  $\zeta_m = \rho_m$ ,  $m = 1, \dots, M$ .

Models (15) and (21) can alternatively be expressed as

$$\theta_{ij}^{(\phi)} = \zeta(\mu_i - \mu_{i+1})(\nu_j - \nu_{j+1}), \quad i = 1, \dots, I-1, \quad j = 1, \dots, J-1, \quad (22)$$

and

$$\theta_{ij}^{(\phi)} = \sum_{m=1}^M \zeta_m (\mu_{im} - \mu_{i+1,m})(\nu_{jm} - \nu_{j+1,m}), \quad i = 1, \dots, I-1, \quad j = 1, \dots, J-1, \quad (23)$$

respectively, where  $\theta_{ij}^{(\phi)}$  is a scaled measure of local dependence, defined for  $i = 1, \dots, I-1$ ,  $j = 1, \dots, J-1$  as

$$\theta_{ij}^{(\phi)}(\boldsymbol{\pi}) = F\left(\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}\right) + F\left(\frac{\pi_{i+1,j+1}}{\pi_{i+1,}\pi_{+,j+1}}\right) - F\left(\frac{\pi_{i+1,j}}{\pi_{i+1,}\pi_{+j}}\right) - F\left(\frac{\pi_{i,j+1}}{\pi_{i+}\pi_{+,j+1}}\right). \quad (24)$$

For  $\phi(x) = x \log x$ , (24) becomes the well-known log local odds ratio  $\log(\theta_{ij})$ , modelled in (13) and from now on denoted as  $\theta_{ij}^{(0)} = \log(\theta_{ij})$ . For the power divergence, we get

$$\theta_{ij}^{(\lambda)}(\boldsymbol{\pi}) = \frac{1}{\lambda} \left[ \left( \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}} \right)^\lambda + \left( \frac{\pi_{i+1,j+1}}{\pi_{i+1,}\pi_{+,j+1}} \right)^\lambda - \left( \frac{\pi_{i+1,j}}{\pi_{i+1,}\pi_{+j}} \right)^\lambda - \left( \frac{\pi_{i,j+1}}{\pi_{i+}\pi_{+,j+1}} \right)^\lambda \right]. \quad (25)$$

**Remark 1.** Kateri and Papaioannou [10] studied properties of the class of  $\phi$ -divergence based AMs. Additionally, they developed a test based on a  $\phi$ -divergence statistic for testing the GOF of  $\phi$ -divergence AMs and studied its efficiency. The choice of the  $\phi$ -function in the test statistic is independent of the  $\phi$ -function used for the model definition. Thus, it serves as a  $\phi$ -divergence based GOF test for the traditional well-known association or correlation models.

In order to understand the role and nature of the scaled measures of local association (24), one can examine a simple  $2 \times 2$  contingency table. In this case, (10) becomes the well-known odds ratio  $\theta (= \theta_{11})$ . Its  $\phi$ -divergence scaled generalization, (24) for  $I = J = 2$ , has been explored by Espendiller and Kateri [12]. For these  $\phi$ -scaled association measures for  $2 \times 2$  tables, asymptotic tests of significance and confidence intervals (CIs) are constructed based on the fact that they are asymptotically normal distributed. An interesting feature of the family of  $\phi$ -scaled association measures for  $2 \times 2$  tables is that when the odds ratio  $\theta$  cannot be finitely estimated (due to sampling zeros), some members of this family provide finite estimates, while, for a subset of them, their variance is also finite. Extensive evaluation studies verified earlier statements in the literature about the low coverage of the log odds ratio CIs when the association is very high ( $\log \theta > 4$ ). In such cases, focusing on the power divergence scaled odds ratios  $\theta^{(\lambda)}$ ,  $\theta^{(1/3)}$  for  $\lambda = 1/3$  is to be preferred, since the corresponding CI is of better coverage when approaching the borders of the parameter space and is in general less conservative than the classical log odds ratio CI [12]. Here, the role of the scale effect on measuring the local dependence will be further clarified in the examples discussed in Section 5.

**Remark 2.** The idea of viewing a model as a departure from a parsimonious reference model, with the property of being the closest to this reference model under certain conditions in terms of the KL divergence, can be adopted for other types of models as well, such as the quasi symmetry (QS) model and the logistic regression, lightening thus a different interpretational aspect of these models. Substitution of the KL divergence by the  $\phi$ -divergence, leads further, in an analogous manner to AMs, to the derivation of generalized  $\phi$ -divergence based classes of QS models [13], ordinal QS [14] and logistic regression models [15]. For example, in case of the QS, the role of the scaled measures (24) take the analogously defined scaled deviations from the model of complete symmetry.

#### 4. Uniform Local Association

We shall focus on the case of uniform local association and the corresponding  $\phi$  divergence scaled model  $U_\phi$ , defined by (15) or (22) with row and column scores satisfying (12). Model  $U_\phi$  is equivalently expressed as

$$\theta_{ij}^{(\phi)} = \zeta \Delta_1 \Delta_2 = \theta^{(\phi)}, \quad i = 1, \dots, I - 1, \quad j = 1, \dots, J - 1 \quad (26)$$

and forms a family of models. Compare to (13) for the U model defined in terms of the usual log odds ratio. The associated probability table under model  $U_\phi$  is uniquely specified by the one-to-one map  $(\pi_r, \pi_c, \theta^{(\phi)}) \mapsto \pi^{U_\phi}$ , not given in a closed-form expression.

The MLEs of the marginal probabilities are the corresponding marginal sample proportions  $\hat{\pi}_r = \mathbf{p}_r = (p_{1+}, \dots, p_{I+})^T$ ,  $\hat{\pi}_c = \mathbf{p}_c = (p_{+1}, \dots, p_{+J})^T$ , while the MLE of  $\theta^{(\phi)}$ ,  $\hat{\theta}^{(\phi)}$ , is not available in explicit form. For a given  $\phi$ -function, model  $U_\phi$  belongs to the family of homogeneous linear predictor (HLP) models [16]. It is straightforward to verify that it satisfies the two conditions of Definition 3 in [16]. In practice, it can be fitted using Lang's mph R-package. The standard U model, denoted in the sequel as  $U_0$ , has the equivalent HLP model expression

$$L(\mathbf{m}_v) = \mathbf{C} \log(\mathbf{m}_v) = \mathbf{X} \boldsymbol{\beta} = \mathbf{1}_{(I-1)(J-1)} \theta^{(0)}, \quad (27)$$

where  $\mathbf{X}$  and  $\boldsymbol{\beta}$  are the model's design matrix and parameter vector (scalar for  $U_0$ ), respectively. Furthermore,  $\theta^{(0)} = \log(\theta)$ ,  $\mathbf{m}_v$  is the  $IJ \times 1$  vector of expected cell frequencies, corresponding to the  $I \times J$  table  $\mathbf{m}$  expanded by rows, and  $\mathbf{C}$  is an  $(I-1)(J-1) \times IJ$  design matrix so that the vector of all  $\log(\text{LOR})$ s is derived, i.e.,  $\mathbf{C} \log(\mathbf{m}_v) = (\theta_{11}^{(0)}, \dots, \theta_{I-1, J-1}^{(0)})^T$ . For more details on inference for model  $U_0$  through HLP models, we refer to [7] (Sections 5.6 and 6.6.4). In Section 6.6.4 of [7], the approach is implemented in R for the example of Table 1 (see Section 5), while an R-function for constructing the design matrix  $\mathbf{C}$  for two-way tables of any size is provided in the web appendix of the book. This approach is easily adjusted for model  $U_\phi$ , by replacing  $\log(\mathbf{m}_v)$  in (27) by  $F(\mathbf{m}_{vs})$ , where  $\mathbf{m}_{vs}$  is the  $IJ \times 1$  vector with entries  $\frac{m_{ij}}{m_{i+} m_{+j}/n} = \frac{\pi_{ij}}{\pi_{i+} \pi_{+j}}$ , expanded by rows.

Under  $U_\phi$ , all  $2 \times 2$  subtables, formed by any successive rows and any successive columns, share the same  $\phi$ -scaled local association  $\theta^{(\phi)}$ . It is of practical interest to have estimators of this common local association, alternative to the MLEs, that are provided in explicit forms. One such estimator, based on the sample version of (17), is given by

$$\tilde{\theta}^{(\phi)} = \zeta(\mathbf{p}, \boldsymbol{\mu}, \nu) \Delta_1 \Delta_2 = \Delta_1 \Delta_2 \sum_{i,j} p_{i+} p_{+j} \mu_i \nu_j F\left(\frac{p_{ij}}{p_{i+} p_{+j}}\right). \quad (28)$$

Another option is

$$\bar{\theta}^{(\phi)} = \frac{1}{(I-1)(J-1)} \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} \theta_{ij}^{(\phi)}(\mathbf{p}), \quad (29)$$

which is the mean of the sample  $\phi$ -scaled local association measures (24). For the power-divergence based models  $U_\lambda$ , estimators (28) and (29) are denoted by  $\tilde{\theta}^{(\lambda)}$  and  $\bar{\theta}^{(\lambda)}$ , respectively. For the U correlation model, derived for  $\lambda = 1$  and denoted thus by  $U_1$ , estimator (28) takes the form

$$\tilde{\theta}^{(1)} = \Delta_1 \Delta_2 \sum_{i,j} \mu_i \nu_j p_{ij} = \Delta_1 \Delta_2 r, \quad (30)$$

where  $r$  is the sample correlation between the row and column scores (compare to (19)).

Under the  $U_\phi$  model,  $\theta^{(\phi)}$  is the single association parameter, measuring the strength of the local association that is uniform across the table. Furthermore,  $\theta^{(\phi)} = 0 \Leftrightarrow \zeta = 0$  if and only if the

independence (I) model holds. Since model I is nested in  $U_\phi$ , the following test of independence, conditional on  $U_\phi$ , can be considered

$$G^2(I|U_\phi) = G^2(I) - G^2(U_\phi), \quad (31)$$

which is asymptotically  $\chi_1^2$  distributed. This test is well-known for the standard  $U_0$  model (see [17] and Section 6.3 in [7]).

**Remark 3.** For model  $U_0$ , Tomizawa [18] proposed a measure of divergence from uniform association, based on the KL-divergence, taking values in the interval  $[0, 1]$  and being equal to 0 if and only if  $U_0$  holds. He constructed also asymptotic confidence interval for this measure, provided  $U_0$  does not hold. Conde and Salicrú [19] extended his work by considering such a measure based on the  $\phi$ -divergence and developed asymptotic inference for it, and also for the case that  $U_0$  holds. Their approach and measures should not be confused with the approach followed here. They aim at detecting departures from  $U_0$  (in favor of a non-uniform association structure) while here we focus on measuring the strength of uniform association, provided that  $U_0$  (or  $U_\phi$ ) holds.

Tomizawa [18] as well as Conde and Salicrú [19] based the estimation of their measures on the following closed-form estimator of  $\theta^{(0)}$  under  $U_0$

$$\theta_*^{(0)} = \log \left( \frac{\sum_{i=1}^{I-1} \sum_{j=1}^{J-1} \theta_{ij}^{(0)}(\mathbf{p})}{(I-1)(J-1)} \right). \quad (32)$$

**Remark 4.** For square contingency tables with commensurable classification variables, analogous to the measure of departure from  $U_0$  (see Remark 3), Tomizawa et al. [20] introduced a measure of departure from complete symmetry relying on the power divergence and Tomizawa [21] a measure of departure from marginal homogeneity. Kateri and Papaioannou [22] extended these measures to corresponding  $\phi$ -divergence based measures and proposed further  $\phi$ -divergence based measures of departure from the QS and triangular symmetry models. The work of Menéndez et al. [23,24] is also related.

## 5. Illustrations

We revisit a classical data set of Grizzle [25], provided in Table 2, which is adequately modelled by the  $U_0$  model (see [18,19]). Our second data set in Table 1 corresponds to a study in [26] and provides strong evidence in favor of  $U_0$  (see [7]). The maximum likelihood estimates of the expected cell frequencies under  $U_0$  are also given in parentheses in Tables 1 and 2.

**Table 1.** Students' survey about cannabis use at the University of Ioannina, Greece (1995). The maximum likelihood estimates of the expected cell frequencies under the  $U_0$  model are given in parentheses.

Alcohol Consumption	I Tried Cannabis...			
	Never	Once or Twice	More Often	Total
at most once/month	204 (204.4)	6 (5.7)	1 (0.9)	211
twice/month	211 (211.4)	13 (13.1)	5 (4.5)	229
twice/week	357 (352.8)	44 (48.8)	38 (37.4)	439
more often	92 (95.3)	34 (29.4)	49 (50.3)	175
Total	864	97	93	1054

**Table 2.** Cross-classification of duodenal ulcer patients according to operation and dumping severity. The maximum likelihood estimates of the expected cell frequencies under the  $U_0$  model are given in parentheses.

Operation	Dumping Severity			
	None	Slight	Moderate	Total
A	61 (62.5)	28 (26.2)	7 (7.3)	96
B	68 (62.9)	23 (30.9)	13 (10.2)	104
C	58 (61.0)	40 (35.3)	12 (13.7)	110
D	53 (53.7)	38 (36.6)	16 (16.7)	107
Total	240	129	48	417

The  $G^2$  test statistics for the  $U_\lambda$  models fitted on these data, for  $\lambda \rightarrow 0$  and  $\lambda = 1/3, 2/3, 1$  are provided in Table 3, along with corresponding  $\hat{\theta}^{(\lambda)}$ ,  $\tilde{\theta}^{(\lambda)}$  and  $\bar{\theta}^{(\lambda)}$  values, i.e., the estimates of the  $\theta^{(\lambda)}$ 's discussed in Section 4. For  $U_0$ , the  $\theta_*^{(0)}$  estimates (32) for Tables 1 and 2 are equal to 0.2890 and 0.7972, respectively.

We observe that all considered  $U_\lambda$  models are of very similar (acceptable) fit for the first example while they differ enormously for the second one. For the data of Table 1, the fit of  $U_0$  is impressive, that of  $U_{1/3}$  acceptable while  $U_{2/3}$  and  $U_1$  are of very bad fit. A reverse situation appears for another data set, given in Table 4. In this case model,  $U_1$  can be accepted while  $U_0$  can not (see Table 5). The maximum likelihood estimates of the expected cell frequencies under  $U_1$  are provided in Table 4 in parentheses.

**Table 3.** Goodness of fit of the  $U_\lambda$  models for the data of Tables 1 and 2 along with estimates of the common  $\lambda$ -scaled local association  $\theta^{(\lambda)}$  under  $U_\lambda$ .

Example in Table 1			
$\lambda$	$G^2$ (p-value)	$\hat{\theta}^{(\lambda)}$ / $\tilde{\theta}^{(\lambda)}$ / $\bar{\theta}^{(\lambda)}$	
0	1.47 (0.917)	0.8026/0.7817/0.7814	
1/3	7.19 (0.207)	0.6857/0.6451/0.6432	
2/3	25.21 (0.000)	0.4720/0.5853/0.5974	
1	40.76 (0.000)	0.3667/0.5732/0.6096	
Example in Table 2			
$\lambda$	$G^2$ (p-value)	$\hat{\theta}^{(\lambda)}$ / $\tilde{\theta}^{(\lambda)}$ / $\bar{\theta}^{(\lambda)}$	
0	4.59 (0.468)	0.1626/0.1665/0.1612	
1/3	4.57 (0.471)	0.1619/0.1638/0.1573	
2/3	4.55 (0.473)	0.1612/0.1616/0.1541	
1	4.52 (0.477)	0.1606/0.1599/0.1515	

**Table 4.** Hypothetical  $4 \times 3$  data table with maximum likelihood estimates of the expected cell frequencies under  $U_1$  (in parentheses).

160 (159.5)	8 (7.8)	1 (1.4)
198 (199.4)	16 (14.7)	14 (13.9)
310 (321.8)	40 (32.9)	50 (45.4)
161 (149.1)	12 (20.4)	30 (33.8)

**Table 5.** Goodness of fit of the  $U_\lambda$  models and maximum likelihood estimates of the common  $\lambda$ -scaled local association  $\theta^{(\lambda)}$  under  $U_\lambda$  for the data of Table 4.

$\lambda$	$G^2$ ( <i>p</i> -Value)	$\hat{\theta}^{(\lambda)}$
0	15.99 (0.007)	0.3082
1/3	13.52 (0.019)	0.3266
2/3	10.64 (0.059)	0.3391
1	7.95 (0.159)	0.3215

Observe (in Table 3) that the closed form estimates for the  $\lambda$ -scaled local associations are close to the corresponding maximum likelihood estimates in case the assumed model is of adequate fit while they diverge for models of bad fit.

## 6. Conclusions

We revealed the role of  $\phi$ -divergence in modelling association in two-way contingency tables and illustrated it for the special case of uniform association in ordinal contingency tables. Targeting at pointing out the potential of this modelling approach and the generated families of models, we avoided presenting technical details, properties and inferential results for these models, which can be found in the initial sources cited.

Crucial quantities are the  $\theta_{ij}^{(\phi)}$ s, the  $\phi$ -scaled measures of local association. The generalized family of  $\phi$ -divergence based AMs enriches the modelling options in CTA, since the pattern of underlying association structure in a table may be simplified and thus described by a more parsimonious model when considering a different scale. A crucial issue, as also pointed out by one of the reviewers, is how to decide on the scale. So far, such a decision is based on trials of various alternative options. A formal approach selecting the scale is missing. In case of a parametric family, like the one based on the power divergence, the problem can be tackled by considering  $\lambda$  as an unknown parameter and estimating it from the data. Such an approach has been followed in the logistic regression set-up by Kateri and Agresti [15].

It is important to realize that, due to the scale difference,  $\theta_{ij}^{(\phi)}$ s are not directly comparable for different  $\phi$ -function (or  $\lambda$ -values in case of the power divergence). Thus, comparisons across different  $\phi$ s (or  $\lambda$ s) are possible only in terms of the corresponding expected cell frequencies or a common measure of local association evaluated on them. AMs can also be considered for other types of generalized odds ratios, like, for example, the global odds ratios. The extension of such models through the  $\phi$ -divergence and their study is the subject of a work in progress. Inference for closed form estimators of the common  $\theta^{(\phi)}$  of the  $U_\phi$  model and comparisons among them is the content of a paper under preparation.

The conditional test of independence (31) can be based not only on  $U_\phi$  but on  $LL_\phi$  models as well. Another 1 *df* test of independence for ordinal classification variables is the linear trend test of Mantel (see [27]). It considers the testing problem  $H_0 : \rho = 0$  vs.  $H_1 : \rho \neq 0$ , where  $\rho$  is the correlation between ordered scores assigned to the categories of the classification variables of the table. It is thus applicable only when the underlying association exhibits a linear trend for the assigned scores. The test of Mantel uses the test statistic  $M^2 = (n - 1)r^2$ , which is under  $H_0$  asymptotically  $\chi_1^2$  distributed. The way this test is linked to the above-mentioned conditional tests, in view also of (19), is interesting to be investigated further.

Throughout this paper, we assumed a multinomial sampling scheme. For the models considered here, the other two classical sampling schemes for contingency tables (independent Poisson and product multinomial) are inferentially equivalent. Furthermore, for ease of presentation, we restricted here to two-way tables. The proposed models extend straightforwardly to multi-way tables. For two- or higher-dimensional tables, the subset of models that are linear in their parameters (i.e., RC-

and  $RC(M)$ -type terms are excluded) belong to the family of homogeneous linear predictor (HLP) models [16] and can thus be fitted using the R-package mph.

**Acknowledgments:** The author thanks the referees for their constructive and very useful comments that improved the paper.

**Conflicts of Interest:** The author declares no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

CTA	contingency table analysis
MLE	maximum likelihood estimator
M $\phi$ Es	minimum $\phi$ -divergence estimators
KL	Kullback–Leibler
GOF	goodness of fit
I	independence model
AMs	association models
LL model	linear by linear association model
R model	row effect association model
C model	column effect association model
RC model	multiplicative row-column effect association model
U model	uniform association model
df	degrees of freedom
LOR	local odds ratio
CA	correspondence analysis
CI	confidence interval
QS	quasi symmetry
HLP	homogeneous linear predictor

## References

1. Stigler, S. The missing early history of contingency tables. *Annales de la Faculté des Sciences de Toulouse* **2002**, *4*, 563–573.
2. Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman & Hall: New York, NY, USA, 2006.
3. Martin, N.; Pardo, N. New families of estimators and test statistics in log-linear models. *J. Multivar. Anal.* **2008**, *99*, 1590–1609.
4. Goodman, L.A. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models and asymmetry models for contingency tables with or without missing entries. *Ann. Stat.* **1985**, *13*, 10–69.
5. Goodman, L.A. Some useful extensions of the usual correspondence analysis and the usual log-linear models approach in the analysis of contingency tables with or without missing entries (with Discussion). *Int. Stat. Rev.* **1986**, *54*, 243–309.
6. Cressie, N.; Read, T.R.C. Multinomial Goodness-of-Fit Tests. *J. R. Stat. Soc. B* **1984**, *46*, 440–464.
7. Kateri, M. *Contingency Table Analysis: Methods and Implementation Using R*; Birkhäuser/Springer: New York, NY, USA, 2014.
8. Greenacre, M. *Correspondence Analysis in Practice*, 2nd ed.; Chapman & Hall: London, UK, 2007.
9. Gilula, Z.; Krieger, A.M.; Ritov, Y. Ordinal association in contingency tables: some interpretive aspects. *J. Am. Stat. Assoc.* **1988**, *83*, 540–545.
10. Kateri, M.; Papaioannou, T.  $f$ -divergence association models. *LRT J. Math. Stat. Sci.* **1995**, *3*, 179–203.
11. Rom, D.; Sarkar, S.K. A generalized model for the analysis of association in ordinal contingency tables. *J. Stat. Plan. Inference* **1992**, *33*, 205–212.
12. Espenheimer, M.; Kateri, M. A family of association measures for  $2 \times 2$  contingency tables based on the  $\phi$ -divergence. *Stat. Methodol.* **2016**, *35*, 45–61.
13. Kateri, M.; Papaioannou, T. Asymmetry models for contingency tables. *J. Am. Stat. Assoc.* **1997**, *92*, 1124–1131.

14. Kateri, M.; Agresti, A. A class of ordinal quasi symmetry models for square contingency tables. *Stat. Probab. Lett.* **2007**, *77*, 598–603.
15. Kateri, M.; Agresti, A. A generalized regression model for a binary response. *Stat. Probab. Lett.* **2010**, *80*, 89–95.
16. Lang, J.B. Homogeneous Linear Predictor Models for Contingency Tables. *J. Am. Stat. Assoc.* **2005**, *100*, 121–134.
17. Goodman, L.A. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Am. Stat. Assoc.* **1981**, *76*, 320–334.
18. Tomizawa, S. Shannon entropy type measure of departure from uniform association in cross-classification having ordered categories. *Stat. Probab. Lett.* **1991**, *11*, 547–550.
19. Conde, J.; Salicrú, M. Uniform association in contingency tables associated to Csiszár divergence. *Stat. Probab. Lett.* **1998**, *37*, 149–154.
20. Tomizawa, S.; Seo, T.; Yamamoto, H. Power-divergence-type measure of departure from symmetry for square tables that have nominal categories. *J. Appl. Stat.* **1998**, *25*, 387–398.
21. Tomizawa, S. Measures of departures from marginal homogeneity for contingency tables with nominal categories. *Statistician* **1995**, *44*, 425–439.
22. Papaioannou, T.; Kateri, M. Measures of symmetry–asymmetry for square contingency tables. In Proceedings of the 13th Conference of the Greek Statistical Institute, Florina, Greece, 3–7 May 2000; pp. 435–444.
23. Menéndez, M.I.; Pardo, J.A.; Pardo, L. Tests based on  $\phi$ -divergences for bivariate symmetry. *Metrika* **2001**, *53*, 15–29.
24. Menéndez, M.L.; Pardo, J.A.; Pardo, L. Tests for bivariate symmetry against ordered alternatives in square contingency tables. *Aust. N. Z. J. Stat.* **2003**, *45*, 115–124.
25. Grizzle, J.E.; Starmer, C.F.; Koch, G.G. Analysis of categorical data by linear models. *Biometrics* **1969**, *25*, 489–504.
26. Marselos, M.; Boutsouris, K.; Liapi, H.; Malamas, M.; Kateri, M.; Papaioannou, T. Epidemiological aspects on the use of cannabis among university students in Greece. *Eur. Addict. Res.* **1997**, *3*, 184–191.
27. Agresti, A. *Categorical Data Analysis*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2013.



© 2018 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Robust and Sparse Regression via $\gamma$ -Divergence

Takayuki Kawashima <sup>1,\*</sup> and Hironori Fujisawa <sup>1,2,3</sup>

<sup>1</sup> Department of Statistical Science, The Graduate University for Advanced Studies, Tokyo 190-8562, Japan; fujisawa@ism.ac.jp

<sup>2</sup> The Institute of Statistical Mathematics, Tokyo 190-8562, Japan

<sup>3</sup> Department of Mathematical Statistics, Nagoya University Graduate School of Medicine, Nagoya 466-8550, Japan

\* Correspondence: t-kawa@ism.ac.jp; Tel.: +81-50-5533-8500

Received: 30 September 2017; Accepted: 9 November 2017; Published: 13 November 2017

**Abstract:** In high-dimensional data, many sparse regression methods have been proposed. However, they may not be robust against outliers. Recently, the use of density power weight has been studied for robust parameter estimation, and the corresponding divergences have been discussed. One such divergence is the  $\gamma$ -divergence, and the robust estimator using the  $\gamma$ -divergence is known for having a strong robustness. In this paper, we extend the  $\gamma$ -divergence to the regression problem, consider the robust and sparse regression based on the  $\gamma$ -divergence and show that it has a strong robustness under heavy contamination even when outliers are heterogeneous. The loss function is constructed by an empirical estimate of the  $\gamma$ -divergence with sparse regularization, and the parameter estimate is defined as the minimizer of the loss function. To obtain the robust and sparse estimate, we propose an efficient update algorithm, which has a monotone decreasing property of the loss function. Particularly, we discuss a linear regression problem with  $L_1$  regularization in detail. In numerical experiments and real data analyses, we see that the proposed method outperforms past robust and sparse methods.

**Keywords:** sparse; robust; divergence; MM algorithm

---

## 1. Introduction

In high-dimensional data, sparse regression methods have been intensively studied. The Lasso [1] is a typical sparse linear regression method with  $L_1$  regularization, but is not robust against outliers. Recently, robust and sparse linear regression methods have been proposed. The robust least angle regression (RLARS) [2] is a robust version of LARS [3], which replaces the sample correlation by a robust estimate of correlation in the update algorithm. The sparse least trimmed squares (sLTS) [4] is a sparse version of the well-known robust linear regression method LTS [5] based on the trimmed loss function with  $L_1$  regularization.

Recently, the robust parameter estimation using density power weight has been discussed by Windham [6], Basu et al. [7], Jones et al. [8], Fujisawa and Eguchi [9], Basu et al. [10], Kanamori and Fujisawa [11], and so on. The density power weight gives a small weight to the terms related to outliers, and then, the parameter estimation becomes robust against outliers. By virtue of this validity, some applications using density power weights have been proposed in signal processing and machine learning [12,13]. Among them, the  $\gamma$ -divergence proposed by Fujisawa and Eguchi [9] is known for having a strong robustness, which implies that the latent bias can be sufficiently small even under heavy contamination. The other robust methods including density power-divergence cannot achieve the above property, and the estimator can be affected by the outlier ratio. In addition, to obtain the robust estimate, an efficient update algorithm was proposed with a monotone decreasing property of the loss function.

In this paper, we propose the robust and sparse regression problem based on the  $\gamma$ -divergence. First, we extend the  $\gamma$ -divergence to the regression problem. Next, we consider a loss function based on the  $\gamma$ -divergence with sparse regularization and propose an update algorithm to obtain the robust and sparse estimate. Fujisawa and Eguchi [9] used a Pythagorean relation on the  $\gamma$ -divergence, but it is not compatible with sparse regularization. Instead of this relation, we use the majorization-minimization algorithm [14]. This idea is deeply considered in a linear regression problem with  $L_1$  regularization. The MM algorithm was also adopted in Hirose and Fujisawa [15] for robust and sparse Gaussian graphical modeling. A tuning parameter selection is proposed using a robust cross-validation. We also show a strong robustness under heavy contamination even when outliers are heterogeneous. Finally, in numerical experiments and real data analyses, we show that our method is computationally efficient and outperforms other robust and sparse methods. The R language software package “gamreg”, which we use to implement our proposed method, can be downloaded at <http://cran.r-project.org/web/packages/gamreg/>.

## 2. Regression Based on $\gamma$ -Divergence

The  $\gamma$ -divergence was defined for two probability density functions, and its properties were investigated by Fujisawa and Eguchi [9]. In this section, the  $\gamma$ -divergence is extended to the regression problem, in other words, defined for two conditional probability density functions.

### 2.1. $\gamma$ -Divergence for Regression

We suppose that  $g(x, y)$ ,  $g(y|x)$  and  $g(x)$  are the underlying probability density functions of  $(x, y)$ ,  $y$  given  $x$  and  $x$ , respectively. Let  $f(y|x)$  be another parametric conditional probability density function of  $y$  given  $x$ . Let us define the  $\gamma$ -cross-entropy for regression by:

$$\begin{aligned} d_\gamma(g(y|x), f(y|x); g(x)) &= -\frac{1}{\gamma} \log \int \left( \int g(y|x) f(y|x)^\gamma dy \right) g(x) dx + \frac{1}{1+\gamma} \log \int \left( \int f(y|x)^{1+\gamma} dy \right) g(x) dx \\ &= -\frac{1}{\gamma} \log \int \int f(y|x)^\gamma g(x, y) dx dy + \frac{1}{1+\gamma} \log \int \left( \int f(y|x)^{1+\gamma} dy \right) g(x) dx \quad \text{for } \gamma > 0. \end{aligned} \quad (1)$$

The  $\gamma$ -divergence for regression is defined by:

$$D_\gamma(g(y|x), f(y|x); g(x)) = -d_\gamma(g(y|x), g(y|x); g(x)) + d_\gamma(g(y|x), f(y|x); g(x)). \quad (2)$$

The  $\gamma$ -divergence for regression was first proposed by Fujisawa and Eguchi [9], and many properties were already shown. However, we adopt the definition (2), which is slightly different from the past one, because (2) satisfies the Pythagorean relation approximately (see Section 4).

**Theorem 1.** *We can show that:*

- (i)  $D_\gamma(g(y|x), f(y|x); g(x)) \geq 0$ ,
- (ii)  $D_\gamma(g(y|x), f(y|x); g(x)) = 0 \Leftrightarrow g(y|x) = f(y|x) \quad (\text{a.e.})$ ,
- (iii)  $\lim_{\gamma \rightarrow 0} D_\gamma(g(y|x), f(y|x); g(x)) = \int D_{KL}(g(y|x), f(y|x)) g(x) dx$ ,

where  $D_{KL}(g(y|x), f(y|x)) = \int g(y|x) \log g(y|x) dy - \int g(y|x) \log f(y|x) dy$ .

The proof is in Appendix A. In what follows, we refer to the regression based on the  $\gamma$ -divergence as the  $\gamma$ -regression.

## 2.2. Estimation for $\gamma$ -Regression

Let  $f(y|x; \theta)$  be the conditional probability density function of  $y$  given  $x$  with parameter  $\theta$ . The target parameter can be considered by:

$$\begin{aligned}\theta_{\gamma}^* &= \operatorname{argmin}_{\theta} D_{\gamma}(g(y|x), f(y|x; \theta); g(x)) \\ &= \operatorname{argmin}_{\theta} d_{\gamma}(g(y|x), f(y|x; \theta); g(x)).\end{aligned}\quad (3)$$

When  $g(y|x) = f(y|x; \theta^*)$ , we have  $\theta_{\gamma}^* = \theta^*$ .

Let  $(x_1, y_1), \dots, (x_n, y_n)$  be the observations randomly drawn from the underlying distribution  $g(x, y)$ . Using the formula (1), the  $\gamma$ -cross-entropy for regression,  $d_{\gamma}(g(y|x), f(y|x; \theta); g(x))$ , can be empirically estimated by:

$$\bar{d}_{\gamma}(f(y|x; \theta)) = -\frac{1}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n f(y_i|x_i; \theta)^{\gamma} \right\} + \frac{1}{1+\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n \int f(y|x_i; \theta)^{1+\gamma} dy \right\}.$$

By virtue of (3), we define the  $\gamma$ -estimator by:

$$\hat{\theta}_{\gamma} = \operatorname{argmin}_{\theta} \bar{d}_{\gamma}(f(y|x; \theta)).\quad (4)$$

In a similar way as in Fujisawa and Eguchi [9], we can show the consistency of  $\hat{\theta}_{\gamma}$  to  $\theta_{\gamma}^*$  under some conditions.

Here, we briefly show why the  $\gamma$ -estimator is robust. Suppose that  $y_1$  is an outlier. The conditional probability density  $f(y_1|x_1; \theta)$  can be expected to be sufficiently small. We see from  $f(y_1|x_1; \theta) \approx 0$  and (4) that:

$$\begin{aligned}&\operatorname{argmin}_{\theta} \bar{d}_{\gamma}(f(y|x; \theta)) \\ &= \operatorname{argmin}_{\theta} -\frac{1}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n f(y_i|x_i; \theta)^{\gamma} \right\} + \frac{1}{1+\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n \int f(y|x_i; \theta)^{1+\gamma} dy \right\} \\ &\approx \operatorname{argmin}_{\theta} -\frac{1}{\gamma} \log \left\{ \frac{1}{n-1} \sum_{i=2}^n f(y_i|x_i; \theta)^{\gamma} \right\} + \frac{1}{1+\gamma} \log \left\{ \frac{1}{n-1} \sum_{i=2}^n \int f(y|x_i; \theta)^{1+\gamma} dy \right\}.\end{aligned}$$

Therefore, the term  $f(y_1|x_1; \theta)$  is naturally ignored in (4). However, for the KL-divergence,  $\log f(y_1|x_1; \theta)$  diverges from  $f(y_1|x_1; \theta) \approx 0$ . That is why the KL-divergence is not robust. The theoretical robust properties are presented in Section 4.

Moreover, the empirical estimation of the  $\gamma$ -cross-entropy with a penalty term can be given by:

$$L_{\gamma}(\theta; \lambda) = \bar{d}_{\gamma}(f(y|x; \theta)) + \lambda P(\theta),$$

where  $P(\theta)$  is a penalty for parameter  $\theta$  and  $\lambda$  is a tuning parameter for the penalty term. As an example of the penalty term, we can consider  $L_1$  (Lasso, Tibshirani 1), elasticnet [16], group Lasso [17], fused Lasso [18], and so on. The sparse  $\gamma$ -estimator can be proposed by:

$$\hat{\theta}_S = \operatorname{argmin}_{\theta} L_{\gamma}(\theta; \lambda).$$

To obtain the minimizer, we propose the iterative algorithm by the majorization-minimization algorithm (MM algorithm) [14].

### 3. Parameter Estimation Procedure

#### 3.1. MM Algorithm for Sparse $\gamma$ -Regression

The MM algorithm is constructed as follows. Let  $h(\eta)$  be the objective function. Let us prepare the majorization function  $h_{MM}$  satisfying:

$$\begin{aligned} h_{MM}(\eta^{(m)} | \eta^{(m)}) &= h(\eta^{(m)}), \\ h_{MM}(\eta | \eta^{(m)}) &\geq h(\eta) \quad \text{for all } \eta, \end{aligned}$$

where  $\eta^{(m)}$  is the parameter of the  $m$ -th iterative step for  $m = 0, 1, 2, \dots$ . Let us consider the iterative algorithm by:

$$\eta^{(m+1)} = \underset{\eta}{\operatorname{argmin}} h_{MM}(\eta | \eta^{(m)}).$$

Then, we can show that the objective function  $h(\eta)$  monotonically decreases at each step, because:

$$\begin{aligned} h(\eta^{(m)}) &= h_{MM}(\eta^{(m)} | \eta^{(m)}) \\ &\geq h_{MM}(\eta^{(m+1)} | \eta^{(m)}) \\ &\geq h(\eta^{(m+1)}). \end{aligned}$$

Note that  $\eta^{(m+1)}$  does not necessarily have to be the minimizer of  $h_{MM}(\eta | \eta^{(m)})$ . We only need:

$$h_{MM}(\eta^{(m)} | \eta^{(m)}) \geq h_{MM}(\eta^{(m+1)} | \eta^{(m)}).$$

We construct the majorization function for the sparse  $\gamma$ -regression by the following inequality:

$$\kappa(z^T \eta) \leq \sum_i \frac{z_i \eta_i^{(m)}}{z^T \eta^{(m)}} \kappa \left[ \eta_i \frac{z^T \eta^{(m)}}{\eta_i^{(m)}} \right], \quad (5)$$

where  $\kappa(u)$  is a convex function,  $z = (z_1, \dots, z_n)^T$ ,  $\eta = (\eta_1, \dots, \eta_n)^T$ ,  $\eta^{(m)} = (\eta_1^{(m)}, \dots, \eta_n^{(m)})^T$ , and  $z_i$ ,  $\eta_i$  and  $\eta_i^{(m)}$  are positive. The inequality (5) holds from Jensen's inequality. Here, we take  $z_i = \frac{1}{n}$ ,  $\eta_i = f(y_i | x_i; \theta)^\gamma$ ,  $\eta_i^{(m)} = f(y_i | x_i; \theta^{(m)})^\gamma$ , and  $\kappa(u) = -\log u$  in (5). We can propose the majorization function as follows:

$$\begin{aligned} h(\theta) &= L_\gamma(\theta; \lambda) \\ &= -\frac{1}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n f(y_i | x_i; \theta)^\gamma \right\} + \frac{1}{1+\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n \int f(y | x_i; \theta)^{1+\gamma} dy \right\} + \lambda P(\theta) \\ &\leq -\frac{1}{\gamma} \sum_{i=1}^n \alpha_i^{(m)} \log \left\{ f(y_i | x_i; \theta)^\gamma \frac{\frac{1}{n} \sum_{l=1}^n f(y_l | x_l; \theta^{(m)})^\gamma}{f(y_i | x_i; \theta^{(m)})^\gamma} \right\} \\ &\quad + \frac{1}{1+\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n \int f(y | x_i; \theta)^{1+\gamma} dy \right\} + \lambda P(\theta) \\ &= -\sum_{i=1}^n \alpha_i^{(m)} \log f(y_i | x_i; \theta) + \frac{1}{1+\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n \int f(y | x_i; \theta)^{1+\gamma} dy \right\} + \lambda P(\theta) \\ &\quad + const \\ &= h_{MM}(\theta | \theta^{(m)}) + const, \end{aligned}$$

where  $\alpha_i^{(m)} = \frac{f(y_i|x_i;\theta^{(m)})^\gamma}{\sum_{l=1}^n f(y_l|x_l;\theta^{(m)})^\gamma}$  and *const* is a term that does not depend on the parameter  $\theta$ .

The first term on the original target function  $h(\theta)$  is a mixture type of densities, which is not easy to optimize, while the first term on  $h_{MM}(\theta|\theta^{(m)})$  is a weighted log-likelihood, which is often easy to optimize.

### 3.2. Sparse $\gamma$ -Linear Regression

Let  $f(y|x;\theta)$  be the conditional density with  $\theta = (\beta_0, \beta, \sigma^2)$ , given by:

$$f(y|x;\theta) = \phi(y; \beta_0 + x^T \beta, \sigma^2),$$

where  $\phi(y;\mu,\sigma^2)$  is the normal density with mean parameter  $\mu$  and variance parameter  $\sigma^2$ . Suppose that  $P(\theta)$  is the  $L_1$  regularization  $\|\beta\|_1$ . After a simple calculation, we have:

$$h_{MM}(\theta|\theta^{(m)}) = \frac{1}{2(1+\gamma)} \log \sigma^2 + \frac{1}{2} \sum_{i=1}^n \alpha_i^{(m)} \frac{(y_i - \beta_0 - x_i^T \beta)^2}{\sigma^2} + \lambda \|\beta\|_1. \quad (6)$$

This function is easy to optimize by an update algorithm. For a fixed value of  $\sigma^2$ , the function  $h_{MM}$  is almost the same as Lasso except for the weight, so that it can be updated using the coordinate decent algorithm with a decreasing property of the loss function. For a fixed value of  $(\beta_0, \beta^T)^T$ , the function  $h_{MM}$  is easy to minimize. Consequently, we can obtain the update algorithm in Algorithm 1 with the decreasing property:

$$h_{MM}(\theta^{(m+1)}|\theta^{(m)}) \leq h_{MM}(\theta^{(m)}|\theta^{(m)}).$$

---

#### Algorithm 1 Sparse $\gamma$ -linear regression.

---

**Require:**  $\beta_0^{(0)}, \beta^{(0)}, \sigma^{2(0)}$

**repeat**  $m = 0, 1, 2, \dots$

$$\alpha_i^{(m)} \leftarrow \frac{\phi(y_i; \beta_0^{(m)} + x_i^T \beta^{(m)}, \sigma^{2(m)})^\gamma}{\sum_{l=1}^n \phi(y_l; \beta_0^{(m)} + x_l^T \beta^{(m)}, \sigma^{2(m)})^\gamma} \quad (i = 1, 2, \dots, n).$$

$$\beta_0^{(m+1)} \leftarrow \sum_{i=1}^n \alpha_i^{(m)} (y_i - x_i^T \beta^{(m)}).$$

**for** **do**  $j = 1, \dots, p$

$$\beta_j^{(m+1)} \leftarrow \frac{S\left(\sum_{i=1}^n \alpha_i^{(m)} (y_i - \beta_0^{(m+1)} - r_{i,-j}^{(m)}) x_{ij}, \sigma^{2(m)} \lambda\right)}{\left(\sum_{i=1}^n \alpha_i^{(m)} x_{ij}^2\right)},$$

where  $S(t, \lambda) = \text{sign}(t)(|t| - \lambda)_+$  and  $r_{i,-j}^{(m)} = \sum_{k \neq j} x_{ik} (\mathbb{1}_{(k < j)} \beta_k^{(m+1)} + \mathbb{1}_{(k > j)} \beta_k^{(m)})$ .

$$\sigma^{2(m+1)} \leftarrow (1+\gamma) \sum_{i=1}^n \alpha_i^{(m)} (y_i - \beta_0^{(m+1)} - x_i^T \beta^{(m+1)})^2.$$

**until** convergence

**Ensure:**  $\hat{\beta}_0, \hat{\beta}, \hat{\sigma}^2$

---

It should be noted that  $h_{MM}$  is convex with respect to parameter  $\beta_0, \beta$  and has the global minimum with respect to parameter  $\sigma^2$ , but the original objective function  $h$  is not convex with respect to them, so that the initial points of Algorithm 1 are important. This issue is discussed in Section 5.4.

In practice, we also use the active set strategy [19] in the coordinate decent algorithm for updating  $\beta^{(m)}$ . The active set consists of the non-zero coordinates of  $\beta^{(m)}$ . Specifically, for a given  $\beta^{(m)}$ , we only update the non-zero coordinates of  $\beta^{(m)}$ , until they are converged. Then, the non-active set parameter estimates are updated once. When they remain zero, the coordinate descent algorithm stops. If some of them do not remain zero, those are added to the active set, and the coordinate descent algorithm continues.

### 3.3. Robust Cross-Validation

In sparse regression, a regularization parameter is often selected via a criterion. Cross-validation is often used for selecting the regularization parameter. Ordinal cross-validation is based on the squared error, and it can also be constructed using the KL-cross-entropy with the normal density. However, the ordinal cross-validation will fail due to outliers. Therefore, we propose the robust cross-validation based on the  $\gamma$ -cross-entropy. Let  $\hat{\theta}_\gamma$  be the robust estimate based on the  $\gamma$ -cross-entropy. The cross-validation based on the  $\gamma$ -cross-entropy can be given by:

$$\text{RoCV}(\lambda) = -\frac{1}{\gamma_0} \log \left\{ \frac{1}{n} \sum_{i=1}^n f(y_i|x_i; \hat{\theta}_\gamma^{[-i]})^{\gamma_0} \right\} + \frac{1}{1+\gamma_0} \log \left\{ \frac{1}{n} \sum_{i=1}^n \int f(y|x_i; \hat{\theta}_\gamma^{[-i]})^{1+\gamma_0} dy \right\},$$

where  $\hat{\theta}_\gamma^{[-i]}$  is the  $\gamma$ -estimator deleting the  $i$ -th observation and  $\gamma_0$  is an appropriate tuning parameter. We can also adopt the  $K$ -fold cross-validation to reduce the computational task [20].

Here, we give a small modification of the above. We often focus only on the mean structure for prediction, not on the variance parameter. Therefore, in this paper,  $\hat{\theta}_\gamma^{[-i]} = (\hat{\beta}_\gamma^{[-i]}, \hat{\sigma}_{\gamma}^{2[-i]})$  is replaced by  $(\hat{\beta}_\gamma^{[-i]}, \hat{\sigma}_{fix}^2)$ . In numerical experiments and real data analyses, we used  $\sigma^{2(0)}$  as  $\sigma_{fix}^2$ .

## 4. Robust Properties

In this section, the robust properties are presented from two viewpoints of latent bias and Pythagorean relation. The latent bias was discussed in Fujisawa and Eguchi [9] and Kanamori and Fujisawa [11], which is described later. Using the results obtained there, the Pythagorean relation is shown in Theorems 2 and 3.

Let  $f^*(y|x) = f_{\theta^*}(y|x) = f(y|x; \theta^*)$  and  $\delta(y|x)$  be the target conditional probability density function and the contamination conditional probability density function related to outliers, respectively. Let  $\epsilon$  and  $\epsilon(x)$  denote the outlier ratios, which are independent of and dependent on  $x$ , respectively. Under homogeneous and heterogeneous contaminations, we suppose that the underlying conditional probability density function can be expressed as:

$$\begin{aligned} g(y|x) &= (1 - \epsilon)f(y|x; \theta^*) + \epsilon\delta(y|x), \\ g(y|x) &= (1 - \epsilon(x))f(y|x; \theta^*) + \epsilon(x)\delta(y|x). \end{aligned}$$

Let:

$$v_{f,\gamma}(x) = \left\{ \int \delta(y|x)f(y|x)^\gamma dy \right\}^{\frac{1}{\gamma}} \quad (\gamma > 0),$$

and let:

$$v_{f,\gamma} = \left\{ \int v_{f,\gamma}(x)^\gamma g(x) dx \right\}^{\frac{1}{\gamma}}.$$

Here, we assume that:

$$v_{f_{\theta^*},\gamma} \approx 0,$$

which implies that  $v_{f_{\theta^*},\gamma}(x) \approx 0$  for any  $x$  (a.e.) and illustrates that the contamination conditional probability density function  $\delta(y|x)$  lies on the tail of the target conditional probability density function  $f(y|x; \theta^*)$ . For example, if  $\delta(y|x)$  is the Dirac function at the outlier  $y_\dagger(x)$  given  $x$ , then we have

$\nu_{f_{\theta^*}, \gamma}(x) = f(y_*(x)|x; \theta^*)$ , which should be sufficiently small because  $y_*(x)$  is an outlier. In this section, we show that  $\theta_\gamma^* - \theta^*$  is expected to be small even if  $\epsilon$  or  $\epsilon(x)$  is not small. To make the discussion easier, we prepare the monotone transformation of the  $\gamma$ -cross-entropy for regression by:

$$\begin{aligned} \tilde{d}_\gamma(g(y|x), f(y|x; \theta); g(x)) \\ = -\exp\{-\gamma d_\gamma(g(y|x), f(y|x; \theta); g(x))\} \\ = -\frac{\int (\int g(y|x)f(y|x; \theta)^\gamma dy) g(x) dx}{\{\int (\int f(y|x; \theta)^{1+\gamma} dy) g(x) dx\}^{\frac{\gamma}{1+\gamma}}}. \end{aligned}$$

#### 4.1. Homogeneous Contamination

Here, we provide the following proposition, which was given in Kanamori and Fujisawa [11].

##### Proposition 1.

$$\begin{aligned} \tilde{d}_\gamma(g(y|x), f(y|x; \theta); g(x)) \\ = (1-\epsilon)\tilde{d}_\gamma(f(y|x; \theta^*), f(y|x; \theta); g(x)) - \frac{\epsilon \nu_{f_{\theta}, \gamma}^\gamma}{\{\int (\int f(y|x; \theta)^{1+\gamma} dy) g(x) dx\}^{\frac{\gamma}{1+\gamma}}}. \end{aligned}$$

Recall that  $\theta_\gamma^*$  and  $\theta^*$  are also the minimizers of  $\tilde{d}_\gamma(g(y|x), f(y|x; \theta); g(x))$  and  $\tilde{d}_\gamma(f(y|x; \theta^*), f(y|x; \theta); g(x))$ , respectively. We can expect  $\nu_{f_{\theta}, \gamma} \approx 0$  from the assumption  $\nu_{f_{\theta^*}, \gamma} \approx 0$  if the tail behavior of  $f(y|x; \theta)$  is close to that of  $f(y|x; \theta^*)$ . We see from Proposition 1 and the condition  $\nu_{f_{\theta}, \gamma} \approx 0$  that:

$$\begin{aligned} \theta_\gamma^* &= \operatorname{argmin}_\theta \tilde{d}_\gamma(g(y|x), f(y|x; \theta); g(x)) \\ &= \operatorname{argmin}_\theta [(1-\epsilon)\tilde{d}_\gamma(f(y|x; \theta^*), f(y|x; \theta); g(x)) \\ &\quad - \frac{\epsilon \nu_{f_{\theta}, \gamma}^\gamma}{\{\int (\int f(y|x; \theta)^{1+\gamma} dy) g(x) dx\}^{\frac{\gamma}{1+\gamma}}}] \\ &\approx \operatorname{argmin}_\theta (1-\epsilon)\tilde{d}_\gamma(f(y|x; \theta^*), f(y|x; \theta); g(x)) \\ &= \theta^*. \end{aligned}$$

Therefore, under homogeneous contamination, it can be expected that the latent bias  $\theta_\gamma^* - \theta^*$  is small even if  $\epsilon$  is not small. Moreover, we can show the following theorem, using Proposition 1.

**Theorem 2.** Let  $\nu = \max\{\nu_{f_{\theta}, \gamma}, \nu_{f_{\theta^*}, \gamma}\}$ . Then, the Pythagorean relation among  $g(y|x)$ ,  $f(y|x; \theta^*)$ ,  $f(y|x; \theta)$  approximately holds:

$$\begin{aligned} D_\gamma(g(y|x), f(y|x; \theta); g(x)) - D_\gamma(g(y|x), f(y|x; \theta^*); g(x)) \\ = D_\gamma(f(y|x; \theta^*), f(y|x; \theta); g(x)) + O(\nu^\gamma). \end{aligned}$$

The proof is in Appendix A. The Pythagorean relation implies that the minimization of the divergence from  $f(y|x; \theta)$  to the underlying conditional probability density function  $g(y|x)$  is approximately the same as that to the target conditional probability density function  $f(y|x; \theta^*)$ . Therefore, under homogeneous contamination, we can see why our proposed method works well in terms of the minimization of the  $\gamma$ -divergence.

#### 4.2. Heterogeneous Contamination

Under heterogeneous contamination, we assume that the parametric conditional probability density function  $f(y|x; \theta)$  is a location-scale family given by:

$$f(y|x; \theta) = \frac{1}{\sigma} s\left(\frac{y - q(x; \xi)}{\sigma}\right),$$

where  $s(y)$  is a probability density function,  $\sigma$  is a scale parameter and  $q(x; \xi)$  is a location function with a regression parameter  $\xi$ , e.g.,  $q(x; \xi) = \xi^T x$ . Then, we can obtain:

$$\begin{aligned} \int f(y|x; \theta)^{1+\gamma} dy &= \int \frac{1}{\sigma^{1+\gamma}} s\left(\frac{y - q(x; \xi)}{\sigma}\right)^{1+\gamma} dy \\ &= \sigma^{-\gamma} \int s(z)^{1+\gamma} dz. \end{aligned}$$

That does not depend on the explanatory variable  $x$ . Here, we provide the following proposition, which was given in Kanamori and Fujisawa [11].

#### Proposition 2.

$$\begin{aligned} \tilde{d}_\gamma(g(y|x), f(y|x; \theta); g(x)) \\ = c \tilde{d}_\gamma(f(y|x; \theta^*), f(y|x; \theta); \tilde{g}(x)) - \frac{\int v_{f_\theta, \gamma}(x)^\gamma \epsilon(x) g(x) dx}{\{\sigma^{-\gamma} \int s(z)^{1+\gamma} dz\}^{\frac{\gamma}{1+\gamma}}}, \end{aligned}$$

where  $c = (1 - \int \epsilon(x) g(x) dx)^{\frac{\gamma}{1+\gamma}}$  and  $\tilde{g}(x) = (1 - \epsilon(x)) g(x)$ .

The second term  $\frac{\int v_{f_\theta, \gamma}(x)^\gamma \epsilon(x) g(x) dx}{\{\sigma^{-\gamma} \int s(z)^{1+\gamma} dz\}^{\frac{\gamma}{1+\gamma}}}$  can be approximated to be zero from the condition  $v_{f_\theta, \gamma} \approx 0$  and  $\epsilon(x) < 1$  as follows:

$$\begin{aligned} \frac{\int v_{f_\theta, \gamma}(x)^\gamma \epsilon(x) g(x) dx}{\{\sigma^{-\gamma} \int s(z)^{1+\gamma} dz\}^{\frac{\gamma}{1+\gamma}}} &< \frac{\int v_{f_\theta, \gamma}(x)^\gamma g(x) dx}{\{\sigma^{-\gamma} \int s(z)^{1+\gamma} dz\}^{\frac{\gamma}{1+\gamma}}} \\ &= \frac{v_{f_\theta, \gamma}^\gamma}{\{\sigma^{-\gamma} \int s(z)^{1+\gamma} dz\}^{\frac{\gamma}{1+\gamma}}} \\ &\approx 0. \end{aligned} \tag{7}$$

We see from Proposition 2 and (7) that:

$$\begin{aligned} \theta_\gamma^* &= \operatorname{argmin}_\theta \tilde{d}_\gamma(g(y|x), f(y|x; \theta); g(x)) \\ &= \operatorname{argmin}_\theta [c \tilde{d}_\gamma(f(y|x; \theta^*), f(y|x; \theta); \tilde{g}(x)) \\ &\quad - \frac{\int v_{f_\theta, \gamma}(x)^\gamma \epsilon(x) g(x) dx}{\{\sigma^{-\gamma} \int s(z)^{1+\gamma} dz\}^{\frac{\gamma}{1+\gamma}}}] \\ &\approx \operatorname{argmin}_\theta c \tilde{d}_\gamma(f(y|x; \theta^*), f(y|x; \theta); \tilde{g}(x)) \\ &= \theta^*. \end{aligned}$$

Therefore, under heterogeneous contamination in a location-scale family, it can be expected that the latent bias  $\theta_\gamma^* - \theta^*$  is small even if  $\epsilon(x)$  is not small. Moreover, we can show the following theorem, using Proposition 2.

**Theorem 3.** Let  $v = \max\{\nu_{f_\theta, \gamma}, \nu_{f_{\theta^*}, \gamma}\}$ . Then, the following relation among  $g(y|x)$ ,  $f(y|x; \theta^*)$ ,  $f(y|x; \theta)$  approximately holds:

$$\begin{aligned} D_\gamma(g(y|x), f(y|x; \theta); g(x)) - D_\gamma(g(y|x), f(y|x; \theta^*); g(x)) \\ = D_\gamma(f(y|x; \theta^*), f(y|x; \theta); \tilde{g}(x)) + O(v^\gamma). \end{aligned}$$

The proof is in Appendix A. The above is slightly different from a conventional Pythagorean relation, because the base measure changes from  $g(x)$  to  $\tilde{g}(x)$  in part. However, it also implies that the minimization of the divergence from  $f(y|x; \theta)$  to the underlying conditional probability density function  $g(y|x)$  is approximately the same as that to the target conditional probability density function  $f(y|x; \theta^*)$ . Therefore, under heterogeneous contamination in a location-scale family, we can see why our proposed method works well in terms of the minimization of the  $\gamma$ -divergence.

#### 4.3. Redescending Property

First, we review a redescending property on M-estimation (see, e.g., [21]), which is often used in robust statistics. Suppose that the estimating equation is given by  $\sum_{i=1}^n \zeta(z_i; \theta) = 0$ . Let  $\hat{\theta}$  be a solution of the estimating equation. The bias caused by outlier  $z_o$  is expressed as  $\hat{\theta}_{n=\infty} - \theta^*$ , where  $\hat{\theta}_{n=\infty}$  is the limiting value of  $\hat{\theta}$  and  $\theta^*$  is the true parameter. We hope the bias is small even if the outlier  $z_o$  exists. Under some conditions, the bias can be approximated to  $\epsilon \text{IF}(z_o; \theta^*)$ , where  $\epsilon$  is a small outlier ratio and  $\text{IF}(z; \theta^*)$  is the influence function. The bias is expected to be small when the influence function is small. The influence function can be expressed as  $\text{IF}(z; \theta^*) = A\zeta(z; \theta^*)$ , where  $A$  is a matrix independent of  $z$ , so that the bias is also expected to be small when  $\zeta(z_o; \theta^*)$  is small. In particular, the estimating equation is said to have a redescending property if  $\zeta(z; \theta^*)$  goes to zero as  $\|z\|$  goes to infinity. This property is favorable in robust statistics, because the bias is expected to be sufficiently small when  $z_o$  is very large.

Here, we prove a redescending property on the sparse  $\gamma$ -linear regression, i.e., when  $f(y|x; \theta) = \phi(y; \beta_0 + x^T \beta, \sigma^2)$  with  $\theta = (\beta_0, \beta, \sigma^2)$  for fixed  $x$ . Recall that the estimate of the sparse  $\gamma$ -linear regression is the minimizer of the loss function:

$$L_\gamma(\theta; \lambda) = -\frac{1}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n \phi(y_i; \beta_0 + x_i^T \beta, \sigma^2)^\gamma \right\} + b_\gamma(\theta; \lambda),$$

where  $b_\gamma(\theta; \lambda) = \frac{1}{1+\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n \int \phi(y; \beta_0 + x_i^T \beta, \sigma^2)^{1+\gamma} dy \right\} + \lambda \|\beta\|_1$ . Then, the estimating equation is given by:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} L_\gamma(\theta; \lambda) \\ &= -\frac{\sum_{i=1}^n \phi(y_i; \beta_0 + x_i^T \beta, \sigma^2)^\gamma s(y_i|x_i; \theta)}{\sum_{i=1}^n \phi(y_i; \beta_0 + x_i^T \beta, \sigma^2)^\gamma} + \frac{\partial}{\partial \theta} b_\gamma(\theta; \lambda), \end{aligned}$$

where  $s(y|x; \theta) = \frac{\partial \log \phi(y; \beta_0 + x^T \beta, \sigma^2)}{\partial \theta}$ . This can be expressed by the M-estimation formula given by:

$$0 = \sum_{i=1}^n \psi(y_i|x_i; \theta),$$

where  $\psi(y|x; \theta) = \phi(y; \beta_0 + x^T \beta, \sigma^2)^\gamma s(y|x; \theta) - \phi(y; \beta_0 + x^T \beta, \sigma^2)^\gamma \frac{\partial}{\partial \theta} b_\gamma(\theta; \lambda)$ . We can easily show that as  $\|y\|$  goes to infinity,  $\phi(y; \beta_0 + x^T \beta, \sigma^2)$  goes to zero and  $\phi(y; \beta_0 + x^T \beta, \sigma^2)s(y|x; \theta)$  also goes

to zero. Therefore, the function  $\psi(y|x; \theta)$  goes to zero as  $||y||$  goes to infinity, so that the estimating equation has a redescending property.

## 5. Numerical Experiment

In this section, we compare our method (sparse  $\gamma$ -linear regression) with the representative sparse linear regression method, the least absolute shrinkage and selection operator (Lasso) [1], and the robust and sparse regression methods, sparse least trimmed squares (sLTS) [4] and robust least angle regression (RLARS) [2].

### 5.1. Regression Models for Simulation

We used the simulation model given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e, \quad e \sim N(0, 0.5^2).$$

The sample size and the number of explanatory variables were set to be  $n = 100$  and  $p = 100, 200$ , respectively. The true coefficients were given by:

$$\begin{aligned} \beta_1 &= 1, \beta_2 = 2, \beta_4 = 4, \beta_7 = 7, \beta_{11} = 11, \\ \beta_j &= 0 \text{ for } j \in \{0, \dots, p\} \setminus \{1, 2, 4, 7, 11\}. \end{aligned}$$

We arranged a broad range of regression coefficients to observe sparsity for various degrees of regression coefficients. The explanatory variables were generated from a normal distribution  $N(0, \Sigma)$  with  $\Sigma = (\rho^{|i-j|})_{1 \leq i,j \leq p}$ . We generated 100 random samples.

Outliers were incorporated into simulations. We investigated two outlier ratios ( $\epsilon = 0.1$  and  $0.3$ ) and two outlier patterns: (a) the outliers were generated around the middle part of the explanatory variable, where the explanatory variables were generated from  $N(0, 0.5^2)$  and the error terms were generated from  $N(20, 0.5^2)$ ; (b) the outliers were generated around the edge part of the explanatory variable, where the explanatory variables were generated from  $N(-1.5, 0.5^2)$  and the error terms were generated from  $N(20, 0.5^2)$ .

### 5.2. Performance Measure

The root mean squared prediction error (RMSPE) and mean squared error (MSE) were examined to verify the predictive performance and fitness of regression coefficient:

$$\begin{aligned} \text{RMSPE}(\hat{\beta}) &= \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^* - x_i^{*T} \hat{\beta})^2}, \\ \text{MSE} &= \frac{1}{p+1} \sum_{j=0}^p (\beta_j^* - \hat{\beta}_j)^2, \end{aligned}$$

where  $(x_i^*, y_i^*)$  ( $i = 1, \dots, n$ ) is the test sample generated from the simulation model without outliers and  $\beta_j^*$ 's are the true coefficients. The true positive rate (TPR) and true negative rate (TNR) were also reported to verify the sparsity:

$$\begin{aligned} \text{TPR}(\hat{\beta}) &= \frac{|\{j \in \{1, \dots, p\} : \hat{\beta}_j \neq 0 \wedge \beta_j^* \neq 0\}|}{|\{j \in \{1, \dots, p\} : \beta_j^* \neq 0\}|}, \\ \text{TNR}(\hat{\beta}) &= \frac{|\{j \in \{1, \dots, p\} : \hat{\beta}_j = 0 \wedge \beta_j^* = 0\}|}{|\{j \in \{1, \dots, p\} : \beta_j^* = 0\}|}. \end{aligned}$$

### 5.3. Comparative Methods

In this subsection, we explain three comparative methods: Lasso, RLARS and sLTS.

Lasso is performed by the R-package “glmnet”. The regularization parameter  $\lambda_{Lasso}$  is selected by grid search via cross-validation in “glmnet”. We used “glmnet” by default.

RLARS is performed by the R-package “robustHD”. This is a robust version of LARS [3]. The optimal model is selected via BIC by default.

sLTS is performed by the R-package “robustHD”. sLTS has the regularization parameter  $\lambda_{sLTS}$  and the fraction parameter  $\alpha$  of squared residuals used for trimmed squares. The regularization parameter  $\lambda_{sLTS}$  is selected by grid search via BIC. The number of grids is 40 by default. However, we considered that this would be small under heavy contamination. Therefore, we used 80 grids under heavy contamination to obtain a good performance. The fraction parameter  $\alpha$  is 0.75 by default. In the case of  $\alpha = 0.75$ , the ratio of outlier is less than 25%. We considered this would be small under heavy contamination and large under low contamination in terms of statistical efficiency. Therefore, we used 0.65, 0.75, 0.85 as  $\alpha$  under low contamination and 0.50, 0.65, 0.75 under heavy contamination.

### 5.4. Details of Our Method

#### 5.4.1. Initial Points

In our method, we need an initial point to obtain the estimate, because we use the iterative algorithm proposed in Section 3.2. The estimate of other conventional robust and sparse regression methods would give a good initial point. For another choice, the estimate of RANSAC (random sample consensus) algorithm would also give a good initial point. In this experiment, we used the estimate of SLTS as an initial point.

#### 5.4.2. How to Choose Tuning Parameters

In our method, we have to choose some tuning parameters. The parameter  $\gamma$  in the  $\gamma$ -divergence was set to 0.1 or 0.5. The parameter  $\gamma_0$  in the robust cross-validation was set to 0.5. In our experience, the result via RoCV is not sensitive to the selection of  $\gamma_0$  when  $\gamma_0$  is large enough, e.g.,  $\gamma_0 = 0.5, 1$ . The parameter  $\lambda$  of  $L_1$  regularization is often selected via grid search. We used 50 grids in the range  $[0.05\lambda_0, \lambda_0]$  with the log scale, where  $\lambda_0$  is an estimate of  $\lambda$ , which would shrink regression coefficients to zero. More specifically, in a similar way as in Lasso, we can derive  $\lambda_0$ , which shrinks the coefficients  $\beta$  to zero in  $h_{MM}(\theta|\theta^{(0)})$  [6] with respect to  $\beta$ , and we used it. This idea was proposed by the R-package “glmnet”.

### 5.5. Result

Table 1 is the low contamination case with Outlier Pattern (a). For the RMSPE, our method outperformed other comparative methods (the oracle value of the RMSPE is 0.5). For the TPR and TNR, sLTS showed a similar performance to our method. Lasso presented the worst performance, because it is sensitive to outliers. Table 2 is the heavy contamination case with Outlier Pattern (a). For the RMSPE, our method outperformed other comparative methods except in the case  $(p, \epsilon, \rho) = (100, 0.3, 0.2)$  for sLTS with  $\alpha = 0.5$ . Lasso also presented a worse performance, and furthermore, sLTS with  $\alpha = 0.75$  showed the worst performance due to a lack of truncation. For the TPR and TNR, our method showed the best performance. Table 3 is the low contamination case with Outlier Pattern (b). For the RMSPE, our method outperformed other comparative methods (the oracle value of the RMSPE is 0.5). For the TPR and TNR, sLTS showed a similar performance to our method. Lasso presented the worst performance, because it is sensitive to outliers. Table 4 is the heavy contamination case with Outlier Pattern (b). For the RMSPE, our method outperformed other comparative methods. sLTS with  $\alpha = 0.5$  showed the worst performance. For the TPR and TNR, it seems that our method showed the best performance. Table 5 is the no contamination case. RLARS showed the best performance, but our method presented comparable performances. In spite of no contamination case, Lasso was clearly

worse than RLARS and our method. This would be because the underlying distribution can generate a large value in simulation, although it is a small probability.

**Table 1.** Outlier Pattern (a) with  $p = 100, 200, \epsilon = 0.1$  and  $\rho = 0.2, 0.5$ . RMSPE, root mean squared prediction error (RMSPE); RLARS, robust least angle regression; sLTS, sparse least trimmed squares.

Methods	$p = 100, \epsilon = 0.1, \rho = 0.2$				$p = 100, \epsilon = 0.1, \rho = 0.5$			
	RMSPE	MSE	TPR	TNR	RMSPE	MSE	TPR	TNR
Lasso	3.04	$9.72 \times 10^{-2}$	0.936	0.909	3.1	$1.05 \times 10^{-1}$	0.952	0.918
RLARS	0.806	$6.46 \times 10^{-3}$	0.936	0.949	0.718	$6.7 \times 10^{-3}$	0.944	0.962
sLTS ( $\alpha = 0.85$ , 80 grids)	0.626	$1.34 \times 10^{-3}$	1.0	0.964	0.599	$1.05 \times 10^{-3}$	1.0	0.966
sLTS ( $\alpha = 0.75$ , 80 grids)	0.651	$1.71 \times 10^{-3}$	1.0	0.961	0.623	$1.33 \times 10^{-3}$	1.0	0.961
sLTS ( $\alpha = 0.65$ , 80 grids)	0.685	$2.31 \times 10^{-3}$	1.0	0.957	0.668	$1.76 \times 10^{-3}$	1.0	0.961
sparse $\gamma$ -linear reg ( $\gamma = 0.1$ )	0.557	$6.71 \times 10^{-4}$	1.0	0.966	0.561	$6.99 \times 10^{-4}$	1.0	0.965
sparse $\gamma$ -linear reg ( $\gamma = 0.5$ )	0.575	$8.25 \times 10^{-4}$	1.0	0.961	0.573	$9.05 \times 10^{-4}$	1.0	0.959
$p = 200, \epsilon = 0.1, \rho = 0.2$				$p = 200, \epsilon = 0.1, \rho = 0.5$				
Methods	RMSPE	MSE	TPR	TNR	RMSPE	MSE	TPR	TNR
Lasso	3.55	$6.28 \times 10^{-2}$	0.904	0.956	3.37	$6.08 \times 10^{-2}$	0.928	0.961
RLARS	0.88	$3.8 \times 10^{-3}$	0.904	0.977	0.843	$4.46 \times 10^{-3}$	0.9	0.986
sLTS ( $\alpha = 0.85$ , 80 grids)	0.631	$7.48 \times 10^{-4}$	1.0	0.972	0.614	$5.77 \times 10^{-4}$	1.0	0.976
sLTS ( $\alpha = 0.75$ , 80 grids)	0.677	$1.03 \times 10^{-3}$	1.0	0.966	0.632	$7.08 \times 10^{-4}$	1.0	0.973
sLTS ( $\alpha = 0.65$ , 80 grids)	0.823	$2.34 \times 10^{-3}$	0.998	0.96	0.7	$1.25 \times 10^{-3}$	1.0	0.967
sparse $\gamma$ -linear reg ( $\gamma = 0.1$ )	0.58	$4.19 \times 10^{-4}$	1.0	0.981	0.557	$3.71 \times 10^{-4}$	1.0	0.977
sparse $\gamma$ -linear reg ( $\gamma = 0.5$ )	0.589	$5.15 \times 10^{-4}$	1.0	0.979	0.586	$5.13 \times 10^{-4}$	1.0	0.977

**Table 2.** Outlier Pattern (a) with  $p = 100, 200, \epsilon = 0.3$  and  $\rho = 0.2, 0.5$ .

Methods	$p = 100, \epsilon = 0.3, \rho = 0.2$				$p = 100, \epsilon = 0.3, \rho = 0.5$			
	RMSPE	MSE	TPR	TNR	RMSPE	MSE	TPR	TNR
Lasso	8.07	$6.72 \times 10^{-1}$	0.806	0.903	8.1	$3.32 \times 10^{-1}$	0.8	0.952
RLARS	2.65	$1.54 \times 10^{-1}$	0.75	0.963	2.09	$1.17 \times 10^{-1}$	0.812	0.966
sLTS ( $\alpha = 0.75$ , 80 grids)	10.4	2.08	0.886	0.709	11.7	2.36	0.854	0.67
sLTS ( $\alpha = 0.65$ , 80 grids)	2.12	$3.66 \times 10^{-1}$	0.972	0.899	2.89	$5.13 \times 10^{-1}$	0.966	0.887
sLTS ( $\alpha = 0.5$ , 80 grids)	1.37	$1.46 \times 10^{-1}$	0.984	0.896	1.53	$1.97 \times 10^{-1}$	0.976	0.909
sparse $\gamma$ -linear reg ( $\gamma = 0.1$ )	1.13	$9.16 \times 10^{-2}$	0.964	0.97	0.961	$5.38 \times 10^{-2}$	0.982	0.977
sparse $\gamma$ -linear reg ( $\gamma = 0.5$ )	1.28	$1.5 \times 10^{-1}$	0.986	0.952	1.00	$8.48 \times 10^{-2}$	0.988	0.958
$p = 200, \epsilon = 0.3, \rho = 0.2$				$p = 200, \epsilon = 0.3, \rho = 0.5$				
Methods	RMSPE	MSE	TPR	TNR	RMSPE	MSE	TPR	TNR
Lasso	8.11	$3.4 \times 10^{-1}$	0.77	0.951	8.02	$6.51 \times 10^{-1}$	0.81	0.91
RLARS	3.6	$1.7 \times 10^{-1}$	0.71	0.978	2.67	$1.02 \times 10^{-1}$	0.76	0.984
sLTS ( $\alpha = 0.75$ , 80 grids)	11.5	1.16	0.738	0.809	11.9	1.17	0.78	0.811
sLTS ( $\alpha = 0.65$ , 80 grids)	3.34	$3.01 \times 10^{-1}$	0.94	0.929	4.22	$4.08 \times 10^{-1}$	0.928	0.924
sLTS ( $\alpha = 0.5$ , 80 grids)	4.02	$3.33 \times 10^{-1}$	0.892	0.903	4.94	$4.44 \times 10^{-1}$	0.842	0.909
sparse $\gamma$ -linear reg ( $\gamma = 0.1$ )	2.03	$1.45 \times 10^{-1}$	0.964	0.924	3.2	$2.86 \times 10^{-1}$	0.94	0.936
sparse $\gamma$ -linear reg ( $\gamma = 0.5$ )	1.23	$7.69 \times 10^{-2}$	0.988	0.942	3.13	$2.98 \times 10^{-1}$	0.944	0.94

**Table 3.** Outlier Pattern (b) with  $p = 100, 200, \epsilon = 0.1$  and  $\rho = 0.2, 0.5$ .

Methods	$p = 100, \epsilon = 0.1, \rho = 0.2$				$p = 100, \epsilon = 0.1, \rho = 0.5$			
	RMSPE	MSE	TPR	TNR	RMSPE	MSE	TPR	TNR
Lasso	2.48	$5.31 \times 10^{-2}$	0.982	0.518	2.84	$5.91 \times 10^{-2}$	0.98	0.565
RLARS	0.85	$6.58 \times 10^{-3}$	0.93	0.827	0.829	$7.97 \times 10^{-3}$	0.91	0.885
sLTS ( $\alpha = 0.85$ , 80 grids)	0.734	$5.21 \times 10^{-3}$	0.998	0.964	0.684	$3.76 \times 10^{-3}$	1.0	0.961
sLTS ( $\alpha = 0.75$ , 80 grids)	0.66	$1.78 \times 10^{-3}$	1.0	0.975	0.648	$1.59 \times 10^{-3}$	1.0	0.961
sLTS ( $\alpha = 0.65$ , 80 grids)	0.734	$2.9 \times 10^{-3}$	1.0	0.96	0.66	$1.74 \times 10^{-3}$	1.0	0.962
sparse $\gamma$ -linear reg ( $\gamma = 0.1$ )	0.577	$8.54 \times 10^{-4}$	1.0	0.894	0.545	$5.44 \times 10^{-4}$	1.0	0.975
sparse $\gamma$ -linear reg ( $\gamma = 0.5$ )	0.581	$7.96 \times 10^{-4}$	1.0	0.971	0.546	$5.95 \times 10^{-4}$	1.0	0.977
$p = 200, \epsilon = 0.1, \rho = 0.2$								
Methods	$p = 200, \epsilon = 0.1, \rho = 0.2$				$p = 200, \epsilon = 0.1, \rho = 0.5$			
	RMSPE	MSE	TPR	TNR	RMSPE	MSE	TPR	TNR
Lasso	2.39	$2.57 \times 10^{-2}$	0.988	0.696	2.57	$2.54 \times 10^{-2}$	0.944	0.706
RLARS	1.01	$5.44 \times 10^{-3}$	0.896	0.923	0.877	$4.82 \times 10^{-3}$	0.898	0.94
sLTS ( $\alpha = 0.85$ , 80 grids)	0.708	$1.91 \times 10^{-3}$	1.0	0.975	0.790	$3.40 \times 10^{-3}$	0.994	0.97
sLTS ( $\alpha = 0.75$ , 80 grids)	0.683	$1.06 \times 10^{-4}$	1.0	0.975	0.635	$7.40 \times 10^{-4}$	1.0	0.977
sLTS ( $\alpha = 0.65$ , 80 grids)	1.11	$1.13 \times 10^{-2}$	0.984	0.956	0.768	$2.60 \times 10^{-3}$	0.998	0.968
sparse $\gamma$ -linear reg ( $\gamma = 0.1$ )	0.603	$5.71 \times 10^{-4}$	1.0	0.924	0.563	$3.78 \times 10^{-3}$	1.0	0.979
sparse $\gamma$ -linear reg ( $\gamma = 0.5$ )	0.592	$5.04 \times 10^{-4}$	1.0	0.982	0.566	$4.05 \times 10^{-3}$	1.0	0.981

**Table 4.** Outlier Pattern (b) with  $p = 100, 200, \epsilon = 0.3$  and  $\rho = 0.2, 0.5$ .

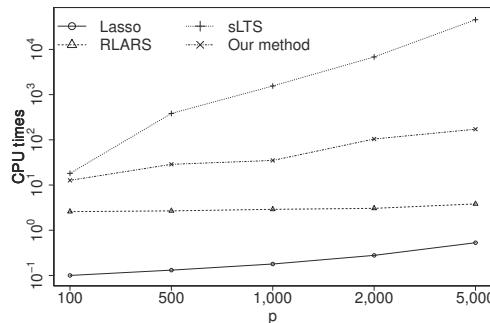
Methods	$p = 100, \epsilon = 0.3, \rho = 0.2$				$p = 100, \epsilon = 0.3, \rho = 0.5$			
	RMSPE	MSE	TPR	TNR	RMSPE	MSE	TPR	TNR
Lasso	2.81	$6.88 \times 10^{-2}$	0.956	0.567	3.13	$7.11 \times 10^{-2}$	0.97	0.584
RLARS	2.70	$7.69 \times 10^{-2}$	0.872	0.789	2.22	$6.1 \times 10^{-2}$	0.852	0.855
sLTS ( $\alpha = 0.75$ , 80 grids)	3.99	$1.57 \times 10^{-1}$	0.856	0.757	4.18	$1.54 \times 10^{-1}$	0.878	0.771
sLTS ( $\alpha = 0.65$ , 80 grids)	3.2	$1.46 \times 10^{-1}$	0.888	0.854	2.69	$1.08 \times 10^{-1}$	0.922	0.867
sLTS ( $\alpha = 0.5$ , 80 grids)	6.51	$4.62 \times 10^{-1}$	0.77	0.772	7.14	$5.11 \times 10^{-1}$	0.844	0.778
sparse $\gamma$ -linear reg ( $\gamma = 0.1$ )	1.75	$3.89 \times 10^{-2}$	0.974	0.725	1.47	$2.66 \times 10^{-2}$	0.976	0.865
sparse $\gamma$ -linear reg ( $\gamma = 0.5$ )	1.68	$3.44 \times 10^{-2}$	0.98	0.782	1.65	$3.58 \times 10^{-2}$	0.974	0.863
$p = 200, \epsilon = 0.3, \rho = 0.2$								
Methods	$p = 200, \epsilon = 0.3, \rho = 0.2$				$p = 200, \epsilon = 0.3, \rho = 0.5$			
	RMSPE	MSE	TPR	TNR	RMSPE	MSE	TPR	TNR
Lasso	2.71	$3.32 \times 10^{-2}$	0.964	0.734	2.86	$3.05 \times 10^{-2}$	0.974	0.728
RLARS	3.03	$4.59 \times 10^{-2}$	0.844	0.876	2.85	$4.33 \times 10^{-2}$	0.862	0.896
sLTS ( $\alpha = 0.75$ , 80 grids)	3.73	$7.95 \times 10^{-2}$	0.864	0.872	4.20	$8.17 \times 10^{-2}$	0.878	0.87
sLTS ( $\alpha = 0.65$ , 80 grids)	4.45	$1.23 \times 10^{-1}$	0.85	0.886	3.61	$8.95 \times 10^{-2}$	0.904	0.908
sLTS ( $\alpha = 0.5$ , 80 grids)	9.05	$4.24 \times 10^{-1}$	0.66	0.853	8.63	$3.73 \times 10^{-1}$	0.748	0.864
sparse $\gamma$ -linear reg ( $\gamma = 0.1$ )	1.78	$1.62 \times 10^{-2}$	0.994	0.731	1.82	$1.62 \times 10^{-2}$	0.988	0.844
sparse $\gamma$ -linear reg ( $\gamma = 0.5$ )	1.79	$1.69 \times 10^{-2}$	0.988	0.79	1.77	$1.51 \times 10^{-2}$	0.996	0.77

**Table 5.** No contamination case with  $p = 100, 200, \epsilon = 0$  and  $\rho = 0.2, 0.5$ .

Methods	$p = 100, \epsilon = 0, \rho = 0.2$				$p = 100, \epsilon = 0, \rho = 0.5$			
	RMSPE	MSE	TPR	TNR	RMSPE	MSE	TPR	TNR
Lasso	0.621	$1.34 \times 10^{-3}$	1.0	0.987	0.621	$1.12 \times 10^{-3}$	1.0	0.987
RLARS	0.551	$7.15 \times 10^{-4}$	0.996	0.969	0.543	$6.74 \times 10^{-4}$	0.996	0.971
sLTS ( $\alpha = 0.75$ , 40 grids)	0.954	$4.47 \times 10^{-3}$	1.0	0.996	0.899	$4.53 \times 10^{-3}$	1.0	0.993
sparse $\gamma$ -linear reg ( $\gamma = 0.1$ )	0.564	$7.27 \times 10^{-4}$	1.0	0.878	0.565	$6.59 \times 10^{-4}$	1.0	0.908
sparse $\gamma$ -linear reg ( $\gamma = 0.5$ )	0.59	$1.0 \times 10^{-3}$	1.0	0.923	0.584	$8.47 \times 10^{-4}$	1.0	0.94
$p = 200, \epsilon = 0, \rho = 0.2$								
Methods	$p = 200, \epsilon = 0, \rho = 0.2$				$p = 200, \epsilon = 0, \rho = 0.5$			
	RMSPE	MSE	TPR	TNR	RMSPE	MSE	TPR	TNR
Lasso	0.635	$7.18 \times 10^{-4}$	1.0	0.992	0.624	$6.17 \times 10^{-4}$	1.0	0.991
RLARS	0.55	$3.63 \times 10^{-4}$	0.994	0.983	0.544	$3.48 \times 10^{-4}$	0.996	0.985
sLTS ( $\alpha = 0.75$ , 40 grids)	1.01	$3.76 \times 10^{-3}$	1.0	0.996	0.909	$2.47 \times 10^{-3}$	1.0	0.996
sparse $\gamma$ -linear reg ( $\gamma = 0.1$ )	0.584	$4.45 \times 10^{-4}$	1.0	0.935	0.573	$3.99 \times 10^{-4}$	1.0	0.938
sparse $\gamma$ -linear reg ( $\gamma = 0.5$ )	0.621	$6.55 \times 10^{-4}$	1.0	0.967	0.602	$5.58 \times 10^{-4}$	1.0	0.966

### 5.6. Computational Cost

In this subsection, we consider the CPU times for Lasso, RLARS, sLTS and our method. The data were generated from the simulation model in Section 5.1. The sample size and the number of explanatory variables were set to be  $n = 100$  and  $p = 100, 500, 1000, 2000, 5000$ , respectively. In Lasso, RLARS and sLTS, all parameters were used by default (see Section 5.3). Our method used the estimate of the RANSAC algorithm as an initial point. The number of candidates for the RANSAC algorithm was set to 1000. The parameters  $\gamma$  and  $\gamma_0$  were set to 0.1 and 0.5, respectively. No method used parallel computing methods. Figure 1 shows the average CPU times over 10 runs in seconds. All results were obtained in R Version 3.3.0 with an Intel Core i7-4790K machine. sLTS shows very high computational cost. RLARS is faster, but does not give a good estimate, as seen in Section 5.5. Our proposed method is fast enough even for  $p = 5000$ .



**Figure 1.** CPU times (in seconds).

## 6. Real Data Analyses

In this section, we use two real datasets to compare our method with comparative methods in real data analysis. We show the best result of comparative methods among some parameter situations (e.g., Section 5.3).

### 6.1. NCI-60 Cancer Cell Panel

We applied our method and comparative methods to regress protein expression on gene expression data at the cancer cell panel of the National Cancer Institute. Experimental conditions were set in the same way as in Alfons et al. [4] as follows. The gene expression data were obtained with an Affymetrix HG-U133A chip and the normalized GCRMA method, resulting in a set of  $p = 22,283$  explanatory variables. The protein expressions based on 162 antibodies were acquired via reverse-phase protein lysate arrays and  $\log_2$  transformed. One observation had to be removed since all values were missing in the gene expression data, reducing the number of observations to  $n = 59$ . Then, the KRT18 antibody was selected as the response variable because it had the largest MAD among 162 antibodies, i.e., KRT18 may include a large number of outliers. Both the protein expressions and the gene expression data can be downloaded via the web application CellMiner (<http://discover.nci.nih.gov/cellminer/>). As a measure of prediction performance, the root trimmed mean squared prediction error (RTMSPE) was computed via leave-one-out cross-validation given by

$$\text{RTMSPE} = \sqrt{\frac{1}{h} \sum_{i=1}^h (e)_{[i:n]}^2},$$

where  $e^2 = ((y_1 - x_1^T \hat{\beta}^{[-1]})^2, \dots, (y_n - x_n^T \hat{\beta}^{[-n]})^2)$  and  $(e)_{[1:n]}^2 \leq \dots \leq (e)_{[n:n]}^2$  are the order statistics of  $e^2$  and  $h = \lfloor (n+1)0.75 \rfloor$ . The choice of  $h$  is important because it is preferable for estimating prediction performance that trimmed squares does not include outliers. We set  $h$  in the same way as in Alfons et al. [4], because the sLTS detected 13 outliers in Alfons et al. [4]. In this experiment, we used the estimate of the RANSAC algorithm as an initial point instead of sLTS because sLTS required high computational cost with such high dimensional data.

Table 6 shows that our method outperformed other comparative methods for the RTMSPE with high dimensional data. Our method presented the smallest RTMSPE with the second smallest number of explanatory variables. RLARS presented the smallest number of explanatory variables, but a much larger RTMSPE than our method.

**Table 6.** Root trimmed mean squared prediction error (RTMSPE) for protein expressions based on the KRT18 antibody (NCI-60 cancer cell panel data), computed from leave-one-out cross-validation.

Methods	RTMSPE	<sup>1</sup> Selected Variables
Lasso	1.058	52
RLARS	0.936	18
sLTS	0.721	33
Our method ( $\gamma = 0.1$ )	0.679	29
Our method ( $\gamma = 0.5$ )	0.700	30

<sup>1</sup> This means the number of non-zero elements.

## 6.2. Protein Homology Dataset

We applied our method and comparative methods to the protein sequence dataset used for KDD-Cup 2004. Experimental conditions were set in the same way as in Khan et al. [2] as follows. The whole dataset consists of  $n = 145,751$  protein sequences, which has 153 blocks corresponding to native protein. Each data point in a particular block is a candidate homologous protein. There were 75 variables in the dataset: the block number (categorical) and 74 measurements of protein features. The first protein feature was used as the response variable. Then, five blocks with a total of  $n = 4141$  protein sequences were selected because they contained the highest proportions of homologous proteins (and hence, the highest proportions of potential outliers). The data of each block were split into two almost equal parts to get a training sample of size  $n_{tra} = 2072$  and a test sample of size  $n_{test} = 2069$ . The number of explanatory variables was  $p = 77$ , consisting of four block indicators (Variables 1–4) and 73 features. The whole protein, training and test dataset can be downloaded from <http://users.ugent.be/~svaelst/software/RLARS.html>. As a measure of prediction performance, the root trimmed mean squared prediction error (RTMSPE) was computed for the test sample given by:

$$\text{RTMSPE} = \sqrt{\frac{1}{h} \sum_{i=1}^h (e)_{[i:n_{test}]}^2},$$

where  $e^2 = ((y_1 - x_1^T \hat{\beta})^2, \dots, (y_{n_{test}} - x_{n_{test}}^T \hat{\beta})^2)$  and  $(e)_{[1:n_{test}]}^2 \leq \dots \leq (e)_{[n_{test}:n_{test}]}^2$  are the order statistics of  $e^2$  and  $h = \lfloor (n_{test} + 1)0.99 \rfloor, \lfloor (n_{test} + 1)0.95 \rfloor$  or  $\lfloor (n_{test} + 1)0.9 \rfloor$ . In this experiment, we used the estimate of sLTS as an initial point.

Table 7 shows that our method outperformed other comparative methods for the RTMSPE. Our method presented the smallest RTMSPE with the largest number of explanatory variables. It might seem that other methods gave a smaller number of explanatory variables than necessary.

**Table 7.** Root trimmed mean squared prediction error in the protein test set.

Methods	Trimming Fraction			
	1%	5%	10%	<sup>1</sup> Selected Variables
Lasso	10.697	9.66	8.729	22
RLARS	10.473	9.435	8.527	27
sLTS	10.614	9.52	8.575	21
Our method ( $\gamma = 0.1$ )	10.461	9.403	8.481	44
Our method ( $\gamma = 0.5$ )	10.463	9.369	8.419	42

<sup>1</sup> This means the number of non-zero elements.

## 7. Conclusions

We proposed robust and sparse regression based on the  $\gamma$ -divergence. We showed desirable robust properties under both homogeneous and heterogeneous contamination. In particular, we presented the Pythagorean relation for the regression case, although it was not shown in Kanamori and Fujisawa [11]. In most of the robust and sparse regression methods, it is difficult to obtain the efficient estimation algorithm, because the objective function is non-convex and non-differentiable. Nonetheless, we succeeded to propose the efficient estimation algorithm, which has a monotone decreasing property of the objective function by using the MM-algorithm. The numerical experiments and real data analyses suggested that our method was superior to comparative robust and sparse linear regression methods in terms of both accuracy and computational costs. However, in numerical experiments, a few results of performance measure “TNR” were a little less than the best results. Therefore, if more sparsity of coefficients is needed, other sparse penalties, e.g., the Smoothly Clipped Absolute Deviations (SCAD) [22] and the Minimax Concave Penalty (MCP)[23], can also be useful.

**Acknowledgments:** This work was supported by a Grant-in-Aid for Scientific Research of the Japan Society for the Promotion of Science.

**Author Contributions:** Takayuki Kawashima and Hironori Fujisawa contributed the theoretical analysis; Takayuki Kawashima performed the experiments; Takayuki Kawashima and Hironori Fujisawa wrote the paper. All authors have read and approved the final manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

**Proof of Theorem 1.** For two non-negative functions  $r(x, y)$  and  $u(x, y)$  and probability density function  $g(x)$ , it follows from Hölder’s inequality that:

$$\int r(x, y)u(x, y)g(x)dxdy \leq \left( \int r(x, y)^\alpha g(x)dxdy \right)^{\frac{1}{\alpha}} \left( \int u(x, y)^\beta g(x)dxdy \right)^{\frac{1}{\beta}},$$

where  $\alpha$  and  $\beta$  are positive constants and  $\frac{1}{\alpha} + \frac{1}{\beta} = 1$ . The equality holds if and only if  $r(x, y)^\alpha = \tau u(x, y)^\beta$  for a positive constant  $\tau$ . Let  $r(x, y) = g(y|x)$ ,  $u(x, y) = f(y|x)^\gamma$ ,  $\alpha = 1 + \gamma$  and  $\beta = \frac{1+\gamma}{\gamma}$ . Then, it holds that:

$$\begin{aligned} & \int \left( \int g(y|x)f(y|x)^\gamma dy \right) dg(x) \\ & \leq \left\{ \int \left( \int g(y|x)^{1+\gamma} dy \right) dg(x) \right\}^{\frac{1}{1+\gamma}} \left\{ \int \left( \int f(y|x)^{1+\gamma} dy \right) dg(x) \right\}^{\frac{\gamma}{1+\gamma}}. \end{aligned}$$

The equality holds if and only if  $g(y|x)^{1+\gamma} = \tau(f(y|x)^\gamma)^{\frac{1+\gamma}{\gamma}}$ , i.e.,  $g(y|x) = f(y|x)$  because  $g(y|x)$  and  $f(y|x)$  are conditional probability density functions. Properties (i), (ii) follow from this inequality, the equality condition and the definition of  $D_\gamma(g(y|x), f(y|x); g(x))$ .

Let us prove Property (iii). Suppose that  $\gamma$  is sufficiently small. Then, it holds that  $f^\gamma = 1 + \gamma \log f + O(\gamma^2)$ . The  $\gamma$ -divergence for regression is expressed by:

$$\begin{aligned}
& D_\gamma(g(y|x), f(y|x); g(x)) \\
&= \frac{1}{\gamma(1+\gamma)} \log \int \left\{ \int g(y|x)(1+\gamma \log g(y|x) + O(\gamma^2)) dy \right\} g(x) dx \\
&\quad - \frac{1}{\gamma} \log \int \left\{ \int g(y|x)(1 + \gamma \log f(y|x) + O(\gamma^2)) dy \right\} g(x) dx \\
&\quad + \frac{1}{1+\gamma} \log \int \left\{ \int f(y|x)(1 + \gamma \log f(y|x) + O(\gamma^2)) dy \right\} g(x) dx \\
&= \frac{1}{\gamma(1+\gamma)} \log \left\{ 1 + \gamma \int \left( \int g(y|x) \log g(y|x) dy \right) g(x) dx + O(\gamma^2) \right\} \\
&\quad - \frac{1}{\gamma} \log \left\{ 1 + \gamma \int \left( \int g(y|x) \log f(y|x) dy \right) g(x) dx + O(\gamma^2) \right\} \\
&\quad + \frac{1}{1+\gamma} \log \left\{ 1 + \gamma \int \left( \int f(y|x) \log f(y|x) dy \right) g(x) dx + O(\gamma^2) \right\} \\
&= \frac{1}{(1+\gamma)} \int \left( \int g(y|x) \log g(y|x) dy \right) g(x) dx \\
&\quad - \int \left( \int g(y|x) \log f(y|x) dy \right) g(x) dx + O(\gamma) \\
&= \int D_{KL}(g(y|x), f(y|x)) g(x) dx + O(\gamma).
\end{aligned}$$

□

**Proof of Theorem 2.** We see that:

$$\begin{aligned}
& \int \left( \int g(y|x) f(y|x; \theta)^\gamma dy \right) g(x) dx \\
&= \int \left( \int \{(1-\epsilon)f(y|x; \theta^*) + \epsilon\delta(y|x)\} f(y|x; \theta)^\gamma dy \right) g(x) dx \\
&= (1-\epsilon) \left\{ \int \left( \int f(y|x; \theta^*) f(y|x; \theta)^\gamma dy \right) g(x) dx \right\} \\
&\quad + \epsilon \left\{ \int \left( \int \delta(y|x) f(y|x; \theta)^\gamma dy \right) g(x) dx \right\}.
\end{aligned}$$

It follows from the assumption  $\epsilon < \frac{1}{2}$  that:

$$\begin{aligned}
& \left\{ \epsilon \int \left( \int \delta(y|x) f(y|x; \theta)^\gamma dy \right) g(x) dx \right\}^{\frac{1}{\gamma}} \\
&< \left\{ \frac{1}{2} \int \left( \int \delta(y|x) f(y|x; \theta)^\gamma dy \right) g(x) dx \right\}^{\frac{1}{\gamma}} \\
&< \left\{ \int \left( \int \delta(y|x) f(y|x; \theta)^\gamma dy \right) g(x) dx \right\}^{\frac{1}{\gamma}} = v_{f_\theta, \gamma}.
\end{aligned}$$

Hence,

$$\begin{aligned} & \int \left( \int g(y|x) f(y|x; \theta)^\gamma dy \right) g(x) dx = \\ & (1-\epsilon) \left\{ \int \left( \int f(y|x; \theta^*) f(y|x; \theta)^\gamma dy \right) g(x) dx \right\} \\ & + O(v_{f_\theta, \gamma}^\gamma). \end{aligned}$$

Therefore, it holds that:

$$\begin{aligned} & d_\gamma(g(y|x), f(y|x; \theta); g(x)) \\ &= -\frac{1}{\gamma} \log \int \left( \int g(y|x) f(y|x; \theta)^\gamma dy \right) g(x) dx \\ &+ \frac{1}{1+\gamma} \log \int \left( \int f(y|x; \theta)^{1+\gamma} dy \right) g(x) dx \\ &= -\frac{1}{\gamma} \log \int \left( \int f(y|x; \theta^*) f(y|x; \theta)^\gamma dy \right) g(x) dx \\ &+ \frac{1}{1+\gamma} \log \int \left( \int f(y|x; \theta)^{1+\gamma} dy \right) g(x) dx \\ &- \frac{1}{\gamma} \log(1-\epsilon) + O(v_{f_\theta, \gamma}^\gamma) \\ &= d_\gamma(f(y|x; \theta^*), f(y|x; \theta); g(x)) \\ &- \frac{1}{\gamma} \log(1-\epsilon) + O(v_{f_\theta, \gamma}^\gamma). \end{aligned}$$

Then, it follows that:

$$\begin{aligned} & D_\gamma(g(y|x), f(y|x; \theta); g(x)) - D_\gamma(g(y|x), f(y|x; \theta^*); g(x)) \\ &- D_\gamma(f(y|x; \theta^*), f(y|x; \theta); g(x)) \\ &= \{-d_\gamma(g(y|x), g(y|x); g(x)) + d_\gamma(g(y|x), f(y|x; \theta); g(x))\} \\ &- \{-d_\gamma(g(y|x), g(y|x); g(x)) + d_\gamma(g(y|x), f(y|x; \theta^*); g(x))\} \\ &- \{-d_\gamma(f(y|x; \theta^*), f(y|x; \theta^*); g(x)) + d_\gamma(f(y|x; \theta^*), f(y|x; \theta); g(x))\} \\ &= d_\gamma(g(y|x), f(y|x; \theta); g(x)) - d_\gamma(f(y|x; \theta^*), f(y|x; \theta); g(x)) \\ &- d_\gamma(g(y|x), f(y|x; \theta^*); g(x)) + d_\gamma(f(y|x; \theta^*), f(y|x; \theta^*); g(x)) \\ &= O(v^\gamma). \end{aligned}$$

□

**Proof of Theorem 3.** We see that:

$$\begin{aligned} & \int \left( \int g(y|x) f(y|x; \theta)^\gamma dy \right) g(x) dx \\ &= \left\{ \int \left( \int f(y|x; \theta^*) f(y|x; \theta)^\gamma dy \right) (1-\epsilon(x)) g(x) dx \right. \\ &\quad \left. + \int \left( \int \delta(y|x) f(y|x; \theta)^\gamma dy \right) \epsilon(x) g(x) dx \right\}. \end{aligned}$$

It follows from the assumption  $\epsilon(x) < \frac{1}{2}$  that:

$$\begin{aligned} & \left\{ \int \left( \int \delta(y|x) f(y|x; \theta)^\gamma dy \right) \epsilon(x) g(x) dx \right\}^{\frac{1}{\gamma}} \\ & < \left\{ \int \left( \int \delta(y|x) f(y|x; \theta)^\gamma dy \right) \frac{g(x)}{2} dx \right\}^{\frac{1}{\gamma}} \\ & < \left\{ \int \left( \int \delta(y|x) f(y|x; \theta)^\gamma dy \right) g(x) dx \right\}^{\frac{1}{\gamma}} = v_{f_{\theta, \gamma}}. \end{aligned}$$

Hence,

$$\begin{aligned} & \int \left( \int g(y|x) f(y|x; \theta)^\gamma dy \right) g(x) dx \\ & = \left\{ \int \left( \int f(y|x; \theta^*) f(y|x; \theta)^\gamma dy \right) (1 - \epsilon(x)) g(x) dx \right\} \\ & \quad + O(v_{f_{\theta, \gamma}}^\gamma). \end{aligned}$$

Therefore, it holds that:

$$\begin{aligned} & d_\gamma(g(y|x), f(y|x; \theta); g(x)) \\ & = -\frac{1}{\gamma} \log \int \left( \int g(y|x) f(y|x; \theta)^\gamma dy \right) g(x) dx \\ & \quad + \frac{1}{1+\gamma} \log \int \left( \int f(y|x; \theta)^{1+\gamma} dy \right) g(x) dx \\ & = -\frac{1}{\gamma} \log \left\{ \int \left( \int f(y|x; \theta^*) f(y|x; \theta)^\gamma dy \right) (1 - \epsilon(x)) g(x) dx \right\} \\ & \quad + O(v_{f_{\theta, \gamma}}^\gamma) + \frac{1}{1+\gamma} \log \int \left( \int f(y|x; \theta)^{1+\gamma} dy \right) g(x) dx \\ & = d_\gamma(f(y|x; \theta^*), f(y|x; \theta); (1 - \epsilon(x)) g(x)) + O(v_{f_{\theta, \gamma}}^\gamma) \\ & \quad - \frac{1}{1+\gamma} \log \int \left( \int f(y|x; \theta)^{1+\gamma} dy \right) (1 - \epsilon(x)) g(x) dx \\ & \quad + \frac{1}{1+\gamma} \log \int \left( \int f(y|x; \theta)^{1+\gamma} dy \right) g(x) dx \\ & = d_\gamma(f(y|x; \theta^*), f(y|x; \theta); (1 - \epsilon(x)) g(x)) \\ & \quad + O(v_{f_{\theta, \gamma}}^\gamma) - \frac{1}{1+\gamma} \log \left\{ 1 - \int \epsilon(x) g(x) dx \right\}. \end{aligned}$$

Then, it follows that:

$$\begin{aligned}
& D_\gamma(g(y|x), f(y|x;\theta); g(x)) \\
& - D_\gamma(g(y|x), f(y|x;\theta^*); g(x)) \\
& - D_\gamma(f(y|x;\theta^*), f(y|x;\theta); (1 - \epsilon(x))g(x)) \\
& = \{-d_\gamma(g(y|x), g(y|x); g(x)) + d_\gamma(g(y|x), f(y|x;\theta); g(x))\} \\
& - \{-d_\gamma(g(y|x), g(y|x); g(x)) + d_\gamma(g(y|x), f(y|x;\theta^*); g(x))\} \\
& - \{-d_\gamma(f(y|x;\theta^*), f(y|x;\theta); (1 - \epsilon(x))g(x)) \\
& \quad + d_\gamma(f(y|x;\theta^*), f(y|x;\theta); (1 - \epsilon(x))g(x))\} \\
& = d_\gamma(g(y|x), f(y|x;\theta); g(x)) \\
& - d_\gamma(f(y|x;\theta^*), f(y|x;\theta); (1 - \epsilon(x))g(x)) \\
& - d_\gamma(g(y|x), f(y|x;\theta^*); g(x)) \\
& \quad + d_\gamma(f(y|x;\theta^*), f(y|x;\theta^*); (1 - \epsilon(x))g(x)) \\
& = O(v^\gamma).
\end{aligned}$$

□

## References

1. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288.
2. Khan, J.A.; Van Aelst, S.; Zamar, R.H. Robust linear model selection based on least angle regression. *J. Am. Stat. Assoc.* **2007**, *102*, 1289–1299.
3. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499.
4. Alfons, A.; Croux, C.; Gelper, S. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Stat.* **2013**, *7*, 226–248.
5. Rousseeuw, P.J. Least Median of Squares Regression. *J. Am. Stat. Assoc.* **1984**, *79*, 871–880.
6. Windham, M.P. Robustifying model fitting. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 599–609.
7. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559.
8. Jones, M.C.; Hjort, N.L.; Harris, I.R.; Basu, A. A Comparison of related density-based minimum divergence estimators. *Biometrika* **2001**, *88*, 865–873.
9. Fujisawa, H.; Eguchi, S. Robust Parameter Estimation with a Small Bias Against Heavy Contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081.
10. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; CRC Press: Boca Raton, FL, USA, 2011.
11. Kanamori, T.; Fujisawa, H. Robust estimation under heavy contamination using unnormalized models. *Biometrika* **2015**, *102*, 559–572.
12. Cichocki, A.; Cruces, S.; Amari, S.I. Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization. *Entropy* **2011**, *13*, 134–170.
13. Samek, W.; Blythe, D.; Müller, K.R.; Kawanabe, M. Robust Spatial Filtering with Beta Divergence. In *Advances in Neural Information Processing Systems 26*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2013; pp. 1007–1015.
14. Hunter, D.R.; Lange, K. A tutorial on MM algorithms. *Am. Stat.* **2004**, *58*, 30–37.
15. Hirose, K.; Fujisawa, H. Robust sparse Gaussian graphical modeling. *J. Multivar. Anal.* **2017**, *161*, 172–190.
16. Zou, H.; Hastie, T. Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320.
17. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B* **2006**, *68*, 49–67.
18. Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; Knight, K. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 91–108.

19. Friedman, J.; Hastie, T.; Höfling, H.; Tibshirani, R. Pathwise coordinate optimization. *Ann. Appl. Stat.* **2007**, *1*, 302–332.
20. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*; Springer: New York, NY, USA, 2010.
21. Maronna, R.A.; Martin, D.R.; Yohai, V.J. *Robust Statistics: Theory and Methods*; John Wiley and Sons: Hoboken, NJ, USA, 2006.
22. Fan, J.; Li, R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360.
23. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942.



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Robust-BD Estimation and Inference for General Partially Linear Models

Chunming Zhang \* and Zhengjun Zhang

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA; zjz@stat.wisc.edu

\* Correspondence: cmzhang@stat.wisc.edu

Received: 10 October 2017; Accepted: 16 November 2017; Published: 20 November 2017

**Abstract:** The classical quadratic loss for the partially linear model (PLM) and the likelihood function for the generalized PLM are not resistant to outliers. This inspires us to propose a class of “robust-Bregman divergence (BD)” estimators of both the parametric and nonparametric components in the general partially linear model (GPLM), which allows the distribution of the response variable to be partially specified, without being fully known. Using the local-polynomial function estimation method, we propose a computationally-efficient procedure for obtaining “robust-BD” estimators and establish the consistency and asymptotic normality of the “robust-BD” estimator of the parametric component  $\beta_0$ . For inference procedures of  $\beta_0$  in the GPLM, we show that the Wald-type test statistic  $W_n$  constructed from the “robust-BD” estimators is asymptotically distribution free under the null, whereas the likelihood ratio-type test statistic  $\Lambda_n$  is not. This provides an insight into the distinction from the asymptotic equivalence (Fan and Huang 2005) between  $W_n$  and  $\Lambda_n$  in the PLM constructed from profile least-squares estimators using the non-robust quadratic loss. Numerical examples illustrate the computational effectiveness of the proposed “robust-BD” estimators and robust Wald-type test in the appearance of outlying observations.

**Keywords:** Bregman divergence; generalized linear model; local-polynomial regression; model check; nonparametric test; quasi-likelihood; semiparametric model; Wald statistic

## 1. Introduction

Semiparametric models, such as the partially linear model (PLM) and generalized PLM, play an important role in statistics, biostatistics, economics and engineering studies [1–5]. For the response variable  $Y$  and covariates  $(X, T)$ , where  $X = (X_1, \dots, X_d)^T \in \mathbb{R}^d$  and  $T \in \mathcal{T} \subseteq \mathbb{R}^D$ , the PLM, which is widely used for continuous responses  $Y$ , describes the model structure according to:

$$Y = X^T \beta_0 + \eta^0(T) + \epsilon, \quad E(\epsilon | X, T) = 0, \quad (1)$$

where  $\beta_0 = (\beta_{1,0}, \dots, \beta_{d,0})^T$  is a vector of unknown parameters and  $\eta^0(\cdot)$  is an unknown smooth function; the generalized PLM, which is more suited to discrete responses  $Y$  and extends the generalized linear model [6], assumes:

$$m(x, t) = E(Y | X = x, T = t) = F^{-1}(x^T \beta_0 + \eta^0(t)), \quad (2)$$

$$Y | (X, T) \sim \text{exponential family of distributions}, \quad (3)$$

where  $F$  is a known link function. Typically, the parametric component  $\beta_0$  is of primary interest, while the nonparametric component  $\eta^0(\cdot)$  serves as a nuisance function. For illustration clarity, this paper focuses on  $D = 1$ . An important application of PLM to brain fMRI data was given in [7] for detecting activated brain voxels in response to external stimuli. There,  $\beta_0$  corresponds to the part of hemodynamic response values, which is the object of primary interest to neuroscientists;  $\eta^0(\cdot)$  is the

slowly drifting baseline of time. Determining whether a voxel is activated or not can be formulated as testing for the linear form of hypotheses,

$$H_0 : A\beta_o = g_0 \quad \text{versus} \quad H_1 : A\beta_o \neq g_0, \quad (4)$$

where  $A$  is a given  $k \times d$  full row rank matrix and  $g_0$  is a known  $k \times 1$  vector.

Estimation of the parametric and nonparametric components of PLM and generalized PLM has received much attention in the literature. On the other hand, the existing work has some limitations: (i) The generalized PLM assumes that  $Y | (X, T)$  follows the distribution in (3), so that the likelihood function is fully available. From the practical viewpoint, results from the generalized PLM are not applicable to situations where the distribution of  $Y | (X, T)$  either departs from (3) or is incompletely known. (ii) Some commonly-used error measures, such as the quadratic loss in PLM for Gaussian-type responses (see for example [7,8]) and the (negative) likelihood function used in the generalized PLM, are not resistant to outliers. The work in [9] studied robust inference based on the kernel regression method for the generalized PLM with a canonical link, based on either the (negative) likelihood or (negative) quasi-likelihood as the error measure, and illustrated numerical examples with the dimension  $d = 1$ . However, the quasi-likelihood is not suitable for the exponential loss function (defined in Section 2.1), commonly used in machine learning and data mining. (iii) The work in [8] developed the inference of (4) for PLM, via the classical quadratic loss as the error measure, and demonstrated that the asymptotic distributions of the likelihood ratio-type statistic and Wald statistic under the null of (4) are both  $\chi_k^2$ . It remains unknown whether this conclusion holds when the tests are constructed based on robust estimators.

Without completely specifying the distribution of  $Y | (X, T)$ , we assume:

$$\text{var}(Y | X = x, T = t) = V(m(x, t)), \quad (5)$$

with a known functional form of  $V(\cdot)$ . We refer to a model specified by (2) and (5) as the “general partially linear model” (GPLM). This paper aims to develop robust estimation of GPLM and robust inference of  $\beta_o$ , allowing the distribution of  $Y | (X, T)$  to be partially specified. To introduce robust estimation, we adopt a broader class of robust error measures, called “robust-Bregman divergence (BD)” developed in [10], for a GLM, in which BD includes the quadratic loss, the (negative) quasi-likelihood, the exponential loss and many other commonly-used error measures as special cases. We propose the “robust-BD estimators” for both the parametric and nonparametric components of the GPLM. Distinct from the explicit-form estimators for PLM using the classical quadratic loss (see [8]), the “robust-BD estimators” for GPLM do not have closed-form expressions, which makes the theoretical derivation challenging. Moreover, the robust-BD estimators, as numerical solutions to non-linear optimization problems, pose key implementation challenges. Our major contributions are given below.

- The robust fitting of the nonparametric component  $\eta^o(\cdot)$  is formulated using the local-polynomial regression technique [11]. See Section 2.3.
- We develop a coordinate descent algorithm for the robust-BD estimator of  $\beta_o$ , which is computationally efficient particularly when the dimension  $d$  is large. See Section 3.
- Theorems 1 and 2 demonstrate that under the GPLM, the consistency and asymptotic normality of the proposed robust-BD estimator for  $\beta_o$  are achieved. See Section 4.
- For robust inference of  $\beta_o$ , we propose a robust version of the Wald-type test statistic  $W_n$ , based on the robust-BD estimators, and justify its validity in Theorems 3–5. It is shown to be asymptotically  $\chi^2$  (central) under the null, thus distribution free, and  $\chi^2$  (noncentral) under the contiguous alternatives. Hence, this result, when applied to the exponential loss, as well as other loss functions in the wider class of BD, is practically feasible. See Section 5.1.

- For robust inference of  $\beta_o$ , we re-examine the likelihood ratio-type test statistic  $\Lambda_n$ , constructed by replacing the negative log-likelihood with the robust-BD. Our Theorem 6 reveals that the asymptotic null distribution of  $\Lambda_n$  is generally not  $\chi^2$ , but a linear combination of independent  $\chi^2$  variables, with weights relying on unknown quantities. Even in the particular case of using the classical-BD, the limit distribution is not invariant with re-scaling the generating function of the BD. Moreover, the limit null distribution of  $\Lambda_n$  (in either the non-robust or robust version) using the exponential loss, which does not belong to the (negative) quasi-likelihood, but falls in BD, is always a weighted  $\chi^2$ , thus limiting its use in practical applications. See Section 5.2.

Simulation studies in Section 6 demonstrate that the proposed class of robust-BD estimators and robust Wald-type test either compare well with or perform better than the classical non-robust counterparts: the former is less sensitive to outliers than the latter, and both perform comparably well for non-contaminated cases. Section 7 illustrates some real data applications. Section 8 ends the paper with brief discussions. Details of technical derivations are relegated to Appendix A.

## 2. Robust-BD and Robust-BD Estimators

This section starts with a brief review of BD in Section 2.1 and “robust-BD” in Section 2.2, followed by the proposed “robust-BD” estimators of  $\eta^o(\cdot)$  and  $\beta_o$  in Sections 2.3 and 2.4.

### 2.1. Classical-BD

To broaden the scope of robust estimation and inference, we consider a class of error measures motivated from the Bregman divergence (BD). For a given concave  $q$ -function, [12] defined a bivariate function,

$$Q_q(v, \mu) = -q(v) + q(\mu) + (v - \mu)q'(\mu). \quad (6)$$

We call  $Q_q$  the BD and call  $q$  the generating  $q$ -function of the BD. For example, a function  $q(\mu) = a\mu - \mu^2$  for some constant  $a$  yields the quadratic loss  $Q_q(Y, \mu) = (Y - \mu)^2$ . For a binary response variable  $Y$ ,  $q(\mu) = \min\{\mu, (1 - \mu)\}$  gives the misclassification loss  $Q_q(Y, \mu) = I\{Y \neq I(\mu > 1/2)\}$ , where  $I(\cdot)$  is an indicator function;  $q(\mu) = -2\{\mu \log(\mu) + (1 - \mu) \log(1 - \mu)\}$  gives the Bernoulli deviance loss log-likelihood  $Q_q(Y, \mu) = -2\{Y \log(\mu) + (1 - Y) \log(1 - \mu)\}$ ;  $q(\mu) = 2 \min\{\mu, (1 - \mu)\}$  results in the hinge loss  $Q_q(Y, \mu) = \max\{1 - (2Y - 1) \text{sign}(\mu - 0.5), 0\}$  of the support vector machine;  $q(\mu) = 2\{\mu(1 - \mu)\}^{1/2}$  yields the exponential loss  $Q_q(Y, \mu) = \exp[-(Y - 0.5) \log\{\mu/(1 - \mu)\}]$  used in AdaBoost [13]. Moreover, [14] showed that if:

$$q(\mu) = \int_a^\mu \frac{s - \mu}{V(s)} ds, \quad (7)$$

with a finite constant  $a$  such that the integral is well defined, then  $Q_q(y, \mu)$  matches the “classical (negative) quasi-likelihood” function.

### 2.2. Robust-BD $\rho_q(y, \mu)$

Let  $r(y, \mu) = (y - \mu)/\sqrt{V(\mu)}$  denote the Pearson residual, which reduces to the standardized residual for linear models. In contrast to the “classical-BD”, denoted by  $Q_q$  in (6), the “robust-BD” developed in [10] for a GLM [6], is formed by:

$$\rho_q(y, \mu) = \int_y^\mu \psi(r(y, s)) \{q''(s)\sqrt{V(s)}\} ds - G(\mu), \quad (8)$$

where  $\psi(r)$  is chosen to be a bounded, odd function, such as the Huber  $\psi$ -function [15],  $\psi(r) = r \min(1, c/|r|)$ , and the bias-correction term,  $G(\mu)$ , entails the Fisher consistency of the parameter estimator and satisfies:

$$G'(\mu) = G'_1(\mu) \{q''(\mu)\sqrt{V(\mu)}\},$$

with

$$G'_1(m(\mathbf{x}, t)) = \mathbb{E}\{\psi(r(Y, m(\mathbf{x}, t))) \mid \mathbf{X} = \mathbf{x}, T = t\}. \quad (9)$$

We make the following discussions regarding features of the “robust-BD”. To facilitate the discussion, we first introduce some necessary notation. Assume that the quantities:

$$p_j(y; \theta) = \frac{\partial^j}{\partial \theta^j} \rho_q(y, F^{-1}(\theta)), \quad j = 0, 1, \dots, \quad (10)$$

exist finitely up to any order required. Then, we have the following expressions,

$$\begin{aligned} p_1(y; \theta) &= \{\psi(r(y, \mu)) - G'_1(\mu)\} \{q''(\mu) \sqrt{V(\mu)}\} / F'(\mu), \\ p_2(y; \theta) &= A_0(y, \mu) + \{\psi(r(y, \mu)) - G'_1(\mu)\} A_1(\mu), \\ p_3(y; \theta) &= A_2(y, \mu) + \{\psi(r(y, \mu)) - G'_1(\mu)\} A'_1(\mu) / F'(\mu), \end{aligned} \quad (11)$$

where  $\mu = F^{-1}(\theta)$ ,

$$A_0(y, \mu) = -\left[\psi'(r(y, \mu)) \left\{1 + \frac{y - \mu}{\sqrt{V(\mu)}} \times \frac{V'(\mu)}{2\sqrt{V(\mu)}}\right\} + G''_1(\mu) \sqrt{V(\mu)}\right] \frac{q''(\mu)}{\{F'(\mu)\}^2},$$

$A_1(\mu) = [\{q^{(3)}(\mu) \sqrt{V(\mu)} + 2^{-1}q''(\mu)V'(\mu)/\sqrt{V(\mu)}\}F'(\mu) - q''(\mu)\sqrt{V(\mu)}F''(\mu)]/\{F'(\mu)\}^3$  and  $A_2(y, \mu) = [\partial A_0(y, \mu)/\partial \mu + \partial\{\psi(r(y, \mu)) - G'_1(\mu)\}/\partial \mu A_1(\mu)]/F'(\mu)$ . Particularly,  $p_1(y; \theta)$  contains  $\psi(r)$ ;  $p_2(y; \theta)$  contains  $\psi(r)$ ,  $\psi'(r)$  and  $\psi'(r)r$ ;  $p_3(y; \theta)$  contains  $\psi(r)$ ,  $\psi'(r)$ ,  $\psi'(r)r$ ,  $\psi''(r)$ ,  $\psi''(r)r$ , and  $\psi''(r)r^2$ , where  $r = r(y, \mu) = (y - \mu)/\sqrt{V(\mu)}$  denotes the Pearson residual. Accordingly,  $\{p_j(y; \theta) : j = 1, 2, 3\}$  depend on  $y$  through  $\psi(r)$  and its derivatives coupled with  $r$ . Then, we observe from (9) and (11) that:

$$\mathbb{E}\{p_1(Y; \mathbf{X}^T \beta_o + \eta^o(T)) \mid \mathbf{X}, T\} = 0. \quad (12)$$

In the particular choice of  $\psi(r) = r$ , it is clearly noticed from (9) that  $G'_1(\cdot) = 0$ , and thus,  $G'(\cdot) = 0$ . In such a case, the proposed “robust-BD”  $\rho_q(y, \mu)$  reduces to the “classical-BD”  $Q_q(y, \mu)$ .

### 2.3. Local-Polynomial Robust-BD Estimator of $\eta^o(\cdot)$

Let  $\{(Y_i, \mathbf{X}_i, T_i)\}_{i=1}^n$  be i.i.d. observations of  $(Y, \mathbf{X}, T)$  captured by the GPLM in (2) and (5), where the dimension  $d \geq 1$  is a finite integer. From (2), it is directly observed that if the true value of  $\beta_o$  is known, then estimating  $\eta^o(\cdot)$  becomes estimating a nonparametric function; conversely, if the actual form of  $\eta^o(\cdot)$  is available, then estimating  $\beta_o$  amounts to estimating a vector parameter.

To motivate the estimation of  $\eta^o(\cdot)$  at a fitting point  $t$ , a proper way to characterize  $\eta^o(t)$  is desired. For any given value of  $\beta$ , define:

$$S(a; t, \beta) = \mathbb{E}\{\rho_q(Y, F^{-1}(\mathbf{X}^T \beta + a)) w_1(\mathbf{X}) \mid T = t\}, \quad (13)$$

where  $a$  is a scalar,  $\rho_q(y, \mu)$  is the “robust-BD” defined in (8), which aims to guard against outlying observations in the response space of  $Y$ , and  $w_1(\cdot) \geq 0$  is a given bounded weight function that downweights high leverage points in the covariate space of  $\mathbf{X}$ . See Sections 6 and 7 for an example of  $w_1(\mathbf{x})$ . Set:

$$\eta_\beta(t) = \arg \min_{a \in \mathbb{R}^1} S(a; t, \beta). \quad (14)$$

Theoretically,  $\eta^o(t) = \eta_{\beta_o}(t)$  will be assumed (in Condition A3) for obtaining asymptotically unbiased estimators of  $\eta^o(\cdot)$ . Such property indeed holds, for example, when a classical quadratic loss combined with an identity link is used in (14). Thus, we call  $\eta_\beta(\cdot)$  the “surrogate function” for  $\eta^o(\cdot)$ .

The characterization of the surrogate function  $\eta_\beta(t)$  in (14) enables us to develop its robust-BD estimator  $\hat{\eta}_\beta(t)$  based on nonparametric function estimation. Assume that

$\eta^o(\cdot)$  is  $(p + 1)$ -times continuously differentiable at the fitting point  $t$ . Denote by  $\mathbf{a}_o(t) = (\eta^o(t), (\eta^o)^{(1)}(t), \dots, (\eta^o)^{(p)}(t)/p!)^T \in \mathbb{R}^{p+1}$  the vector consisting of  $\eta^o(t)$  along with its (re-scaled) derivatives. For observed covariates  $T_i$  close to the point  $t$ , the Taylor expansion implies that:

$$\begin{aligned}\eta^o(T_i) &\approx \eta^o(t) + (T_i - t)(\eta^o)^{(1)}(t) + \dots + (T_i - t)^p(\eta^o)^{(p)}(t)/p! \\ &= \mathbf{t}_i(t)^T \mathbf{a}_o(t),\end{aligned}\quad (15)$$

where  $\mathbf{t}_i(t) = (1, (T_i - t), \dots, (T_i - t)^p)^T$ . For any given value of  $\beta$ , let  $\hat{\mathbf{a}}(t; \beta) = (\hat{a}_0(t; \beta), \hat{a}_1(t; \beta), \dots, \hat{a}_p(t; \beta))^T$  be the minimizer of the criterion function,

$$S_n(\mathbf{a}; t, \beta) = \frac{1}{n} \sum_{i=1}^n \rho_q(Y_i, F^{-1}(\mathbf{X}_i^T \beta + \mathbf{t}_i(t)^T \mathbf{a})) w_1(\mathbf{X}_i) K_h(T_i - t), \quad (16)$$

with respect to  $\mathbf{a} \in \mathbb{R}^{p+1}$ , where  $K_h(\cdot) = K(\cdot/h)/h$  is re-scaled from a kernel function  $K$  and  $h > 0$  is termed a bandwidth parameter. The first entry of  $\hat{\mathbf{a}}(t; \beta)$  supplies the local-polynomial robust-BD estimator  $\hat{\eta}_\beta(t)$  of  $\eta_\beta(t)$ , i.e.,

$$\hat{\eta}_\beta(t) = \mathbf{e}_{1,p+1}^T \left\{ \arg \min_{\mathbf{a} \in \mathbb{R}^{p+1}} S_n(\mathbf{a}; t, \beta) \right\}, \quad (17)$$

where  $\mathbf{e}_{j,p+1}$  denotes the  $j$ -th column of a  $(p + 1) \times (p + 1)$  identity matrix.

It is noted that the reliance of  $\hat{\eta}_\beta(t)$  on  $\beta$  does not guarantee its consistency to  $\eta^o(t)$ . Nonetheless, it is anticipated from the uniform consistency of  $\hat{\eta}_\beta$  in Lemma 1 that  $\hat{\eta}_\beta(t)$  will offer a valid estimator of  $\eta^o(t)$ , provided that  $\hat{\beta}$  consistently estimates  $\beta_o$ . Section 2.4 will discuss our proposed robust-BD estimator  $\hat{\beta}$ . Furthermore, Lemma 1 will assume (in Condition A1) that  $\eta_\beta(t)$  is the unique minimizer of  $S(\mathbf{a}; t, \beta)$  with respect to  $\mathbf{a}$ .

**Remark 1.** The case of using the “kernel estimation”, or locally-constant estimation, corresponds to the choice of degree  $p = 0$  in (15). In that case, the criterion function in (16) and the estimator in (17) reduce to:

$$S_n(\mathbf{a}; t, \beta) = \frac{1}{n} \sum_{i=1}^n \rho_q(Y_i, F^{-1}(\mathbf{X}_i^T \beta + \mathbf{a})) w_1(\mathbf{X}_i) K_h(T_i - t), \quad (18)$$

$$\hat{\eta}_\beta(t) = \arg \min_{\mathbf{a} \in \mathbb{R}^d} S_n(\mathbf{a}; t, \beta), \quad (19)$$

respectively.

#### 2.4. Robust-BD Estimator of $\beta_o$

For any given value of  $\beta$ , define:

$$J(\beta, \eta_\beta) = E\{\rho_q(Y, F^{-1}(\mathbf{X}^T \beta + \eta_\beta(T))) w_2(\mathbf{X})\}, \quad (20)$$

where  $\eta_\beta(\cdot)$  is as defined in (14) and  $w_2(\cdot)$  plays the same role as  $w_1(\cdot)$  in (13). Theoretically, it is anticipated that:

$$\beta_o = \arg \min_{\beta \in \mathbb{R}^d} J(\beta, \eta_\beta), \quad (21)$$

which holds for example in the case where a classical quadratic loss combined with an identity link is used. To estimate  $\beta_o$ , it is natural to replace (20) by its sample-based criterion,

$$J_n(\beta, \hat{\eta}_\beta) = \frac{1}{n} \sum_{i=1}^n \rho_q(Y_i, F^{-1}(\mathbf{X}_i^T \beta + \hat{\eta}_\beta(T_i))) w_2(\mathbf{X}_i), \quad (22)$$

where  $\hat{\eta}_\beta(\cdot)$  is as defined in (17). Hence, a parametric estimator of  $\beta_o$  is provided by:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^d} J_n(\beta, \hat{\eta}_\beta). \quad (23)$$

Finally, the estimator of  $\eta^o(\cdot)$  is given by:

$$\hat{\eta}(\cdot) = \hat{\eta}_{\hat{\beta}}(\cdot).$$

To achieve asymptotic normality of  $\hat{\beta}$ , Theorem 2 assumes (in Condition A2) that  $\beta_o$  is the unique minimizer in (21), a standard condition for consistent  $M$ -estimators [16].

As a comparison, it is seen that  $w_1(\cdot)$  in (16) is used to robustify covariates  $X_i$  in estimating  $\eta^o(\cdot)$ ,  $w_2(\cdot)$  in (22) is used to robustify covariates  $X_i$  in estimating  $\beta_o$  and  $\rho_q(\cdot, \cdot)$  serves to robustify the responses  $Y_i$  in both estimating procedures.

### 3. Two-Step Iterative Algorithm for Robust-BD Estimation

In a special case of using the classical quadratic loss combined with an identity link function, the robust-BD estimators for parametric and nonparametric components have explicit expressions,

$$\hat{\beta} = (\tilde{\mathbf{X}}^T \mathbf{w}_2 \tilde{\mathbf{X}})^{-1} (\tilde{\mathbf{X}}^T \mathbf{w}_2 \tilde{\mathbf{y}}), \quad (\hat{\eta}(T_1), \dots, \hat{\eta}(T_n))^T = S_h(\mathbf{y} - \mathbf{X}\hat{\beta}), \quad (24)$$

where  $\mathbf{w}_2 = \text{diag}(w_2(X_1), \dots, w_2(X_n))$ ,  $\tilde{\mathbf{y}} = (\mathbf{I} - S_h)\mathbf{y}$ ,  $\tilde{\mathbf{X}} = (\mathbf{I} - S_h)\mathbf{X}$ , with  $\mathbf{I}$  being an identity matrix,  $\mathbf{y} = (Y_1, \dots, Y_n)^T$ ,  $\mathbf{X} = (X_1, \dots, X_n)^T$  the design matrix,

$$S_h = \begin{pmatrix} \mathbf{e}_{1,p+1}^T [\{\mathbf{T}(T_1)\}^T \mathbf{W}_{w_1;K}(T_1) \mathbf{T}(T_1)]^{-1} \{\mathbf{T}(T_1)\}^T \mathbf{W}_{w_1;K}(T_1) \\ \vdots \\ \mathbf{e}_{1,p+1}^T [\{\mathbf{T}(T_n)\}^T \mathbf{W}_{w_1;K}(T_n) \mathbf{T}(T_n)]^{-1} \{\mathbf{T}(T_n)\}^T \mathbf{W}_{w_1;K}(T_n) \end{pmatrix},$$

and:

$$\mathbf{T}(t) = (\mathbf{t}_1(t), \dots, \mathbf{t}_n(t))^T, \quad \mathbf{W}_{w_1;K}(t) = \text{diag}\{w_1(X_i) K_h(T_i - t) : i = 1, \dots, n\}.$$

When  $w_1(x) = w_2(x) \equiv 1$ , (24) reduces to the “profile least-squares estimators” of [8].

In other cases, robust-BD estimators from (17) and (23) do not have closed-form expressions and need to be solved numerically, which are computationally challenging and intensive. We now discuss a two-step robust proposal for iteratively estimating  $\beta_o$  and  $\eta^o(\cdot)$ . Let  $\hat{\beta}^{[k-1]}$  and  $\{\hat{\eta}^{[k-1]}(T_i)\}_{i=1}^n$  denote the estimates in the  $(k-1)$ -th iteration, where  $\hat{\eta}^{[k-1]}(\cdot) = \hat{\eta}_{\hat{\beta}^{[k-1]}}(\cdot)$ . The  $k$ -th iteration consists of two steps below.

Step 1: Instead of solving (23) directly, we propose to solve a surrogate optimization problem,  $\hat{\beta}^{[k]} = \arg \min_{\beta \in \mathbb{R}^d} J_n(\beta, \hat{\eta}^{[k-1]})$ . This minimizer approximates  $\hat{\beta}$ .

Step 2: Obtain  $\hat{\eta}^{[k]}(T_i) = \hat{\eta}_{\hat{\beta}^{[k]}}(T_i)$ ,  $i = 1, \dots, n$ , where  $\hat{\eta}_\beta(t)$  is defined in (17).

The algorithm terminates provided that  $\|\hat{\beta}^{[k]} - \hat{\beta}^{[k-1]}\|$  is below some pre-specified threshold value, and all  $\{\hat{\eta}^{[k]}(T_i)\}_{i=1}^n$  stabilize.

#### 3.1. Step 1

For the above two-step algorithm, we first elaborate on the procedure of acquiring  $\hat{\beta}^{[k]}$  in Step 1, by extending the coordinate descent (CD) iterative algorithm [17] designed for penalized estimation to our current robust-BD estimation, which is computationally efficient. For any given value of  $\eta$ ,

by Taylor expansion, around some initial estimate  $\beta^*$  (for example,  $\hat{\beta}^{[k-1]}$ ), we obtain the weighted quadratic approximation,

$$p_q(Y_i, F^{-1}(\mathbf{X}_i^T \beta + \eta)) \approx \frac{1}{2} s_i^I (Z_i^I - \mathbf{X}_i^T \beta)^2 + C_i,$$

where  $C_i$  is a constant not depending on  $\beta$ ,

$$\begin{aligned} s_i^I &= p_2(Y_i; \mathbf{X}_i^T \beta^* + \eta), \\ Z_i^I &= \mathbf{X}_i^T \beta^* - p_1(Y_i; \mathbf{X}_i^T \beta^* + \eta) / p_2(Y_i; \mathbf{X}_i^T \beta^* + \eta), \end{aligned}$$

with  $p_j(y; \theta)$  defined in (10). Hence,

$$\begin{aligned} J_n(\beta, \eta) &= \frac{1}{n} \sum_{i=1}^n p_q(Y_i, F^{-1}(\mathbf{X}_i^T \beta + \eta)) w_2(\mathbf{X}_i) \\ &\approx \frac{1}{2} \sum_{i=1}^n \left\{ n^{-1} s_i^I w_2(\mathbf{X}_i) \right\} (Z_i^I - \mathbf{X}_i^T \beta)^2 + \text{constant}. \end{aligned}$$

Thus it suffices to conduct minimization of  $\sum_{i=1}^n s_i^I w_2(\mathbf{X}_i) (Z_i^I - \mathbf{X}_i^T \beta)^2$  with respect to  $\beta$ , using a coordinate descent (CD) updating procedure. Suppose that the current estimate is  $\hat{\beta}^{\text{old}} = (\hat{\beta}_1^{\text{old}}, \dots, \hat{\beta}_d^{\text{old}})^T$ , with the current residual vector  $\hat{r}^{\text{old}} = (\hat{r}_1^{\text{old}}, \dots, \hat{r}_n^{\text{old}})^T = z^I - \mathbf{X}\hat{\beta}^{\text{old}}$ , where  $z^I = (Z_1^I, \dots, Z_n^I)^T$  is the vector of pseudo responses. Adopting the Newton–Raphson algorithm, the estimate of the  $j$ -th coordinate based on the previous estimate  $\hat{\beta}_j^{\text{old}}$  is updated to:

$$\hat{\beta}_j^{\text{new}} = \hat{\beta}_j^{\text{old}} + \frac{\sum_{i=1}^n \{s_i^I w_2(\mathbf{X}_i)\} \hat{r}_i^{\text{old}} X_{i,j}}{\sum_{i=1}^n \{s_i^I w_2(\mathbf{X}_i)\} X_{i,j}^2}.$$

As a result, the residuals due to such an update are updated to:

$$\hat{r}_i^{\text{new}} = \hat{r}_i^{\text{old}} - X_{i,j} (\hat{\beta}_j^{\text{new}} - \hat{\beta}_j^{\text{old}}), \quad i = 1, \dots, n.$$

Cycling through  $j = 1, \dots, d$ , we obtain the estimate  $\hat{\beta}^{\text{new}} = (\hat{\beta}_1^{\text{new}}, \dots, \hat{\beta}_d^{\text{new}})^T$ . Now, we set  $\eta = \hat{\eta}^{[k-1]}$  and  $\beta^* = \hat{\beta}^{[k-1]}$ . Iterate the process of weighted quadratic approximation followed by the CD updating, for a number of times, until the estimate  $\hat{\beta}^{\text{new}}$  stabilizes to the solution  $\hat{\beta}^{[k]}$ .

The validity of  $\hat{\beta}^{[k]}$  in Step 1 converging to the true parameter  $\beta_o$  is justified as follows. (i) Standard results for M-estimation [16] indicate that the minimizer of  $J_n(\beta, \eta_{\beta_o})$  is consistent with  $\beta_o$ . (ii) According to our Theorem 1 (ii) in Section 4.1,  $\sup_{t \in \mathcal{T}} |\hat{\eta}_{\hat{\beta}}(t) - \eta_{\beta_o}(t)| \xrightarrow{P} 0$  for a compact set  $\mathcal{T}$ , where  $\xrightarrow{P}$  stands for convergence in probability. Using derivations similar to those of (A4) gives  $\sup_{\beta \in \mathcal{K}} |J_n(\beta, \hat{\eta}_{\hat{\beta}}) - J_n(\beta, \eta_{\beta_o})| \xrightarrow{P} 0$  for any compact set  $\mathcal{K}$ . Thus, minimizing  $J_n(\beta, \hat{\eta}_{\hat{\beta}})$  is asymptotically equivalent to minimizing  $J_n(\beta, \eta_{\beta_o})$ . (iii) Similarly, provided that  $\hat{\beta}^{[k-1]}$  is close to  $\hat{\beta}$ , minimizing  $J_n(\beta, \hat{\eta}_{\hat{\beta}^{[k-1]}})$  is asymptotically equivalent to minimizing  $J_n(\beta, \hat{\eta}_{\hat{\beta}})$ . Assembling these three results with the definition of  $\hat{\beta}^{[k]}$  yields:

$$\begin{aligned} \hat{\beta}^{[k]} &= \arg \min_{\beta} J_n(\beta, \hat{\eta}_{\hat{\beta}^{[k-1]}}) \\ &= \arg \min_{\beta} J_n(\beta, \hat{\eta}_{\hat{\beta}}) + o_p(1) \\ &= \arg \min_{\beta} J_n(\beta, \eta_{\beta_o}) + o_p(1) \\ &= \beta_o + o_p(1). \end{aligned}$$

### 3.2. Step 2

In Step 2, obtaining  $\hat{\eta}_{\beta}(t)$  for any given values of  $\beta$  and  $t$  is equivalent to minimizing  $S_n(\mathbf{a}; t, \beta)$  in (16). Notice that the dimension  $(p + 1)$  of  $\mathbf{a}$  is typically low, with degrees  $p = 0$  or  $p = 1$  being the most commonly used in practice. Hence, the minimizer of  $S_n(\mathbf{a}; t, \beta)$  can be obtained by directly applying the Newton–Raphson iteration: for  $k = 0, 1, \dots$ ,

$$\mathbf{a}^{[k+1]}(t; \beta) = \mathbf{a}^{[k]}(t; \beta) - \left\{ \frac{\partial^2 S_n(\mathbf{a}; t, \beta)}{\partial \mathbf{a} \partial \mathbf{a}^T} \Big|_{\mathbf{a}=\mathbf{a}^{[k]}(t; \beta)} \right\}^{-1} \frac{\partial S_n(\mathbf{a}; t, \beta)}{\partial \mathbf{a}} \Big|_{\mathbf{a}=\mathbf{a}^{[k]}(t; \beta)},$$

where  $\mathbf{a}^{[k]}(t; \beta)$  denotes the estimate in the  $k$ -th iteration, and:

$$\begin{aligned} \frac{\partial S_n(\mathbf{a}; t, \beta)}{\partial \mathbf{a}} &= \frac{1}{n} \sum_{i=1}^n p_1(Y_i; \mathbf{X}_i^T \beta + \mathbf{t}_i(t)^T \mathbf{a}) \mathbf{t}_i(t) w_1(\mathbf{X}_i) K_h(T_i - t), \\ \frac{\partial^2 S_n(\mathbf{a}; t, \beta)}{\partial \mathbf{a} \partial \mathbf{a}^T} &= \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \mathbf{X}_i^T \beta + \mathbf{t}_i(t)^T \mathbf{a}) \mathbf{t}_i(t) \mathbf{t}_i(t)^T w_1(\mathbf{X}_i) K_h(T_i - t). \end{aligned}$$

The iterations terminate until the estimate  $\hat{\eta}^{[k+1]}(t) = e_{1,p+1}^T \mathbf{a}^{[k+1]}(t; \beta)$  stabilizes.

Our numerical studies of the robust-BD estimation indicate that (i) the kernel regression method can be both faster and stabler than the local-linear method; (ii) to estimate the nonparametric component  $\eta^o(\cdot)$ , the local-linear method outperforms the kernel method, especially at the edges of points  $\{T_i\}_{i=1}^n$ ; (iii) for the performance of the robust estimation of  $\beta_o$ , which is of major interest, there is a relatively negligible difference between choices of using the kernel and local-linear methods in estimating nonparametric components.

## 4. Asymptotic Property of the Robust-BD Estimators

This section investigates the asymptotic behavior of robust-BD estimators  $\hat{\beta}$  and  $\hat{\eta}_{\beta}$ , under regularity conditions. The consistency of  $\hat{\beta}$  to  $\beta_o$  and uniform consistency of  $\hat{\eta}_{\beta}$  to  $\eta^o$  are given in Theorem 1; the asymptotic normality of  $\hat{\beta}$  is obtained in Theorem 2. For the sake of exposition, the asymptotic results will be derived using local-linear estimation with degree  $p = 1$ . Analogous results can be obtained for local-polynomial methods with lengthier technical details and are omitted.

We assume that  $T \in \mathcal{T}$ , and let  $\mathcal{T}_0 \subseteq \mathcal{T}$  be a compact set. For any continuous function  $v : \mathcal{T} \mapsto \mathbb{R}$ , define  $\|v\|_\infty = \sup_{t \in \mathcal{T}} |v(t)|$  and  $\|v\|_{\mathcal{T}_0, \infty} = \sup_{t \in \mathcal{T}_0} |v(t)|$ . For a matrix  $M$ , the smallest and largest eigenvalues are denoted by  $\lambda_j(M)$ ,  $\lambda_{\min}(M)$  and  $\lambda_{\max}(M)$ , respectively. Let  $\|M\| = \sup_{\|\mathbf{x}\|=1} \|M\mathbf{x}\| = \{\lambda_{\max}(M^T M)\}^{1/2}$  be the matrix  $L_2$  norm. Denote by  $\xrightarrow{P}$  convergence in probability and  $\xrightarrow{D}$  convergence in distribution.

### 4.1. Consistency

We first present Lemma 1, which states the uniform consistency of  $\hat{\eta}_{\beta}(\cdot)$  to the surrogate function  $\eta_{\beta}(\cdot)$ . Theorem 1 gives the consistency of  $\hat{\beta}$  and  $\hat{\eta}_{\beta}$ .

**Lemma 1** (For the non-parametric surrogate  $\eta_{\beta}(\cdot)$ ). *Let  $\mathcal{K} \subseteq \mathbb{R}^d$  and  $\mathcal{T}_0 \subseteq \mathcal{T}$  be compact sets. Assume Condition A1 and Condition B in the Appendix. If  $n \rightarrow \infty$ ,  $h \rightarrow 0$ ,  $nh \rightarrow \infty$ ,  $\log(1/h)/(nh) \rightarrow 0$ , then  $\sup_{\beta \in \mathcal{K}} \|\hat{\eta}_{\beta} - \eta_{\beta}\|_{\mathcal{T}_0, \infty} \xrightarrow{P} 0$ .*

**Theorem 1** (For  $\beta_o$  and  $\eta^o(\cdot)$ ). *Assume conditions in Lemma 1.*

(i) *If there exists a compact set  $\mathcal{K}_1$  such that  $\lim_{n \rightarrow \infty} P(\hat{\beta} \in \mathcal{K}_1) = 1$  and Condition A2 holds, then  $\hat{\beta} \xrightarrow{P} \beta_o$ .*

(ii) Moreover, if Condition A3 holds, then  $\|\widehat{\eta}_{\beta} - \eta^o\|_{T_0, \infty} \xrightarrow{P} 0$ .

#### 4.2. Asymptotic Normality

The asymptotic normality of  $\widehat{\beta}$  is provided in Theorem 2.

**Theorem 2** (For the parametric part  $\beta_o$ ). Assume Conditions A and Condition B in the Appendix. If  $n \rightarrow \infty$ ,  $nh^4 \rightarrow 0$  and  $\log(1/h)/(nh^2) \rightarrow 0$ , then:

$$\sqrt{n}(\widehat{\beta} - \beta_o) \xrightarrow{D} N(\mathbf{0}, \mathbf{H}_0^{-1} \Omega_0^* \mathbf{H}_0^{-1}),$$

where:

$$\mathbf{H}_0 = E \left[ p_2(Y; \mathbf{X}^T \beta_o + \eta^o(T)) \left\{ \mathbf{X} + \frac{\partial \eta_\beta(T)}{\partial \beta} \Big|_{\beta=\beta_o} \right\} \left\{ \mathbf{X} + \frac{\partial \eta_\beta(T)}{\partial \beta} \Big|_{\beta=\beta_o} \right\}^T w_2(\mathbf{X}) \right], \quad (25)$$

and:

$$\begin{aligned} \Omega_0^* &= E \left( p_1^2(Y; \mathbf{X}^T \beta_o + \eta^o(T)) \left[ \left\{ \mathbf{X} + \frac{\partial \eta_\beta(T)}{\partial \beta} \Big|_{\beta=\beta_o} \right\} w_2(\mathbf{X}) - \frac{\gamma(T)}{g_2(T; T, \beta_o)} w_1(\mathbf{X}) \right] \right. \\ &\quad \times \left. \left[ \left\{ \mathbf{X} + \frac{\partial \eta_\beta(T)}{\partial \beta} \Big|_{\beta=\beta_o} \right\} w_2(\mathbf{X}) - \frac{\gamma(T)}{g_2(T; T, \beta_o)} w_1(\mathbf{X}) \right]^T \right) \end{aligned} \quad (26)$$

with:

$$\begin{aligned} \gamma(t) &= E \left[ p_2(Y; \mathbf{X}^T \beta_o + \eta^o(t)) \left\{ \mathbf{X} + \frac{\partial \eta_\beta(t)}{\partial \beta} \Big|_{\beta=\beta_o} \right\} w_2(\mathbf{X}) \Big| T = t \right], \\ g_2(t; t, \beta) &= E \{ p_2(Y; \mathbf{X}^T \beta + \eta_\beta(t)) w_1(\mathbf{X}) \mid T = t \}. \end{aligned}$$

From Condition A1, (13) and (14), we can show that if  $w_1(\cdot) \equiv Cw_2(\cdot)$  for some constant  $C \in (0, \infty)$ , then  $\gamma(t) = 0$ . In that case,  $\Omega_0^* = \Omega_0$ , where:

$$\Omega_0 = E \left[ p_1^2(Y; \mathbf{X}^T \beta_o + \eta^o(T)) \left\{ \mathbf{X} + \frac{\partial \eta_\beta(T)}{\partial \beta} \Big|_{\beta=\beta_o} \right\} \left\{ \mathbf{X} + \frac{\partial \eta_\beta(T)}{\partial \beta} \Big|_{\beta=\beta_o} \right\}^T w_2^2(\mathbf{X}) \right]. \quad (27)$$

Consider the conventional PLM in (1), estimated using the classical quadratic loss, identity link and  $w_1(\cdot) = w_2(\cdot) \equiv 1$ . If  $\text{var}(\epsilon \mid X, T) \equiv \sigma^2$ , then  $\mathbf{H}_0^{-1} \Omega_0 \mathbf{H}_0^{-1} = \sigma^2 [E\{\text{var}(X \mid T)\}]^{-1}$ , and thus, the result of Theorem 2 agrees with that in [18].

**Remark 2.** Theorem 2 implies the root- $n$  convergence rate of  $\widehat{\beta}$ . This differs from  $\widehat{\eta}_{\beta}(t)$ , which converges at some rate incorporating both the sample size  $n$  and the bandwidth  $h$ , as seen in the proofs of Lemma 1 and Theorem 2.

#### 5. Robust Inference for $\beta_o$ Based on BD

In many statistical applications, we will check whether or not a subset of explanatory variables used is statistically significant. Specific examples include:

$$\begin{aligned} H_0 : \beta_{j_0} &= 0, & \text{for } j = j_0, \\ H_0 : \beta_{j_0} &= 0, & \text{for } j = j_1, \dots, j_2. \end{aligned}$$

These forms of linear hypotheses for  $\beta_o$  can be more generally formulated as: (4).

### 5.1. Wald-Type Test $W_n$

We propose a robust version of the Wald-type test statistic,

$$W_n = n(\mathbf{A}\hat{\beta} - g_0)^T (\mathbf{A}\hat{\mathbf{H}}_0^{-1}\hat{\Omega}_0^*\hat{\mathbf{H}}_0^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\beta} - g_0), \quad (28)$$

based on the robust-BD estimator  $\hat{\beta}$  proposed in Section 2.4, where  $\hat{\Omega}_0^*$  and  $\hat{\mathbf{H}}_0$  are estimates of  $\Omega_0^*$  and  $\mathbf{H}_0$  satisfying  $\hat{\mathbf{H}}_0^{-1}\hat{\Omega}_0^*\hat{\mathbf{H}}_0^{-1} \xrightarrow{P} \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$ . For example,

$$\hat{\mathbf{H}}_0 = \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \mathbf{X}_i^T \hat{\beta} + \hat{\eta}_{\hat{\beta}}(T_i)) \left\{ \mathbf{X}_i + \frac{\partial \hat{\eta}_{\hat{\beta}}(T_i)}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right\} \left\{ \mathbf{X}_i + \frac{\partial \hat{\eta}_{\hat{\beta}}(T_i)}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right\}^T w_2(\mathbf{X}_i),$$

and:

$$\begin{aligned} \hat{\Omega}_0^* &= \frac{1}{n} \sum_{i=1}^n p_1^2(Y_i; \mathbf{X}_i^T \hat{\beta} + \hat{\eta}_{\hat{\beta}}(T_i)) \left[ \left\{ \mathbf{X}_i + \frac{\partial \hat{\eta}_{\hat{\beta}}(T_i)}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right\} w_2(\mathbf{X}_i) - \frac{\hat{\gamma}(T_i)}{\hat{g}_2(T_i; T_i, \hat{\beta})} w_1(\mathbf{X}_i) \right] \\ &\quad \times \left[ \left\{ \mathbf{X}_i + \frac{\partial \hat{\eta}_{\hat{\beta}}(T_i)}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right\} w_2(\mathbf{X}_i) - \frac{\hat{\gamma}(T_i)}{\hat{g}_2(T_i; T_i, \hat{\beta})} w_1(\mathbf{X}_i) \right]^T, \end{aligned}$$

fulfill the requirement, where:

$$\begin{aligned} \frac{\partial \hat{\eta}_{\hat{\beta}}(t)}{\partial \beta} &= -\frac{\sum_{k=1}^n p_2(Y_k; \mathbf{X}_k^T \hat{\beta} + \hat{\eta}_{\hat{\beta}}(t)) \mathbf{X}_k w_1(\mathbf{X}_k) K_h(T_k - t)}{\sum_{k=1}^n p_2(Y_k; \mathbf{X}_k^T \hat{\beta} + \hat{\eta}_{\hat{\beta}}(t)) w_1(\mathbf{X}_k) K_h(T_k - t)}, \\ \hat{\gamma}(t) &= \frac{1}{n} \sum_{k=1}^n p_2(Y_k; \mathbf{X}_k^T \hat{\beta} + \hat{\eta}_{\hat{\beta}}(t)) \left\{ \mathbf{X}_k + \frac{\partial \hat{\eta}_{\hat{\beta}}(t)}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right\} w_2(\mathbf{X}_k) K_h(T_k - t), \\ \hat{g}_2(t; t, \beta) &= \frac{1}{n} \sum_{k=1}^n p_2(Y_k; \mathbf{X}_k^T \hat{\beta} + \hat{\eta}_{\hat{\beta}}(t)) w_1(\mathbf{X}_k) K_h(T_k - t). \end{aligned}$$

Again, we can verify that if  $w_1(\cdot) \equiv Cw_2(\cdot)$  for some constant  $C \in (0, \infty)$  and  $\hat{\eta}_{\hat{\beta}}(t)$  is obtained from kernel estimation method, then  $\hat{\gamma}(t) = \mathbf{0}$ , and hence,  $\hat{\Omega}_0^* = \hat{\Omega}_0$ , where:

$$\hat{\Omega}_0 = \frac{1}{n} \sum_{i=1}^n p_1^2(Y_i; \mathbf{X}_i^T \hat{\beta} + \hat{\eta}_{\hat{\beta}}(T_i)) \left\{ \mathbf{X}_i + \frac{\partial \hat{\eta}_{\hat{\beta}}(T_i)}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right\} \left\{ \mathbf{X}_i + \frac{\partial \hat{\eta}_{\hat{\beta}}(T_i)}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right\}^T w_2^2(\mathbf{X}_i).$$

Theorem 3 justifies that under the null,  $W_n$  would for large  $n$  be distributed as  $\chi_k^2$ , thus asymptotically distribution-free.

**Theorem 3** (Wald-type test based on robust-BD under  $H_0$ ). *Assume conditions in Theorem 2, and  $\hat{\mathbf{H}}_0^{-1}\hat{\Omega}_0^*\hat{\mathbf{H}}_0^{-1} \xrightarrow{P} \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$  in (28). Then, under  $H_0$  in (4), we have that:*

$$W_n \xrightarrow{D} \chi_k^2.$$

Theorem 4 indicates that  $W_n$  has a non-trivial local power detecting contiguous alternatives approaching the null at the rate  $n^{-1/2}$ :

$$H_{1n} : \mathbf{A}\beta_o - g_0 = c/\sqrt{n} \{1 + o(1)\}, \quad (29)$$

where  $c = (c_1, \dots, c_k)^T \neq \mathbf{0}$ .

**Theorem 4** (Wald-type test based on robust-BD under  $H_{1n}$ ). *Assume conditions in Theorem 2, and  $\hat{\mathbf{H}}_0^{-1}\hat{\Omega}_0^*\hat{\mathbf{H}}_0^{-1} \xrightarrow{P} \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$  in (28). Then, under  $H_{1n}$  in (29),  $W_n \xrightarrow{D} \chi_k^2(\tau^2)$ , where  $\tau^2 = c^T (\mathbf{A}\mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}\mathbf{A}^T)^{-1} c > 0$ .*

To appreciate the discriminating power of  $W_n$  in assessing the significance, the asymptotic power is analyzed. Theorem 5 manifests that under the fixed alternative  $H_1$ ,  $W_n \xrightarrow{P} +\infty$  at the rate  $n$ . Thus,  $W_n$  has the power approaching to one against fixed alternatives.

**Theorem 5** (Wald-type test based on robust-BD under  $H_1$ ). *Assume conditions in Theorem 2, and  $\widehat{\mathbf{H}}_0^{-1}\widehat{\Omega}_0^*\widehat{\mathbf{H}}_0^{-1} \xrightarrow{P} \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$  in (28). Then, under  $H_1$  in (4),  $n^{-1}W_n \geq \lambda_{\max}^{-1}(\mathbf{A}\mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}\mathbf{A}^T)\|\mathbf{A}\beta_o - g_0\|^2 + o_p(1)$ .*

For the conventional PLM in (1) estimated using the non-robust quadratic loss, [8] showed the asymptotic equivalence between the Wald-type test and likelihood ratio-type test. Our results in the next Section 5.2 reveal that such equivalence is violated when estimators are obtained using the robust loss functions.

### 5.2. Likelihood Ratio-Type Test $\Lambda_n$

This section explores the degree to which the likelihood ratio-type test is extended to the “robust-BD” for testing the null hypothesis in (4) for the GPLM. The robust-BD test statistic is:

$$\Lambda_n = 2n \left\{ \min_{\beta \in \mathbb{R}^d: \mathbf{A}\beta = g_0} J_n(\beta, \widehat{\eta}_\beta) - J_n(\widehat{\beta}, \widehat{\eta}_{\widehat{\beta}}) \right\}, \quad (30)$$

where  $\widehat{\beta}$  is the robust-BD estimator for  $\beta_o$  developed in Section 2.4.

Theorem 6 indicates that the limit distribution of  $\Lambda_n$  under  $H_0$  is a linear combination of independent chi-squared variables, with weights relying on some unknown quantities, thus not distribution free.

**Theorem 6** (Likelihood ratio-type test based on robust-BD under  $H_0$ ). *Assume conditions in Theorem 2.*

- (i) *Under  $H_0$  in (4), we obtain:*

$$\Lambda_n \xrightarrow{D} \sum_{j=1}^k \lambda_j \{(\mathbf{A}\mathbf{H}_0^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{V}_0\mathbf{A}^T)\} Z_j^2,$$

where  $\mathbf{V}_0 = \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$  and  $\{Z_j\}_{j=1}^k \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ .

- (ii) *Moreover, if  $\psi(r) = r$ ,  $w_1(x) = w_2(x) \equiv 1$ , and the generating q-function of BD satisfies:*

$$q''(m(x, t)) = -\frac{C}{V(m(x, t))}, \quad \text{for a constant } C > 0, \quad (31)$$

*then under  $H_0$  in (4), we have that  $\Lambda_n/C \xrightarrow{D} \chi_k^2$ .*

Theorem 7 states that  $\Lambda_n$  has non-trivial local power for identifying contiguous alternatives approaching the null at rate  $n^{-1/2}$  and that  $\Lambda_n \xrightarrow{P} +\infty$  at the rate  $n$  under  $H_1$ , thus having the power approaching to one against fixed alternatives.

**Theorem 7** (Likelihood ratio-type test based on robust-BD under  $H_{1n}$  and  $H_1$ ). *Assume conditions in Theorem 2. Let  $\mathbf{V}_0 = \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$  and  $\lambda_j = \lambda_j\{(\mathbf{A}\mathbf{H}_0^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{V}_0\mathbf{A}^T)\}$ ,  $j = 1, \dots, k$ .*

- (i) *Under  $H_{1n}$  in (29),  $\Lambda_n \xrightarrow{D} \sum_{j=1}^k (\sqrt{\lambda_j} Z_j + e_{j,k}^T S c)^2$ , where  $\{Z_j\}_{j=1}^k \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ , and  $S$  is a matrix satisfying  $S^T S = (\mathbf{A}\mathbf{H}_0^{-1}\mathbf{A}^T)^{-1}$  and  $S(\mathbf{A}\mathbf{V}_0\mathbf{A}^T)S^T = \text{diag}(\lambda_1, \dots, \lambda_k)$ .*
- (ii) *Under  $H_1$  in (4),  $n^{-1}\Lambda_n \geq c\|\mathbf{A}\beta_o - g_0\|^2 + o_p(1)$  for a constant  $c > 0$ .*

### 5.3. Comparison between $W_n$ and $\Lambda_n$

In summary, the test  $W_n$  has some advantages over the test  $\Lambda_n$ . First, the asymptotic null distribution of  $W_n$  is distribution-free, whereas the asymptotic null distribution of  $\Lambda_n$  in general depends on unknown quantities. Second,  $W_n$  is invariant with re-scaling the generating  $q$ -function of the BD, but  $\Lambda_n$  is not. Third, the computational expense of  $W_n$  is much more reduced than that of  $\Lambda_n$ , partly because the integration operations for  $\rho_q$  are involved in  $\Lambda_n$ , but not in  $W_n$ , and partly because  $\Lambda_n$  requires both unrestricted and restricted parameter estimates, while  $W_n$  is useful in cases where restricted parameter estimates are difficult to compute. Thus,  $W_n$  will be focused on in numerical studies of Section 6.

## 6. Simulation Study

We conduct simulation evaluations of the performance of robust-BD estimation methods for general partially linear models. We use the Huber  $\psi$ -function  $\psi(\cdot)$  with  $c = 1.345$ . The weight functions are chosen to be  $w_1(x) = w_2(x) = 1/\{1 + \sum_{j=1}^d (\frac{x_j - m_j}{s_j})^2\}^{1/2}$ , where  $x = (x_1, \dots, x_d)^T$ ,  $m_j$  and  $s_j$  denote the sample median and sample median absolute deviation of  $\{X_{i,j} : i = 1, \dots, n\}$  respectively,  $j = 1, \dots, d$ . As a comparison, the classical non-robust estimation counterparts correspond to using  $\psi(r) = r$  and  $w_1(x) = w_2(x) \equiv 1$ . Throughout the numerical work, the Epanechnikov kernel function  $K(t) = 0.75 \max(1 - t^2, 0)$  is used. All these choices (among many others) are for feasibility; the issues on the trade-off between robustness and efficiency are not pursued further in the paper.

The following setup is used in the simulation studies. The sample size is  $n = 200$ , and the number of replications is 500. (Incorporating a nonparametric component in the GPLM desires a larger  $n$  when the number of covariates increases for better numerical performance.) Local-linear robust-BD estimation is illustrated with the bandwidth parameter  $h$  to be 20% of the interval length of the variable  $T$ . Results using other data-driven choices of  $h$  are similar and are omitted.

### 6.1. Bernoulli Responses

We generate observations  $\{(X_i, T_i, Y_i)\}_{i=1}^n$  randomly from the model,

$$Y | (X, T) \sim \text{Bernoulli}(m(X, T)), \quad X \sim N(\mathbf{0}, \Sigma), \quad T \sim \text{Uniform}(0, 1),$$

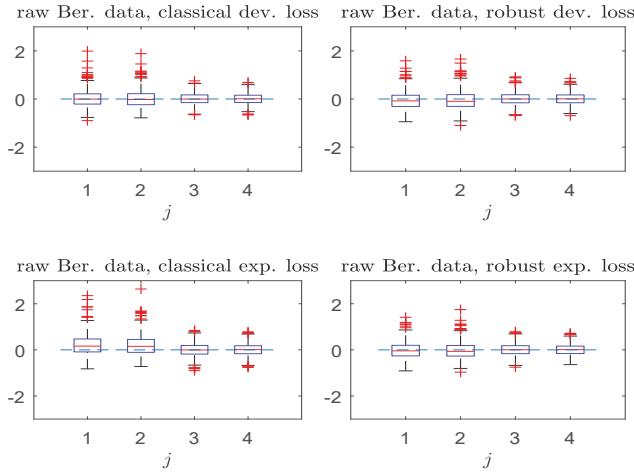
where  $\Sigma = (\sigma_{jk})$  with  $\sigma_{jk} = 0.2^{|j-k|}$ , and  $X$  is independent of  $T$ . The link function is  $\text{logit}\{m(x, t)\} = x^T \beta_o + \eta^o(t)$ , where  $\beta_o = (2, 2, 0, 0)^T$  and  $\eta^o(t) = 2 \sin\{\pi(1 + 2t)\}$ . Both the deviance and exponential loss functions are employed as the BD.

For each generated dataset from the true model, we create a contaminated dataset, where 10 data points  $(X_{i,j}, Y_i)$  are contaminated as follows: they are replaced by  $(X_{i,j}^*, Y_i^*)$ , where  $Y_i^* = 1 - Y_i$ ,  $i = 1, \dots, 5$ ,

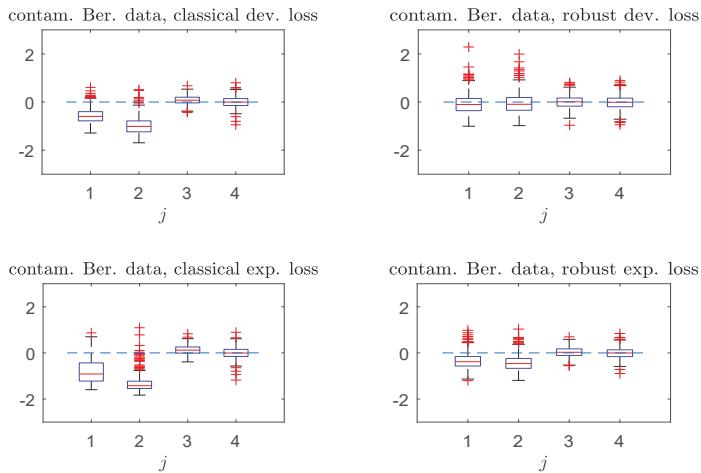
$$\begin{aligned} X_{1,2}^* &= 5 \text{sign}(U_1 - 0.5), & X_{2,2}^* &= 5 \text{sign}(U_2 - 0.5), & X_{3,2}^* &= 5 \text{sign}(U_3 - 0.5), \\ X_{4,4}^* &= 5 \text{sign}(U_4 - 0.5), & X_{5,1}^* &= 5 \text{sign}(U_5 - 0.5), & X_{6,2}^* &= 5 \text{sign}(U_6 - 0.5), \\ X_{7,3}^* &= 5 \text{sign}(U_7 - 0.5), & X_{8,4}^* &= 5 \text{sign}(U_8 - 0.5), & X_{9,2}^* &= 5 \text{sign}(U_9 - 0.5), \\ X_{10,3}^* &= 5 \text{sign}(U_{10} - 0.5), \end{aligned}$$

with  $\{U_i\} \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1)$ .

Figures 1 and 2 compare the boxplots of  $(\hat{\beta}_j - \beta_{j,0})$ ,  $j = 1, \dots, d$ , based on the non-robust and robust-BD estimates, where the deviance loss and exponential loss are used as the BD in the top and bottom panels respectively. As seen from Figure 1 in the absence of contamination, both non-robust and robust methods perform comparably well. Besides, the bias in non-robust methods using the exponential loss (with  $p_2(y; \theta)$  unbounded) is larger than that of the deviance loss (with  $p_2(y; \theta)$  bounded). In the presence of contamination, Figure 2 reveals that the robust method is more effective in decreasing the estimation bias without excessively increasing the estimation variance.

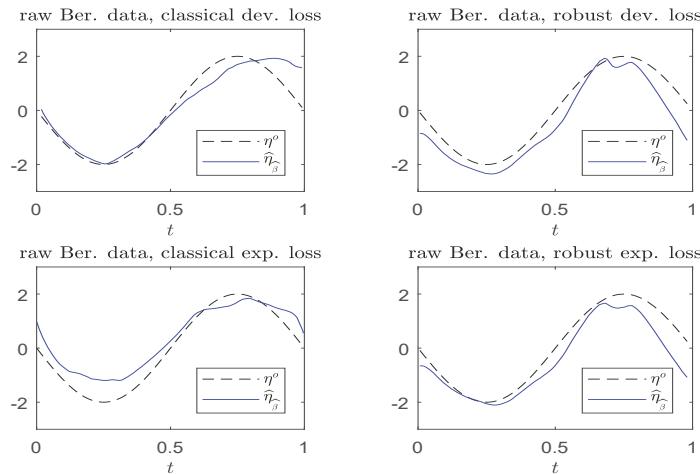


**Figure 1.** Simulated Bernoulli response data without contamination. Boxplots of  $(\hat{\beta}_j - \beta_{j,0})$ ,  $j = 1, \dots, d$  (from left to right). (Left panels): non-robust method; (right panels): robust method.

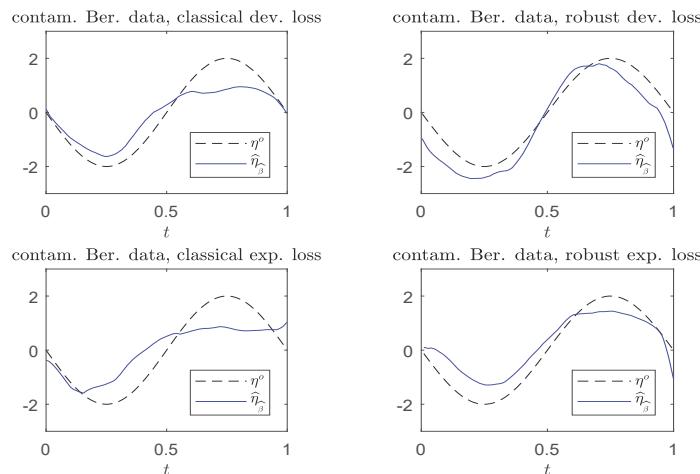


**Figure 2.** Simulated Bernoulli response data with contamination. The captions are identical to those in Figure 1.

For each replication, we calculate  $\text{MSE}(\hat{\eta}) = n^{-1} \sum_{i=1}^n \{\hat{\eta}_{\hat{\beta}}(t_i) - \eta^0(t_i)\}^2$ . Figures 3 and 4 compare the plots of  $\hat{\eta}_{\hat{\beta}}(t)$  from typical samples, using non-robust and robust-BD estimates, where the deviance loss and exponential loss are used as the BD in the top and bottom panels, respectively. There, the typical sample in each panel is selected in a way such that its MSE value corresponds to the 50-th percentile among the MSE-ranked values from 500 replications. These fitted curves reveal little difference between using the robust and non-robust methods, in the absence of contamination. For contaminated cases, robust estimates perform slightly better than non-robust estimates. Moreover, the boundary bias issue arising from the curve estimates at the edges using the local constant method can be ameliorated by using the local-linear method.



**Figure 3.** Simulated Bernoulli response data without contamination. Plots of  $\eta^o(t)$  and  $\hat{\eta}_\beta(t)$ . (Left panels): non-robust method; (right panels): robust method.



**Figure 4.** Simulated Bernoulli response data with contamination. Plots of  $\eta^o(t)$  and  $\hat{\eta}_\beta(t)$ . (Left panels): non-robust method; (right panels): robust method.

## 6.2. Gaussian Responses

We generate independent observations  $\{(X_i, T_i, Y_i)\}_{i=1}^n$  from  $(X, T, Y)$  satisfying:

$$Y \mid (X, T) \sim N(m(X, T), \sigma^2), \quad (X, \Phi^{-1}(T)) \sim N(\mathbf{0}, \Sigma),$$

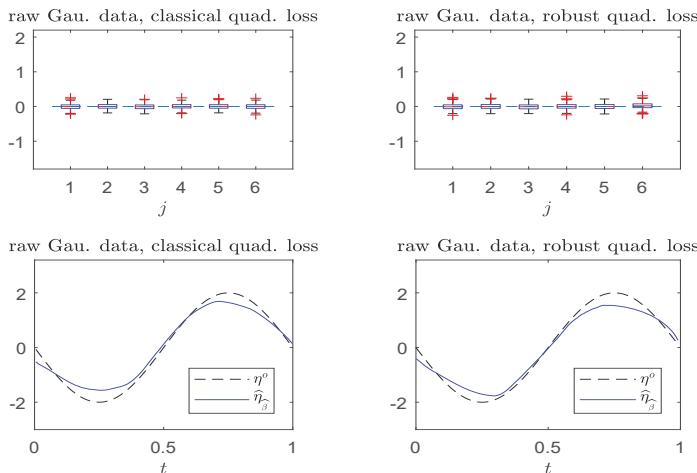
where  $\sigma = 1$ ,  $\Sigma = (\sigma_{jk})$  with  $\sigma_{jk} = 0.2^{|j-k|}$ ,  $\Phi$  denotes the CDF of the standard normal distribution. The link function is  $m(x, t) = x^T \beta_o + \eta^o(t)$ , where  $\beta_o = (2, -2, 1, -1, 0, 0)^T$  and  $\eta^o(t) = 2 \sin\{\pi(1 + 2t)\}$ . The quadratic loss is utilized as the BD.

For each dataset simulated from the true model, a contaminated data-set is created, where 10 data points  $(X_{i,j}, Y_i)$  are subject to contamination. They are replaced by  $(X_{i,j}^*, Y_i^*)$ , where  $Y_i^* = Y_i I\{|Y_i - m(X_i, T_i)|/\sigma > 2\} + 15 I\{|Y_i - m(X_i, T_i)|/\sigma \leq 2\}$ ,  $i = 1, \dots, 10$ ,

$$\begin{aligned} X_{1,2}^* &= 5 \operatorname{sign}(U_1 - 0.5), & X_{2,2}^* &= 5 \operatorname{sign}(U_2 - 0.5), & X_{3,2}^* &= 5 \operatorname{sign}(U_3 - 0.5), \\ X_{4,4}^* &= 5 \operatorname{sign}(U_4 - 0.5), & X_{5,6}^* &= 5 \operatorname{sign}(U_5 - 0.5), & X_{6,1}^* &= 5 \operatorname{sign}(U_6 - 0.5), \\ X_{7,2}^* &= 5 \operatorname{sign}(U_7 - 0.5), & X_{8,3}^* &= 5 \operatorname{sign}(U_8 - 0.5), & X_{9,4}^* &= 5 \operatorname{sign}(U_9 - 0.5), \\ X_{10,5}^* &= 5 \operatorname{sign}(U_{10} - 0.5), \end{aligned}$$

with  $\{U_i\} \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1)$ .

Figures 5 and 6 compare the boxplots of  $(\hat{\beta}_j - \beta_{j,o})$ ,  $j = 1, \dots, d$ , on the top panels, and plots of  $\hat{\eta}_{\hat{\beta}}(t)$  from typical samples, on the bottom panels, using the non-robust and robust-BD estimates. The typical samples are selected similar to those in Section 6.1. The simulation results in Figure 5 indicate that the robust method performs, as well as the non-robust method for estimating both the parameter vector and non-parametric curve in non-contaminated cases. Figure 6 reveals that the robust estimates are less sensitive to outliers than the non-robust counterparts. Indeed, the non-robust method yields a conceivable bias for parametric estimation, and non-parametric estimation is worse than that of the robust method.

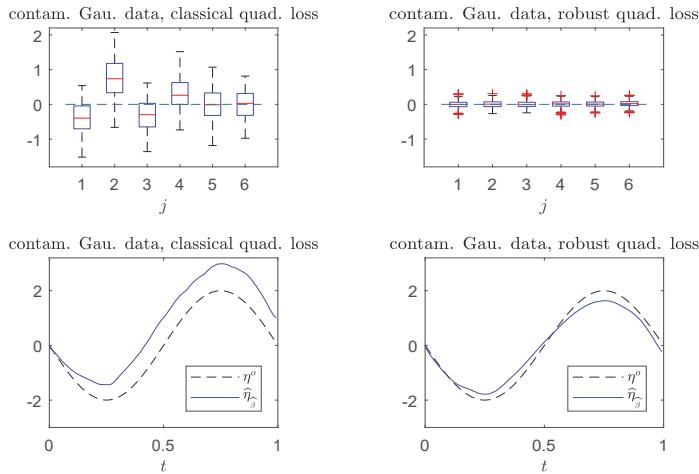


**Figure 5.** Simulated Gaussian response data without contamination. Top panels: boxplots of  $(\hat{\beta}_j - \beta_{j,o})$ ,  $j = 1, \dots, d$  (from left to right). Bottom panels: plots of  $\eta^o(t)$  and  $\hat{\eta}_{\beta}(t)$ . (Left panels): non-robust method; (right panels): robust method.

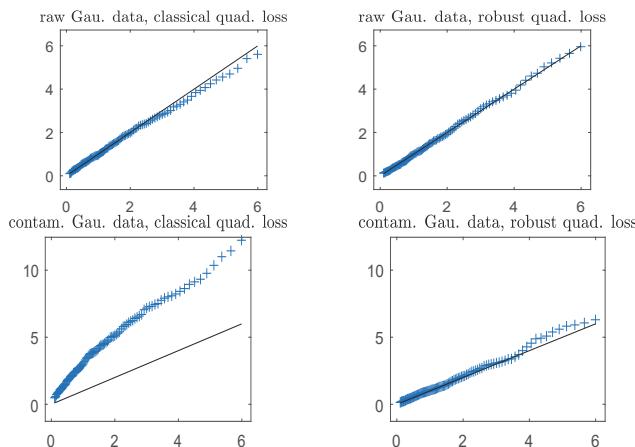
Figure 7 gives the QQ plots of the (first to 95-th) percentiles of the Wald-type statistic  $W_n$  versus those of the  $\chi_2^2$  distribution for testing the null hypothesis:

$$H_0 : \beta_{2,o} = -2 \text{ and } \beta_{4,o} = -1. \quad (32)$$

The plots depict that in both clean and contaminated cases, the robust  $W_n$  (in right panels) closely follows the  $\chi_2^2$  distribution, lending support to Theorem 3. On the other hand, the non-robust  $W_n$  agrees well with the  $\chi_2^2$  distribution in clean data; the presence of a small number of outlying data points severely distorts the sampling distribution of the non-robust  $W_n$  (in the bottom left panel) from the  $\chi_2^2$  distribution, yielding inaccurate levels of the test.

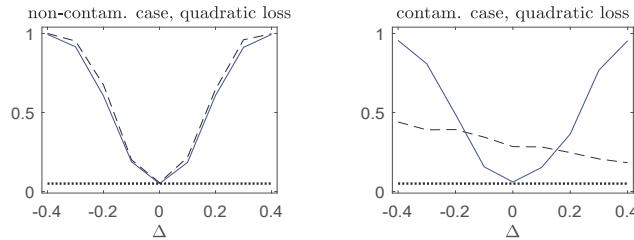


**Figure 6.** Simulated Gaussian response data with contamination. Top panels: boxplots of  $(\hat{\beta}_j - \beta_{j,0})$ ,  $j = 1, \dots, d$  (from left to right). Bottom panels: plots of  $\eta^o(t)$  and  $\hat{\eta}_\beta(t)$ . (**Left panels**): non-robust method; (**right panels**): robust method.



**Figure 7.** Simulated Gaussian response data with contamination. Empirical quantiles (on the  $y$ -axis) of the Wald-type statistics  $W_n$  versus quantiles (on the  $x$ -axis) of the  $\chi^2$  distribution. Solid line: the 45 degree reference line. (**Left panels**): non-robust method; (**right panels**): robust method.

To assess the stability of the power of the Wald-type test for testing the hypothesis (32), we evaluate the power in a sequence of alternatives with parameters  $\beta_0 + \Delta c$  for each given  $\Delta$ , where  $c = \beta_0 + (1, \dots, 1)^T$ . Figure 8 plots the empirical rejection rates of the null model in the non-contaminated case and the contaminated case. The price to pay for the robust  $W_n$  is a little loss of power in the non-contaminated cases. However, under contamination, a very different behavior is observed. The observed power curve of the robust  $W_n$  is close to those attained in the non-contaminated case. On the contrary, the non-robust  $W_n$  is less informative, since its power curve is much lower than that of the robust  $W_n$  against the alternative hypotheses with  $\Delta \neq 0$ , but higher than the nominal level at the null hypothesis with  $\Delta = 0$ .



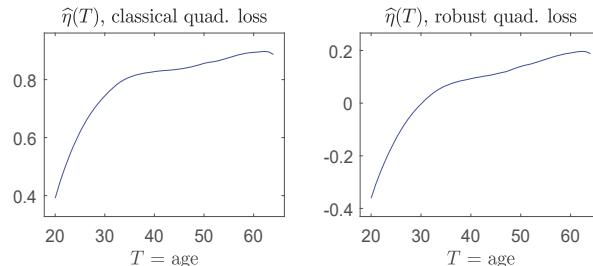
**Figure 8.** Observed power curves of tests for the Gaussian response data. The dashed line corresponds to the non-robust Wald-type test  $W_n$ ; the solid line corresponds to the robust  $W_n$ ; the dotted line indicates the 5% nominal level. (**Left panels**): non-contaminated case; (**right panels**): contaminated case.

## 7. Real Data Analysis

Two real datasets are analyzed. In both cases, the quadratic loss is set to be the BD, and the nonparametric function is fitted via local-linear regression method, where the bandwidth parameter is chosen to be 25% of the interval length of the variable  $T$ . Choices of the Huber  $\psi$ -function and weight functions are identical to those in Section 6.

### 7.1. Example 1

The dataset studied in [19] consists of 2447 observations on three variables,  $\log(\text{wage})$ , age and education, for women. It is of interest to learn how wages change with years of age and years of education. It is anticipated to find an increasing regression function of  $Y = \log(\text{wage})$  in  $T = \text{age}$  as well as in  $X_1 = \text{education}$ . We fit a partially linear model  $Y = \eta(T) + \beta_1 X_1 + \epsilon$ . Profiles of the fitted nonparametric functions  $\hat{\eta}(\cdot)$  in Figure 9 indeed exhibit the overall upward trend in age. The coefficient estimate is  $\hat{\beta}_1 = 0.0809$  with standard error 0.0042 using the non-robust method, and is  $\hat{\beta}_1 = 0.1334$  with standard error 0.0046 by means of the robust method. It is seen that robust estimates are similar to the non-robust counterparts. Our evaluation, based on both the non-robust and robust methods, supports the predicted result in theoretical and empirical literature in socio-economical studies.



**Figure 9.** The dataset in [19]. (**Left panels**): estimate of  $\eta(T)$  via the non-robust quadratic loss; (**right panels**): estimate of  $\eta(T)$  via the robust quadratic loss.

### 7.2. Example 2

We analyze an employee dataset (Example 11.3 of [20]) of the Fifth National Bank of Springfield, based on year 1995 data. The bank, whose name has been changed, was charged in court with that its female employees received substantially smaller salaries than its male employees. For each of its 208 employees, the dataset consists of seven variables, EduLev (education level), JobGrade (job grade), YrHired (year that an employee was hired), YrBorn (year that an employee was born), Female

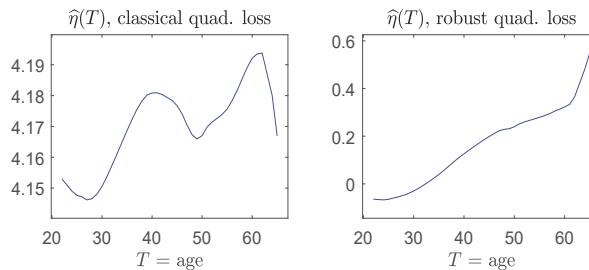
(indicator of being female), YrsPrior (years of work experience at another bank before working at the Fifth National bank), and Salary (current annual salary in thousands of dollars).

To explain variation in salary, we fit a partial linear model,  $Y = \eta(T) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$ , for  $Y = \log(\text{Salary})$ ,  $T = \text{Age}$ ,  $X_1 = \text{Female}$ ,  $X_2 = \text{YrHired}$ ,  $X_3 = \text{EducLev}$ ,  $X_4 = \text{JobGrade}$  and  $X_5 = \text{YrsPrior}$ , where  $\text{Age} = 95 - \text{YrBorn}$  is age. Table 1 presents parameter estimates and their standard errors (given within brackets), along with  $p$ -values calculated from the Wald-type test  $W_n$ . Figure 10 depicts the estimated nonparametric functions.

It is interesting to note that for this dataset, results from using the robust and non-robust methods make a difference in drawing conclusions. For example, from Table 1, the non-robust method gives the estimate of parameter  $\beta_1$  for gender to be below zero, which may be interpreted as the evidence of discrimination against female employees in salary and lends support to the plaintiff. In contrast, the robust method yields  $\hat{\beta}_1 > 0$ , which does not indicate that gender has an adverse effect. (A similar conclusion made from penalized-likelihood was obtained in Section 4.1 of [21]). Moreover, the estimated nonparametric functions  $\hat{\eta}(\cdot)$  obtained from non-robust and robust methods are qualitatively different: the former method does not deliver a monotone increasing pattern with Age, whereas the latter method does. Whether or not the difference was caused by outlying observations will be an interesting issue to be investigated.

**Table 1.** Parameter estimates and  $p$ -values for partially linear model of the dataset in [20]

Variable	Classical-BD Estimation		Robust-BD Estimation	
	Estimate (s.e.)	$p$ -Value of $W_n$	Estimate (s.e.)	$p$ -Value of $W_n$
Female	−0.0491 (0.0232)	0.0339	0.0530 (0.0323)	0.1010
YrHired	−0.0093 (0.0026)	0.0005	0.0359 (0.0086)	0.0000
EducLev	0.0179 (0.0079)	0.0228	−0.0133 (0.0131)	0.3103
JobGrade	0.0899 (0.0075)	0.0000	0.1672 (0.0168)	0.0000
YrsPrior	0.0033 (0.0023)	0.1528	−0.0050 (0.0061)	0.4104



**Figure 10.** The dataset in [20]. (Left panel): estimate of  $\eta(T)$  via the non-robust quadratic loss; (right panel): estimate of  $\eta(T)$  via the robust quadratic loss.

## 8. Discussion

Over the past two decades, nonparametric inference procedures for testing hypotheses concerning nonparametric regression functions have been developed extensively. See [22–26] and the references therein. The work on the generalized likelihood ratio test [24] offers light into nonparametric inference, based on function estimation under nonparametric models, using the quadratic loss function as the error measure. These works do not directly deal with the robust procedure. Exploring the inference on nonparametric functions, such as  $\eta^0(t)$  in GPLM associated with a scalar variable  $T$  and the additive structure  $\sum_{d=1}^D \eta_d^0(t_d)$  as in [27] with a vector variable  $T = (T_1, \dots, T_D)$ , estimated via the “robust-BD” as the error measure, when there are possible outlying data points, will be the future work.

This paper utilizes the class BD of loss functions, the optimal choice of which depends on specific settings and criteria. For e.g., regression and classification will utilize different loss functions, and thus further study on optimality is desirable.

Some recent work on partially linear models in econometrics includes [28–30]. There, the nonparametric function is approximated via linear expansions, with the number of coefficients diverging with  $n$ . Developing inference procedures to be resistant to outliers could be of interest.

**Acknowledgments:** The authors thank the two referees for insightful comments and suggestions. The research is supported by the U.S. NSF Grants DMS-1712418, DMS-1505367, CMMI-1536978, DMS-1308872, the Wisconsin Alumni Research Foundation and the National Natural Science Foundation of China, grants 11690014.

**Author Contributions:** C.Z. conceived and designed the experiments; C.Z. analyzed the data; Z.Z. contributed to discussions and analysis tools; C.Z. wrote the paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Proofs of Main Results

Throughout the proof,  $C$  represents a generic finite constant. We impose some regularity conditions, which may not be the weakest, but facilitate the technical derivations.

Notation:

For integers  $j \geq 0$ ,  $\mu_j(K) = \int u^j K(u) du$ ;  $c_p = (\mu_{p+1}(K), \dots, \mu_{2p+1}(K))^T$ ;  $\mathcal{S} = (\mu_{j+k-2}(K))_{1 \leq j, k \leq p+1}$ . Define:  $\eta(\mathbf{x}, t) = F(m(\mathbf{x}, t)) = \mathbf{x}^T \boldsymbol{\beta}_o + \eta^o(t)$ ;  $\eta_i = \eta(\mathbf{X}_i, T_i)$ . Set  $\eta_i(t; \boldsymbol{\beta}) = \mathbf{X}_i^T \boldsymbol{\beta} + \eta_\beta(t) + \sum_{k=1}^p (T_i - t)^k \eta_\beta^{(k)}(t) / k!$ ;  $g_1(\tau; t, \boldsymbol{\beta}) = E\{p_1(Y_i; \eta_i(t; \boldsymbol{\beta})) w_1(\mathbf{X}_i) \mid T_i = \tau\}$ ;  $g_2(\tau; t, \boldsymbol{\beta}) = E\{p_2(Y_i; \eta_i(t; \boldsymbol{\beta})) w_1(\mathbf{X}_i) \mid T_i = \tau\}$ .

Condition A:

- A1.  $\eta_\beta(t)$  is the unique minimizer of  $S(a; t, \boldsymbol{\beta})$  with respect to  $a \in \mathbb{R}^1$ .
- A2.  $\boldsymbol{\beta}_o \in \mathbb{R}^d$  is the unique minimizer of  $J(\boldsymbol{\beta}, \eta_\beta)$  with respect to  $\boldsymbol{\beta}$ , where  $d \geq 1$ .
- A3.  $\eta'(\cdot) = \eta_{\boldsymbol{\beta}_o}'(\cdot)$ .

Condition B:

- B1. The function  $\rho_q(y, \mu)$  is continuous and bounded. The functions  $p_1(y; \theta)$ ,  $p_2(y; \theta)$ ,  $p_3(y; \theta)$ ,  $w_1(\cdot)$  and  $w_2(\cdot)$  are bounded;  $p_2(y; \theta)$  is continuous in  $\theta$ .
- B2. The kernel function  $K$  is Lipschitz continuous, a symmetric probability density function with bounded support. The matrix  $\mathcal{S}$  is positive definite.
- B3. The marginal density  $f_T(t)$  of  $T$  is a continuous function, uniformly bounded away from zero and  $\infty$  for  $t \in \mathcal{T}_0$ .
- B4. The function  $S(a; t, \boldsymbol{\beta})$  is continuous and  $\eta_\beta(t)$  is a continuous function of  $(t, \boldsymbol{\beta})$ .
- B5. Assume  $g_2(\tau; t, \boldsymbol{\beta})$  is continuous in  $\tau$ ;  $g_2(t; t, \boldsymbol{\beta})$  is continuous in  $t \in \mathcal{T}_0$ .
- B6. Functions  $\eta_\beta(t)$  and  $\eta^o(t)$  are  $(p+1)$ -times continuously differentiable at  $t$ .
- B7. The link function  $F(\cdot)$  is monotone increasing and a bijection,  $F^{(3)}(\cdot)$  is continuous, and  $F^{(1)}(\cdot) > 0$ . The matrix  $\text{var}(\mathbf{X} \mid T = t)$  is positive definite for a.e.  $t$ .
- B8. The matrix  $\mathbf{H}_0$  in (25) is invertible;  $\Omega_0^*$  in (26) is positive-definite.
- B9.  $\hat{\eta}_\beta(t)$  and  $\eta_\beta(t)$  are continuously differentiable with respect to  $(t, \boldsymbol{\beta})$ , and twice continuously differentiable with respect to  $\boldsymbol{\beta}$  such that for any  $1 \leq j, k \leq d$ ,  $\frac{\partial^2}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k} \eta_\beta(t) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o}$  is bounded. Furthermore, for any  $1 \leq j, k \leq d$ ,  $\frac{\partial^2}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k} \eta_\beta(t)$  satisfies the equicontinuity condition:

$$\forall \varepsilon > 0, \exists \delta_\varepsilon > 0 : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_o\| < \delta_\varepsilon \implies \left\| \frac{\partial^2}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k} \eta_\beta \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_1} - \frac{\partial^2}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k} \eta_\beta \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} \right\|_\infty < \varepsilon.$$

Note that Conditions A, B2–B5 and B8–B9 were similarly used in [9]. Conditions B1 and B7 follow [10]. Condition B6 is due to the local  $p$ -th-degree polynomial regression estimation.

**Proof of Lemma 1:** From Condition A1, we obtain  $E\{p_1(Y; X^T \beta + \eta_\beta(t))w_1(X) \mid T = t\} = 0$  and  $E\{p_2(Y; X^T \beta + \eta_\beta(t))w_1(X) \mid T = t\} > 0$ , i.e.,

$$g_1(t; t, \beta) = E\{p_1(Y; X^T \beta + \eta_\beta(t))w_1(X) \mid T = t\} = 0, \quad (\text{A1})$$

$$g_2(t; t, \beta) = E\{p_2(Y; X^T \beta + \eta_\beta(t))w_1(X) \mid T = t\} > 0. \quad (\text{A2})$$

Define by  $\eta_\beta^{(0,\dots,p)}(t) = (\eta_\beta(t), \eta_\beta^{(1)}(t), \dots, \eta_\beta^{(p)}(t)/p!)^T$  the vector of  $\eta_\beta(t)$  along with re-scaled derivatives with respect to  $t$  up to the order  $p$ . Note that:

$$\begin{aligned} \eta_i(t; \beta) &= X_i^T \beta + \sum_{k=0}^p (T_i - t)^k \frac{\eta_\beta^{(k)}(t)}{k!} \\ &= X_i^T \beta + \mathbf{t}_i(t)^T \eta_\beta^{(0,\dots,p)}(t) \\ &= X_i^T \beta + \{H^{-1} \mathbf{t}_i(t)\}^T H \eta_\beta^{(0,\dots,p)}(t) \\ &= X_i^T \beta + \mathbf{t}_i^*(t)^T H \eta_\beta^{(0,\dots,p)}(t), \end{aligned}$$

where  $H = \text{diag}\{(1, h, \dots, h^p)\}$  and  $\mathbf{t}_i^*(t) = H^{-1} \mathbf{t}_i(t) = (1, (T_i - t)/h, \dots, (T_i - t)^p/h^p)^T$  denotes the re-scaled  $\mathbf{t}_i(t)$ . Then:

$$\begin{aligned} &X_i^T \beta + \mathbf{t}_i(t)^T \mathbf{a} \\ &= X_i^T \beta + \mathbf{t}_i^*(t)^T H \mathbf{a} \\ &= X_i^T \beta + \mathbf{t}_i^*(t)^T H \eta_\beta^{(0,\dots,p)}(t) + \mathbf{t}_i^*(t)^T H \{\mathbf{a} - \eta_\beta^{(0,\dots,p)}(t)\} \\ &= \eta_i(t; \beta) + \mathbf{t}_i^*(t)^T H \{\mathbf{a} - \eta_\beta^{(0,\dots,p)}(t)\}. \end{aligned}$$

Hence, we rewrite (16) as:

$$S_n(\mathbf{a}; t, \beta) = \frac{1}{n} \sum_{i=1}^n \rho_q(Y_i, F^{-1}(\eta_i(t; \beta) + \mathbf{t}_i^*(t)^T H \{\mathbf{a} - \eta_\beta^{(0,\dots,p)}(t)\})) w_1(X_i) K_h(T_i - t).$$

Therefore,  $\hat{\mathbf{a}}(t, \beta)$  minimizing  $S_n(\mathbf{a}; t, \beta)$  is equivalent to the one minimizing:

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \left\{ \rho_q(Y_i, F^{-1}(\eta_i(t; \beta) + \mathbf{t}_i^*(t)^T H \{\mathbf{a} - \eta_\beta^{(0,\dots,p)}(t)\})) \right. \\ &\quad \left. - \rho_q(Y_i, F^{-1}(\eta_i(t; \beta))) \right\} w_1(X_i) K_h(T_i - t) \end{aligned}$$

with respect to  $\mathbf{a}$ . It follows that  $\hat{\mathbf{a}}^*(t, \beta)$ , defined by  $\hat{\mathbf{a}}^*(t, \beta) = \sqrt{nh} H \{\hat{\mathbf{a}}(t, \beta) - \eta_\beta^{(0,\dots,p)}(t)\}$ , minimizes:

$$\begin{aligned} G_n(\mathbf{a}^*; t, \beta) &= nh \left[ \frac{1}{n} \sum_{i=1}^n \left\{ \rho_q(Y_i, F^{-1}(\eta_i(t; \beta) + \{a_n \mathbf{t}_i^*(t)^T \mathbf{a}^*\})) - \rho_q(Y_i, F^{-1}(\eta_i(t; \beta))) \right\} \right. \\ &\quad \left. w_1(X_i) K_h(T_i - t) \right] \end{aligned}$$

with respect to  $\mathbf{a}^* \in \mathbb{R}^{p+1}$ , where  $a_n = 1/\sqrt{nh}$ . Note that for any fixed  $\mathbf{a}^*$ ,  $|\mathbf{t}_i^*(t)^T \mathbf{a}^*| \leq C$ . By Taylor expansion,

$$\begin{aligned} G_n(\mathbf{a}^*; t, \beta) &= nh \left( a_n \left[ \frac{1}{n} \sum_{i=1}^n p_1(Y_i; \eta_i(t; \beta)) \{ \mathbf{t}_i^*(t)^T \mathbf{a}^* \} w_1(X_i) K_h(T_i - t) \right] \right. \\ &\quad \left. + a_n^2 \frac{1}{2} \left[ \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \eta_i(t; \beta)) \{ \mathbf{t}_i^*(t)^T \mathbf{a}^* \}^2 w_1(X_i) K_h(T_i - t) \right] \right) \end{aligned}$$

$$+ a_n^3 \frac{1}{6} \left[ \frac{1}{n} \sum_{i=1}^n p_3(Y_i; \eta_i^*(t; \beta)) \{ \mathbf{t}_i^*(t)^T \mathbf{a}^* \}^3 w_1(\mathbf{X}_i) K_h(T_i - t) \right] \\ = I_{n,1} + I_{n,2} + I_{n,3},$$

where  $\eta_i^*(t; \beta)$  is located between  $\eta_i(t; \beta)$  and  $\eta_i(t; \beta) + \{a_n \mathbf{t}_i^*(t)^T \mathbf{a}^*\}$ . We notice that:

$$I_{n,1} \equiv \sqrt{nh} \mathbf{W}_n(t; \beta)^T \mathbf{a}^*,$$

where:

$$\mathbf{W}_n(t; \beta) = \frac{1}{n} \sum_{i=1}^n p_1(Y_i; \eta_i(t; \beta)) \mathbf{t}_i^*(t) w_1(\mathbf{X}_i) K_h(T_i - t);$$

also, Lemma A1 implies:

$$I_{n,2} = nha_n^2 \frac{1}{2} \mathbf{a}^{*T} \left[ \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \eta_i(t; \beta)) \{ \mathbf{t}_i^*(t) \mathbf{t}_i^*(t)^T \} w_1(\mathbf{X}_i) K_h(T_i - t) \right] \mathbf{a}^* \\ = \frac{1}{2} \mathbf{a}^{*T} \mathbf{S}_2(t; \beta) \mathbf{a}^* + o_p(1),$$

where:

$$\mathbf{S}_2(t; \beta) = g_2(t; t; \beta) f_T(t) \mathcal{S} \succ 0$$

by (A2), Condition B2 and B5; and (by using  $X_n = O_p(E(|X_n|))$ ):

$$I_{n,3} \leq C O_p(nha_n^2) = O_p(1/\sqrt{nh}) = o_p(1).$$

Then:

$$G_n(\mathbf{a}^*; t; \beta) = \sqrt{nh} \mathbf{W}_n(t; \beta)^T \mathbf{a}^* + \frac{1}{2} \mathbf{a}^{*T} \mathbf{S}_2(t; \beta) \mathbf{a}^* + o_p(1),$$

where  $\mathbf{a}^{*T} \mathbf{S}_2(t; \beta) \mathbf{a}^* = (\mathbf{a}^{*T} \mathcal{S} \mathbf{a}^*) g_2(t; t; \beta) f_T(t)$  is continuous in  $t \in T_0$  by B3 and B5.

We now examine  $\mathbf{W}_n(t; \beta)$ . Note that:

$$\text{var}\{\mathbf{W}_n(t; \beta)\} = \frac{1}{n} \text{var}\{p_1(Y_i; \eta_i(t; \beta)) \mathbf{t}_i^*(t) w_1(\mathbf{X}_i) K_h(T_i - t)\} \\ \leq \frac{1}{n} E[p_1^2(Y_i; \eta_i(t; \beta)) \{ \mathbf{t}_i^*(t) \mathbf{t}_i^*(t)^T \} w_1^2(\mathbf{X}_i) \{ K_h(T_i - t) \}^2] \\ \leq \frac{C}{n} E\left[\frac{1}{h^2} \left\{ K\left(\frac{T_i - t}{h}\right) \right\}^2\right] \\ = \frac{C}{nh}.$$

To evaluate  $E\{\mathbf{W}_n(t; \beta)\}$ , it is easy to see that for each  $j \in \{0, 1, \dots, p\}$ ,

$$\mathbf{e}_{j+1,p+1}^T E\{\mathbf{W}_n(t; \beta)\} = E\{p_1(Y_i; \eta_i(t; \beta)) \mathbf{e}_{j+1,p+1}^T \mathbf{t}_i^*(t) w_1(\mathbf{X}_i) K_h(T_i - t)\} \\ = E\left\{p_1(Y_i; \eta_i(t; \beta)) \left(\frac{T_i - t}{h}\right)^j w_1(\mathbf{X}_i) K_h(T_i - t)\right\} \\ = E\left[E\{p_1(Y_i; \eta_i(t; \beta)) w_1(\mathbf{X}_i) | T_i\} \left(\frac{T_i - t}{h}\right)^j K_h(T_i - t)\right] \\ = E\left\{g_1(T_i; t; \beta) \left(\frac{T_i - t}{h}\right)^j K_h(T_i - t)\right\} \\ = \int g_1(y; t; \beta) \left(\frac{y - t}{h}\right)^j \frac{1}{h} K\left(\frac{y - t}{h}\right) f_T(y) dy \\ = \int g_1(t + hx; t; \beta) x^j K(x) f_T(t + hx) dx.$$

Note that by Taylor expansion,

$$\eta_{\beta}(t + hx) = \sum_{k=0}^p (hx)^k \frac{\eta_{\beta}^{(k)}(t)}{k!} + (hx)^{p+1} \frac{\eta_{\beta}^{(p+1)}(t)}{(p+1)!} + o(h^{p+1}).$$

This combined with the facts (A1) and (A2) give that:

$$\begin{aligned} & g_1(t + hx; t, \beta) \\ &= E \left\{ p_1 \left( Y; \mathbf{X}^T \beta + \sum_{k=0}^p (hx)^k \frac{\eta_{\beta}^{(k)}(t)}{k!} \right) w_1(\mathbf{X}) \mid T = t + hx \right\} \\ &= E \left[ p_1(Y; \mathbf{X}^T \beta + \eta_{\beta}(t + hx)) w_1(\mathbf{X}) \right. \\ &\quad \left. + p_2(Y; \mathbf{X}^T \beta + \eta_{\beta}(t + hx)) \left\{ \sum_{k=0}^p (hx)^k \frac{\eta_{\beta}^{(k)}(t)}{k!} - \eta_{\beta}(t + hx) \right\} w_1(\mathbf{X}) \mid T = t + hx \right] \\ &\quad + o(h^{p+1}) \\ &= g_1(t + hx; t + hx, \beta) - (hx)^{p+1} \frac{\eta_{\beta}^{(p+1)}(t)}{(p+1)!} g_2(t + hx; t + hx, \beta) + o(h^{p+1}) \\ &= -(hx)^{p+1} \frac{\eta_{\beta}^{(p+1)}(t)}{(p+1)!} g_2(t + hx; t + hx, \beta) + o(h^{p+1}). \end{aligned}$$

Thus, using the continuity of  $g_2(t; t, \beta)$  and  $f_T(t)$  in  $t$ , we obtain:

$$E\{W_n(t, \beta)\} = -c_p \frac{\eta_{\beta}^{(p+1)}(t)}{(p+1)!} g_2(t; t, \beta) f_T(t) h^{p+1} + o(h^{p+1})$$

uniformly in  $(t, \beta)$ . Thus, we conclude that  $\sqrt{nh} W_n(t, \beta) = O_p(1)$  when  $nh^{2p+3} = O(1)$ .

By Lemma A2,

$$\sup_{a^* \in \Theta, t \in \mathcal{T}_0, \beta \in \mathcal{K}} \left| G_n(a^*; t, \beta) - \sqrt{nh} W_n(t, \beta)^T a^* - \frac{1}{2} a^{*T} \mathbf{S}_2(t, \beta) a^* \right| = o_p(1).$$

This along with Lemma A.1 of [18] yields:

$$\sup_{t \in \mathcal{T}_0, \beta \in \mathcal{K}} \|\hat{a}^*(t, \beta) + \{\mathbf{S}_2(t, \beta)\}^{-1} \sqrt{nh} W_n(t, \beta)\| = o_p(1),$$

the first entry of which satisfies:

$$\sup_{t \in \mathcal{T}_0, \beta \in \mathcal{K}} |\sqrt{nh} \{\hat{\eta}_{\beta}(t) - \eta_{\beta}(t)\} + e_{1,p+1}^T \{\mathbf{S}_2(t, \beta)\}^{-1} \sqrt{nh} W_n(t, \beta)| = o_p(1),$$

namely,  $\sup_{t \in \mathcal{T}_0, \beta \in \mathcal{K}} |\hat{\eta}_{\beta}(t) - \eta_{\beta}(t) + e_{1,p+1}^T \{\mathbf{S}_2(t, \beta)\}^{-1} W_n(t, \beta)| = o_p(1/\sqrt{nh})$ . By [31],  $\sup_{t \in \mathcal{T}_0, \beta \in \mathcal{K}} \|W_n(t, \beta) - E\{W_n(t, \beta)\}\| = O_p(\{\frac{\log(1/h)}{nh}\}^{1/2})$ . Furthermore,

$$\{\mathbf{S}_2(t, \beta)\}^{-1} E\{W_n(t, \beta)\} = -\mathcal{S}^{-1} c_p \frac{\eta_{\beta}^{(p+1)}(t)}{(p+1)!} h^{p+1} + o(h^{p+1})$$

uniformly in  $(t, \beta)$ . Therefore,

$$\sup_{t \in \mathcal{T}_0, \beta \in \mathcal{K}} \left| \hat{\eta}_{\beta}(t) - \eta_{\beta}(t) - e_{1,p+1}^T \mathcal{S}^{-1} c_p \frac{\eta_{\beta}^{(p+1)}(t)}{(p+1)!} h^{p+1} \right| = o_p(1).$$

This yields:

$$\begin{aligned} & \sup_{\beta \in \mathcal{K}} \sup_{t \in \mathcal{T}_0} \left| \hat{\eta}_{\beta}(t) - \eta_{\beta}(t) - \mathbf{e}_{1,p+1}^T \mathcal{S}^{-1} \mathbf{c}_p \frac{\eta_{\beta}^{(p+1)}(t)}{(p+1)!} h^{p+1} \right| \\ & \leq \sup_{t \in \mathcal{T}_0, \beta \in \mathcal{K}} \left| \hat{\eta}_{\beta}(t) - \eta_{\beta}(t) - \mathbf{e}_{1,p+1}^T \mathcal{S}^{-1} \mathbf{c}_p \frac{\eta_{\beta}^{(p+1)}(t)}{(p+1)!} h^{p+1} \right| = o_p(1). \end{aligned}$$

Note that for  $p = 1$ ,  $\mathbf{e}_{1,p+1}^T \mathcal{S}^{-1} \mathbf{c}_p = \mu_2(K)$ . This completes the proof.  $\square$

**Lemma A1.** Assume Condition B in the Appendix. If  $n \rightarrow \infty$ ,  $h \rightarrow 0$  and  $nh \rightarrow \infty$ , then for given  $t \in \mathcal{T}_0$  and  $\beta \in \mathcal{K}$ ,

$$\frac{1}{n} \sum_{i=1}^n p_2(Y_i; \eta_i(t; \beta)) \{ \mathbf{t}_i^*(t) \mathbf{t}_i^*(t)^T \} w_1(\mathbf{X}_i) K_h(T_i - t) = \mathbf{S}_2(t, \beta) + o_p(1),$$

where  $\mathbf{S}_2(t, \beta) = g_2(t; t, \beta) f_T(t) \mathcal{S}$ , with  $\mathcal{S} = (\mu_{j+k-2}(K))_{1 \leq j, k \leq p+1}$  and  $\mu_j(K) = \int u^j K(u) du$ ,  $j = 0, 1, \dots, 2p$ .

**Proof.** Recall the  $(p+1) \times (p+1)$  matrix  $\mathbf{t}_i^*(t) \mathbf{t}_i^*(t)^T = ((\frac{T_i-t}{h})^{j+k-2})_{1 \leq j, k \leq p+1}$ . Set  $X_j = \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \eta_i(t; \beta)) (\frac{T_i-t}{h})^j w_1(\mathbf{X}_i) K_h(T_i - t)$  for  $j = 0, 1, \dots, 2p$ . We observe that:

$$\begin{aligned} E(X_j) &= \frac{1}{n} \sum_{i=1}^n E \left[ E \{ p_2(Y_i; \eta_i(t; \beta)) w_1(\mathbf{X}_i) | T_i \} \left( \frac{T_i-t}{h} \right)^j K_h(T_i - t) \right] \\ &= \frac{1}{n} \sum_{i=1}^n E \left\{ g_2(T_i; t, \beta) \left( \frac{T_i-t}{h} \right)^j K_h(T_i - t) \right\} \\ &= E \left\{ g_2(T; t, \beta) \left( \frac{T-t}{h} \right)^j K_h(T-t) \right\} \\ &= \int g_2(y; t, \beta) \left( \frac{y-t}{h} \right)^j \frac{1}{h} K \left( \frac{y-t}{h} \right) f_T(y) dy \\ &= \int g_2(t+hx; t, \beta) x^j K(x) f_T(t+hx) dx \\ &= g_2(t; t, \beta) f_T(t) \mu_j(K) + o(1), \end{aligned}$$

using the continuity of  $g_2(\tau; t, \beta)$  in  $\tau$  and  $f_T(t)$  in  $t$ . Similarly,

$$\begin{aligned} \text{var}(X_j) &= \frac{1}{n^2} \sum_{i=1}^n \text{var} \left\{ p_2(Y_i; \eta_i(t; \beta)) \left( \frac{T_i-t}{h} \right)^j w_1(\mathbf{X}_i) K_h(T_i - t) \right\} \\ &\leq \frac{1}{n^2} \sum_{i=1}^n E \left[ p_2^2(Y_i; \eta_i(t; \beta)) \left( \frac{T_i-t}{h} \right)^{2j} w_1^2(\mathbf{X}_i) \{ K_h(T_i - t) \}^2 \right] \\ &\leq \frac{C}{nh}. \end{aligned}$$

This completes the proof.  $\square$

**Lemma A2.** Assume Condition B. If  $n \rightarrow \infty$ ,  $h \rightarrow 0$ ,  $nh \rightarrow \infty$ ,  $\log(1/h)/(nh) \rightarrow 0$ , then  $\sup_{\alpha^* \in \Theta, t \in \mathcal{T}_0, \beta \in \mathcal{K}} |G_n(\alpha^*; t, \beta) - \sqrt{nh} W_n(t, \beta)^T \alpha^* - 2^{-1} \alpha^{*T} \mathbf{S}_2(t, \beta) \alpha^*| = o_p(1)$ , with a compact set  $\Theta \subseteq \mathbb{R}^{p+1}$ .

**Proof.** Let  $D_n(\alpha^*; t, \beta) = G_n(\alpha^*; t, \beta) - \sqrt{nh} W_n(t, \beta)^T \alpha^*$ . Note that:

$$\begin{aligned} & D_n(\alpha^*; t, \beta) \\ &= nh \left[ \frac{1}{n} \sum_{i=1}^n \rho_q(Y_i, F^{-1}(\eta_i(t; \beta) + \{ a_n \mathbf{t}_i^*(t)^T \alpha^* \})) w_1(\mathbf{X}_i) K_h(T_i - t) \right. \\ & \quad \left. - \frac{1}{n} \sum_{i=1}^n \rho_q(Y_i, F^{-1}(\eta_i(t; \beta))) w_1(\mathbf{X}_i) K_h(T_i - t) \right] \end{aligned}$$

$$\begin{aligned}
& -\frac{1}{n} \sum_{i=1}^n p_1(Y_i; \eta_i(t; \beta)) \{a_n \mathbf{t}_i^*(t)^T \mathbf{a}^*\} w_1(\mathbf{X}_i) K_h(T_i - t) \\
& = \frac{1}{2} \mathbf{a}^{*T} \left[ \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \tilde{\eta}_i(t; \beta)) \{\mathbf{t}_i^*(t) \mathbf{t}_i^*(t)^T\} w_1(\mathbf{X}_i) K_h(T_i - t) \right] \mathbf{a}^*,
\end{aligned}$$

where  $a_n = 1/\sqrt{n h}$  and  $\tilde{\eta}_i(t; \beta)$  is between  $\eta_i(t; \beta)$  and  $\eta_i(t; \beta) + \{a_n \mathbf{t}_i^*(t)^T \mathbf{a}^*\}$ . Then:

$$\begin{aligned}
& |D_n(\mathbf{a}^*; t, \beta) - 2^{-1} \mathbf{a}^{*T} \mathbf{S}_2(t, \beta) \mathbf{a}^*| \\
& = \frac{1}{2} \left| \mathbf{a}^{*T} \left[ \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \tilde{\eta}_i(t; \beta)) \{\mathbf{t}_i^*(t) \mathbf{t}_i^*(t)^T\} w_1(\mathbf{X}_i) K_h(T_i - t) - \mathbf{S}_2(t, \beta) \right] \mathbf{a}^* \right| \\
& \leq \|\mathbf{a}^*\|^2 \left| \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \tilde{\eta}_i(t; \beta)) \{\mathbf{t}_i^*(t) \mathbf{t}_i^*(t)^T\} w_1(\mathbf{X}_i) K_h(T_i - t) - \mathbf{S}_2(t, \beta) \right|.
\end{aligned}$$

The proof completes by applying [31].  $\square$

**Proof of Theorem 1.** Before showing Theorem 1, we need Proposition A1 (whose proof is omitted), where the following notation will be used. Denote by  $\mathcal{C}^1(\mathcal{T})$  the set of continuously differentiable functions in  $\mathcal{T}$ . Let  $\mathcal{V}(\beta)$  denote the neighborhood of  $\beta \in \mathcal{K}$ . Let  $\mathcal{H}_\delta(\beta)$  denote the neighborhood of  $\eta_\beta$  such that  $\mathcal{V}(\beta) \subseteq \mathcal{K}$  and  $\mathcal{H}_\delta(\beta) = \{u \in \mathcal{C}^1(\mathcal{T}) : \|u - \eta_\beta\|_\infty \leq \delta, \|\frac{\partial}{\partial t} u - \frac{\partial}{\partial t} \eta_\beta\|_\infty \leq \delta\}$ .

**Proposition A1.** Let  $\{(Y_i, \mathbf{X}_i, T_i)\}_{i=1}^n$  be independent observations of  $(Y, \mathbf{X}, T)$  modeled by (2) and (5). Assume that a random variable  $T$  is distributed on  $\mathcal{T}$ . Let  $\mathcal{K}$  and  $\mathcal{H}_1(\beta)$  be compact sets,  $g(\cdot, \cdot) : \mathbb{R}^2 \rightarrow \mathbb{R}$  be a continuous and bounded function,  $W(\mathbf{x}, t) : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$  be such that  $E\{|W(\mathbf{X}, T)|\} < \infty$  and  $\eta_\beta(t) = \eta(t, \beta) : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$  be a continuous function of  $(t, \beta)$ . Then:

- (i)  $E\{g(Y; \mathbf{X}^T \theta + v(T)) W(\mathbf{X}, T)\} \rightarrow E\{g(Y; \mathbf{X}^T \beta + \eta_\beta(T)) W(\mathbf{X}, T)\}$  as  $\|\theta - \beta\| + \|v - \eta_\beta\|_\infty \rightarrow 0$ ;
- (ii)  $\sup_{\theta \in \mathcal{K}} |n^{-1} \sum_{i=1}^n g(Y_i; \mathbf{X}_i^T \theta + \eta_\theta(T_i)) W(\mathbf{X}_i, T_i) - E\{g(Y; \mathbf{X}^T \theta + \eta_\theta(T)) W(\mathbf{X}, T)\}| \xrightarrow{P} 0$  as  $n \rightarrow \infty$ ;
- (iii) if, in addition,  $\mathcal{T}$  is compact and  $\eta_\beta \in \mathcal{C}^1(\mathcal{T})$ , then  $\sup_{\theta \in \mathcal{K}, v \in \mathcal{H}_1(\beta)} |n^{-1} \sum_{i=1}^n g(Y_i; \mathbf{X}_i^T \theta + v(T_i)) W(\mathbf{X}_i, T_i) - E\{g(Y; \mathbf{X}^T \theta + v(T)) W(\mathbf{X}, T)\}| \xrightarrow{P} 0$  as  $n \rightarrow \infty$ .

For part (i), we first show that for any compact set  $\mathcal{K}$  in  $\mathbb{R}^d$ ,

$$\sup_{\beta \in \mathcal{K}} |J_n(\beta, \hat{\eta}_\beta) - J(\beta, \eta_\beta)| \xrightarrow{P} 0. \quad (\text{A3})$$

It suffices to show  $\sup_{\beta \in \mathcal{K}} |J_n(\beta, \eta_\beta) - J(\beta, \eta_\beta)| \xrightarrow{P} 0$ , which follows from Proposition A1 (ii), and:

$$\sup_{\beta \in \mathcal{K}} |J_n(\beta, \hat{\eta}_\beta) - J_n(\beta, \eta_\beta)| \xrightarrow{P} 0. \quad (\text{A4})$$

To show (A4), we note that for any  $\varepsilon > 0$ , let  $\mathcal{T}_0$  be a compact set such that  $P(T_i \notin \mathcal{T}_0) < \varepsilon$ . Then:

$$\begin{aligned}
& J_n(\beta, \hat{\eta}_\beta) - J_n(\beta, \eta_\beta) \\
& = \frac{1}{n} \sum_{i=1}^n \{ \rho_q(Y_i, F^{-1}(\mathbf{X}_i^T \beta + \hat{\eta}_\beta(T_i))) - \rho_q(Y_i, F^{-1}(\mathbf{X}_i^T \beta + \eta_\beta(T_i))) \} w_2(\mathbf{X}_i) I(T_i \in \mathcal{T}_0) \\
& \quad + \frac{1}{n} \sum_{i=1}^n \{ \rho_q(Y_i, F^{-1}(\mathbf{X}_i^T \beta + \hat{\eta}_\beta(T_i))) - \rho_q(Y_i, F^{-1}(\mathbf{X}_i^T \beta + \eta_\beta(T_i))) \} w_2(\mathbf{X}_i) I(T_i \notin \mathcal{T}_0).
\end{aligned}$$

For  $T_i \in \mathcal{T}_0$ , by the mean-value theorem,

$$\begin{aligned}
& |\rho_q(Y_i, F^{-1}(\mathbf{X}_i^T \beta + \hat{\eta}_\beta(T_i))) - \rho_q(Y_i, F^{-1}(\mathbf{X}_i^T \beta + \eta_\beta(T_i)))| \\
& = |\mathbf{p}_1(Y_i; \mathbf{X}_i^T \beta + \eta_{i,\beta}^*) \{\hat{\eta}_\beta(T_i) - \eta_\beta(T_i)\}|
\end{aligned}$$

$$\leq \|p_1(\cdot; \cdot)\|_\infty \sup_{\beta \in \mathcal{K}} \|\hat{\eta}_\beta - \eta_\beta\|_{\mathcal{T}_0; \infty},$$

where  $\eta_{i,\beta}^*$  is located between  $\hat{\eta}_\beta(T_i)$  and  $\eta_\beta(T_i)$ . For  $T_i \notin \mathcal{T}_0$ , it follows that:

$$\begin{aligned} & |\rho_q(Y_i, F^{-1}(X_i^T \beta + \hat{\eta}_\beta(T_i))) - \rho_q(Y_i, F^{-1}(X_i^T \beta + \eta_\beta(T_i)))| \\ & \leq 2\|\rho_q(\cdot, \cdot)\|_\infty. \end{aligned}$$

Hence,

$$\begin{aligned} |J_n(\beta, \hat{\eta}_\beta) - J_n(\beta, \eta_\beta)| & \leq \left\{ \|p_1(\cdot; \cdot)\|_\infty \sup_{\beta \in \mathcal{K}} \|\hat{\eta}_\beta - \eta_\beta\|_{\mathcal{T}_0; \infty} + 2\|\rho_q(\cdot, \cdot)\|_\infty T_n^*\right\} \|w_2\|_\infty \\ & \leq 2\varepsilon, \end{aligned}$$

where the last inequality is entailed by Lemma 1 and the law of large numbers for  $T_n^* = n^{-1} \sum_{i=1}^n I(T_i \notin \mathcal{T}_0)$ . This completes the proof of (A3). The proof of  $\hat{\beta} \xrightarrow{P} \beta_o$  follows from combining Lemma A-1 of [1] with (A3) and Condition A2.

Part (ii) follows from Lemma 1, Part (i) and Condition B5 for  $\eta_\beta(t)$ .  $\square$

**Proof of Theorem 2.** Similar to the proof of Lemma 1, it can be shown that  $|\hat{\eta}_\beta(t) - \eta_\beta(t) + e_{1,p+1}^T \{S_2(t, \beta)\}^{-1} \frac{1}{n} \sum_{i=1}^n p_1(Y_i; \eta_i(t; \beta)) t_i^*(t) w_1(X_i) K_h(T_i - t)| = O_p(h^2 a_n + a_n^2 \sqrt{\log(1/h)})$ . Note that for  $p = 1$ ,

$$\begin{aligned} e_{1,p+1}^T \{S_2(t, \beta)\}^{-1} t_i^*(t) & = \frac{1}{g_2(t; t, \beta) f_T(t)} (1, 0) \begin{pmatrix} 1 & 0 \\ 0 & 1/\mu_2(K) \end{pmatrix} \begin{pmatrix} 1 \\ (T_i - t)/h \end{pmatrix} \\ & = \frac{1}{g_2(t; t, \beta) f_T(t)}. \end{aligned}$$

Thus:

$$\left| \hat{\eta}_\beta(t) - \eta_\beta(t) + \frac{1}{n f_T(t) g_2(t; t, \beta)} \sum_{i=1}^n p_1(Y_i; \eta_i(t; \beta)) w_1(X_i) K_h(T_i - t) \right| = O_p(h^2 a_n + a_n^2 \sqrt{\log(1/h)}).$$

Consider  $\hat{\beta}$  defined in (23). Note that:

$$\begin{aligned} X_i^T \beta + \hat{\eta}_\beta(T_i) & = X_i^T \beta_o + X_i^T (\beta - \beta_o) + \hat{\eta}_{(\beta - \beta_o) + \beta_o}(T_i) \\ & = X_i^T \beta_o + c_n X_i^T \{\sqrt{n}(\beta - \beta_o)\} + \hat{\eta}_{c_n \{\sqrt{n}(\beta - \beta_o)\} + \beta_o}(T_i), \end{aligned}$$

where  $c_n = 1/\sqrt{n}$ . Then,  $\hat{\theta} = \sqrt{n}(\hat{\beta} - \beta_o)$  minimizes:

$$\begin{aligned} J_n(\theta) & = n \left[ \frac{1}{n} \sum_{i=1}^n \left\{ \rho_q(Y_i, F^{-1}(X_i^T \beta_o + c_n X_i^T \theta + \hat{\eta}_{c_n \theta + \beta_o}(T_i))) w_2(X_i) \right. \right. \\ & \quad \left. \left. - \rho_q(Y_i, F^{-1}(X_i^T \beta_o + \hat{\eta}_{\beta_o}(T_i))) w_2(X_i) \right\} \right] \end{aligned}$$

with respect to  $\theta$ . By Taylor expansion,

$$\begin{aligned} & J_n(\theta) \\ & = n \left( \frac{1}{n} \sum_{i=1}^n p_1(Y_i; X_i^T \beta_o + \hat{\eta}_{\beta_o}(T_i)) [c_n X_i^T \theta + \{\hat{\eta}_{c_n \theta + \beta_o}(T_i) - \hat{\eta}_{\beta_o}(T_i)\}] w_2(X_i) \right. \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{2n} \sum_{i=1}^n p_2(Y_i; \mathbf{X}_i^T \boldsymbol{\beta}_o + \hat{\eta}_{\boldsymbol{\beta}_o}(T_i)) [c_n \mathbf{X}_i^T \boldsymbol{\theta} + \{\hat{\eta}_{c_n \boldsymbol{\theta} + \boldsymbol{\beta}_o}(T_i) - \hat{\eta}_{\boldsymbol{\beta}_o}(T_i)\}]^2 w_2(\mathbf{X}_i) \\
& + \frac{1}{6n} \sum_{i=1}^n p_3(Y_i; \eta_i^*) [c_n \mathbf{X}_i^T \boldsymbol{\theta} + \{\hat{\eta}_{c_n \boldsymbol{\theta} + \boldsymbol{\beta}_o}(T_i) - \hat{\eta}_{\boldsymbol{\beta}_o}(T_i)\}]^3 w_2(\mathbf{X}_i) \\
& = I_{n,1} + I_{n,2} + I_{n,3},
\end{aligned}$$

where  $\eta_i^*$  is located between  $\mathbf{X}_i^T \boldsymbol{\beta}_o + \hat{\eta}_{\boldsymbol{\beta}_o}(T_i)$  and  $\mathbf{X}_i^T \boldsymbol{\beta}_o + c_n \mathbf{X}_i^T \boldsymbol{\theta} + \hat{\eta}_{c_n \boldsymbol{\theta} + \boldsymbol{\beta}_o}(T_i)$ ,

$$\begin{aligned}
I_{n,1} &= \sum_{i=1}^n p_1(Y_i; \mathbf{X}_i^T \boldsymbol{\beta}_o + \hat{\eta}_{\boldsymbol{\beta}_o}(T_i)) \left\{ c_n \mathbf{X}_i^T \boldsymbol{\theta} + \frac{\partial \hat{\eta}_{\boldsymbol{\beta}}(T_i)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n} c_n \boldsymbol{\theta} \right\} w_2(\mathbf{X}_i) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n p_1(Y_i; \mathbf{X}_i^T \boldsymbol{\beta}_o + \hat{\eta}_{\boldsymbol{\beta}_o}(T_i)) \left\{ \mathbf{X}_i + \frac{\partial \hat{\eta}_{\boldsymbol{\beta}}(T_i)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n} \right\}^T \boldsymbol{\theta} w_2(\mathbf{X}_i) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n p_1(Y_i; \mathbf{X}_i^T \boldsymbol{\beta}_o + \hat{\eta}_{\boldsymbol{\beta}_o}(T_i)) \left\{ \mathbf{X}_i + \frac{\partial \hat{\eta}_{\boldsymbol{\beta}}(T_i)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} \right\}^T \boldsymbol{\theta} w_2(\mathbf{X}_i) + o_p(1), \\
I_{n,2} &= \frac{1}{2} \boldsymbol{\theta}^T \left[ \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \mathbf{X}_i^T \boldsymbol{\beta}_o + \hat{\eta}_{\boldsymbol{\beta}_o}(T_i)) \right. \\
&\quad \left. \left\{ \mathbf{X}_i + \frac{\partial \hat{\eta}_{\boldsymbol{\beta}}(T_i)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n} \right\} \left\{ \mathbf{X}_i + \frac{\partial \hat{\eta}_{\boldsymbol{\beta}}(T_i)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n} \right\}^T w_2(\mathbf{X}_i) \right] \boldsymbol{\theta} \\
&= \frac{1}{2} \boldsymbol{\theta}^T \mathbf{B}_2 \boldsymbol{\theta} + o_p(1), \\
I_{n,3} &= o_p(1),
\end{aligned}$$

with  $\boldsymbol{\beta}_n$  located between  $\boldsymbol{\beta}_o$  and  $c_n \boldsymbol{\theta} + \boldsymbol{\beta}_o$ , and  $\mathbf{B}_2 = \mathbf{H}_0$  following Lemma 1, Condition A3 and Proposition A1. Thus:

$$J_n(\boldsymbol{\theta}) = I_{n,1}^* \boldsymbol{\theta} + \frac{1}{2} \boldsymbol{\theta}^T \mathbf{B}_2 \boldsymbol{\theta} + o_p(1), \quad (\text{A5})$$

where  $I_{n,1}^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n p_1(Y_i; \mathbf{X}_i^T \boldsymbol{\beta}_o + \hat{\eta}_{\boldsymbol{\beta}_o}(T_i)) \left\{ \mathbf{X}_i + \frac{\partial \hat{\eta}_{\boldsymbol{\beta}}(T_i)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} \right\} w_2(\mathbf{X}_i)$ . Note that:

$$\begin{aligned}
I_{n,1}^* &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ p_1(Y_i; \mathbf{X}_i^T \boldsymbol{\beta}_o + \eta_{\boldsymbol{\beta}_o}(T_i)) \left\{ \mathbf{X}_i + \frac{\partial \eta_{\boldsymbol{\beta}}(T_i)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} \right\} w_2(\mathbf{X}_i) \right. \\
&\quad \left. + p_2(Y_i; \mathbf{X}_i^T \boldsymbol{\beta}_o + \eta_{\boldsymbol{\beta}_o}(T_i)) \left\{ \mathbf{X}_i + \frac{\partial \eta_{\boldsymbol{\beta}}(T_i)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} \right\} w_2(\mathbf{X}_i) \{\hat{\eta}_{\boldsymbol{\beta}_o}(T_i) - \eta_{\boldsymbol{\beta}_o}(T_i)\} \right. \\
&\quad \left. + \frac{1}{2} p_3(Y_i; \eta_i^{**}) \left\{ \mathbf{X}_i + \frac{\partial \eta_{\boldsymbol{\beta}}(T_i)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} \right\} w_2(\mathbf{X}_i) \{\hat{\eta}_{\boldsymbol{\beta}_o}(T_i) - \eta_{\boldsymbol{\beta}_o}(T_i)\}^2 \right] \\
&= T_{n,1} + T_{n,2} + T_{n,3},
\end{aligned}$$

where  $\eta_i^{**}$  is between  $\mathbf{X}_i^T \boldsymbol{\beta}_o + \hat{\eta}_{\boldsymbol{\beta}_o}(T_i)$  and  $\mathbf{X}_i^T \boldsymbol{\beta}_o + \eta_{\boldsymbol{\beta}_o}(T_i)$ ,

$$\begin{aligned}
T_{n,3} &= o_p(1), \\
T_{n,2} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n p_2(Y_i; \mathbf{X}_i^T \boldsymbol{\beta}_o + \eta_{\boldsymbol{\beta}_o}(T_i)) \left\{ \mathbf{X}_i + \frac{\partial \eta_{\boldsymbol{\beta}}(T_i)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} \right\} w_2(\mathbf{X}_i) \\
&\quad \times \frac{(-1)}{n f_T(T_i) g_2(T_i; T_i, \boldsymbol{\beta}_o)} \sum_{j=1}^n p_1(Y_j; \eta_j(T_i; \boldsymbol{\beta}_o)) w_1(\mathbf{X}_j) K_h(T_j - T_i) \\
&= -\frac{1}{\sqrt{n}} \sum_{j=1}^n \frac{p_1(Y_j; \eta_j) w_1(\mathbf{X}_j)}{g_2(T_j; T_j, \boldsymbol{\beta}_o)} E \left[ p_2(Y_j; \eta_j) \left\{ \mathbf{X}_j + \frac{\partial \eta_{\boldsymbol{\beta}}(T_j)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} \right\} w_2(\mathbf{X}_j) \Big| T_j \right] \\
&\equiv -\frac{1}{\sqrt{n}} \sum_{j=1}^n p_1(Y_j; \eta_j) \frac{\gamma(T_j)}{g_2(T_j; T_j, \boldsymbol{\beta}_o)} w_1(\mathbf{X}_j),
\end{aligned}$$

with:

$$\gamma(t) = E \left[ p_2(Y; \eta(X, T)) \left\{ X + \frac{\partial \eta_{\beta}(T)}{\partial \beta} \Big|_{\beta=\beta_o} \right\} w_2(X) \Big| T = t \right].$$

Therefore,

$$I_{n,1}^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n p_1(Y_i; \eta_i) \left[ \left\{ X_i + \frac{\partial \eta_{\beta}(T_i)}{\partial \beta} \Big|_{\beta=\beta_o} \right\} w_2(X_i) - \frac{\gamma(T_i)}{g_2(T_i; T_i, \beta_o)} w_1(X_i) \right] + o_p(1).$$

By the central limit theorem,

$$I_{n,1}^* \xrightarrow{\mathcal{D}} N(\mathbf{0}, \Omega_0^*), \quad (\text{A6})$$

where:

$$\begin{aligned} \Omega_0^* &= E \left( p_1^2(Y; \eta(X, T)) \left[ \left\{ X + \frac{\partial \eta_{\beta}(T)}{\partial \beta} \Big|_{\beta=\beta_o} \right\} w_2(X) - \frac{\gamma(T)}{g_2(T; T, \beta_o)} w_1(X) \right] \right. \\ &\quad \left. \left[ \left\{ X + \frac{\partial \eta_{\beta}(T)}{\partial \beta} \Big|_{\beta=\beta_o} \right\} w_2(X) - \frac{\gamma(T)}{g_2(T; T, \beta_o)} w_1(X) \right]^T \right). \end{aligned}$$

From (A5) and (A6),  $\hat{\theta} = -B_2^{-1} I_{n,1}^* + o_p(1)$ . This implies that  $\sqrt{n}(\hat{\beta} - \beta_o) \xrightarrow{\mathcal{D}} N(\mathbf{0}, H_0^{-1} \Omega_0^* H_0^{-1})$ .  $\square$

**Proof of Theorem 3.** Denote  $V_0 = H_0^{-1} \Omega_0^* H_0^{-1}$  and  $\hat{V}_n = \hat{H}_0^{-1} \hat{\Omega}_0^* \hat{H}_0^{-1}$ . Note that  $A\hat{\beta} - g_0 = A(\hat{\beta} - \beta_o) + (A\beta_o - g_0)$ . Thus:

$$\begin{aligned} &(A\hat{V}_n A^T)^{-1/2} \sqrt{n}(A\hat{\beta} - g_0) \\ &= (A\hat{V}_n A^T)^{-1/2} \{A\sqrt{n}(\hat{\beta} - \beta_o)\} + (A\hat{V}_n A^T)^{-1/2} \{\sqrt{n}(A\beta_o - g_0)\} \\ &\equiv I_1 + I_2, \end{aligned}$$

which implies that  $W_n = \|I_1 + I_2\|^2$ . Arguments for Theorem 2 give  $I_1 \xrightarrow{\mathcal{D}} N(\mathbf{0}, I_k)$ . Under  $H_0$  in (4),  $I_2 \equiv \mathbf{0}$  and thus  $(I_1 + I_2) \xrightarrow{\mathcal{D}} N(\mathbf{0}, I_k)$ , which completes the proof.  $\square$

**Proof of Theorem 4.** Follow the notation and proof in Theorem 3. Under  $H_{1n}$  in (29),  $I_2 \xrightarrow{P} (AV_0 A^T)^{-1/2} c$  and thus  $(I_1 + I_2) \xrightarrow{\mathcal{D}} N((AV_0 A^T)^{-1/2} c, I_k)$ . This completes the proof.  $\square$

**Proof of Theorem 5.** Following the notation and proof in Theorem 3,  $W_n = \|I_1\|^2 + 2I_1^T I_2 + \|I_2\|^2$ . We see that  $\|I_1\|^2 \xrightarrow{\mathcal{D}} \chi_k^2$ . Under  $H_1$  in (4),  $I_2 = (AV_0 A^T)^{-1/2} \sqrt{n}(A\beta_o - g_0)\{1 + o_p(1)\}$ , which means  $\|I_2\|^2 = n(A\beta_o - g_0)^T (AV_0 A^T)^{-1} (A\beta_o - g_0)\{1 + o_p(1)\}$  and thus  $I_1^T I_2 = O_p(\sqrt{n})$ . Hence,  $n^{-1} W_n \geq \lambda_{\min}\{(AV_0 A^T)^{-1}\} \|A\beta_o - g_0\|^2 + o_p(1)$ . This completes the proof.  $\square$

**Proof of Theorem 6.** Denote  $J_n(\beta) = J_n(\beta, \hat{\eta}_{\beta})$ . For the matrix  $A$  in (4), there exists a  $(d-k) \times d$  matrix  $B$  satisfying  $BB^T = I_{d-k}$  and  $AB^T = \mathbf{0}$ . Therefore,  $A\beta = g_0$  is equivalent to  $\beta = B^T \gamma + b_0$  for some vector  $\gamma \in \mathbb{R}^{d-k}$  and  $b_0 = A^T(AA^T)^{-1} g_0$ . Then, minimizing  $J_n(\beta)$  subject to  $A\beta = g_0$  is equivalent to minimizing  $J_n(B^T \gamma + b_0)$  with respect to  $\gamma$ , and we denote by  $\hat{\gamma}$  the minimizer. Furthermore, under  $H_0$  in (4), we have  $\beta_o = B^T \gamma_0 + b_0$  for  $\gamma_0 = B\beta_o$ , and  $\hat{\gamma} - \gamma_0 \xrightarrow{P} \mathbf{0}$ .

For Part (i), using the Taylor expansion around  $\hat{\beta}$ , we get:

$$J_n(B^T \hat{\gamma} + b_0) - J_n(\hat{\beta}) = \frac{1}{2n} \{ \sqrt{n}(B^T \hat{\gamma} + b_0 - \hat{\beta}) \}^T J_n''(\tilde{\beta}) \{ \sqrt{n}(B^T \hat{\gamma} + b_0 - \hat{\beta}) \}, \quad (\text{A7})$$

where  $\tilde{\beta}$  is between  $B^T\hat{\gamma} + \mathbf{b}_0$  and  $\hat{\beta}$ . We now discuss  $B^T\hat{\gamma} + \mathbf{b}_0 - \hat{\beta}$ . From the proof in Theorem 2,  $(\hat{\beta} - \beta_o) = -\mathbf{H}_0^{-1}\mathbf{J}'_n(\beta_o)\{1 + o_p(1)\}$ , where  $\mathbf{J}'_n(\beta_o) = \{I_{n,1}^* + o_p(1)\}/\sqrt{n}$ . Similar arguments deduce  $\hat{\gamma} - \gamma_0 = -(B\mathbf{H}_0 B^T)^{-1}B\mathbf{J}'_n(\beta_o)\{1 + o_p(1)\}$ . Thus, under  $H_0$  in (4),

$$B^T\hat{\gamma} + \mathbf{b}_0 - \hat{\beta} = B^T(\hat{\gamma} - \gamma_0) - (\hat{\beta} - \beta_o) = \mathbf{H}_0^{-1/2}P_{\mathbf{H}_0^{-1/2}\mathbf{A}^T}\mathbf{H}_0^{-1/2}\mathbf{J}'_n(\beta_o)\{1 + o_p(1)\},$$

and thus by (A6),

$$\sqrt{n}(B^T\hat{\gamma} + \mathbf{b}_0 - \hat{\beta}) \xrightarrow{\mathcal{D}} \mathbf{H}_0^{-1/2}P_{\mathbf{H}_0^{-1/2}\mathbf{A}^T}\mathbf{H}_0^{-1/2}\Omega_0^{*1/2}\mathbf{Z}, \quad (\text{A8})$$

where  $\mathbf{Z} = (Z_1, \dots, Z_d)^T \sim N(\mathbf{0}, \mathbf{I}_d)$ . Combining the fact  $\mathbf{J}''_n(\tilde{\beta}) \xrightarrow{P} \mathbf{H}_0$ , (A7) and (A8) gives:

$$\begin{aligned} \Lambda_n &= \{\sqrt{n}(B^T\hat{\gamma} + \mathbf{b}_0 - \hat{\beta})\}^T \mathbf{H}_0 \{\sqrt{n}(B^T\hat{\gamma} + \mathbf{b}_0 - \hat{\beta})\} \{1 + o_p(1)\} \\ &\xrightarrow{\mathcal{D}} \mathbf{Z}^T \Omega_0^{*1/2} \mathbf{H}_0^{-1/2} P_{\mathbf{H}_0^{-1/2}\mathbf{A}^T} \mathbf{H}_0^{-1/2} \Omega_0^{*1/2} \mathbf{Z} \\ &= \sum_{j=1}^d \lambda_j (\Omega_0^{*1/2} \mathbf{H}_0^{-1/2} P_{\mathbf{H}_0^{-1/2}\mathbf{A}^T} \mathbf{H}_0^{-1/2} \Omega_0^{*1/2}) Z_j^2 \\ &= \sum_{j=1}^k \lambda_j \{(\mathbf{A}\mathbf{H}_0^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{V}_0\mathbf{A}^T)\} Z_j^2. \end{aligned} \quad (\text{A9})$$

This proves Part (i).

For Part (ii), using  $\psi(r) = r$ ,  $w_1(x) = w_2(x) \equiv 1$  and (31), we obtain  $\Omega_0^* = \Omega_0 = C\mathbf{H}_0$ , and thus,  $\mathbf{A}\mathbf{V}_0\mathbf{A}^T = C(\mathbf{A}\mathbf{H}_0^{-1}\mathbf{A}^T)$ . Thus, (A9) =  $C \sum_{j=1}^k Z_j^2 \sim C\chi_k^2$ , which completes the proof.  $\square$

**Proof of Theorem 7.** The proofs are similar to those used in Theorem 4 and Theorems 5 and 6. The lengthy details are omitted.  $\square$

## References

1. Andrews, D. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* **1994**, *62*, 43–72.
2. Robinson, P.M. Root-n consistent semiparametric regression. *Econometrica* **1988**, *56*, 931–954.
3. Speckman, P. Kernel smoothing in partial linear models. *J. R. Statist. Soc. B* **1988**, *50*, 413–436.
4. Yatchew, A. An elementary estimator of the partial linear model. *Econ. Lett.* **1997**, *57*, 135–143.
5. Fan, J.; Li, R. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Am. Stat. Assoc.* **2004**, *99*, 710–723.
6. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman & Hall: London, UK, 1989.
7. Zhang, C.M.; Yu, T. Semiparametric detection of significant activation for brain fMRI. *Ann. Stat.* **2008**, *36*, 1693–1725.
8. Fan, J.; Huang, T. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **2005**, *11*, 1031–1057.
9. Boente, G.; He, X.; Zhou, J. Robust estimates in generalized partially linear models. *Ann. Stat.* **2006**, *34*, 2856–2878.
10. Zhang, C.M.; Guo, X.; Cheng, C.; Zhang, Z.J. Robust-BD estimation and inference for varying-dimensional general linear models. *Stat. Sin.* **2014**, *24*, 653–673.
11. Fan, J.; Gijbels, I. *Local Polynomial Modeling and Its Applications*; Chapman and Hall: London, UK, 1996.
12. Brègman, L.M. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 620–631.
13. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2001.
14. Zhang, C.M.; Jiang, Y.; Shang, Z. New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Can. J. Stat.* **2009**, *37*, 119–139.
15. Huber, P. Robust estimation of a location parameter. *Ann. Math. Statist.* **1964**, *35*, 73–101.
16. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 1998.

17. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22.
18. Carroll, R.; Fan, J.; Gijbels, I.; Wand, M. Generalized partially linear single-index models. *J. Am. Stat. Assoc.* **1997**, *92*, 477–489.
19. Mukarjee, H.; Stern, S. Feasible nonparametric estimation of multiargument monotone functions. *J. Am. Stat. Assoc.* **1994**, *89*, 77–80.
20. Albright, S.C.; Winston, W.L.; Zappe, C.J. *Data Analysis and Decision Making with Microsoft Excel*; Duxbury Press: Pacific Grove, CA, USA, 1999.
21. Fan, J.; Peng, H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.* **2004**, *32*, 928–961.
22. Dette, H. A consistent test for the functional form of a regression based on a difference of variance estimators. *Ann. Stat.* **1999**, *27*, 1012–1050.
23. Dette, H.; von Lieres und Wilkau, C. Testing additivity by kernel-based methods. *Bernoulli* **2001**, *7*, 669–697.
24. Fan, J.; Zhang, C.M.; Zhang, J. Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.* **2001**, *29*, 153–193.
25. Hong, Y.M.; Lee, Y.J. A loss function approach to model specification testing and its relative efficiency. *Ann. Stat.* **2013**, *41*, 1166–1203.
26. Zheng, J.X. A consistent test of functional form via nonparametric estimation techniques. *J. Econ.* **1996**, *75*, 263–289.
27. Opsomer, J.D.; Ruppert D. A root-n consistent backfitting estimator for semiparametric additive modeling. *J. Comput. Graph. Stat.* **1999**, *8*, 715–732.
28. Belloni, A.; Chernozhukov, V.; Hansen, C. Inference on treatment effects after selection amongst high-dimensional controls. *Rev. Econ. Stud.* **2014**, *81*, 608–650.
29. Cattaneo, M.D.; Jansson, M.; Newey, W.K. Alternative asymptotics and the partially linear model with many regressors. *Econ. Theory* **2016**, *1*–25.
30. Cattaneo, M.D.; Jansson, M.; Newey, W.K. Treatment effects with many covariates and heteroskedasticity. *arXiv* **2015**, arXiv:1507.02493.
31. Mack, Y.P.; Silverman, B.W. Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **1982**, *61*, 405–415.



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Robust Estimation for the Single Index Model Using Pseudodistances

Aida Toma <sup>1,2,\*</sup> and Cristinca Fulga <sup>1</sup>

<sup>1</sup> Department of Applied Mathematics, Bucharest Academy of Economic Studies, 010374 Bucharest, Romania; cristinca.fulga@gmail.com

<sup>2</sup> “Gh. Mihoc - C. Iacob” Institute of Mathematical Statistics and Applied Mathematics, Romanian Academy, 010071 Bucharest, Romania

\* Correspondence: aida\_toma@yahoo.com

Received: 31 March 2018; Accepted: 14 May 2018; Published: 17 May 2018



**Abstract:** For portfolios with a large number of assets, the single index model allows for expressing the large number of covariances between individual asset returns through a significantly smaller number of parameters. This avoids the constraint of having very large samples to estimate the mean and the covariance matrix of the asset returns, which practically would be unrealistic given the dynamic of market conditions. The traditional way to estimate the regression parameters in the single index model is the maximum likelihood method. Although the maximum likelihood estimators have desirable theoretical properties when the model is exactly satisfied, they may give completely erroneous results when outliers are present in the data set. In this paper, we define minimum pseudodistance estimators for the parameters of the single index model and using them we construct new robust optimal portfolios. We prove theoretical properties of the estimators, such as consistency, asymptotic normality, equivariance, robustness, and illustrate the benefits of the new portfolio optimization method for real financial data.

**Keywords:** minimum divergence methods; robustness; single index model

---

## 1. Introduction

The problem of portfolio optimization in the mean-variance approach depends on a large number of parameters that need to be estimated on the basis of relatively small samples. Due to the dynamics of market conditions, only a short period of market history can be used for estimation of the model's parameters. In order to reduce the number of parameters that need to be estimated, the single index model proposed by Sharpe (see [1,2]) can be used. The traditional estimators for parameters of the single index model are based on the maximum likelihood method. These estimators have optimal properties for normally distributed variables, but they may give completely erroneous results in the presence of outlying observations. Since the presence of outliers in financial asset returns is a frequently occurring phenomenon, robust estimates for the parameters of the single index model are necessary in order to provide robust and optimal portfolios.

Our contribution to robust portfolio optimization through the single index model is based on using minimum pseudodistance estimators.

The interest on statistical methods based on information measures and particularly on divergences has grown substantially in recent years. It is a known fact that, for a wide variety of models, statistical methods based on divergence measures have some optimal properties in relation to efficiency, but especially in relation to robustness, representing viable alternatives to the classical methods. We refer to the monographs of Pardo [3] and Basu et al. [4] for an excellent presentation of such methods, for their importance and applications.

We can say that the minimum pseudodistance methods for estimation go to the same category as the minimum divergence methods. The minimum divergence estimators are defined by minimizing some appropriate divergence between the assumed theoretical model and the true model corresponding to the data. Depending on the choice of the divergence, minimum divergence estimators can afford considerable robustness with a minimal loss of efficiency. The classical minimum divergence methods require nonparametric density estimation, which imply some difficulties such as the bandwidth selection. In order to avoid the nonparametric density estimation in minimum divergence estimation methods, some proposals have been made in [5–7] and robustness properties of such estimators have been studied in [8,9].

The pseudodistances that we use in the present paper were originally introduced in [6], where they are called "type-0" divergences, and corresponding minimum divergence estimators have been studied. They are also obtained (using a cross entropy argument) and extensively studied in [10] where they are called  $\gamma$ -divergences. They are also introduced in [11] in the context of decomposable pseudodistances. By its very definition, a pseudodistance satisfies two properties, namely the nonnegativity and the fact that the pseudodistance between two probability measures equals to zero if and only if the two measures are equal. The divergences are moreover characterized by the information processing property, i.e., by the complete invariance with respect to statistically sufficient transformations of the observation space (see [11], p. 617). In general, a pseudodistance may not satisfy this property. We adopted the term pseudodistance for this reason, but in the literature we can also meet the other terms above. The minimum pseudodistance estimators for general parametric models have been presented in [12] and consist of minimization of an empirical version of a pseudodistance between the assumed theoretical model and the true model underlying the data. These estimators have the advantages of not requiring any prior smoothing and conciliate robustness with high efficiency, usually requiring distinct techniques.

In this paper, we define minimum pseudodistance estimators for the parameters of the single index model and using them we construct new robust optimal portfolios. We study properties of the estimators, such as, consistency, asymptotic normality, robustness and equivariance and illustrate the benefits of the proposed portfolio optimization method through examples for real financial data.

We mention that we define minimum pseudodistance estimators, and prove corresponding theoretical properties, for the parameters of the simple linear regression model (35), associated with the single index model. However, in a very similar way, we can define minimum pseudodistance estimators and obtain the same theoretical results for the more general linear regression model  $Y_j = X_j^T \beta + e_j$ ,  $j = 1, \dots, n$ , where the errors  $e_j$  are i.i.d. normal variables with mean zero and variance  $\sigma^2$ ,  $X_j = (X_{j1}, \dots, X_{jp})^T$  is the vector of independent variables corresponding to the  $j$ -th observation and  $\beta = (\beta_1, \dots, \beta_p)^T$  represents the regression coefficients.

The rest of the paper is organized as follows. In Section 2, we present the problem of robust estimation for some portfolio optimization models. In Section 3, we present the proposed approach. We define minimum pseudodistance estimators for regression parameters corresponding to the single index model and obtain corresponding estimating equations. Some asymptotic properties and equivariance properties of these estimators are studied. The robustness issue for estimators is considered through the influence function analysis. Using minimum pseudodistance estimators, new optimal portfolios are defined. Section 4 presents numerical results illustrating the performance of the proposed methodology. Finally, the proofs of the theorems are provided in the Appendix A.

## 2. The Single Index Model

Portfolio selection represents the problem of allocating a given capital over a number of available assets in order to maximize the return of the investment while minimizing the risk. We consider a portfolio formed by a collection of  $N$  assets. The returns of the assets are given by the random vector  $X := (X_1, \dots, X_N)^T$ . Usually, it is supposed that  $X$  follows a multivariate normal distribution  $N_N(\mu, \Sigma)$ , with  $\mu$  being the vector containing the mean returns of the assets and  $\Sigma = (\sigma_{ij})$  the

covariance matrix of the assets returns. Let  $w := (w_1, \dots, w_N)^T$  be the vector of weights associated with the portfolio, where  $w_i$  is the proportion of capital invested in the asset  $i$ . Then, the total return of the portfolio is defined by the random variable

$$w^T X = w_1 X_1 + \dots + w_N X_N. \quad (1)$$

The mean and the variance of the portfolio return are given by

$$R(w) := w^T \mu, \quad (2)$$

$$S(w) := w^T \Sigma w. \quad (3)$$

A classical approach for portfolio selection is the mean-variance optimization introduced by Markowitz [13]. For a given investor's risk aversion  $\lambda > 0$ , the mean-variance optimization gives the optimal portfolio  $w^*$ , solution of the problem

$$\arg \max_w \{R(w) - \frac{\lambda}{2} S(w)\}, \quad (4)$$

with the constraint  $w^T e_N = 1$ ,  $e_N$  being the  $N$ -dimensional vector of ones. The solution of the optimization problem (4) is explicit, the optimal portfolio weights for a given value of  $\lambda$  being

$$w^* = \frac{1}{\lambda} \Sigma^{-1} (\mu - \eta e_N), \quad (5)$$

where

$$\eta = \frac{e_N^T \Sigma^{-1} \mu - \lambda}{e_N^T \Sigma^{-1} e_N}. \quad (6)$$

This is the case when short selling is allowed. When short selling is not allowed, we have a supplementary constraint in the optimization problem, namely all the weights  $w_i$  are positive.

Another classical approach for portfolio selection is to minimize the portfolio risk defined by the portfolio variance, under given constraints. This means determining the optimal portfolio  $w^*$  as a solution of the optimization problem

$$\arg \min_w S(w), \quad (7)$$

subject to  $R(w) = w^T \mu \geq \mu_0$ , for a given value  $\mu_0$  of the portfolio return.

However, the mean-variance analysis has been criticized for being sensitive to estimation errors of the mean and the covariance of the assets returns. For both optimization problems above, estimations of the input parameters  $\mu$  and  $\Sigma$  are necessary. The quality and hence the usefulness of the results of the portfolio optimization problem critically depend on the quality of the statistical estimates for these input parameters. The mean vector and the covariance matrix of the returns are in practice estimated by the maximum likelihood estimators under the multivariate normal assumption. When the model is exactly satisfied, the maximum likelihood estimators have optimal properties, being the most efficient. On the other hand, in the presence of outlying observations, these estimators may give completely erroneous results and consequently the weights of the corresponding optimal portfolio may be completely misleading. It is a known fact that outliers frequently occur in asset returns, where an outlier is defined to be an unusually large value well separated from the bulk of the returns. Therefore, robust alternatives to the classical approaches need to be carefully analyzed.

For an overview on the robust methods for portfolio optimization, using robust estimators of the mean and covariance matrix in the Markowitz's model, we refer to [14]. We also cite the methods proposed by Vaz-de Melo and Camara [15], Perret-Gentil and Victoria-Feser [16], Welsch and Zhou [17], DeMiguel and Nogales [18], and Toma and Leoni-Aubin [19].

On the other hand, in portfolio analysis, one is sometimes faced with two conflicting demands. Good quality statistical estimates require a large sample size. When estimating the covariance matrix, the sample size must be larger than the number of different elements of the matrix. For example, for a portfolio involving 100 securities, this would mean observations from 5050 trading days, which is about 20 years. From a practical point of view, considering such large samples is not adequate for the considered problem. Since the market conditions change rapidly, very old observations would lead to irrelevant estimates for the current or future market conditions. In addition, in some situations, the number of assets could even be much larger than the sample size of exploitable historical data. Therefore, estimating the covariance matrix of asset returns is challenging due to the high dimensionality and also to the heavy-tailedness of asset return data. It is a known fact that extreme events are typical in financial asset prices, leading to heavy-tailed asset returns. One way to treat these problems is to use the single index model.

The single index model (see [1]) allows us to express the large number of covariances between the returns of the individual assets through a significantly smaller number of parameters. This is possible under the hypothesis that the correlation between two assets is strictly given by their dependence on a common market index. The return of each asset  $i$  is expressed under the form

$$X_i = \alpha_i + \beta_i X_M + e_i, \quad (8)$$

where  $X_M$  is the random variable representing the return of the market index,  $e_i$  are zero mean random variables representing error terms and  $\alpha_i, \beta_i$  are new parameters to be estimated. It is supposed that the  $e_i$ 's are independent and also that the  $e_i$ 's are independent of  $x_M$ . Thus,  $E(e_i) = 0$ ,  $E(e_i e_j) = 0$  and  $E(e_i x_M) = 0$  for all  $i$  and all  $j \neq i$ .

The intercept in Equation (35) represents the asset's expected return when the market index return is zero. The slope coefficient  $\beta_i$  represents the asset's sensitivity to the index, namely the impact of a unit change in the return of the index. The error  $e_i$  is the return variation that cannot be explained by the index.

The following notations are also used:

$$\sigma_i^2 := \text{Var}(e_i), \quad \mu_M := E(X_M), \quad \sigma_M^2 := \text{Var}(X_M).$$

Using Equation (35), the components of the parameters  $\mu$  and  $\Sigma$  from the models (4) and (7) are given by

$$\mu_i = \alpha_i + \beta_i \mu_M, \quad (9)$$

$$\sigma_{ii} = \beta_i^2 \sigma_M^2 + \sigma_i^2, \quad (10)$$

$$\sigma_{ij} = \beta_i \beta_j \sigma_M^2. \quad (11)$$

Both variances and covariances are determined by the assets' betas and sigmas and by the standard deviation of the market index. Thus, the  $N(N + 1)/2$  different elements of the covariance matrix  $\Sigma$  can be expressed by  $2N + 1$  parameters  $\beta_i, \sigma_i, \sigma_M$ . This is a significant reduction of the number of parameters that need to be estimated.

The traditional estimators for parameters of the single index model are based on the maximum likelihood method. These estimators have optimal properties for normally distributed variables, but they may give completely erroneous results in the presence of outlying observations. Therefore, robust estimates for the parameters of the single index model are necessary in order to provide robust and optimal portfolios.

### 3. Robust Estimators for the Single Index Model and Robust Portfolios

#### 3.1. Definitions of the Estimators

Consider the linear regression model

$$X = \alpha + \beta X_M + e. \quad (12)$$

Suppose we have i.i.d. two-dimensional random vectors  $Z_j = (X_{Mj}, X_j)$ ,  $j = 1, \dots, n$ , such that  $X_j = \alpha + \beta X_{Mj} + e_j$ . The random variables  $e_j$ ,  $j = 1, \dots, n$ , are i.i.d. with  $\mathcal{N}(0, \sigma)$  and independent on the  $X_{Mj}$ ,  $j = 1, \dots, n$ .

The classical estimators for the unknown parameters  $\alpha, \beta, \sigma$  of the linear regression model are the maximum likelihood estimators (MLE). The classical MLE estimators perform well if the model hypotheses are satisfied exactly and may otherwise perform poorly. It is well known that the MLE are not robust, since a small fraction of outliers, even one outlier may have an important effect inducing significant errors on the estimates. Therefore, robust alternatives of the MLE should be considered, in order to propose robust estimates for the single index model, leading then to robust portfolio weights.

In order to robustly estimate the unknown parameters  $\alpha, \beta, \sigma$ , suppressing the outsized effects of outliers, we use the approach based on pseudodistance minimization.

For two probability measures  $P, Q$  admitting densities  $p$ , respectively,  $q$  with respect to the Lebesgue measure, we consider the following family of pseudodistances (also called  $\gamma$ -divergences in some articles) of orders  $\gamma > 0$

$$R_\gamma(P, Q) := \frac{1}{(1+\gamma)} \ln \left( \int p^\gamma dP \right) + \frac{1}{\gamma(1+\gamma)} \ln \left( \int q^\gamma dQ \right) - \frac{1}{\gamma} \ln \left( \int p^\gamma dQ \right), \quad (13)$$

satisfying the limit relation

$$R_\gamma(P, Q) \rightarrow R_0(P, Q) := \int \ln \frac{q}{p} dQ \text{ for } \gamma \downarrow 0.$$

Note that  $R_0(P, Q)$  is the well-known modified Kullback–Leibler divergence. Minimum pseudodistance estimators for parametric models, using the family (13), have been studied by [6,10,11]. We also mention that pseudodistances (13) have also been used for defining optimal robust M-estimators with the Hampel’s infinitesimal approach in [20].

For the linear regression model, we consider the joint distribution of the entire data, the explanatory variable  $X_M$  being random together with the response variable  $X$ , and write a pseudodistance between a theoretical model and the data. Let  $P_\theta$ , with  $\theta = (\alpha, \beta, \sigma)$ , be the probability measure associated with the theoretical model given by the random vector  $(X_M, X)$ , where  $X = \alpha + \beta X_M + e$  with  $e \sim \mathcal{N}(0, \sigma)$ ,  $e$  independent on  $X_M$ , and  $Q$  the probability measure associated with the data. Denote by  $p_\theta$ , respectively,  $q$ , the corresponding densities. For  $\gamma > 0$ , the pseudodistance between  $P_\theta$  and  $Q$  is defined by

$$R_\gamma(P_\theta, Q) := \frac{1}{(1+\gamma)} \ln \left( \int p_\theta^\gamma(x_M, x) dP_\theta(x_M, x) \right) + \frac{1}{\gamma(1+\gamma)} \ln \left( \int q^\gamma(x_M, x) dQ(x_M, x) \right) - \frac{1}{\gamma} \ln \left( \int p_\theta^\gamma(x_M, x) dQ(x_M, x) \right). \quad (14)$$

Using the change of variables  $(x_M, x) \rightarrow (u, v) := (x_M, x - \alpha - \beta x_M)$  and taking into account that  $f(u, v) := p_\theta(u, v + \alpha + \beta u)$  is the density of  $(X_M, e)$ , since  $X_M$  and  $e$  are independent, we can write

$$\int p_\theta^\gamma(x_M, x) dP_\theta(x_M, x) = \int p_M^{\gamma+1}(u) du \cdot \int \phi_\sigma^{\gamma+1}(v) dv, \quad (15)$$

$$\int p_\theta^\gamma(x_M, x) dQ(x_M, x) = \int p_M^\gamma(x_M) \cdot \phi_\sigma^\gamma(x - \alpha - \beta x_M) dQ(x_M, x), \quad (16)$$

where  $p_M$  is the density of  $X_M$  and  $\phi_\sigma$  is the density of the random variable  $e \sim \mathcal{N}(0, \sigma)$ . Then,

$$\begin{aligned} R_\gamma(P_\theta, Q) &= \frac{1}{(1+\gamma)} \ln \left( \int p_M^{\gamma+1}(u) du \right) + \frac{1}{(1+\gamma)} \ln \left( \int \phi_\sigma^{\gamma+1}(v) dv \right) \\ &\quad + \frac{1}{\gamma(1+\gamma)} \ln \left( \int q^\gamma(x_M, x) dQ(x_M, x) \right) \\ &\quad - \frac{1}{\gamma} \ln \left( \int p_M^\gamma(x_M) \cdot \phi_\sigma^\gamma(x - \alpha - \beta x_M) dQ(x_M, x) \right). \end{aligned}$$

Notice that the first and the third terms in the pseudodistance  $R_\gamma(P_\theta, Q)$  do not depend on  $\theta$  and hence are not included in the minimization process. The parameter  $\theta_0 := (\alpha_0, \beta_0, \sigma_0)$  of interest is then given by

$$\begin{aligned} (\alpha_0, \beta_0, \sigma_0) &:= \arg \min_{\alpha, \beta, \sigma} R_\gamma(P_\theta, Q) \\ &= \arg \min_{\alpha, \beta, \sigma} \left\{ \frac{1}{(1+\gamma)} \ln \left( \int \phi_\sigma^{\gamma+1}(v) dv \right) - \frac{1}{\gamma} \ln \left( \int p_M^\gamma(x_M) \cdot \phi_\sigma^\gamma(x - \alpha - \beta x_M) dQ(x_M, x) \right) \right\}. \end{aligned} \quad (17)$$

Suppose now that an i.i.d. sample  $Z_1, \dots, Z_n$  is available from the true model. For a given  $\gamma > 0$ , we define a minimum pseudodistance estimator of  $\theta_0 = (\alpha_0, \beta_0, \sigma_0)$  by minimizing an empirical version of the objective function in Equation (17). This empirical version is obtained by replacing  $p_M(x_M)$  with the empirical density function  $\hat{p}_M(x_M) = \frac{1}{n} \sum_{i=1}^n \delta(x_M - X_{Mi})$ , where  $\delta(\cdot)$  is the Dirac delta function, and  $Q$  with the empirical measure corresponding to the sample. More precisely, we define  $\hat{\theta} := (\hat{\alpha}, \hat{\beta}, \hat{\sigma})$

$$\begin{aligned} (\hat{\alpha}, \hat{\beta}, \hat{\sigma}) &:= \arg \min_{\alpha, \beta, \sigma} \left\{ \frac{1}{(1+\gamma)} \ln \left( \int \phi_\sigma^{\gamma+1}(v) dv \right) - \frac{1}{\gamma} \ln \left( \int \hat{p}_M^\gamma(x_M) \cdot \phi_\sigma^\gamma(x - \alpha - \beta x_M) dP_n(x_M, x) \right) \right\} \\ &= \arg \min_{\alpha, \beta, \sigma} \left\{ \frac{1}{(1+\gamma)} \ln \left( \int \phi_\sigma^{\gamma+1}(v) dv \right) - \frac{1}{\gamma} \ln \left( \frac{1}{n^{\gamma+1}} \sum_{j=1}^n \phi_\sigma^\gamma(X_j - \alpha - \beta X_{Mj}) \right) \right\}, \end{aligned} \quad (18)$$

or equivalently

$$\begin{aligned} (\hat{\alpha}, \hat{\beta}, \hat{\sigma}) &= \arg \max_{\alpha, \beta, \sigma} \sum_{j=1}^n \frac{\phi_\sigma^\gamma(X_j - \alpha - \beta X_{Mj})}{[\int \phi_\sigma^{\gamma+1}(v) dv]^{\gamma/(\gamma+1)}} \\ &= \arg \max_{\alpha, \beta, \sigma} \sum_{j=1}^n \sigma^{-\gamma/(\gamma+1)} \exp \left( -\frac{\gamma}{2} \left( \frac{X_j - \alpha - \beta X_{Mj}}{\sigma} \right)^2 \right). \end{aligned}$$

Differentiating with respect to  $\alpha, \beta, \sigma$ , the estimators  $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$  are solutions of the system

$$\sum_{j=1}^n \exp \left( -\frac{\gamma}{2} \left( \frac{X_j - \alpha - \beta X_{Mj}}{\sigma} \right)^2 \right) \left( \frac{X_j - \alpha - \beta X_{Mj}}{\sigma} \right) = 0, \quad (19)$$

$$\sum_{j=1}^n \exp \left( -\frac{\gamma}{2} \left( \frac{X_j - \alpha - \beta X_{Mj}}{\sigma} \right)^2 \right) \left( \frac{X_j - \alpha - \beta X_{Mj}}{\sigma} \right) X_{Mj} = 0, \quad (20)$$

$$\sum_{j=1}^n \exp \left( -\frac{\gamma}{2} \left( \frac{X_j - \alpha - \beta X_{Mj}}{\sigma} \right)^2 \right) \left[ \left( \frac{X_j - \alpha - \beta X_{Mj}}{\sigma} \right)^2 - \frac{1}{\gamma+1} \right] = 0. \quad (21)$$

Note that, for  $\gamma = 0$ , the solution of this system is nothing but the maximum likelihood estimator of  $(\alpha, \beta, \sigma)$ . Therefore, the estimating Equations (19)–(21) are generalizations of the maximum likelihood score equations. The tuning parameter  $\gamma$  associated with the pseudodistance controls the trade-off between robustness and efficiency of the minimum pseudodistance estimators.

We can also write that  $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma})$  is a solution of

$$\sum_{j=1}^n \Psi(Z_j, \hat{\theta}) = 0 \text{ or } \int \Psi(z, \hat{\theta}) dP_n(z) = 0, \quad (22)$$

where

$$\Psi(z, \theta) = \left( \phi \left( \frac{x - \alpha - \beta x_M}{\sigma} \right), \phi \left( \frac{x - \alpha - \beta x_M}{\sigma} \right) x_M, \chi \left( \frac{x - \alpha - \beta x_M}{\sigma} \right) \right)^T, \quad (23)$$

with  $z = (x_M, x)$ ,  $\theta = (\alpha, \beta, \sigma)$ ,  $\phi(t) = \exp(-\frac{\gamma}{2}t^2)t$  and  $\chi(t) = \exp(-\frac{\gamma}{2}t^2)[t^2 - \frac{1}{\gamma+1}]$ .

When the measure  $Q$  corresponding to the data pertain to the theoretical model, hence  $Q = P_{\theta_0}$ , it holds that

$$\int \Psi(z, \theta_0) dP_{\theta_0}(z) = 0. \quad (24)$$

Thus, we can consider  $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma})$  as a Z-estimator of  $\theta_0 = (\alpha_0, \beta_0, \sigma_0)$ , which allows for adapting in the present context asymptotic results from the general theory of Z-estimators (see [21]).

**Remark 1.** In the case when the density  $p_M$  is known, by replacing  $Q$  with the empirical measure  $P_n$  in Equation (17), a new class of estimators of  $(\alpha_0, \beta_0, \sigma_0)$  can be obtained. These estimators can also be written under the form of Z-estimators, using the same reasoning as above. The results of Theorems 1–4 below could be adapted for these new estimators, and moreover all the influence functions of these estimators would be redescending bounded. However, in practice, the density of the index return is not known. Therefore, we will work with the class of minimum pseudodistance estimators as defined above.

### 3.2. Asymptotic Properties

In order to prove the consistency of the estimators, we use their definition (22) as Z-estimators.

#### 3.2.1. Consistency

**Theorem 1.** Assume that, for any  $\epsilon > 0$ , the following condition for the separability of solution holds

$$\inf_{\theta \in M} \left\| \int \psi(z, \theta) dP_{\theta_0}(z) \right\| > 0 = \left\| \int \psi(z, \theta_0) dP_{\theta_0}(z) \right\|, \quad (25)$$

where  $M := \{\theta \text{ s.t. } \|\theta - \theta_0\| \geq \epsilon\}$ . Then,  $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma})$  converges in probability to  $\theta_0 = (\alpha_0, \beta_0, \sigma_0)$ .

#### 3.2.2. Asymptotic Normality

Assume that  $Z_1, \dots, Z_n$  are i.i.d. two-dimensional random vectors having the common probability distribution  $P_{\theta_0}$ . For  $\gamma > 0$  fixed, let  $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma})$  be a sequence of estimators of the unknown parameter  $\theta_0 = (\alpha_0, \beta_0, \sigma_0)$ , solution of

$$\sum_{j=1}^n \Psi(Z_j, \hat{\theta}) = 0, \quad (26)$$

where

$$\Psi(z, \theta) = \left( \sigma^2 \phi \left( \frac{x - \alpha - \beta x_M}{\sigma} \right), \sigma^2 \phi \left( \frac{x - \alpha - \beta x_M}{\sigma} \right) x_M, \sigma^2 \chi \left( \frac{x - \alpha - \beta x_M}{\sigma} \right) \right)^T, \quad (27)$$

with  $z = (x_M, x)$ ,  $\theta = (\alpha, \beta, \sigma)$ ,  $\phi(t) = \exp(-\frac{\gamma}{2}t^2)t$  and  $\chi(t) = \exp(-\frac{\gamma}{2}t^2)[t^2 - \frac{1}{\gamma+1}]$ . Note that the estimators  $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma})$  defined by Equations (19)–(21), or equivalently by (22), are also solutions of the system (26). Using the function (27) for defining the estimators allows for obtaining the asymptotic normality, only imposing the consistency condition of the estimators, without other supplementary assumptions that are usually imposed in the case of Z-estimators.

**Theorem 2.** Assume that  $\hat{\theta} \rightarrow \theta_0$  in probability. Then,

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow \mathcal{N}_3(0, B^{-1}A(B^{-1})^T) \quad (28)$$

in distribution, where  $A = E(\Psi(Z, \theta_0)\Psi(Z, \theta_0)^T)$  and  $B = E(\Psi(Z, \theta_0))$ , with  $\Psi$  defined by (27),  $\Psi$  being the matrix with elements  $\Psi_{ik} = \frac{\partial \Psi_i}{\partial \theta_k}$ .

After some calculations, we obtain the asymptotic covariance matrix of  $\hat{\theta}$  having the form

$$\sigma_0^2 \frac{(\gamma+1)^3}{(2\gamma+1)^{3/2}} \begin{pmatrix} \frac{\mu_M^2 + \sigma_M^2}{\sigma_M^2} & \frac{-\mu_M}{\sigma_M^2} & 0 \\ \frac{-\mu_M}{\sigma_M^2} & \frac{1}{\sigma_M^2} & 0 \\ 0 & 0 & \frac{3\gamma^2+4\gamma+2}{4(2\gamma+1)} \end{pmatrix}.$$

It follows that  $\hat{\beta}$  and  $\hat{\sigma}$  are asymptotically independent; in addition,  $\hat{\alpha}$  and  $\hat{\sigma}$  are asymptotically independent.

### 3.3. Influence Functions

In order to describe stability properties of the estimators, we use the following well-known concepts from the theory of robust statistics. A map  $T$ , defined on a set of probability measures and parameter space valued, is a statistical functional corresponding to an estimator  $\hat{\theta}$  of the parameter  $\theta$ , if  $\hat{\theta} = T(P_n)$ ,  $P_n$  being the empirical measure pertaining to the sample. The influence function of  $T$  at  $P_\theta$  is defined by

$$\text{IF}(z; T, P_\theta) := \left. \frac{\partial T(\tilde{P}_{\varepsilon z})}{\partial \varepsilon} \right|_{\varepsilon=0},$$

where  $\tilde{P}_{\varepsilon z} := (1 - \varepsilon)P_\theta + \varepsilon\delta_z$ ,  $\delta_z$  being the Dirac measure putting all mass at  $z$ . As a consequence, the influence function describes the linearized asymptotic bias of a statistic under a single point contamination of the model  $P_\theta$ . An unbounded influence function implies an unbounded asymptotic bias of a statistic under single point contamination of the model. Therefore, a natural robustness requirement on a statistical functional is the boundedness of its influence function.

For  $\gamma > 0$  fixed and a given probability measure  $P$ , the statistical functionals  $\alpha(P)$ ,  $\beta(P)$  and  $\sigma(P)$ , corresponding to the minimum pseudodistance estimators  $\hat{\alpha}$ ,  $\hat{\beta}$  and  $\hat{\sigma}$ , are defined by the solution of the system

$$\int \Psi(z, T(P))dP(z) = 0, \quad (29)$$

with  $\Psi$  defined by (23) and  $T(P) := (\alpha(P), \beta(P), \sigma(P))$ , whenever this solution exists.

When  $P = P_\theta$  corresponds to the considered theoretical model, the solution of system (29) is  $T(P_\theta) = \theta = (\alpha, \beta, \sigma)$ .

**Theorem 3.** The influence functions corresponding to the estimators  $\hat{\alpha}$ ,  $\hat{\beta}$  and  $\hat{\sigma}$  are respectively given by

$$IF(x_{M0}, x_0; \alpha, P_\theta) = \sigma(\gamma + 1)^{3/2} \phi\left(\frac{x_0 - \alpha - \beta x_{M0}}{\sigma}\right) \left[1 - \frac{(x_{M0} - E(X_M))E(X_M)}{Var(X_M)}\right], \quad (30)$$

$$IF(x_{M0}, x_0; \beta, P_\theta) = \sigma(\gamma + 1)^{3/2} \phi\left(\frac{x_0 - \alpha - \beta x_{M0}}{\sigma}\right) \frac{x_{M0} - E(X_M)}{Var(X_M)}, \quad (31)$$

$$IF(x_{M0}, x_0; \sigma, P_\theta) = \frac{\sigma(\gamma + 1)^{5/2}}{2} \chi\left(\frac{x_0 - \alpha - \beta x_{M0}}{\sigma}\right). \quad (32)$$

Since  $\chi$  is redescending,  $\hat{\sigma}$  has a bounded influence function and hence it is a redescending B-robust estimator. On the other hand,  $IF(x_{M0}, x_0, \alpha, P)$  and  $IF(x_{M0}, x_0, \beta, P)$  will tend to infinity only when  $x_{M0}$  tends to infinity and  $|\frac{x_0 - \alpha - \beta x_{M0}}{\sigma}| \leq k$ , for some  $k$ . Hence, these influence functions are bounded with respect to partial outliers or leverage points (outlying values of the independent variable). This means that large outliers with respect to  $x_M$ , or with respect to  $x$ , will have a reduced influence on the estimates. However, the influence functions are clearly unbounded for  $\gamma = 0$ , which corresponds to the non-robust maximum likelihood estimators.

### 3.4. Equivariance of the Regression Coefficients' Estimators

If an estimator is equivariant, it means that it transforms "properly" in some sense. Rousseeuw and Leroy [22] (p. 116) discuss three important equivariance properties for a regression estimator: regression equivariance, scale equivariance and affine equivariance. These are desirable properties since they allow one to know how the estimates change under different types of transformations of the data. Regression equivariance means that any additional linear dependence is reflected in the regression vector accordingly. The regression equivariance is routinely used when studying regression estimators. It allows for assuming, without loss generality, any value for the parameter  $(\alpha, \beta)$  for proving asymptotic properties or describing Monte-Carlo studies. An estimator being scale equivariant means that the fit produced by it is independent of the choice of measurement unit for the response variable. The affine equivariance is useful because it means that changing to a different co-ordinate system for the explanatory variable will not affect the estimate. It is known that the maximum likelihood estimator of the regression coefficients satisfies all these three properties. We show that the minimum pseudodistance estimators of the regression coefficients satisfy all the three equivariance properties, for all  $\gamma > 0$ .

**Theorem 4.** For all  $\gamma > 0$ , the minimum pseudodistance estimators  $(\hat{\alpha}, \hat{\beta})^T$  of the regression coefficients  $(\alpha, \beta)^T$  are regression equivariant, scale equivariant and affine equivariant.

On the other hand, the objective function in the definition of the estimators depends on data only through the summation

$$\sum_{j=1}^n \sigma^{-\gamma/(\gamma+1)} \exp\left(-\frac{\gamma}{2} \left(\frac{X_j - \alpha - \beta X_{Mj}}{\sigma}\right)^2\right), \quad (33)$$

which is permutation invariant. Thus, the corresponding estimators of the regression coefficients and of the error standard deviation are permutation invariant, therefore the ordering of data does not affect the estimators.

The minimum pseudodistance estimators are also equivariant with respect to reparametrizations. If  $\theta = (\alpha, \beta, \sigma)$  and the model is reparametrized to  $Y = Y(\theta)$  with a one-to-one transformation, then the minimum pseudodistance estimator of  $Y$  is simply  $\hat{Y} = Y(\hat{\theta})$ , in terms of the minimum pseudodistance estimator  $\hat{\theta}$  of  $\theta$ , for the same  $\gamma$ .

### 3.5. Robust Portfolios Using Minimum Pseudodistance Estimators

The robust estimation of the parameters  $\alpha_i, \beta_i, \sigma_i$  from the single index model given by (35), using minimum pseudodistance estimators, together with the robust estimation of  $\mu_M$  and  $\sigma_M$  lead to robust estimates of  $\mu$  and  $\Sigma$ , on the basis of relations (9)–(11). Since we do not model the explanatory variable  $X_M$  in a specific way, we estimate  $\mu_M$  and the standard deviation  $\sigma_M$  using as robust estimators the median, respectively the median absolute deviation. Then, the portfolio weights, obtained as solutions of the optimization problems (4) or (7) with input parameters robustly estimated, will also be robust. This methodology leads to new optimal robust portfolios. In the next section, on the basis of real financial data, we illustrate this new methodology and compare it with the traditional method based on maximum likelihood estimators.

## 4. Applications

### 4.1. Comparisons of the Minimum Pseudodistance Estimators with Other Robust Estimators for the Linear Regression Model

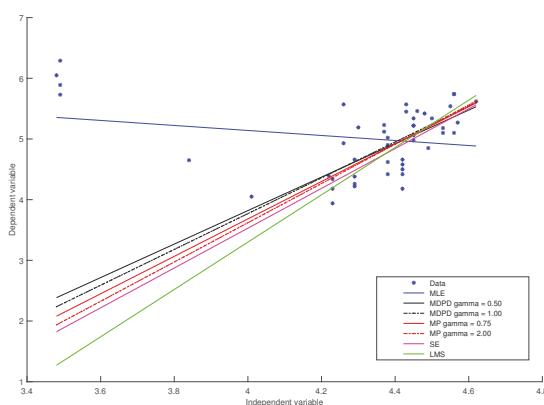
In order to illustrate the performance of the minimum pseudodistance estimators for the simple linear regression model, we compare them with the least median of squares (LMS) estimator (see [22,23]), with S-estimators (SE) (see [24]) and with the minimum density power divergence (MDPD) estimators (see [25]), estimators that are known to have a good behavior from the robustness point of view.

We considered a data set that comes from astronomy, namely the data from the Hertzsprung–Russell diagram of the star clusters CYG OB1 containing 47 stars in the direction of Cygnus. For these data, the independent variable is the logarithm of the effective temperature at the surface of the star and the dependent variable is the logarithm of its light intensity. The data are given in Rousseeuw and Leroy [22] (p. 27), who underlined that there are two groups of points: the majority, following a steep band, and four stars clearly forming a separate group from the rest of the data. These four stars are known as giants in astronomy. Thus, these outliers are not recording errors, but represents leverage points coming from a different group.

The estimates of the regression coefficients and of error standard deviation obtained with minimum pseudodistance estimators for several values of  $\gamma$  are given in Table 1 and some of the fitted models are plotted in Figure 1. For comparison, in Table 1, we also give estimates obtained with S-estimators based on the Tukey biweighted function, these estimates being taken from [24], as well as estimations obtained with minimum density power divergence methods for several values of the tuning parameter, and estimates obtained with the least median of squares method, all these estimates being taken from [25]. The MLE estimates, given on the first line of Table 1, are significantly affected by the four leverage points. On the other hand, like the robust least median of squares estimator, the robust S-estimators and some minimum density power divergence estimators, the minimum pseudodistance estimators with  $\gamma \geq 0.32$  can successfully ignore outliers. In addition, the minimum pseudodistance estimators with  $\gamma \geq 0.5$  give robust fits that are closer to the fits generated by the least median of squares estimates or by the S-estimates than the fits generated by the minimum density power divergence estimates.

**Table 1.** The parameter estimates for the linear regression model for the Hertzsprung–Russell data using several minimum pseudodistance (MP) methods, several minimum density power divergence (MDPD) methods, the least median of squares (LMS) method, S-estimators and the MLE method.  $\gamma$  represents tuning parameter.

MLE Estimates			
	$\alpha$	$\beta$	$\sigma$
6.79	−0.41	0.55	
MP Estimates			
$\gamma$	$\alpha$	$\beta$	$\sigma$
0.01	6.79	−0.41	0.55
0.1	6.81	−0.41	0.56
0.25	6.86	−0.42	0.58
0.3	6.88	−0.42	0.59
0.31	6.89	−0.43	0.59
0.32	−6.81	2.66	0.39
0.35	−7.16	2.74	0.38
0.4	−7.62	2.85	0.38
0.5	−8.17	2.97	0.37
0.75	−8.65	3.08	0.38
1	−8.84	3.12	0.39
1.2	−8.94	3.15	0.40
1.5	−9.08	3.18	0.41
2	−9.31	3.23	0.43
MDPD Estimates			
$\gamma$	$\alpha$	$\beta$	$\sigma$
0.1	6.78	−0.41	0.60
0.25	−5.16	2.30	0.42
0.5	−7.22	2.76	0.40
0.8	−7.89	2.91	0.40
1	−8.03	2.95	0.41
S-Estimates			
	$\alpha$	$\beta$	$\sigma$
−9.59	3.28	—	—
LMS Estimates			
	$\alpha$	$\beta$	$\sigma$
−12.30	3.90	—	—



**Figure 1.** Plots of the Hertzsprung–Russell data and fitted regression lines using MLE, minimum density power divergence (MDPD) methods for several values of  $\gamma$ , minimum pseudodistance (MP) methods for several values of  $\gamma$ , S-estimators (SE) and the least median of squares (LMS) method.

#### 4.2. Robust Portfolios Using Minimum Pseudodistance Estimators

In order to illustrate the performance of the proposed robust portfolio optimization method, we considered real data sets for the Russell 2000 index and for 50 stocks from its components. The stocks are listed in Appendix B. We selected daily return data for the Russell 2000 index and for all these stocks from 2 January 2013 to 30 June 2016. The data were retrieved from Yahoo Finance.

The data has been divided by quarter, in total 14 quarters for index and each stock. For each quarter, on the basis of data corresponding to the index, we estimated  $\mu_M$  and the standard deviation  $\sigma_M$  using as robust estimators the median (MED), respectively the median absolute deviation (MAD) defined by

$$MAD := \frac{1}{0.6745} \cdot MED(|X_i - MED(X_i)|). \quad (34)$$

We also estimated  $\mu_M$  and  $\sigma_M$  classically, using sample mean and sample variance. Then, for each quarter and each of the 50 stocks, we estimated  $\alpha$ ,  $\beta$  and  $\sigma$  from the regression model using robust minimum pseudodistance estimators, respectively the classical MLE estimators. Then, on the basis of relations (9), (10) and (11), we estimated  $\mu$  and  $\Sigma$  first using the robust estimates and then the classical estimates, all being previously computed.

Once the input parameters for the portfolio optimization procedure were estimated, for each quarter, we determined efficient frontiers, for both robust estimates and classical estimates. In both cases, the efficient frontier is determined as follows. Firstly, the range of returns is determined as the interval comprised between the return of the portfolio of global minimum risk (variance) and the maximum value of the return of a feasible portfolio, where the feasible region is

$$X = \left\{ w \in \mathbb{R}^N \mid w^T e_N = 1, w_k \geq 0, k \in \{1, \dots, 50\} \right\}$$

and  $N = 50$ . We trace each efficient frontier in 100 points; therefore, the range of returns is divided, in each case, in ninety-nine sub-intervals with

$$\mu_1 < \mu_2 < \dots < \mu_{100},$$

where  $\mu_1$  is the return of the portfolio of global minimum variance and  $\mu_{100}$  is the maximum return for the feasible region  $X$ . We determined  $\mu_1$  and  $\mu_{100}$  using robust estimates of  $\mu$  and  $\Sigma$  (for the robust frontier) and then using classical estimates (for the classical frontier). In each case, 100 optimization problems are solved:

$$\begin{aligned} & \arg \min_{w \in \mathbb{R}^N} S(w) \\ & w_k \geq 0, k \in \{1, \dots, 50\} \\ & w^T e_N = 1 \\ & R(w) \geq \mu_i, \end{aligned}$$

where  $i \in \{1, \dots, 100\}$ .

In Figure 2, for eight quarters (the first four quarters and the last four quarters), we present efficient frontiers corresponding to the optimal minimum variance portfolios based on the robust minimum pseudodistance estimates with  $\gamma = 0.5$ , respectively based on the classical estimates. Thus, on the  $ox$ -axis, we consider the portfolio risk (given by the portfolio standard deviation) and, on the  $oy$ -axis, we represent the portfolio return. We notice that, in comparison with the classical method based on MLE, the proposed robust method provides optimal portfolios that have higher returns for the same level of risk (standard deviation). Indeed, for each quarter, the robust frontier is situated above the classical one, the standard deviations of the robust portfolios being smaller compared with those of the classical portfolios. We obtained similar results for the other quarters and for other choices of the tuning parameter  $\gamma$ , corresponding to the minimum pseudodistance estimators, too.

We also illustrate the empirical performance of the proposed optimal portfolios through an out-of-sample analysis, by using the Sharpe ratio as out-of-sample measure. For this analysis, we apply a “rolling-horizon” procedure as presented in [18]. First, we choose a window over which to perform the estimation. We denote the length of the estimation window by  $\tau < T$ , where  $T$  is the size of the entire data set. Then, using the data in the first estimation window, we compute the weights for the considered portfolios. We repeat this procedure for the next window, by including the data for the next day and dropping the data for the earliest day. We continue doing this until the end of the data set is reached. At the end of this process, we have generated  $T - \tau$  portfolio weight vectors for each strategy, which are the vectors  $w_t^k$  for  $t \in \{\tau, \dots, T - 1\}$ ,  $k$  denoting the strategy. For a strategy  $k$ ,  $w_t^k$  has the components  $w_{j,t}^k$ , where  $w_{j,t}^k$  denotes the portfolio weight in asset  $j$  chosen at the time  $t$ .

The out-of-sample return at the time  $t + 1$ , corresponding to the strategy  $k$ , is defined as  $(w_t^k)^T X_{t+1}$ ,  $X_{t+1} := (X_{1,t+1}, \dots, X_{N,t+1})^T$  representing the data at the time  $t + 1$ . For each strategy  $k$ , using these out-of-sample returns, the out-of-sample mean and the out-of-sample variance are defined by

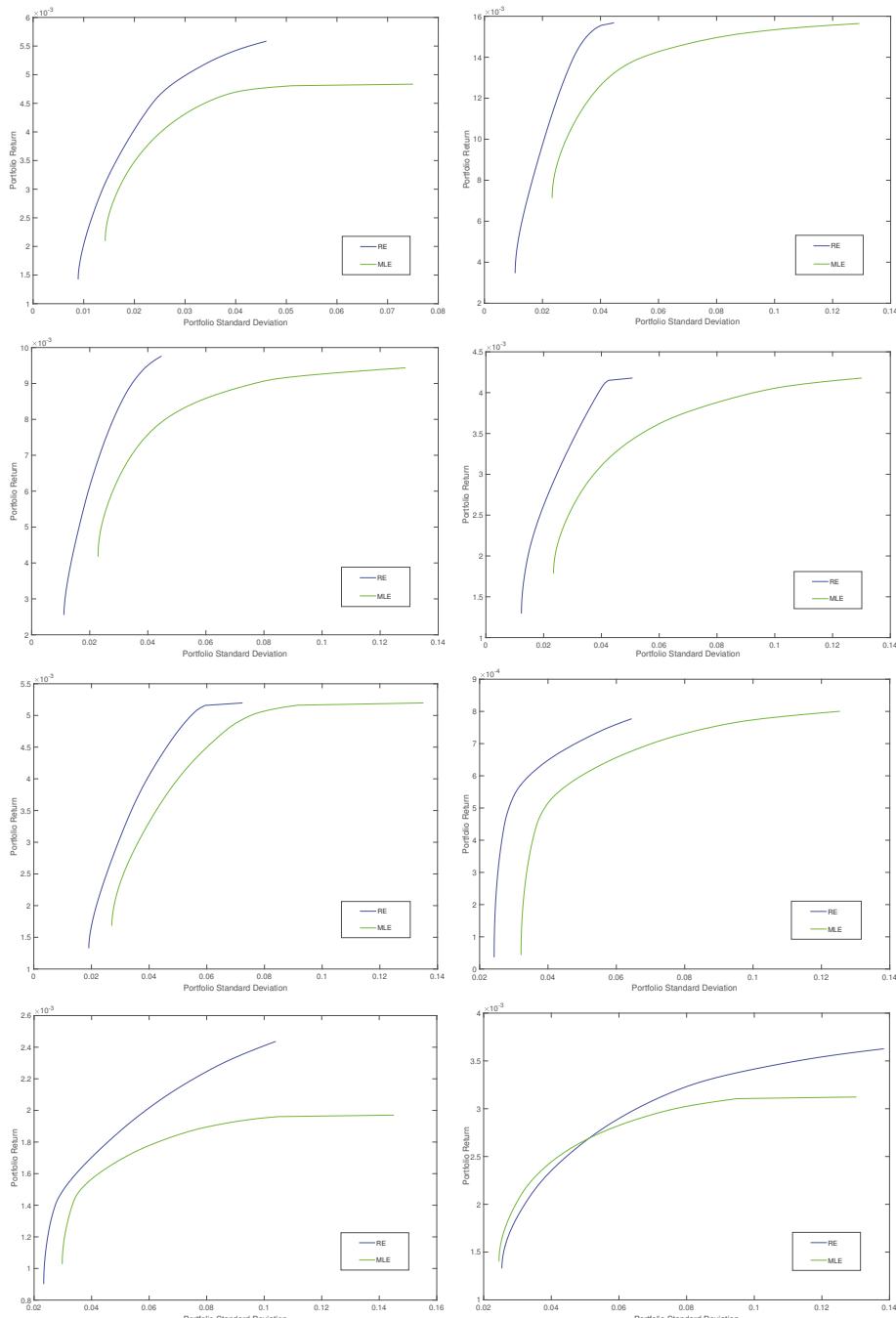
$$\hat{\mu}^k = \frac{1}{T - \tau} \sum_{t=\tau}^{T-1} (w_t^k)^T X_{t+1} \quad \text{and} \quad (\hat{\sigma}^k)^2 = \frac{1}{T - \tau - 1} \sum_{t=\tau}^{T-1} ((w_t^k)^T X_{t+1} - \hat{\mu}^k)^2 \quad (35)$$

and the out-of-sample Sharpe ratio is defined by

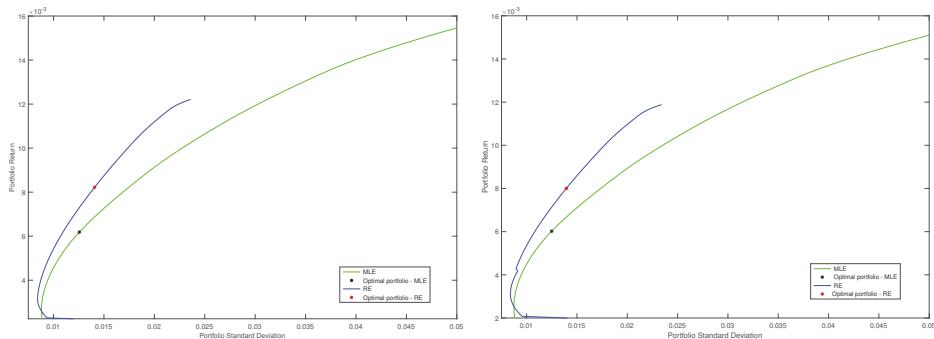
$$\widehat{SR}^k = \frac{\hat{\mu}^k}{\hat{\sigma}^k}. \quad (36)$$

In this example, we considered the data set corresponding to the quarters 13 and 14. The size of the entire data set was  $T = 126$  and the length of the estimation window was  $\tau = 63$  points. For the data from the first window, classical and robust efficient frontiers were traced, following all the steps that we explained in the first part of this subsection. More precisely, we considered the classical efficient frontier corresponding to the optimal minimum variance portfolios based on MLE and three robust frontiers, corresponding to the optimal minimum variance portfolios using robust minimum pseudodistance estimations with  $\gamma = 1$ ,  $\gamma = 1.2$  and  $\gamma = 1.5$ , respectively. Then, on each frontier, we chose the optimal portfolio associated with the maximal value of the ratio between the portfolio return and portfolio standard deviation. These four optimal portfolios represent the strategies that we compared in the out-of-sample analysis. For each of these portfolios, we computed the out-of-sample returns for the next time (next day). Then, we repeated all these procedures for the next window, and so on until the end of the data set has been reached. In the spirit of [18] Section 5, using (35) and (36), we computed out-of-sample means, out-of-sample variances and out-of-sample Sharpe ratios for each strategy. The out-of-sample means and out-of-sample variances were annualized, and we also considered a benchmark rate of 1.5 %. In this way, we obtained the following values for the out-of-sample Sharpe ratio:  $\widehat{SR} = 0.22$  for the optimal portfolio based on MLE,  $\widehat{SR} = 0.74$  for the optimal portfolio based on minimum pseudodistance estimations with  $\gamma = 1$ ,  $\widehat{SR} = 0.71$  for the optimal portfolio based on minimum pseudodistance estimations with  $\gamma = 1.2$  and  $\widehat{SR} = 0.29$  for the optimal portfolio based on minimum pseudodistance estimations with  $\gamma = 1.5$ . In Figure 3, we illustrate efficient frontiers for the windows 7 and 8, as well as the optimal portfolios chosen on each frontier.

This example shows that the optimal minimum variance portfolios based on robust minimum pseudodistance estimations in the single index model may attain higher Sharpe ratios than the traditional optimal minimum variance portfolios given by the single index model using MLE.



**Figure 2.** Efficient frontiers, classical (MLE) vs. robust corresponding to  $\gamma = 0.5$  (RE), for eight quarters (the first four quarters and the last four quarters).



**Figure 3.** Efficient frontiers, classical (MLE) vs. robust corresponding to  $\gamma = 1$  (RE), and optimal portfolios chosen on frontiers, for the windows 7 (left) and 8 (right).

The obtained numerical results show that, for the single index model, the presented robust technique for portfolio optimization yields better results than the classical method based on MLE, in the sense that it leads to larger returns for the same value of risk in the case when outliers or atypical observations are present in the data set. The considered data sets contain such outliers. This is often the case for the considered problem, since outliers frequently occur in asset returns data. However, when there are no outliers in the data set, the classical method based on MLE is more efficient than the robust ones and therefore may lead to better results.

## 5. Conclusions

When outliers or atypical observations are present in the data set, the new portfolio optimization method based on robust minimum pseudodistance estimates yields better results than the classical single index method based on MLE estimates, in the sense that it leads to larger returns for smaller risks. In literature, there exist various methods for robust estimation in regression models. In the present paper, we proposed the method based on the minimum pseudodistance approach, which suppose to solve a simple optimization problem. In addition, from a theoretical point of view, these estimators have attractive properties, such as being redescending robust, consistent, equivariant and asymptotically normally distributed. The comparison with other known robust estimators of the regression parameters, such as the least median of squares estimators, the S-estimators or the minimum density power divergence estimators, shows that the minimum pseudodistance estimators represent an attractive alternative that may be considered in other applications too.

**Author Contributions:** A.T. designed the methodology, obtained the theoretical results and wrote the paper. A.T. and C.F. conceived the application part. C.F. implemented the methods in MATLAB and obtained the numerical results. Both authors have read and approved the final manuscript.

**Acknowledgments:** This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS-UEFISCDI, project number PN-II-RU-TE-2012-3-0007.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Proof of the Results

**Proof of Theorem 1.** Since the functions  $\phi$  and  $\chi$  are redescending bounded functions, for a compact neighborhood  $N_{\theta_0}$  of  $\theta_0$ , it holds

$$\int \sup_{\theta \in N_{\theta_0}} \|\Psi(z, \theta)\| dP_{\theta_0}(z) < \infty. \quad (\text{A1})$$

Since  $\theta \mapsto \Psi(z, \theta)$  is continuous, by the uniform law of large numbers, (A1) implies

$$\sup_{\theta \in N_{\theta_0}} \left\| \int \psi(z, \theta) dP_n(z) - \int \psi(z, \theta) dP_{\theta_0}(z) \right\| \rightarrow 0 \quad (\text{A2})$$

in probability.

Then, (A2) together with assumption (25) assure the convergence in probability of  $\hat{\theta}$  toward  $\theta_0$ . The arguments are the same as those from van der Vaart [21], Theorem 5.9, p. 46.  $\square$

**Proof of Theorem 2.** First, note that  $\Psi$  defined by (27) is twice differentiable with respect to  $\theta$  with bounded derivatives. The matrix  $\Psi(z, \theta)$  has the form

$$\begin{pmatrix} -\sigma\phi' \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) & -\sigma\phi' \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) x_M & 2\sigma\phi \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) - \sigma\phi' \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) \\ -\sigma\phi' \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) x_M & -\sigma\phi' \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) x_M^2 & 2\sigma\phi \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) x_M - \sigma\phi' \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) x_M \\ -\sigma\chi' \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) & -\sigma\chi' \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) x_M & 2\sigma\chi \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) - \sigma\chi' \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) \left( \frac{x-\alpha-\beta x_M}{\sigma} \right) \end{pmatrix}$$

with  $\phi'(t) = [1 - \gamma t^2] \exp(-\frac{\gamma}{2} t^2)$  and  $\chi'(t) = [\frac{3\gamma+2}{\gamma+1} t - \gamma t^3] \exp(-\frac{\gamma}{2} t^2)$ . Since  $\phi(t), \chi(t), \phi'(t), \chi'(t)$  are redescending bounded functions, for  $\theta = \theta_0$ , it holds

$$|\Psi_{ik}(z, \theta_0)| \leq K(z) \text{ with } E(K(Z)) < \infty. \quad (\text{A3})$$

In addition, a simple calculation shows that each component  $\frac{\partial \Psi_i}{\partial \theta_k \partial \theta_l}$  is a bounded function, since it can be expressed through the functions  $\phi(t), \chi(t), \phi'(t), \chi'(t), \phi''(t), \chi''(t)$ , which are redescending bounded functions. In addition, bounds that can be established for each component  $\frac{\partial \Psi_i}{\partial \theta_k \partial \theta_l}$  do not depend on the parameter  $\theta$ .

For each  $i$ , call  $\ddot{\Psi}_i$  the matrix with elements  $\frac{\partial \Psi_i}{\partial \theta_k \partial \theta_l}$  and  $C_n(z, \theta)$  the matrix with its  $i$ -th raw equal to  $(\hat{\theta} - \theta_0)^T \ddot{\Psi}_i(z, \theta)$ . Using a Taylor expansion, we get

$$0 = \sum_{j=1}^n \Psi(Z_j, \hat{\theta}) = \sum_{j=1}^n \{\Psi(Z_j, \theta_0) + \dot{\Psi}(Z_j, \theta_0)(\hat{\theta} - \theta_0) + \frac{1}{2} C_n(Z_j, \theta_j)(\hat{\theta} - \theta_0)\}. \quad (\text{A4})$$

Therefore,

$$0 = A_n + (B_n + \bar{C}_n)(\hat{\theta} - \theta_0) \quad (\text{A5})$$

with

$$A_n = \frac{1}{n} \sum_{j=1}^n \Psi(Z_j, \theta_0), \quad B_n = \frac{1}{n} \sum_{j=1}^n \dot{\Psi}(Z_j, \theta_0), \quad \bar{C}_n = \frac{1}{2n} \sum_{j=1}^n C_n(Z_j, \theta_j) \quad (\text{A6})$$

i.e.,  $\bar{C}_n$  is the matrix with its  $i$ -th raw equal to  $(\hat{\theta} - \theta_0)^T \ddot{\Psi}_i$ , where

$$\ddot{\Psi}_i^- = \frac{1}{2n} \sum_{j=1}^n \ddot{\Psi}_i(Z_j, \theta_j), \quad (\text{A7})$$

which is bounded by a constant that does not depend on  $\theta$ , according to the arguments mentioned above. Since  $\hat{\theta} - \theta_0 \rightarrow 0$  in probability, this implies that  $\bar{C}_n \rightarrow 0$  in probability.

We have

$$\sqrt{n}(\hat{\theta} - \theta_0) = -(B_n + \bar{C}_n)^{-1} \sqrt{n} A_n. \quad (\text{A8})$$

Note that, for  $j = 1, \dots, n$ , the vectors  $\Psi(Z_j, \theta_0)$  are i.i.d. with mean zero and the covariance matrix  $A$ , and the matrices  $\dot{\Psi}(Z_j, \theta_0)$  are i.i.d. with mean  $B$ . Hence, when  $n \rightarrow \infty$ , using (A3), the law of

large numbers implies that  $B_n \rightarrow B$  in probability, which implies  $B_n + \bar{C}_n \rightarrow B$  in probability, which is nonsingular. Then, the multivariate central limit theorem implies  $\sqrt{n}A_n \rightarrow \mathcal{N}_3(0, A)$  in distribution.

Then,

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow \mathcal{N}_3(0, B^{-1}A(B^{-1})^T) \quad (\text{A9})$$

in distribution, according to the multivariate Slutzki's Lemma.  $\square$

**Proof of Theorem 3.** The system (29) can be written as

$$\begin{aligned} \int \phi \left( \frac{x - \alpha(P) - \beta(P)x_M}{\sigma(P)} \right) dP(x_M, x) &= 0, \\ \int \phi \left( \frac{x - \alpha(P) - \beta(P)x_M}{\sigma(P)} \right) x_M dP(x_M, x) &= 0, \\ \int \chi \left( \frac{x - \alpha(P) - \beta(P)x_M}{\sigma(P)} \right) dP(x_M, x) &= 0. \end{aligned}$$

We consider the contaminated model  $\tilde{P}_{\varepsilon, x_{M0}, x_0} := (1 - \varepsilon)P_\theta + \varepsilon\delta_{(x_{M0}, x_0)}$ , where  $\delta_{(x_{M0}, x_0)}$  is the Dirac measure putting all mass in the point  $(x_{M0}, x_0)$ , which we simply denote here by  $\tilde{P}_\varepsilon$ . Then, it holds

$$(1 - \varepsilon) \int \phi \left( \frac{x - \alpha(\tilde{P}_\varepsilon) - \beta(\tilde{P}_\varepsilon)x_M}{\sigma(\tilde{P}_\varepsilon)} \right) dP_\theta(x_M, x) + \varepsilon\phi \left( \frac{x_0 - \alpha(\tilde{P}_\varepsilon) - \beta(\tilde{P}_\varepsilon)x_{M0}}{\sigma(\tilde{P}_\varepsilon)} \right) = 0, \quad (\text{A10})$$

$$(1 - \varepsilon) \int \phi \left( \frac{x - \alpha(\tilde{P}_\varepsilon) - \beta(\tilde{P}_\varepsilon)x_M}{\sigma(\tilde{P}_\varepsilon)} \right) x_M dP_\theta(x_M, x) + \varepsilon\phi \left( \frac{x_0 - \alpha(\tilde{P}_\varepsilon) - \beta(\tilde{P}_\varepsilon)x_{M0}}{\sigma(\tilde{P}_\varepsilon)} \right) x_{M0} = 0, \quad (\text{A11})$$

$$(1 - \varepsilon) \int \chi \left( \frac{x - \alpha(\tilde{P}_\varepsilon) - \beta(\tilde{P}_\varepsilon)x_M}{\sigma(\tilde{P}_\varepsilon)} \right) dP_\theta(x_M, x) + \varepsilon\chi \left( \frac{x_0 - \alpha(\tilde{P}_\varepsilon) - \beta(\tilde{P}_\varepsilon)x_{M0}}{\sigma(\tilde{P}_\varepsilon)} \right) = 0. \quad (\text{A12})$$

Derivating the first equation with respect to  $\varepsilon$  and taking the derivatives in  $\varepsilon = 0$ , we obtain

$$\begin{aligned} \int \phi' \left( \frac{x - \alpha - \beta x_M}{\sigma} \right) \left[ \frac{1}{\sigma} (-IF(x_{M0}, x_0, \alpha, P_\theta) - x_M IF(x_{M0}, x_0, \beta, P_\theta)) \right. \\ \left. - \frac{x - \alpha - \beta x_M}{\sigma^2} IF(x_{M0}, x_0, \sigma, P_\theta) \right] dP_\theta(x_M, x) + \phi \left( \frac{x_0 - \alpha - \beta x_{M0}}{\sigma} \right) = 0. \end{aligned}$$

After some calculations, we obtain the relation

$$-\frac{1}{\sigma(\gamma+1)^{3/2}} IF(x_{M0}, x_0, \alpha, P_\theta) - \frac{1}{\sigma(\gamma+1)^{3/2}} IF(x_{M0}, x_0, \beta, P_\theta) E(X_M) + \phi \left( \frac{x_0 - \alpha - \beta x_{M0}}{\sigma} \right) = 0. \quad (\text{A13})$$

Similarly, derivating with respect to  $\varepsilon$  Equations (A11) and (A12) and taking the derivatives in  $\varepsilon = 0$ , we get

$$\begin{aligned} -\frac{1}{\sigma(\gamma+1)^{3/2}} E(X_M) IF(x_{M0}, x_0, \alpha, P_\theta) - \frac{1}{\sigma(\gamma+1)^{3/2}} E(X_M^2) IF(x_{M0}, x_0, \beta, P_\theta) \\ + \phi \left( \frac{x_0 - \alpha - \beta x_{M0}}{\sigma} \right) x_{M0} = 0 \end{aligned} \quad (\text{A14})$$

and

$$-\frac{2}{\sigma(\gamma+1)^{5/2}} IF(x_{M0}, x_0, \sigma, P_\theta) + \chi \left( \frac{x_0 - \alpha - \beta x_{M0}}{\sigma} \right) = 0. \quad (\text{A15})$$

Solving the system formed with the Equations (A13)–(A15), we find the expressions for the influence functions.  $\square$

**Proof of Theorem 4.** In the following, we simply denote by  $X_{Mj}$  the vector  $(1, X_{Mj})^T$ . Then,

$$\begin{aligned} & (\hat{\alpha}, \hat{\beta})^T(\{(X_{Mj}, X_j) : j = 1, \dots, n\}) \\ &= \arg_{(\alpha, \beta)^T} \max_{(\alpha, \beta, \sigma)} \sum_{j=1}^n \sigma^{-\gamma/(\gamma+1)} \exp \left( -\frac{\gamma}{2} \left( \frac{X_j - X_{Mj}^T(\alpha, \beta)^T}{\sigma} \right)^2 \right). \end{aligned}$$

For any two-dimensional column vector  $v$ , we have

$$\begin{aligned} & (\hat{\alpha}, \hat{\beta})^T(\{(X_{Mj}, X_j + X_{Mj}^T v) : j = 1, \dots, n\}) \\ &= \arg_{(\alpha, \beta)^T} \max_{(\alpha, \beta, \sigma)} \sum_{j=1}^n \sigma^{-\gamma/(\gamma+1)} \exp \left( -\frac{\gamma}{2} \left( \frac{X_j + X_{Mj}^T v - X_{Mj}^T(\alpha, \beta)^T}{\sigma} \right)^2 \right) \\ &= \arg_{(\alpha, \beta)^T} \max_{(\alpha, \beta, \sigma)} \sum_{j=1}^n \sigma^{-\gamma/(\gamma+1)} \exp \left( -\frac{\gamma}{2} \left( \frac{X_j - X_{Mj}^T((\alpha, \beta)^T - v)}{\sigma} \right)^2 \right) \\ &= \arg_{((\alpha, \beta)^T - v)} \max_{((\alpha, \beta)^T - v)^T, \sigma} \sum_{j=1}^n \sigma^{-\gamma/(\gamma+1)} \exp \left( -\frac{\gamma}{2} \left( \frac{X_j - X_{Mj}^T((\alpha, \beta)^T - v)}{\sigma} \right)^2 \right) + v \\ &= (\hat{\alpha}, \hat{\beta})^T(\{(X_{Mj}, X_j) : j = 1, \dots, n\}) + v, \end{aligned}$$

which show that  $(\hat{\alpha}, \hat{\beta})^T$  is regression equivariant.

For any constant  $c \neq 0$ , we have

$$\begin{aligned} & (\hat{\alpha}, \hat{\beta})^T(\{(X_{Mj}, cX_j) : j = 1, \dots, n\}) \\ &= \arg_{(\alpha, \beta)^T} \max_{(\alpha, \beta, \sigma)} \sum_{j=1}^n \sigma^{-\gamma/(\gamma+1)} \exp \left( -\frac{\gamma}{2} \left( \frac{cX_j - X_{Mj}^T(\alpha, \beta)^T}{\sigma} \right)^2 \right) \\ &= \arg_{(\alpha, \beta)^T} \max_{(\alpha, \beta, \sigma)} \sum_{j=1}^n c^{-\gamma/(\gamma+1)} (\sigma/c)^{-\gamma/(\gamma+1)} \exp \left( -\frac{\gamma}{2} \left( \frac{X_j - X_{Mj}^T((\alpha, \beta)^T/c)}{(\sigma/c)} \right)^2 \right) \\ &= c \cdot \arg_{(\alpha/c, \beta/c)^T} \max_{(\alpha/c, \beta/c, \sigma/c)} \sum_{j=1}^n c^{-\gamma/(\gamma+1)} (\sigma/c)^{-\gamma/(\gamma+1)} \exp \left( -\frac{\gamma}{2} \left( \frac{X_j - X_{Mj}^T((\alpha, \beta)^T/c)}{(\sigma/c)} \right)^2 \right) \\ &= c \cdot (\hat{\alpha}, \hat{\beta})^T(\{(X_{Mj}, X_j) : j = 1, \dots, n\}). \end{aligned}$$

This implies that the estimator  $(\hat{\alpha}, \hat{\beta}) = (\hat{\alpha}, \hat{\beta})(\{(X_{Mj}, X_j) : j = 1, \dots, n\})$  is scale equivariant. Now, for any two-dimensional square matrix  $A$ , we get

$$\begin{aligned} & (\hat{\alpha}, \hat{\beta})^T(\{(A^T X_{Mj}, X_j) : j = 1, \dots, n\}) \\ &= \arg_{(\alpha, \beta)^T} \max_{(\alpha, \beta, \sigma)} \sum_{j=1}^n \sigma^{-\gamma/(\gamma+1)} \exp \left( -\frac{\gamma}{2} \left( \frac{X_j - X_{Mj}^T A(\alpha, \beta)^T}{\sigma} \right)^2 \right) \\ &= A^{-1} \arg_{A(\alpha, \beta)^T} \max_{((\alpha, \beta) A^T, \sigma)} \sum_{j=1}^n \sigma^{-\gamma/(\gamma+1)} \exp \left( -\frac{\gamma}{2} \left( \frac{X_j - X_{Mj}^T(A(\alpha, \beta)^T)}{\sigma} \right)^2 \right) \\ &= A^{-1} \cdot (\hat{\alpha}, \hat{\beta})^T(\{(X_{Mj}, X_j) : j = 1, \dots, n\}), \end{aligned}$$

which show the affine equivariance of the estimator  $(\hat{\alpha}, \hat{\beta}) = (\hat{\alpha}, \hat{\beta})(\{(X_{Mj}, X_j) : j = 1, \dots, n\})$ .  $\square$

## Appendix B. The 50 Stocks and Their Abbreviations

1. Asbury Automotive Group, Inc. (ABG)
2. Arctic Cat Inc. (ACAT)
3. American Eagle Outfitters, Inc. (AEO)
4. AK Steel Holding Corporation (AKS)
5. Albany Molecular Research, Inc. (AMRI)
6. The Andersons, Inc. (ANDE)
7. ARMOUR Residential REIT, Inc. (ARR)
8. BJ's Restaurants, Inc. (BJRI)
9. Brooks Automation, Inc. (BRKS)
10. Caleres, Inc. (CAL)
11. Cincinnati Bell Inc. (CBB)
12. Calgon Carbon Corporation (CCC)
13. Coeur Mining, Inc. (CDE)
14. Cohen & Steers, Inc. (CNS)
15. Cray Inc. (CRAY)
16. Cirrus Logic, Inc. (CRUS)
17. Covenant Transportation Group, Inc. (CVTI)
18. EarthLink Holdings Corp. (ELNK)
19. Gray Television, Inc. (GTN)
20. Triple-S Management Corporation (GTS)
21. Getty Realty Corp. (GYT)
22. Hecla Mining Company (HL)
23. Harmonic Inc. (HLIT)
24. Ligand Pharmaceuticals Incorporated (LGND)
25. Louisiana-Pacific Corporation (LPX)
26. Lattice Semiconductor Corporation (LSCC)
27. ManTech International Corporation (MANT)
28. MiMedx Group, Inc. (MDXG)
29. Medifast, Inc. (MED)
30. Mentor Graphics Corporation (MENT)
31. Mistras Group, Inc. (MG)
32. Mesa Laboratories, Inc. (MLAB)
33. Meritor, Inc. (MTOR)
34. Monster Worldwide, Inc. (MWW)
35. Nektar Therapeutics (NKTR)
36. Osiris Therapeutics, Inc. (OSIR)
37. PennyMac Mortgage Investment Trust (PMT)
38. Paratek Pharmaceuticals, Inc. (PRTK)
39. Repligen Corporation (RGEN)
40. Rigel Pharmaceuticals, Inc. (RIGL)
41. Schnitzer Steel Industries, Inc. (SCHN)
42. comScore, Inc. (SCOR)
43. Safeguard Scientifics, Inc. (SFE)
44. Silicon Graphics International (SGI)
45. Sagent Pharmaceuticals, Inc. (SGNT)
46. Semtech Corporation (SMTC)
47. Sapiens International Corporation N.V. (SPNS)
48. Sarepta Therapeutics, Inc. (SRPT)
49. Take-Two Interactive Software, Inc. (TTWO)
50. Park Sterling Corporation (PSTB)

## References

- Sharpe, W.F. A simplified model to portfolio analysis. *Manag. Sci.* **1963**, *9*, 277–293. [[CrossRef](#)]
- Alexander, G.J.; Sharpe, W.F.; Bailey, J.V. *Fundamentals of Investments*; Prentice-Hall: Upper Saddle River, NJ, USA, 2000.
- Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman & Hall: Boca Raton, FL, USA, 2006.
- Basu, A.; Shioya, H.; Park, C. *Statistical Inference: the Minimum Pseudodistance Approach*; CRC Press: Boca Raton, FL, USA, 2011.
- Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimizing a density power divergence. *Biometrika* **1998**, *85*, 549–559. [[CrossRef](#)]
- Jones, M.C.; Hjort, N.L.; Harris, I.R.; Basu, A. A comparison of related density-based minimum divergence estimators. *Biometrika* **2001**, *88*, 865–873. [[CrossRef](#)]
- Broniatowski, M.; Keziou, A. Parametric estimation and tests through divergences and the duality technique. *J. Multivar. Anal.* **2009**, *100*, 16–36. [[CrossRef](#)]
- Toma, A.; Leoni-Aubin, S. Robust tests based on dual divergence estimators and saddlepoint approximations. *J. Multivar. Anal.* **2010**, *101*, 1143–1155. [[CrossRef](#)]
- Toma, A.; Broniatowski, M. Dual divergence estimators and tests: Robustness results. *J. Multivar. Anal.* **2011**, *102*, 20–36. [[CrossRef](#)]
- Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081. [[CrossRef](#)]
- Broniatowski, M.; Vajda, I. Several applications of divergence criteria in continuous families. *Kybernetika* **2012**, *48*, 600–636.
- Broniatowski, M.; Toma, A.; Vajda, I. Decomposable pseudodistances and applications in statistical estimation. *J. Stat. Plan. Inference* **2012**, *142*, 2574–2585. [[CrossRef](#)]
- Markowitz, H.M. Mean-variance analysis in portfolio choice and capital markets. *J. Finance* **1952**, *7*, 77–91.
- Fabozzi, F.J.; Huang, D.; Zhou, G. Robust portfolios: contributions from operations research and finance. *Ann. Oper. Res.* **2010**, *176*, 191–220. [[CrossRef](#)]
- Vaz-de Melo, B.; Camara, R.P. Robust multivariate modeling in finance. *Int. J. Manag. Finance* **2005**, *4*, 12–23. [[CrossRef](#)]
- Perret-Gentil, C.; Victoria-Feser, M.P. *Robust Mean-Variance Portfolio Selection*; FAME Research Paper, No. 140; 2005. Available online: [papers.ssrn.com/sol3/papers.cfm?abstract\\_id=721509](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=721509) (accessed on 28 February 2018).
- Welsch, R.E.; Zhou, X. Application of robust statistics to asset allocation models. *Revstat. Stat. J.* **2007**, *5*, 97–114.
- DeMiguel, V.; Nogales, F.J. Portfolio selection with robust estimation. *Oper. Res.* **2009**, *57*, 560–577. [[CrossRef](#)]
- Toma, A.; Leoni-Aubin, S. Robust portfolio optimization using pseudodistances. *PLoS ONE* **2015**, *10*, 1–26. [[CrossRef](#)] [[PubMed](#)]
- Toma, A.; Leoni-Aubin, S. Optimal robust M-estimators using Renyi pseudodistances. *J. Multivar. Anal.* **2013**, *115*, 359–373. [[CrossRef](#)]
- Van der Vaart, A. *Asymptotic Statistics*; Cambridge University Press: New York, NY, USA, 1998.
- Rousseeuw, P.J.; Leroy, A.M. *Robust Regression and Outlier Detection*; John Wiley & Sons: Hoboken, NJ, USA, 2005.
- Andersen, R. *Modern Methods for Robust Regression*; SAGE Publications, Inc.: Los Angeles, CA, USA, 2008.
- Rousseeuw, P.J.; Yohai, V. Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis*; Franke, J., Hardle, W., Martin, D., Eds.; Springer: New York, NY, USA, 1984; pp. 256–272, ISBN 978-0-387-96102-6.
- Ghosh, A.; Basu, A. Robust estimations for independent, non-homogeneous observations using density power divergence with applications to linear regression. *Electron. J. Stat.* **2013**, *7*, 2420–2456. [[CrossRef](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Robust Inference after Random Projections via Hellinger Distance for Location-Scale Family

Lei Li <sup>1</sup>, Anand N. Vidyashankar <sup>1,\*</sup>, Guoqing Diao <sup>1</sup> and Ejaz Ahmed <sup>2</sup>

<sup>1</sup> Department of Statistics, George Mason University, Fairfax, VA 22030, USA; lli14@masonlive.gmu.edu (L.L.); gdiao@gmu.edu (G.D.)

<sup>2</sup> Department of Mathematics and Statistics, Brock University, St. Catharines, ON L2S 3A1, Canada; sahmed5@brocku.ca

\* Correspondence: avidyash@gmu.edu

Received: 6 February 2019; Accepted: 24 March 2019; Published: 29 March 2019



**Abstract:** Big data and streaming data are encountered in a variety of contemporary applications in business and industry. In such cases, it is common to use random projections to reduce the dimension of the data yielding compressed data. These data however possess various anomalies such as heterogeneity, outliers, and round-off errors which are hard to detect due to volume and processing challenges. This paper describes a new robust and efficient methodology, using Hellinger distance, to analyze the compressed data. Using large sample methods and numerical experiments, it is demonstrated that a routine use of robust estimation procedure is feasible. The role of double limits in understanding the efficiency and robustness is brought out, which is of independent interest.

**Keywords:** compressed data; Hellinger distance; representation formula; iterated limits; influence function; consistency; asymptotic normality; location-scale family

## 1. Introduction

Streaming data are commonly encountered in several business and industrial applications leading to the so-called Big Data. These are commonly characterized using four V's: velocity, volume, variety, and veracity. Velocity refers to the speed of data processing while volume refers to the amount of data. Variety refers to various types of data while veracity refers to uncertainty and imprecision in data. It is believed that veracity is due to data inconsistencies, incompleteness, and approximations. Whatever be the real cause, it is hard to identify and pre-process data for veracity in a big data setting. The issues are even more complicated when the data are streaming.

A consequence of the data veracity is that statistical assumptions used for analytics tend to be inaccurate. Specifically, considerations such as model misspecification, statistical efficiency, robustness, and uncertainty assessment—which are standard part of a statistical toolkit—cannot be routinely carried out due to storage limitations. Statistical methods that facilitate simultaneous addressal of twin problems of volume and veracity would enhance the value of the big data. While health care industry and financial industries would be the prime benefactors of this technology, the methods can be routinely applied in a variety of problems that use big data for decision making.

We consider a collection of  $n$  ( $n$  is of the order of at least  $10^6$ ) observations, assumed to be independent and identically distributed (i.i.d.), from a probability distribution  $f(\cdot)$  belonging to a location-scale family; that is,

$$f(x; \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right), \quad \mu \in \mathbb{R}, \sigma > 0.$$

We denote by  $\Theta$  the parameter space and without loss of generality take it as compact since otherwise it can be re-parametrized in a such a way that the resulting parameter space is compact (see [1]).

The purpose of this paper is to describe a methodology for joint robust and efficient estimation of  $\mu$  and  $\sigma^2$  that takes into account (i) storage issues, (ii) potential model misspecifications, and (iii) presence of aberrant outliers. These issues—which are more likely to occur when dealing with massive amounts of data—if not appropriately accounted in the methodological development, can lead to inaccurate inference and misleading conclusions. On the other hand, incorporating them in the existing methodology may not be feasible due to a computational burden.

Hellinger distance-based methods have long been used to handle the dual issue of robustness and statistical efficiency. Since the work of [1,2] statistical methods that invoke alternative objective functions which converge to the objective function under the posited model have been developed and the methods have been shown to possess efficiency and robustness. However, their routine use in the context of big data problems is not feasible due to the complexity in the computations and other statistical challenges. Recently, a class of algorithms—referred to as *Divide and Conquer*—have been developed to address some of these issues in the context of likelihood. These algorithms consist in distributing the data across multiple processors and, in the context of the problem under consideration, estimating the parameters from each processor separately and then combining them to obtain an overall estimate. The algorithm assumes availability of several processors, *with substantial processing power*, to solve the complex problem at hand. Since robust procedures involve complex iterative computations—invoking the increased demand for several high-speed processors and enhanced memory—routine use of available analytical methods in a big data setting is challenging. Maximum likelihood method of estimation in the context of location-scale family of distributions has received much attention in the literature ([3–7]). It is well-known that the maximum likelihood estimators (MLE) of location-scale families may not exist unless the defining function  $f(\cdot)$  satisfies certain regularity conditions. Hence, it is natural to ask if other methods of estimation such as minimum Hellinger distance estimator (MHDE) under weaker regularity conditions. This manuscript provides a first step towards addressing this question. Random projections and sparse random projections are being increasingly used to “compress data” and then use the resulting compressed data for inference. The methodology, primarily developed by computer scientists, is increasingly gaining attention among the statistical community and is investigated in a variety of recent work ([8–12]). In this manuscript, we describe a Hellinger distance-based methodology for robust and efficient estimation after the use of random projections for compressing i.i.d data belonging to the location-scale family. The proposed method consists in reducing the dimension of the data to facilitate the ease of computations and simultaneously maintain robustness and efficiency when the posited model is correct. While primarily developed to handle big and streaming data, the approach can also be used to handle privacy issues in a variety of applications [13].

The rest of the paper is organized as follows: Section 2 provides background on minimum Hellinger distance estimation; Section 3 is concerned with the development of Hellinger distance-based methods for compressed data obtained after using random projections; additionally, it contains the main results and their proofs. Section 4 contains results of the numerical experiments and also describes an algorithm for implementation of the proposed methods. Section 5 contains a real data example from financial analytics. Section 6 is concerned with discussions and extensions. Section 7 contains some concluding remarks.

## 2. Background on Minimum Hellinger Distance Estimation

Ref. [1] proposed minimum Hellinger distance (MHD) estimation for i.i.d. observations and established that MHD estimators (MHDE) are simultaneously robust and first-order efficient under the true model. Other researchers have investigated related estimators, for example, [14–20]. These authors establish that when the model is correct, the MHDE is asymptotically equivalent to the

maximum likelihood estimator (MLE) in a variety of independent and dependent data settings. For a comprehensive discussion of minimum divergence theory see [21].

We begin by recalling that the Hellinger distance between two probability densities is the  $L^2$  distance between the square root of the densities. Specifically, let, for  $p \geq 1$ ,  $\|\cdot\|_p$  denote the  $L^p$  norm defined by

$$\|h\|_p = \left\{ \int |h|^p \right\}^{1/p}.$$

The Hellinger distance between the densities  $f(\cdot)$  and  $g(\cdot)$  is given by

$$H^2(f(\cdot), g(\cdot)) = \|f^{1/2}(\cdot) - g^{1/2}(\cdot)\|_2^2.$$

Let  $f(\cdot|\theta)$  denote the density of  $\mathbb{R}^d$  valued independent and identically distributed random variables  $X_1, \dots, X_n$ , where  $\theta \in \Theta \subset \mathbb{R}^p$ ; let  $g_n(\cdot)$  be a nonparametric density estimate (typically a kernel density estimator). The Hellinger distance between  $f(\cdot|\theta)$  and  $g_n(\cdot)$  is then

$$H^2(f(\cdot|\theta), g_n(\cdot)) = \|f^{1/2}(\cdot|\theta) - g_n^{1/2}(\cdot)\|_2^2.$$

The MHDE is a mapping  $T(\cdot)$  from the set of all densities to  $\mathbb{R}^p$  defined as follows:

$$\theta_g = T(g) = \underset{\theta \in \Theta}{\operatorname{argmin}} H^2(f(\cdot|\theta), g(\cdot)). \quad (1)$$

Please note that the above minimization problem is equivalent to maximizing  $\mathcal{A}(f(\cdot|\theta), g(\cdot)) = \int f^{1/2}(x|\theta)g^{1/2}(x)dx$ . Hence MHDE can alternatively be defined as

$$\theta_g = \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{A}(f(\cdot|\theta), g(\cdot)).$$

To study the robustness of MHDE, ref. [1] showed that to assess the robustness of a functional with respect to the gross-error model it is necessary to examine the  $\alpha$ -influence curve rather than the influence curve, except when the influence curve provides a uniform approximation to the  $\alpha$ -influence curve. Specifically, the  $\alpha$ -influence function ( $\text{IF}_\alpha(\theta, z)$ ) is defined as follows: for  $\theta \in \Theta$ , let  $f_{\alpha,\theta,z} = (1 - \alpha)f(\cdot|\theta) + \alpha\eta_z$ , where  $\eta_z$  denotes the uniform density on the interval  $(z - \epsilon, z + \epsilon)$ , where  $\epsilon > 0$  is small,  $\alpha \in (0, 1)$ ,  $z \in \mathbb{R}$ ; the  $\alpha$ -influence function is then defined to be

$$\text{IF}_\alpha(\theta, z) = \frac{T(f_{\alpha,\theta,z}) - \theta}{\alpha}, \quad (2)$$

where  $T(f_{\alpha,\theta,z})$  is the functional for the model with density  $f_{\alpha,\theta,z}(\cdot)$ . Equation (2) represents a complete description of the behavior of the estimator in the presence of contamination, up to the shape of the contaminating density. If  $\text{IF}_\alpha(\theta, z)$  is a bounded function of  $z$  such that  $\lim_{z \rightarrow \infty} \text{IF}_\alpha(\theta, z) = 0$ , for every  $\theta \in \Theta$ , then the functional  $T$  is robust at  $f(\cdot|\theta)$  against 100%  $\alpha$  contamination by gross errors at arbitrary large value  $z$ . The influence function can be obtained by letting  $\alpha \rightarrow 0$ . Under standard regularity conditions, the minimum divergence estimators (MDE) are first order efficient and have the same influence function as the MLE under the model, which is often unbounded. Hence the robustness of these estimators cannot be explained through their influence functions. In contrast, the  $\alpha$ -influence function of the estimators are often bounded, continuous functions of the contaminating point. Finally, this approach often leads to high breakdown points in parametric estimation. Other explanations can also be found in [22,23].

Ref. [1] showed that the MHDE of location has a breakdown point equal to 50%. Roughly speaking, the breakdown point is the smallest fraction of data that, when strategically placed, can cause an estimator to take arbitrary values. Ref. [24] obtained breakdown results for MHDE of multivariate

location and covariance. They showed that the affine-invariant MHDE for multivariate location and covariance has a breakdown point of at least 25%. Ref. [18] showed that the MHDE has 50% breakdown in some discrete models.

### 3. Hellinger Distance Methodology for Compressed Data

In this section we describe the Hellinger distance-based methodology as applied to the compressed data. Since we are seeking to model the streaming independent and identically distributed data, we denote by  $J$  the number of observations in a fixed time-interval (for instance, every ten minutes, or every half-hour, or every three hours). Let  $B$  denote the total number of time intervals. Alternatively,  $B$  could also represent the number of sources from which the data are collected. Then, the incoming data can be expressed as  $\{X_{jl}, 1 \leq j \leq J; 1 \leq l \leq B\}$ . Throughout this paper, we assume that the density of  $X_{jl}$  belongs to a location-scale family and is given by  $f(x; \theta^*) = \frac{1}{\sigma^*} f(\frac{x - \mu^*}{\sigma^*})$ , where  $\theta^* = (\mu^*, \sigma^*)$ . A typical example is a data store receiving data from multiple sources, for instance financial or healthcare organizations, where information from multiple sources across several hours are used to monitor events of interest such as cumulative usage of certain financial instruments or drugs.

#### 3.1. Random Projections

Let  $R_l = (r_{ijl})$  be a  $S \times J$  matrix, where  $S$  is the number of compressed observations in each time interval,  $S \ll J$ , and  $r_{ijl}$ 's are independent and identically distributed random variables and assumed to be independent of  $\{X_{jl}, j = 1, 2, \dots, J; 1 \leq l \leq B\}$ . Let

$$\tilde{Y}_{il} = \sum_{j=1}^J r_{ijl} X_{jl}$$

and set  $\tilde{Y}_l = (\tilde{Y}_{l1}, \dots, \tilde{Y}_{lS})'$ ; in matrix form this can be expressed as  $\tilde{Y}_l = R_l X_l$ . The matrix  $R_l$  is referred to as the *sensing matrix* and  $\{\tilde{Y}_{il}, i = 1, 2, \dots, S; l = 1, 2, \dots, B\}$  is referred to as the *compressed data*. The total number of compressed observations  $m = SB$  is much smaller than the number of original observations  $n = JB$ . We notice here that  $R_l$ 's are independent and identically distributed random matrices of order  $S \times J$ . Referring to each time interval or a source as a group, the following Table 1 is a tabular representation of the compressed data.

**Table 1.** Illustration of Data Reduction Mechanism, Here  $r_{il}^* = (r_{il}, \omega_{il})$ .

	Grp 1	Grp 2	...	Grp B		Grp 1	Grp 2	...	Grp B
Original Data	$X_{11}$	$X_{12}$	...	$X_{1B}$	Compressed Data	$(\tilde{Y}_{11}, r_{11}^*)$	$(\tilde{Y}_{12}, r_{12}^*)$	...	$(\tilde{Y}_{1B}, r_{1B}^*)$
	$X_{21}$	$X_{22}$	...	$X_{2B}$		$(\tilde{Y}_{21}, r_{21}^*)$	$(\tilde{Y}_{22}, r_{22}^*)$	...	$(\tilde{Y}_{2B}, r_{2B}^*)$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\xrightarrow{S \ll J}$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$X_{J1}$	$X_{J2}$	...	$X_{JB}$		$(\tilde{Y}_{S1}, r_{S1}^*)$	$(\tilde{Y}_{S2}, r_{S2}^*)$	...	$(\tilde{Y}_{SB}, r_{SB}^*)$

In random projections literature, the distribution of  $r_{ijl}$  is typically taken to be Gaussian; but other distributions such as Rademacher distribution, exponential distribution and extreme value distributions are also used (for instance, see [25]). In this paper, we do not make any strong distributional assumptions on  $r_{ijl}$ . We only assume that  $E[r_{ijl}] = 1$  and  $Var[r_{ijl}] = \gamma_0^2$ , where  $E[\cdot]$  represents the expectation of the random variable and  $Var[\cdot]$  represents the variance of the random variable. Additionally, we denote the density of  $r_{ijl}$  by  $q(\cdot)$ .

We next return to the storage issue. When  $S = 1$  and  $r_{ijl} = 1$ ,  $\tilde{Y}_{il}$  is a sum of  $J$  random variables. In this case, one retains (stores) only the sum of  $J$  observations and robust estimates of  $\theta^*$  are sought using the sum of observations. In other situations, that is when  $r_{ijl}$  are not degenerate

at 1, the distribution of  $\tilde{Y}_{il}$  is complicated. Indeed, even if  $r_{ijl}$  are assumed to be normally distributed, the marginal distribution of  $\tilde{Y}_{il}$  is complicated. The conditional distribution is  $\tilde{Y}_{il}$  (given  $r_{ijl}$ ) is a weighted sum of location scale distributions and does not have a useful closed form expression. Hence, in general for these problems the MLE method is *not feasible*. We denote by  $\omega_{il}^2 = \sum_{j=1}^J r_{ijl}^2$  and work with the random variables  $Y_{il} \equiv \omega_{il}^{-1} \tilde{Y}_{il}$ . We denote the true density of  $Y_{il}$  to be  $h_J(\cdot|\theta^*, \gamma_0)$ . Also, when  $\gamma_0 = 0$  (which implies  $r_{ijl} \equiv 1$ ) we denote the true density of  $Y_{il}$  by  $h^{*J}(\cdot|\theta^*)$  to emphasize that the true density is a convolution of  $J$  independent and identically distributed random variables.

### 3.2. Hellinger Distance Method for Compressed Data

In this section, we describe the Hellinger distance-based method for estimating the parameters of the location scale family using the compressed data. As described in the last section, let  $\{X_{jl}, j = 1, 2, \dots, J; l = 1, 2, \dots, B\}$  be a doubly indexed collection of independent and identically distributed random variables with true density  $\frac{1}{\sigma^*} f\left(\frac{\cdot - \mu^*}{\sigma^*}\right)$ . Our goal is to estimate  $\theta^* = (\mu^*, \sigma^{2*})$  using the compressed data  $\{Y_{il}, i = 1, 2, \dots, S; l = 1, 2, \dots, B\}$ . We re-emphasize here that the density of  $Y_{il}$  depends additionally on  $\gamma_0$ , the variance of the sensing random variables  $r_{ijl}$ .

To formulate the Hellinger-distance estimation method, let  $\mathcal{G}$  be a class of densities metrized by the  $L_1$  distance. Let  $\{h_J(\cdot|\theta, \gamma_0); \theta \in \Theta\}$  be a parametric family of densities. The Hellinger distance functional  $T$  is a measurable mapping mapping from  $\mathcal{G}$  to  $\Theta$ , defined as follows:

$$\begin{aligned} T(g) &\equiv \arg \min_{\theta} \int_{\mathbb{R}} \left( g^{\frac{1}{2}}(y) - h_J^{\frac{1}{2}}(y|\theta, \gamma_0) \right)^2 dy \\ &= \arg \min_{\theta} HD^2(g, h_J(\cdot|\theta, \gamma_0)) = \theta_g^*(\gamma_0). \end{aligned}$$

When  $g(\cdot) = h_J(\cdot|\theta^*, \gamma_0)$ , then under additional assumptions  $\theta_g^*(\gamma_0) = \theta^*(\gamma_0)$ . Since minimizing the Hellinger-distance is equivalent to maximizing the affinity, it follows that

$$T(g) = \arg \max_{\theta} \mathcal{A}(g, h_J(\cdot|\theta, \gamma_0)), \text{ where}$$

$$\mathcal{A}(g, h_J(\cdot|\theta, \gamma_0)) \equiv \int_{\mathbb{R}} g^{\frac{1}{2}}(y) h_J^{\frac{1}{2}}(y|\theta, \gamma_0) dy.$$

It is worth noticing here that

$$\mathcal{A}(g, h_J(\cdot|\theta, \gamma_0)) = 1 - \frac{1}{2} HD^2(g, h_J(\cdot|\theta, \gamma_0)). \quad (3)$$

To obtain the Hellinger distance estimator of the true unknown parameters  $\theta^*$ , expectedly we choose the parametric family  $h_J(\cdot|\theta, \gamma_0)$  to be density of  $Y_{il}$  and  $g(\cdot)$  to be a non-parametric  $L_1$  consistent estimator  $g_B(\cdot)$  of  $h_J(\cdot|\theta, \gamma_0)$ . Thus, the MHDE of  $\theta_B^*$  is given by

$$\hat{\theta}_B(\gamma_0) = \arg \max_{\theta} \mathcal{A}(g_B, h_J(\cdot|\theta, \gamma_0)) = T(g_B).$$

In the notation above, we emphasize the dependence of the estimator on the variance of the projecting random variables. We notice here that the solution to (1) may not be unique. In such cases, we choose one of the solutions in a measurable manner.

The choice of the density estimate, typically employed in the literature is the kernel density estimate. However, in the setting of the compressed data investigated here, there are  $S$  observations per group. These  $S$  observations are, conditioned on  $r_{ijl}$  independent; however they are marginally

dependent (if  $S > 1$ ). In the case when  $S > 1$ , we propose the following formula for  $g_B(\cdot)$ . First, we consider the estimator

$$g_B^{(i)}(y) = \frac{1}{Bc_B} \sum_{l=1}^B K\left(\frac{y - Y_{il}}{c_B}\right), \quad i = 1, 2, \dots, S.$$

With this choice, the MHDE of  $\theta_B^*$  is given by, for  $1 \leq i \leq S$ ,

$$\hat{\theta}_{i,B}(\gamma_0) = \arg \max_{\theta} \mathcal{A}\left(g_B^{(i)}, h_J(\cdot|\theta, \gamma_0)\right). \quad (4)$$

The above estimate of the density chooses  $i^{th}$  observation from each group and obtains the kernel density estimator using the  $B$  independent and identically distributed compressed observations. This is one choice for the estimator. Of course, alternatively, one could obtain  $S^B$  different estimators by choosing different combinations of observations from each group.

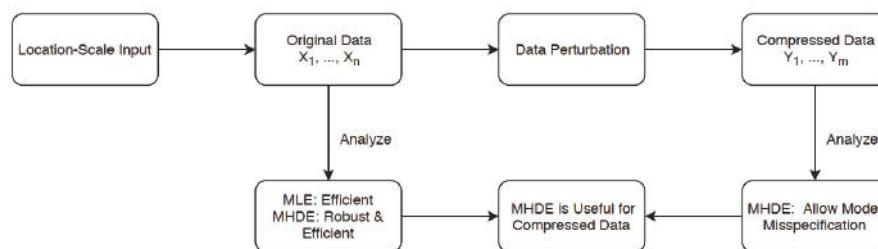
It is well-known that the estimator is almost surely  $L_1$  consistent for  $h_J(\cdot|\theta^*, \gamma_0)$  as long as  $c_B \rightarrow 0$  and  $Bc_B \rightarrow \infty$  as  $B \rightarrow \infty$ . Hence, under additional regularity and identifiability conditions and further conditions on the bandwidth  $c_B$ , existence, uniqueness, consistency and asymptotic normality of  $\hat{\theta}_{i,B}(\gamma_0)$ , for fixed  $\gamma_0$ , follows from the existing results in the literature.

When  $\gamma_0 = 0$  and  $r_{ijl} \equiv 1$ , as explained previously, the true density is a  $J$ -fold convolution of  $f(\cdot|\theta^*)$ , it is natural to ask the following question: if one lets  $\gamma_0 \rightarrow 0$ , will the asymptotic results converge to what one would obtain by taking  $\gamma_0 = 0$ . We refer to this property as a *continuity property* in  $\gamma_0$  of the procedure. Furthermore, it is natural to wonder if these asymptotic properties can be established uniformly in  $\gamma_0$ . If that is the case, then one can also allow  $\gamma_0$  to depend on  $B$ . This idea has an intuitive appeal since one can choose the parameters of the sensing random variables to achieve an optimal inferential scheme. We address some of these issues in the next subsection.

Finally, we emphasize here that while we do not require  $S > 1$ , in applications involving streaming data and privacy problems  $S$  tends to greater than one. In problems where the variance of sensing variables are large, one can obtain an overall estimator by averaging  $\hat{\theta}_{i,B}(\gamma_0)$  over various choices of  $1 \leq i \leq S$ ; that is,

$$\theta_B(\gamma_0) = \frac{1}{S} \sum_{i=1}^S \hat{\theta}_{i,B}(\gamma_0). \quad (5)$$

The averaging improves the accuracy of the estimator in small compressed samples (data not presented). For this reason, we provide results for this general case, even though our simulation and theoretical results demonstrate that for some problems considered in this paper,  $S$  can be taken to be one. We now turn to our main results which are presented in the next subsection. The following Figure 1 provides a overview of our work.



**Figure 1.** MLE vs. MHDE after Data Compression.

### 3.3. Main Results

In this section we state our main results concerning the asymptotic properties of the MHDE of compressed data  $Y_{il}$ . We emphasize here that we only store  $\{(\tilde{Y}_{il}, r_{il}, \omega_{il}^2) : i = 1, 2, \dots, S; l = 1, 2, \dots, B\}$ . Specifically, we establish the continuity property in  $\gamma_0$  of the proposed methods by establishing the existence of the iterated limits. This provides a first step in establishing the double limit. The first proposition is well-known and is concerned with the existence and uniqueness of MHDE for the location-scale family defined in (4) using compressed data.

**Proposition 1.** Assume that  $h_J(\cdot|\theta, \gamma_0)$  is a continuous density function. Assume further that if  $\theta_1 \neq \theta_2$ . Then for every  $\gamma_0 \geq 0$ ,  $h_J(y|\theta_1, \gamma_0) \neq h_J(y|\theta_2, \gamma_0)$  on a set of positive Lebesgue measure, the MHDE in (4) exists and is unique.

**Proof.** The proof follows from Theorem 2.2 of [20] since, without loss of generality,  $\Theta$  is taken to be compact and the density function  $h_J(\cdot|\theta, \gamma_0)$  is continuous in  $\theta$ .  $\square$

**Consistency:** We next turn our attention to consistency. As explained previously, under regularity conditions for each fixed  $\gamma_0$ , the MHDE  $\hat{\theta}_{i,B}(\gamma_0)$  is consistent for  $\theta^*(\gamma_0)$ . The next result says that under additional conditions, the consistency property of MHDE is continuous in  $\gamma_0$ .

**Proposition 2.** Let  $h_J(\cdot|\theta, \gamma_0)$  be a continuous probability density function satisfying the conditions of Proposition 1. Assume that

$$\lim_{\gamma_0 \rightarrow 0} \sup_{\theta \in \Theta} \int_{\mathbb{R}} |h_J(y|\theta, \gamma_0) - h^{*J}(y|\theta)| dy = 0. \quad (6)$$

Then, with probability one (wp1) the iterated limits also exist and equals  $\theta^*$ ; that is, for  $1 \leq i \leq S$ ,

$$\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} \hat{\theta}_{i,B}(\gamma_0) = \lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} \hat{\theta}_{i,B}(\gamma_0) = \theta^*.$$

**Proof.** Without loss of generality let  $\Theta$  be compact since otherwise it can be embedded into a compact set as described in [1]. Since  $f(\cdot)$  is continuous in  $\theta$  and  $g(\cdot)$  is continuous in  $\gamma_0$ , it follows that  $h_J(\cdot|\theta, \gamma_0)$  is continuous in  $\theta$  and  $\gamma_0$ . Hence by Theorem 1 of [1] for every fixed  $\gamma_0 \geq 0$  and  $1 \leq i \leq S$ ,

$$\lim_{B \rightarrow \infty} \hat{\theta}_{i,B}(\gamma_0) = \theta^*(\gamma_0).$$

Thus, to verify the convergence of  $\theta^*(\gamma_0)$  to  $\theta^*$  as  $\gamma_0 \rightarrow 0$ , we first establish, using (6), that

$$\lim_{\gamma_0 \rightarrow 0} \sup_{\theta \in \Theta} |\mathcal{A}(h_J(\cdot|\theta, \gamma_0), h^{*J}(\cdot|\theta)) - 1| = 0.$$

To this end, we first notice that

$$\sup_{\theta \in \Theta} HD^2(h_J(\cdot|\theta, \gamma_0), h^{*J}(\cdot|\theta)) \leq \sup_{\theta \in \Theta} \int_{\mathbb{R}} |(h_J(y|\theta, \gamma_0) - h^{*J}(y|\theta))| dy.$$

Hence, using (3),

$$\begin{aligned} \sup_{\theta \in \Theta} |\mathcal{A}(h_J(\cdot|\theta, \gamma_0), h^{*J}(\cdot|\theta)) - 1| &= \frac{1}{2} \sup_{\theta \in \Theta} HD^2(h_J(\cdot|\theta, \gamma_0), h^{*J}(\cdot|\theta)) \\ &\rightarrow 0 \text{ as } \gamma_0 \rightarrow 0. \end{aligned}$$

Hence,

$$\lim_{\gamma_0 \rightarrow 0} \mathcal{A}(h_J(\cdot|\theta^*(\gamma_0), \gamma_0), h^{*J}(\cdot|\theta^*(\gamma_0))) = 1.$$

Also, by continuity,

$$\lim_{\gamma_0 \rightarrow 0} \mathcal{A}(h^{*J}(\cdot|\theta^*(\gamma_0), \gamma_0), h^{*J}(\cdot|\theta^*)) = 1,$$

which, in turn implies that

$$\lim_{\gamma_0 \rightarrow 0} \mathcal{A}(h_J(\cdot|\theta^*(\gamma_0), \gamma_0), h^{*J}(\cdot|\theta^*)) = 1.$$

Thus existence of the iterated limit first as  $B \rightarrow \infty$  and then  $\gamma_0 \rightarrow 0$  follows using compactness of  $\Theta$  and the identifiability of the model. As for the other iterated limit, again notice that for each  $1 \leq i \leq S$ ,  $\mathcal{A}(g_B^{(i)}, h_J(\cdot|\theta, \gamma_0))$  converges to  $\mathcal{A}(g_B^{(i)}, h^{*J}(\cdot|\theta))$  with probability one as  $\gamma_0$  converges to 0. The result then follows again by an application of Theorem 1 of [20].  $\square$

**Remark 1.** Verification of condition (6) seems to be involved even in the case of standard Gaussian random variables and standard Gaussian sensing random variables. Indeed in this case, the density of  $h_J(\cdot|\theta, \gamma_0)$  is a  $J$ -fold convolution of a Bessel function of second kind. It may be possible to verify the condition (6) using the properties of these functions and compactness of the parameter space  $\Theta$ . However, if one is focused only on weak-consistency, it is an immediate consequence of Theorems 1 and 2 below and condition (6) is not required. Finally, it is worth mentioning here that the convergence in (6) without uniformity over  $\Theta$  is a consequence of convergence in probability of  $r_{ijl}$  to 1 and Glick's Theorem.

**Asymptotic limit distribution:** We now proceed to investigate the limit distribution of  $\theta_B^*(\gamma_0)$  as  $B \rightarrow \infty$  and  $\gamma_0 \rightarrow 0$ . It is well-known that for fixed  $\gamma_0 \geq 0$ , after centering and scaling,  $\theta_B^*(\gamma_0)$  has a limiting Gaussian distribution, under appropriate regularity conditions (see for instance [20]). However to evaluate the iterated limits as  $\gamma_0 \rightarrow 0$  and  $B \rightarrow \infty$ , additional refinements of the techniques in [20] are required. To this end, we start with additional notations. Let  $s_J(\cdot|\theta, \gamma_0) = h_J^{\frac{1}{2}}(\cdot|\theta, \gamma_0)$  and let the score function be denoted by  $u_J(\cdot|\theta, \gamma_0) \equiv \nabla \log h_J(\cdot|\theta, \gamma_0) = \left( \frac{\partial \log h_J(\cdot|\theta, \gamma_0)}{\partial \mu}, \frac{\partial \log h_J(\cdot|\theta, \gamma_0)}{\partial \sigma} \right)'$ . Also, the Fisher information  $I(\theta(\gamma_0))$  is given by

$$I(\theta(\gamma_0)) = \int_{\mathbb{R}} u_J(y|\theta, \gamma_0) u_J'(y|\theta, \gamma_0) h_J(y|\theta, \gamma_0) dy.$$

In addition, let  $\dot{s}_J(\cdot|\theta, \gamma_0)$  be the gradient of  $s_J(\cdot|\theta, \gamma_0)$  with respect to  $\theta$ , and  $\ddot{s}_J(\cdot|\theta, \gamma_0)$  is the second derivative matrix of  $s_J(\cdot|\theta, \gamma_0)$  with respect to  $\theta$ . In addition, let  $t_J(\cdot|\theta) = h^{*J\frac{1}{2}}(\cdot|\theta)$  and  $v_J(\cdot|\theta) = \nabla \log h^{*J}(\cdot|\theta)$ . Furthermore, let  $Y_{il}^*$  denote  $Y_{il}$  when  $\gamma_0 \equiv 0$ . Please note that in this case,  $Y_{il} = Y_{1l}$  for all  $i = 1, 2, \dots, S$ . The corresponding kernel density estimate of  $Y_{il}^*$  is given by

$$g_B^*(y) = \frac{1}{B c_B} \sum_{l=1}^B K \left( \frac{y - Y_{il}^*}{c_B} \right). \quad (7)$$

We emphasize here that we suppress  $i$  on the LHS of the above equation since  $g_B^{(i)*}(\cdot)$  are equal for all  $1 \leq i \leq S$ .

The iterated limit distribution involves additional regularity conditions which are stated in the Appendix. The first step towards this aim is a representation formula which expresses the quantity of interest, *viz.*,  $\sqrt{B} (\hat{\theta}_{i,B}(\gamma_0) - \theta^*(\gamma_0))$  as a sum of two terms, one involving sums of compressed i.i.d. random variables and the other involving remainder terms that converge to 0 at a specific rate.

This expression will appear in different guises in the rest of the manuscript and will play a critical role in the proofs.

### 3.4. Representation Formula

Before we state the lemma, we first provide two crucial assumptions that allow differentiating the objective function and interchanging the differentiation and integration:

#### **Model assumptions on $h_J(\cdot|\theta, \gamma_0)$**

- (D1)  $h_J(\cdot|\theta, \gamma_0)$  is twice continuously differentiable in  $\theta$ .
- (D2) Assume further that  $\|\nabla s_J(\cdot|\theta, \gamma_0)\|_2$  is continuous and bounded.

**Lemma 1.** *Assume that the conditions (D1) and (D2) hold. Then for every  $1 \leq i \leq S$  and  $\gamma_0 \geq 0$ , the following holds:*

$$B^{\frac{1}{2}} (\hat{\theta}_{i,B}(\gamma_0) - \theta^*(\gamma_0))' = A_{1B}(\gamma_0) + A_{2B}(\gamma_0), \quad \text{where} \quad (8)$$

$$A_{1B}(\gamma_0) = B^{\frac{1}{2}} D_B^{-1}(\tilde{\theta}_{i,B}(\gamma_0)) T_B(\gamma_0), \quad A_{2B}(\gamma_0) = B^{\frac{1}{2}} D_B^{-1}(\tilde{\theta}_{i,B}(\gamma_0)) R_B(\gamma_0), \quad (9)$$

$$\tilde{\theta}_{i,B}(\gamma_0) \in U_B(\theta'(\gamma_0)), \quad U_B(\theta'(\gamma_0)) = \{\theta' : \theta'(\gamma_0) = t\theta^*(\gamma_0) + (1-t)\hat{\theta}_{i,B}(\gamma_0), t \in [0, 1]\}, \quad (10)$$

$$\begin{aligned} D_B(\theta(\gamma_0)) &= -\frac{1}{2} \int_{\mathbb{R}} \dot{u}_J(y|\theta, \gamma_0) s_J(y|\theta, \gamma_0) g_B^{(i)\frac{1}{2}}(y) dy \\ &\quad -\frac{1}{4} \int_{\mathbb{R}} u_J(y|\theta, \gamma_0) u'_J(y|\theta, \gamma_0) s_J(y|\theta, \gamma_0) g_B^{(i)\frac{1}{2}}(y) dy \\ &\equiv D_{1B}(\theta(\gamma_0)) + D_{2B}(\theta(\gamma_0)), \end{aligned} \quad (11)$$

$$T_B(\gamma_0) \equiv \frac{1}{4} \int_{\mathbb{R}} u_J(y|\theta^*, \gamma_0) \left( h_J(y|\theta^*, \gamma_0) - g_B^{(i)}(y) \right) dy, \quad \text{and} \quad (12)$$

$$R_B(\gamma_0) \equiv \frac{1}{4} \int_{\mathbb{R}} u_J(y|\theta^*, \gamma_0) \left( h_J^{\frac{1}{2}}(y|\theta^*, \gamma_0) - g_B^{(i)\frac{1}{2}}(y) \right)^2 dy. \quad (13)$$

**Proof.** By algebra, note that  $s_J(y|\theta, \gamma_0) = \frac{1}{2} u_J(y|\theta, \gamma_0) s_J(y|\theta, \gamma_0)$ . Furthermore, the second partial derivative of  $s_J(\cdot|\theta, \gamma_0)$  is given by  $\ddot{s}_J(y|\theta, \gamma_0) = \frac{1}{2} \dot{u}_J(y|\theta, \gamma_0) s_J(y|\theta, \gamma_0) + \frac{1}{4} u_J(y|\theta, \gamma_0) u'_J(y|\theta, \gamma_0) s_J(y|\theta, \gamma_0)$ . Now using (D1) and (D2) and partially differentiating  $HD_B^2(\theta(\gamma_0)) \equiv HD^2(g_B^{(i)}(\cdot), h_J(\cdot|\theta, \gamma_0))$  with respect to  $\theta$  and setting it equal to 0, the estimating equations for  $\theta^*(\gamma_0)$  is

$$\nabla HD_B^2(\theta^*(\gamma_0)) = 0. \quad (14)$$

Let  $\hat{\theta}_{i,B}(\gamma_0)$  be the solution to (14). Now applying first order Taylor expansion of (14) we get

$$\nabla HD_B^2(\theta^*(\gamma_0)) = \nabla HD_B^2(\hat{\theta}_{i,B}(\gamma_0)) + D_B(\tilde{\theta}_{i,B}(\gamma_0)) (\hat{\theta}_{i,B}(\gamma_0) - \theta^*(\gamma_0)),$$

where  $\tilde{\theta}_{i,B}(\gamma_0)$  is defined in (10), and  $D_B(\cdot)$  is given by

$$\begin{aligned} D_B(\boldsymbol{\theta}(\gamma_0)) &= -\frac{1}{2} \int_{\mathbb{R}} \dot{u}_J(y|\boldsymbol{\theta}, \gamma_0) s_J(y|\boldsymbol{\theta}, \gamma_0) g_B^{(i)\frac{1}{2}}(y) dy \\ &\quad -\frac{1}{4} \int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}, \gamma_0) u'_J(y|\boldsymbol{\theta}, \gamma_0) s_J(y|\boldsymbol{\theta}, \gamma_0) g_B^{(i)\frac{1}{2}}(y) dy \\ &\equiv D_{1B}(\boldsymbol{\theta}(\gamma_0)) + D_{2B}(\boldsymbol{\theta}(\gamma_0)), \end{aligned}$$

and  $\nabla HD_B^2(\cdot)$  is given by

$$\nabla HD_B^2(\boldsymbol{\theta}(\gamma_0)) = -\frac{1}{2} \int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}, \gamma_0) s_J(y|\boldsymbol{\theta}, \gamma_0) \left( h_J^{\frac{1}{2}}(y|\boldsymbol{\theta}^*, \gamma_0) - g_B^{(i)\frac{1}{2}}(y) \right) dy.$$

Thus,

$$(\hat{\theta}_{i,B}(\gamma_0) - \boldsymbol{\theta}^*(\gamma_0))' = D_B^{-1}(\tilde{\theta}_{i,B}(\gamma_0)) \nabla HD_B^2(\boldsymbol{\theta}^*(\gamma_0)).$$

By using the identity,  $b^{\frac{1}{2}} - a^{\frac{1}{2}} = (2a^{\frac{1}{2}})^{-1} ((b-a) - (b^{\frac{1}{2}} - a^{\frac{1}{2}})^2)$ ,  $\nabla HD_B^2(\boldsymbol{\theta}^*(\gamma_0))$  can be expressed as the difference of  $T_B(\gamma_0)$  and  $R_B(\gamma_0)$ , where

$$T_B(\gamma_0) \equiv \frac{1}{4} \int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}^*, \gamma_0) \left( h_J(y|\boldsymbol{\theta}^*, \gamma_0) - g_B^{(i)}(y) \right) dy,$$

and

$$R_B(\gamma_0) \equiv \frac{1}{4} \int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}^*, \gamma_0) \left( h_J^{\frac{1}{2}}(y|\boldsymbol{\theta}^*, \gamma_0) - g_B^{(i)\frac{1}{2}}(y) \right)^2 dy.$$

Hence,

$$B^{\frac{1}{2}} (\hat{\theta}_{i,B}(\gamma_0) - \boldsymbol{\theta}^*(\gamma_0))' = A_{1B}(\gamma_0) + A_{2B}(\gamma_0),$$

where  $A_{1B}(\gamma_0)$  and  $A_{2B}(\gamma_0)$  are given in (9).  $\square$

**Remark 2.** In the rest of the manuscript, we will refer to  $A_{2B}(\gamma_0)$  as the remainder term in the representation formula.

We now turn to the first main result of the manuscript, namely a central limit theorem for  $\hat{\theta}_{i,B}(\gamma_0)$  as first  $B \rightarrow \infty$  and then  $\gamma_0 \rightarrow 0$ . As a first step, we note that the Fisher information of the density  $h^{*J}(\cdot|\boldsymbol{\theta})$  is given by

$$I(\boldsymbol{\theta}) = \int_{\mathbb{R}} v_J(y|\boldsymbol{\theta}) v'_J(y|\boldsymbol{\theta}) h^{*J}(y|\boldsymbol{\theta}) dy. \quad (15)$$

Next we state the assumptions needed in the proof of Theorem 1. We separate these conditions as (i) model assumptions, (ii) kernel assumptions, (iii) regularity conditions, (iv) conditions that allow comparison of original data and compressed data.

#### Model assumptions on $h^{*J}(\cdot|\boldsymbol{\theta})$

(D1')  $h^{*J}(\cdot|\boldsymbol{\theta})$  is twice continuously differentiable in  $\boldsymbol{\theta}$ .

(D2') Assume further that  $\|\nabla t_J(\cdot|\boldsymbol{\theta})\|_2$  is continuous and bounded.

#### Kernel assumptions

(B1)  $K(\cdot)$  is symmetric about 0 on a compact support and bounded in  $L_2$ . We denote the support of  $K(\cdot)$  by  $Supp(K)$ .

**(B2)** The bandwidth  $c_B$  satisfies  $c_B \rightarrow 0$ ,  $B^{\frac{1}{2}}c_B \rightarrow \infty$ ,  $B^{\frac{1}{2}}c_B^2 \rightarrow 0$ .

### Regularity conditions

**(M1)** The function  $u_J(\cdot|\theta, \gamma_0)s_J(\cdot|\theta, \gamma_0)$  is continuously differentiable and bounded in  $L_2$  at  $\theta^*$ .

**(M2)** The function  $u_J(\cdot|\theta, \gamma_0)s_J(\cdot|\theta, \gamma_0)$  is continuous and bounded in  $L_2$  at  $\theta^*$ . In addition, assume that

$$\lim_{B \rightarrow \infty} \int_{\mathbb{R}} (\dot{u}_J(y|\theta_{i,B}, \gamma_0)s_J(y|\theta_{i,B}, \gamma_0) - \dot{u}_J(y|\theta^*, \gamma_0)s_J(y|\theta^*, \gamma_0))^2 dy = 0.$$

**(M3)** The function  $u_J(\cdot|\theta, \gamma_0)u'_J(\cdot|\theta, \gamma_0)s_J(\cdot|\theta, \gamma_0)$  is continuous and bounded in  $L_2$  at  $\theta^*$ ; also,

$$\lim_{B \rightarrow \infty} \int_{\mathbb{R}} (u_J(y|\hat{\theta}_{i,B}, \gamma_0)u'_J(y|\hat{\theta}_{i,B}, \gamma_0)s_J(y|\theta_{i,B}, \gamma_0) - u_J(y|\theta^*, \gamma_0)u'_J(y|\theta^*, \gamma_0)s_J(y|\theta^*, \gamma_0))^2 dy = 0.$$

**(M4)** Let  $\{\alpha_B : B \geq 1\}$  be a sequence diverging to infinity. Assume that

$$\lim_{B \rightarrow \infty} B \sup_{t \in \text{Supp}(K)} P_{\theta^*(\gamma_0)}(|\Delta - c_B t| > \alpha_B) = 0,$$

where  $\text{Supp}(K)$  is the support of the kernel density  $K(\cdot)$  and  $\Delta$  is a generic random variable with density  $h_J(\cdot|\theta^*, \gamma_0)$ .

**(M5)** Let

$$M_B = \sup_{|y| \leq \alpha_B} \sup_{t \in \text{Supp}(K)} \left| \frac{h_J(y - tc_B|\theta^*, \gamma_0)}{h_J(y|\theta^*, \gamma_0)} \right|.$$

Assume  $\sup_{B \geq 1} M_B < \infty$ .

**(M6)** The score function has a regular central behavior relative to the smoothing constants, i.e.,

$$\lim_{B \rightarrow \infty} (B^{\frac{1}{2}}c_B)^{-1} \int_{-\alpha_B}^{\alpha_B} u_J(y|\theta^*, \gamma_0) dy = 0.$$

Furthermore,

$$\lim_{B \rightarrow \infty} (B^{\frac{1}{2}}c_B^4) \int_{-\alpha_B}^{\alpha_B} u_J(y|\theta^*, \gamma_0) dy = 0.$$

**(M7)** The density functions are smooth in an  $L_2$  sense; i.e.,

$$\lim_{B \rightarrow \infty} \sup_{t \in \text{Supp}(K)} \int_{\mathbb{R}} (u_J(y + c_B t|\theta^*, \gamma_0) - u_J(y|\theta^*, \gamma_0))^2 h_J(y|\theta^*, \gamma_0) dy = 0.$$

**(M1')** The function  $v_J(\cdot|\theta)t_J(\cdot|\theta)$  is continuously differentiable and bounded in  $L_2$  at  $\theta^*$ .

**(M2')** The function  $\dot{v}_J(\cdot|\theta)t_J(\cdot|\theta)$  is continuous and bounded in  $L_2$  at  $\theta^*$ . In addition, assume that

$$\lim_{B \rightarrow \infty} \int_{\mathbb{R}} (\dot{v}_J(y|\theta_B)t_J(y|\theta_B) - \dot{v}_J(y|\theta^*)t_J(y|\theta^*))^2 dy = 0.$$

**(M3')** The function  $v_J(\cdot|\theta)v'_J(\cdot|\theta)t_J(\cdot|\theta)$  is continuous and bounded in  $L_2$  at  $\theta^*$ . also,

$$\lim_{B \rightarrow \infty} \int_{\mathbb{R}} (v_J(y|\hat{\theta}_{i,B})v'_J(y|\hat{\theta}_{i,B})t_J(y|\hat{\theta}_{i,B}) - v_J(y|\theta^*)v'_J(y|\theta^*)t_J(y|\theta^*))^2 dy = 0.$$

### Assumptions comparing models for original and compressed data

(O1) For all  $\theta \in \Theta$ ,

$$\lim_{\gamma_0 \rightarrow 0} \int_{\mathbb{R}} \left( u_J(y|\theta, \gamma_0) u'_J(y|\theta, \gamma_0) s_J(y|\theta, \gamma_0) - v_J(y|\theta) v'_J(y|\theta) t_J(y|\theta) \right)^2 dy = 0.$$

(O2) For all  $\theta \in \Theta$ ,

$$\lim_{\gamma_0 \rightarrow 0} \int_{\mathbb{R}} \left( \dot{u}_J(y|\theta, \gamma_0) s_J(y|\theta, \gamma_0) - \dot{v}_J(y|\theta) t_J(y|\theta) \right)^2 dy = 0.$$

**Theorem 1.** Assume that the conditions (B1)–(B2), (D1)–(D2), (D1')–(D2'), (M1)–(M7), (M1')–(M3'), and (O1)–(O2) hold. Then, for every  $1 \leq i \leq S$ , the following holds:

$$\lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} P \left( \sqrt{B} (\hat{\theta}_{i,B}(\gamma_0) - \theta^*(\gamma_0)) \leq x \right) = P(G \leq x),$$

where  $G$  is a bivariate Gaussian random variable with mean 0 and variance  $I^{-1}(\theta^*)$ , where  $I(\theta)$  is defined in (15).

Before we embark on the proof of Theorem 1, we first discuss the assumptions. Assumptions (B1) and (B2) are standard assumptions on the kernel and the bandwidth and are typically employed when investigating the asymptotic behavior of divergence-based estimators (see for instance [1]). Assumptions (M1)–(M7) and (M1')–(M3') are regularity conditions which are concerned essentially with  $L_2$  continuity and boundedness of the scores and their derivatives. Assumptions (O1)–(O2) allow for comparison of  $u_J(\cdot|\theta, \gamma_0)$  and  $v_J(\cdot|\theta)$ . Returning to the proof of Theorem 1, using representation formula, we will first show that  $\lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} P(A_{1B}(\gamma_0) \leq x) = P(G \leq x)$ , and then prove that  $\lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} A_{2B}(\gamma_0) = 0$  in probability. We start with the following proposition.

**Proposition 3.** Assume that the conditions (B1), (D1)–(D2), (M1)–(M3), (M1')–(M3'), (M7) and (O1)–(O2) hold. Then,

$$\lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} P(A_{1B}(\gamma_0) \leq x) = P(G \leq x),$$

where  $G$  is given in Theorem 1.

We divide the proof of Proposition 3 into two lemmas. In the first lemma we will show that

$$\lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} D_B(\tilde{\theta}_{i,B}(\gamma_0)) = \frac{1}{4} I(\theta^*).$$

Next in the second lemma we will show that first letting  $B \rightarrow \infty$  and then allowing  $\gamma_0 \rightarrow 0$ ,

$$4B^{\frac{1}{2}} T_B(\gamma_0) \xrightarrow{d} N(0, I(\theta^*)).$$

We start with the first part.

**Lemma 2.** Assume that the conditions (D1)–(D2), (D1')–(D2'), (M1)–(M3), (M1')–(M3') and (O1)–(O2) hold. Then, with probability one, the following prevails:

$$\lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} D_B(\tilde{\theta}_{i,B}(\gamma_0)) = \frac{1}{4} I(\theta^*).$$

**Proof.** Using representation formula in Lemma 1. First fix  $\gamma_0 > 0$ . It suffices to show

$$\lim_{B \rightarrow \infty} D_{1B}(\tilde{\theta}_{i,B}(\gamma_0)) = \frac{1}{2}I(\theta^*(\gamma_0)), \quad \text{and} \quad \lim_{B \rightarrow \infty} D_{2B}(\tilde{\theta}_{i,B}(\gamma_0)) = -\frac{1}{4}I(\theta^*(\gamma_0)).$$

We begin with  $D_{1B}(\tilde{\theta}_{i,B}(\gamma_0))$ . By algebra,  $D_{1B}(\tilde{\theta}_{i,B}(\gamma_0))$  can be expressed as

$$D_{1B}(\tilde{\theta}_{i,B}(\gamma_0)) = D_{1B}^{(1)}(\tilde{\theta}_{i,B}(\gamma_0)) + D_{1B}^{(2)}(\tilde{\theta}_{i,B}(\gamma_0)) + D_{1B}^{(3)}(\theta^*(\gamma_0)), \quad \text{where}$$

$$D_{1B}^{(1)}(\tilde{\theta}_{i,B}(\gamma_0)) = -\frac{1}{2} \int_{\mathbb{R}} \dot{u}_J(y|\tilde{\theta}_{i,B}, \gamma_0) s_J(y|\tilde{\theta}_{i,B}, \gamma_0) \left( g_B^{(i)\frac{1}{2}}(y) - s_J(y|\theta^*, \gamma_0) \right) dy,$$

$$D_{1B}^{(2)}(\tilde{\theta}_{i,B}(\gamma_0)) = -\frac{1}{2} \int_{\mathbb{R}} (\dot{u}_J(y|\tilde{\theta}_{i,B}, \gamma_0) s_J(y|\tilde{\theta}_{i,B}, \gamma_0) - \dot{u}_J(y|\theta^*, \gamma_0) s_J(y|\theta^*, \gamma_0)) h_J^{\frac{1}{2}}(y|\theta^*, \gamma_0) dy,$$

$$\text{and} \quad D_{1B}^{(3)}(\theta^*(\gamma_0)) = -\frac{1}{2} \int_{\mathbb{R}} \dot{u}_J(y|\theta^*, \gamma_0) h_J(y|\theta^*, \gamma_0) dy = \frac{1}{2} I(\theta^*(\gamma_0)).$$

It suffices to show that as  $B \rightarrow \infty$ ,  $D_{1B}^{(1)}(\tilde{\theta}_{i,B}(\gamma_0)) \rightarrow 0$ , and  $D_{1B}^{(2)}(\tilde{\theta}_{i,B}(\gamma_0)) \rightarrow 0$ . We first consider  $D_{1B}^{(1)}(\tilde{\theta}_{i,B}(\gamma_0))$ . By Cauchy-Schwarz inequality and assumption (M2), it follows that there exists  $0 < C_1 < \infty$ ,

$$\begin{aligned} |D_{1B}^{(1)}(\tilde{\theta}_{i,B}(\gamma_0))| &\leq \frac{1}{2} \left\{ \int_{\mathbb{R}} (\dot{u}_J(y|\tilde{\theta}_{i,B}, \gamma_0) s_J(y|\tilde{\theta}_{i,B}, \gamma_0))^2 dy \right\}^{\frac{1}{2}} \left\{ \int_{\mathbb{R}} \left( g_B^{(i)\frac{1}{2}}(y) - s_J(y|\theta^*, \gamma_0) \right)^2 dy \right\}^{\frac{1}{2}} \\ &\leq C_1 \left\{ \int_{\mathbb{R}} \left( g_B^{(i)\frac{1}{2}}(y) - s_J(y|\theta^*, \gamma_0) \right)^2 dy \right\}^{\frac{1}{2}} \rightarrow 0, \end{aligned}$$

where the last convergence follows from the  $L_1$  convergence of  $g_B^{(i)}(\cdot)$  and  $h_J(\cdot|\theta^*, \gamma_0)$ . Hence, as  $B \rightarrow \infty$ ,  $D_{1B}^{(1)}(\tilde{\theta}_{i,B}(\gamma_0)) \rightarrow 0$ . Next we consider  $D_{1B}^{(2)}(\tilde{\theta}_{i,B}(\gamma_0))$ . Again, by Cauchy-Schwarz inequality and assumption (M2), it follows that  $D_{1B}^{(2)}(\tilde{\theta}_{i,B}(\gamma_0)) \rightarrow 0$ . Hence  $D_{1B}(\tilde{\theta}_{i,B}(\gamma_0)) \rightarrow \frac{1}{2}I(\theta^*(\gamma_0))$ . Turning to  $D_{2B}(\tilde{\theta}_{i,B}(\gamma_0))$ , by similar argument, using Cauchy-Schwarz inequality and assumption (M3), it follows that  $D_{2B}(\tilde{\theta}_{i,B}(\gamma_0)) \rightarrow -\frac{1}{4}I(\theta^*(\gamma_0))$ . Thus, to complete the proof, it is enough to show that

$$\lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} D_{1B}(\tilde{\theta}_{i,B}(\gamma_0)) = \frac{1}{2}I(\theta^*) \quad \text{and} \quad \lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} D_{2B}(\tilde{\theta}_{i,B}(\gamma_0)) = -\frac{1}{4}I(\theta^*). \quad (16)$$

We start with the first term of (16). Let

$$\mathcal{J}_1(\gamma_0) = \int_{\mathbb{R}} \dot{u}_J(y|\theta^*, \gamma_0) h_J(y|\theta^*, \gamma_0) dy - \int_{\mathbb{R}} \dot{v}_J(y|\theta^*) h_J^*(y|\theta^*) dy.$$

We will show that  $\lim_{\gamma_0 \rightarrow 0} \mathcal{J}_1(\gamma_0) = 0$ . By algebra, the difference of the above two terms can be expressed as the sum of  $\mathcal{J}_{11}(\gamma_0)$  and  $\mathcal{J}_{12}(\gamma_0)$ , where

$$\mathcal{J}_{11}(\gamma_0) = \int_{\mathbb{R}} (\dot{u}_J(y|\theta^*, \gamma_0) s_J(y|\theta^*, \gamma_0) - \dot{v}_J(y|\theta^*) t_J(y|\theta^*)) s_J(y|\theta^*, \gamma_0) dy, \quad \text{and}$$

$$\mathcal{J}_{12}(\gamma_0) = \int_{\mathbb{R}} \dot{v}_J(y|\theta^*) t_J(y|\theta^*) (s_J(y|\theta^*, \gamma_0) - t_J(y|\theta^*)) dy.$$

$\mathcal{J}_{11}(\gamma_0)$  converges to zero by Cauchy-Schwarz inequality and assumption (O2), and  $\mathcal{J}_{12}(\gamma_0)$  converges to zero by Cauchy-Schwarz inequality, assumption (M2') and Scheffe's theorem. Next we consider the second term of (16). Let

$$\mathcal{J}_2(\gamma_0) = \int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}^*, \gamma_0) u'_J(y|\boldsymbol{\theta}^*, \gamma_0) h_J(y|\boldsymbol{\theta}^*, \gamma_0) dy - \int_{\mathbb{R}} v_J(y|\boldsymbol{\theta}^*) v'_J(y|\boldsymbol{\theta}^*) h^{*J}(y|\boldsymbol{\theta}^*) dy.$$

We will show that  $\lim_{\gamma_0 \rightarrow 0} \mathcal{J}_2(\gamma_0) = 0$ . By algebra, the difference of the above two terms can be expressed as the sum of  $\mathcal{J}_{21}(\gamma_0)$  and  $\mathcal{J}_{22}(\gamma_0)$ , where

$$\mathcal{J}_{21}(\gamma_0) = \int_{\mathbb{R}} \left( u_J(y|\boldsymbol{\theta}^*, \gamma_0) u'_J(y|\boldsymbol{\theta}^*, \gamma_0) s_J(y|\boldsymbol{\theta}^*, \gamma_0) - v_J(y|\boldsymbol{\theta}^*) v'_J(y|\boldsymbol{\theta}^*) t_J(y|\boldsymbol{\theta}^*) \right) s_J(y|\boldsymbol{\theta}^*, \gamma_0) dy,$$

$$\text{and } \mathcal{J}_{22}(\gamma_0) = \int_{\mathbb{R}} v_J(y|\boldsymbol{\theta}^*) v'_J(y|\boldsymbol{\theta}^*) t_J(y|\boldsymbol{\theta}^*) (s_J(y|\boldsymbol{\theta}^*, \gamma_0) - t_J(y|\boldsymbol{\theta}^*)) dy.$$

$\mathcal{J}_{11}(\gamma_0)$  converges to zero by Cauchy-Schwarz inequality and assumption (O1), and  $\mathcal{J}_{12}(\gamma_0)$  converges to zero by Cauchy-Schwarz inequality, assumption (M3') and Scheffe's theorem. Therefore the lemma holds.  $\square$

**Lemma 3.** Assume that the conditions (B1), (D1)–(D2), (D1')–(D2'), (M1)–(M3), (M3'), (M7) and (O1)–(O2) hold. Then, first letting  $B \rightarrow \infty$ , and then  $\gamma_0 \rightarrow 0$ ,

$$4B^{\frac{1}{2}} T_B(\gamma_0) \xrightarrow{d} N(0, I(\boldsymbol{\theta}^*)).$$

**Proof.** First fix  $\gamma_0 > 0$ . Please note that using  $\int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}^*, \gamma_0) h_J(y|\boldsymbol{\theta}^*, \gamma_0) dy = 0$ , we have that

$$\begin{aligned} 4B^{\frac{1}{2}} T_B(\gamma_0) &= B^{\frac{1}{2}} \int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}^*, \gamma_0) g_B^{(i)}(y) dy \\ &= B^{\frac{1}{2}} \int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}^*, \gamma_0) \frac{1}{B} \sum_{l=1}^B \frac{1}{c_B} K\left(\frac{y - Y_{il}}{c_B}\right) dy \\ &= B^{\frac{1}{2}} \frac{1}{B} \sum_{l=1}^B \int_{\mathbb{R}} u_J(Y_{il} + c_B t|\boldsymbol{\theta}^*, \gamma_0) K(t) dt. \end{aligned}$$

Therefore,

$$4B^{\frac{1}{2}} T_B(\gamma_0) - B^{\frac{1}{2}} \frac{1}{B} \sum_{l=1}^B u_J(Y_{il}|\boldsymbol{\theta}^*, \gamma_0) = B^{\frac{1}{2}} \frac{1}{B} \sum_{l=1}^B \int_{\mathbb{R}} (u_J(Y_{il} + c_B t|\boldsymbol{\theta}^*, \gamma_0) - u_J(Y_{il}|\boldsymbol{\theta}^*, \gamma_0)) K(t) dt.$$

Since  $Y_{il}$ 's are i.i.d. across  $l$ , using Cauchy-Schwarz inequality and assumption (B1), we can show that there exists  $0 < C < \infty$ ,

$$\begin{aligned} E \left[ 4B^{\frac{1}{2}} T_B - B^{\frac{1}{2}} \frac{1}{B} \sum_{l=1}^B u_J(Y_{il}|\boldsymbol{\theta}^*, \gamma_0) \right]^2 &= E \left[ \int_{\mathbb{R}} (u_J(Y_{i1} + c_B t|\boldsymbol{\theta}^*, \gamma_0) - u_J(Y_{i1}|\boldsymbol{\theta}^*, \gamma_0)) K(t) dt \right]^2 \\ &\leq CE \left[ \left\{ \int_{\mathbb{R}} (u_J(Y_{i1} + c_B t|\boldsymbol{\theta}^*, \gamma_0) - u_J(Y_{i1}|\boldsymbol{\theta}^*, \gamma_0))^2 dt \right\}^{\frac{1}{2}} \right]^2 \\ &\leq CE \left[ \int_{\mathbb{R}} (u_J(Y_{i1} + c_B t|\boldsymbol{\theta}^*, \gamma_0) - u_J(Y_{i1}|\boldsymbol{\theta}^*, \gamma_0))^2 dt \right] \\ &= C \int_{\mathbb{R}} \int_{\mathbb{R}} (u_J(y + c_B t|\boldsymbol{\theta}^*, \gamma_0) - u_J(y|\boldsymbol{\theta}^*, \gamma_0))^2 h_J(y|\boldsymbol{\theta}^*, \gamma_0) dy dt, \end{aligned}$$

converging to zero as  $B \rightarrow \infty$  by assumption (M7). Also, the limiting distribution of  $4B^{\frac{1}{2}}T_B(\gamma_0)$  is  $N(0, I(\theta^*(\gamma_0)))$  as  $B \rightarrow \infty$ . Now let  $\gamma_0 \rightarrow 0$ . It is enough to show that as  $\gamma_0 \rightarrow 0$  the density of  $N(0, I(\theta^*(\gamma_0)))$  converges to the density of  $N(0, I(\theta^*))$ . To this end, it suffices to show that  $\lim_{\gamma_0 \rightarrow 0} I(\theta^*(\gamma_0)) = I(\theta^*)$ . However, this is established in Lemma 2. Combining the results, the lemma follows.  $\square$

**Proof of Proposition 3.** The proof of Proposition 3 follows immediately by combining Lemmas 2 and 3.  $\square$

We now turn to establishing that the remainder term in the representation formula converges to zero.

**Lemma 4.** Assume that the assumptions (B1)–(B2), (M1)–(M6) hold. Then

$$\lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} A_{2B}(\gamma_0) = 0 \quad \text{in probability.}$$

**Proof.** Using Lemma 2, it is sufficient to show that  $B^{\frac{1}{2}}R_B$  converges to 0 in probability as  $B \rightarrow \infty$ . Let

$$d_J(y|\theta^*(\gamma_0)) = g_B^{(i)\frac{1}{2}}(y) - s_J(y|\theta^*, \gamma_0).$$

Please note that

$$d_J^2(y|\theta^*(\gamma_0)) \leq 2 \left\{ \left( h_J(y|\theta^*, \gamma_0) - E[g_B^{(i)}(y)] \right)^2 + \left( E[g_B^{(i)}(y)] - g_B^{(i)}(y) \right)^2 \right\} h_J^{-1}(y|\theta^*, \gamma_0).$$

Then

$$\begin{aligned} |R_B(\gamma_0)| &\leq \frac{1}{2} \int_{\mathbb{R}} |u_J(y|\theta^*, \gamma_0)| d_J^2(y|\theta^*(\gamma_0)) dy \\ &\leq \frac{1}{2} \int_{-\alpha_B}^{\alpha_B} |u_J(y|\theta^*, \gamma_0)| d_J^2(y|\theta^*(\gamma_0)) dy + \frac{1}{2} \int_{|y| \geq \alpha_B} |u_J(y|\theta^*, \gamma_0)| d_J^2(y|\theta^*(\gamma_0)) dy \\ &\equiv R_{1B}(\gamma_0) + R_{2B}(\gamma_0). \end{aligned}$$

We first deal with  $R_{1B}(\gamma_0)$ , which can be expressed as the sum of  $R_{1B}(\gamma_0)$  and  $R_{2B}(\gamma_0)$ , where

$$R_{1B}^{(1)}(\gamma_0) = \int_{-\alpha_B}^{\alpha_B} |u_J(y|\theta^*, \gamma_0)| \left( h_J(y|\theta^*, \gamma_0) - E[g_B^{(i)}(y)] \right)^2 h_J^{-1}(y|\theta^*, \gamma_0) dy, \quad (17)$$

$$\text{and } R_{1B}^{(2)}(\gamma_0) = \int_{-\alpha_B}^{\alpha_B} |u_J(y|\theta^*, \gamma_0)| \left( E[g_B^{(i)}(y)] - g_B^{(i)}(y) \right)^2 h_J^{-1}(y|\theta^*, \gamma_0) dy.$$

Now consider  $R_{1B}^{(2)}$ . Let  $\epsilon > 0$  be arbitrary but fixed. Then, by Markov's inequality,

$$\begin{aligned} P\left(B^{\frac{1}{2}}R_{1B}^{(2)} > \epsilon\right) &\leq \epsilon^{-1} B^{\frac{1}{2}} E[R_{1B}^{(2)}] \\ &\leq \epsilon^{-1} B^{\frac{1}{2}} \int_{\alpha_B}^{\alpha_B} |u_J(y|\theta^*, \gamma_0)| \left( Var[g_B^{(i)}(y)] \right) h_J^{-1}(y|\theta^*, \gamma_0) dy. \end{aligned} \quad (18)$$

Now since  $Y_i^t$ s are independent and identically distributed across  $i$ , it follows that

$$Var[g_B^{(i)}(y)] \leq \frac{1}{Bc_B} \int_{\mathbb{R}} K^2(t) h_J(y - tc_B|\theta^*, \gamma_0) dt. \quad (19)$$

Now plugging (19) into (18), interchanging the order of integration (using Tonelli's Theorem), we get

$$P\left(B^{\frac{1}{2}}R_{1B}^{(2)} > \epsilon\right) \leq C \left(B^{\frac{1}{2}}c_B\right)^{-1} \int_{-\alpha_B}^{\alpha_B} |u_J(y|\theta^*, \gamma_0)| dy \rightarrow 0,$$

where  $C$  is a universal constant, and the last convergence follows from conditions (M5)–(M6). We now deal with  $R_{1B}^{(1)}$ . To this end, we need to calculate  $\left(E\left[g_B^{(i)}(y)\right] - h_J(y|\theta^*, \gamma_0)\right)^2$ . Using change of variables, two-step Taylor approximation, and assumption (B1), we get

$$\begin{aligned} E\left[g_B^{(i)}(y)\right] - h_J(y|\theta^*, \gamma_0) &= \int_{\mathbb{R}} K(t) (h_J(y - tc_B|\theta^*, \gamma_0) - h_J(y|\theta^*, \gamma_0)) dt \\ &= \int_{\mathbb{R}} K(t) \frac{(tc_B)^2}{2} h_J''(y_B^*(t)|\theta^*, \gamma_0) dt. \end{aligned} \quad (20)$$

Now plugging in (20) into (17) and using conditions (M3) and (M6), we get

$$B^{\frac{1}{2}}R_{1B}^{(1)}(\gamma_0) \leq CB^{\frac{1}{2}}c_B^4 \int_{-\alpha_B}^{\alpha_B} |u_J(y|\theta^*, \gamma_0)| dy. \quad (21)$$

Convergence of (21) to 0 now follows from condition (M6). We next deal with  $R_{2B}(\gamma_0)$ . To this end, by writing our the square term of  $d_J(\cdot|\theta^*(\gamma_0))$ , we have

$$B^{\frac{1}{2}}R_{2B}(\gamma_0) = \int_{|y| \geq \alpha_B} |u_J(y|\theta^*, \gamma_0)| \left( h_J(y|\theta^*, \gamma_0) + g_B^{(i)}(y) - s_J(y|\theta^*, \gamma_0)g_B^{(i)\frac{1}{2}}(y) \right) dy. \quad (22)$$

We will show that the RHS of (22) converges to 0 as  $B \rightarrow \infty$ . We begin with the first term. Please note that by Cauchy-Schwarz inequality,

$$B \left( \int_{|y| \geq \alpha_B} |u_J(y|\theta^*, \gamma_0)| h_J(y|\theta^*, \gamma_0) dy \right)^2 \leq \left\{ \int_{\mathbb{R}} u_J(y|\theta^*, \gamma_0) u'_J(y|\theta^*, \gamma_0) h_J(y|\theta^*, \gamma_0) dy \right\}_{\{BP_{\theta^*(\gamma_0)}(|\Delta| \geq \alpha_B)\}},$$

the last term converges to 0 by (M4). As for the second term, note that, a.s., by Cauchy-Schwarz inequality,

$$\left( \int_{|y| \geq \alpha_B} |u_J(y|\theta^*, \gamma_0)| g_B^{(i)}(y) dy \right)^2 \leq \int_{|y| \geq \alpha_B} u_J(y|\theta^*, \gamma_0) u'_J(y|\theta^*, \gamma_0) g_B^{(i)}(y) dy.$$

Now taking the expectation and using Cauchy-Schwarz inequality, one can show that

$$BE \left[ \int_{|y| \geq \alpha_m} |u_J(y|\theta^*, \gamma_0)| g_B^{(i)}(y) dy \right]^2 \leq a_B \int_{\mathbb{R}} K(t) \int_{\mathbb{R}} u_J(y|\theta^*, \gamma_0) u'_J(y|\theta^*, \gamma_0) h_J(y - c_B t|\theta^*, \gamma_0) dy dt,$$

where  $a_B = B \sup_{z \in \text{Supp}(K)} P_{\theta^*}(|\Delta - c_B z| > \alpha_B)$ . The convergence to 0 of the RHS of above inequality now follows from condition (M4). Finally, by another application of the Cauchy-Schwarz inequality,

$$BE \left[ \int_{|y| \geq \alpha_m} |u_J(y|\theta^*, \gamma_0)| g_B^{(i)\frac{1}{2}}(y) s_J(y|\theta^*, \gamma_0) dy \right] \leq a_B \int_{\mathbb{R}} u_J(y - c_B t|\theta^*, \gamma_0) u'_J(y - c_B t|\theta^*, \gamma_0) h_J(y|\theta^*, \gamma_0) dy.$$

The convergence of RHS of above inequality to zero follows from (M4). Now the lemma follows.  $\square$

**Proof of Theorem 1.** Recall that

$$B^{\frac{1}{2}} (\hat{\theta}_{i,B}(\gamma_0) - \theta^*(\gamma_0))' = A_{1B}(\gamma_0) + A_{2B}(\gamma_0),$$

where  $A_{1B}(\gamma_0)$  and  $A_{2B}(\gamma_0)$  are given in (9). Proposition 3 shows that  $\lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} A_{1B}(\gamma_0) = N(0, I^{-1}(\theta^*))$ ; while Lemma 4 shows that  $\lim_{\gamma_0 \rightarrow 0} \lim_{B \rightarrow \infty} A_{2B}(\gamma_0) = 0$  in probability. The result follows from Slutsky's theorem.  $\square$

We next show that by interchanging the limits, namely first allowing  $\gamma_0$  to converge to 0 and then letting  $B \rightarrow \infty$  the limit distribution of  $\hat{\theta}_{i,B}(\gamma_0)$  is Gaussian with the same covariance matrix as Theorem 1. We begin with additional assumptions required in the proof of the theorem.

#### Regularity conditions

(M4') Let  $\{\alpha_B : B \geq 1\}$  be a sequence diverging to infinity. Assume that

$$\lim_{B \rightarrow \infty} B \sup_{t \in \text{Supp}(K)} P_{\theta^*}(|\Delta - c_B t| > \alpha_B) = 0,$$

where  $\text{Supp}(K)$  is the support of the kernel density  $K(\cdot)$  and  $\Delta$  is a generic random variable with density  $h^{*J}(\cdot | \theta^*)$ .

(M5') Let

$$M_B = \sup_{|y| \leq \alpha_B} \sup_{t \in \text{Supp}(K)} \left| \frac{h^{*J}(y - tc_B | \theta^*)}{h^{*J}(y | \theta^*)} \right|.$$

Assume that  $\sup_{B \geq 1} M_B < \infty$ .

(M6') The score function has a regular central behavior relative to the smoothing constants, i.e.,

$$\lim_{B \rightarrow \infty} (B^{\frac{1}{2}} c_B)^{-1} \int_{-\alpha_B}^{\alpha_B} v_J(y | \theta^*) dy = 0.$$

Furthermore,

$$\lim_{B \rightarrow \infty} (B^{\frac{1}{2}} c_B^4) \int_{-\alpha_B}^{\alpha_B} v_J(y | \theta^*) dy = 0.$$

(M7') The density functions are smooth in an  $L_2$  sense; i.e.,

$$\lim_{B \rightarrow \infty} \sup_{t \in \text{Supp}(K)} \int_{\mathbb{R}} (v_J(y + c_B t | \theta^*) - v_J(y | \theta^*))^2 h^{*J}(y | \theta^*) dy = 0.$$

#### Assumptions comparing models for original and compressed data

(V1) Assume that  $\lim_{\gamma_0 \rightarrow 0} \sup_y |u_J(y | \theta^*, \gamma_0) - v_J(y | \theta^*)| = 0$ .

(V2)  $v_J(\cdot | \theta)$  is  $L_1$  continuous in the sense that  $X_n \xrightarrow{P} X$  implies that  $E[v_J(X_n | \theta) - v_J(X | \theta)] = 0$ , where the expectation is with respect to distribution  $K(\cdot)$ .

(V3) Assume that for all  $\theta \in \Theta$ ,  $\int_{\mathbb{R}} \nabla h^{*J}(y | \theta) dy < \infty$ .

(V4) Assume that for all  $\theta \in \Theta$ ,  $\lim_{\gamma_0 \rightarrow 0} \sup_y \left| \frac{s_J(y | \theta, \gamma_0)}{t_J(y | \theta)} - 1 \right| = 0$ .

**Theorem 2.** Assume that the conditions (B1)–(B2), (D1')–(D2'), (M1')–(M7'), (O1)–(O2) and (V1)–(V4) hold. Then,

$$\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} P \left( \sqrt{B} (\hat{\theta}_{i,B}(\gamma_0) - \theta^*(\gamma_0)) \leq x \right) = P(G \leq x),$$

where  $G$  is a bivariate Gaussian random variable with mean 0 and variance  $I^{-1}(\boldsymbol{\theta}^*)$ .

We notice that in the above Theorem 2 that we use conditions (V2)–(V4) which are regularity conditions on the scores of the  $J$ -fold convolution of  $f(\cdot)$  while (V1) facilitates comparison of the scores of the densities of the compressed data and that of the  $J$ -fold convolution. As before, we will first establish (a):

$$\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} \mathbf{P}(A_{1B}(\gamma_0) \leq x) = \mathbf{P}(G \leq x),$$

and then (b):  $\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} A_{2B}(\gamma_0) = 0$  in probability. We start with the proof of (a).

**Proposition 4.** Assume that the conditions (B1)–(B2), (D1')–(D2'), (M1')–(M3'), (M7'), (O1)–(O2), and (V1)–(V2) hold. Then,

$$\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} \mathbf{P}(A_{1B}(\gamma_0) \leq x) = \mathbf{P}(G \leq x).$$

We divide the proof of Proposition 4 into two lemmas. In the first lemma, we will show that

$$\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} D_B(\tilde{\boldsymbol{\theta}}_{i,B}(\gamma_0)) = \frac{1}{4} I(\boldsymbol{\theta}^*).$$

In the second lemma, we will show that first let  $\gamma_0 \rightarrow 0$ , then let  $B \rightarrow \infty$ ,

$$4B^{\frac{1}{2}} T_B(\gamma_0) \xrightarrow{d} N(0, I(\boldsymbol{\theta}^*)).$$

**Lemma 5.** Assume that the conditions (B1)–(B2), (D1')–(D2'), (M1')–(M3'), (O1)–(O2), and (V1)–(V2) hold. Then,

$$\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} D_B(\tilde{\boldsymbol{\theta}}_{i,B}(\gamma_0)) = \frac{1}{4} I(\boldsymbol{\theta}^*). \quad (23)$$

**Proof.** First fix  $B$ . Recall that

$$\begin{aligned} D_B(\boldsymbol{\theta}(\gamma_0)) &= -\frac{1}{2} \int_{\mathbb{R}} \dot{u}_J(y|\boldsymbol{\theta}, \gamma_0) s_J(y|\boldsymbol{\theta}, \gamma_0) (g_B^{(i)}(y))^{\frac{1}{2}} dy \\ &\quad -\frac{1}{4} \int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}, \gamma_0) u'_J(y|\boldsymbol{\theta}, \gamma_0) s_J(y|\boldsymbol{\theta}, \gamma_0) (g_B^{(i)}(y))^{\frac{1}{2}} dy \\ &\equiv D_{1B}(\boldsymbol{\theta}(\gamma_0)) + D_{2B}(\boldsymbol{\theta}(\gamma_0)). \end{aligned}$$

By algebra,  $D_{1B}(\tilde{\boldsymbol{\theta}}_{i,B}(\gamma_0))$  can be expressed as the sum of  $H_{1B}^{(1)}$ ,  $H_{1B}^{(2)}$ ,  $H_{1B}^{(3)}$ ,  $H_{1B}^{(4)}$  and  $H_{1B}^{(5)}$ , where

$$H_{1B}^{(1)} = -\frac{1}{2} \int_{\mathbb{R}} [\dot{u}_J(y|\tilde{\boldsymbol{\theta}}_{i,B}, \gamma_0) s_J(y|\tilde{\boldsymbol{\theta}}_{i,B}, \gamma_0) - \dot{v}_J(y|\tilde{\boldsymbol{\theta}}_{i,B}) t_J(y|\tilde{\boldsymbol{\theta}}_{i,B})] g_B^{(i)\frac{1}{2}}(y) dy,$$

$$H_{1B}^{(2)} = -\frac{1}{2} \int_{\mathbb{R}} [\dot{v}_J(y|\tilde{\boldsymbol{\theta}}_{i,B}) t_J(y|\tilde{\boldsymbol{\theta}}_{i,B}) - \dot{v}_J(y|\boldsymbol{\theta}^*) t_J(y|\boldsymbol{\theta}^*)] g_B^{(i)\frac{1}{2}}(y) dy,$$

$$H_{1B}^{(3)} = -\frac{1}{2} \int_{\mathbb{R}} \dot{v}_J(y|\boldsymbol{\theta}^*) t_J(y|\boldsymbol{\theta}^*) \left[ g_B^{(i)\frac{1}{2}}(y) - h_J^{\frac{1}{2}}(y|\boldsymbol{\theta}^*, \gamma_0) \right] dy,$$

$$H_{1B}^{(4)} = -\frac{1}{2} \int_{\mathbb{R}} \dot{v}_J(y|\boldsymbol{\theta}^*) t_J(y|\boldsymbol{\theta}^*) [s_J(y|\boldsymbol{\theta}^*, \gamma_0) - t_J(y|\boldsymbol{\theta}^*)] dy, \quad \text{and} \quad H_{1B}^{(5)} = \frac{1}{2} I(\boldsymbol{\theta}^*).$$

We will show that

$$\lim_{\gamma_0 \rightarrow 0} D_{1B}(\tilde{\theta}_{i,B}(\gamma_0)) = H_{1B}^{(2)} + \lim_{\gamma_0 \rightarrow 0} H_{1B}^{(3)} + H_{1B}^{(5)}, \quad (24)$$

where

$$\lim_{\gamma_0 \rightarrow 0} H_{1B}^{(3)} = -\frac{1}{2} \int_{\mathbb{R}} \dot{v}_J(y|\theta^*) t_J(y|\theta^*) \left[ g_B^{*(\frac{1}{2})}(y) - t_J(y|\theta^*) \right] dy \quad \text{and} \quad (25)$$

$g_B^*(\cdot)$  is given in (7). First consider  $H_{1B}^{(1)}$ . It converges to zero as  $\gamma_0 \rightarrow 0$  by Cauchy-Schwarz inequality and assumption (O2). Next we consider  $H_{1B}^{(3)}$ . We will first show that

$$\lim_{\gamma_0 \rightarrow 0} -\frac{1}{2} \int_{\mathbb{R}} \dot{v}_J(y|\theta^*) t_J(y|\theta^*) g_B^{(i)\frac{1}{2}}(y) dy = -\frac{1}{2} \int_{\mathbb{R}} \dot{v}_J(y|\theta^*) t_J(y|\theta^*) g_B^{*\frac{1}{2}}(y) dy.$$

To this end, notice that by Cauchy-Schwarz inequality and boundedness of  $\dot{v}_J(y|\theta^*) t_J(y|\theta^*)$  in  $L_2$ , it follows that there exists a constant  $C$  such that

$$\begin{aligned} \left| \int_{\mathbb{R}} \dot{v}_J(y|\theta^*) t_J(y|\theta^*) \left[ g_B^{(i)\frac{1}{2}}(y) - g_B^{*\frac{1}{2}}(y) \right] dy \right| &\leq C \left\{ \int_{\mathbb{R}} \left( g_B^{(i)\frac{1}{2}}(y) - g_B^{*\frac{1}{2}}(y) \right)^2 dy \right\}^{\frac{1}{2}} \\ &\leq C \left\{ \int_{\mathbb{R}} |g_B^{(i)}(y) - g_B^*(y)| dy \right\}^{\frac{1}{2}}. \end{aligned}$$

It suffices to show that  $g_B^{(i)}(\cdot)$  converges to  $g_B^*(\cdot)$  in  $L_1$ . Since

$$\int_{\mathbb{R}} |g_B^{(i)}(y) - g_B^*(y)| dy = 2 - 2 \int_{\mathbb{R}} \min \{g_B^{(i)}(y), g_B^*(y)\} dy,$$

and  $\min \{g_B^{(i)}(y), g_B^*(y)\} \leq g_B^*(y)$ , by dominated convergence theorem,  $g_B^{(i)}(\cdot) \xrightarrow{L_1} g_B^*(\cdot)$ . Next we will show that

$$\lim_{\gamma_0 \rightarrow 0} -\frac{1}{2} \int_{\mathbb{R}} \dot{v}_J(y|\theta^*) t_J(y|\theta^*) s_J(y|\theta^*, \gamma_0) dy = -\frac{1}{2} \int_{\mathbb{R}} \dot{v}_J(y|\theta^*) t_J(y|\theta^*) t_J(y|\theta^*) dy.$$

In addition, by Cauchy-Schwarz inequality, boundedness of  $\dot{v}_J(y|\theta^*) t_J(y|\theta^*)$  in  $L_2$  and Scheffe's theorem, we have that  $\int_{\mathbb{R}} \dot{v}_J(y|\theta^*) h_J^{\frac{1}{2}}(y|\theta^*, \gamma_0) (s_J(y|\theta^*, \gamma_0) - t_J(y|\theta^*)) dy$  converges to zero as  $\gamma_0 \rightarrow 0$ . Next we consider  $H_{1B}^{(4)}$ . It converges to zero by Cauchy-Schwarz inequality and assumption (M2'). Thus (24) holds. Now let  $B \rightarrow \infty$ , we will show that  $\lim_{B \rightarrow \infty} H_{1B}^{(2)} = 0$  and  $\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} H_{1B}^{(3)} = 0$ . First consider  $\lim_{B \rightarrow \infty} H_{1B}^{(2)}$ . It converges to zero by Cauchy-Schwarz inequality and assumption (M2'). Next we consider  $\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} H_{1B}^{(3)}$ . It converges to zero by Cauchy-Schwarz inequality and  $L_1$  convergence of  $g_B^*(\cdot)$  and  $h^{*J}(\cdot|\theta^*)$ . Therefore  $\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} D_{1B}(\tilde{\theta}_{i,B}(\gamma_0)) = \frac{1}{2} I(\theta^*)$ .

We now turn to show that  $\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} D_{2B}(\tilde{\theta}_{i,B}(\gamma_0)) = -\frac{1}{4} I(\theta^*)$ . First fix  $B$  and express  $D_{2B}(\tilde{\theta}_{i,B}(\gamma_0))$  as the sum of  $H_{2B}^{(1)}, H_{2B}^{(2)}, H_{2B}^{(3)}, H_{2B}^{(4)}$ , and  $H_{2B}^{(5)}$ , where

$$H_{2B}^{(1)} = -\frac{1}{4} \int_{\mathbb{R}} \left[ u_J(y|\tilde{\theta}_{i,B}, \gamma_0) u'_J(y|\tilde{\theta}_{i,B}, \gamma_0) s_J(y|\tilde{\theta}_{i,B}, \gamma_0) - v_J(y|\tilde{\theta}_{i,B}) v'_J(y|\tilde{\theta}_{i,B}) t_J(y|\tilde{\theta}_{i,B}) \right] g_B^{(i)\frac{1}{2}}(y) dy,$$

$$H_{2B}^{(2)} = -\frac{1}{4} \int_{\mathbb{R}} \left[ v_J(y|\tilde{\theta}_{i,B}) v'_J(y|\tilde{\theta}_{i,B}) t_J(y|\tilde{\theta}_{i,B}) - v_J(y|\theta^*) v'_J(y|\theta^*) t_J(y|\theta^*) \right] g_B^{(i)\frac{1}{2}}(y) dy,$$

$$H_{2B}^{(3)} = -\frac{1}{4} \int_{\mathbb{R}} v_J(y|\theta^*) v'_J(y|\theta^*) t_J(y|\theta^*) \left[ g_B^{(i)\frac{1}{2}}(y) - h_J^{\frac{1}{2}}(y|\theta^*, \gamma_0) \right] dy,$$

$$H_{2B}^{(4)} = -\frac{1}{4} \int_{\mathbb{R}} v_J(y|\theta^*) v'_J(y|\theta^*) t_J(y|\theta^*) [s_J(y|\theta^*, \gamma_0) - t_J(y|\theta^*)] dy, \quad \text{and} \quad H_{2B}^{(5)} = -\frac{1}{4} I(\theta^*).$$

We will show that

$$\lim_{\gamma_0 \rightarrow 0} D_{2B}(\tilde{\theta}_{i,B}(\gamma_0)) = H_{2B}^{(2)} + \lim_{\gamma_0 \rightarrow 0} H_{2B}^{(3)} + H_{2B}^{(5)}, \quad \text{where} \quad (26)$$

$$\lim_{\gamma_0 \rightarrow 0} H_{2B}^{(3)} = -\frac{1}{2} \int_{\mathbb{R}} v_J(y|\theta^*) v'_J(y|\theta^*) t_J(y|\theta^*) \left[ g_B^{*\frac{1}{2}}(y) - t_J(y|\theta^*) \right] dy. \quad (27)$$

First consider  $H_{2B}^{(1)}$ . It converges to zero as  $\gamma_0 \rightarrow 0$  by Cauchy-Schwarz inequality and assumption (O1). Next consider  $H_{2B}^{(3)}$ . By similar argument as above and boundedness of  $v_J^2(y|\theta^*) t_J(y|\theta^*)$ , it follows that (27) holds. Next consider  $H_{2B}^{(4)}$ . It converges to zero as  $\gamma_0 \rightarrow 0$  by Cauchy-Schwarz inequality and assumption (M3'). Now let  $B \rightarrow \infty$ , we will show that  $\lim_{B \rightarrow \infty} H_{2B}^{(2)} = 0$  and  $\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} H_{2B}^{(3)} = 0$ . First consider  $H_{2B}^{(2)}$ . It converges to zero by Cauchy-Schwarz inequality and assumption (M3') as  $B \rightarrow \infty$ . Finally consider  $\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} H_{2B}^{(3)}$ . It converges to zero by Cauchy-Schwarz inequality and  $L_1$  convergence of  $g_B^*(\cdot)$  and  $h^* J(\cdot|\theta^*)$ . Thus  $\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} D_{2B}(\tilde{\theta}_{i,B}(\gamma_0)) = -\frac{1}{4} I(\theta^*)$ . Now letting  $B \rightarrow \infty$ , the proof of (23) follows using arguments similar to the one in Lemma 2.  $\square$

**Lemma 6.** Assume that the conditions (B1)–(B2), (D1')–(D2'), (M1')–(M3'), (M7'), (O1)–(O2), and (V1)–(V2) hold. Then, first letting  $B \rightarrow \infty$ , and then letting  $\gamma_0 \rightarrow 0$ ,

$$4B^{\frac{1}{2}} T_B(\gamma_0) \xrightarrow{d} N(0, I(\theta^*)). \quad (28)$$

**Proof.** First fix  $B$ . We will show that as  $\gamma_0 \rightarrow 0$ ,

$$4B^{\frac{1}{2}} T_B(\gamma_0) \xrightarrow{d} \int_{\mathbb{R}} v_J(y|\theta^*) g_B^*(y) dy.$$

First observe that

$$4B^{\frac{1}{2}} T_B(\gamma_0) - \int_{\mathbb{R}} v_J(y|\theta^*) g_B^*(y) dy = \int_{\mathbb{R}} [u_J(y|\theta^*, \gamma_0) - v_J(y|\theta^*)] g_B^{(i)}(y) dy \quad (29)$$

$$+ \int_{\mathbb{R}} v_J(y|\theta^*) \left[ g_B^{(i)}(y) - g_B^*(y) \right] dy. \quad (30)$$

We will show that the RHS of (29) converges to zero as  $\gamma_0 \rightarrow 0$  and the RHS of (30) converges to zero in probability as  $\gamma_0 \rightarrow 0$ . First consider the RHS of (29). Since

$$\int_{\mathbb{R}} [u_J(y|\theta^*, \gamma_0) - v_J(y|\theta^*)] g_B^{(i)}(y) dy \leq \int_{\mathbb{R}} \sup_y |u_J(y|\theta^*, \gamma_0) - v_J(y|\theta^*)| g_B^{(i)}(y) dy,$$

which converges to zero as  $\gamma_0 \rightarrow 0$  by assumption (V1). Next consider the RHS of (30). Since

$$\int_{\mathbb{R}} v_J(y|\theta^*) \left[ g_B^{(i)}(y) - g_B^*(y) \right] dy = \frac{1}{B} \sum_{l=1}^B \int_{\mathbb{R}} [v_J(Y_{il} + uc_B) - v_J(Y_{il}^* + uc_B)] K(u) du.$$

By assumption (V2), it follows that as  $\gamma_0 \rightarrow 0$ , (30) converges to zero in probability. Now letting  $B \rightarrow \infty$ , we have

$$B^{\frac{1}{2}} \int_{\mathbb{R}} v_J(y|\theta^*) g_B^*(y) dy - B^{\frac{1}{2}} \frac{1}{B} \sum_{l=1}^B v_J(Y_{il}^*|\theta^*) = B^{\frac{1}{2}} \frac{1}{B} \sum_{l=1}^B \int_{\mathbb{R}} (v_J(Y_{il}^* + c_B t|\theta^*) - v_J(Y_{il}^*|\theta^*)) K(t) dt,$$

and

$$\begin{aligned}
E \left[ B^{\frac{1}{2}} \int_{\mathbb{R}} v_J(y|\theta^*) g_B^*(y) dy - B^{\frac{1}{2}} \frac{1}{B} \sum_{l=1}^B v_J(Y_{il}^*|\theta^*) \right]^2 &= E \left[ B^{\frac{1}{2}} \frac{1}{B} \sum_{l=1}^B \int_{\mathbb{R}} (v_J(Y_{il}^* + c_B t|\theta^*) - v_J(Y_{il}^*|\theta^*)) K(t) dt \right]^2 \\
&\leq C E \left[ \int_{\mathbb{R}} (v_J(Y_{i1}^* + c_B t|\theta^*) - v_J(Y_{i1}^*|\theta^*))^2 dt \right] \\
&= C \int_{\mathbb{R}} \int_{\mathbb{R}} (v_J(y + c_B t|\theta^*) - v_J(y|\theta^*))^2 h^{*J}(y|\theta^*) dy dt \\
&\rightarrow 0 \quad \text{as } B \rightarrow \infty,
\end{aligned}$$

where the last convergence follows by assumption (M7'). Hence, using the Central limit theorem for independent and identically distributed random variables it follows that the limiting distribution of  $B^{\frac{1}{2}} \int_{\mathbb{R}} v_J(y|\theta^*) g_B^*(y) dy$  is  $N(0, I(\theta^*))$ , proving the lemma.  $\square$

**Proof of Proposition 4.** The proof of Proposition 4 follows by combining Lemmas 5 and 6.  $\square$

**Lemma 7.** Assume that the conditions (M1')–(M6') and (V1)–(V4) hold. Then,

$$\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} A_{2B}(\gamma_0) = 0 \quad \text{in probability.}$$

**Proof.** First fix  $B$ . Let

$$\mathcal{H}_B(\gamma_0) = \int_{\mathbb{R}} u_J(y|\theta^*, \gamma_0) \left[ h_J^{\frac{1}{2}}(y|\theta^*, \gamma_0) - g_B^{(i)}(y) \right]^2 dy - \int_{\mathbb{R}} v_J(y|\theta^*) [t_J(y|\theta^*) - g_B^*(y)]^2 dy.$$

we will show that as  $\gamma_0 \rightarrow 0$ ,  $\mathcal{H}_B(\gamma_0) \rightarrow 0$ . By algebra,  $\mathcal{H}_B(\gamma_0)$  can be written as the sum of  $\mathcal{H}_{1B}(\gamma_0)$  and  $\mathcal{H}_{2B}(\gamma_0)$ , where

$$\mathcal{H}_{1B}(\gamma_0) = \int_{\mathbb{R}} (u_J(y|\theta^*, \gamma_0) - v_J(y|\theta^*)) \left[ h_J^{\frac{1}{2}}(y|\theta^*, \gamma_0) - g_B^{(i)}(y) \right]^2 dy, \quad \text{and}$$

$$\mathcal{H}_{2B}(\gamma_0) = \int_{\mathbb{R}} v_J(y|\theta^*) \left[ h_J^{\frac{1}{2}}(y|\theta^*, \gamma_0) - g_B^{(i)}(y) \right]^2 dy.$$

First consider  $\mathcal{H}_{1B}(\gamma_0)$ . It is bounded above by  $C \sup_y |u_J(y|\theta^*, \gamma_0) - v_J(y|\theta^*)|$ , which converges to zero as  $\gamma_0 \rightarrow 0$  by assumption (V1), where  $C$  is a constant. Next consider  $\mathcal{H}_{2B}(\gamma_0)$ . We will show that  $\mathcal{H}_{2B}(\gamma_0)$  converges to

$$\int_{\mathbb{R}} v_J(y|\theta^*) [t_J(y|\theta^*, \gamma_0) - g_B^*(y)]^2 dy.$$

In fact, the difference of  $\mathcal{H}_{2B}(\gamma_0)$  and the above formula can be expressed as the sum of  $\mathcal{H}_{2B}^{(1)}(\gamma_0)$ ,  $\mathcal{H}_{2B}^{(2)}(\gamma_0)$ , and  $\mathcal{H}_{2B}^{(3)}(\gamma_0)$ , where

$$\mathcal{H}_{2B}^{(1)}(\gamma_0) = \int_{\mathbb{R}} v_J(y|\theta^*) \left( h_J(y|\theta^*, \gamma_0) - h^{*J}(y|\theta^*) \right) dy,$$

$$\mathcal{H}_{2B}^{(2)}(\gamma_0) = \int_{\mathbb{R}} v_J(y|\theta^*) \left( g_B^{(i)}(y) - g_B^*(y) \right) dy, \quad \text{and}$$

$$\mathcal{H}_{2B}^{(3)}(\gamma_0) = \int_{\mathbb{R}} v_J(y|\theta^*) \left( h_J^{\frac{1}{2}}(y|\theta^*, \gamma_0) g_B^{(i)}(y) - t_J(y|\theta^*, \gamma_0) g_B^*(y) \right) dy.$$

First consider  $\mathcal{H}_{2B}^{(1)}(\gamma_0)$ . Please note that

$$\begin{aligned} |\mathcal{H}_{2B}^{(1)}(\gamma_0)| &\leq \int_{\mathbb{R}} |\nabla h^{*J}(y|\boldsymbol{\theta}^*)| \left| \frac{h_J(y|\boldsymbol{\theta}^*, \gamma_0)}{h^{*J}(y|\boldsymbol{\theta}^*)} - 1 \right| dy \\ &\leq \left\{ \left( \sup_y \left| \frac{s_J(y|\boldsymbol{\theta}, \gamma_0)}{t_J(y|\boldsymbol{\theta})} - 1 \right| \right)^2 + 2 \sup_y \left| \frac{s_J(y|\boldsymbol{\theta}, \gamma_0)}{t_J(y|\boldsymbol{\theta})} - 1 \right| \right\} \int_{\mathbb{R}} |\nabla h^{*J}(y|\boldsymbol{\theta}^*)| dy, \end{aligned}$$

which converges to 0 as  $\gamma_0 \rightarrow 0$  by assumptions (V3) and (V4). Next we consider  $\mathcal{H}_{2B}^{(2)}(\gamma_0)$ . Since

$$\mathcal{H}_{2B}^{(2)}(\gamma_0) = \frac{1}{B} \sum_{l=1}^B \int_{\mathbb{R}} (v_J(Y_{il} + uc_B|\boldsymbol{\theta}^*) - v_J(Y_{il}^* + uc_B|\boldsymbol{\theta}^*)) K(u) du,$$

which converges to zero as  $\gamma_0 \rightarrow 0$  due to assumption (V2). Finally consider  $\mathcal{H}_{2B}^{(3)}(\gamma_0)$ , which can be expressed as the sum of  $\mathcal{L}_{1B}(\gamma_0)$  and  $\mathcal{L}_{2B}$ , where

$$\mathcal{L}_{1B}(\gamma_0) = \int_{\mathbb{R}} v_J(y|\boldsymbol{\theta}^*) \left( h_J^{\frac{1}{2}}(y|\boldsymbol{\theta}^*, \gamma_0) - t_J(y|\boldsymbol{\theta}^*) \right) g_B^{(i)\frac{1}{2}}(y) dy, \quad \text{and}$$

$$\mathcal{L}_{2B} = \int_{\mathbb{R}} v_J(y|\boldsymbol{\theta}^*) t_J(y|\boldsymbol{\theta}^*) \left( g_B^{(i)\frac{1}{2}}(y) - g_B^{*\frac{1}{2}}(y) \right) dy.$$

First consider  $\mathcal{L}_{1B}(\gamma_0)$ . Notice that

$$|\mathcal{L}_{1B}(\gamma_0)| \leq \sup_y \left| \frac{s_J(y|\boldsymbol{\theta}, \gamma_0)}{t_J(y|\boldsymbol{\theta})} - 1 \right| \int_{\mathbb{R}} v_J(y|\boldsymbol{\theta}^*) t_J(y|\boldsymbol{\theta}^*) g_B^{(i)\frac{1}{2}}(y) dy \rightarrow 0,$$

where the last convergence follows by Cauchy-Schwarz inequality and assumption (V4). Next we consider  $\mathcal{L}_{2B}$ . By Cauchy-Schwarz inequality, it is bounded above by

$$\left\{ \int_{\mathbb{R}} v_J(y|\boldsymbol{\theta}^*) v'_J(y|\boldsymbol{\theta}^*) h^{*J}(y|\boldsymbol{\theta}^*) dy \right\}^{\frac{1}{2}} \left\{ \int_{\mathbb{R}} \left( g_B^{(i)\frac{1}{2}}(y) - g_B^{*\frac{1}{2}}(y) \right)^2 dy \right\}^{\frac{1}{2}}. \quad (31)$$

Equation (31) converges to zero as  $\gamma_0 \rightarrow 0$  by boundedness of  $\int_{\mathbb{R}} v_J(y|\boldsymbol{\theta}^*) v'_J(y|\boldsymbol{\theta}^*) h^{*J}(y|\boldsymbol{\theta}^*) dy$  and  $L_1$  convergence between  $g_B^{(i)}(\cdot)$  and  $g_B^*(\cdot)$ , where the  $L_1$  convergence has already been established in Lemma 5. Now letting  $B \rightarrow \infty$ , following similar argument as Lemma 4 and assumptions (M1')–(M6'), the lemma follows.  $\square$

**Proof of Theorem 2.** Recall that

$$B^{\frac{1}{2}} (\hat{\theta}_{i,B}(\gamma_0) - \boldsymbol{\theta}^*(\gamma_0))' = A_{1B}(\gamma_0) + A_{2B}(\gamma_0).$$

Proposition 4 shows that first letting  $\gamma_0 \rightarrow 0$ , then  $B \rightarrow \infty$ ,  $A_{1B}(\gamma_0) \xrightarrow{d} N(0, I^{-1}(\boldsymbol{\theta}^*))$ ; while Lemma 7 shows that  $\lim_{B \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} A_{2B}(\gamma_0) = 0$  in probability. The theorem follows from Slutsky's theorem.  $\square$

**Remark 3.** The above two theorems (Theorems 1 and 2) do not immediately imply the double limit exists. This requires stronger conditions and more delicate calculations and will be considered elsewhere.

### 3.5. Robustness of MHDE

In this section, we describe the robustness properties of MHDE for compressed data. Accordingly, let  $h_{J,\alpha,z}(\cdot|\theta, \gamma_0) \equiv (1-\alpha)h_J(\cdot|\theta, \gamma_0) + \alpha\eta_z$ , where  $\eta_z$  denotes the uniform density on the interval  $(z-\epsilon, z+\epsilon)$ , where  $\epsilon > 0$  is small,  $\theta \in \Theta$ ,  $\alpha \in (0, 1)$ , and  $z \in \mathbb{R}$ . Also, let  $s_{J,\alpha,z}(y|\theta, \gamma_0) = h_{J,\alpha,z}^{\frac{1}{2}}(y|\theta, \gamma_0)$ ,  $u_{J,\alpha,z}(y|\theta, \gamma_0) = \nabla \log h_{J,\alpha,z}(y|\theta, \gamma_0)$ ,  $h_{\alpha,z}^{*J}(\cdot|\theta) \equiv (1-\alpha)h^{*J}(\cdot|\theta) + \alpha\eta_z$ ,  $s_{\alpha,z}^{*J}(\cdot|\theta) = h_{\alpha,z}^{*J\frac{1}{2}}(\cdot|\theta)$ , and  $u_{\alpha,z}^{*J} = \nabla \log h_{\alpha,z}^{*J}(\cdot|\theta)$ . Before we state the theorem, we describe certain additional assumptions—which are essentially  $L_2$ —continuity conditions—that are needed in the proof.

#### Model assumptions for robustness analysis

**(O3)** For  $\alpha \in [0, 1]$  and all  $\theta \in \Theta$ ,

$$\lim_{\gamma_0 \rightarrow 0} \int_{\mathbb{R}} \left( \dot{u}_{J,\alpha,z}(y|\theta, \gamma_0) s_{J,\alpha,z}(y|\theta, \gamma_0) - \dot{u}_{\alpha,z}^{*J}(y|\theta) s_{\alpha,z}^{*J}(y|\theta) \right)^2 dy = 0.$$

**(O4)** For  $\alpha \in [0, 1]$  and all  $\theta \in \Theta$ ,

$$\lim_{\gamma_0 \rightarrow 0} \int_{\mathbb{R}} \left( u_{J,\alpha,z}(y|\theta, \gamma_0) u'_{J,\alpha,z}(y|\theta, \gamma_0) s_{J,\alpha,z}(y|\theta, \gamma_0) - u_{\alpha,z}^{*J}(y|\theta) u_{\alpha,z}^{*J'}(y|\theta) s_{\alpha,z}^{*J}(y|\theta) \right)^2 dy = 0.$$

**Theorem 3.** (i) Let  $\alpha \in (0, 1)$ , and assume that for all  $\theta \in \Theta$ , and assume that the assumptions of Proposition 1 hold, also assume that  $T(h_{J,\alpha,z}(\cdot|\theta, \gamma_0))$  is unique for all  $z$ . Then,  $T(h_{J,\alpha,z}(\cdot|\theta, \gamma_0))$  is a bounded, continuous function of  $z$  and

$$\lim_{\gamma_0 \rightarrow 0} \lim_{|z| \rightarrow \infty} T(h_{J,\alpha,z}(\cdot|\theta, \gamma_0)) = \theta; \quad (32)$$

(ii) Assume further that the conditions (V1), (M2)-(M3), and (O3)-(O4) hold. Then,

$$\lim_{\gamma_0 \rightarrow 0} \lim_{\alpha \rightarrow 0} \alpha^{-1} [T(h_{J,\alpha,z}(\cdot|\theta, \gamma_0)) - \theta] = [I(\theta)]^{-1} \int_{\mathbb{R}} [\eta_z(y) v_I(y|\theta)] dy,$$

**Proof.** Let  $\theta_z(\gamma_0)$  denote  $T(h_{J,\alpha,z}(\cdot|\theta, \gamma_0))$  and let  $\theta_z$  denote  $T(h_{\alpha,z}^{*J}(\cdot|\theta))$ . We first show that (32) holds. Let  $\gamma_0 \geq 0$  be fixed. Then, by triangle inequality,

$$\lim_{|z| \rightarrow \infty} |\theta_z(\gamma_0) - \theta| \leq \lim_{|z| \rightarrow \infty} |\theta_z(\gamma_0) - \theta(\gamma_0)| + \lim_{|z| \rightarrow \infty} |\theta(\gamma_0) - \theta|. \quad (33)$$

We will show that the first term of RHS of (33) is equal to zero. Suppose that it is not zero, without loss of generality, by going to a subsequence if necessary, we may assume that  $\theta_z \rightarrow \theta_1 \neq \theta$  as  $|z| \rightarrow \infty$ . Since  $\theta_z(\gamma_0)$  minimizes  $HD^2(h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_z, \gamma_0))$ , it follows that

$$HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_z, \gamma_0)) \leq HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0)) \quad (34)$$

for every  $\theta' \in \Theta$ . We now show that as  $|z| \rightarrow \infty$ ,

$$HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_z, \gamma_0)) \rightarrow HD^2((1-\alpha)h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_1, \gamma_0)). \quad (35)$$

To this end, note that as  $|z| \rightarrow \infty$ , for every  $y$ ,

$$h_{J,\alpha,z}(y|\theta, \gamma_0) \rightarrow (1-\alpha)h_J(y|\theta, \gamma_0), \quad \text{and} \quad h_J(y|\theta_z, \gamma_0) \rightarrow h_J(y|\theta_1, \gamma_0)$$

Therefore, as  $|z| \rightarrow \infty$ ,

$$\left| HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_z, \gamma_0)) - HD^2((1-\alpha)h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_1, \gamma_0)) \right| \leq 2(Q_1 + Q_2),$$

where

$$Q_1 = \int_{\mathbb{R}} \left| h_{J,\alpha,z}^{\frac{1}{2}}(y|\theta, \gamma_0) - ((1-\alpha)h_J(y|\theta, \gamma_0))^{\frac{1}{2}} \right| (h_J(y|\theta_z, \gamma_0))^{\frac{1}{2}} dy,$$

$$Q_2 = \int_{\mathbb{R}} \left| h_J^{\frac{1}{2}}(y|\theta_z, \gamma_0) - (h_J(y|\theta_1, \gamma_0))^{\frac{1}{2}} \right| ((1-\alpha)h_J(y|\theta, \gamma_0))^{\frac{1}{2}} dy.$$

Now, by Cauchy-Schwarz inequality and Scheffe's theorem, it follows that as  $|z| \rightarrow \infty$ ,  $Q_1 \rightarrow 0$  and  $Q_2 \rightarrow 0$ . Therefore, (35) holds. By Equations (34) and (35), we have

$$HD^2((1-\alpha)h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_1, \gamma_0)) \leq HD^2((1-\alpha)h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0)) \quad (36)$$

for every  $\theta' \in \Theta$ . Now consider

$$HIF(\alpha, h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0)) \equiv \int_{\mathbb{R}} \left( [(1-\alpha)\delta(h_J(\cdot|\theta, \gamma_0), h_J(y|\theta', \gamma_0)) + 1]^{\frac{1}{2}} - 1 \right)^2 h_J(y|\theta', \gamma_0) dy,$$

where  $\delta(h_J(\cdot|\theta, \gamma_0), h_J(y|\theta', \gamma_0)) = \frac{h_J(y|\theta, \gamma_0)}{h_J(y|\theta', \gamma_0)} - 1$ . Since  $G^*(\delta) = \left[ ((1-\alpha)\delta + 1)^{\frac{1}{2}} - 1 \right]^2$  is a non-negative and strictly convex function with  $\delta = 0$  as the unique point of minimum. Hence  $HIF(\alpha, h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0)) > 0$  unless  $\delta(h_J(\cdot|\theta, \gamma_0), h_J(y|\theta', \gamma_0)) = 0$  on a set of Lebesgue measure zero, which by the model identifiability assumption, is true if and only if  $\theta' = \theta$ . Since  $\theta_1 \neq \theta$ , it follows that

$$HIF(\alpha, h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_1, \gamma_0)) > HIF(\alpha, h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)).$$

Since  $HIF(\alpha, h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0)) = HD^2((1-\alpha)h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0)) - \alpha$ . This implies that

$$HD^2((1-\alpha)h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_1, \gamma_0)) > HD^2((1-\alpha)h_J(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0)),$$

which contradicts (36). The continuity of  $\theta_z$  follows from Proposition 2 and the boundedness follows from the compactness of  $\Theta$ . Now let  $\gamma_0 \rightarrow 0$ , the second term of RHS of (33) converges to zero by Proposition 2.

We now turn to part (ii) of the Theorem. First fix  $\gamma_0 \geq 0$ . Since  $\theta_z$  minimizes  $H^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), t(\gamma_0))$  over  $\Theta$ . By Taylor expansion of  $HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_z, \gamma_0))$  around  $\theta$ , we get

$$\begin{aligned} 0 = \nabla HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_z, \gamma_0)) &= HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) \\ &\quad + (\theta_z - \theta)D(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_z^*, \gamma_0)), \end{aligned}$$

where  $\theta_z^*(\gamma_0)$  is a point between  $\theta$  and  $\theta_z$ ,

$$\nabla HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) = \frac{1}{2} \int_{\mathbb{R}} u_{J,\alpha,z}(y|\theta, \gamma_0) s_{J,\alpha,z}(y|\theta, \gamma_0) (s_{J,\alpha,z}(y|\theta, \gamma_0) - s_J(y|\theta, \gamma_0)) dy, \quad (37)$$

and  $D(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0))$  can be expressed the sum of  $D_1(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0))$  and  $D_2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0))$ , where

$$D_1(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0)) = \frac{1}{2} \int_{\mathbb{R}} \dot{u}_{J,\alpha,z}(y|\theta, \gamma_0) s_{J,\alpha,z}(y|\theta, \gamma_0) s_J(y|\theta', \gamma_0) dy \quad \text{and} \quad (38)$$

$$D_2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta', \gamma_0)) = \frac{1}{4} \int_{\mathbb{R}} u_{J,\alpha,z}(y|\theta, \gamma_0) u'_{J,\alpha,z}(y|\theta, \gamma_0) s_{J,\alpha,z}(y|\theta, \gamma_0) s_J(y|\theta', \gamma_0) dy. \quad (39)$$

Therefore,

$$\alpha^{-1}(\theta_z - \theta) = -\alpha^{-1}D^{-1}(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_z^*, \gamma_0))\nabla HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)).$$

We will show that

$$\lim_{\alpha \rightarrow 0} D(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_z^*, \gamma_0)) = -\frac{1}{4}I(\theta(\gamma_0)), \quad \text{and} \quad (40)$$

$$\lim_{\alpha \rightarrow 0} \alpha^{-1}\nabla HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) = \frac{1}{4} \int_{\mathbb{R}} [\eta_z(y)u_J(y|\theta, \gamma_0)] dy. \quad (41)$$

We will first establish (40). Please note that as  $\alpha \rightarrow 0$ , by definition  $\theta_z(\alpha) \rightarrow \theta$ . Thus,  $\lim_{\alpha \rightarrow 0} \theta_z^*(\alpha) = \theta$ . In addition, by assumptions (O3) and (O4),  $D(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta_z, \gamma_0))$  is continuous in  $\theta_z$ . Therefore, to prove (40), it suffices to show that

$$\lim_{\alpha \rightarrow 0} D_1(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) = -\frac{1}{2} \int_{\mathbb{R}} \dot{u}_{J,\alpha,z}(y|\theta, \gamma_0)h_J(y|\theta, \gamma_0)dy = -\frac{1}{2}I(\theta(\gamma_0)), \quad \text{and} \quad (42)$$

$$\lim_{\alpha \rightarrow 0} D_2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) = \frac{1}{4} \int_{\mathbb{R}} u_{J,\alpha,z}(y|\theta, \gamma_0)u'_{J,\alpha,z}(y|\theta, \gamma_0)h_J(y|\theta, \gamma_0)dy = \frac{1}{4}I(\theta(\gamma_0)). \quad (43)$$

We begin with  $D_2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0))$ . Notice that

$$\lim_{\alpha \rightarrow 0} s_{J,\alpha,z}(y|\theta, \gamma_0) = s_J(y|\theta, \gamma_0), \quad \lim_{\alpha \rightarrow 0} u_{J,\alpha,z}(y|\theta, \gamma_0) = u_J(y|\theta, \gamma_0), \quad \text{and}$$

$$\lim_{\alpha \rightarrow 0} \dot{u}_{J,\alpha,z}(y|\theta, \gamma_0) = \dot{u}_J(y|\theta, \gamma_0).$$

Thus,

$$\lim_{\alpha \rightarrow 0} \dot{u}_{J,\alpha,z}(y|\theta, \gamma_0)s_{J,\alpha,z}(y|\theta, \gamma_0)s_J(y|\theta, \gamma_0) = \dot{u}_J(y|\theta, \gamma_0)h_J(y|\theta, \gamma_0).$$

In addition, in order to pass the limit inside the integral, note that, for every component of matrix  $u_{J,\alpha,z}(\cdot|\theta, \gamma_0)u'_{J,\alpha,z}(\cdot|\theta, \gamma_0)$ , we have

$$\begin{aligned} |u_{J,\alpha,z}(y|\theta, \gamma_0)u'_{J,\alpha,z}(y|\theta, \gamma_0)| &= \left| \left( \frac{(1-\alpha)\nabla h_{J,\alpha,z}(y|\theta, \gamma_0)}{(1-\alpha)h_{J,\alpha,z}(y|\theta, \gamma_0) + \alpha\eta_z(y)} \right) \left( \frac{(1-\alpha)\nabla h_{J,\alpha,z}(y|\theta, \gamma_0)}{(1-\alpha)h_{J,\alpha,z}(y|\theta, \gamma_0) + \alpha\eta_z(y)} \right)' \right| \\ &= \left| \left( \frac{\nabla h_{J,\alpha,z}(y|\theta, \gamma_0)}{h_{J,\alpha,z}(y|\theta, \gamma_0) + \frac{\alpha}{1-\alpha}\eta_z(y)} \right) \left( \frac{\nabla h_{J,\alpha,z}(y|\theta, \gamma_0)}{h_{J,\alpha,z}(y|\theta, \gamma_0) + \frac{\alpha}{1-\alpha}\eta_z(y)} \right)' \right| \\ &\leq \left| \left( \frac{\nabla h_{J,\alpha,z}(y|\theta, \gamma_0)}{h_{J,\alpha,z}(y|\theta, \gamma_0)} \right) \left( \frac{\nabla h_{J,\alpha,z}(y|\theta, \gamma_0)}{h_{J,\alpha,z}(y|\theta, \gamma_0)} \right)' \right| = |u_J(y|\theta, \gamma_0)u'_J(y|\theta, \gamma_0)|, \end{aligned}$$

where  $|\cdot|$  represents the absolute function for each component of the matrix, and

$$|s_{J,\alpha,z}(y|\theta, \gamma_0)| \leq [h_J(y|\theta, \gamma_0) + \eta_z(y)]^{\frac{1}{2}}.$$

Now choosing the dominating function

$$m_J^{(1)}(y|\theta, \gamma_0) = \left| u_J(y|\theta, \gamma_0)u'_J(y|\theta, \gamma_0) \right| [h_J(y|\theta, \gamma_0) + \eta_z(y)]^{\frac{1}{2}} s_J(y|\theta, \gamma_0)$$

and applying Cauchy-Schwarz inequality, we obtain that there exists a constant  $C$  such that

$$\int_{\mathbb{R}} \left| m_J^{(1)}(y|\boldsymbol{\theta}, \gamma_0) \right| dy \leq C \left\{ \int_{\mathbb{R}} \left( u_J(y|\boldsymbol{\theta}, \gamma_0) u'_J(y|\boldsymbol{\theta}, \gamma_0) s_J(y|\boldsymbol{\theta}, \gamma_0) \right)^2 dy \right\}^{\frac{1}{2}},$$

which is finite by assumption **(M2)**. Hence, by the dominated convergence theorem, (43) holds. Turning to (42), notice that for each component of the matrix  $\dot{u}_{J,\alpha,z}(y|\boldsymbol{\theta}, \gamma_0)$ ,

$$\begin{aligned} |\dot{u}_{J,\alpha,z}(y|\boldsymbol{\theta}, \gamma_0)| &= \left| \frac{\ddot{h}_J(y|\boldsymbol{\theta}, \gamma_0) [h_J(y|\boldsymbol{\theta}, \gamma_0) + \frac{\alpha}{1-\alpha} \eta_z(y)] - (\nabla h_J(y|\boldsymbol{\theta}, \gamma_0)) (\nabla h_J(y|\boldsymbol{\theta}, \gamma_0))'}{(h_J(y|\boldsymbol{\theta}, \gamma_0) + \frac{\alpha}{1-\alpha} \eta_z(y))^2} \right| \\ &\leq \left| \frac{\ddot{h}_J(y|\boldsymbol{\theta}, \gamma_0)}{h_J(y|\boldsymbol{\theta}, \gamma_0)} \right| + \left| \frac{(\nabla h_J(y|\boldsymbol{\theta}, \gamma_0)) (\nabla h_J(y|\boldsymbol{\theta}, \gamma_0))'}{h_J^2(y|\boldsymbol{\theta}, \gamma_0)} \right|, \end{aligned}$$

where  $|\cdot|$  denotes the absolute function for each component. Now choosing the dominating function

$$m_J^{(2)}(y|\boldsymbol{\theta}, \gamma_0) = \left( \left| \frac{\ddot{h}_J(y|\boldsymbol{\theta}, \gamma_0)}{h_J(y|\boldsymbol{\theta}, \gamma_0)} \right| + \left| \frac{(\nabla h_J(y|\boldsymbol{\theta}, \gamma_0)) (\nabla h_J(y|\boldsymbol{\theta}, \gamma_0))'}{h_J^2(y|\boldsymbol{\theta}, \gamma_0)} \right| \right) [h_J(y|\boldsymbol{\theta}, \gamma_0) + \eta_z(y)]^{\frac{1}{2}} s_J(y|\boldsymbol{\theta}, \gamma_0),$$

and applying the Cauchy-Schwarz inequality it follows, using **(M3)**, that

$$\int_{\mathbb{R}} \left| m_J^{(2)}(y|\boldsymbol{\theta}, \gamma_0) \right| dy < \infty.$$

Finally, by the dominated convergence theorem, it follows that

$$\lim_{\alpha \rightarrow 0} D_1(h_{J,\alpha,z}(\cdot|\boldsymbol{\theta}, \gamma_0), h_J(\cdot|\boldsymbol{\theta}_z^*, \gamma_0)) = -\frac{1}{2} I(\boldsymbol{\theta}(\gamma_0)).$$

Therefore (40) follows. It remains to show that (41) holds. To this end, note that

$$\nabla H D^2(h_{J,\alpha,z}(\cdot|\boldsymbol{\theta}, \gamma_0), h_J(\cdot|\boldsymbol{\theta}, \gamma_0)) = -\frac{1}{2} \int_{\mathbb{R}} s_{J,\alpha,z}(y|\boldsymbol{\theta}, \gamma_0) u_{J,\alpha,z}(y|\boldsymbol{\theta}, \gamma_0) s_J(y|\boldsymbol{\theta}, \gamma_0) dy.$$

Now taking partial derivative of  $H D^2(h_{J,\alpha,z}(\cdot|\boldsymbol{\theta}, \gamma_0), h_J(\cdot|\boldsymbol{\theta}, \gamma_0))$  with respect to  $\alpha$ , it can be expressed as the sum of  $\mathcal{U}_1$ ,  $\mathcal{U}_2$  and  $\mathcal{U}_3$ , where

$$\mathcal{U}_1 = -\frac{1}{4} \int_{\mathbb{R}} \frac{-h_J(y|\boldsymbol{\theta}, \gamma_0) + \eta_z(y)}{s_{J,\alpha,z}(y|\boldsymbol{\theta}, \gamma_0)} u_{J,\alpha,z}(y|\boldsymbol{\theta}, \gamma_0) s_J(y|\boldsymbol{\theta}, \gamma_0) dy,$$

$$\mathcal{U}_2 = -\frac{1}{2} \int_{\mathbb{R}} s_{J,\alpha,z}(y|\boldsymbol{\theta}, \gamma_0) \frac{-\nabla h_J(y|\boldsymbol{\theta}, \gamma_0) h_{J,\alpha,z}(y|\boldsymbol{\theta}, \gamma_0)}{h_{J,\alpha,z}^2(y|\boldsymbol{\theta}, \gamma_0)} s_J(y|\boldsymbol{\theta}, \gamma_0) dy, \quad \text{and}$$

$$\mathcal{U}_3 = -\frac{1}{2} \int_{\mathbb{R}} s_{J,\alpha,z}(y|\boldsymbol{\theta}, \gamma_0) \frac{-(1-\alpha) \nabla h_J(y|\boldsymbol{\theta}, \gamma_0) (-h_J(y|\boldsymbol{\theta}, \gamma_0) + \eta_z(y))}{h_{J,\alpha,z}^2(y|\boldsymbol{\theta}, \gamma_0)} s_J(y|\boldsymbol{\theta}, \gamma_0) dy.$$

By dominated convergence theorem (using similar idea as above to find dominating functions), we have

$$\lim_{\alpha \rightarrow 0} \frac{\partial \nabla H D^2(h_{J,\alpha,z}(\cdot|\boldsymbol{\theta}, \gamma_0), h_J(\cdot|\boldsymbol{\theta}, \gamma_0))}{\partial \alpha} = \frac{1}{4} \int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}, \gamma_0) \eta_z(y) dy.$$

Hence, by L'Hospital rule, (41) holds. It remains to show that

$$\lim_{\gamma_0 \rightarrow 0} \lim_{\alpha \rightarrow 0} D(h_{J,\alpha,z}(\cdot|\boldsymbol{\theta}, \gamma_0), h_J(\cdot|\boldsymbol{\theta}, \gamma_0)) = -\frac{1}{4} I(\boldsymbol{\theta}), \quad \text{and} \quad (44)$$

$$\lim_{\gamma_0 \rightarrow 0} \lim_{\alpha \rightarrow 0} \alpha^{-1} \nabla H D^2(h_{J,\alpha,z}(\cdot|\boldsymbol{\theta}, \gamma_0), h_J(\cdot|\boldsymbol{\theta}, \gamma_0)) = \frac{1}{4} \int_{\mathbb{R}} [\eta_z(y) v_J(y|\boldsymbol{\theta})] dy. \quad (45)$$

We start with (44). Since for fixed  $\gamma_0 \geq 0$ , by the above argument, it follows that

$$\lim_{\alpha \rightarrow 0} D(h_{J,\alpha,z}(\cdot|\boldsymbol{\theta}, \gamma_0), h_J(\cdot|\boldsymbol{\theta}, \gamma_0)) = -\frac{1}{4} I(\boldsymbol{\theta}(\gamma_0)) = -\frac{1}{4} \int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}, \gamma_0) u'_J(y|\boldsymbol{\theta}, \gamma_0) h_J(y|\boldsymbol{\theta}, \gamma_0) dy,$$

it is enough to show

$$\lim_{\gamma_0 \rightarrow 0} \int_{\mathbb{R}} u_J(y|\boldsymbol{\theta}, \gamma_0) u'_J(y|\boldsymbol{\theta}, \gamma_0) h_J(y|\boldsymbol{\theta}, \gamma_0) dy = \int_{\mathbb{R}} v_J(y|\boldsymbol{\theta}) v'_J(y|\boldsymbol{\theta}) h^{*J}(y|\boldsymbol{\theta}) dy,$$

which is proved in Lemma 2. Hence (44) holds. Next we prove (45). By the argument used to establish (40), it is enough to show that

$$\lim_{\gamma_0 \rightarrow 0} \int_{\mathbb{R}} [\eta_z(y) u_J(y|\boldsymbol{\theta}, \gamma_0)] dy = \int_{\mathbb{R}} [\eta_z(y) v_J(y|\boldsymbol{\theta})] dy. \quad (46)$$

However,

$$\int_{\mathbb{R}} \eta_z(y) [u_J(y|\boldsymbol{\theta}, \gamma_0) - v_J(y|\boldsymbol{\theta})] dy \leq \sup_y |u_J(y|\boldsymbol{\theta}, \gamma_0) - v_J(y|\boldsymbol{\theta})|,$$

and the RHS of the above inequality converges to zero as  $\gamma_0 \rightarrow 0$  from assumption (V1). Hence (46) holds. This completes the proof.  $\square$

Our next result is concerned with the behavior of the  $\alpha$ -influence function when  $\gamma_0 \rightarrow 0$  first and then  $|z| \rightarrow \infty$  or  $\alpha \rightarrow 0$ . The following three additional assumptions will be used in the proof of part (ii) of Theorem 4.

#### Model assumptions for robustness analysis

(O5) For  $\alpha \in [0, 1]$  and all  $\boldsymbol{\theta} \in \Theta$ ,  $\dot{u}_{\alpha,z}^{*J}(y|\boldsymbol{\theta}) s_{\alpha,z}^{*J}(y|\boldsymbol{\theta})$  is bounded in  $L_2$ .

(O6) For  $\alpha \in [0, 1]$  and all  $\boldsymbol{\theta} \in \Theta$ ,  $u_{\alpha,z}^{*J}(y|\boldsymbol{\theta}) u_{\alpha,z}^{*J'}(y|\boldsymbol{\theta}) s_{\alpha,z}^{*J}(y|\boldsymbol{\theta})$  is bounded in  $L_2$ .

(O7) For  $\alpha \in [0, 1]$  and all  $\boldsymbol{\theta} \in \Theta$ ,

$$\lim_{\gamma_0 \rightarrow 0} \int_{\mathbb{R}} \left( s_{J,\alpha,z}(y|\boldsymbol{\theta}, \gamma_0) u_{J,\alpha,z}(y|\boldsymbol{\theta}, \gamma_0) - s_{\alpha,z}^{*J}(y|\boldsymbol{\theta}) u_{\alpha,z}^{*J}(y|\boldsymbol{\theta}) \right)^2 dy = 0.$$

**Theorem 4.** (i) Let  $\alpha \in (0, 1)$ , and assume that for all  $\boldsymbol{\theta} \in \Theta$ , assume that the assumptions of Proposition 1 hold, also assume that  $T(h_{J,\alpha,z}(\cdot|\boldsymbol{\theta}, \gamma_0))$  is unique for all  $z$ . Then,  $T(h_{J,\alpha,z}(\cdot|\boldsymbol{\theta}, \gamma_0))$  is a bounded, continuous function of  $z$  such that

$$\lim_{|z| \rightarrow \infty} \lim_{\gamma_0 \rightarrow 0} T(h_{J,\alpha,z}(\cdot|\boldsymbol{\theta}, \gamma_0)) = \boldsymbol{\theta};$$

(ii) Assume further that the conditions (O3)–(O7) hold. Then,

$$\lim_{\alpha \rightarrow 0} \lim_{\gamma_0 \rightarrow 0} \alpha^{-1} [T(h_{J,\alpha,z}(\cdot|\boldsymbol{\theta}, \gamma_0)) - \boldsymbol{\theta}] = [I(\boldsymbol{\theta})]^{-1} \int_{\mathbb{R}} [\eta_z(y) v_J(y|\boldsymbol{\theta})] dy.$$

**Proof.** Let  $\theta_z(\gamma_0)$  denote  $T(h_{J,\alpha,z}(\cdot|\theta, \gamma_0))$  and let  $\theta_z$  denote  $T(h_{\alpha,z}^{*J}(\cdot|\theta))$ . First fix  $z \in \mathbb{R}$ ; then by the triangular inequality,

$$\lim_{\gamma_0 \rightarrow 0} |\theta_z(\gamma_0) - \theta| \leq \lim_{\gamma_0 \rightarrow 0} |\theta_z(\gamma_0) - \theta_z| + \lim_{\gamma_0 \rightarrow 0} |\theta_z - \theta|. \quad (47)$$

The first term of RHS of (47) is equal to zero due to proposition 2. Now let  $|z| \rightarrow \infty$ , then the second term on the RHS of (47) converges to zero using similar argument as Theorem 3 with density converging to  $h_{\alpha,z}^{*J}(\cdot|\theta)$ . This completes the proof of I). Turning to (ii), we will prove that

$$\lim_{\alpha \rightarrow 0} \lim_{\gamma_0 \rightarrow 0} D(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) = -\frac{1}{4} I(\theta), \quad (48)$$

$$\lim_{\alpha \rightarrow 0} \lim_{\gamma_0 \rightarrow 0} \alpha^{-1} \nabla H D^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) = \frac{1}{4} \int_{\mathbb{R}} [\eta_z(y) v_J(y|\theta)] dy. \quad (49)$$

Recall from the proof of part (ii) of Theorem 3 that

$$\begin{aligned} D(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) &= \frac{1}{2} \int_{\mathbb{R}} \dot{u}_{J,\alpha,z}(y|\theta, \gamma_0) s_{J,\alpha,z}(y|\theta, \gamma_0) s_J(y|\theta, \gamma_0) \\ &\quad + \frac{1}{4} \int_{\mathbb{R}} u_{J,\alpha,z}(y|\theta, \gamma_0) u'_{J,\alpha,z}(y|\theta, \gamma_0) s_{J,\alpha,z}(y|\theta, \gamma_0) s_J(y|\theta, \gamma_0) \\ &\equiv D_1(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) + D_2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)). \end{aligned}$$

We will now show that for fixed  $\alpha \in (0, 1)$

$$\lim_{\gamma_0 \rightarrow 0} D_1(h_{J,\alpha,z}(\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) = \frac{1}{2} \int_{\mathbb{R}} \dot{u}_{\alpha,z}^{*J}(y|\theta) s_{\alpha,z}^{*J}(y|\theta) t_J(y|\theta) dy, \quad \text{and} \quad (50)$$

$$\lim_{\gamma_0 \rightarrow 0} D_2(h_{J,\alpha,z}(\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) = \frac{1}{4} \int_{\mathbb{R}} u_{\alpha,z}^{*J}(y|\theta) u_{\alpha,z}^{*J'}(y|\theta) s_{\alpha,z}^{*J}(y|\theta) t_J(y|\theta) dy. \quad (51)$$

We begin with (50). A standard calculation shows that  $D_1(h_{J,\alpha,z}(\theta, \gamma_0), u_{\alpha,z}^{*J}(y|\theta))$  can be expressed as the sum of  $\mathcal{D}_{11}$ ,  $\mathcal{D}_{12}$  and  $\mathcal{D}_{13}$ , where

$$\mathcal{D}_{11} = \frac{1}{2} \int_{\mathbb{R}} \left( \dot{u}_{J,\alpha,z}(y|\theta, \gamma_0) s_{J,\alpha,z}(y|\theta, \gamma_0) - \dot{u}_{\alpha,z}^{*J}(y|\theta) s_{\alpha,z}^{*J}(y|\theta) \right) s_J(y|\theta, \gamma_0) dy,$$

$$\mathcal{D}_{12} = \frac{1}{2} \int_{\mathbb{R}} \dot{u}_{\alpha,z}^{*J}(y|\theta) s_{\alpha,z}^{*J}(y|\theta) (s_J(y|\theta, \gamma_0) - t_J(y|\theta)) dy, \quad \text{and}$$

$$\mathcal{D}_{13} = \frac{1}{2} \int_{\mathbb{R}} \dot{u}_{\alpha,z}^{*J}(y|\theta) s_{\alpha,z}^{*J}(y|\theta) t_J(y|\theta) dy.$$

It can be seen that  $\mathcal{D}_{11}$  converges to zero as  $\gamma_0 \rightarrow 0$  by Cauchy-Schwarz inequality and assumption (O3); also,  $\mathcal{D}_{12}$  converges to zero as  $\gamma_0 \rightarrow 0$  by Cauchy-Schwarz inequality, assumption (O5) and Scheffe's theorem. Hence (50) follows. Similarly (51) follows as  $\gamma_0 \rightarrow 0$  by Cauchy-Schwarz inequality, assumption (O4), assumption (O6) and Scheffe's theorem.

Now let  $\alpha \rightarrow 0$ . Using the same idea as in Theorem 3 to find dominating functions, one can apply the dominated convergence Theorem to establish that

$$\lim_{\alpha \rightarrow 0} \frac{1}{2} \int_{\mathbb{R}} \dot{u}_{\alpha,z}^{*J}(y|\theta) s_{\alpha,z}^{*J}(y|\theta) t_J(y|\theta) dy = -\frac{1}{2} I(\theta), \quad \text{and}$$

$$\lim_{\alpha \rightarrow 0} \frac{1}{4} \int_{\mathbb{R}} u_{\alpha,z}^{*J}(y|\theta) u_{\alpha,z}^{*J'}(y|\theta) s_{\alpha,z}^{*J}(y|\theta) t_J(y|\theta) dy = \frac{1}{4} I(\theta).$$

Hence (48) follows. Finally, it remains to establish (49). First fix  $\alpha \in (0, 1)$ ; we will show that

$$\lim_{\gamma_0 \rightarrow 0} \nabla HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0)) = -\frac{1}{2} \int_{\mathbb{R}} s_{\alpha,z}^{*J}(y|\theta) u_{\alpha,z}^{*J}(y|\theta) t_J(y|\theta) dy. \quad (52)$$

Please  $\nabla HD^2(h_{J,\alpha,z}(\cdot|\theta, \gamma_0), h_J(\cdot|\theta, \gamma_0))$  can be expressed as the sum of  $\mathcal{T}_1$ ,  $\mathcal{T}_2$  and  $\mathcal{T}_3$ , where

$$\mathcal{T}_1 = -\frac{1}{2} \int_{\mathbb{R}} \left( s_{J,\alpha,z}(y|\theta, \gamma_0) u_{J,\alpha,z}(y|\theta, \gamma_0) - s_{\alpha,z}^{*J}(y|\theta) u_{\alpha,z}^{*J}(y|\theta) \right) s_J(y|\theta, \gamma_0) dy,$$

$$\mathcal{T}_2 = -\frac{1}{2} \int_{\mathbb{R}} s_{\alpha,z}^{*J}(y|\theta) u_{\alpha,z}^{*J}(y|\theta) (s_J(y|\theta, \gamma_0) - t_J(y|\theta)) dy, \quad \text{and}$$

$$\mathcal{T}_3 = -\frac{1}{2} \int_{\mathbb{R}} s_{\alpha,z}^{*J}(y|\theta) u_{\alpha,z}^{*J}(y|\theta) t_J(y|\theta) dy.$$

It can be seen that  $\mathcal{T}_1$  converges to zero as  $\gamma_0 \rightarrow 0$  by Cauchy-Schwarz inequality and assumption (O7);  $\mathcal{T}_2$  converges to zero as  $\gamma_0 \rightarrow 0$  by Cauchy-Schwarz inequality, boundedness of  $u_{\alpha,z}^{*J}(\cdot) s_{\alpha,z}^{*J}(\cdot)$  in  $L_2$ , and Scheffe's theorem. Therefore, (52) holds. Finally, letting  $\alpha \rightarrow 0$  and using the same idea as in Theorem 3 to find the dominating function, it follows by the dominated convergence theorem and L'Hospital rule that (49) holds. This completes the proof of the Theorem.  $\square$

**Remark 4.** Theorems 3 and 4 do not imply that the double limit exists. This is beyond the scope of this paper.

In the next section, we describe the implementation details and provide several simulation results in support of our methodology.

#### 4. Implementation and Numerical Results

In this section, we apply the proposed MHD based methods to estimate the unknown parameters  $\theta = (\mu, \sigma^2)$  using the compressed data. We set  $J = 10,000$  and  $B = 100$ . All simulations are based on 5000 replications. We consider the Gaussian kernel and Epanechnikov kernel for the nonparametric density estimation. The Gaussian kernel is given by

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

and the Epanechnikov kernel is given by

$$K(x) = \frac{3}{4} (1 - x^2) \mathbf{1}_{(|x| \leq 1)}.$$

We generate  $\mathbf{X}$  and uncontaminated compressed data  $\tilde{\mathbf{Y}}$  in the following way:

- Step 1. Generate  $\mathbf{X}_l$ , where  $X_{jl} \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$ .
- Step 2. Generate  $\mathbf{R}_l$ , where  $r_{ijl} \stackrel{i.i.d.}{\sim} N(1, \gamma_0^2)$ .
- Step 3. Generate the uncontaminated  $\tilde{\mathbf{Y}}_l$  by calculating  $\tilde{\mathbf{Y}}_l = \mathbf{R}_l \mathbf{X}_l$ .

#### 4.1. Objective Function

In practice, we store the compressed data  $(\tilde{\mathbf{Y}}_l, \mathbf{r}_l, \omega_l)$  for all  $1 \leq l \leq B$ . Hence if  $X_{jl}$  follows Normal distribution with mean  $\mu$  and variance  $\sigma^2$ , the form of the marginal density of the compressed data, *viz.*,  $Y_{il}$  is complicated and does not have a closed form expression. However, for large  $J$ , using the local limit theorem its density can be approximated by Gaussian density with mean  $\sqrt{J}\mu$  and variance  $\sigma^2 + \gamma_0^2(\mu^2 + \sigma^2)$ . Hence, we work with  $U_{il}$ , where  $U_{il} = \frac{\tilde{Y}_{il} - \mu r_{il}}{\omega_{il}}$ . Please note that with this transformation,  $E[U_{il}] = 0$  and  $Var[U_{il}] = \sigma^2$ . Hence, the kernel density estimate of the unknown true density is given by

$$g_B^{(i)}(y|\mu) = \frac{1}{Bc_B} \sum_{l=1}^B K\left(\frac{y - U_{il}}{c_B}\right).$$

The difference between the kernel density estimate and the one proposed here is that we include the unknown parameter  $\mu$  in the kernel. Additionally, this allows one to incorporate  $(\mathbf{r}_l, \omega_l)$  into the kernel. Consequently, only the scale parameter  $\sigma$  is part of the parametric model. Using the local limit theorem, we approximate the true parametric model by  $\phi(\cdot|\sigma)$ , where  $\phi(\cdot|\sigma)$  is the density of  $N(0, \sigma^2)$ . Hence, the objective function is

$$\Psi(i, \theta) \equiv \mathcal{A}(g_B^{(i)}(\cdot|\mu), \phi(\cdot|\sigma)) = \int_{\mathbb{R}} g_B^{(i)\frac{1}{2}}(y|\mu) \phi^{\frac{1}{2}}(y|\sigma) dy;$$

and, the estimator is given by

$$\hat{\theta}_B(\gamma_0) = \frac{1}{S} \sum_{i=1}^S \hat{\theta}_{i,B}(\gamma_0), \quad \text{where } \hat{\theta}_{i,B}(\gamma_0) = \underset{\theta \in \Theta}{\operatorname{argmax}} \Psi(i, \theta).$$

It is clear that  $\hat{\theta}_B(\gamma_0)$  is a consistent estimator of  $\theta^*$ . In the next subsection, we use Quasi-Newton method with Broyden-Fletcher-Goldfarb-Shanno (BFGS) update to estimate  $\theta$ . Quasi-Newton method is appealing since (i) it replaces the complicated calculation of the Hessian matrix with an approximation which is easier to compute ( $\Delta_k(\theta)$  given in the next subsection) and (ii) gives more flexible step size  $t$  (compared to the Newton-Raphson method), ensuring that it does not “jump” too far at every step and hence guaranteeing convergence of the estimating equation. The BFGS update ( $H_k$ ) is a popular method for approximating the Hessian matrix via gradient evaluations. The step size  $t$  is determined using Backtracking line search algorithm described in Algorithm 2. The algorithms are given in detail in the next subsection. Our analysis also includes the case where  $S \equiv 1$  and  $r_{ijl} \equiv 1$ . In this case, as explained previously, one obtains significant reduction in storage and computational complexity. Finally, we emphasize here that the density estimate contains  $\mu$  and is not parameter free as is typical in classical MHDE analysis. In the next subsection, we describe an algorithm to implement our method.

#### 4.2. Algorithm

As explained previously, we use the Quasi-Newton Algorithm with BFGS update to obtain  $\hat{\theta}_{\text{MHDE}}$ . To describe this method, consider the objective function (suppressing  $i$ )  $\Psi(\theta)$ , which is twice continuously differentiable. Let the initial value of  $\theta$  be  $\theta^{(0)} = (\mu^{(0)}, \sigma^{(0)})$  and  $H_0 = I$ , where  $I$  is the identity matrix.

**Algorithm 1:** The Quasi-Newton Algorithm.

---

Set  $k = 1$ .

**repeat**

1. Calculate  $\Delta_k(\theta) = -H_{k-1}^{-1}\nabla\Psi(\theta^{(k-1)})$ , where  $\nabla\Psi(y; \theta^{k-1})$  is the first derivative of  $\Psi(\theta)$  with respect to  $\theta$  at  $(k-1)$ th step.
2. Determine the step length parameter  $t$  via backtracking line search.
3. Compute  $\theta^{(k)} = \theta^{(k-1)} + t\Delta_k(\theta)$ .
4. Compute  $H_k$ , where the BFGS update is

$$H_k = H_{k-1} + \frac{q_{k-1}q_{k-1}^T}{q_{k-1}^T d_{k-1}} - \frac{H_{k-1}d_{k-1}d_{k-1}^T H_{k-1}^T}{d_{k-1}^T H_{k-1} d_{k-1}},$$

where

$$d_{k-1} = \theta^{(k)} - \theta^{(k-1)},$$

$$q_{k-1} = \nabla\Psi(\theta^{(k)}) - \nabla\Psi(\theta^{(k-1)}).$$

5. Compute  $e_k = |\Psi(\theta^{(k)}) - \Psi(\theta^{(k-1)})|$ .
6. Set  $k = k + 1$ .

**until**  $(e_k) < \text{threshold}$ .

---

**Remark 5.** In step 1, one can directly use the Inverse update for  $H_k^{-1}$  as follows:

$$H_k^{-1} = \left( I - \frac{d_{k-1}q_{k-1}^T}{q_{k-1}^T d_{k-1}} \right) H_{k-1}^{-1} \left( I - \frac{q_{k-1}d_{k-1}^T}{q_{k-1}^T d_{k-1}} \right) + \frac{d_{k-1}d_{k-1}^T}{q_{k-1}^T d_{k-1}}.$$

**Remark 6.** In step 2, the step size  $t$  should satisfy the Wolfe conditions:

$$\begin{aligned} \Psi\left(y; \theta^{(k)} + t\Delta_k\right) &\leq \Psi\left(\theta^{(k)}\right) + u_1 t \nabla\Psi^T\left(\theta^{(k)}\right) \Delta_k, \\ \nabla\Psi\left(\theta^{(k)} + t\Delta_k\right) &\geq u_2 \nabla\Psi^T\left(\theta^{(k)}\right) \Delta_k, \end{aligned}$$

where  $u_1$  and  $u_2$  are constants with  $0 < u_1 < u_2 < 1$ . The first condition requires that  $t$  sufficiently decrease the objective function. The second condition ensures that the step size is not too small. The Backtracking line search algorithm proceeds as follows (see [26]):

**Algorithm 2:** The Backtracking Line Search Algorithm.

---

Given a descent direction  $\Delta(\theta)$  for  $\Psi$  at  $\theta$ ,  $\zeta \in (0, 0.5)$ ,  $\kappa \in (0, 1)$ .  $t := 1$ .

```

while  $\Psi(\theta + t\Delta\theta) > \Psi(\theta) + \zeta t \nabla\Psi(\theta)^T \Delta\theta$ ,
do
     $t := \kappa t$ .
end while

```

---

### 4.3. Initial Values

The initial value for  $\theta$  are taken to be

$$\begin{aligned}\mu^{(0)} &= \text{median}(\tilde{Y}_{il}) / J, \\ \sigma^{(0)} &= 1.48 \times \text{median}(|\tilde{Y}_{il} - \text{median}(\tilde{Y}_{il})|) / B.\end{aligned}$$

Another choice of the initial value for  $\sigma$  is:

$$\hat{\sigma}^{(0)} = \sqrt{\frac{(\widehat{\text{Var}}[\tilde{Y}_{il}] - \gamma_0^2 \mu)}{\gamma_0^2 + \mu_0^2}}, \quad (53)$$

where  $\widehat{\text{Var}}[\tilde{Y}_{il}]$  is an empirical estimate of the variance of  $\tilde{Y}_1$ .

**Bandwidth Selection:** A key issue in implementing the above method of estimation is the choice of the bandwidth. We express the bandwidth in the form  $h_B = c_B s_B$ , where  $c_B \in \{0.3, 0.4, 0.5, 0.7, 0.9\}$ , and  $s_B$  is set equal to  $1.48 \times \text{median}(|\tilde{Y}_{il} - \text{median}(\tilde{Y}_{il})|) / B$ .

In all the tables below, we report the average (Ave), standard deviation (StD) and mean square error (MSE) to assess the performance of the proposed methods.

### 4.4. Analyses Without Contamination

From Tables 2–5, we let true  $\mu = 2, \sigma = 1$ , and take the kernel to be Gaussian kernel. In Table 2, we compare the estimates of the parameters as the dimension of the compressed data  $S$  increases. In this table, we allow  $S$  to take values in the set  $\{1, 2, 5, 10\}$ . Also, we let the number of groups  $B = 100$ , the bandwidth is chosen to be  $c_B = 0.3$ , and  $\gamma_0 = 0.1$ . In addition, in Table 2,  $S^* = 1$  means that  $S = 1$  with  $\gamma_0 \equiv 0$ .

**Table 2.** MHDE as the dimension  $S$  changes for compressed data  $\tilde{Y}$  using Gaussian kernel.

	$\hat{\mu}$			$\hat{\sigma}$		
	Ave	StD $\times 10^3$	MSE $\times 10^3$	Ave	StD $\times 10^3$	MSE $\times 10^3$
$S^* = 1$	2.000	1.010	0.001	1.016	74.03	5.722
$S = 1$	2.000	1.014	0.001	1.018	74.22	5.844
$S = 2$	2.000	1.005	0.001	1.019	73.81	5.832
$S = 5$	2.000	0.987	0.001	1.017	74.16	5.798
$S = 10$	2.000	0.995	0.001	1.019	71.87	5.525

From Table 2 we observe that as  $S$  increases, the estimates for  $\mu$  and  $\sigma$  remain stable. The case  $S^* = 1$  is interesting, since even by storing the sum we are able to obtain point estimates which are close to the true value. In Table 3, we choose  $S = 1, B = 100$  and  $c_B = 0.3$  and compare the estimates as  $\gamma_0$  changes from 0.01 to 1.00. We can see that as  $\gamma_0$  increases, the estimate for  $\mu$  remains stable, whereas the bias, standard deviation and MSE for  $\sigma$  increase.

**Table 3.** MHDE as  $\gamma_0$  changes for compressed data  $\tilde{Y}$  using Gaussian kernel.

	$\hat{\mu}$			$\hat{\sigma}$		
	Ave	StD $\times 10^3$	MSE $\times 10^3$	Ave	StD $\times 10^3$	MSE $\times 10^3$
$\gamma_0 = 0.00$	2.000	1.010	0.001	1.016	74.03	5.722
$\gamma_0 = 0.01$	2.000	1.017	0.001	1.015	74.83	5.814
$\gamma_0 = 0.10$	2.000	1.023	0.001	1.021	72.80	5.717
$\gamma_0 = 0.50$	2.000	1.119	0.001	1.076	72.59	11.08
$\gamma_0 = 1.00$	2.000	1.399	0.002	1.226	82.21	57.75

In Table 4, we fix  $S = 1, B = 100$  and  $\gamma_0 = 0.1$  and allow the bandwidth  $c_B$  to increase. Also,  $c_B^* = 0.30$  means that the bandwidth is chosen as 0.30 with  $\gamma_0 \equiv 0$ . Notice that in this case when  $c_B = 0.9 B^{\frac{1}{2}} c_B = 9$  while  $B^{\frac{1}{2}} c_B^2 = 8.1$  which is not small as is required in assumption (B2). We notice again that as  $c_B$  decreases, the estimates of  $\mu$  and  $\sigma$  are close to the true value with small MSE and StD.

**Table 4.** MHDE as the bandwidth  $c_B$  changes for compressed data  $\tilde{Y}$  using Gaussian kernel.

	$\hat{\mu}$			$\hat{\sigma}$		
	Ave	StD $\times 10^3$	MSE $\times 10^3$	Ave	StD $\times 10^3$	MSE $\times 10^3$
$c_B^* = 0.30$	2.000	1.010	0.001	1.016	74.03	5.722
$c_B = 0.30$	2.000	1.014	0.001	1.018	74.22	5.844
$c_B = 0.40$	2.000	1.015	0.001	1.063	79.68	10.26
$c_B = 0.50$	2.000	1.014	0.001	1.108	82.33	18.33
$c_B = 0.70$	2.000	1.004	0.001	1.212	93.96	53.64
$c_B = 0.90$	2.000	1.009	0.001	1.346	110.5	132.2

In Table 5, we let  $S = 1, c_B = 0.3$  and  $\gamma_0 = 0.1$  and let the number of groups  $B$  increase. This table implies that as  $B$  increases, the estimate performs better in terms of bias, standard deviation and MSE.

**Table 5.** MHDE as  $B$  changes for compressed data  $\tilde{Y}$  using Gaussian kernel with  $\gamma_0 = 0.1$ .

	$\hat{\mu}$			$\hat{\sigma}$		
	Ave	StD $\times 10^3$	MSE $\times 10^3$	Ave	StD $\times 10^3$	MSE $\times 10^3$
$B = 20$	2.000	2.205	0.005	1.739	378.5	688.6
$B = 50$	2.000	1.409	0.002	1.136	125.2	34.17
$B = 100$	2.000	1.010	0.001	1.016	74.03	5.722
$B = 500$	2.000	0.455	0.000	0.972	32.63	1.873

In Table 6, we set  $\gamma_0 \equiv 0$  and keep other settings same as Table 5. This table implies that as  $B$  increases, the estimate performs better in terms of bias, standard deviation and MSE. Furthermore, the standard deviation and MSE are slightly smaller than the results in Table 5.

**Table 6.** MHDE as  $B$  changes for compressed data  $\tilde{Y}$  using Gaussian kernel with  $\gamma_0 = 0$ .

	$\hat{\mu}$			$\hat{\sigma}$		
	Ave	StD $\times 10^3$	MSE $\times 10^3$	Ave	StD $\times 10^3$	MSE $\times 10^3$
$B = 20$	2.000	2.282	0.005	1.749	381.4	706.0
$B = 50$	2.000	1.440	0.002	1.148	125.2	37.42
$B = 100$	2.000	1.014	0.001	1.018	74.22	5.844
$B = 500$	2.000	0.465	0.000	0.973	31.33	1.692

We next move on to investigating the effect of other sensing variables. In the following table, we use Gamma model to generate the additive matrix  $R_l$ . Specifically, the mean of Gamma random variable is set as  $\alpha_0 \beta_0 = 1$ , and the variance  $var \equiv \alpha_0 \beta_0^2$  is chosen from the set  $\{0, 0.01^2, 0.01, 0.25, 1.00\}$  which are also the variances in Table 3.

From Table 7, notice that using Gamma sensing variable yields similar results as Gaussian sensing variable. Our next example considers the case when the mean of the sensing variable is not equal to one and the sensing variable is taken to have a discrete distribution. Specifically, we use Bernoulli sensing variables with parameter  $p$ . Moreover, we fix  $S = 1$  and let  $pJ = S$ . Therefore  $p = 1/J$ . Hence as  $J$

increases, the variance decreases. Now notice that in this case the mean of sensing variable is  $p$  instead of 1. In addition,  $E[\tilde{Y}_{il}] = \mu$  and  $Var[\tilde{Y}_{il}] = \sigma^2 + \mu^2(1 - \frac{1}{J})$ . Hence we set the initial value as

$$\begin{aligned}\mu^{(0)} &= \text{median}(\tilde{Y}_{il}), \\ \sigma^{(0)} &= 1.48 \times \text{median}(|\tilde{Y}_{il} - \text{median}(\tilde{Y}_{il})|).\end{aligned}$$

Additionally, we take  $B = 100$ ,  $c_B = 0.30$  and  $s_B$  to be  $1.48 \times \text{median}(|\tilde{Y}_{il} - \text{median}(\tilde{Y}_{il})|)$ .

**Table 7.** MHDE as variance changes for compressed data  $\tilde{Y}$  using Gaussian kernel under Gamma sensing variable.

	$\hat{\mu}$			$\hat{\sigma}$		
	Ave	StD $\times 10^3$	MSE $\times 10^3$	Ave	StD $\times 10^3$	MSE $\times 10^3$
$var = 0.00$	2.000	1.010	0.001	1.016	74.03	5.722
$var = 0.01^2$	2.000	1.005	0.001	1.016	74.56	5.806
$var = 0.01$	2.000	1.006	0.001	1.018	73.70	5.762
$var = 0.25$	2.000	1.120	0.001	1.078	73.70	11.56
$var = 1.00$	2.000	1.438	0.001	1.228	81.94	58.48

Table 8 shows that MHD method also performs well with Bernoulli sensing variable, although the bias of  $\sigma$ , standard deviation and mean square error for both estimates are larger than those using Gaussian sensing variable and Gamma sensing variable.

**Table 8.** MHDE as  $J$  changes for compressed data  $\tilde{Y}$  using Gaussian kernel under Bernoulli sensing variable.

	$\hat{\mu}$			$\hat{\sigma}$		
	Ave	StD $\times 10^3$	MSE $\times 10^3$	Ave	StD $\times 10^3$	MSE $\times 10^3$
$J = 10$	2.000	104.9	11.01	1.215	97.78	55.79
$J = 100$	1.998	104.5	10.93	1.201	104.5	51.26
$J = 1000$	1.998	104.7	10.96	1.195	106.6	49.36
$J = 5000$	2.001	103.9	10.80	1.200	105.7	51.20
$J = 10000$	1.996	105.1	11.07	1.196	104.4	49.16

#### 4.5. Robustness and Model Misspecification

In this section, we provide a numerical assessment of the robustness of the proposed methodology. To this end, let

$$f_{\alpha,\eta}(x|\theta) = (1 - \alpha)f(x|\theta) + \alpha\eta(x),$$

where  $\eta(x)$  is a contaminating component,  $\alpha \in [0, 1]$ . We generate the contaminated reduced data  $\mathbf{Y}$  in the following way:

- Step 1. Generate  $\mathbf{X}_l$ , where  $X_{jl} \stackrel{i.i.d.}{\sim} N(2, 1)$ .
- Step 2. Generate  $\mathbf{R}_l$ , where  $r_{ijl} \stackrel{i.i.d.}{\sim} N(1, \gamma_0^2)$ .
- Step 3. Generate uncontaminated  $\tilde{Y}_l$  by calculating  $\tilde{Y}_l = \mathbf{R}_l \mathbf{X}_l$ .
- Step 4. Generate contaminated  $\tilde{Y}_{il}^c$ , where  $\tilde{Y}_{il}^c = \tilde{Y}_{il} + \eta(x)$  with probability  $\alpha$ , and  $\tilde{Y}_{il}^c = \tilde{Y}_{il}$  with probability  $1 - \alpha$ .

In the above description, the contamination with outliers is within blocks. A conceptual issue that one encounters is the meaning of outliers in this setting. Specifically, a data point which is an

outlier in the original data set may not remain an outlier in the reduced data and vice-versa. Hence the concepts such as breakdown point and influence function need to be carefully studied. The tables below present one version of the robustness exhibited by the proposed method. In Tables 9 and 10, we set  $J = 10^4$ ,  $B = 100$ ,  $S = 1$ ,  $\gamma_0 = 0.1$ ,  $c_B = 0.3$ ,  $\eta = 1000$ . In addition,  $\alpha^* = 0$  means that  $\alpha = 0$  with  $\gamma_0 \equiv 0$ .

**Table 9.** MHDE as  $\alpha$  changes for contaminated data  $\tilde{Y}$  using Gaussian kernel.

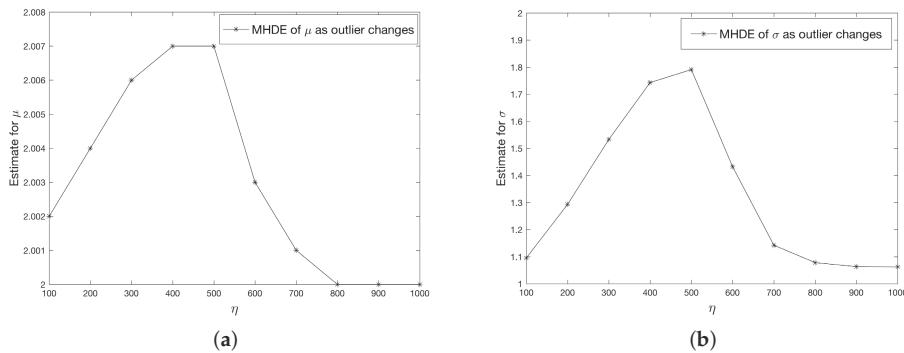
	$\hat{\mu}$			$\hat{\sigma}$		
	Ave	$StD \times 10^3$	$MSE \times 10^3$	Ave	$StD \times 10^3$	$MSE \times 10^3$
$\alpha^* = 0.00$	2.000	1.010	0.001	1.016	74.03	5.722
$\alpha = 0.00$	2.000	1.014	0.001	1.018	74.22	5.844
$\alpha = 0.01$	2.000	1.002	0.001	1.022	74.89	6.079
$\alpha = 0.05$	2.000	1.053	0.001	1.023	77.86	6.599
$\alpha = 0.10$	2.000	1.086	0.001	1.034	79.30	7.350
$\alpha = 0.20$	2.000	1.146	0.001	1.073	93.45	14.06
$\alpha = 0.30$	2.001	7.205	0.054	1.264	688.2	542.5
$\alpha = 0.40$	2.026	21.60	1.100	3.454	1861	9480
$\alpha = 0.50$	2.051	14.00	2.600	4.809	1005	15513

**Table 10.** MHDE as  $\alpha$  changes for contaminated data  $\tilde{Y}$  using Epanechnikov kernel.

	$\hat{\mu}$			$\hat{\sigma}$		
	Ave	$StD \times 10^3$	$MSE \times 10^3$	Ave	$StD \times 10^3$	$MSE \times 10^3$
$\alpha^* = 0.00$	2.000	0.972	0.001	1.008	73.22	5.425
$\alpha = 0.00$	2.000	1.014	0.001	1.018	74.22	5.844
$\alpha = 0.01$	2.000	0.978	0.001	1.028	107.4	12.19
$\alpha = 0.05$	2.000	1.264	0.002	1.025	108.7	12.35
$\alpha = 0.10$	2.000	1.202	0.001	1.008	114.7	13.09
$\alpha = 0.20$	2.000	1.263	0.002	1.046	129.8	18.76
$\alpha = 0.30$	2.001	5.098	0.026	1.104	557.8	318.9
$\alpha = 0.40$	2.021	21.80	0.900	3.004	1973	7870
$\alpha = 0.50$	2.051	10.21	3.000	4.893	720.4	15669

From the above Table we observe that, even under 50% contamination the estimate of the mean remains stable; however, the estimate of the variance is affected at high-levels of contamination (beyond 30%). An interesting and important issue is to investigate the role of  $\gamma_0$  on the breakdown point of the estimator.

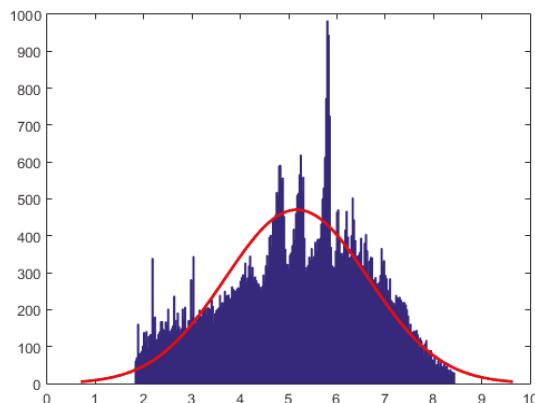
Finally, we investigate the bias in MHDE as a function of the values of the outlier. The graphs below (Figure 2) describe the changes to MHDE when outlier values ( $\eta$ ) increase. Here we set  $S = 1$ ,  $B = 100$ ,  $\gamma_0 = 0.1$ . In addition, we let  $\alpha = 0.2$ , and  $\eta$  to take values from  $\{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ . We can see that as  $\eta$  increases, both  $\hat{\mu}$  and  $\hat{\sigma}$  increase up to  $\eta = 500$  then decrease, although  $\hat{\mu}$  does not change too much. This phenomenon is because when the outlier value is small (or closer to the observations), then it may not be considered as an “outlier” by the MHD method. However, as the outlier values move “far enough” from other values, then the estimate for  $\mu$  and  $\sigma$  remain the stable.



**Figure 2.** Comparison of estimates of  $\mu$  (a) and  $\sigma$  (b) as outlier changes.

## 5. Example

In this section we describe an analysis of data from financial analytics, using the proposed methods. The data are from a bank (a cash and credit card issuer) in Taiwan and the targets of analyses were credit card holders of the bank. The research focused on the case of customers' default payments. The data set (see [27] for details) contains 180,000 observations and includes information on twenty five variables such as default payments, demographic factors, credit data, history of payment, and billing statements of credit card clients from April 2005 to September 2005. Ref. [28] study machine learning methods for evaluating the probability of default. Here, we work with the first three months of data containing 90,000 observations concerning bill payments. For our analyses we remove zero payments and negative payment from the data set and perform a logarithmic transformation of the bill payments. Since the log-transformed data was multi-modal and exhibited features of a mixture of normal distributions, we work with the log-transformed data with values in the range (6.1, 13). Next, we performed the Box-Cox transformation to the log-transformed data. This transformation identifies the best transformation that yields approximately normal distribution (which belongs to the location-scale family). Specifically, let  $L$  denote the log-transformed data in range (6.1, 13), then the data after Box-Cox transformation is given by  $X = (L^2 - 1) / 19.9091$ . The histogram for  $X$  is given in Figure 3. The number of observations at the end of data processing was 70,000.



**Figure 3.** The histogram of credit payment data after Box-Cox transformation to Normality.

Our goal is to estimate the average bill payment during the first three months. For this, we will apply the proposed method. In this analysis, we assume that the target model for  $X$  is Gaussian and split the data, randomly, into  $B = 100$  blocks yielding  $J = 700$  observations per block.

In Table 11, “est” represents the estimator, “95% CI” stands for 95% confidence interval for the estimator. When analyzing the whole data and choosing bandwidth as  $c_n = 0.30$ , we get the MHDE of  $\mu$  to be  $\hat{\mu} = 5.183$  with 95% confidence interval (5.171, 5.194), and the MHDE of  $\sigma$  as  $\hat{\sigma} = 1.425$  with confidence interval (1.418, 1.433).

In Table 11, we choose the bandwidth as  $c_B = 0.30$ . Also,  $S^* = 1$  represents the case where  $S = 1$  and  $\gamma_0 \equiv 0$ . In all other settings, we keep  $\gamma_0 = 0.1$ . We observe that all estimates are similar as  $S$  changes.

**Table 11.** MHDE from the real data analysis.

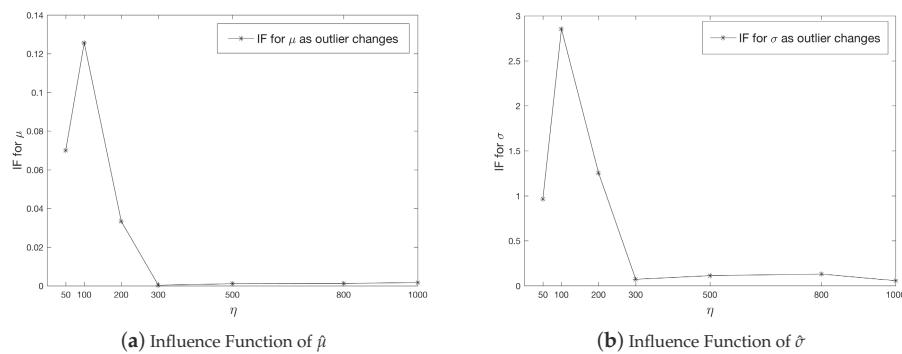
		$\hat{\mu}$	$\hat{\sigma}$
$S^* = 1$	est	5.171	1.362
	95% CI	(4.904, 5.438)	(1.158, 1.540)
$S = 1$	est	5.171	1.391
	95% CI	(4.898, 5.443)	(1.183, 1.572)
$S = 5$	est	5.172	1.359
	95% CI	(4.905, 5.438)	(1.155, 1.535)
$S = 10$	est	5.171	1.372
	95% CI	(4.902, 5.440)	(1.167, 1.551)
$S = 20$	est	5.171	1.388
	95% CI	(4.899, 5.443)	(1.180, 1.569)

Next we study the robustness of MHDE for this data by investigating the relative bias and studying the influence function. Specifically, we first reduce the dimension from  $J = 700$  to  $S = 1$  for each of the  $B = 100$  blocks and obtain the compressed data  $\tilde{Y}$ ; next, we generate the contaminated reduced data  $\tilde{Y}_{il}^c$  from step 4 in Section 4.5. Also, we set  $\alpha = 0.20, \gamma_0 = 0.20$ ; the kernel is taken to be to be Epanechnikov density with bandwidth  $c_B = 0.30$ .  $\eta(x)$  is assumed to takes values in  $\{50, 100, 200, 300, 500, 800, 1000\}$  (note that the approximate mean of  $\tilde{Y}$  is around 3600). Let  $T_{\text{MHD}}$  be the Hellinger distance functional. The influence function given by

$$\text{IF}(\alpha; T, \tilde{Y}) = \frac{T_{\text{MHD}}(\tilde{Y}^c) - T_{\text{MHD}}(\tilde{Y})}{\alpha},$$

which we use to assess the robustness. The graphs shown below (Figure 4) illustrate how the influence function changes as the outlier values increase. We observe that for both estimates ( $\hat{\mu}$  and  $\hat{\sigma}$ ), the influence function first increase and then decrease fast. From  $\eta(x) = 300$ , the influence functions remain stable and are close to zero, which clearly indicate that MHDE is stable.

**Additional Analyses:** The histogram in Figure 3 suggests that, may be a mixture of normal distributions may fit the log and Box-Cox transformed data better than the normal distribution. For this reason, we calculated the Hellinger distance between four component mixture (chosen using BIC criteria) and the normal distribution and this was determined to be 0.0237, approximately. Thus, the normal distribution (which belongs to the location-scale family) can be viewed as a misspecified target distribution; admittedly, one does lose information about the components of the mixture distribution due to model misspecification. However, since our goal was to estimate the overall mean and variance the proposed estimate seems to possess the properties described in the manuscript.

Figure 4. Influence function of  $\hat{\mu}$  (a) and  $\hat{\sigma}$  (b) for MHDE.

## 6. Discussion and Extensions

The results in the manuscript focus on the iterated limit theory for MHDE of the compressed data obtained from a location-scale family. Two pertinent questions arise: (i) is it easy to extend this theory to MHDE of compressed data arising from *non location-scale* family of distributions? and (ii) is it possible to extend the theory from iterated limits to a double limit? Turning to (i), we note that the heuristic for considering the location-scale family comes from the fact that the first and the second moment are consistently estimable for partially observed random walks (see [29,30]). This is related to the size of  $J$  and can be of exponential order. For such large  $J$ , other moments may not be consistently estimable. Hence, the entire theory goes through as long as one is considering parametric models  $f(\cdot|\theta)$ , where  $\theta = \mathcal{W}(\mu, \sigma^2)$ , for a known function  $\mathcal{W}(\cdot, \cdot)$ . The case in point is the Gamma distribution which can be re-parametrized in terms of the first two moments.

As for (ii), it is well-known that existence and equality of iterated limits for real sequences does not imply the existence of the double limit unless additional uniformity of convergence holds (see [31] for instance). Extension of this notion for distributional convergence requires additional assumptions and are investigated in a different manuscript wherein more general divergences are also considered.

## 7. Concluding Remarks

In this paper we proposed the Hellinger distance-based method to obtain robust estimates for mean and variance in a location-scale model using compressed data. Our extensive theoretical investigations and simulations show the usefulness of the methodology and hence can be applied in a variety of scientific settings. Several theoretical and practical questions concerning robustness in a big data setting arise. For instance, the effect of the variability in the  $R$  matrix and its effect on outliers are important issues that need further investigation. Furthermore, statistical properties such as uniform consistency and uniform asymptotic normality under different choices for the distribution of  $R$  would be useful. These are under investigation by the authors.

**Author Contributions:** The problem was conceived by E.A., A.N.V. and G.D. L.L. is a student of A.N.V., and worked on theoretical and simulation details with inputs from all members at different stages.

**Funding:** The authors thank George Mason University Libraries for support with the article processing fees; Ahmed's research is supported by a grant from NSERC.

**Acknowledgments:** The authors thank the anonymous reviewers for a careful reading of the manuscript and several useful suggestions that improved the readability of the paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

MHDE	Minimum Hellinger Distance Estimator
MHD	Minimum Hellinger Distance
i.i.d.	independent and identically distributed
MLE	Maximum Likelihood Estimator
CI	Confidence Interval
IF	Influence Function
RHS	Right Hand Side
LHS	Left Hand Side
BFGS	Broyden-Fletcher-Goldfarb-Shanno
var	Variance
StD	Standard Deviation
MSE	Mean Square Error

## References

1. Beran, R. Minimum Hellinger distance estimates for parametric models. *Ann. Stat.* **1977**, *5*, 445–463. [[CrossRef](#)]
2. Lindsay, B.G. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Stat.* **1994**, *22*, 1081–1114. [[CrossRef](#)]
3. Fisher, R.A. Two new properties of mathematical likelihood. *Proc. R. Soc. Lond. Ser. A* **1934**, *144*, 285–307. [[CrossRef](#)]
4. Pitman, E.J.G. The estimation of the location and scale parameters of a continuous population of any given form. *Biometrika* **1939**, *30*, 391–421. [[CrossRef](#)]
5. Gupta, A.; Székely, G. On location and scale maximum likelihood estimators. *Proc. Am. Math. Soc.* **1994**, *120*, 585–589. [[CrossRef](#)]
6. Duerinckx, M.; Ley, C.; Swan, Y. Maximum likelihood characterization of distributions. *Bernoulli* **2014**, *20*, 775–802. [[CrossRef](#)]
7. Teicher, H. Maximum likelihood characterization of distributions. *Ann. Math. Stat.* **1961**, *32*, 1214–1222. [[CrossRef](#)]
8. Thanei, G.A.; Heinze, C.; Meinshausen, N. Random projections for large-scale regression. In *Big and Complex Data Analysis*; Springer: Berlin, Germany, 2017; pp. 51–68.
9. Slawski, M. Compressed least squares regression revisited. In *Artificial Intelligence and Statistics*; Addison-Wesley: Boston, MA, USA, 2017; pp. 1207–1215.
10. Slawski, M. On principal components regression, random projections, and column subsampling. *Electron. J. Stat.* **2018**, *12*, 3673–3712. [[CrossRef](#)]
11. Raskutti, G.; Mahoney, M.W. A statistical perspective on randomized sketching for ordinary least-squares. *J. Mach. Learn. Res.* **2016**, *17*, 7508–7538.
12. Ahfock, D.; Astle, W.J.; Richardson, S. Statistical properties of sketching algorithms. *arXiv* **2017**, arXiv:1706.03665.
13. Vidyashankar, A.; Hanlon, B.; Lei, L.; Doyle, L. Anonymized Data: Trade off between Efficiency and Privacy. **2018**, preprint.
14. Woodward, W.A.; Whitney, P.; Eslinger, P.W. Minimum Hellinger distance estimation of mixture proportions. *J. Stat. Plan. Inference* **1995**, *48*, 303–319. [[CrossRef](#)]
15. Basu, A.; Harris, I.R.; Basu, S. Minimum distance estimation: The approach using density-based distances. In *Robust Inference, Handbook of Statistics*; Elsevier: Amsterdam, The Netherlands, 1997; Volume 15, pp. 21–48.
16. Hooker, G.; Vidyashankar, A.N. Bayesian model robustness via disparities. *Test* **2014**, *23*, 556–584. [[CrossRef](#)]
17. Sriram, T.; Vidyashankar, A. Minimum Hellinger distance estimation for supercritical Galton–Watson processes. *Stat. Probab. Lett.* **2000**, *50*, 331–342. [[CrossRef](#)]
18. Simpson, D.G. Minimum Hellinger distance estimation for the analysis of count data. *J. Am. Stat. Assoc.* **1987**, *82*, 802–807. [[CrossRef](#)]

19. Simpson, D.G. Hellinger deviance tests: Efficiency, breakdown points, and examples. *J. Am. Stat. Assoc.* **1989**, *84*, 107–113. [[CrossRef](#)]
20. Cheng, A.; Vidyashankar, A.N. Minimum Hellinger distance estimation for randomized play the winner design. *J. Stat. Plan. Inference* **2006**, *136*, 1875–1910. [[CrossRef](#)]
21. Basu, A.; Shiota, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapman and Hall/CRC: London, UK, 2011.
22. Bhandari, S.K.; Basu, A.; Sarkar, S. Robust inference in parametric models using the family of generalized negative exponential dispatches. *Aust. N. Z. J. Stat.* **2006**, *48*, 95–114. [[CrossRef](#)]
23. Ghosh, A.; Harris, I.R.; Maji, A.; Basu, A.; Pardo, L. A generalized divergence for statistical inference. *Bernoulli* **2017**, *23*, 2746–2783. [[CrossRef](#)]
24. Tamura, R.N.; Boos, D.D. Minimum Hellinger distance estimation for multivariate location and covariance. *J. Am. Stat. Assoc.* **1986**, *81*, 223–229. [[CrossRef](#)]
25. Li, P. Estimators and tail bounds for dimension reduction in  $\ell^{\alpha}$  ( $0 < \alpha \leq 2$ ) using stable random projections. In Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, USA, 20–22 January 2008; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2008; pp. 10–19.
26. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
27. Lichman, M. UCI Machine Learning Repository. Available online: <https://archive.ics.uci.edu/ml/index.php> (accessed on 29 March 2019).
28. Yeh, I.C.; Lien, C.H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl.* **2009**, *36*, 2473–2480. [[CrossRef](#)]
29. Guttorp, P.; Lockhart, R.A. Estimation in sparsely sampled random walks. *Stoch. Process. Appl.* **1989**, *31*, 315–320, doi:10.1016/0304-4149(89)90095-1. [[CrossRef](#)]
30. Guttorp, P.; Siegel, A.F. Consistent estimation in partially observed random walks. *Ann. Stat.* **1985**, *13*, 958–969. doi:10.1214/aos/1176349649. [[CrossRef](#)]
31. Apostol, T.M. *Mathematical Analysis*; Addison Wesley Publishing Company: Boston, MA, USA, 1974.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Robustness Property of Robust-BD Wald-Type Test for Varying-Dimensional General Linear Models

Xiao Guo <sup>1,\*</sup> and Chunming Zhang <sup>2</sup>

<sup>1</sup> Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

<sup>2</sup> Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA; cmzhang@stat.wisc.edu

\* Correspondence: xiaoguo@ustc.edu.cn; Tel.: +86-55163603167

Received: 12 January 2018; Accepted: 1 March 2018; Published: 5 March 2018

**Abstract:** An important issue for robust inference is to examine the stability of the asymptotic level and power of the test statistic in the presence of contaminated data. Most existing results are derived in finite-dimensional settings with some particular choices of loss functions. This paper re-examines this issue by allowing for a diverging number of parameters combined with a broader array of robust error measures, called “*robust-BD*”, for the class of “general linear models”. Under regularity conditions, we derive the influence function of the *robust-BD* parameter estimator and demonstrate that the *robust-BD* Wald-type test enjoys the robustness of validity and efficiency asymptotically. Specifically, the asymptotic level of the test is stable under a small amount of contamination of the null hypothesis, whereas the asymptotic power is large enough under a contaminated distribution in a neighborhood of the contiguous alternatives, thus lending supports to the utility of the proposed *robust-BD* Wald-type test.

**Keywords:** Bregman divergence; general linear model; hypothesis testing; influence function; robust; Wald-type test

---

## 1. Introduction

The class of varying-dimensional “general linear models” [1], including the conventional generalized linear model (GLM in [2]), is flexible and powerful for modeling a large variety of data and plays an important role in many statistical applications. In the literature, it has been extensively studied that the conventional maximum likelihood estimator for the GLM is nonrobust; for example, see [3,4]. To enhance the resistance to outliers in applications, many efforts have been made to obtain robust estimators. For example, Noh et al. [5] and Künsch et al. [6] developed robust estimator for the GLM, and Stefanski et al. [7], Bianco et al. [8] and Croux et al. [9] studied robust estimation for the logistic regression model with the deviance loss as the error measure.

Besides robust estimation for the GLM, robust inference is another important issue, which, however, receives relatively less attention. Basically, the study of robust testing includes two aspects: (i) establishing the stability of the asymptotic level under small departures from the null hypothesis (i.e., robustness of “validity”); and (ii) demonstrating that the asymptotic power is sufficiently large under small departures from specified alternatives (i.e., robustness of “efficiency”). In the literature, robust inference has been conducted for different models. For example, Heritier et al. [10] studied the robustness properties of the Wald, score and likelihood ratio tests based on M estimators for general parametric models. Cantoni et al. [11] developed a test statistic based on the robust deviance, and conducted robust inference for the GLM using quasi-likelihood as the loss function. A robust Wald-type test for the logistic regression model is studied in [12]. Ronchetti et al. [13] concerned the robustness property for the generalized method of moments estimators. Basu et al. [14]

proposed robust tests based on the density power divergence (DPD) measure for the equality of two normal means. Robust tests for parameter change have been studied using the density-based divergence method in [15,16]. However, the aforementioned methods based on the GLM mostly focus on situations where the number of parameters is fixed and the loss function is specific.

Zhang et al. [1] developed robust estimation and testing for the “general linear model” based on a broader array of error measures, namely Bregman divergence, allowing for a diverging number of parameters. The Bregman divergence includes a wide class of error measures as special cases, e.g., the (negative) quasi-likelihood in regression, the deviance loss and exponential loss in machine learning practice, among many other commonly used loss functions. Zhang et al. [1] studied the consistency and asymptotic normality of their proposed *robust*-BD parameter estimator and demonstrated the asymptotic distribution of the Wald-type test constructed from *robust*-BD estimators. Naturally, it remains an important issue to examine the robustness property of the *robust*-BD Wald-type test [1] in the varying-dimensional case, i.e., whether the test still has stable asymptotic level and power, in the presence of contaminated data.

This paper aims to demonstrate the robustness property of the *robust*-BD Wald-type test in [1]. Nevertheless, it is a nontrivial task to address this issue. Although the local stability for the Wald-type tests have been established for the M estimators [10], generalized method of moment estimators [13], minimum density power divergence estimator [17] and general M estimators under random censoring [18], their results for finite-dimensional settings are not directly applicable to our situations with a diverging number of parameters. Under certain regularity conditions, we provide rigorous theoretical derivation for robust testing based on the Wald-type test statistic. The essential results are approximations of the asymptotic level and power under contaminated distributions of the data in a small neighborhood of the null and alternative hypotheses, respectively.

- Specifically, we show in Theorem 1 that, if the influence function of the estimator is bounded, then the asymptotic level of the test is also bounded under a small amount of contamination.
- We also demonstrate in Theorem 2 that, if the contamination belongs to a neighborhood of the contiguous alternatives, then the asymptotic power is also stable.

Hence, we contribute to establish the robustness of validity and efficiency for the *robust*-BD Wald-type test for the “general linear model” with a diverging number of parameters.

The rest of the paper is organized as follows. Section 2 reviews the Bregman divergence (BD), *robust*-BD estimation and the Wald-type test statistic proposed in [1]. Section 3 derives the influence function of the *robust*-BD estimator and studies the robustness properties of the asymptotic level and power of the Wald-type test under a small amount of contamination. Section 4 conducts the simulation studies. The technical conditions and proofs are given in Appendix A. A list of notations and symbols is provided in Appendix B.

We will introduce some necessary notations. In the following,  $C$  and  $c$  are generic finite constants which may vary from place to place, but do not depend on the sample size  $n$ . Denote by  $E_K(\cdot)$  the expectation with respect to the underlying distribution  $K$ . For a positive integer  $q$ , let  $\mathbf{0}_q = (0, \dots, 0)^T \in \mathbb{R}^q$  be a  $q \times 1$  zero vector and  $\mathbf{I}_q$  be the  $q \times q$  identity matrix. For a vector  $\mathbf{v} = (v_1, \dots, v_q)^T \in \mathbb{R}^q$ , the  $L_1$  norm is  $\|\mathbf{v}\|_1 = \sum_{i=1}^q |v_i|$ ,  $L_2$  norm is  $\|\mathbf{v}\|_2 = (\sum_{i=1}^q v_i^2)^{1/2}$  and the  $L_\infty$  norm is  $\|\mathbf{v}\|_\infty = \max_{i=1, \dots, q} |v_i|$ . For a  $q \times q$  matrix  $A$ , the  $L_2$  and Frobenius norms of  $A$  are  $\|A\|_2 = \{\lambda_{\max}(A^T A)\}^{1/2}$  and  $\|A\|_F = \sqrt{\text{tr}(A A^T)}$ , respectively, where  $\lambda_{\max}(\cdot)$  denotes the largest eigenvalue of a matrix and  $\text{tr}(\cdot)$  denotes the trace of a matrix.

## 2. Review of Robust-BD Estimation and Inference for “General Linear Models”

This section briefly reviews the *robust*-BD estimation and inference methods for the “general linear model” developed in [1]. Let  $\{(X_{n1}, Y_1), \dots, (X_{nn}, Y_n)\}$  be i.i.d. observations from some underlying distribution  $(X_n, Y)$  with  $X_n = (X_1, \dots, X_{p_n})^T \in \mathbb{R}^{p_n}$  the explanatory variables and  $Y$  the response

variable. The dimension  $p_n$  is allowed to diverge with the sample size  $n$ . The “general linear model” is given by

$$m(\mathbf{x}_n) \equiv E(Y | \mathbf{X}_n = \mathbf{x}_n) = F^{-1}(\tilde{\mathbf{x}}_n^T \tilde{\boldsymbol{\beta}}_{n,0}), \quad (1)$$

and

$$\text{var}(Y | \mathbf{X}_n = \mathbf{x}_n) = V(m(\mathbf{x}_n)), \quad (2)$$

where  $F$  is a known link function,  $\tilde{\boldsymbol{\beta}}_{n,0} \in \mathbb{R}^{p_n+1}$  is the vector of unknown true regression parameters,  $\tilde{\mathbf{x}}_n = (1, \mathbf{x}_n^T)^T$  and  $V(\cdot)$  is a known function. Note that the conventional generalized linear model (GLM) satisfying Equations (1) and (2) assumes that  $Y | \mathbf{X}_n = \mathbf{x}_n$  follows a particular distribution in the exponential family. However, our “general linear model” does not require explicit form of distributions of the response. Hence, the “general linear model” includes the GLM as a special case. For notational simplicity, denote  $\mathbf{Z}_n = (\mathbf{X}_n^T, Y)^T$  and  $\tilde{\mathbf{Z}}_n = (\tilde{\mathbf{X}}_n^T, Y)^T$ .

Bregman divergence (BD) is a class of error measures, which is introduced in [19] and covers a wide range of loss functions. Specifically, Bregman divergence is defined as a bivariate function,

$$Q_q(v, \mu) = -q(v) + q(\mu) + (v - \mu)q'(\mu),$$

where  $q(\cdot)$  is the concave generating  $q$ -function. For example,  $q(\mu) = a\mu - \mu^2$  for a constant  $a$  corresponds to the quadratic loss  $Q_a(Y, \mu) = (Y - \mu)^2$ . For a binary response variable  $Y$ ,  $q(\mu) = \min\{\mu, 1 - \mu\}$  gives the misclassification loss  $Q_q(Y, \mu) = I\{Y \neq I(\mu > 0.5)\}$ ;  $q(\mu) = -2\{\mu \log(\mu) + (1 - \mu) \log(1 - \mu)\}$  gives Bernoulli deviance loss  $Q_q(Y, \mu) = -2\{Y \log(\mu) + (1 - Y) \log(1 - \mu)\}$ ;  $q(\mu) = 2\min\{\mu, 1 - \mu\}$  gives the hinge loss  $Q_q(Y, \mu) = \max\{1 - (2Y - 1)\text{sign}(\mu - 0.5), 0\}$  for the support vector machine;  $q(\mu) = 2\{\mu(1 - \mu)\}^{1/2}$  yields the exponential loss  $Q_q(Y, \mu) = \exp[-(Y - 0.5) \log\{\mu/(1 - \mu)\}]$  used in AdaBoost [20]. Furthermore, Zhang et al. [21] showed that if

$$q(\mu) = \int_a^\mu \frac{s - \mu}{V(s)} ds, \quad (3)$$

where  $a$  is a finite constant such that the integral is well-defined, then  $Q_q(y, \mu)$  is the “classical (negative) quasi-likelihood” function  $-Q_{QL}(y, \mu)$  with  $\partial Q_{QL}(y, \mu)/\partial \mu = (y - \mu)/V(\mu)$ .

To obtain a robust estimator based on BD, Zhang et al. [1] developed the *robust*-BD loss function

$$\rho_q(y, \mu) = \int_y^\mu \psi(r(y, s))\{q''(s)\sqrt{V(s)}\} ds - G(\mu), \quad (4)$$

where  $\psi(\cdot)$  is a bounded odd function, such as the Huber  $\psi$ -function [22],  $r(y, s) = (y - s)/\sqrt{V(s)}$  denotes the Pearson residual and  $G(\mu)$  is the bias-correction term satisfying

$$G'(\mu) = G'_1(\mu)\{q''(\mu)\sqrt{V(\mu)}\},$$

with

$$G'_1(m(\mathbf{x}_n)) = E\{\psi(r(Y, m(\mathbf{x}_n))) | \mathbf{X}_n = \mathbf{x}_n\}.$$

Based on *robust*-BD, the estimator of  $\tilde{\boldsymbol{\beta}}_{n,0}$  proposed in [1] is defined as

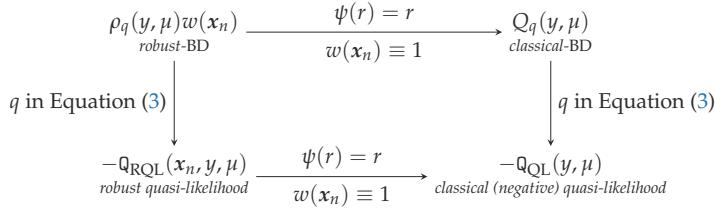
$$\hat{\tilde{\boldsymbol{\beta}}} = \arg \min_{\tilde{\boldsymbol{\beta}}} \left\{ \frac{1}{n} \sum_{i=1}^n \rho_q(Y_i, F^{-1}(\tilde{\mathbf{X}}_{ni}^T \tilde{\boldsymbol{\beta}})) w(\mathbf{X}_{ni}) \right\}, \quad (5)$$

where  $w(\cdot) \geq 0$  is a known bounded weight function which downweights the high leverage points.

In [11], the “robust quasi-likelihood estimator” of  $\hat{\beta}_{n,0}$  is formulated according to the “robust quasi-likelihood function” defined as

$$\begin{aligned} & \mathbb{Q}_{\text{RQL}}(\mathbf{x}_n, y, \mu) \\ &= \left\{ \int_{\mu_0}^{\mu} \psi(r(y, s)) / \sqrt{V(s)} ds \right\} w(\mathbf{x}_n) - \frac{1}{n} \sum_{j=1}^n \int_{\mu_0}^{\mu_j} \left[ E\{\psi(r(Y_j, s)) | \mathbf{X}_{nj}\} / \sqrt{V(s)} ds \right] w(\mathbf{X}_{nj}), \end{aligned}$$

where  $\mu = F^{-1}(\tilde{\mathbf{x}}_n^T \tilde{\beta})$  and  $\mu_j = \mu_j(\tilde{\beta}) = F^{-1}(\tilde{\mathbf{X}}_{nj}^T \tilde{\beta})$ ,  $j = 1, \dots, n$ . To describe the intuition of the “robust-BD”, we use the following diagram from [1], which illustrates the relation among the “robust-BD”, “classical-BD”, “robust quasi-likelihood” and “classical (negative) quasi-likelihood”.



For the *robust-BD*, assume that

$$p_j(y; \theta) = \frac{\partial^j}{\partial \theta^j} \rho_q(y, F^{-1}(\theta)), \quad j = 0, 1, \dots,$$

exist finitely up to any order required. For example, for  $j = 1$ ,

$$p_1(y; \theta) = \{\psi(r(y, \mu)) - G'_1(\mu)\} \{q''(\mu) \sqrt{V(\mu)}\} / F'(\mu), \quad (6)$$

where  $\mu = F^{-1}(\theta)$ . Explicit expressions for  $p_j(y; \theta)$  ( $j = 2, 3$ ) can be found in Equation (3.7) of [1]. Then, the estimation equation for  $\hat{\beta}$  is

$$\frac{1}{n} \sum_{i=1}^n \psi_{\text{RBD}}(\mathbf{Z}_{ni}; \tilde{\beta}) = \mathbf{0},$$

where the score vector is

$$\psi_{\text{RBD}}(z_n; \tilde{\beta}) = p_1(y; \theta) w(\mathbf{x}_n) \tilde{\mathbf{x}}_n, \quad (7)$$

with  $\theta = \tilde{\mathbf{x}}_n^T \tilde{\beta}$ . The consistency and asymptotic normality of  $\hat{\beta}$  have been studied in [1]; see Theorems 1 and 2 therein.

Furthermore, to conduct statistical inference for the “general linear model”, the following hypotheses are considered,

$$H_0 : A_n \tilde{\beta}_{n,0} = \mathbf{g}_0 \quad \text{versus} \quad H_1 : A_n \tilde{\beta}_{n,0} \neq \mathbf{g}_0, \quad (8)$$

where  $A_n$  is a given  $k \times (p_n + 1)$  matrix such that  $A_n A_n^T \rightarrow \mathbb{G}$  with  $\mathbb{G}$  being a  $k \times k$  positive-definite matrix, and  $\mathbf{g}_0$  is a known  $k \times 1$  vector.

To perform the test of Equation (8), Zhang et al. [1] proposed the Wald-type test statistic,

$$W_n = n(A_n \hat{\beta} - \mathbf{g}_0)^T (A_n \hat{\mathbf{H}}_n^{-1} \hat{\Omega}_n \hat{\mathbf{H}}_n^{-1} A_n^T)^{-1} (A_n \hat{\beta} - \mathbf{g}_0), \quad (9)$$

constructed from the *robust*-BD estimator  $\hat{\beta}$  in Equation (5), where

$$\begin{aligned}\hat{\Omega}_n &= \frac{1}{n} \sum_{i=1}^n p_1^2(Y_i; \tilde{X}_{ni}^T \hat{\beta}) w^2(\mathbf{X}_{ni}) \tilde{X}_{ni} \tilde{X}_{ni}^T, \\ \hat{\mathbf{H}}_n &= \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \tilde{X}_{ni}^T \hat{\beta}) w(\mathbf{X}_{ni}) \tilde{X}_{ni} \tilde{X}_{ni}^T.\end{aligned}$$

The asymptotic distributions of  $W_n$  under the null and alternative hypotheses have been developed in [1]; see Theorems 4–6 therein.

On the other hand, the issue on the robustness of  $W_n$ , used for possibly contaminated data, remains unknown. Section 3 of this paper will address this issue with detailed derivations.

### 3. Robustness Properties of $W_n$ in Equation (9)

This section derives the influence function of the *robust*-BD Wald-type test and studies the influence of a small amount of contamination on the asymptotic level and power of the test. The proofs of the theoretical results are given in Appendix A.

Denote by  $K_{n,0}$  the true distribution of  $Z_n$  following the “general linear model” characterized by Equations (1) and (2). To facilitate the discussion of robustness properties, we consider the  $\epsilon$ -contamination,

$$K_{n,\epsilon} = \left(1 - \frac{\epsilon}{\sqrt{n}}\right) K_{n,0} + \frac{\epsilon}{\sqrt{n}} J, \quad (10)$$

where  $J$  is an arbitrary distribution and  $\epsilon > 0$  is a constant. Then,  $K_{n,\epsilon}$  is a contaminated distribution of  $Z_n$  with the amount of contamination converging to 0 at rate  $1/\sqrt{n}$ . Denote by  $\mathbb{K}_n$  the empirical distribution of  $\{Z_{ni}\}_{i=1}^n$ .

For a generic distribution  $K$  of  $Z_n$ , define

$$\begin{aligned}\ell_K(\tilde{\beta}) &= E_K\{\rho_q(Y, F^{-1}(\tilde{X}_n^T \tilde{\beta})) w(\mathbf{X}_n)\}, \\ \mathcal{S}_K &= \{\tilde{\beta} : E_K\{\psi_{\text{RBD}}(Z_n; \tilde{\beta})\} = 0\},\end{aligned} \quad (11)$$

where  $\rho_q(\cdot, \cdot)$  and  $\psi_{\text{RBD}}(\cdot, \cdot)$  are defined in Equations (4) and (7), respectively. It’s worth noting that the solution to  $E_K\{\psi_{\text{RBD}}(Z_n; \tilde{\beta})\} = 0$  may not be unique, i.e.,  $\mathcal{S}_K$  may contain more than one element. We then define a functional for the estimator of  $\tilde{\beta}_{n,0}$  as follows,

$$T(K) = \arg \min_{\tilde{\beta} \in \mathcal{S}_K} \|\tilde{\beta} - \tilde{\beta}_{n,0}\|. \quad (12)$$

From the result of Lemma A1 in Appendix A,  $T(K_{n,\epsilon})$  is the unique local minimizer of  $\ell_{K_{n,\epsilon}}(\tilde{\beta})$  in the  $\sqrt{p_n/n}$ -neighborhood of  $\tilde{\beta}_{n,0}$ . Particularly,  $T(K_{n,0}) = \tilde{\beta}_{n,0}$ . Similarly, from Lemma A2 in Appendix A,  $T(\mathbb{K}_n)$  is the unique local minimizer of  $\ell_{\mathbb{K}_n}(\tilde{\beta})$  which satisfies  $\|T(\mathbb{K}_n) - \tilde{\beta}_{n,0}\| = O_p(\sqrt{p_n/n})$ .

From [23] (Equation (2.1.6) on pp. 84), the influence function of  $T(\cdot)$  at  $K_{n,0}$  is defined as

$$\text{IF}(z_n; T, K_{n,0}) = \frac{\partial}{\partial t} T((1-t)K_{n,0} + t\Delta_{z_n}) \Big|_{t=0} = \lim_{t \downarrow 0} \frac{T((1-t)K_{n,0} + t\Delta_{z_n}) - \tilde{\beta}_{n,0}}{t},$$

where  $\Delta_{z_n}$  is the probability measure which puts mass 1 at the point  $z_n$ . Since the dimension of  $T(\cdot)$  diverges with  $n$ , its influence function is defined for each fixed  $n$ . From Lemma A8 in Appendix A, under certain regularity conditions, the influence function exists and has the following expression:

$$\text{IF}(z_n; T, K_{n,0}) = -\mathbf{H}_n^{-1} \psi_{\text{RBD}}(z_n; \tilde{\beta}_{n,0}), \quad (13)$$

where  $\mathbf{H}_n = \mathbb{E}_{K_{n,0}}\{\mathbf{p}_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\boldsymbol{\beta}}_{n,0})w(\mathbf{X}_n)\tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T\}$ . The form of the influence function for diverging  $p_n$  in Equation(13) coincides with that in [23,24] for fixed  $p_n$ .

In our theoretical derivations, approximations of the asymptotic level and power of  $W_n$  will involve the following matrices:

$$\begin{aligned}\Omega_n &= \mathbb{E}_{K_{n,0}}\{\mathbf{p}_1^2(Y; \tilde{\mathbf{X}}_n^T \tilde{\boldsymbol{\beta}}_{n,0})w^2(\mathbf{X}_n)\tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T\}, \\ U_n &= A_n \mathbf{H}_n^{-1} \Omega_n \mathbf{H}_n^{-1} A_n^T.\end{aligned}$$

### 3.1. Asymptotic Level of $W_n$ under Contamination

We now investigate the asymptotic level of the Wald-type test  $W_n$  under the  $\epsilon$ -contamination.

**Theorem 1.** Assume Conditions A0–A9 and B4 in Appendix A. Suppose  $p_n^6/n \rightarrow 0$  as  $n \rightarrow \infty$ ,  $\sup_n \mathbb{E}_f(\|w(\mathbf{X}_n)\tilde{\mathbf{X}}_n\|) \leq C$ . Denote by  $\alpha(K_{n,\epsilon})$  the level of  $W_n = n\{A_n \mathbf{T}(\mathbb{K}_n) - g_0\}^T (A_n \mathbf{H}_n^{-1} \tilde{\Omega}_n \tilde{\mathbf{H}}_n^{-1} A_n^T)^{-1} \{A_n \mathbf{T}(\mathbb{K}_n) - g_0\}$  when the underlying distribution is  $K_{n,\epsilon}$  in Equation (10) and by  $\alpha_0$  the nominal level. Under  $H_0$  in Equation (8), it follows that

$$\limsup_{n \rightarrow \infty} \alpha(K_{n,\epsilon}) = \alpha_0 + \epsilon^2 \mu_k D + o(\epsilon^2) \quad \text{as } \epsilon \rightarrow 0,$$

where

$$D = \limsup_{n \rightarrow \infty} \|U_n^{-1/2} A_n \mathbb{E}_f\{\text{IF}(\mathbf{Z}_n; \mathbf{T}, K_{n,0})\}\|^2 < \infty,$$

$\mu_k = -\frac{\partial}{\partial \delta} H_k(\eta_{1-\alpha_0}; \delta)|_{\delta=0}$ ,  $H_k(\cdot; \delta)$  is the cumulative distribution function of a  $\chi_k^2(\delta)$  distribution, and  $\eta_{1-\alpha_0}$  is the  $1 - \alpha_0$  quantile of the central  $\chi_k^2$  distribution.

Theorem 1 indicates that if the influence function for  $T(\cdot)$  is bounded, then the asymptotic level of  $W_n$  under the  $\epsilon$ -contamination is also bounded and close to the nominal level when  $\epsilon$  is sufficiently small. As a comparison, the robustness property in [10] of the Wald-type test is studied based on M-estimator for general parametric models with a fixed dimension  $p_n$ . They assumed certain conditions that guarantee Fréchet differentiability which further implies the existence of the influence function and the asymptotic normality of the corresponding estimator. However, in the set-ups of our paper, it's difficult to check those conditions, due to the use of Bregman divergence and the diverging dimension  $p_n$ . Hence, the assumptions we make in Theorem 1 are different from those in [10], and are comparatively mild and easy to check. Moreover, the result of Theorem 1 cannot be easily derived from that of [10].

In Theorem 1,  $p_n$  is allowed to diverge with  $p_n^6/n = o(1)$ , which is slower than that in [1] with  $p_n^5/n = o(1)$ . Theoretically, the assumption  $p_n^5/n = o(1)$  is required to obtain the asymptotic distribution of  $W_n$  in [1]. Furthermore, to derive the limit distribution of  $W_n$  under the  $\epsilon$ -contamination, assumption  $p_n^6/n = o(1)$  is needed (see Lemma A7 in Appendix A). Hence, the reason that our assumption is stronger than that in [1] is the consideration of the  $\epsilon$ -contamination of the data. Practically, due to the advancement of technology and different forms of data gathering, large dimension becomes a common characteristic and hence the varying-dimensional model has a wide range of applications, e.g., brain imaging data, financial data, web term-document data and gene expression data. Even some of the classical settings, e.g. the Framingham heart study with  $n = 25,000$  and  $p_n = 100$ , can be viewed as varying-dimensional cases.

As an illustration, we apply the general result of Theorem 1 to the special case of a point mass contamination.

**Corollary 1.** With the notations in Theorem 1, assume Conditions A0–A9 in Appendix A,  $\sup_{x_n \in \mathbb{R}^{p_n}} \|w(x_n)x_n\| \leq C$  and  $\sup_{\mu \in \mathbb{R}} |q''(\mu)|\sqrt{V(\mu)/F'(\mu)} \leq C$ .

- (i) If  $p_n \equiv p$ ,  $A_n \equiv A$ ,  $\tilde{\beta}_{n,0} \equiv \tilde{\beta}_0$ ,  $K_{n,0} \equiv K_0$  and  $U_n \equiv U$  are fixed, then, for  $K_{n,\epsilon} = (1 - \epsilon/\sqrt{n})K_0 + \epsilon/\sqrt{n}\Delta_z$  with  $z \in \mathbb{R}^p$  a fixed point, under  $H_0$  in Equation (8), it follows that

$$\sup_{z \in \mathbb{R}^p} \lim_{n \rightarrow \infty} \alpha(K_{n,\epsilon}) = \alpha_0 + \epsilon^2 \mu_k D_1 + o(\epsilon^2) \quad \text{as } \epsilon \rightarrow 0,$$

where

$$D_1 = \sup_{z \in \mathbb{R}^p} \|U^{-1/2} A \text{IF}(z; T, K_0)\|^2 < \infty.$$

- (ii) If  $p_n$  diverges with  $p_n^6/n \rightarrow 0$ , for  $K_{n,\epsilon} = (1 - \epsilon/\sqrt{n})K_{n,0} + \epsilon/\sqrt{n}\Delta_{z_n}$  with  $z_n \in \mathbb{R}^{p_n}$  a sequence of deterministic points, then, under  $H_0$  in Equation (8),

$$\sup_{C_0 > 0} \sup_{z_n \in S_{C_0}} \limsup_{n \rightarrow \infty} \alpha(K_{n,\epsilon}) = \alpha_0 + \epsilon^2 \mu_k D_2 + o(\epsilon^2) \quad \text{as } \epsilon \rightarrow 0,$$

where  $S_{C_0} = \{z_n = (x_n^T, y)^T : \|x_n\|_\infty \leq C_0\}$ ,  $C_0 > 0$  is a constant and

$$D_2 = \sup_{C_0 > 0} \sup_{z_n \in S_{C_0}} \limsup_{n \rightarrow \infty} \|U_n^{-1/2} A_n \text{IF}(z_n; T, K_{n,0})\|^2 < \infty.$$

In Corollary 1, conditions  $\sup_{x_n \in \mathbb{R}^{p_n}} \|w(x_n)x_n\| \leq C$  and  $\sup_{\mu \in \mathbb{R}} |q''(\mu)\sqrt{V(\mu)}/F'(\mu)| \leq C$  are needed to guarantee the boundedness of the score function in Equation (7). Particularly, the function  $w(x_n)$  downweights the high leverage points and can be chosen as, e.g.,  $w(x_n) = 1/(1 + \|x_n\|)$ . The condition  $\sup_{\mu \in \mathbb{R}} |q''(\mu)\sqrt{V(\mu)}/F'(\mu)| \leq C$  is needed to bound Equation (6), and is satisfied in many situations.

- For example, for the linear model with  $q(\mu) = a\mu - \mu^2$ ,  $V(\mu) = \sigma^2$  and  $F(\mu) = \mu$ , where  $a$  and  $\sigma^2$  are constants, we observe  $|q''(\mu)\sqrt{V(\mu)}/F'(\mu)| = 2\sigma \leq C$ .
- Another example is the logistic regression model with binary response and  $q(\mu) = -2\{\mu \log(\mu) + (1-\mu) \log(1-\mu)\}$  (corresponding to Bernoulli deviance loss),  $V(\mu) = \mu(1-\mu)$ ,  $F(\mu) = \log\{\mu/(1-\mu)\}$ . In this case,  $|q''(\mu)\sqrt{V(\mu)}/F'(\mu)| = 2\{\mu(1-\mu)\}^{1/2} \leq C$  since  $\mu \in [0, 1]$ . Likewise, if  $q(\mu) = 2\{\mu(1-\mu)\}^{1/2}$  (for the exponential loss), then  $|q''(\mu)\sqrt{V(\mu)}/F'(\mu)| = 1/2$ .

Furthermore, the bound on  $\psi(\cdot)$  is useful to control deviations in the  $Y$ -space, which ensures the stability of the robust-BD test if  $Y$  is arbitrarily contaminated.

Concerning the dimensionality  $p_n$ , Corollary 1 reveals the following implications. If  $p_n$  is fixed, then the asymptotic level of  $W_n$  under the  $\epsilon$ -contamination is uniformly bounded for all  $z \in \mathbb{R}^p$ , which implies the robustness of validity of the test. This result coincides with that in Proposition 5 of [10]. When  $p_n$  diverges, the asymptotic level is still stable if the point contamination satisfies  $\|x_n\|_\infty \leq C_0$ , where  $C_0 > 0$  is an arbitrary constant. Although this condition may not be the weakest, it still covers a wide range of point mass contaminations.

### 3.2. Asymptotic Power of $W_n$ under Contamination

Now, we will study the asymptotic power of  $W_n$  under a sequence of contiguous alternatives of the form

$$H_{1n} : A_n \tilde{\beta}_{n,0} - g_0 = n^{-1/2} c, \tag{14}$$

where  $c = (c_1, \dots, c_k)^T \neq \mathbf{0}$  is fixed.

**Theorem 2.** Assume Conditions A0–A9 and B4 in Appendix A. Suppose  $p_n^6/n \rightarrow 0$  as  $n \rightarrow \infty$ ,  $\sup_n E_I(\|w(X_n)\tilde{X}_n\|) \leq C$ . Denote by  $\beta(K_{n,\epsilon})$  the power of  $W_n = n\{A_n T(\mathbb{K}_n) -$

$\mathbf{g}_0\}^T(\mathbf{A}_n\widehat{\mathbf{H}}_n^{-1}\widehat{\Omega}_n\widehat{\mathbf{H}}_n^{-1}\mathbf{A}_n^T)^{-1}\{\mathbf{A}_n\mathbf{T}(\mathbb{K}_n) - \mathbf{g}_0\}$  when the underlying distribution is  $K_{n,\epsilon}$  in Equation (10) and by  $\beta_0$  the nominal power. Under  $H_{1n}$  in Equation (14), it follows that

$$\liminf_{n \rightarrow \infty} \beta(K_{n,\epsilon}) = \beta_0 + \epsilon v_k B + o(\epsilon) \quad \text{as } \epsilon \rightarrow 0,$$

where

$$B = \liminf_{n \rightarrow \infty} 2c^T U_n^{-1} A_n E_J\{\text{IF}(\mathbf{Z}_n; \mathbf{T}, K_{n,0})\},$$

with  $|B| < \infty$ ,  $v_k = -\frac{\partial}{\partial \delta} H_k(\eta_{1-\alpha_0}; \delta)|_{\delta=c^T U_n^{-1} c}$  and  $H_k(\cdot; \delta)$  and  $\eta_{1-\alpha_0}$  being defined in Theorem 1.

The result for the asymptotic power is similar in spirit to that for the level. From Theorem 2, if the influence function is bounded, the asymptotic power is also bounded from below and close to the nominal power under a small amount of contamination. This means that the robust-BD Wald-type test enjoys the robustness of efficiency. In addition, the property of the asymptotic power can be obtained for a point mass contamination.

**Corollary 2.** With the notations in Theorem 2, assume Conditions A0–A9 in Appendix A,  $\sup_{x_n \in \mathbb{R}^{p_n}} \|w(x_n)x_n\| \leq C$  and  $\sup_{\mu \in \mathbb{R}} |q''(\mu)\sqrt{V(\mu)}/F'(\mu)| \leq C$ .

- (i) If  $p_n \equiv p$ ,  $A_n \equiv A$ ,  $\tilde{\beta}_{n,0} \equiv \tilde{\beta}_0$ ,  $K_{n,0} \equiv K_0$  and  $U_n \equiv U$  are fixed, then, for  $K_{n,\epsilon} = (1 - \epsilon/\sqrt{n})K_0 + \epsilon/\sqrt{n}\Delta_z$  with  $z \in \mathbb{R}^p$  a fixed point, under  $H_{1n}$  in Equation (14), it follows that

$$\inf_{z \in \mathbb{R}^p} \lim_{n \rightarrow \infty} \beta(K_{n,\epsilon}) = \beta_0 + \epsilon v_k B_1 + o(\epsilon) \quad \text{as } \epsilon \rightarrow 0,$$

where

$$B_1 = \inf_{z \in \mathbb{R}^p} 2c^T U^{-1} A \text{IF}(z; \mathbf{T}, K_0),$$

with  $|B_1| < \infty$ .

- (ii) If  $p_n$  diverges with  $p_n^6/n \rightarrow 0$ , for  $K_{n,\epsilon} = (1 - \epsilon/\sqrt{n})K_{n,0} + \epsilon/\sqrt{n}\Delta_{z_n}$  with  $z_n \in \mathbb{R}^{p_n}$  a sequence of deterministic points, then, under  $H_{1n}$  in Equation (14),

$$\inf_{C_0 > 0} \inf_{z_n \in S_{C_0}} \liminf_{n \rightarrow \infty} \beta(K_{n,\epsilon}) = \beta_0 + \epsilon v_k B_2 + o(\epsilon) \quad \text{as } \epsilon \rightarrow 0,$$

where  $S_{C_0} = \{z_n = (x_n^T, y)^T : \|x_n\|_\infty \leq C_0\}$ ,  $C_0 > 0$  is a constant and

$$B_2 = \inf_{C_0 > 0} \inf_{z_n \in S_{C_0}} \liminf_{n \rightarrow \infty} 2c^T U_n^{-1} A_n \text{IF}(\mathbf{Z}_n; \mathbf{T}, K_{n,0}),$$

with  $|B_2| < \infty$ .

#### 4. Simulation

Regarding the practical utility of  $W_n$ , numerical studies concerning the empirical level and power of  $W_n$  under a fixed amount of contamination have been conducted in Section 6 of [1]. To support the theoretical results in our paper, we conduct new simulations to check the robustness of validity and efficiency of  $W_n$ . Specifically, we will examine the empirical level and power of the test statistic as  $\epsilon$  varies.

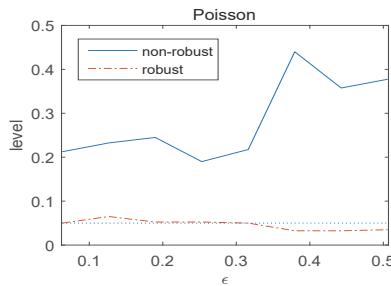
The robust-BD estimation utilizes the Huber  $\psi$ -function  $\psi_c(\cdot)$  with  $c = 1.345$  and the weight function  $w(X_n) = 1/(1 + \|X_n\|)$ . Comparisons are made with the classical non-robust counterparts corresponding to using  $\psi(r) = r$  and  $w(x_n) \equiv 1$ . For each situation below, we set  $n = 1000$  and conduct 400 replications.

#### 4.1. Overdispersed Poisson Responses

Overdispersed Poisson counts  $Y$ , satisfying  $\text{var}(Y|X_n = x_n) = 2m(x_n)$ , are generated via a negative Binomial( $m(x_n), 1/2$ ) distribution. Let  $p_n = \lfloor 4(n^{1/5.5} - 1) \rfloor$  and  $\tilde{\beta}_{n,0} = (0, 2, 0, \dots, 0)^T$ , where  $\lfloor \cdot \rfloor$  denotes the floor function. Generate  $X_{ni} = (X_{i,1}, \dots, X_{i,p_n})^T$  by  $X_{ij} \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[-0.5, 0.5]$ . The log link function is considered and the (negative) quasi-likelihood is utilized as the BD, generated by the  $q$ -function in Equation (3) with  $V(\mu) = \mu$ . The estimator and test statistic are calculated by assuming  $Y$  follows Poisson distribution.

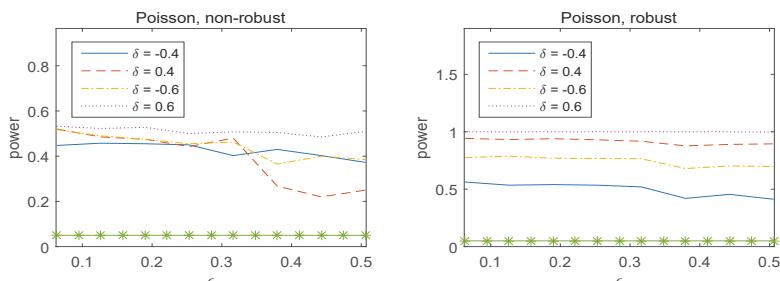
The data are contaminated by  $X_{i,\text{mod}(i,p_n-1)+1}^* = 3\text{sign}(U_i - 0.5)$  and  $Y_i^* = Y_i\mathbf{I}(Y_i > 20) + 20\mathbf{I}(Y_i \leq 20)$  for  $i = 1, \dots, k$ , with  $k \in \{2, 4, 6, 8, 10, 12, 14, 16\}$  the number of contaminated data points, where  $\text{mod}(a, b)$  is the modulo operation “ $a$  modulo  $b$ ” and  $\{U_i\} \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$ . Then, the proportion of contaminated data,  $k/n$ , is equal to  $\epsilon/\sqrt{n}$  as in Equation (10), which implies  $\epsilon = k/\sqrt{n}$ .

Consider the null hypothesis  $H_0 : A_n\tilde{\beta}_{n,0} = 0$  with  $A_n = (0, 0, 0, 1, 0, \dots, 0)$ . Figure 1 plots the empirical level of  $W_n$  versus  $\epsilon$ . We observe that the asymptotic nominal level 0.05 is approximately retained by the robust Wald-type test. On the other hand, under contaminations, the non-robust Wald-type test breaks in level, showing high sensitivity to the presence of outliers.



**Figure 1.** Observed level of  $W_n$  versus  $\epsilon$  for overdispersed Poisson responses. The dotted line indicates the 5% significance level.

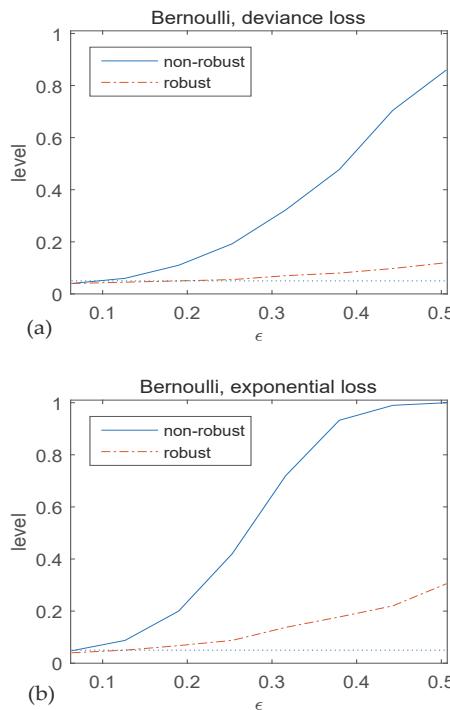
To assess the stability of the power of the test, we generate the original data from the true model, but with the true parameter  $\tilde{\beta}_{n,0}$  replaced by  $\tilde{\beta}_n = \tilde{\beta}_{n,0} + \delta c$  with  $\delta \in \{-0.4, 0.4, -0.6, 0.6\}$  and  $c = (1, \dots, 1)^T$  a vector of ones. Figure 2 plots the empirical rejection rates of the null model, which implies that the robust Wald-type test has sufficiently large power to detect the alternative hypothesis. In addition, the power of the robust method is generally larger than that of the non-robust method.



**Figure 2.** Observed power of  $W_n$  versus  $\epsilon$  for overdispersed Poisson responses. The statistics in the left panel correspond to non-robust method and those in the right panel are for robust method. The asterisk line indicates the 5% significance level.

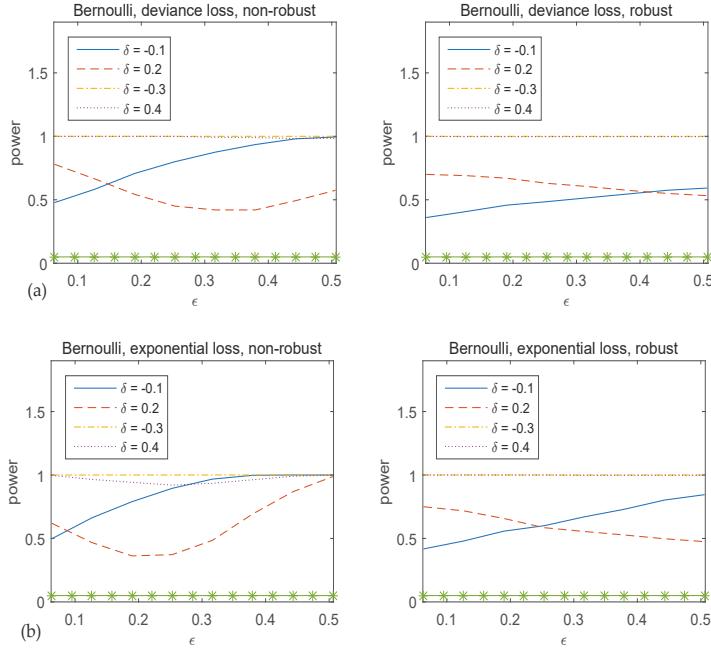
#### 4.2. Bernoulli Responses

We generate data with two classes from the model,  $Y|X_n = x_n \sim \text{Bernoulli}\{m(x_n)\}$ , where  $\text{logit}\{m(x_n)\} = \tilde{x}_n^T \tilde{\beta}_{n,0}$ . Let  $p_n = 2$ ,  $\tilde{\beta}_{n,0} = (0, 1, 1)^T$  and  $X_{ni} \stackrel{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{I}_{p_n})$ . The null hypothesis is  $H_0 : \tilde{\beta}_{n,0} = (0, 1, 1)^T$ . Both the deviance loss and the exponential loss are employed as the BD. We contaminate the data by setting  $X_{i,1}^* = 2 + i/8$  and  $Y_i^* = 0$  for  $i = 1, \dots, k$  with  $k \in \{2, 4, 6, 8, 10, 12, 14, 16\}$ . To investigate the robustness of validity of  $W_n$ , we plot the observed level versus  $\epsilon$  in Figure 3. We find that the level of the non-robust method diverges fast as  $\epsilon$  increases. It's also clear that the empirical level of the robust method is close to the nominal level when  $\epsilon$  is small and increases slightly with  $\epsilon$ , which coincides with our results in Theorem 1.



**Figure 3.** Observed level of  $W_n$  versus  $\epsilon$  for Bernoulli responses. The statistics in (a) use deviance loss and those in (b) use exponential loss. The dotted line indicates the 5% significance level.

To assess the stability of the power of  $W_n$ , we generate the original data from the true model, but with the true parameter  $\tilde{\beta}_{n,0}$  replaced by  $\tilde{\beta}_n = \tilde{\beta}_{n,0} + \delta c$  with  $\delta \in \{-0.1, 0.2, -0.3, 0.4\}$  and  $c = (1, \dots, 1)^T$  a vector of ones. Figure 4 plots the power of the Wald-type test versus  $\epsilon$ , which implies that the robust method has sufficiently large power, and hence supports the theoretical results in Theorem 2.



**Figure 4.** Observed power of  $W_n$  versus  $\epsilon$  for Bernoulli responses. The top panels correspond to deviance loss while the bottom panels are for exponential loss. The statistics in the left panels are calculated using non-robust method and those in the right panels are from robust method. The asterisk line indicates the 5% significance level.

**Acknowledgments:** We thank the two referees for insightful comments and suggestions. Chunming Zhang's research is supported by the U.S. NSF Grants DMS-1712418, DMS-1521761, the Wisconsin Alumni Research Foundation and the National Natural Science Foundation of China, grants 11690014. Xiao Guo's research is supported by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China, grants 11601500, 11671374 and 11771418.

**Author Contributions:** Chunming Zhang conceived and designed the experiments; Xiao Guo performed the experiments; Xiao Guo analyzed the data; Chunming Zhang contributed to analysis tools; Chunming Zhang and Xiao Guo wrote the paper.

**Conflicts of Interest:** The authors declare no conflicts of interest.

#### Appendix A. Conditions and Proofs of Main Results

We first introduce some necessary notations used in the proof.

**Notations.** For arbitrary distributions  $K$  and  $K'$  of  $Z_n$ , define

$$\begin{aligned}\Omega_{n,K,T(K')} &= E_K\{p_1^2(Y; \tilde{X}_n^T T(K')) w^2(X_n) \tilde{X}_n \tilde{X}_n^T\}, \\ H_{n,K,T(K')} &= E_K\{p_2(Y; \tilde{X}_n^T T(K')) w(X_n) \tilde{X}_n \tilde{X}_n^T\}.\end{aligned}$$

Therefore,  $\Omega_n = \Omega_{n,K_{n,0},\tilde{\beta}_{n,0}}$ ,  $H_n = H_{n,K_{n,0},\tilde{\beta}_{n,0}}$ ,  $\hat{\Omega}_n = \Omega_{n,\mathbb{K}_n,T(\mathbb{K}_n)}$  and  $\hat{H}_n = H_{n,\mathbb{K}_n,T(\mathbb{K}_n)}$ . For notational simplicity, let  $\Omega_{n,\epsilon} = \Omega_{n,K_{n,\epsilon},T(K_{n,\epsilon})}$  and  $H_{n,\epsilon} = H_{n,K_{n,\epsilon},T(K_{n,\epsilon})}$ .

Define the following matrices,

$$\begin{aligned}U(K_{n,\epsilon}) &= A_n H_{n,\epsilon}^{-1} \Omega_{n,\epsilon} H_{n,\epsilon}^{-1} A_n^T, \\ U(\mathbb{K}_n) &= A_n \hat{H}_n^{-1} \hat{\Omega}_n \hat{H}_n^{-1} A_n^T.\end{aligned}$$

The following conditions are needed in the proof, which are adopted from [1].

**Condition A.**

- A0.  $\sup_{n \geq 1} \|\tilde{\beta}_{n,0}\|_1 < \infty$ .
- A1.  $w(\cdot)$  is a bounded function. Assume that  $\psi(r)$  is a bounded, odd function, and twice differentiable, such that  $\psi'(r)$ ,  $\psi'(r)r$ ,  $\psi''(r)$ ,  $\psi''(r)r$  and  $\psi''(r)r^2$  are bounded;  $V(\cdot) > 0$ ,  $V^{(2)}$  is continuous.
- A2.  $q^{(4)}(\cdot)$  is continuous, and  $q^{(2)}(\cdot) < 0$ .  $G_1^{(3)}$  is continuous.
- A3.  $F(\cdot)$  is monotone and a bijection,  $F^{(3)}(\cdot)$  is continuous, and  $F^{(1)}(\cdot) \neq 0$ .
- A4.  $\|\mathbf{X}_n\|_\infty \leq C$  almost surely if the underlying distribution is  $K_{n,0}$ .
- A5.  $E_{K_{n,0}}(\tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T)$  exists and is nonsingular.
- A6. There is a large enough open subset of  $\mathbb{R}^{p_n+1}$  which contains  $\tilde{\beta}_{n,0}$ , such that  $F^{-1}(\tilde{x}_n^T \tilde{\beta})$  is bounded for all  $\tilde{\beta}$  in the subset and all  $\tilde{x}_n$  such that  $\|\tilde{x}_n\|_\infty \leq C$ , where  $C > 0$  is a large enough constant.
- A7.  $\mathbf{H}_n$  is positive definite, with eigenvalues uniformly bounded away from 0.
- A8.  $\Omega_n$  is positive definite, with eigenvalues uniformly bounded away from 0.
- A9.  $\|\mathbf{H}_n^{-1} \Omega_n\|$  is bounded away from  $\infty$ .

**Condition B.**

- B4.  $\|\mathbf{X}_n\|_\infty \leq C$  almost surely if the underlying distribution is  $J$ .

The following Lemmas A1–A9 are needed to prove the main theoretical results in this paper.

**Lemma A1** ( $\|T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\|$ ). *Assume Conditions A0–A7 and B4. For  $K_{n,\epsilon}$  in Equation (10),  $\ell_K(\cdot)$  in Equation (11) and  $T(\cdot)$  in Equation (12), if  $p_n^4/n \rightarrow 0$  as  $n \rightarrow \infty$ , then  $T(K_{n,\epsilon})$  is a local minimizer of  $\ell_{K_{n,\epsilon}}(\tilde{\beta})$  such that  $\|T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\| = O(\sqrt{p_n/n})$ . Furthermore,  $T(K_{n,\epsilon})$  is unique.*

**Proof.** We follow the idea of the proof in [25]. Let  $r_n = \sqrt{p_n/n}$  and  $\tilde{\mathbf{u}}_n = (u_0, u_1, \dots, u_{p_n})^T \in \mathbb{R}^{p_n+1}$ . First, we show that there exists a sufficiently large constant  $C$  such that, for large  $n$ , we have

$$\inf_{\|\tilde{\mathbf{u}}_n\|=C} \ell_{K_{n,\epsilon}}(\tilde{\beta}_{n,0} + r_n \tilde{\mathbf{u}}_n) > \ell_{K_{n,\epsilon}}(\tilde{\beta}_{n,0}). \quad (\text{A1})$$

To show Equation (A1), consider

$$\begin{aligned} \ell_{K_{n,\epsilon}}(\tilde{\beta}_{n,0} + r_n \tilde{\mathbf{u}}_n) - \ell_{K_{n,\epsilon}}(\tilde{\beta}_{n,0}) &= E_{K_{n,\epsilon}} \{ \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0} + r_n \tilde{\mathbf{X}}_n^T \tilde{\mathbf{u}}_n)) w(\mathbf{X}_n) \\ &\quad - \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0})) w(\mathbf{X}_n) \} \\ &\equiv I_1, \end{aligned}$$

where  $\|\tilde{\mathbf{u}}_n\| = C$ .

By Taylor expansion,

$$I_1 = I_{1,1} + I_{1,2} + I_{1,3}, \quad (\text{A2})$$

where

$$\begin{aligned} I_{1,1} &= r_n E_{K_{n,\epsilon}} \{ p_1(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n^T \} \tilde{\mathbf{u}}_n, \\ I_{1,2} &= r_n^2 / 2 E_{K_{n,\epsilon}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) (\tilde{\mathbf{X}}_n^T \tilde{\mathbf{u}}_n)^2 \}, \\ I_{1,3} &= r_n^3 / 6 E_{K_{n,\epsilon}} \{ p_3(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}^*) w(\mathbf{X}_n) (\tilde{\mathbf{X}}_n^T \tilde{\mathbf{u}}_n)^3 \}, \end{aligned}$$

for  $\tilde{\beta}_n^*$  located between  $\tilde{\beta}_{n,0}$  and  $\tilde{\beta}_{n,0} + r_n \tilde{\mathbf{u}}_n$ . Hence

$$|I_{1,1}| \leq r_n \|E_{K_{n,\epsilon}} \{ p_1(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n^T \} \| \|\tilde{\mathbf{u}}_n\|$$

$$\begin{aligned}
&= r_n \frac{\epsilon}{\sqrt{n}} \| \mathbb{E}_I \{ p_1(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \} \| \| \tilde{\mathbf{u}}_n \| \\
&\leq C r_n \sqrt{p_n/n} \| \tilde{\mathbf{u}}_n \|,
\end{aligned}$$

since  $\| \mathbb{E}_I \{ p_1(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \} \| = O(\sqrt{p_n})$  and  $\mathbb{E}_{K_{n,0}} \{ p_1(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \} = \mathbf{0}$ . For  $I_{1,2}$  in Equation (A2),

$$\begin{aligned}
I_{1,2} &= \frac{r_n^2}{2} \mathbb{E}_{K_{n,0}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) (\tilde{\mathbf{X}}_n^T \tilde{\mathbf{u}}_n)^2 \} \\
&\quad + \frac{r_n^2}{2} [ \mathbb{E}_{K_{n,\epsilon}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) (\tilde{\mathbf{X}}_n^T \tilde{\mathbf{u}}_n)^2 \} - \mathbb{E}_{K_{n,0}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) (\tilde{\mathbf{X}}_n^T \tilde{\mathbf{u}}_n)^2 \} ] \\
&\equiv I_{1,2,1} + I_{1,2,2},
\end{aligned}$$

where  $I_{1,2,1} = 2^{-1} r_n^2 \tilde{\mathbf{u}}_n^T \mathbf{H}_n \tilde{\mathbf{u}}_n$ . Meanwhile, we have

$$\begin{aligned}
&|I_{1,2,2}| \\
&\leq r_n^2 \| \mathbb{E}_{K_{n,\epsilon}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \} - \mathbb{E}_{K_{n,0}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \} \|_F \| \tilde{\mathbf{u}}_n \|^2 \\
&= r_n^2 \frac{\epsilon}{\sqrt{n}} \| \mathbb{E}_I \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \} - \mathbb{E}_{K_{n,0}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \} \|_F \| \tilde{\mathbf{u}}_n \|^2 \\
&\leq C r_n^2 p_n \| \tilde{\mathbf{u}}_n \|^2 / \sqrt{n},
\end{aligned}$$

where  $\| \mathbb{E}_I \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \} \|_F = O(p_n)$  and  $\| \mathbb{E}_{K_{n,0}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \} \|_F = O(p_n)$ . Thus,

$$I_{1,2} = 2^{-1} r_n^2 \tilde{\mathbf{u}}_n^T \mathbf{H}_n \tilde{\mathbf{u}}_n + O(r_n^2 p_n / \sqrt{n}) \| \tilde{\mathbf{u}}_n \|^2. \quad (\text{A3})$$

For  $I_{1,3}$  in Equation (A2), we observe that

$$|I_{1,3}| \leq C r_n^3 \mathbb{E}_{K_{n,\epsilon}} \{ |p_3(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_n^*)| |w(\mathbf{X}_n)| |\tilde{\mathbf{X}}_n^T \tilde{\mathbf{u}}_n|^3 \} = O(r_n^3 p_n^{3/2}) \| \tilde{\mathbf{u}}_n \|^3.$$

We can choose some large  $C$  such that  $I_{1,1}$ ,  $I_{1,2,2}$  and  $I_{1,3}$  are all dominated by the first term of  $I_{1,2}$  in Equation (A3), which is positive by the eigenvalue assumption. This implies Equation (A1). Therefore, there exists a local minimizer of  $\ell_{K_{n,\epsilon}}(\tilde{\beta})$  in the  $\sqrt{p_n/n}$  neighborhood of  $\tilde{\beta}_{n,0}$ , and denote this minimizer by  $\tilde{\beta}_{n,\epsilon}$ .

Next, we show that the local minimizer  $\tilde{\beta}_{n,\epsilon}$  of  $\ell_{K_{n,\epsilon}}(\tilde{\beta})$  is unique in the  $\sqrt{p_n/n}$  neighborhood of  $\tilde{\beta}_{n,0}$ . For all  $\tilde{\beta}$  such that  $\|\tilde{\beta} - \tilde{\beta}_{n,0}\| = O(n^{-1/4} p_n^{-1/2})$ ,

$$\begin{aligned}
\mathbb{E}_{K_{n,\epsilon}} \left\| \frac{\partial}{\partial \tilde{\beta}} \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta})) w(\mathbf{X}_n) \right\| &= \mathbb{E}_{K_{n,\epsilon}} \| p_1(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \| \leq C \sqrt{p_n} \\
\mathbb{E}_{K_{n,\epsilon}} \left\| \frac{\partial^2}{\partial \tilde{\beta}^2} \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta})) w(\mathbf{X}_n) \right\| &= \mathbb{E}_{K_{n,\epsilon}} \| p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \| \leq C p_n
\end{aligned}$$

and hence,

$$\begin{aligned}
\frac{\partial}{\partial \tilde{\beta}} \mathbb{E}_{K_{n,\epsilon}} \{ \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta})) w(\mathbf{X}_n) \} &= \mathbb{E}_{K_{n,\epsilon}} \left\{ \frac{\partial}{\partial \tilde{\beta}} \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta})) w(\mathbf{X}_n) \right\} \\
\frac{\partial^2}{\partial \tilde{\beta}^2} \mathbb{E}_{K_{n,\epsilon}} \{ \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta})) w(\mathbf{X}_n) \} &= \mathbb{E}_{K_{n,\epsilon}} \left\{ \frac{\partial^2}{\partial \tilde{\beta}^2} \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta})) w(\mathbf{X}_n) \right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{\partial^2}{\partial \tilde{\beta}^2} \mathbb{E}_{K_{n,\epsilon}} \{ \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta})) w(\mathbf{X}_n) \} \\
&= \mathbb{E}_{K_{n,\epsilon}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \}
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_{K_{n,0}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \} \\
&\quad + \mathbb{E}_{K_{n,0}} [\{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}) - p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) \} w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T] \\
&\quad + [\mathbb{E}_{K_{n,\epsilon}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \} - \mathbb{E}_{K_{n,0}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \}] \\
&= I_1^* + I_2^* + I_3^*.
\end{aligned}$$

We know that the minimum eigenvalues of  $I_1^*$  are uniformly bounded away from 0,

$$\begin{aligned}
\|I_2^*\| &= \|\mathbb{E}_{K_{n,0}} \{ p_3(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}^{***}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \tilde{\mathbf{X}}_n^T (\tilde{\beta} - \tilde{\beta}_{n,0}) \} \| \leq C p_n / n^{1/4} = o(1) \\
\|I_3^*\| &\leq \epsilon / \sqrt{n} [\|\mathbb{E}_{K_{n,0}} \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \}\| + \|\mathbb{E}_f \{ p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \}\|] \\
&\leq C p_n / \sqrt{n} = o(1).
\end{aligned}$$

Hence, for  $n$  large enough,  $\frac{\partial^2}{\partial \tilde{\beta}^2} \mathbb{E}_{K_{n,\epsilon}} \{ \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta})) w(\mathbf{X}_n) \}$  is positive definite for all  $\tilde{\beta}$  such that  $\|\tilde{\beta} - \tilde{\beta}_{n,0}\| = O(n^{-1/4} p_n^{-1/2})$ . Therefore, there exists a unique minimizer of  $\ell_{K_{n,\epsilon}}(\tilde{\beta})$  in the  $n^{-1/4} p_n^{-1/2}$  neighborhood of  $\tilde{\beta}_{n,0}$  which covers  $\tilde{\beta}_{n,\epsilon}$ . From

$$\begin{aligned}
0 &= \frac{\partial}{\partial \tilde{\beta}} \mathbb{E}_{K_{n,\epsilon}} \{ \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta})) w(\mathbf{X}_n) \} \Big|_{\tilde{\beta}=\tilde{\beta}_{n,\epsilon}} = \mathbb{E}_{K_{n,\epsilon}} \left\{ \frac{\partial}{\partial \tilde{\beta}} \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta})) \Big|_{\tilde{\beta}=\tilde{\beta}_{n,\epsilon}} w(\mathbf{X}_n) \right\} \\
&= \mathbb{E}_{K_{n,\epsilon}} \{ p_1(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,\epsilon}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \},
\end{aligned}$$

we know  $T(K_{n,\epsilon}) = \tilde{\beta}_{n,\epsilon}$ . From the definition of  $T(\cdot)$ , it's easy to see that  $T(K_{n,\epsilon})$  is unique.  $\square$

**Lemma A2** ( $\|T(\mathbb{K}_n) - T(K_{n,\epsilon})\|$ ). Assume Conditions A0–A7 and B4. For  $K_{n,\epsilon}$  in Equation (10),  $\ell_K(\cdot)$  in Equation (11) and  $T(\cdot)$  in Equation (12), if  $p_n^4/n \rightarrow 0$  as  $n \rightarrow \infty$  and the distribution of  $(\mathbf{X}_n, Y)$  is  $K_{n,\epsilon}$ , then there exists a unique local minimizer  $\hat{\tilde{\beta}}_n$  of  $\ell_{\mathbb{K}_n}(\tilde{\beta})$  such that  $\|\hat{\tilde{\beta}}_n - T(K_{n,\epsilon})\| = O_p(\sqrt{p_n/n})$ . Furthermore,  $\|\hat{\tilde{\beta}}_n - \tilde{\beta}_{n,0}\| = O_p(\sqrt{p_n/n})$  and  $T(\mathbb{K}_n) = \hat{\tilde{\beta}}$ .

**Proof.** Let  $r_n = \sqrt{p_n/n}$  and  $\tilde{\mathbf{u}}_n = (u_0, u_1, \dots, u_{p_n})^T \in \mathbb{R}^{p_n+1}$ . To show the existence of the estimator, it suffices to show that for any given  $\kappa > 0$ , there exists a sufficiently large constant  $C_\kappa$  such that, for large  $n$  we have

$$P \left\{ \inf_{\|\tilde{\mathbf{u}}_n\|=C_\kappa} \ell_{\mathbb{K}_n}(T(K_{n,\epsilon}) + r_n \tilde{\mathbf{u}}_n) > \ell_{\mathbb{K}_n}(T(K_{n,\epsilon})) \right\} \geq 1 - \kappa. \quad (\text{A4})$$

This implies that with probability at least  $1 - \kappa$ , there exists a local minimizer  $\hat{\tilde{\beta}}_n$  of  $\ell_{\mathbb{K}_n}(\tilde{\beta})$  in the ball  $\{T(K_{n,\epsilon}) + r_n \tilde{\mathbf{u}}_n : \|\tilde{\mathbf{u}}_n\| \leq C_\kappa\}$ . To show Equation (A4), consider

$$\begin{aligned}
\ell_{\mathbb{K}_n}(T(K_{n,\epsilon}) + r_n \tilde{\mathbf{u}}_n) - \ell_{\mathbb{K}_n}(T(K_{n,\epsilon})) &= \frac{1}{n} \sum_{i=1}^n \{ \rho_q(Y_i, F^{-1}(\tilde{\mathbf{X}}_{ni}^T (T(K_{n,\epsilon}) + r_n \tilde{\mathbf{u}}_n))) w(\mathbf{X}_{ni}) \\
&\quad - \rho_q(Y_i, F^{-1}(\tilde{\mathbf{X}}_{ni}^T T(K_{n,\epsilon}))) w(\mathbf{X}_{ni}) \} \\
&\equiv I_1,
\end{aligned}$$

where  $\|\tilde{\mathbf{u}}_n\| = C_\kappa$ .

By Taylor expansion,

$$I_1 = I_{1,1} + I_{1,2} + I_{1,3}, \quad (\text{A5})$$

where

$$I_{1,1} = r_n/n \sum_{i=1}^n p_1(Y_i; \tilde{\mathbf{X}}_{ni}^T T(K_{n,\epsilon})) w(\mathbf{X}_{ni}) \tilde{\mathbf{X}}_{ni}^T \tilde{\mathbf{u}}_n,$$

$$\begin{aligned} I_{1,2} &= r_n^2 / (2n) \sum_{i=1}^n p_2(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) (\tilde{\mathbf{X}}_{ni}^T \tilde{\mathbf{u}}_n)^2, \\ I_{1,3} &= r_n^3 / (6n) \sum_{i=1}^n p_3(Y_i; \tilde{\mathbf{X}}_{ni}^T \tilde{\beta}_n^*) w(\mathbf{X}_{ni}) (\tilde{\mathbf{X}}_{ni}^T \tilde{\mathbf{u}}_n)^3 \end{aligned}$$

for  $\tilde{\beta}_n^*$  located between  $\mathbf{T}(K_{n,\epsilon})$  and  $\mathbf{T}(K_{n,\epsilon}) + r_n \tilde{\mathbf{u}}_n$ .

Since  $\|\mathbf{T}(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\| = O(\sqrt{p_n/n}) = o(1)$ , the large open set considered in Condition A6 contains  $\mathbf{T}(K_{n,\epsilon})$  when  $n$  is large enough, say  $n \geq N$  where  $N$  is a positive constant. Therefore, for any fixed  $n \geq N$ , there exists a bounded open subset of  $\mathbb{R}^{p_n+1}$  containing  $\mathbf{T}(K_{n,\epsilon})$  such that for all  $\tilde{\beta}$  in this set,  $\|p_1(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\| \leq C \|\tilde{\mathbf{X}}_n\|$  which is integrable with respect to  $K_{n,\epsilon}$ , where  $C$  is a positive constant. Thus, for  $n \geq N$ ,

$$\mathbf{0} = \frac{\partial}{\partial \tilde{\beta}} \mathbb{E}_{K_{n,\epsilon}} \{ \rho_q(Y, F^{-1}(\tilde{\mathbf{X}}_n^T \tilde{\beta})) w(\mathbf{X}_n) \} \Big|_{\tilde{\beta}=\mathbf{T}(K_{n,\epsilon})} = \mathbb{E}_{K_{n,\epsilon}} \{ p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \}. \quad (\text{A6})$$

Hence,

$$|I_{1,1}| \leq r_n \left\| \frac{1}{n} \sum_{i=1}^n p_1(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) \tilde{\mathbf{X}}_{ni} \right\| \|\tilde{\mathbf{u}}_n\| = O_P(r_n \sqrt{p_n/n}) \|\tilde{\mathbf{u}}_n\|.$$

For  $I_{1,2}$  in Equation (A5),

$$\begin{aligned} I_{1,2} &= \frac{r_n^2}{2n} \sum_{i=1}^n \mathbb{E}_{K_{n,\epsilon}} \{ p_2(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) (\tilde{\mathbf{X}}_{ni}^T \tilde{\mathbf{u}}_n)^2 \} \\ &\quad + \frac{r_n^2}{2n} \sum_{i=1}^n [p_2(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) (\tilde{\mathbf{X}}_{ni}^T \tilde{\mathbf{u}}_n)^2 \\ &\quad - \mathbb{E}_{K_{n,\epsilon}} \{ p_2(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) (\tilde{\mathbf{X}}_{ni}^T \tilde{\mathbf{u}}_n)^2 \}] \\ &\equiv I_{1,2,1} + I_{1,2,2}, \end{aligned}$$

where  $I_{1,2,1} = 2^{-1} r_n^2 \tilde{\mathbf{u}}_n^T \mathbf{H}_{n,\epsilon} \tilde{\mathbf{u}}_n$ . Meanwhile, we have

$$\begin{aligned} |I_{1,2,2}| &\leq \frac{r_n^2}{2} \left\| \frac{1}{n} \sum_{i=1}^n [p_2(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) \tilde{\mathbf{X}}_{ni} \tilde{\mathbf{X}}_{ni}^T \right. \\ &\quad \left. - \mathbb{E}_{K_{n,\epsilon}} \{ p_2(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) \tilde{\mathbf{X}}_{ni} \tilde{\mathbf{X}}_{ni}^T \}] \right\|_F \|\tilde{\mathbf{u}}_n\|^2 \\ &= r_n^2 O_P(p_n / \sqrt{n}) \|\tilde{\mathbf{u}}_n\|^2. \end{aligned}$$

Thus,

$$I_{1,2} = 2^{-1} r_n^2 \tilde{\mathbf{u}}_n^T \mathbf{H}_{n,\epsilon} \tilde{\mathbf{u}}_n + O_P(r_n^2 p_n / \sqrt{n}) \|\tilde{\mathbf{u}}_n\|^2. \quad (\text{A7})$$

For  $I_{1,3}$  in Equation (A5), we observe that

$$|I_{1,3}| \leq C r_n^3 \frac{1}{n} \sum_{i=1}^n |p_3(Y_i; \tilde{\mathbf{X}}_{ni}^T \tilde{\beta}_n^*)| w(\mathbf{X}_{ni}) |\tilde{\mathbf{X}}_{ni}^T \tilde{\mathbf{u}}_n|^3 = O_P(r_n^3 p_n^{3/2}) \|\tilde{\mathbf{u}}_n\|^3.$$

We will show that the minimum eigenvalue of  $\mathbf{H}_{n,\epsilon}$  is uniformly bounded away from 0.  $\mathbf{H}_{n,\epsilon} = (1 - \epsilon / \sqrt{n}) \mathbf{H}_{n,K_{n,0},T(K_{n,\epsilon})} + \epsilon / \sqrt{n} \mathbf{H}_{n,J,T(K_{n,\epsilon})}$ . Note

$$\begin{aligned} &\|\mathbf{H}_{n,K_{n,0},T(K_{n,\epsilon})} - \mathbf{H}_n\| \\ &= \|\mathbb{E}_{K_{n,0}} [\{p_2(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) - p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0})\} w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T]\| \\ &= \|\mathbb{E}_{K_{n,0}} [\{p_3(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_n^*) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\}\}]\| = O(p_n^2 / \sqrt{n}). \end{aligned}$$

Since the eigenvalues of  $\mathbf{H}_n$  are uniformly bounded away from 0, so are those of  $\mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}$  and  $\mathbf{H}_{n,\epsilon}$ .

We can choose some large  $C_\kappa$  such that  $I_{1,1}$  and  $I_{1,3}$  are both dominated by the first term of  $I_{1,2}$  in Equation (A7), which is positive by the eigenvalue assumption. This implies Equation (A4).

Next we show the uniqueness of  $\hat{\beta}$ . For all  $\tilde{\beta}$  such that  $\|\tilde{\beta} - T(K_{n,\epsilon})\| = O(n^{-1/4} p_n^{-1/2})$ ,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \tilde{X}_{ni}^T \tilde{\beta}) w(\mathbf{X}_{ni}) \tilde{X}_{ni} \tilde{X}_{ni}^T \\ = & E_{K_{n,0}} \{ p_2(Y; \tilde{X}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{X}_n \tilde{X}_n^T \} \\ & + E_{K_{n,0}} [\{ p_2(Y; \tilde{X}_n^T \tilde{\beta}) - p_2(Y; \tilde{X}_n^T \tilde{\beta}_{n,0}) \} w(\mathbf{X}_n) \tilde{X}_n \tilde{X}_n^T] \\ & + [E_{K_{n,\epsilon}} \{ p_2(Y; \tilde{X}_n^T \tilde{\beta}) w(\mathbf{X}_n) \tilde{X}_n \tilde{X}_n^T \} - E_{K_{n,0}} \{ p_2(Y; \tilde{X}_n^T \tilde{\beta}) w(\mathbf{X}_n) \tilde{X}_n \tilde{X}_n^T \}] \\ & + \left[ \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \tilde{X}_{ni}^T \tilde{\beta}) w(\mathbf{X}_{ni}) \tilde{X}_{ni} \tilde{X}_{ni}^T - E_{K_{n,\epsilon}} \{ p_2(Y; \tilde{X}_n^T \tilde{\beta}) w(\mathbf{X}_n) \tilde{X}_n \tilde{X}_n^T \} \right] \\ = & I_1^* + I_2^* + I_3^* + I_4^*. \end{aligned}$$

We know that the minimum eigenvalues of  $I_1^*$  are uniformly bounded away from 0. Following the proof of Lemma A1, we have  $\|I_2^*\| = o(1)$  and  $\|I_3^*\| = o(1)$ . It's easy to see  $\|I_4^*\| = O_p(p_n / \sqrt{n})$ .

Hence, for  $n$  large enough,  $\frac{\partial^2}{\partial \beta^2} \ell_{\mathbb{K}_n}(\tilde{\beta})$  is positive definite with high probability for all  $\tilde{\beta}$  such that  $\|\tilde{\beta} - \tilde{\beta}_{n,0}\| = O(n^{-1/4} p_n^{-1/2})$ . Therefore, there exists a unique minimizer of  $\ell_{\mathbb{K}_n}(\tilde{\beta})$  in the  $n^{-1/4} p_n^{-1/2}$  neighborhood of  $T(K_{n,\epsilon})$  which covers  $\hat{\beta}$ .  $\square$

**Lemma A3** ( $\|A_n \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\}\|$ ). Assume Conditions A0–A7 and B4. For  $K_{n,\epsilon}$  in Equation (10) and  $T(\cdot)$  in Equation (12), if  $p_n^5/n \rightarrow 0$  as  $n \rightarrow \infty$ , the distribution of  $(\mathbf{X}_n, Y)$  is  $K_{n,\epsilon}$  and  $E_J(\|w(\mathbf{X}_n)\mathbf{X}_n\|) \leq C$ , then

$$\sqrt{n} A_n \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\} = O(1),$$

where  $A_n$  is any given  $k \times (p_n + 1)$  matrix such that  $A_n A_n^T \rightarrow \mathbb{G}$ , with  $\mathbb{G}$  being a  $k \times k$  positive-definite matrix and  $k$  is a fixed integer.

**Proof.** Taylor's expansion yields

$$\begin{aligned} \mathbf{0} &= E_{K_{n,\epsilon}} \{ p_1(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{X}_n \} \\ &= E_{K_{n,\epsilon}} \{ p_1(Y; \tilde{X}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{X}_n \} \\ &\quad + E_{K_{n,\epsilon}} \{ p_2(Y; \tilde{X}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{X}_n \tilde{X}_n^T \} \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\} \\ &\quad + 1/2 E_{K_{n,\epsilon}} \{ p_3(Y; \tilde{X}_n^T \tilde{\beta}_n^*) w(\mathbf{X}_n) \tilde{X}_n [\tilde{X}_n^T \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\}]^2 \} \\ &= I_1 + I_2 \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\} + I_3, \end{aligned}$$

where  $\tilde{\beta}_n^*$  lies between  $T(K_{n,\epsilon})$  and  $\tilde{\beta}_{n,0}$ . Below, we will show

$$\|I_1\| = O(1/\sqrt{n}), \quad \|I_2 - \mathbf{H}_n\| = O(p_n / \sqrt{n}), \quad \|I_3\| = O(p_n^{5/2}/n).$$

First,  $\|I_1\| = \epsilon / \sqrt{n} \|E_J\{p_1(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{X}_n\}\| \leq C\epsilon / \sqrt{n} E_J(\|w(\mathbf{X}_n)\mathbf{X}_n\|) = O(1/\sqrt{n})$ . Following the proof of  $I_3^*$  in Lemma A1,  $\|I_2 - \mathbf{H}_n\| = O(p_n / \sqrt{n})$ . Since  $\|T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\| = O(\sqrt{p_n/n})$ , we have  $\|I_3\| = O(p_n^{5/2}/n)$ .

Therefore,  $\sqrt{n} A_n \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\} = -\sqrt{n} A_n \mathbf{H}_n^{-1} I_1 + o(1)$ , which completes the proof.  $\square$

**Lemma A4** (asymptotic normality of  $\mathbf{T}(\mathbb{K}_n) - \mathbf{T}(K_{n,\epsilon})$ ). Assume Conditions A0–A8 and B4. If  $p_n^5/n \rightarrow 0$  as  $n \rightarrow \infty$  and the distribution of  $(\mathbf{X}_n, Y)$  is  $K_{n,\epsilon}$ , then

$$\sqrt{n}\{U(K_{n,\epsilon})\}^{-1/2}A_n\{\mathbf{T}(\mathbb{K}_n) - \mathbf{T}(K_{n,\epsilon})\} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_k),$$

where  $U(K_{n,\epsilon}) = A_n \mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1} A_n^T$ ,  $A_n$  is any given  $k \times (p_n + 1)$  matrix such that  $A_n A_n^T \rightarrow \mathbb{G}$ , with  $\mathbb{G}$  being a  $k \times k$  positive-definite matrix,  $k$  is a fixed integer.

**Proof.** We will first show that

$$\mathbf{T}(\mathbb{K}_n) - \mathbf{T}(K_{n,\epsilon}) = -\frac{1}{n} \mathbf{H}_{n,\epsilon}^{-1} \sum_{i=1}^n p_1(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) \tilde{\mathbf{X}}_{ni} + o_P(n^{-1/2}). \quad (\text{A8})$$

From  $\frac{\partial \ell_{\mathbb{K}_n}(\tilde{\beta})}{\partial \tilde{\beta}}|_{\tilde{\beta}=\mathbf{T}(\mathbb{K}_n)} = \mathbf{0}$ , Taylor's expansion yields

$$\begin{aligned} \mathbf{0} &= \left\{ \frac{1}{n} \sum_{i=1}^n p_1(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) \tilde{\mathbf{X}}_{ni} \right\} \\ &\quad + \left\{ \frac{1}{n} \sum_{i=1}^n p_2(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) \tilde{\mathbf{X}}_{ni} \tilde{\mathbf{X}}_{ni}^T \right\} \{\mathbf{T}(\mathbb{K}_n) - \mathbf{T}(K_{n,\epsilon})\} \\ &\quad + \frac{1}{2n} \sum_{i=1}^n p_3(Y_i; \tilde{\mathbf{X}}_{ni}^T \tilde{\beta}_n^*) w(\mathbf{X}_{ni}) [\tilde{\mathbf{X}}_{ni}^T \{\mathbf{T}(\mathbb{K}_n) - \mathbf{T}(K_{n,\epsilon})\}]^2 \tilde{\mathbf{X}}_{ni} \\ &\equiv \left\{ \frac{1}{n} \sum_{i=1}^n p_1(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) \tilde{\mathbf{X}}_{ni} \right\} + I_2 \{\mathbf{T}(\mathbb{K}_n) - \mathbf{T}(K_{n,\epsilon})\} + I_3, \end{aligned} \quad (\text{A9})$$

where  $\tilde{\beta}_n^*$  lies between  $\mathbf{T}(K_{n,\epsilon})$  and  $\mathbf{T}(\mathbb{K}_n)$ . Below, we will show

$$\|I_2 - \mathbf{H}_{n,\epsilon}\| = O_P(p_n/\sqrt{n}), \quad \|I_3\| = O_P(p_n^{5/2}/n).$$

Similar arguments for the proof of  $I_{1,2}$  of Lemma A2, we have  $\|I_2 - \mathbf{H}_{n,\epsilon}\| = O_P(p_n/\sqrt{n})$ .

Second, a similar proof used for  $I_{1,3}^*$  in Equation (A5) gives  $\|I_3\| = O_P(p_n^{5/2}/n)$ .

Third, by Equation (A9) and  $\|\mathbf{T}(\mathbb{K}_n) - \mathbf{T}(K_{n,\epsilon})\| = O_P(\sqrt{p_n/n})$ , we see that

$$\mathbf{H}_{n,\epsilon}\{\mathbf{T}(\mathbb{K}_n) - \mathbf{T}(K_{n,\epsilon})\} = -\frac{1}{n} \sum_{i=1}^n p_1(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) \tilde{\mathbf{X}}_{ni} + \mathbf{u}_n,$$

where  $\|\mathbf{u}_n\| = O_P(p_n^{5/2}/n) = o_P(n^{-1/2})$ . From the proof of Lemma A2, the eigenvalues of  $\mathbf{H}_{n,\epsilon}$  are uniformly bounded away from 0 and we complete the proof of Equation (A8).

Following the proof for the bounded eigenvalues of  $\mathbf{H}_{n,\epsilon}$  in Lemma A2, we can show that the eigenvalues of  $\Omega_{n,\epsilon}$  are uniformly bounded away from 0. Hence, the eigenvalues of  $\mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1}$  are uniformly bounded away from 0, as are the eigenvalues of  $U(K_{n,\epsilon})$ . From Equation (A8), we see that

$$A_n\{\mathbf{T}(\mathbb{K}_n) - \mathbf{T}(K_{n,\epsilon})\} = -\frac{1}{n} A_n \mathbf{H}_{n,\epsilon}^{-1} \sum_{i=1}^n p_1(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) \tilde{\mathbf{X}}_{ni} + o_P(n^{-1/2}).$$

It follows that

$$\sqrt{n}\{U(K_{n,\epsilon})\}^{-1/2}A_n\{\mathbf{T}(\mathbb{K}_n) - \mathbf{T}(K_{n,\epsilon})\} = \sum_{i=1}^n \mathbf{R}_{ni} + o_P(1),$$

where  $\mathbf{R}_{ni} = -n^{-1/2}\{U(K_{n,\epsilon})\}^{-1/2}A_n \mathbf{H}_{n,\epsilon}^{-1} p_1(Y_i; \tilde{\mathbf{X}}_{ni}^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_{ni}) \tilde{\mathbf{X}}_{ni}$ . Following (A6) in Lemma A2, one can show that  $E_{K_{n,\epsilon}}(\mathbf{R}_{ni}) = \mathbf{0}$  for  $n$  large enough.

To show  $\sum_{i=1}^n \mathbf{R}_{ni} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_k)$ , we apply the Lindeberg-Feller central limit theorem in [26]. Specifically, we check (I)  $\sum_{i=1}^n \text{cov}_{K_{n,\epsilon}}(\mathbf{R}_{ni}) \rightarrow \mathbf{I}_k$ ; (II)  $\sum_{i=1}^n E_{K_{n,\epsilon}}(\|\mathbf{R}_{ni}\|^{2+\delta}) = o(1)$  for some  $\delta > 0$ . Condition (I) is straightforward since  $\sum_{i=1}^n \text{cov}_{K_{n,\epsilon}}(\mathbf{R}_{ni}) = \{U(K_{n,\epsilon})\}^{-1/2} U(K_{n,\epsilon}) \{U(K_{n,\epsilon})\}^{-1/2} = \mathbf{I}_k$ . To check condition (II), we can show that  $E_{K_{n,\epsilon}}(\|\mathbf{R}_{ni}\|^{2+\delta}) = O((p_n/n)^{(2+\delta)/2})$ . This yields  $\sum_{i=1}^n E_{K_{n,\epsilon}}(\|\mathbf{R}_{ni}\|^{2+\delta}) \leq O(p_n^{(2+\delta)/2}/n^{\delta/2}) = o(1)$ . Hence

$$\sqrt{n}\{U(K_{n,\epsilon})\}^{-1/2} A_n \{T(\mathbb{K}_n) - T(K_{n,\epsilon})\} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_k).$$

Thus, we complete the proof.  $\square$

**Lemma A5** (asymptotic covariance matrices  $U(K_{n,\epsilon})$  and  $U_n$ ). *Assume Conditions A0–A9 and B4. If  $p_n^4/n \rightarrow 0$  as  $n \rightarrow \infty$ , then*

$$\|U_n^{-1/2}\{U(K_{n,\epsilon})\}^{1/2} - \mathbf{I}_k\| = O(p_n/n^{1/4}),$$

where  $U(K_{n,\epsilon}) = A_n \mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1} A_n^T$ ,  $A_n$  is any given  $k \times (p_n + 1)$  matrix such that  $A_n A_n^T \rightarrow \mathbb{G}$ , with  $\mathbb{G}$  being a  $k \times k$  positive-definite matrix, and  $k$  is a fixed integer.

**Proof.** Note that

$$\begin{aligned} & \| \{U(K_{n,\epsilon})\}^{1/2} - U_n^{1/2} \|^2 \leq \|U(K_{n,\epsilon}) - U_n\| \\ & \leq \|\mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1} - \mathbf{H}_n^{-1} \Omega_n \mathbf{H}_n^{-1}\| \|A_n\|_F^2. \end{aligned}$$

Since  $\|A_n\|_F^2 \rightarrow \text{tr}(\mathbb{G})$ , it suffices to prove that  $\|\mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1} - \mathbf{H}_n^{-1} \Omega_n \mathbf{H}_n^{-1}\| = O(p_n^2/\sqrt{n})$ .

First, we prove  $\|\mathbf{H}_{n,\epsilon} - \mathbf{H}_n\| = O(p_n^2/\sqrt{n})$ . Note that

$$\begin{aligned} \mathbf{H}_{n,\epsilon} - \mathbf{H}_n &= E_{K_{n,\epsilon}}[\{p_2(Y; \tilde{\mathbf{X}}_n^T T(K_{n,\epsilon})) - p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0})\} w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T] \\ &\quad + [E_{K_{n,\epsilon}}\{p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T\} - \mathbf{H}_n] \\ &= E_{K_{n,\epsilon}}[p_3(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}^*) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \tilde{\mathbf{X}}_n^T \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\}] \\ &\quad + [E_{K_{n,\epsilon}}\{p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T\} - \mathbf{H}_n] \\ &\equiv I_1 + I_2. \end{aligned}$$

We know that  $\|I_1\| = O(p_n^2/\sqrt{n})$  and  $\|I_2\| = O(p_n/\sqrt{n})$ . Thus,  $\|I_1\| = O(p_n^2/\sqrt{n})$ .

Second, we show  $\|\Omega_{n,\epsilon} - \Omega_n\| = O(p_n^2/\sqrt{n})$ . It is easy to see that

$$\begin{aligned} \Omega_{n,\epsilon} - \Omega_n &= E_{K_{n,\epsilon}}[\{p_1^2(Y; \tilde{\mathbf{X}}_n^T T(K_{n,\epsilon})) - p_1^2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0})\} w^2(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T] \\ &\quad + [E_{K_{n,\epsilon}}\{p_1^2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0}) w^2(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T\} - \Omega_n] \\ &= \Delta_{1,1} + \Delta_{1,2}, \end{aligned}$$

where  $\|\Delta_{1,1}\| = O(p_n^2/\sqrt{n})$  and  $\|\Delta_{1,2}\| = O(p_n/\sqrt{n})$ . We observe that  $\|\Omega_{n,\epsilon} - \Omega_n\| = O(p_n^2/\sqrt{n})$ .

Third, we show  $\|\mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1} - \mathbf{H}_n^{-1} \Omega_n \mathbf{H}_n^{-1}\| = O(p_n^2/\sqrt{n})$ . Note  $\mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1} - \mathbf{H}_n^{-1} \Omega_n \mathbf{H}_n^{-1} = L_1 + L_2 + L_3$ , where  $L_1 = \mathbf{H}_{n,\epsilon}^{-1} (\Omega_{n,\epsilon} - \Omega_n) \mathbf{H}_{n,\epsilon}^{-1}$ ,  $L_2 = \mathbf{H}_{n,\epsilon}^{-1} (\mathbf{H}_n - \mathbf{H}_{n,\epsilon}) \mathbf{H}_n^{-1} \Omega_n \mathbf{H}_n^{-1}$  and  $L_3 = \mathbf{H}_n^{-1} \Omega_n \mathbf{H}_{n,\epsilon}^{-1} (\mathbf{H}_n - \mathbf{H}_{n,\epsilon}) \mathbf{H}_n^{-1}$ . Under Conditions A7 and A9, it is straightforward to see that  $\|\mathbf{H}_{n,\epsilon}^{-1}\| = O(1)$ ,  $\|\mathbf{H}_n^{-1}\| = O(1)$  and  $\|\mathbf{H}_n^{-1} \Omega_n\| = O(1)$ . Since  $\|L_1\| \leq \|\mathbf{H}_{n,\epsilon}^{-1}\| \|\Omega_{n,\epsilon} - \Omega_n\| \|\mathbf{H}_{n,\epsilon}^{-1}\|$ , we conclude  $\|L_1\| = O(p_n^2/\sqrt{n})$ , and similarly  $\|L_2\| = O(p_n^2/\sqrt{n})$  and  $\|L_3\| = O(p_n^2/\sqrt{n})$ . Hence,  $\|\mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1} - \mathbf{H}_n^{-1} \Omega_n \mathbf{H}_n^{-1}\| = O(p_n^2/\sqrt{n})$ .

Thus, we can conclude that  $\|U(K_{n,\epsilon}) - U_n\| = O(p_n^2/\sqrt{n})$  and that the eigenvalues of  $U(K_{n,\epsilon})$  and  $U_n$  are uniformly bounded away from 0 and  $\infty$ . Consequently,  $\|\{U(K_{n,\epsilon})\}^{1/2} - U_n^{1/2}\| = O(p_n/n^{1/4})$  and proof is finished.  $\square$

**Lemma A6** (asymptotic covariance matrices  $U(\mathbb{K}_n)$  and  $U(K_{n,\epsilon})$ ). Assume Conditions A0–A9 and B4. If  $p_n^4/n \rightarrow 0$  as  $n \rightarrow \infty$  and the distribution of  $(\mathbf{X}_n, Y)$  is  $K_{n,\epsilon}$ , then

$$\|\{U(\mathbb{K}_n)\}^{-1/2}\{U(K_{n,\epsilon})\}^{1/2} - \mathbf{I}_k\| = O_P(p_n/n^{1/4}),$$

where  $U(K_{n,\epsilon}) = A_n \mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1} A_n^T$ ,  $U(\mathbb{K}_n) = A_n \widehat{\mathbf{H}}_n^{-1} \widehat{\Omega}_n \widehat{\mathbf{H}}_n^{-1} A_n^T$ ,  $A_n$  is any given  $k \times (p_n + 1)$  matrix such that  $A_n A_n^T \rightarrow \mathbb{G}$ , with  $\mathbb{G}$  being a  $k \times k$  positive-definite matrix, and  $k$  is a fixed integer.

**Proof.** Note that  $\|\{U(\mathbb{K}_n)\}^{1/2} - \{U(K_{n,\epsilon})\}^{1/2}\|^2 \leq \|U(\mathbb{K}_n) - U(K_{n,\epsilon})\| \leq \|\widehat{\mathbf{H}}_n^{-1} \widehat{\Omega}_n \widehat{\mathbf{H}}_n^{-1} - \mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1}\| \|A_n\|_F^2$ . Since  $\|A_n\|_F^2 \rightarrow \text{tr}(\mathbb{G})$ , it suffices to prove that  $\|\widehat{\mathbf{H}}_n^{-1} \widehat{\Omega}_n \widehat{\mathbf{H}}_n^{-1} - \mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1}\| = O_P(p_n^2/\sqrt{n})$ .

Following the proof of Proposition 1 in [1], we can show that  $\|\widehat{\mathbf{H}}_n - \mathbf{H}_{n,\epsilon}\| = O_P(p_n^2/\sqrt{n})$  and  $\|\widehat{\Omega}_n - \Omega_{n,\epsilon}\| = O_P(p_n^2/\sqrt{n})$ .

To show  $\|\widehat{\mathbf{H}}_n^{-1} \widehat{\Omega}_n \widehat{\mathbf{H}}_n^{-1} - \mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1}\| = O_P(p_n^2/\sqrt{n})$ , note  $\widehat{\mathbf{H}}_n^{-1} \widehat{\Omega}_n \widehat{\mathbf{H}}_n^{-1} - \mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1} = L_1 + L_2 + L_3$ , where  $L_1 = \widehat{\mathbf{H}}_n^{-1} (\widehat{\Omega}_n - \Omega_{n,\epsilon}) \widehat{\mathbf{H}}_n^{-1}$ ,  $L_2 = \widehat{\mathbf{H}}_n^{-1} (\mathbf{H}_{n,\epsilon} - \widehat{\mathbf{H}}_n) \mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \widehat{\mathbf{H}}_n^{-1}$  and  $L_3 = \mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \widehat{\mathbf{H}}_n^{-1} (\mathbf{H}_{n,\epsilon} - \widehat{\mathbf{H}}_n) \mathbf{H}_{n,\epsilon}^{-1}$ . Following the proof in Lemma A2, it is straightforward to verify that  $\|\mathbf{H}_{n,\epsilon}^{-1}\| = O(1)$ ,  $\|\widehat{\mathbf{H}}_n^{-1}\| = O_P(1)$ . In addition,  $\|\mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon}\| = \|(\mathbf{H}_{n,\epsilon}^{-1} - \mathbf{H}_n^{-1}) \Omega_{n,\epsilon} + \mathbf{H}_n^{-1} (\Omega_{n,\epsilon} - \Omega_n) + \mathbf{H}_n^{-1} \Omega_n\| \leq \|\mathbf{H}_{n,\epsilon}^{-1}\| \|\mathbf{H}_{n,\epsilon} - \mathbf{H}_n\| \|\mathbf{H}_n^{-1}\| \|\Omega_{n,\epsilon}\| + \|\mathbf{H}_n^{-1}\| \|\Omega_{n,\epsilon} - \Omega_n\| + \|\mathbf{H}_n^{-1} \Omega_n\| = O(1)$ .

Since  $\|L_1\| \leq \|\widehat{\mathbf{H}}_n^{-1}\| \|\widehat{\Omega}_n - \Omega_{n,\epsilon}\| \|\widehat{\mathbf{H}}_n^{-1}\|$ , we conclude  $\|L_1\| = O_P(p_n^2/\sqrt{n})$ , and similarly  $\|L_2\| = O_P(p_n^2/\sqrt{n})$  and  $\|L_3\| = O_P(p_n^2/\sqrt{n})$ . Hence,  $\|\widehat{\mathbf{H}}_n^{-1} \widehat{\Omega}_n \widehat{\mathbf{H}}_n^{-1} - \mathbf{H}_{n,\epsilon}^{-1} \Omega_{n,\epsilon} \mathbf{H}_{n,\epsilon}^{-1}\| = O_P(p_n^2/\sqrt{n})$ .

Thus, we can conclude that  $\|U(\mathbb{K}_n) - U(K_{n,\epsilon})\| = O_P(p_n^2/\sqrt{n})$  and the eigenvalues of  $U(\mathbb{K}_n)$  are uniformly bounded away from 0 and  $\infty$  with probability tending to 1. Noting that  $\|\{U(\mathbb{K}_n)\}^{1/2} - \{U(K_{n,\epsilon})\}^{1/2}\|^2 \leq \|U(\mathbb{K}_n) - U(K_{n,\epsilon})\|$ .  $\square$

**Lemma A7** (asymptotic distribution of test statistic). Assume Conditions A0–A9 and B4. If  $p_n^6/n \rightarrow 0$  as  $n \rightarrow \infty$  and the distribution of  $(\mathbf{X}_n, Y)$  is  $K_{n,\epsilon}$ , then

$$\sqrt{n}[\{U(\mathbb{K}_n)\}^{-1/2} A_n \{T(\mathbb{K}_n) - \tilde{\beta}_{n,0}\} - U_n^{-1/2} A_n \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\}] \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_k),$$

where  $A_n$  is any given  $k \times (p_n + 1)$  matrix such that  $A_n A_n^T \rightarrow \mathbb{G}$ , with  $\mathbb{G}$  being a  $k \times k$  positive-definite matrix, and  $k$  is a fixed integer.

**Proof.** Note that

$$\begin{aligned} & \sqrt{n}[\{U(\mathbb{K}_n)\}^{-1/2} A_n \{T(\mathbb{K}_n) - \tilde{\beta}_{n,0}\} - U_n^{-1/2} A_n \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\}] \\ &= \sqrt{n}[\{U(\mathbb{K}_n)\}^{-1/2} A_n \{T(\mathbb{K}_n) - T(K_{n,\epsilon})\} \\ &\quad + \sqrt{n}[\{U(\mathbb{K}_n)\}^{-1/2} - \{U(K_{n,\epsilon})\}^{-1/2}] A_n \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\} \\ &\quad + \sqrt{n}[\{U(K_{n,\epsilon})\}^{-1/2} - U_n^{-1/2}] A_n \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\}] \\ &\equiv \text{I} + \text{II} + \text{III}. \end{aligned}$$

For term I, we obtain from Lemma A4 that  $\sqrt{n}[\{U(K_{n,\epsilon})\}^{-1/2} A_n (T(\mathbb{K}_n) - T(K_{n,\epsilon}))] \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_k)$ . From Lemma A6, we get  $\|\{U(\mathbb{K}_n)\}^{-1/2} \{U(K_{n,\epsilon})\}^{1/2} - \mathbf{I}_k\| = o_P(1)$ . Thus, by Slutsky theorem,

$$\text{I} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_k). \tag{A10}$$

For term II, we see from Lemma A6 that

$$\|\{U(\mathbb{K}_n)\}^{-1/2} - \{U(K_{n,\epsilon})\}^{-1/2}\| = O_P(p_n/n^{1/4}).$$

Since

$$\|A_n \{T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\}\| \leq \|A_n\| \|\mathbf{T}(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\| = O(\sqrt{p_n/n}).$$

Thus,

$$\|\Pi\| \leq \sqrt{n} \|\{U(\mathbb{K}_n)\}^{-1/2} - \{U(K_{n,\epsilon})\}^{-1/2}\| \|A_n\| \|T(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\| = O_P(p_n^{3/2}/n^{1/4}). \quad (\text{A11})$$

Similarly,  $\|\text{III}\| = o_P(1)$ . Combining (A10) and (A11) with Slutsky theorem completes the proof.  $\square$

**Lemma A8** (Influence Function IF). *Assume Conditions A1–A8 and B4. For any fixed sample size n,*

$$= -\mathbf{H}_{n,K_{t_0},T(K_{t_0})}^{-1} [\mathbb{E}_J \{\psi_{\text{RBD}}(\mathbf{Z}_n; T(K_{t_0}))\} - \mathbb{E}_{K_{n,0}} \{\psi_{\text{RBD}}(\mathbf{Z}_n; T(K_{t_0}))\}],$$

where  $K_{t_0} = (1-t_0)K_{n,0} + t_0 J$  and  $t_0$  is a positive constant such that  $t_0 \leq c/p_n^2$  with  $c > 0$  a sufficiently small constant. In addition,  $\|\mathbf{H}_{n,K_{t_0},T(K_{t_0})}^{-1}\| \leq C$  uniformly for all  $n$  and  $t_0$  such that  $t_0 \leq c/p_n^2$  with  $c > 0$  a sufficiently small constant.

**Proof.** We follow the proof of Theorem 5.1 in [27]. Note

$$\begin{aligned} & \lim_{t \rightarrow t_0} \frac{T((1-t)K_{n,0} + tJ) - T((1-t_0)K_{n,0} + t_0 J)}{t - t_0} \\ &= \lim_{\Delta \rightarrow 0} \frac{T(K_{t_0} + \Delta(J - K_{n,0})) - T(K_{t_0})}{\Delta}, \end{aligned}$$

where  $\Delta = t - t_0$ .

It suffices to prove that for any sequence  $\{\Delta_j\}_{j=1}^\infty$  such that  $\lim_{j \rightarrow \infty} \Delta_j = 0$ , we have

$$\begin{aligned} & \lim_{j \rightarrow \infty} \frac{T(K_{t_0} + \Delta_j(J - K_{n,0})) - T(K_{t_0})}{\Delta_j} \\ &= -\mathbf{H}_{n,K_{t_0},T(K_{t_0})}^{-1} [\mathbb{E}_J \{\psi_{\text{RBD}}(\mathbf{Z}_n; T(K_{t_0}))\} - \mathbb{E}_{K_{n,0}} \{\psi_{\text{RBD}}(\mathbf{Z}_n; T(K_{t_0}))\}]. \end{aligned}$$

Following similar proofs in Lemma A1, we can show that for  $t_0$  sufficiently small,

$$\|\tilde{\beta}_{n,0} - T(K_{t_0})\| \leq C t_0 \sqrt{p_n}. \quad (\text{A12})$$

Next we will show that the eigenvalues of  $\mathbf{H}_{n,K_{t_0},T(K_{t_0})}$  are bounded away from 0.

$$\begin{aligned} & \mathbf{H}_{n,K_{t_0},T(K_{t_0})} = (1-t_0)\mathbf{H}_{n,K_{n,0},T(K_{t_0})} + t_0 \mathbf{H}_{n,J,T(K_{t_0})} \\ &= (1-t_0)\mathbf{H}_n + t_0 \mathbf{H}_{n,J,\tilde{\beta}_{n,0}} + (1-t_0)\{\mathbf{H}_{n,K_{n,0},T(K_{t_0})} - \mathbf{H}_n\} \\ &\quad + t_0\{\mathbf{H}_{n,J,T(K_{t_0})} - \mathbf{H}_{n,J,\tilde{\beta}_{n,0}}\} = (1-t_0)I_1 + t_0 I_2 + I_3 + I_4. \end{aligned}$$

First,

$$\begin{aligned} \|I_3\| &\leq C \mathbb{E}_{K_{n,0}} \|\{p_2(Y; \tilde{\mathbf{X}}_n^T T(K_{t_0})) - p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0})\} w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T\| \\ &\leq Cp_n^{3/2} \|T(K_{t_0}) - \tilde{\beta}_{n,0}\| \leq Cp_n^{3/2} t_0. \end{aligned}$$

Similarly,  $\|I_2\| \leq Cp_n t_0$  and  $\|I_4\| \leq Cp_n^2 t_0^2$ . Since the eigenvalues of  $I_1$  are bounded away from zero,  $\|I_2\|$ ,  $\|I_3\|$  and  $\|I_4\|$  could be sufficiently small, we conclude that for  $t_0 \leq c/p_n^2$  when  $c$  is sufficiently small, the eigenvalues of  $\mathbf{H}_{n,K_{t_0},T(K_{t_0})}$  are uniformly bounded away from 0.

Define  $K_j = K_{t_0} + \Delta_j(J - K_{n,0})$ . Following similar arguments for (A6) in Lemma A2, for  $j$  large enough,  $E_{K_j}\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_j))\} = \mathbf{0}$ . We will only consider  $j$  large enough below. The two term Taylor expansion yields

$$\mathbf{0} = E_{K_j}\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_j))\} = E_{K_j}\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_{t_0}))\} + \mathbf{H}_{n, K_j, \tilde{\beta}_j^*}\{\mathbf{T}(K_j) - \mathbf{T}(K_{t_0})\}, \quad (\text{A13})$$

where  $\tilde{\beta}_j^*$  lies between  $\mathbf{T}(K_{t_0})$  and  $\mathbf{T}(K_j)$ .

Thus, from (A13) and the fact  $E_{K_j}\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_{t_0}))\} = \Delta_j[E_J\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_{t_0}))\} - E_{K_{n,0}}\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_{t_0}))\}]$ , we have

$$\begin{aligned} \mathbf{0} &= E_{K_j}\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_{t_0}))\} + \mathbf{H}_{n, K_{t_0}, T(K_{t_0})}\{\mathbf{T}(K_j) - \mathbf{T}(K_{t_0})\} \\ &\quad + \{\mathbf{H}_{n, K_j, \tilde{\beta}_j^*} - \mathbf{H}_{n, K_{t_0}, T(K_{t_0})}\}\{\mathbf{T}(K_j) - \mathbf{T}(K_{t_0})\} \\ &= \Delta_j[E_J\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_{t_0}))\} - E_{K_{n,0}}\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_{t_0}))\}] \\ &\quad + \mathbf{H}_{n, K_{t_0}, T(K_{t_0})}\{\mathbf{T}(K_j) - \mathbf{T}(K_{t_0})\} + (\mathbf{H}_{n, K_j, \tilde{\beta}_j^*} - \mathbf{H}_{n, K_{t_0}, T(K_{t_0})})\{\mathbf{T}(K_j) - \mathbf{T}(K_{t_0})\}, \end{aligned}$$

and we obtain that

$$\begin{aligned} &\mathbf{T}(K_j) - \mathbf{T}(K_{t_0}) \\ &= -\Delta_j \mathbf{H}_{n, K_{t_0}, T(K_{t_0})}^{-1} [E_J\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_{t_0}))\} - E_{K_{n,0}}\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_{t_0}))\}] \\ &\quad - \mathbf{H}_{n, K_j, \tilde{\beta}_j^*}^{-1} \{\mathbf{H}_{n, K_j, \tilde{\beta}_j^*} - \mathbf{H}_{n, K_{t_0}, T(K_{t_0})}\}\{\mathbf{T}(K_j) - \mathbf{T}(K_{t_0})\}. \end{aligned} \quad (\text{A14})$$

Next, we will show that  $\|\mathbf{H}_{n, K_j, \tilde{\beta}_j^*} - \mathbf{H}_{n, K_{t_0}, T(K_{t_0})}\| = o(1)$  as  $j \rightarrow \infty$  for any fixed  $n$ . Since  $\|\tilde{\beta}_j^* - \mathbf{T}(K_{t_0})\| \leq \|\mathbf{T}(K_j) - \mathbf{T}(K_{t_0})\| = O(\Delta_j)$ ,

$$\begin{aligned} &\|\mathbf{H}_{n, K_j, \tilde{\beta}_j^*} - \mathbf{H}_{n, K_{t_0}, \tilde{\beta}_j^*}\| \\ &= \Delta_j \|E_J\{p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_j^*) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T\} - E_{K_{n,0}}\{p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_j^*) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T\}\| \\ &= O(\Delta_j) = o(1) \quad \text{as } j \rightarrow \infty, \end{aligned} \quad (\text{A15})$$

and also,

$$\begin{aligned} &\|\mathbf{H}_{n, K_{t_0}, \tilde{\beta}_j^*} - \mathbf{H}_{n, K_{t_0}, T(K_{t_0})}\| \\ &= \|E_{K_{t_0}}[p_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_j^*) - p_2(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{t_0}))] w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T\| \\ &= o(1) \quad \text{as } j \rightarrow \infty. \end{aligned} \quad (\text{A16})$$

From Equations (A15) and (A16),

$$\|\mathbf{H}_{n, K_j, \tilde{\beta}_j^*} - \mathbf{H}_{n, K_{t_0}, T(K_{t_0})}\| = o(1) \quad \text{as } j \rightarrow \infty$$

which, together with Equations (A12) and (A14), implies that

$$\|\mathbf{T}(K_j) - \mathbf{T}(K_{t_0}) + \Delta_j \mathbf{H}_{n, K_{t_0}, T(K_{t_0})}^{-1} [E_J\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_{t_0}))\} - E_{K_{n,0}}\{\psi_{\text{RBD}}(\mathbf{Z}_n; \mathbf{T}(K_{t_0}))\}]\| = o(\Delta_j).$$

This completes the proof.  $\square$

**Lemma A9.** Assume Conditions A1–A8 and B4 and  $\sup_n E_J(\|w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\|) \leq C$ . Let  $H_k(\cdot; \delta)$  be the cumulative distribution function of  $\chi_k^2(\delta)$  distribution with  $\delta$  the noncentrality parameter. Denote  $\delta(\epsilon) = n \|U_n^{-1/2} \{A_n \mathbf{T}(K_{n,\epsilon}) - \mathbf{g}_0\}\|^2$ . Let  $b(\epsilon) = -H_k(x; \delta(\epsilon))$ . Then, for any fixed  $x > 0$ ,  $\sup_{\epsilon \in [0, C]} \limsup_{n \rightarrow \infty} |b^{(3)}(\epsilon)| \leq C$  under  $H_0$  and  $\sup_{\epsilon \in [0, C]} \limsup_{n \rightarrow \infty} |b''(\epsilon)| \leq C$  under  $H_{1n}$ .

**Proof.** Since  $b(\epsilon) = -H_k(x; \delta(\epsilon))$ , we have

$$\begin{aligned} b'(\epsilon) &= -\frac{\partial}{\partial \epsilon} H_k(x; \delta(\epsilon)) = \left\{ -\frac{\partial}{\partial \delta} H_k(x; \delta) \Big|_{\delta=\delta(\epsilon)} \right\} \left\{ \frac{\partial \delta(\epsilon)}{\partial \epsilon} \right\} \\ b''(\epsilon) &= \left\{ -\frac{\partial^2}{\partial \delta^2} H_k(x; \delta) \Big|_{\delta=\delta(\epsilon)} \right\} \left\{ \frac{\partial \delta(\epsilon)}{\partial \epsilon} \right\}^2 + \left\{ -\frac{\partial}{\partial \delta} H_k(x; \delta) \Big|_{\delta=\delta(\epsilon)} \right\} \left\{ \frac{\partial^2 \delta(\epsilon)}{\partial \epsilon^2} \right\} \\ b^{(3)}(\epsilon) &= \left\{ -\frac{\partial^3}{\partial \delta^3} H_k(x; \delta) \Big|_{\delta=\delta(\epsilon)} \right\} \left\{ \frac{\partial \delta(\epsilon)}{\partial \epsilon} \right\}^3 \\ &\quad + 3 \left\{ -\frac{\partial^2}{\partial \delta^2} H_k(x; \delta) \Big|_{\delta=\delta(\epsilon)} \right\} \left\{ \frac{\partial \delta(\epsilon)}{\partial \epsilon} \right\} \left\{ \frac{\partial^2 \delta(\epsilon)}{\partial \epsilon^2} \right\} \\ &\quad + \left\{ -\frac{\partial}{\partial \delta} H_k(x; \delta) \Big|_{\delta=\delta(\epsilon)} \right\} \left\{ \frac{\partial^3 \delta(\epsilon)}{\partial \epsilon^3} \right\}. \end{aligned}$$

To complete the proof, we only need to show that  $\partial^i / \partial \delta^i H_k(x; \delta) |_{\delta=\delta(\epsilon)}$  and  $\partial^i \delta(\epsilon) / \partial \epsilon^i$  ( $i = 1, 2, 3$ ) are bounded as  $n \rightarrow \infty$  for all  $\epsilon \in [0, C]$ . Note that

$$H_k(x; \delta) = e^{-\delta/2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} \frac{\gamma(j+k/2, x/2)}{\Gamma(j+k/2)},$$

where  $\Gamma(\cdot)$  is the Gamma function, and  $\gamma(\cdot, \cdot)$  is the lower incomplete gamma function  $\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt$ , which satisfies  $\gamma(s, x) = (s-1)\gamma(s-1, x) - x^{s-1}e^{-x}$ . Therefore,

$$\begin{aligned} \frac{\partial}{\partial \delta} H_k(x; \delta) &= -\frac{e^{-\delta/2}}{2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} \frac{\gamma(j+k/2, x/2)}{\Gamma(j+k/2)} + \frac{e^{-\delta/2}}{2} \sum_{j=1}^{\infty} \frac{(\delta/2)^{j-1}}{(j-1)!} \frac{\gamma(j+k/2, x/2)}{\Gamma(j+k/2)} \\ &= \frac{1}{2} e^{-\delta/2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} \left\{ -\frac{\gamma(j+k/2, x/2)}{\Gamma(j+k/2)} + \frac{\gamma(j+1+k/2, x/2)}{\Gamma(j+1+k/2)} \right\}. \end{aligned}$$

Since

$$\begin{aligned} \frac{\gamma(j+1+k/2, x/2)}{\Gamma(j+1+k/2)} &= \frac{(j+k/2)\gamma(j+k/2, x/2) - (x/2)^{j+k/2}e^{-x/2}}{\Gamma(j+1+k/2)} \\ &= \frac{\gamma(j+k/2, x/2)}{\Gamma(j+k/2)} - \frac{(x/2)^{j+k/2}e^{-x/2}}{\Gamma(j+1+k/2)}, \end{aligned}$$

we have

$$\begin{aligned} \frac{\partial}{\partial \delta} H_k(x; \delta) &= -\frac{1}{2} e^{-\delta/2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} \frac{(x/2)^{j+k/2}e^{-x/2}}{\Gamma(j+1+k/2)} \\ \frac{\partial^2}{\partial \delta^2} H_k(x; \delta) &= \frac{1}{4} e^{-\delta/2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} \frac{(x/2)^{j+k/2}e^{-x/2}}{\Gamma(j+1+k/2)} - \frac{1}{4} e^{-\delta/2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} \frac{(x/2)^{j+1+k/2}e^{-x/2}}{\Gamma(j+2+k/2)} \\ &= \frac{1}{4} (x/2)^{k/2} e^{-x/2} e^{-\delta/2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} \left\{ \frac{(x/2)^j}{\Gamma(j+1+k/2)} - \frac{(x/2)^{j+1}}{\Gamma(j+2+k/2)} \right\} \\ &= \frac{1}{4} (x/2)^{k/2} e^{-x/2} e^{-\delta/2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} \frac{(x/2)^j}{\Gamma(j+1+k/2)} \left\{ 1 - \frac{(x/2)}{j+1+k/2} \right\} \\ \frac{\partial^3}{\partial \delta^3} H_k(x; \delta) &= -\frac{1}{8} (x/2)^{k/2} e^{-x/2} e^{-\delta/2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} \frac{(x/2)^j}{\Gamma(j+1+k/2)} \left\{ 1 - \frac{(x/2)}{j+1+k/2} \right\} \\ &\quad + \frac{1}{8} (x/2)^{k/2} e^{-x/2} e^{-\delta/2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} \frac{(x/2)^{j+1}}{\Gamma(j+2+k/2)} \left\{ 1 - \frac{(x/2)}{j+2+k/2} \right\} \\ &= \frac{1}{8} (x/2)^{k/2} e^{-x/2} e^{-\delta/2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} \frac{(x/2)^j}{\Gamma(j+1+k/2)} \end{aligned}$$

$$\cdot \left[ \frac{(x/2)}{j+1+k/2} \left\{ 1 - \frac{(x/2)}{j+2+k/2} \right\} - \left\{ 1 - \frac{(x/2)}{j+1+k/2} \right\} \right].$$

From the results of Lemma A3, that  $|\delta(\epsilon)|$  is bounded as  $n \rightarrow \infty$  for all  $\epsilon \in [0, C]$  under both  $H_0$  and  $H_{1n}$ , so are  $\partial^i/\partial\epsilon^i H_k(x; \delta)|_{\delta=\delta(\epsilon)}$  ( $i = 1, 2, 3$ ). Now, we consider the derivatives of  $\delta(\epsilon)$ ,

$$\begin{aligned}\frac{\partial \delta(\epsilon)}{\partial \epsilon} &= 2n \left\{ A_n \frac{\partial T(K_{n,\epsilon})}{\partial \epsilon} \right\}^T U_n^{-1} \{ A_n T(K_{n,\epsilon}) - g_0 \} \\ \frac{\partial^2 \delta(\epsilon)}{\partial \epsilon^2} &= 2n \left\{ A_n \frac{\partial T(K_{n,\epsilon})}{\partial \epsilon} \right\}^T U_n^{-1} \left\{ A_n \frac{\partial T(K_{n,\epsilon})}{\partial \epsilon} \right\} \\ &\quad + 2n \left\{ A_n \frac{\partial^2 T(K_{n,\epsilon})}{\partial \epsilon^2} \right\}^T U_n^{-1} \{ A_n T(K_{n,\epsilon}) - g_0 \} \\ \frac{\partial^3 \delta(\epsilon)}{\partial \epsilon^3} &= 6n \left\{ A_n \frac{\partial^2 T(K_{n,\epsilon})}{\partial \epsilon^2} \right\}^T U_n^{-1} \left\{ A_n \frac{\partial T(K_{n,\epsilon})}{\partial \epsilon} \right\} \\ &\quad + 2n \left\{ A_n \frac{\partial^3 T(K_{n,\epsilon})}{\partial \epsilon^3} \right\}^T U_n^{-1} \{ A_n T(K_{n,\epsilon}) - g_0 \}.\end{aligned}$$

To complete the proof, we only need to show that  $\sqrt{n} \|\partial^i/\partial\epsilon^i T(K_{n,\epsilon})\|$  ( $i = 1, 2, 3$ ) are bounded as  $n \rightarrow \infty$  for all  $\epsilon \in [0, C]$ , and  $\sqrt{n} \|A_n T(K_{n,\epsilon}) - g_0\|$  is bounded under  $H_0$  and  $H_{1n}$  as  $n \rightarrow \infty$  for all  $\epsilon \in [0, C]$ . The result for  $\sqrt{n} \|A_n T(K_{n,\epsilon}) - g_0\|$  is straightforward from Lemma A3.

First, for the first order derivative of  $T(K_{n,\epsilon})$ ,

$$\begin{aligned}&\sqrt{n} \frac{\partial}{\partial \epsilon} T(K_{n,\epsilon}) \\ &= -H_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1} [E_J \{ p_1(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(X_n) \tilde{X}_n \} - E_{K_{n,0}} \{ p_1(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(X_n) \tilde{X}_n \} ].\end{aligned}$$

Since  $\|H_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1}\| \leq C$ ,  $\|E_J \{ p_1(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(X_n) \tilde{X}_n \} \| \leq C E_J \|w(X_n) \tilde{X}_n\| \leq C$  and

$$\begin{aligned}&\|E_{K_{n,0}} \{ p_1(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(X_n) \tilde{X}_n \}\| \\ &= \|E_{K_{n,0}} \{ p_1(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(X_n) \tilde{X}_n \} - E_{K_{n,0}} \{ p_1(Y; \tilde{X}_n^T \tilde{\beta}_{n,0}) w(X_n) \tilde{X}_n \}\| \\ &= \|E_{K_{n,0}} [p_2(Y; \tilde{X}_n^T \tilde{\beta}^*) w(X_n) \tilde{X}_n \tilde{X}_n^T \{ T(K_{n,\epsilon}) - \tilde{\beta}_{n,0} \}]\| \\ &\leq Cp_n^{3/2}/\sqrt{n},\end{aligned}$$

we conclude that  $\sqrt{n} \|\partial/\partial\epsilon T(K_{n,\epsilon})\|$  is uniformly bounded for all  $\epsilon \in [0, C]$  as  $n \rightarrow \infty$ .

Second, for the second order derivative of  $T(K_{n,\epsilon})$ ,

$$\begin{aligned}&\sqrt{n} \frac{\partial^2}{\partial \epsilon^2} T((1 - \epsilon/\sqrt{n}) K_{n,0} + \epsilon/\sqrt{n} I) \\ &= -\frac{\partial H_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1}}{\partial \epsilon} \\ &\quad \cdot [E_J \{ p_1(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(X_n) \tilde{X}_n \} - E_{K_{n,0}} \{ p_1(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(X_n) \tilde{X}_n \}] \\ &\quad - H_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1} \\ &\quad \cdot \frac{\partial}{\partial \epsilon} [E_J \{ p_1(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(X_n) \tilde{X}_n \} - E_{K_{n,0}} \{ p_1(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(X_n) \tilde{X}_n \}]\end{aligned}$$

with

$$\begin{aligned}\frac{\partial}{\partial \epsilon} H_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1} &= -H_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1} \frac{\partial H_{n,K_{n,\epsilon},T(K_{n,\epsilon})}}{\partial \epsilon} H_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1}, \\ \frac{\partial H_{n,K_{n,\epsilon},T(K_{n,\epsilon})}}{\partial \epsilon} &= -\frac{1}{\sqrt{n}} E_{K_{n,0}} \{ p_2(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(X_n) \tilde{X}_n \tilde{X}_n^T \} \\ &\quad + \frac{1}{\sqrt{n}} E_J \{ p_2(Y; \tilde{X}_n^T T(K_{n,\epsilon})) w(X_n) \tilde{X}_n \tilde{X}_n^T \}\end{aligned}$$

$$\begin{aligned}
& + (1 - \epsilon / \sqrt{n}) E_{K_{n,0}} \left\{ p_3(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \tilde{\mathbf{X}}_n \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}) \right\} \\
& + \epsilon / \sqrt{n} E_J \left\{ p_3(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \tilde{\mathbf{X}}_n \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}) \right\}.
\end{aligned}$$

Therefore,  $\|\partial/\partial \epsilon \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1}\| \leq C \|\partial/\partial \epsilon \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}\| \leq Cp_n^{3/2}/\sqrt{n}$ . In addition,

$$\begin{aligned}
& \left\| \frac{\partial}{\partial \epsilon} [E_J \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\} - E_{K_{n,0}} \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\}] \right\| \\
& = \left\| E_J \left\{ p_2(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}) \right\} \right. \\
& \quad \left. - E_{K_{n,0}} \left\{ p_2(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}) \right\} \right\| \\
& \leq Cp_n / \sqrt{n}.
\end{aligned}$$

Therefore,  $\|\sqrt{n} \frac{\partial^2}{\partial \epsilon^2} \mathbf{T}((1 - \epsilon / \sqrt{n}) K_{n,0} + \epsilon / \sqrt{n} J)\| = o(1)$  for all  $\epsilon \in [0, C]$ .

Finally, for the third order derivative of  $\mathbf{T}(K_{n,\epsilon})$ ,

$$\begin{aligned}
& \sqrt{n} \frac{\partial^3}{\partial \epsilon^3} \mathbf{T}((1 - \epsilon / \sqrt{n}) K_{n,0} + \epsilon / \sqrt{n} J) \\
& = - \frac{\partial^2 \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1}}{\partial \epsilon^2} \\
& \quad \cdot [E_J \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\} - E_{K_{n,0}} \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\}] \\
& \quad - 2 \frac{\partial \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1}}{\partial \epsilon} \\
& \quad \cdot \frac{\partial}{\partial \epsilon} [E_J \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\} - E_{K_{n,0}} \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\}] \\
& \quad - \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1} \\
& \quad \cdot \frac{\partial^2}{\partial \epsilon^2} [E_J \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\} - E_{K_{n,0}} \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\}].
\end{aligned}$$

Note:

$$\begin{aligned}
& \frac{\partial^2}{\partial \epsilon^2} \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1} \\
& = - \frac{\partial \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1}}{\partial \epsilon} \frac{\partial \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}}{\partial \epsilon} \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1} \\
& \quad - \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1} \frac{\partial^2 \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}}{\partial \epsilon^2} \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1} \\
& \quad - \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1} \frac{\partial \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}}{\partial \epsilon} \frac{\partial \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1}}{\partial \epsilon},
\end{aligned}$$

where

$$\begin{aligned}
& \frac{\partial^2}{\partial \epsilon^2} \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1} \\
& = - \frac{2}{\sqrt{n}} E_{K_{n,0}} \left\{ p_3(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \tilde{\mathbf{X}}_n \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}) \right\} \\
& \quad + \frac{2}{\sqrt{n}} E_J \left\{ p_3(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \tilde{\mathbf{X}}_n \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}) \right\} \\
& \quad + (1 - \epsilon / \sqrt{n}) E_{K_{n,0}} \left\{ p_4(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T (\tilde{\mathbf{X}}_n \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}))^2 \right\} \\
& \quad + \epsilon / \sqrt{n} E_J \left\{ p_4(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T (\tilde{\mathbf{X}}_n \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}))^2 \right\}.
\end{aligned}$$

Hence,  $\|\frac{\partial^2}{\partial \epsilon^2} \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}\| \leq Cp_n^2/n$  which implies that  $\|\frac{\partial^2}{\partial \epsilon^2} \mathbf{H}_{n,K_{n,\epsilon},T(K_{n,\epsilon})}^{-1}\| = o(1)$  for all  $\epsilon \in [0, C]$ . In addition,

$$\begin{aligned} & \frac{\partial^2}{\partial \epsilon^2} [\mathbb{E}_J \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\} - \mathbb{E}_{K_{n,0}} \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\}] \\ &= \frac{\partial}{\partial \epsilon} \left[ \mathbb{E}_J \left\{ p_2(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}) \right\} \right. \\ &\quad \left. - \mathbb{E}_{K_{n,0}} \left\{ p_2(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}) \right\} \right] \\ &= \mathbb{E}_J \left\{ p_3(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \left( \tilde{\mathbf{X}}_n^T \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}) \right)^2 \right\} \\ &\quad + \mathbb{E}_J \left\{ p_2(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \frac{\partial^2}{\partial \epsilon^2} \mathbf{T}(K_{n,\epsilon}) \right\} \\ &\quad - \mathbb{E}_{\tilde{\beta}_{n,0}} \left\{ p_3(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \left( \tilde{\mathbf{X}}_n^T \frac{\partial}{\partial \epsilon} \mathbf{T}(K_{n,\epsilon}) \right)^2 \right\} \\ &\quad - \mathbb{E}_{\tilde{\beta}_{n,0}} \left\{ p_2(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T \frac{\partial^2}{\partial \epsilon^2} \mathbf{T}(K_{n,\epsilon}) \right\}. \end{aligned}$$

Hence,  $\|\frac{\partial^2}{\partial \epsilon^2} [\mathbb{E}_J \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\} - \mathbb{E}_{K_{n,0}} \{p_1(Y; \tilde{\mathbf{X}}_n^T \mathbf{T}(K_{n,\epsilon})) w(\mathbf{X}_n) \tilde{\mathbf{X}}_n\}]\| \leq Cp_n/\sqrt{n}$ . Therefore,  $\|\sqrt{n} \frac{\partial^2}{\partial \epsilon^3} \mathbf{T}((1-\epsilon/\sqrt{n})K_{n,0} + \epsilon/\sqrt{n}J)\| = o(1)$  for all  $\epsilon \in [0, C]$ . Hence, we complete the proof.  $\square$

**Proof of Theorem 1.** We follow the idea of the proof in [10]. Lemma A7 implies that the Wald-type test statistic  $W_n$  is asymptotically noncentral  $\chi_k^2$  with noncentrality parameter  $\delta(\epsilon) = n \|U_n^{-1/2} \{A_n \mathbf{T}(K_{n,\epsilon}) - g_0\}\|^2$ . Therefore,  $\alpha(K_{n,\epsilon}) = P(W_n > \eta_{1-\alpha_0}|H_0) = 1 - H_k(\eta_{1-\alpha_0}; \delta(\epsilon)) + h(n, \epsilon)$  where  $h(n, \epsilon) = \alpha(K_{n,\epsilon}) - 1 + H_k(\eta_{1-\alpha_0}; \delta(\epsilon)) \rightarrow 0$  as  $n \rightarrow \infty$  for any fixed  $\epsilon$ . Let  $b(\epsilon) = -H_k(\eta_{1-\alpha_0}; \delta(\epsilon))$ . Then, for  $\epsilon$  close to 0, we have

$$\begin{aligned} & \alpha(K_{n,\epsilon}) - \alpha_0 = b(\epsilon) - b(0) + h(n, \epsilon) - h(n, 0) \\ &= \epsilon b'(0) + \frac{1}{2} \epsilon^2 b''(0) + \frac{1}{6} \epsilon^3 b^{(3)}(\epsilon^*) + h(n, \epsilon) - h(n, 0), \end{aligned} \tag{A17}$$

where  $0 < \epsilon^* < \epsilon$ . Note that under  $H_0$

$$b'(0) = \mu_k \left\{ \frac{\partial \delta(\epsilon)}{\partial \epsilon} \Big|_{\epsilon=0} \right\} = 2\mu_k n \left\{ A_n \frac{\partial \mathbf{T}(K_{n,\epsilon})}{\partial \epsilon} \right\}^T \Big|_{\epsilon=0} U_n^{-1} \{A_n \tilde{\beta}_{n,0} - g_0\} = 0.$$

From Lemma A8, under  $H_0$

$$\frac{\partial \mathbf{T}(K_{n,\epsilon})}{\partial \epsilon} \Big|_{\epsilon=0} = 1/\sqrt{n} \mathbb{E}_J \{\text{IF}(\mathbf{Z}_n; \mathbf{T}, K_{n,0})\}.$$

Thus,

$$b''(0) = \mu_k \left\{ \frac{\partial^2 \delta(\epsilon)}{\partial \epsilon^2} \Big|_{\epsilon=0} \right\} = 2\mu_k \|U_n^{-1/2} A_n \mathbb{E}_J \{\text{IF}(\mathbf{Z}_n; \mathbf{T}, K_{n,0})\}\|^2.$$

Since from Lemma A8,  $\text{IF}(\mathbf{z}_n; \mathbf{T}, K_{n,0}) = -\mathbf{H}_n^{-1} \mathbb{E}_J \{\psi_{\text{RBD}}(\mathbf{z}_n; \tilde{\beta}_{n,0})\}$  is uniformly bounded, we have

$$D = \limsup_{n \rightarrow \infty} \|U_n^{-1/2} A_n \mathbb{E}_J \{\text{IF}(\mathbf{Z}_n; \mathbf{T}, K_{n,0})\}\|^2 < \infty.$$

From Equation (A17)

$$\limsup_{n \rightarrow \infty} \alpha(K_{n,\epsilon}) = \alpha_0 + \epsilon^2 \mu_k D + o(\epsilon^2),$$

since  $\sup_{\epsilon \in [0,C]} \limsup_{n \rightarrow \infty} |b^{(3)}(\epsilon)| \leq C$  from Lemma A9. We complete the proof.  $\square$

**Proof of Corollary 1.** For Part (i), following the proof of Theorem 1, for any fixed  $z$ ,

$$\lim_{n \rightarrow \infty} \alpha(K_{n,\epsilon}) = \alpha_0 + \epsilon^2 \mu_k \|U^{-1/2} A \text{IF}(z; T, K_0)\|^2 + d(z, \epsilon),$$

where  $d(z, \epsilon) = o(\epsilon^2)$ . From the assumption that  $\sup_{x \in \mathbb{R}^p} \|w(x)x\| \leq C$  and  $\sup_{\mu \in \mathbb{R}} |q''(\mu)| \sqrt{V(\mu)/F'(\mu)} \leq C$ , we know  $D_1 \leq \infty$ . Following the proof of Lemma A9,  $\sup_{z \in \mathbb{R}} |d(z, \epsilon)| = o(\epsilon^2)$ . We finished the proof of part (i).

Part (ii) is straightforward by applying Theorem 1 with  $J = \Delta_{z_n}$ .  $\square$

**Proof of Theorem 2.** Lemma A7 implies that

$$\begin{aligned} \sqrt{n}[\{U(\mathbb{K}_n)\}^{-1/2}\{A_n \mathbf{T}(\mathbb{K}_n) - g_0\} - \{U(\mathbb{K}_n)\}^{-1/2}(A_n \tilde{\beta}_{n,0} - g_0) \\ - U_n^{-1/2} A_n \{\mathbf{T}(K_{n,\epsilon}) - \tilde{\beta}_{n,0}\}] \xrightarrow{\mathcal{L}} (\mathbf{0}, \mathbf{I}_k). \end{aligned}$$

From Lemmas A5 and A6,

$$\sqrt{n}[\{U(\mathbb{K}_n)\}^{-1/2}\{A_n \mathbf{T}(\mathbb{K}_n) - g_0\} - U_n^{-1/2}\{A_n \mathbf{T}(K_{n,\epsilon}) - g_0\}] \xrightarrow{\mathcal{L}} (\mathbf{0}, \mathbf{I}_k).$$

Then,  $W_n$  is asymptotically  $\chi_k^2(\delta(\epsilon))$  with  $\delta(\epsilon) = n\|U_n^{-1/2}\{A_n \mathbf{T}(K_{n,\epsilon}) - g_0\}\|^2$  under  $H_{1n}$ . Therefore,  $\beta(K_{n,\epsilon}) = P(W_n > \eta_{1-\alpha_0} | H_{1n}) = 1 - H_k(\eta_{1-\alpha_0}; \delta(\epsilon)) + h(n, \epsilon)$ , where  $h(n, \epsilon) = \beta(K_{n,\epsilon}) - 1 + H_k(\eta_{1-\alpha_0}; \delta(\epsilon)) \rightarrow 0$  as  $n \rightarrow \infty$  for any fixed  $\epsilon$ . Let  $b(\epsilon) = -H_k(\eta_{1-\alpha_0}; \delta(\epsilon))$ . Then, for  $\epsilon$  close to 0, we have

$$\begin{aligned} \beta(K_{n,\epsilon}) - \beta_0 &= b(\epsilon) - b(0) + h(n, \epsilon) - h(n, 0) \\ &= \epsilon b'(0) + \frac{1}{2} \epsilon^2 b''(\epsilon^*) + h(n, \epsilon) - h(n, 0), \end{aligned} \tag{A18}$$

where  $0 < \epsilon^* < \epsilon$ . Note that under  $H_{1n}$ ,  $\delta(0) = n\|U_n^{-1/2}(A_n \tilde{\beta}_{n,0} - g_0)\|^2 = \mathbf{c}^T U_n^{-1} \mathbf{c}$ . Then,

$$\begin{aligned} b'(0) &= \left. \frac{-\partial H_k(\eta_{1-\alpha_0}; \delta)}{\partial \delta} \right|_{\delta=\delta(0)} \left. \frac{\partial \delta(\epsilon)}{\partial \epsilon} \right|_{\epsilon=0} = 2\nu_k n \left\{ A_n \frac{\partial \mathbf{T}(K_{n,\epsilon})}{\partial \epsilon} \right\}^T \Big|_{\epsilon=0} U_n^{-1} \{A_n \tilde{\beta}_{n,0} - g_0\} \\ &= 2\nu_k \sqrt{n} \left\{ A_n \frac{\partial \mathbf{T}(K_{n,\epsilon})}{\partial \epsilon} \right\}^T \Big|_{\epsilon=0} U_n^{-1} \mathbf{c}. \end{aligned}$$

From Lemma A8,

$$\left. \frac{\partial \mathbf{T}(K_{n,\epsilon})}{\partial \epsilon} \right|_{\epsilon=0} = 1/\sqrt{n} E_J \{\text{IF}(\mathbf{Z}_n; \mathbf{T}, K_{n,0})\},$$

and hence,

$$b'(0) = 2\nu_k \mathbf{c}^T U_n^{-1} A_n E_J \{\text{IF}(\mathbf{Z}_n; \mathbf{T}, K_{n,0})\}.$$

Since  $\sup_{\epsilon \in [0,C]} \limsup_{n \rightarrow \infty} |b''(\epsilon)| \leq C$  under  $H_{1n}$  by Lemma A9, we have  $\liminf_{n \rightarrow \infty} 1/2\epsilon^2 b''(\epsilon^*) = o(\epsilon)$  as  $\epsilon \rightarrow 0$ .

Since from Lemma A8,  $\text{IF}(\mathbf{z}_n; \mathbf{T}, K_{n,0}) = -\mathbf{H}_n^{-1} E_J \{\psi_{\text{RBD}}(\mathbf{z}_n; \tilde{\beta}_{n,0})\}$  is uniformly bounded,

$$|B| = |\liminf_{n \rightarrow \infty} 2\epsilon^2 b''(\epsilon^*)| < \infty.$$

From Equation (A18), we complete the proof.  $\square$

**Proof of Corollary 2.** The proof is similar to that for Corollary 1, using the results in Theorem 2.  $\square$

## Appendix B. List of Notations and Symbols

- $A_n$ :  $k \times (p_n + 1)$  matrix in hypotheses Equations (8) and (14)
- $c$ :  $k$  dimensional vector in  $H_{1n}$  in Equation (14)
- $F(\cdot)$ : link function
- $G$ : bias-correction term in “robust-BD”
- $\mathbb{G}$ : limit of  $A_n A_n^T$ , i.e.  $A_n A_n^T \xrightarrow{n \rightarrow \infty} \mathbb{G}$
- $\mathbf{H}_n$ :  $\mathbf{H}_n = E_{K_{n,0}}\{\mathbf{p}_2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0})w(\mathbf{X}_n)\tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T\}$
- $\text{IF}(\cdot, \cdot)$ : influence function
- $J$ : an arbitrary distribution in the contamination of Equation (10)
- $K_{n,0}$ : true parametric distribution of  $Z_n$
- $K_{n,\epsilon}$ :  $K_{n,0} + (1 - \frac{\epsilon}{\sqrt{n}})J$ ,  $\epsilon$ -contamination in Equation (10)
- $\mathbb{K}_n$ : empirical distribution of  $\{\mathbf{Z}_{ni}\}_{i=1}^n$
- $\ell_K(\cdot)$ : expectation of robust-BD in Equation (11)
- $m(\cdot)$ : conditional mean of  $Y$  given  $\mathbf{X}_n$  in Equation (1)
- $n$ : sample size
- $p_n$ : dimension of  $\beta$
- $p_i(\cdot, \cdot)$ :  $i$ th order derivative of robust-BD
- $q(\cdot)$ : generating  $q$ -function of BD
- $T(\cdot)$ : vector, a functional of estimator in Equation (12)
- $U_n$ :  $U_n = A_n \mathbf{H}_n^{-1} \Omega_n \mathbf{H}_n^{-1} A_n^T$
- $V(\cdot)$ : conditional variance of  $Y$  given  $\mathbf{X}_n$  in Equation (2)
- $W_n$ : Wald-type test statistic in Equation (9)
- $w(\cdot)$ : weight function
- $\mathbf{X}_n$ : explanatory variables
- $Y$ : response variable
- $\mathbf{Z}_n = (\mathbf{X}_n^T, Y)^T$
- $\alpha(\cdot)$ : level of the test
- $\beta(\cdot)$ : power of the test
- $\tilde{\beta}_{n,0}$ : true regression parameter
- $\Delta_{z_n}$ : probability measure which puts mass 1 at the point  $z_n$
- $\epsilon$ : amount of contamination in Equation (10), positive constant
- $\psi_{\text{RBD}}(\cdot, \cdot)$ : score vector in Equation (7)
- $\Omega_n$ :  $\Omega_n = E_{K_{n,0}}\{\mathbf{p}_1^2(Y; \tilde{\mathbf{X}}_n^T \tilde{\beta}_{n,0})w^2(\mathbf{X}_n)\tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^T\}$
- $\rho_q(\cdot, \cdot)$ : robust-BD in Equation (4)

## References

1. Zhang, C.M.; Guo, X.; Cheng, C.; Zhang, Z.J. Robust-BD estimation and inference for varying-dimensional general linear models. *Stat. Sin.* **2012**, *24*, 653–673.
2. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman & Hall: London, UK, 1989.
3. Morgenthaler, S. Least-absolute-deviations fits for generalized linear models. *Biometrika* **1992**, *79*, 747–754.
4. Ruckstuhl, A.F.; Welsh, A.H. Robust fitting of the binomial model. *Ann. Stat.* **2001**, *29*, 1117–1136.
5. Noh, M.; Lee, Y. Robust modeling for inference from generalized linear model classes. *J. Am. Stat. Assoc.* **2007**, *102*, 1059–1072.
6. Künsch, H.R.; Stefanski, L.A.; Carroll, R.J. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *J. Am. Stat. Assoc.* **1989**, *84*, 460–466.

7. Stefanski, L.A.; Carroll, R.J.; Ruppert, D. Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika* **1986**, *73*, 413–424.
8. Bianco, A.M.; Yohai, V.J. Robust estimation in the logistic regression model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods*; Springer: New York, NY, USA, 1996; pp. 17–34.
9. Croux, C.; Haesbroeck, G. Implementing the Bianco and Yohai estimator for logistic regression. *Comput. Stat. Data Anal.* **2003**, *44*, 273–295.
10. Heritier, S.; Ronchetti, E. Robust bounded-influence tests in general parametric models. *J. Am. Stat. Assoc.* **1994**, *89*, 897–904.
11. Cantoni, E.; Ronchetti, E. Robust inference for generalized linear models. *J. Am. Stat. Assoc.* **2001**, *96*, 1022–1030.
12. Bianco, A.M.; Martínez, E. Robust testing in the logistic regression model. *Comput. Stat. Data Anal.* **2009**, *53*, 4095–4105.
13. Ronchetti, E.; Trojani, F. Robust inference with GMM estimators. *J. Econom.* **2001**, *101*, 37–69.
14. Basu, A.; Mandal, N.; Martin, N.; Pardo, L. Robust tests for the equality of two normal means based on the density power divergence. *Metrika* **2015**, *78*, 611–634.
15. Lee, S.; Na, O. Test for parameter change based on the estimator minimizing density-based divergence measures. *Ann. Inst. Stat. Math.* **2005**, *57*, 553–573.
16. Kang, J.; Song, J. Robust parameter change test for Poisson autoregressive models. *Stat. Probab. Lett.* **2015**, *104*, 14–21.
17. Basu, A.; Ghosh, A.; Martin, N.; Pardo, L. Robust Wald-type tests for non-homogeneous observations based on minimum density power divergence estimator. *ArXiv Pre-print* **2017**, arXiv:1707.02333.
18. Ghosh, A.; Basu, A.; Pardo, L. Robust Wald-type tests under random censoring. *ArXiv* **2017**, arXiv:1708.09695.
19. Brégman, L.M. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comput. Math. Math. Phys.* **1967**, *7*, 620–631.
20. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: Berlin, Germany, 2001.
21. Zhang, C.M.; Jiang, Y.; Shang, Z. New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Can. J. Stat.* **2009**, *37*, 119–139.
22. Huber, P. Robust estimation of a location parameter. *Ann. Math. Statist.* **1964**, *35*, 73–101.
23. Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.J.; Stahel, W.A. *Robust Statistics: The Application Based on Influence Function*; John Wiley: New York, NY, USA, 1986.
24. Hampel, F.R. The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.* **1974**, *69*, 383–393.
25. Fan, J.; Peng, H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.* **2004**, *32*, 928–961.
26. van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 1998.
27. Clarke, B.R. Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Ann. Stat.* **1983**, *11*, 1196–1205.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

# Robust Relative Error Estimation

Kei Hirose <sup>1,2,\*</sup> and Hiroki Masuda <sup>3</sup>

<sup>1</sup> Institute of Mathematics for Industry, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

<sup>2</sup> RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

<sup>3</sup> Faculty of Mathematics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan;  
hiroki@math.kyushu-u.ac.jp

\* Correspondence: hirose@imi.kyushu-u.ac.jp or mail@keihirose.com

Received: 11 July 2018; Accepted: 20 August 2018; Published: 24 August 2018

**Abstract:** Relative error estimation has been recently used in regression analysis. A crucial issue of the existing relative error estimation procedures is that they are sensitive to outliers. To address this issue, we employ the  $\gamma$ -likelihood function, which is constructed through  $\gamma$ -cross entropy with keeping the original statistical model in use. The estimating equation has a redescending property, a desirable property in robust statistics, for a broad class of noise distributions. To find a minimizer of the negative  $\gamma$ -likelihood function, a majorize-minimization (MM) algorithm is constructed. The proposed algorithm is guaranteed to decrease the negative  $\gamma$ -likelihood function at each iteration. We also derive asymptotic normality of the corresponding estimator together with a simple consistent estimator of the asymptotic covariance matrix, so that we can readily construct approximate confidence sets. Monte Carlo simulation is conducted to investigate the effectiveness of the proposed procedure. Real data analysis illustrates the usefulness of our proposed procedure.

**Keywords:**  $\gamma$ -divergence; relative error estimation; robust estimation

---

## 1. Introduction

In regression analysis, many analysts use the (penalized) least squares estimation, which aims at minimizing the mean squared prediction error [1]. On the other hand, the relative (percentage) error is often more useful and/or adequate than the mean squared error. For example, in econometrics, the comparison of prediction performance between different stock prices with different units should be made by relative error; we refer to [2,3] among others. Additionally, the prediction error of photovoltaic power production or electricity consumption is evaluated by not only mean squared error but also relative error (see, e.g., [4]). We refer to [5] regarding the usefulness and importance of the relative error.

In relative error estimation, we minimize a loss function based on the relative error. An advantage of using such a loss function is that it is scale free or unit free. Recently, several researchers have proposed various loss functions based on relative error [2,3,6–9]. Some of these procedures have been extended to the nonparametric model [10] and random effect model [11]. The relative error estimation via the  $L_1$  regularization, including the least absolute shrinkage and operator (lasso; [12]), and the group lasso [13], have also been proposed by several authors [14–16], to allow for the analysis of high-dimensional data.

In practice, a response variable  $y (> 0)$  can turn out to be extremely large or close to zero. For example, the electricity consumption of a company may be low during holidays and high on exceptionally hot days. These responses may often be considered to be outliers, to which the relative error estimator is sensitive because the loss function diverges when  $y \rightarrow \infty$  or  $y \rightarrow 0$ . Therefore, a relative error estimation that is robust against outliers must be considered. Recently, Chen et al. [8] discussed the robustness of various relative error estimation procedures by investigating the corresponding distributions, and concluded that the distribution of least product relative error

estimation (LPRE) proposed by [8] has heavier tails than others, implying that the LPRE might be more robust than others in practical applications. However, our numerical experiments show that the LPRE is not as robust as expected, so that the robustification of the LPRE is yet to be investigated from the both theoretical and practical viewpoints.

To achieve a relative error estimation that is robust against outliers, this paper employs the  $\gamma$ -likelihood function for regression analysis by Kawashima and Fujisawa [17], which is constructed by the  $\gamma$ -cross entropy [18]. The estimating equation is shown to have a redescending property, a desirable property in robust statistics literature [19]. To find a minimizer of the negative  $\gamma$ -likelihood function, we construct a majorize-minimization (MM) algorithm. The loss function of our algorithm at each iteration is shown to be convex, although the original negative  $\gamma$ -likelihood function is nonconvex. Our algorithm is guaranteed to decrease the objective function at each iteration. Moreover, we derive the asymptotic normality of the corresponding estimator together with a simple consistent estimator of the asymptotic covariance matrix, which enables us to straightforwardly create approximate confidence sets. Monte Carlo simulation is conducted to investigate the performance of our proposed procedure. An analysis of electricity consumption data is presented to illustrate the usefulness of our procedure. Supplemental material includes our R package `rree` (robust relative error estimation), which implements our algorithm, along with a sample program of the `rree` function.

The reminder of this paper is organized as follows: Section 2 reviews several relative error estimation procedures. In Section 3, we propose a relative error estimation that is robust against outliers via the  $\gamma$ -likelihood function. Section 4 presents theoretical properties: the redescending property of our method and the asymptotic distribution of the estimator, the proof of the latter being deferred to Appendix A. In Section 5, the MM algorithm is constructed to find the minimizer of the negative  $\gamma$ -likelihood function. Section 6 investigates the effectiveness of our proposed procedure via Monte Carlo simulations. Section 7 presents the analysis on electricity consumption data. Finally, concluding remarks are given in Section 8.

## 2. Relative Error Estimation

Suppose that  $x_i = (x_{i1}, \dots, x_{ip})^T$  ( $i = 1, \dots, n$ ) are predictors and  $y = (y_1, \dots, y_n)^T$  is a vector of positive responses. Consider the multiplicative regression model

$$y_i = \exp(x_i^T \beta) \varepsilon_i = \exp\left(\sum_{j=1}^p x_{ij} \beta_j\right) \varepsilon_i, \quad (i = 1, \dots, n), \quad (1)$$

where  $\beta = (\beta_1, \dots, \beta_p)^T$  is a  $p$ -dimensional coefficient vector, and  $\varepsilon_i$  are positive random variables. Predictors  $x_i \in \mathbb{R}^p$  may be random and serially dependent, while we often set  $x_{i1} = 1$ , that is, incorporate the intercept in the exponent. The parameter space  $\mathcal{B} \subset \mathbb{R}^p$  of  $\beta$  is a bounded convex domain such that  $\beta_0 \in \mathcal{B}$ . We implicitly assume that the model is correctly specified, so that there exists a true parameter  $\beta_0 = (\beta_{1,0}, \dots, \beta_{p,0}) \in \mathcal{B}$ . We want to estimate  $\beta_0$  from a sample  $\{(x_i, y_i), i = 1, \dots, n\}$ .

We first remark that the condition  $x_{i1} = 1$  ensures that the model (1) is scale-free regarding variables  $\varepsilon_i$ , which is an essentially different nature from the linear regression model  $y_i = x_i^T \beta + \varepsilon_i$ . Specifically, multiplying a positive constant  $\sigma$  to  $\varepsilon_i$  results in the translation of the intercept in the exponent:

$$y_i = \exp(x_i^T \beta) \sigma \varepsilon_i = \exp(\log \sigma + x_i^T \beta) \varepsilon_i$$

so that the change from  $\varepsilon_i$  to  $\sigma \varepsilon_i$  is equivalent to that from  $\beta_1$  to  $\beta_1 + \log \sigma$ . See Remark 1 on the distribution of  $\varepsilon_1$ .

To provide a simple expression of the loss functions based on the relative error, we write

$$t_i = t_i(\beta) = \exp(x_i^T \beta), \quad (i = 1, \dots, n).$$

Chen et al. [6,8] pointed out that the loss criterion for relative error may depend on  $|y_i - t_i|/y_i$  and / or  $|(y_i - t_i)/t_i|$ . These authors also proposed general relative error (GRE) criteria, defined as

$$G(\beta) = \sum_{i=1}^n g\left(\left|\frac{y_i - t_i}{y_i}\right|, \left|\frac{y_i - t_i}{t_i}\right|\right), \quad (2)$$

where  $g : [0, \infty) \times [0, \infty) \rightarrow [0, \infty)$ . Most of the loss functions based on the relative error are included in the GRE. Park and Stefanski [2] considered a loss function  $g(a, b) = a^2$ . It may highly depend on a small  $y_i$  because it includes  $1/y_i^2$  terms, and then the estimator can be numerically unstable. Consistency and asymptotic normality may not be established under general regularity conditions [8]. The loss functions based on  $g(a, b) = \max\{a, b\}$  [3] and  $g(a, b) = a + b$  (least absolute relative error estimation, [6]) can have desirable asymptotic properties [3,6]. However, the minimization of the loss function can be challenging, in particular for high-dimensional data, when the function is nonsmooth or nonconvex.

In practice, the following two criteria would be useful:

**Least product relative error estimation (LPRE)** Chen et al. [8] proposed the LPRE given by  $g(a, b) = ab$ . The LPRE tries to minimize the product  $|1 - t_i/y_i| \times |1 - y_i/t_i|$ , not necessarily both terms at once.

**Least squared-sum relative error estimation (LSRE)** Chen et al. [8] considered the LSRE given by  $g(a, b) = a^2 + b^2$ . The LSRE aims to minimize both  $|1 - t_i/y_i|$  and  $|1 - y_i/t_i|$  through sum of squares  $(1 - t_i/y_i)^2 + (1 - y_i/t_i)^2$ .

The loss functions of LPRE and LSRE are smooth and convex, and also possess desirable asymptotic properties [8]. The above-described GRE criteria and their properties are summarized in Table 1. Particularly, the “convexity” in the case of  $g(a, b) = a + b$  holds when  $\varepsilon_i > 0$ ,  $\varepsilon_i \neq 1$ , and  $\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$  is positive definite, since the Hessian matrix of the corresponding  $G(\beta)$  is  $\sum_{i=1}^n |\varepsilon_i - \varepsilon_i^{-1}| \mathbf{x}_i \mathbf{x}_i^T$  a.s.

**Table 1.** Several examples of general relative error (GRE) criteria and their properties. “Likelihood” in the second column means the existence of a likelihood function that corresponds to the loss function. The properties of “Convexity” and “Smoothness” in the last two columns respectively indicate those with respect to  $\beta$  of the corresponding loss function.

$g(a, b)$	Likelihood	Convexity	Smoothness
$a^2$			✓
$a + b$	✓	✓	
$\max\{a, b\}$	✓		
$ab$	✓	✓	✓
$a^2 + b^2$	✓	✓	✓

Although not essential, we assume that the variables  $\varepsilon_i$  in Equation (1) are i.i.d. with common density function  $h$ . As in Chen et al. [8], we consider the following class of  $h$  associated with  $g$ :

$$h(\varepsilon) := \frac{C(g)}{\varepsilon} \exp\{-\rho(\varepsilon)\} I_+(\varepsilon), \quad (3)$$

where

$$\rho(\varepsilon) = \rho(\varepsilon; g) := g\left(\left|1 - \frac{1}{\varepsilon}\right|, |1 - \varepsilon|\right),$$

and  $C(g)$  is a normalizing constant ( $\int h(\varepsilon) d\varepsilon = 1$ ) and  $I_+$  denotes the indicator function of set  $(0, \infty)$ . Furthermore, we assume the symmetry property  $g(a, b) = g(b, a)$ ,  $a, b \geq 0$ , from which it follows that  $\varepsilon_1 \sim \varepsilon_1^{-1}$ . The latter property is necessary for a score function to be associated with the gradient

of a GRE loss function, hence being a martingale with respect to a suitable filtration, which often entails estimation efficiency. Indeed, the asymmetry of  $g(a, b)$  (i.e.,  $g(a, b) \neq g(b, a)$ ) may produce a substantial bias in the estimation [3]. The entire set of our regularity conditions will be shown in Section 4.3. The conditions therein concerning  $g$  are easily verified for both LPRE and LSRE.

In this paper, we implicitly suppose that  $i = 1, \dots, n$  denote “time” indices. As usual, in order to deal with cases of non-random and random predictors in a unified manner, we employ the partial-likelihood framework. Specifically, in the expression of the joint density (with the obvious notation for the densities)

$$f(\mathbf{x}_1, \dots, \mathbf{x}_n, y_1, \dots, y_n) \\ = \left\{ f(\mathbf{x}_1) \prod_{i=2}^n f(\mathbf{x}_i | \mathbf{x}_1, \dots, \mathbf{x}_{i-1}, y_1, \dots, y_{i-1}) \right\} \left\{ f(y_1 | \mathbf{x}_1) \prod_{i=2}^n f(y_i | \mathbf{x}_1, \dots, \mathbf{x}_i, y_1, \dots, y_{i-1}) \right\},$$

we ignore the first product  $\{\dots\}$  and only look at the second one  $\{\dots\}$ , which is defined as the partial likelihood. We further assume that the  $i$ th-stage noise  $\varepsilon_i$  is independent of  $(\mathbf{x}_1, \dots, \mathbf{x}_i, y_1, \dots, y_{i-1})$ , so that, in view of Equation (1), we have

$$f(y_i | \mathbf{x}_1, \dots, \mathbf{x}_i, y_1, \dots, y_{i-1}) = f(y_i | \mathbf{x}_i), \quad i = 1, \dots, n.$$

The density function of response  $y$  given  $\mathbf{x}_i$  is

$$f(y | \mathbf{x}_i; \boldsymbol{\beta}) = \exp(-\mathbf{x}_i^T \boldsymbol{\beta}) h\left(y \exp(-\mathbf{x}_i^T \boldsymbol{\beta})\right) \quad (4)$$

$$= \frac{1}{t_i} h\left(\frac{y}{t_i}\right) \quad (5)$$

From Equation (3), we see that the maximum likelihood estimator (MLE) based on the error distribution in Equation (5) is obtained by the minimization of Equation (2). For example, the density functions of LPRE and LSRE are

$$\text{LPRE : } f(y | \mathbf{x}_i) = \frac{1}{2K_0(2)} y^{-1} \exp\left(-\frac{y}{t_i} - \frac{t_i}{y}\right), \quad (y > 0), \quad (6)$$

$$\text{LSRE : } f(y | \mathbf{x}_i) = C_{LSRE} y^{-1} \exp\left\{-\left(1 - \frac{t_i}{y}\right)^2 - \left(1 - \frac{y}{t_i}\right)^2\right\}, \quad (y > 0),$$

where  $K_\nu(z)$  denotes the modified Bessel function of third kind with index  $\nu \in \mathbb{R}$ :

$$K_\nu(z) = \frac{z^\nu}{2^{\nu+1}} \int_0^\infty t^{-\nu-1} \exp\left(-t - \frac{z^2}{4t}\right) dt$$

and  $C_{LSRE}$  is a constant term. Constant terms are numerically computed as  $K_0(2) \approx 0.1139$  and  $C_{LSRE} \approx 0.911411$ . Density (6) is a special case of the generalized inverse Gaussian distribution (see, e.g., [20]).

**Remark 1.** We assume that the noise density  $h$  is fully specified in the sense that, given  $g$ , the density  $h$  does not involve any unknown quantity. However, this is never essential. For example, for the LPRE defined by Equation (6), we could naturally incorporate one more parameter  $\sigma > 0$  into  $h$ , the resulting form of  $h(\varepsilon)$  being

$$\varepsilon \mapsto \frac{1}{2K_0(\sigma)} \varepsilon^{-1} \exp\left\{-\frac{\sigma}{2} \left(\varepsilon + \frac{1}{\varepsilon}\right)\right\} I_+(\varepsilon).$$

Then, we can verify that the distributional equivalence  $\varepsilon_1 \sim \varepsilon_1^{-1}$  holds whatever the value of  $\sigma$  is. Particularly, the estimation of parameter  $\sigma$  does make statistical sense and, indeed, it is possible to deduce the

asymptotic normality of the joint maximum-(partial-) likelihood estimator of  $(\beta, \sigma)$ . In this paper, we do not pay attention to such a possible additional parameter, but instead regard it (whenever it exists) as a nuisance parameter, as in the noise variance in the least-squares estimation of a linear regression model.

### 3. Robust Estimation via $\gamma$ -Likelihood

In practice, outliers can often be observed. For example, the electricity consumption data can have the outliers on extremely hot days. The estimation methods via GRE criteria, including LPRE and LSRE, are not robust against outliers, because the corresponding density functions are not generally heavy-tailed. Therefore, a relative error estimation method that is robust against the outliers is needed. To achieve this, we consider minimizing the negative  $\gamma$ -(partial-)likelihood function based on the  $\gamma$ -cross entropy [17].

We now define the negative  $\gamma$ -(partial-)likelihood function by

$$\ell_{\gamma,n}(\beta) = -\frac{1}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n f(y_i | \mathbf{x}_i; \beta)^\gamma \right\} + \frac{1}{1+\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n \int_0^\infty f(y | \mathbf{x}_i; \beta)^{1+\gamma} dy \right\}, \quad (7)$$

where  $\gamma > 0$  is a parameter that controls the degrees of robustness;  $\gamma \rightarrow 0$  corresponds to the negative log-likelihood function, and robustness is enhanced as  $\gamma$  increases. On the other hand, a too large  $\gamma$  can decrease the efficiency of the estimator [18]. In practice, the value of  $\gamma$  may be selected by a cross-validation based on  $\gamma$ -cross entropy (see, e.g., [18,21]). We refer to Kawashima and Fujisawa [22] for more recent observations on comparison of the  $\gamma$ -divergences between Fujisawa and Eguchi [18] and Kawashima and Fujisawa [17].

There are several likelihood functions which yield robust estimation. Examples include the  $L_q$ -likelihood [23], and the likelihood based on the density power divergence [24], referred to as  $\beta$ -likelihood. It is shown that the  $\gamma$ -likelihood, the  $L_q$ -likelihood, and the  $\beta$ -likelihood are closely related. The negative  $\beta$ -likelihood function  $\ell_{\alpha,n}(\beta)$  and the negative  $L_q$ -likelihood function  $\ell_{q,n}(\beta)$  are, respectively, expressed as

$$\ell_{\alpha,n}(\beta) = -\frac{1}{\alpha} \frac{1}{n} \sum_{i=1}^n f(y_i | \mathbf{x}_i; \beta)^\alpha + \frac{1}{1+\alpha} \frac{1}{n} \sum_{i=1}^n \int_0^\infty f(y | \mathbf{x}_i; \beta)^{1+\alpha} dy, \quad (8)$$

$$\ell_{q,n}(\beta) = -\sum_{i=1}^n \frac{f(y_i | \mathbf{x}_i; \beta)^{1-q} - 1}{1-q}. \quad (9)$$

The difference between  $\gamma$ -likelihood and  $\beta$ -likelihood is just the existence of the logarithm on  $\ell_{\gamma,n}(\beta)$ . Furthermore, substituting  $q = 1 - \alpha$  into Equation (9) gives us

$$\ell_{q,n}(\beta) = -\frac{1}{\alpha} \sum_{i=1}^n f(y_i | \mathbf{x}_i; \beta)^\alpha + \text{const.}$$

Therefore, the minimization of the negative  $L_q$ -likelihood function is equivalent to minimization of the negative  $\beta$ -likelihood function without second term in the right side of Equation (8). Note that the  $\gamma$ -likelihood has the redescending property, a desirable property in robust statistics literature, as shown in Section 4.2. Moreover, it is known that the  $\gamma$ -likelihood is the essentially unique divergence that is robust against heavy contamination (see [18] for details). On the other hand, we have not shown whether the  $L_q$ -likelihood and/or the  $\beta$ -likelihood have the redescending property or not.

The integration  $\int f(y | \mathbf{x}_i; \beta)^{1+\gamma} dy$  in the second term on the right-hand side of Equation (7) is

$$\int_0^\infty f(y | \mathbf{x}_i; \beta)^{1+\gamma} dy = \frac{1}{t_i^{1+\gamma}} \int_0^\infty \left\{ h\left(\frac{y}{t_i}\right)\right\}^{1+\gamma} dy =: t_i^{-\gamma} C(\gamma, h),$$

where

$$C(\gamma, h) := \int_0^\infty h(v)^{1+\gamma} dv \quad (10)$$

is a constant term, which is assumed to be finite. Then, Equation (7) is expressed as

$$\ell_{\gamma,n}(\beta) = \underbrace{-\frac{1}{\gamma} \log \left\{ \sum_{i=1}^n f(y_i | x_i; \beta)^\gamma \right\}}_{=: \ell_1(\beta)} + \underbrace{\frac{1}{1+\gamma} \log \left\{ \sum_{i=1}^n t_i^{-\gamma} \right\}}_{=: \ell_2(\beta)} + C_0(\gamma, h), \quad (11)$$

where  $C_0(\gamma, h)$  is a constant term free from  $\beta$ . We define the maximum  $\gamma$ -likelihood estimator to be any element such that

$$\hat{\beta}_\gamma \in \operatorname{argmin} \ell_{\gamma,n}.$$

#### 4. Theoretical Properties

##### 4.1. Technical Assumptions

Let  $\xrightarrow{p}$  denote the convergence in probability.

**Assumption 1 (Stability of the predictor).** There exists a probability measure  $\pi(dx)$  on the state space  $\mathcal{X}$  of the predictors and positive constants  $\delta, \delta' > 0$  such that

$$\frac{1}{n} \sum_{i=1}^n |x_i|^3 \exp(\delta' |x_i|^{1+\delta}) = O_p(1),$$

and that

$$\frac{1}{n} \sum_{i=1}^n \eta(x_i) \xrightarrow{p} \int_{\mathcal{X}} \eta(x) \pi(dx), \quad n \rightarrow \infty,$$

where the limit is finite for any measurable  $\eta$  satisfying that

$$\sup_{x \in \mathbb{R}^p} \frac{|\eta(x)|}{(1 + |x|^3) \exp(\delta' |x|^{1+\delta})} < \infty.$$

**Assumption 2 (Noise structure).** The a.s. positive i.i.d. random variables  $\varepsilon_1, \varepsilon_2, \dots$  have a common positive density  $h$  of the form (3):

$$h(\varepsilon) = \frac{C(g)}{\varepsilon} \exp\{-\rho(\varepsilon)\} I_+(\varepsilon),$$

for which the following conditions hold.

1. Function  $g : [0, \infty) \times [0, \infty) \rightarrow [0, \infty)$  is three times continuously differentiable on  $(0, \infty)$  and satisfies that

$$g(a, b) = g(b, a), \quad a, b \geq 0.$$

2. There exist constants  $\kappa_0, \kappa_\infty > 0$ , and  $c > 1$  such that

$$\frac{1}{c} (\varepsilon^{-\kappa_0} \vee \varepsilon^{\kappa_\infty}) \leq \rho(\varepsilon) \leq c (\varepsilon^{-\kappa_0} \vee \varepsilon^{\kappa_\infty})$$

for every  $\varepsilon > 0$ .

3. There exist constants  $c_0, c_\infty \geq 0$  such that

$$\sup_{\varepsilon > 0} (\varepsilon^{-c_0} \vee \varepsilon^{c_\infty})^{-1} \max_{k=1,2,3} |\partial_\varepsilon^k \rho(\varepsilon)| < \infty.$$

Here and in the sequel, for a variable  $a$ , we denote by  $\partial_a^k$  the  $k$ th-order partial differentiation with respect to  $a$ .

Assumption 1 is necessary to identify the large-sample stochastic limits of the several key quantities in the proofs: without them, we will not be able to deduce an explicit asymptotic normality result. Assumption 2 holds for many cases, including the LPRE and the LSRE (i.e.,  $g(a, b) = ab$  and  $a^2 + b^2$ ), while excluding  $g(a, b) = a^2$  and  $g(a, b) = b^2$ . The smoothness condition on  $h$  on  $(0, \infty)$  is not essential and could be weakened in light of the  $M$ -estimation theory ([25], Chapter 5). Under these assumptions, we can deduce the following statements.

- $h$  is three times continuously differentiable on  $(0, \infty)$ , and for each  $\alpha > 0$ ,

$$\int_0^\infty h^\alpha(\varepsilon)d\varepsilon < \infty \quad \text{and} \quad \max_{k=0,1,2,3} \sup_{\varepsilon>0} \left| \partial_\varepsilon^k \{h(\varepsilon)^\alpha\} \right| < \infty.$$

- For each  $\gamma > 0$  and  $\alpha > 0$  (recall that the value of  $\gamma > 0$  is given),

$$\lim_{\varepsilon \downarrow 0} h(\varepsilon)^\gamma |u_h(\varepsilon)|^\alpha = \lim_{\varepsilon \uparrow \infty} h(\varepsilon)^\gamma |u_h(\varepsilon)|^\alpha = 0, \quad (12)$$

where

$$u_h(z) := 1 + z \partial_z \log h(z) = 1 + z \frac{h'(z)}{h(z)}.$$

The verifications are straightforward hence omitted.

Finally, we impose the following assumption:

**Assumption 3 (Identifiability).** We have  $\beta = \beta_0$  if

$$\rho(e^{-x^T \beta} y) = \rho(e^{-x^T \beta_0} y) \quad \pi(dx) \otimes \lambda_+(dy)\text{-a.e. } (x, y),$$

where  $\lambda_+$  denotes the Lebesgue measure on  $(0, \infty)$ .

#### 4.2. Redescending Property

The estimating function based on the negative  $\gamma$ -likelihood function is given by

$$\sum_{i=1}^n \psi(y_i|x_i; \beta) = \mathbf{0}.$$

In our model, we consider not only too large  $y_i$ s but also too small  $y_i$ s as outliers: the estimating equation is said to have the redescending property if

$$\lim_{y \rightarrow \infty} \psi(y|x; \beta_0) = \lim_{y \rightarrow +0} \psi(y|x; \beta_0) = \mathbf{0}$$

for each  $x$ . The redescending property is known as a desirable property in robust statistics literature [19]. Here, we show the proposed procedure has the redescending property.

The estimating equation based on the negative  $\gamma$ -likelihood function is

$$-\frac{\sum_{i=1}^n f(y_i|x_i; \beta)^\gamma s(y_i|x_i; \beta)}{\sum_{j=1}^n f(y_j|x_j; \beta)^\gamma} + \frac{\partial}{\partial \beta} \ell_2(\beta) = \mathbf{0},$$

where

$$s(y|x; \beta) = \frac{\partial \log f(y|x; \beta)}{\partial \beta}.$$

We have expression

$$\psi(y|x; \beta) = f(y|x; \beta)^\gamma \left\{ s(y|x; \beta) - \frac{\partial}{\partial \beta} \ell_2(\beta) \right\}.$$

Note that  $\frac{\partial}{\partial \beta} \ell_2(\beta)$  is free from  $y$ . For each  $(x, \beta)$ , direct computations give the estimate

$$|\psi(y|x; \beta)| \leq C(x; \beta) h\left(\exp(-x^T \beta)y\right)^\gamma \left| u_h\left(\exp(-x^T \beta)y\right) \right| \quad (13)$$

for some constant  $C(x; \beta)$  free from  $y$ . Hence, Equation (12) combined with the inequality (13) leads to the redescending property.

#### 4.3. Asymptotic Distribution

Recall Equation (10) for the definition of  $C(\gamma, h)$  and let

$$\begin{aligned} C_1(\gamma, h) &:= \int_0^\infty \varepsilon h(\varepsilon)^\gamma h'(\varepsilon) d\varepsilon, \\ C_2(\gamma, h) &:= \int_0^\infty u_h(\varepsilon)^2 h(\varepsilon)^{2\gamma+1} d\varepsilon, \\ \Pi_k(\gamma) &:= \int x^{\otimes k} \exp(-\gamma x^T \beta_0) \pi(dx), \quad k = 0, 1, 2, \end{aligned}$$

where  $x^{\otimes 0} := 1 \in \mathbb{R}$ ,  $x^{\otimes 1} := x \in \mathbb{R}^p$ , and  $x^{\otimes 2} := xx^T \in \mathbb{R}^p \otimes \mathbb{R}^p$ ; Assumptions 1 and 2 ensure that all these quantities are finite for each  $\gamma > 0$ . Moreover,

$$\begin{aligned} H'_\gamma(\beta_0) &:= \iint f(y|x; \beta_0)^{\gamma+1} dy \pi(dx) = C(\gamma, h) \Pi_0(\gamma), \\ H''_\gamma(\beta_0) &:= \iint f(y|x; \beta_0)^{\gamma+1} s(y|x; \beta_0) dy \pi(dx) = -\{C(\gamma, h) + C_1(\gamma, h)\} \Pi_1(\gamma), \\ \Delta_\gamma(\beta_0) &:= C(\gamma, h)^2 C_2(\gamma, h) \Pi_0(\gamma)^2 \Pi_2(2\gamma) \\ &\quad + \{C(\gamma, h) + C_1(\gamma, h)\}^2 C(2\gamma, h) \Pi_0(2\gamma) \Pi_1(\gamma)^{\otimes 2} \\ &\quad - 2C(\gamma, h) \{C(\gamma, h) + C_1(\gamma, h)\} \{C(2\gamma, h) + C_1(2\gamma, h)\} \Pi_0(\gamma) \Pi_1(2\gamma) \Pi_1(\gamma)^T, \quad (14) \\ J_\gamma(\beta_0) &:= C(\gamma, h) C_2(\gamma/2, h) \Pi_0(\gamma) \Pi_2(\gamma) - \{C(\gamma, h) + C_1(\gamma, h)\}^2 \Pi_1(\gamma)^{\otimes 2}. \quad (15) \end{aligned}$$

We are assuming that density  $h$  and tuning parameter  $\gamma$  are given a priori, hence we can (numerically) compute constants  $C(\gamma, h)$ ,  $C_1(\gamma, h)$ , and  $C_2(\gamma, h)$ . In the following, we often omit  $"(\beta_0)"$  from the notation.

Let  $\xrightarrow{\mathcal{L}}$  denote the convergence in distribution.

**Theorem 1.** Under Assumptions 1–3, we have

$$\sqrt{n} (\hat{\beta}_\gamma - \beta_0) \xrightarrow{\mathcal{L}} N_p (\mathbf{0}, J_\gamma^{-1} \Delta_\gamma J_\gamma^{-1}). \quad (16)$$

The asymptotic covariance matrix can be consistently estimated through expressions (14) and (15) with quantities  $\Pi_k(\gamma)$  therein replaced by the empirical estimates:

$$\hat{\Pi}_{k,n}(\gamma) := \frac{1}{n} \sum_{i=1}^n x_i^{\otimes k} \exp(-\gamma x_i^T \hat{\beta}_\gamma) \xrightarrow{p} \Pi_k(\gamma), \quad k = 0, 1, 2. \quad (17)$$

The proof of Theorem 1 will be given in Appendix A. Note that, for  $\gamma \rightarrow 0$ , we have  $C(\gamma, h) \rightarrow 1$ ,  $C_1(\gamma, h) \rightarrow -1$ , and  $C_2(\gamma, h) \rightarrow \int_0^\infty u_h(\varepsilon)^2 h(\varepsilon) d\varepsilon$ , which in particular entails  $H'_\gamma \rightarrow 1$  and  $H''_\gamma \rightarrow \mathbf{0}$ . Then, both  $\Delta_\gamma$  and  $J_\gamma$  tend to the Fisher information matrix

$$\mathcal{I}_0 := \iint s(y|x; \beta_0)^{\otimes 2} f(y|x; \beta_0) \pi(dx) dy = \int_0^\infty u_h(\varepsilon)^2 h(\varepsilon) d\varepsilon \int x^{\otimes 2} \pi(dx)$$

as  $\gamma \rightarrow 0$ , so that the asymptotic distribution  $N_p(\mathbf{0}, J_\gamma^{-1} \Delta_\gamma J_\gamma^{-1})$  becomes  $N_p(\mathbf{0}, \mathcal{I}_0^{-1})$ , the usual one of the MLE.

We also note that, without details, we could deduce a density-power divergence (also known as the  $\beta$ -divergence [26]) counterpart to Theorem 1 similarly but with slightly lesser computation cost; in that case, we consider the objective function  $\ell_{\alpha,n}(\beta)$  defined by Equation (8) instead of the  $\gamma$ -(partial-)likelihood (7). See Basu et al. [24] and Jone et al. [21] for details of the density-power divergence.

## 5. Algorithm

Even if the GRE criterion in Equation (2) is a convex function, the negative  $\gamma$ -likelihood function is nonconvex. Therefore, it is difficult to find a global minimum. Here, we derive the MM (majorize-minimization) algorithm to obtain a local minimum. The MM algorithm monotonically decreases the objective function at each iteration. We refer to Hunter and Lange [27] for a concise account of the MM algorithm.

Let  $\beta^{(t)}$  be the value of the parameter at the  $t$ th iteration. The negative  $\gamma$ -likelihood function in Equation (11) consists of two nonconvex functions,  $\ell_1(\beta)$  and  $\ell_2(\beta)$ . The majorization functions of  $\ell_j(\beta)$ , say  $\tilde{\ell}_j(\beta|\beta^{(t)})$  ( $j = 1, 2$ ), are constructed so that the optimization of  $\min_{\beta} \tilde{\ell}_j(\beta|\beta^{(t)})$  is much easier than that of  $\min_{\beta} \ell_j(\beta)$ . The majorization functions must satisfy the following inequalities:

$$\tilde{\ell}_j(\beta|\beta^{(t)}) \geq \ell_j(\beta), \quad (18)$$

$$\tilde{\ell}_j(\beta^{(t)}|\beta^{(t)}) = \ell_j(\beta^{(t)}). \quad (19)$$

Here, we construct majorization functions  $\tilde{\ell}_j(\beta|\beta^{(t)})$  for  $j = 1, 2$ .

### 5.1. Majorization Function for $\ell_1(\beta)$

Let

$$w_i^{(t)} = \frac{f(y_i|x_i; \beta^{(t)})^\gamma}{\sum_{j=1}^n f(y_j|x_j; \beta^{(t)})^\gamma}, \quad (20)$$

$$r_i^{(t)} = \sum_{j=1}^n f(y_j|x_j; \beta^{(t)})^\gamma \frac{f(y_i|x_i; \beta^{(t)})^\gamma}{f(y_i|x_i; \beta^{(t)})^\gamma}. \quad (21)$$

Obviously,  $\sum_{i=1}^n w_i^{(t)} = 1$  and  $w_i^{(t)} r_i^{(t)} = f(y_i|x_i; \beta^{(t)})^\gamma$ . Applying Jensen's inequality to  $y = -\log x$ , we obtain inequality

$$-\log \left( \sum_{i=1}^n w_i^{(t)} r_i^{(t)} \right) \leq -\sum_{i=1}^n w_i^{(t)} \log r_i^{(t)}. \quad (22)$$

Substituting Equation (20) and Equation (21) into Equation (22) gives

$$\ell_1(\beta) \leq -\sum_{i=1}^n w_i^{(t)} \log f(y_i|x_i; \beta) + C,$$

where  $C = \frac{1}{\gamma} \sum_i w_i^{(t)} \log w_i^{(t)}$ . Denoting

$$\tilde{\ell}_1(\beta|\beta^{(t)}) = - \sum_{i=1}^n w_i^{(t)} \log f(y_i|x_i;\beta) + C, \quad (23)$$

we observe that Equation (23) satisfies Equation (18) and Equation (19). It is shown that  $\tilde{\ell}_1(\beta|\beta^{(t)})$  is a convex function if the original relative error loss function is convex. Particularly, the majorization functions  $\tilde{\ell}_1(\beta|\beta^{(t)})$  based on LPRE and LSRE are both convex.

### 5.2. Majorization Function for $\ell_2(\beta)$

Let  $\theta_i = -\gamma x_i^T \beta$ . We view  $\ell_2(\beta)$  as a function of  $\theta = (\theta_1, \dots, \theta_n)^T$ . Let

$$s(\theta) := \log \left( \sum_{i=1}^n t_i^{-\gamma} \right) = \log \left( \sum_{i=1}^n \exp(\theta_i) \right). \quad (24)$$

By taking the derivative of  $s(\theta)$  with respect to  $\theta$ , we have

$$\frac{\partial s(\theta)}{\partial \theta_i} = \pi_i, \quad \frac{\partial^2 s(\theta)}{\partial \theta_j \partial \theta_i} = \pi_i \delta_{ij} - \pi_i \pi_j,$$

where  $\pi_i = \exp(\theta_i) / \{\sum_{k=1}^n \exp(\theta_k)\}$ . Note that  $\sum_{i=1}^n \pi_i = 1$  for any  $\theta$ .

The Taylor expansion of  $s(\theta)$  at  $\theta = \theta^{(t)}$  is expressed as

$$s(\theta) = s(\theta^{(t)}) + \pi^{(t)T}(\theta - \theta^{(t)}) + \frac{1}{2}(\theta - \theta^{(t)})^T \frac{\partial^2 s(\theta^*)}{\partial \theta \partial \theta^T} (\theta - \theta^{(t)}), \quad (25)$$

where  $\pi^{(t)} = (\pi_1^{(t)}, \dots, \pi_n^{(t)})^T$  and  $\theta^*$  is an  $n$ -dimensional vector located between  $\theta$  and  $\theta^{(t)}$ . We define an  $n \times n$  matrix  $B$  as follows:

$$B := \frac{1}{2} \left( I - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right).$$

It follows from [28] that, in the matrix sense,

$$\frac{\partial^2 s(\theta)}{\partial \theta \partial \theta^T} \leq B \quad (26)$$

for any  $\theta$ . Combining Equation (25) and Equation (26), we have

$$s(\theta) \leq s(\theta^{(t)}) + \pi^{(t)T}(\theta - \theta^{(t)}) + \frac{1}{2}(\theta - \theta^{(t)})^T B (\theta - \theta^{(t)}). \quad (27)$$

Substituting Equation (24) into Equation (27) gives

$$\begin{aligned} \log \left\{ \sum_{i=1}^n \exp(-\gamma x_i^T \beta) \right\} &\leq \log \left\{ \sum_{i=1}^n \exp(-\gamma x_i^T \beta^{(t)}) \right\} - \gamma \pi^{(t)T} X (\beta - \beta^{(t)}) \\ &\quad + \frac{\gamma^2}{2} (\beta - \beta^{(t)})^T X^T B X (\beta - \beta^{(t)}), \end{aligned}$$

where  $X = (x_1, \dots, x_n)^T$ . The majorization function of  $\ell_2(\beta)$  is then constructed by

$$\tilde{\ell}_2(\beta|\beta^{(t)}) = \frac{\gamma^2}{2(1+\gamma)} \beta^T X^T B X \beta - \frac{\gamma}{1+\gamma} \beta^T (X^T \pi^{(t)} + \gamma X^T B X \beta^{(t)}) + C, \quad (28)$$

where  $C$  is a constant term free from  $\beta$ . We observe that  $\tilde{\ell}_2(\beta|\beta^{(t)})$  in Equation (28) satisfies Equation (18) and Equation (19). It is shown that  $\tilde{\ell}_2(\beta|\beta^{(t)})$  is a convex function because  $X^T BX$  is positive semi-definite.

### 5.3. MM Algorithm for Robust Relative Error Estimation

In Sections 5.1 and 5.2, we have constructed the majorization functions for both  $\ell_1(\beta)$  and  $\ell_2(\beta)$ . The MM algorithm based on these majorization functions is detailed in Algorithm 1. The majorization function  $\tilde{\ell}_1(\beta|\beta^{(t)}) + \tilde{\ell}_2(\beta|\beta^{(t)})$  is convex if the original relative error loss function is convex. Particularly, the majorization functions of LPRE and LSRE are both convex.

---

**Algorithm 1** Algorithm of robust relative error estimation.

---

- 1:  $t \leftarrow 0$
- 2: Set an initial value of parameter vector  $\beta^{(0)}$ .
- 3: **while**  $\beta^{(t)}$  is converged **do**
- 4:   Update the weights by Equation (20)
- 5:   Update  $\beta$  by

$$\beta^{(t+1)} \leftarrow \arg \min_{\beta} \{ \tilde{\ell}_1(\beta|\beta^{(t)}) + \tilde{\ell}_2(\beta|\beta^{(t)}) \},$$

where  $\tilde{\ell}_1(\beta|\beta^{(t)})$  and  $\tilde{\ell}_2(\beta|\beta^{(t)})$  are given by Equation (23) and Equation (28), respectively.

- 6:    $t \leftarrow t + 1$
  - 7: **end while**
- 

**Remark 2.** Instead of the MM algorithm, one can directly use the quasi-Newton method, such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm to minimize the negative  $\gamma$ -likelihood function. In our experience, the BFGS algorithm is faster than the MM algorithm but is more sensitive to an initial value than the MM algorithm. The strengths of BFGS and MM algorithms would be shared by using the following hybrid algorithm:

1. We first conduct the MM algorithm with a small number of iterations.
2. Then, the BFGS algorithm is conducted. We use the estimate obtained by the MM algorithm as an initial value of the BFGS algorithm.

The stabilities of the MM algorithm is investigated through the real data analysis in Section 7.

**Remark 3.** To deal with high-dimensional data, we often use the  $L_1$  regularization, such as the lasso [12], elastic net [29], and Smoothly Clipped Absolute Deviation (SCAD) [30]. In robust relative error estimation, the loss function based on the lasso is expressed as

$$\ell_{\gamma,n}(\beta) + \lambda \sum_{j=1}^p |\beta_j|, \quad (29)$$

where  $\lambda > 0$  is a regularization parameter. However, the loss function in Equation (29) is non-convex and non-differentiable. Instead of directly minimizing the non-convex loss function in Equation (29), we may use the MM algorithm; the following convex loss function is minimized at each iteration:

$$\tilde{\ell}_1(\beta|\beta^{(t)}) + \tilde{\ell}_2(\beta|\beta^{(t)}) + \lambda \sum_{j=1}^p |\beta_j|. \quad (30)$$

The minimization of Equation (30) can be realized by the alternating direction method of multipliers algorithm [14] or the coordinate descent algorithm with quadratic approximation of  $\tilde{\ell}_1(\beta|\beta^{(t)}) + \tilde{\ell}_2(\beta|\beta^{(t)})$  [31].

## 6. Monte Carlo Simulation

### 6.1. Setting

We consider the following two simulation models as follows:

$$\begin{aligned}\text{Model 1: } \beta_0 &= (1, 1, 1)^T, \\ \text{Model 2: } \beta_0 &= (\underbrace{0.5, \dots, 0.5}_6, \underbrace{0, \dots, 0}_{45})^T.\end{aligned}$$

The number of observations is set to be  $n = 200$ . For each model, we generate  $T=10,000$  datasets of predictors  $x_i$  ( $i = 1, \dots, n$ ) according to  $N(\mathbf{0}, (1 - \rho)I + \rho\mathbf{1}\mathbf{1}^T)$ . Here, we consider the case of  $\rho = 0.0$  and  $\rho = 0.6$ . Responses  $y_i$  are generated from the mixture distribution

$$(1 - \delta)f(y|x_i; \beta_0) + \delta q(y) \quad (i = 1, \dots, n),$$

where  $f(y|x; \beta_0)$  is a density function corresponding to the LPRE defined as Equation (6),  $q(y)$  is a density function of distribution of outliers, and  $\delta$  ( $0 \leq \delta < 1$ ) is an outlier ratio. The outlier ratio is set to be  $\delta = 0, 0.05, 0.1$ , and  $0.2$  in this simulation. We assume that  $q(y)$  follows a log-normal distribution (pdf:  $q(y) = 1/(\sqrt{2\pi}y\sigma) \exp\{-(\log y - \mu)^2/(2\sigma^2)\}$ ) with  $(\mu, \sigma) = (\pm 5, 1)$ . When  $\mu = 5$ , the outliers take extremely large values. On the other hand, when  $\mu = -5$ , the data values of outliers are nearly zero.

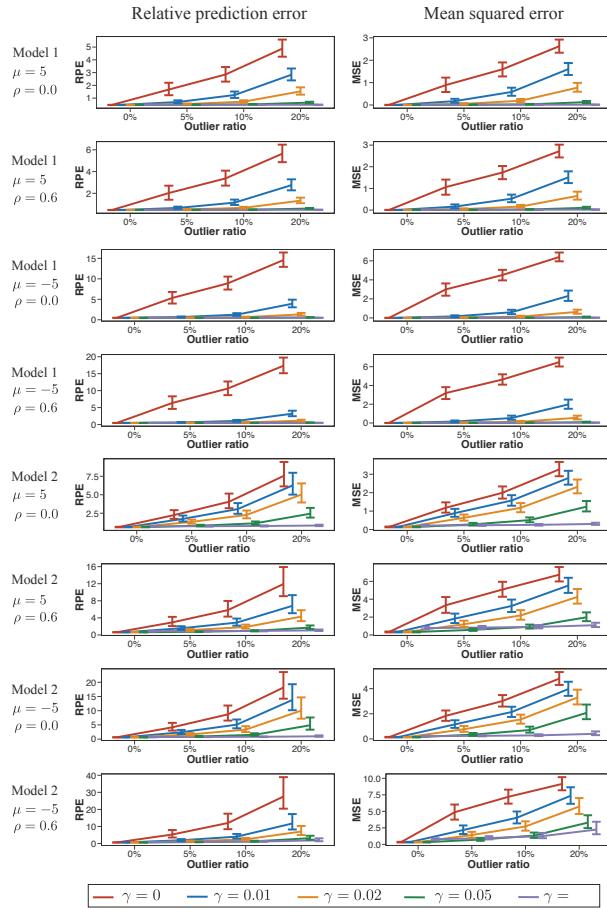
### 6.2. Investigation of Relative Prediction Error and Mean Squared Error of the Estimator

To investigate the performance of our proposed procedure, we use the relative prediction error (RPE) and the mean square error (MSE) for the  $t$ th dataset, defined as

$$RPE(t) = \sum_{i=1}^n \frac{[y_i^{\text{new}}(t) - \exp\{x_i(t)^T \hat{\beta}(t)\}]^2}{y_i^{\text{new}}(t) \exp\{x_i(t)^T \hat{\beta}(t)\}}, \quad (31)$$

$$MSE(t) = \|\hat{\beta}(t) - \beta_0\|^2, \quad (32)$$

respectively, where  $\hat{\beta}(t)$  is an estimator obtained from the dataset  $\{(x_i(t), y_i(t)); i = 1, \dots, n\}$ , and  $y_i^{\text{new}}(t)$  is an observation from  $y_i^{\text{new}}(t)|x_i(t)$ . Here,  $y_i^{\text{new}}(t)|x_i(t)$  follows a distribution of  $f(y|x_i(t); \beta_0)$  and is independent of  $y_i(t)|x_i(t)$ . Figure 1 shows the median and error bar of  $\{RPE(1), \dots, RPE(T)\}$  and  $\{MSE(1), \dots, MSE(T)\}$ . The error bars are delineated by the 25th and 75th percentiles.



**Figure 1.** Median and error bar of relative prediction error (RPE) in Equation (31) and mean squared error (MSE) of  $\beta$  in Equation (32) when parameters of the log-normal distribution (distribution of outliers) are  $(\mu, \sigma) = (\pm 5, 1)$ . The error bars are delineated by 25th and 75th percentiles.

We observe the following tendencies from the results in Figure 1:

- As the outlier ratio increases, the performance becomes worse in all cases. Interestingly, the length of the error bar of RPE increases as the outlier ratio increases.
- The proposed method becomes robust against outliers as the value of  $\gamma$  increases. We observe that a too large  $\gamma$ , such as  $\gamma = 10$ , leads to extremely poor RPE and MSE because most observations are regarded as outliers. Therefore, the not too large  $\gamma$ , such as the  $\gamma = 0.5$  used here, generally results in better estimation accuracy than the MLE.
- The cases of  $\rho = 0.6$ , where the predictors are correlated, are worse than those of  $\rho = 0$ . Particularly, when  $\gamma = 0$ , the value of RPE of  $\rho = 0.6$  becomes large on the large outlier ratio. However, increasing  $\gamma$  has led to better estimation performance uniformly.
- The results for different simulation models on the same value of  $\gamma$  are generally different, which implies the appropriate value of  $\gamma$  may change according to the data generating mechanisms.

### 6.3. Investigation of Asymptotic Distribution

The asymptotic distribution is derived under the assumption that the true distribution of  $y|x_i$  follows  $f(y|x_i; \beta_0)$ , that is,  $\delta = 0$ . However, we expect that, when  $\gamma$  is sufficiently large and  $\delta$  is moderate, the asymptotic distribution may approximate the true distribution well, a point underlined by Fujisawa and Eguchi ([18], Theorem 5.1) in the case of i.i.d. data. We investigate whether the asymptotic distribution given by Equation (16) appropriately works when there exist outliers.

The asymptotic covariance matrix in Equation (16) depends on  $C(\gamma, h)$ ,  $C_1(\gamma, h)$ , and  $C_2(\gamma, h)$ . For the LPRE, simple calculations provide

$$\begin{aligned} C(\gamma, h) &= \int_0^\infty h(x)^{1+\gamma} dx = \frac{K_\gamma(2+2\gamma)}{2^\gamma K_0(2)^{1+\gamma}}, \\ C_1(\gamma, h) &= \int_0^\infty xh(x)^\gamma h'(x) dx = -\frac{K_\gamma(2+2\gamma)}{(1+\gamma)2^\gamma K_0(2)^{1+\gamma}}, \\ C_2(\gamma, h) &= \int_0^\infty u(x)^2 h(x)^{2\gamma+1} dx \\ &= \frac{2^{1-2\gamma}}{(2\gamma+1)^2 K_0(2)^{2\gamma+1}} \left\{ \gamma(2\gamma+1)K_{2\gamma-2}(4\gamma+2) + (1+\gamma+2\gamma^2)K_{2\gamma-1}(4\gamma+2) \right\}. \end{aligned}$$

The Bessel function of third kind,  $K(\cdot)$ , can be numerically computed, and then we obtain the values of  $C(\gamma, h)$ ,  $C_1(\gamma, h)$ , and  $C_2(\gamma, h)$ .

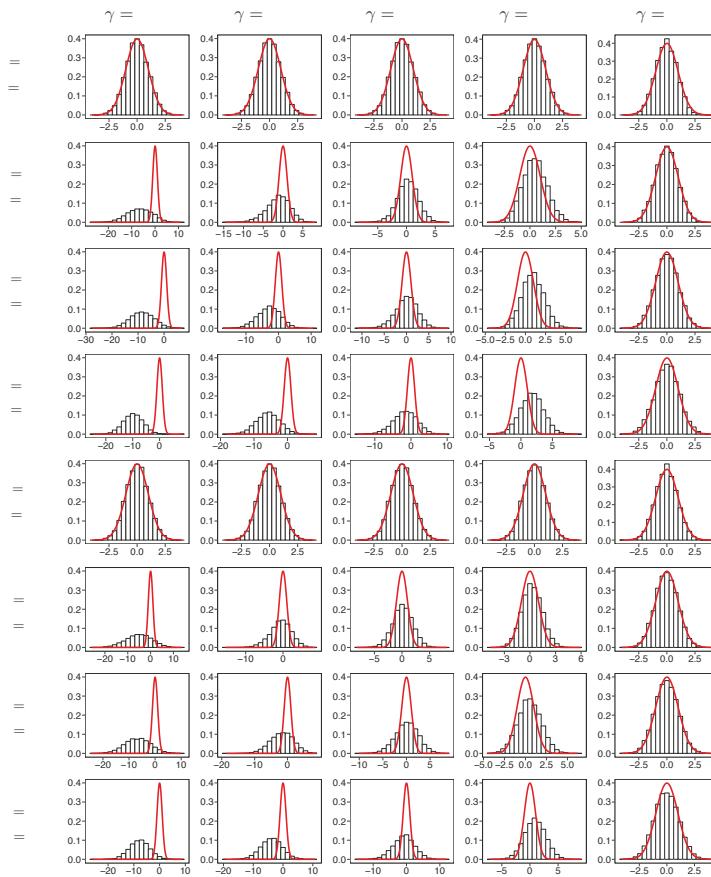
Let  $z = (z_1, \dots, z_p)^T$  be

$$z := \sqrt{n} \left\{ \text{diag} \left( J_\gamma^{-1} \Delta_\gamma J_\gamma^{-1} \right) \right\}^{-\frac{1}{2}} (\hat{\beta}_\gamma - \beta_0).$$

Equation (16) implies that

$$z_j \xrightarrow{\mathcal{L}} N(0, 1), \quad (j = 1, \dots, p).$$

We expect that the histogram of  $z_j$  obtained by the simulation would approximate the density function of the standard normal distribution when there are no (or a few) outliers. When there exists a significant number of outliers, the asymptotic distribution of  $z_j$  may not be  $N(0, 1)$  but is expected to be close to  $N(0, 1)$  for large  $\gamma$ . Figure 2 shows the histograms of T=10,000 samples of  $z_2$  along with the density function of the standard normal distribution for  $\mu = 5$  in Model 1.



**Figure 2.** Histograms of  $T = 100,000$  samples of  $z_2$  along with the density function of standard normal distribution for  $\mu = 5$  in Model 1.

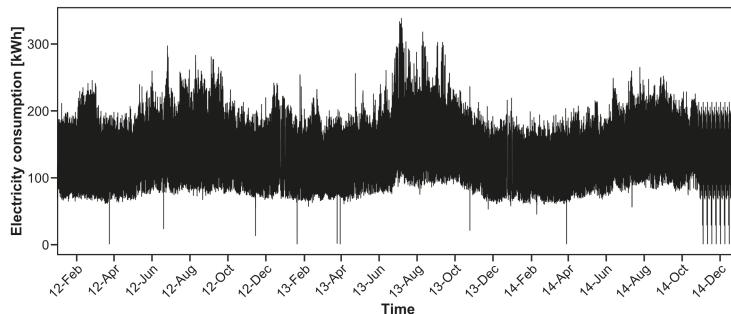
When there are no outliers, the distribution of  $z_2$  is close to the standard normal distribution whatever the value of  $\gamma$  is selected. When the outlier ratio is large, the histogram of  $z_2$  is far from the density function of  $N(0, 1)$  for a small  $\gamma$ . However, when the value of  $\gamma$  is large, the histogram of  $z_2$  is close to the density function of  $N(0, 1)$ , which implies the asymptotic distribution in Equation (16) appropriately approximates the distribution of estimators even when there exist outliers. We observe that the result of the asymptotic distributions for other  $z_j$ s shows a similar tendency to that of  $z_2$ .

## 7. Real Data Analysis

We apply the proposed method to electricity consumption data from the UCI (University of California, Irvine) Machine Learning repository [32]. The dataset consists of 370 household electricity consumption observations from January 2011 to December 2014. The electricity consumption is in kWh at 15-minute intervals. We consider the problem of prediction of the electricity consumption for next day by using past electricity consumption. The prediction of the day ahead electricity consumption is needed when we trade electricity on markets, such as the European Power Exchange (EPEX) day ahead market (<https://www.epexpath.com/en/market-data/dayaheadauction>) and the Japan Power Exchange (JEPX) day ahead market (<http://www.jepx.org/english/index.html>). In the JEPX market,

when the prediction value of electricity consumption  $\hat{y}_t$  is smaller than actual electricity consumption  $y_t$ , the price of the amount of  $y_t - \hat{y}_t$  becomes “imbalance price”, which is usually higher than the ordinary price. For details, please refer to Sioshansi and Pfaffenberger [33].

To investigate the effectiveness of the proposed procedure, we choose one household that includes small positive values of electricity consumption. The consumption data for 25 December 2014 were deleted because the corresponding data values are zero. We predict the electricity consumption from January 2012 to December 2014 (the data in 2011 are used only for estimating the parameter). The actual electricity consumption data from January 2012 to December 2014 are depicted in Figure 3.

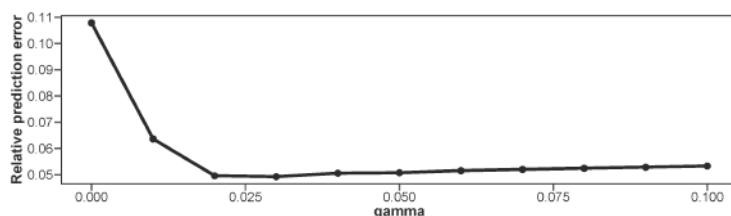


**Figure 3.** Electricity consumption from January 2012 to December 2014 for one of the 370 households.

We observe that several data values are close to zero from Figure 3. Particularly, from October to December 2014, several spikes exist that attain nearly zero values. In this case, the estimation accuracy is poor with ordinary GRE criteria, as shown in our numerical simulation in the previous section.

We assume the multiplicative regression model in Equation (1) to predict electricity consumption. Let  $y_t$  denote the electricity consumption at  $t$  ( $t = 1, \dots, T$ ). The number of observations is  $T = (365 \times 3 + 366 - 1) \times 96 = 146,160$ . Here, 96 is the number of measurements in one day because electricity demand is expressed in 15-minute intervals. We define  $x_t$  as  $x_t = (y_{t-d}, \dots, y_{t-dq})^T$ , where  $d = 96$ . In our model, the electricity consumption at  $t$  is explained by the electricity consumption of the past  $q$  days for the same period. We set  $q = 5$  for data analysis and use past  $n = 100$  days of observations to estimate the model.

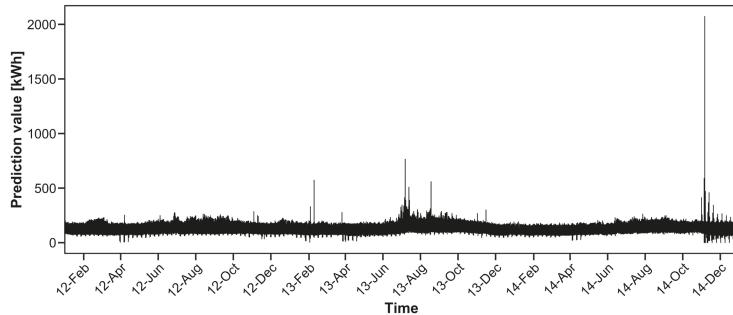
The model parameters are estimated by robust LPRE. The values of  $\gamma$  are set to be regular sequences from 0 to 0.1, with increments of 0.01. To minimize the negative  $\gamma$ -likelihood function, we apply our proposed MM algorithm. As the electricity consumption pattern on weekdays is known to be completely different from that on weekends, we make predictions for weekdays and weekends separately. The results of the relative prediction error are depicted in Figure 4.



**Figure 4.** Relative prediction error for various values of  $\gamma$  for household electricity consumption data.

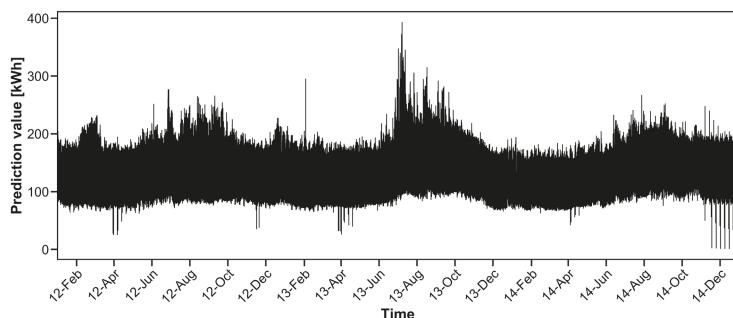
The relative prediction error is large when  $\gamma = 0$  (i.e., ordinary LPRE estimation). The minimum value of relative prediction error is 0.049 and the corresponding value of  $\gamma$  is  $\gamma = 0.03$ . When we set a too large value of  $\gamma$ , efficiency decreases and the relative prediction error might increase.

Figure 5 shows the prediction value when  $\gamma = 0$ . We observe that there exist several extremely large prediction values (e.g., 8 July 2013 and 6 November 2014) due to the model parameters, which are heavily affected by the nearly zero values of electricity consumption.



**Figure 5.** Prediction value based on least product relative error (LPRE) loss for household electricity consumption data.

Figure 6 shows the prediction values when  $\gamma = 0.03$ . Extremely large prediction values are not observed and the prediction values are similar to the actual electricity demand in Figure 3. Therefore, our proposed procedure is robust against outliers.



**Figure 6.** Prediction value based on the proposed method with  $\gamma = 0.03$  for household electricity consumption data.

Additionally, we apply the Yule–Walker method, one of the most popular estimation procedures in the autoregressive (AR) model. Note that the Yule–Walker method does not regard a small positive value of  $y_t$  as an outlier, so that we do not have to conduct the robust AR model for this dataset. The relative prediction error of the Yule–Walker is 0.123, which is larger than that of our proposed method (0.049).

Furthermore, to investigate the stabilities of the MM algorithm described in Section 5, we also apply the BFGS method to obtain the minimizer of the negative  $\gamma$ -likelihood function. The optim function in R is used to implement the BFGS method. With the BFGS method, relative prediction errors diverge when  $\gamma \geq 0.03$ . Consequently, the MM algorithm is more stable than the BFGS algorithm for this dataset.

## 8. Discussion

We proposed a relative error estimation procedure that is robust against outliers. The proposed procedure is based on the  $\gamma$ -likelihood function, which is constructed by  $\gamma$ -cross entropy [18]. We showed that the proposed method has the redescending property, a desirable property in robust statistics literature. The asymptotic normality of the corresponding estimator together with a simple consistent estimator of the asymptotic covariance matrix are derived, which allows the construction of approximate confidence sets. Besides the theoretical results, we have constructed an efficient algorithm, in which we minimize a convex loss function at each iteration. The proposed algorithm monotonically decreases the objective function at each iteration.

Our simulation results showed that the proposed method performed better than the ordinary relative error estimation procedures in terms of prediction accuracy. Furthermore, the asymptotic distribution of the estimator yielded a good approximation, with an appropriate value of  $\gamma$ , even when outliers existed. The proposed method was applied to electricity consumption data, which included small positive values. Although the ordinary LPRE was sensitive to small positive values, our method was able to appropriately eliminate the negative effect of these values.

In practice, variable selection is one of the most important topics in regression analysis. The ordinary AIC (Akaike information criterion, Akaike [34]) cannot be directly applied to our proposed method because the AIC aims at minimizing the Kullback–Leibler divergence, whereas our method aims at minimizing the  $\gamma$ -divergence. As a future research topic, it would be interesting to derive the model selection criterion for evaluating a model estimated by the  $\gamma$ -likelihood method.

High-dimensional data analysis is also an important topic in statistics. In particular, the sparse estimation, such as the lasso [12], is a standard tool to deal with high-dimensional data. As shown in Remark 3, our method may be extended to  $L_1$  regularization. An important point in the regularization procedure is the selection of a regularization parameter. Hao et al. [14] suggested using the BIC (Bayesian information criterion)-type criterion of Wang et al. [35,36] for the ordinary LPRE estimator. It would also be interesting to consider the problem of regularization parameter selection in high-dimensional robust relative error estimation.

In regression analysis, we may formulate two types of  $\gamma$ -likelihood functions: Fujisawa and Eguchi's formulation [18] and Kawashima and Fujisawa's formulation [17]. Kawashima and Fujisawa [22] reported that the difference of performance occurs when the outlier ratio depends on the explanatory variable. In multiplicative regression model in Equation (1), the responses  $y_i$  highly depend on the exploratory variables  $x_i$  compared with the ordinary linear regression model because  $y_i$  is an exponential function of  $x_{ij}$ . As a result, the comparison of the above two formulations of the  $\gamma$ -likelihood functions would be important from both theoretical and practical points of view.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/1099-4300/20/9/632/s1>

**Author Contributions:** K.H. proposed the algorithm, made the R package free, conducted the simulation study, and analyzed the real data; H.M. derived the asymptotics; K.H. and H.M. wrote the paper.

**Funding:** This research was funded by the Japan Society for the Promotion of Science KAKENHI 15K15949, and the Center of Innovation Program (COI) from Japan Science and Technology Agency (JST), Japan (K.H.), and JST CREST Grant Number JPMJCR14D7 (H.M.)

**Acknowledgments:** The authors would like to thank anonymous reviewers for the constructive and helpful comments that improved the quality of the paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Proof of Theorem 1

All of the asymptotics will be taken under  $n \rightarrow \infty$ . We write  $a_n \lesssim b_n$  if there exists a universal constant  $c > 0$  such that  $a_n \leq cb_n$  for every  $n$  large enough. For any random functions  $X_n$  and  $X_0$  on  $\overline{\mathcal{B}}$ , we denote  $X_n(\beta) \xrightarrow{P} X_0(\beta)$  if  $\sup_{\beta \in \overline{\mathcal{B}}} |X_n(\beta) - X_0(\beta)| \xrightarrow{P} 0$ ; below, we will simply write  $\sup_{\beta}$  for  $\sup_{\beta \in \overline{\mathcal{B}}}$ .

First, we state a preliminary lemma, which will be repeatedly used in the sequel.

**Lemma A1.** Let  $\eta(x; \beta)$  and  $\zeta(x, y; \beta)$  be vector-valued measurable functions satisfying that

$$\begin{aligned} \sup_{\beta} \max_{k \in \{0,1\}} |\partial_{\beta}^k \eta(x; \beta)| &\leq \bar{\eta}(x), \\ \sup_{\beta} \max_{k \in \{0,1\}} |\partial_{\beta}^k \zeta(x, y; \beta)| &\leq \bar{\zeta}(x, y), \end{aligned}$$

for some  $\bar{\eta}$  and  $\bar{\zeta}$  such that

$$\bar{\eta} + \int_0^\infty \bar{\zeta}(\cdot, y) dy \in \bigcap_{q>0} L^q(\pi).$$

Then,

$$\frac{1}{n} \sum_{i=1}^n \eta(x_i; \beta) \xrightarrow{p} \int_{\mathcal{X}} \eta(x; \beta) \pi(dx), \quad (\text{A1})$$

$$\frac{1}{n} \sum_{i=1}^n \zeta(x_i, y_i; \beta) \xrightarrow{p} \int_0^\infty \int_{\mathcal{X}} \zeta(x, y; \beta) f(y|x, \beta_0) \pi(dx) dy. \quad (\text{A2})$$

**Proof.** Equation (A1) is a special case of Equation (A2), hence we only show the latter. Observe that

$$\begin{aligned} &\sup_{\beta} \left| \frac{1}{n} \sum_{i=1}^n \zeta(x_i, y_i; \beta) - \iint \zeta(x, y; \beta) f(y|x, \beta_0) \pi(dx) dy \right| \\ &\leq \frac{1}{\sqrt{n}} \sup_{\beta} \left| \sum_{i=1}^n \frac{1}{\sqrt{n}} \left( \zeta(x_i, y_i; \beta) - \int \zeta(x_i, y; \beta) f(y|x_i, \beta_0) dy \right) \right| \\ &\quad + \sup_{\beta} \left| \frac{1}{n} \sum_{i=1}^n \int \zeta(x_i, y; \beta) f(y|x_i, \beta_0) dy - \iint \zeta(x, y; \beta) f(y|x, \beta_0) \pi(dx) dy \right| \\ &=: \frac{1}{\sqrt{n}} \sup_{\beta} |M_n(\beta)| + \sup_{\beta} |C_n(\beta)|. \end{aligned}$$

For the first term, let us recall the Sobolev inequality ([37], Section 10.2):

$$E \left( \sup_{\beta} |M_n(\beta)|^q \right) \lesssim \sup_{\beta} E \{ |M_n(\beta)|^q \} + \sup_{\beta} E \{ |\partial_{\beta} M_n(\beta)|^q \} \quad (\text{A3})$$

for  $q > p$ . The summands of  $M_n(\beta)$  trivially form a martingale difference array with respect to the filtration  $\mathcal{F}_i := \sigma(x_j; j \leq i)$ ,  $i \in \mathbb{N}$ : since we are assuming that the conditional distribution of  $y_i$  given  $\{(x_i, x_{i-1}, x_{i-2}, \dots), (y_{i-1}, y_{i-2}, \dots)\}$  equals that given  $x_i$  (Sections 2 and 3), each summand of  $M_n(\beta)$  equals  $\frac{1}{\sqrt{n}} (\zeta(x_i, y_i; \beta) - E\{\zeta(x_i, y_i; \beta) | \mathcal{F}_{i-1}\})$ . Hence, by means of the Burkholder's inequality for martingales, we obtain, for  $q > p \vee 2$ ,

$$\sup_{\beta} E \{ |M_n(\beta)|^q \} \lesssim \sup_{\beta} \frac{1}{n} \sum_{i=1}^n E \left\{ \left| \left( \zeta(x_i, y_i; \beta) - \int \zeta(x_i, y; \beta) f(y|x_i, \beta_0) dy \right) \right|^q \right\} < \infty.$$

We can take the same route for the summands of  $\partial_{\beta} M_n(\beta)$  to conclude that  $\sup_{\beta} E \{ |\partial_{\beta} M_n(\beta)|^q \} < \infty$ . These estimates combined with Equation (A3) then lead to the conclusion that

$$\sup_{\beta} |M_n(\beta)| = O_p(1).$$

As for the other term, we have  $C_n(\beta) \xrightarrow{p} 0$  for each  $\beta$  and also

$$\sup_n E \left( \sup_\beta |\partial_\beta C_n(\beta)| \right) < \infty.$$

The latter implies the tightness of the family  $\{C_n(\beta)\}_n$  of continuous random functions on the compact set  $\overline{\mathcal{B}}$ , thereby entailing that  $C_n(\beta) \xrightarrow{p} 0$ . The proof is complete.  $\square$

#### Appendix A.1. Consistency

Let  $f_i(\beta) := f(y_i | x_i; \beta)$  for brevity and

$$\begin{aligned} A_{\gamma,n}(\beta) &:= \frac{1}{n} \sum_{i=1}^n f_i(\beta)^\gamma, \\ \bar{A}_{\gamma,n}(\beta) &:= \frac{1}{n} \sum_{i=1}^n \int f(y | x_i; \beta)^{\gamma+1} dy = C(\gamma, h) \frac{1}{n} \sum_{i=1}^n \exp(-\gamma x_i^T \beta). \end{aligned}$$

By means of Lemma A1, we have

$$\begin{aligned} A_{\gamma,n}(\beta) &\xrightarrow{p} A_\gamma(\beta) := \iint f(y | x; \beta)^\gamma f(y | x, \beta_0) \pi(dx) dy, \\ \bar{A}_{\gamma,n}(\beta) &\xrightarrow{p} \bar{A}_\gamma(\beta) := \iint f(y | x; \beta)^{\gamma+1} \pi(dx) dy = C(\gamma, h) \int \exp(-\gamma x^T \beta) \pi(dx). \end{aligned}$$

Since  $\inf_\beta \{A_\gamma(\beta) \wedge \bar{A}_\gamma(\beta)\} > 0$ , we see that taking the logarithm preserves the uniformity of the convergence in probability: for the  $\gamma$ -likelihood function (7), it holds that

$$\ell_{\gamma,n}(\beta) \xrightarrow{p} \ell_{\gamma,0}(\beta) := -\frac{1}{\gamma} \log \{A_\gamma(\beta)\} + \frac{1}{1+\gamma} \log \{\bar{A}_\gamma(\beta)\}. \quad (\text{A4})$$

The limit equals the  $\gamma$ -cross entropy from  $g(\cdot|\cdot) = f(\cdot|\cdot; \beta_0)$  to  $f(\cdot|\cdot; \beta)$ . We have  $\ell_{\gamma,0}(\beta) \geq \ell_{\gamma,0}(\beta_0)$ , the equality holding if and only if  $f(\cdot|\cdot; \beta_0) = f(\cdot|\cdot; \beta)$  (see [17], Theorem 1). By Equation (4), the latter condition is equivalent to  $\rho(e^{-x^T \beta_0} y) = \rho(e^{-x^T \beta} y)$ , followed by  $\beta = \beta_0$  from Assumption 3. This, combined with Equation (A4) and the argmin theorem (cf. [25], Chapter 5), concludes the consistency  $\hat{\beta}_\gamma \xrightarrow{p} \beta_0$ . Note that we do not need Assumption 3 if  $\ell_{\gamma,n}$  is a.s. convex, which generally may not be the case for  $\gamma > 0$ .

#### Appendix A.2. Asymptotic Normality

First, we note that Assumption 2 ensures that, for every  $\alpha > 0$ , there corresponds a function  $\bar{F}_\alpha \in L^1(f(y | x, \beta_0) \pi(dx) dy)$  such that

$$\max_{k=0,1,2,3} \sup_\beta \left| \partial_\beta^k \{f(y | x, \beta)^\alpha\} \right| \leq \bar{F}_\alpha(x, y).$$

This estimate will enable us to interchange the order of  $\partial_\beta$  and the  $dy$ -Lebesgue integration, repeatedly used below without mention.

Let  $s_i(\beta) = s(y_i | x_i; \beta)$ , and

$$\begin{aligned} S_{\gamma,n}(\beta) &:= \frac{1}{n} \sum_{i=1}^n f_i(\beta)^\gamma s_i(\beta), \\ \bar{S}_{\gamma,n}(\beta) &:= \frac{1}{n} \sum_{i=1}^n \int f(y | x_i; \beta)^{\gamma+1} s(y | x_i; \beta) dy. \end{aligned}$$

Then, the  $\gamma$ -likelihood equation  $\partial_{\beta} \ell_{\gamma,n}(\beta) = \mathbf{0}$  is equivalent to

$$\Psi_{\gamma,n}(\beta) := \bar{A}_{\gamma,n}(\beta)S_{\gamma,n}(\beta) - A_{\gamma,n}(\beta)\bar{S}_{\gamma,n}(\beta) = \mathbf{0}.$$

By the consistency of  $\hat{\beta}_{\gamma}$ , we have  $P(\hat{\beta}_{\gamma} \in \mathcal{B}) \rightarrow 1$ ; hence  $P\{\Psi_{\gamma,n}(\hat{\beta}_{\gamma}) = \mathbf{0}\} \rightarrow 1$  as well, for  $\mathcal{B}$  is open. Therefore, virtually defining  $\hat{\beta}_{\gamma}$  to be  $\beta_0 \in \mathcal{B}$  if  $\Psi_{\gamma,n}(\hat{\beta}_{\gamma}) = \mathbf{0}$  has no root, we may and do proceed as if  $\Psi_{\gamma,n}(\hat{\beta}_{\gamma}) = \mathbf{0}$  a.s. Because of the Taylor expansion

$$\left( - \int_0^1 \partial_{\beta} \Psi_{\gamma,n} (\beta_0 + s(\hat{\beta}_n - \beta_0)) ds \right) \sqrt{n} (\hat{\beta}_{\gamma} - \beta_0) = \sqrt{n} \Psi_{\gamma,n}(\beta_0)$$

to conclude Equation (16), it suffices to show that (recall the definitions (14) and (15))

$$\sqrt{n} \Psi_{\gamma,n}(\beta_0) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, \Delta_{\gamma}), \quad (\text{A5})$$

$$-\partial_{\beta} \Psi_{\gamma,n}(\hat{\beta}'_n) \xrightarrow{p} J_{\gamma} \quad \text{for every } \hat{\beta}'_n \xrightarrow{p} \beta_0. \quad (\text{A6})$$

First, we prove Equation (A5). By direct computations and Lemma A1, we see that

$$\begin{aligned} \sqrt{n} \Psi_{\gamma,n} &= \bar{A}_{\gamma,n} \sqrt{n} (S_{\gamma,n} - \bar{S}_{\gamma,n}) - \sqrt{n} (A_{\gamma,n} - \bar{A}_{\gamma,n}) \bar{S}_{\gamma,n} \\ &= \sum_{i=1}^n \frac{1}{\sqrt{n}} \left\{ H'_{\gamma} \left( f_i^{\gamma} s_i - \int f(y|x_i; \beta_0)^{\gamma+1} s(y|x_i; \beta_0) dy \right) - \left( f_i^{\gamma} - \int f(y|x_i; \beta_0)^{\gamma+1} dy \right) H''_{\gamma} \right\} \\ &=: \sum_{i=1}^n \chi_{\gamma,i}. \end{aligned}$$

The sequence  $(\chi_{\gamma,i})_{i \leq n}$  is an  $(\mathcal{F}_j)$ -martingale-difference array. It is easy to verify the Lapunov condition:

$$\exists \alpha > 0, \quad \sup_n \sup_{i \leq n} E(|\chi_{\gamma,i}|^{2+\alpha}) < \infty.$$

Hence, the martingale central limit theorem concludes Equation (A5) if we show the following convergence of the quadratic characteristic:

$$\frac{1}{n} \sum_{i=1}^n E(\chi_{\gamma,i}^{\otimes 2} | \mathcal{F}_{i-1}) \xrightarrow{p} \Delta_{\gamma}.$$

This follows upon observation that

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n E(\chi_{\gamma,i}^{\otimes 2} | \mathcal{F}_{i-1}) \\ &= (H'_{\gamma})^2 \frac{1}{n} \sum_{i=1}^n \text{var}(f_j^{\gamma} s_j | \mathcal{F}_{i-1}) + (H''_{\gamma})^{\otimes 2} \frac{1}{n} \sum_{i=1}^n \text{var}(f_j^{\gamma} | \mathcal{F}_{i-1}) - 2H'_{\gamma} \frac{1}{n} \sum_{i=1}^n \text{cov}(f_j^{\gamma} s_j, f_j^{\gamma} | \mathcal{F}_{i-1}) H''_{\gamma} \\ &= (H'_{\gamma})^2 \left\{ \iint f(y|x; \beta_0)^{2\gamma+1} s(y|x; \beta_0)^{\otimes 2} dy \pi(dx) - (H''_{\gamma})^{\otimes 2} \right\} \\ &\quad + (H''_{\gamma})^{\otimes 2} \left\{ \iint f(y|x; \beta_0)^{2\gamma+1} dy \pi(dx) - (H'_{\gamma})^2 \right\} \\ &\quad - 2H'_{\gamma} \left\{ \iint f(y|x; \beta_0)^{2\gamma+1} s(y|x; \beta_0) dy \pi(dx) - H'_{\gamma} H''_{\gamma} \right\} (H''_{\gamma})^T + o_p(1) \\ &= (H'_{\gamma})^2 \iint f(y|x; \beta_0)^{2\gamma+1} s(y|x; \beta_0)^{\otimes 2} dy \pi(dx) \\ &\quad + (H''_{\gamma})^{\otimes 2} \iint f(y|x; \beta_0)^{2\gamma+1} dy \pi(dx) - 2H'_{\gamma} \iint f(y|x; \beta_0)^{2\gamma+1} s(y|x; \beta_0) dy \pi(dx) (H''_{\gamma})^T + o_p(1) \\ &= \Delta_{\gamma} + o_p(1) \end{aligned}$$

invoke the expression (4) for the last equality.

Next, we show Equation (A6). Under the present regularity condition, we can deduce that

$$\sup_{\beta} |\partial_{\beta}^2 \Psi_{\gamma,n}(\beta)| = O_p(1).$$

It therefore suffices to verify that  $-\partial_{\beta} \Psi_{\gamma,n}(\beta_0) \xrightarrow{p} J_{\gamma}(\beta_0) = J_{\gamma}$ . This follows from a direct computation of  $-\partial_{\beta} \Psi_{\gamma,n}(\beta_0)$ , combined with the applications of Lemma A1 (note that  $\bar{A}_{\gamma,n}$  and  $A_{\gamma,n}$  have the same limit in probability):

$$\begin{aligned} -\partial_{\beta} \Psi_{\gamma,n}(\beta_0) &= A_{\gamma,n} \left( \frac{1}{n} \sum_{i=1}^n \int f(y|x_i; \beta_0)^{\gamma+1} s(y|x_i; \beta_0)^{\otimes 2} dy \right) - \bar{S}_{\gamma,n} S_{\gamma,n}^T \\ &\quad + \gamma \left( S_{\gamma,n} \bar{S}_{\gamma,n}^T - \bar{S}_{\gamma,n} S_{\gamma,n}^T \right) \\ &\quad + \gamma \left\{ A_{\gamma,n} \left( \frac{1}{n} \sum_{i=1}^n \int f(y|x_i; \beta_0)^{\gamma+1} s(y|x_i; \beta_0)^{\otimes 2} dy \right) - \bar{A}_{\gamma,n} \left( \frac{1}{n} \sum_{i=1}^n f_i^{\gamma} s_i^{\otimes 2} \right) \right\} \\ &\quad + \left\{ A_{\gamma,n} \left( \frac{1}{n} \sum_{i=1}^n \int f(y|x_i; \beta_0)^{\gamma+1} \partial_{\beta} s(y|x_i; \beta_0) dy \right) - \bar{A}_{\gamma,n} \left( \frac{1}{n} \sum_{i=1}^n f_i^{\gamma} \partial_{\beta} s_i \right) \right\} \\ &= A_{\gamma,n} \left( \frac{1}{n} \sum_{i=1}^n \int f(y|x_i; \beta_0)^{\gamma+1} s(y|x_i; \beta_0)^{\otimes 2} dy \right) - \bar{S}_{\gamma,n} S_{\gamma,n}^T + o_p(1) \\ &= H'_{\gamma} \iint f(y|x; \beta_0)^{\gamma+1} s(y|x; \beta_0)^{\otimes 2} dy \pi(dx) - (H''_{\gamma})^{\otimes 2} + o_p(1) \\ &= J_{\gamma} + o_p(1). \end{aligned}$$

### Appendix A.3. Consistent Estimator of the Asymptotic Covariance Matrix

Thanks to the stability assumptions on the sequence  $x_1, x_2, \dots$ , we have

$$\frac{1}{n} \sum_{i=1}^n x_i^{\otimes k} \exp(-\gamma x_i^T \beta_0) \xrightarrow{p} \Pi_k(\gamma), \quad k = 0, 1, 2.$$

Moreover, for  $\delta, \delta' > 0$  given in Assumption 1, we have

$$\begin{aligned} &\left| \frac{1}{n} \sum_{i=1}^n x_i^{\otimes k} \exp(-\gamma x_i^T \beta_0) - \frac{1}{n} \sum_{i=1}^n x_i^{\otimes k} \exp(-\gamma x_i^T \hat{\beta}_{\gamma}) \right| \\ &\lesssim \left( \frac{1}{n} \sum_{i=1}^n |x_i|^{k+1} \exp(\delta' |x_i|^{1+\delta}) \right) |\hat{\beta}_{\gamma} - \beta_0| = O_p(1) |\hat{\beta}_{\gamma} - \beta_0| \xrightarrow{p} 0. \end{aligned}$$

These observations are enough to conclude Equation (17).

## References

1. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2009.
2. Park, H.; Stefanski, L.A. Relative-error prediction. *Stat. Probab. Lett.* **1998**, *40*, 227–236. [[CrossRef](#)]
3. Ye, J. Price Models and the Value Relevance of Accounting Information. *SSRN Electronic Journal* 2007. Available online: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1003067](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1003067) (accessed on 20 August 2018).
4. Van der Meer, D.W.; Widén, J.; Munkhammar, J. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renew. Sust. Energ. Rev.* **2018**, *81*, 1484–1512. [[CrossRef](#)]

5. Mount, J. Relative error distributions, without the heavy tail theatrics. 2016. Available online: <http://www.win-vector.com/blog/2016/09/relative-error-distributions-without-the-heavy-tail-theatrics/> (accessed on 20 August 2018).
6. Chen, K.; Guo, S.; Lin, Y.; Ying, Z. Least Absolute Relative Error Estimation. *J. Am. Stat. Assoc.* **2010**, *105*, 1104–1112. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Li, Z.; Lin, Y.; Zhou, G.; Zhou, W. Empirical likelihood for least absolute relative error regression. *TEST* **2013**, *23*, 86–99. [\[CrossRef\]](#)
8. Chen, K.; Lin, Y.; Wang, Z.; Ying, Z. Least product relative error estimation. *J. Multivariate Anal.* **2016**, *144*, 91–98. [\[CrossRef\]](#)
9. Ding, H.; Wang, Z.; Wu, Y. A relative error-based estimation with an increasing number of parameters. *Commun. Stat. Theory Methods* **2017**, *47*, 196–209. [\[CrossRef\]](#)
10. Demongeot, J.; Hamie, A.; Laksaci, A.; Rachdi, M. Relative-error prediction in nonparametric functional statistics: Theory and practice. *J. Multivariate Anal.* **2016**, *146*, 261–268. [\[CrossRef\]](#)
11. Wang, Z.; Chen, Z.; Chen, Z. H-relative error estimation for multiplicative regression model with random effect. *Comput. Stat.* **2018**, *33*, 623–638. [\[CrossRef\]](#)
12. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Methodol.* **1996**, *58*, 267–288.
13. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B Stat. Methodol.* **2006**, *68*, 49–67. [\[CrossRef\]](#)
14. Hao, M.; Lin, Y.; Zhao, X. A relative error-based approach for variable selection. *Comput. Stat. Data Anal.* **2016**, *103*, 250–262. [\[CrossRef\]](#)
15. Liu, X.; Lin, Y.; Wang, Z. Group variable selection for relative error regression. *J. Stat. Plan. Inference* **2016**, *175*, 40–50. [\[CrossRef\]](#)
16. Xia, X.; Liu, Z.; Yang, H. Regularized estimation for the least absolute relative error models with a diverging number of covariates. *Comput. Stat. Data Anal.* **2016**, *96*, 104–119. [\[CrossRef\]](#)
17. Kawashima, T.; Fujisawa, H. Robust and Sparse Regression via  $\gamma$ -Divergence. *Entropy* **2017**, *19*, 608. [\[CrossRef\]](#)
18. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Anal.* **2008**, *99*, 2053–2081. [\[CrossRef\]](#)
19. Maronna, R.; Martin, D.; Yohai, V. *Robust Statistics*; John Wiley & Sons: Chichester, UK, 2006.
20. Koudou, A.E.; Ley, C. Characterizations of GIG laws: A survey. *Probab. Surv.* **2014**, *11*, 161–176. [\[CrossRef\]](#)
21. Jones, M.C.; Hjort, N.L.; Harris, I.R.; Basu, A. A comparison of related density-based minimum divergence estimators. *Biometrika* **2001**, *88*, 865–873. [\[CrossRef\]](#)
22. Kawashima, T.; Fujisawa, H. On Difference between Two Types of  $\gamma$ -divergence for Regression. 2018. Available online: <https://arxiv.org/abs/1805.06144> (accessed on 20 August 2018).
23. Ferrari, D.; Yang, Y. Maximum Lq-likelihood estimation. *Ann. Stat.* **2010**, *38*, 753–783. [\[CrossRef\]](#)
24. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559, doi:10.1093/biomet/85.3.549. [\[CrossRef\]](#)
25. Van der Vaart, A.W. *Asymptotic Statistics*; Vol. 3, Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: Cambridge, UK, 1998.
26. Eguchi, S.; Kano, Y. Robustifying maximum likelihood estimation by psi-divergence. *ISM Research Memorandum*. 2001. Available online: <https://www.researchgate.net/profile/ShintoEguchi/publication/228561230Robustifyingmaximumlikelihoodestimationbypsi-divergence/links545d65910cf2c1a63bfa63e6.pdf> (accessed on 20 August 2018).
27. Hunter, D.R.; Lange, K. A tutorial on MM algorithms. *Am. Stat.* **2004**, *58*, 30–37. [\[CrossRef\]](#)
28. Böhning, D. Multinomial logistic regression algorithm. *Ann. Inst. Stat. Math.* **1992**, *44*, 197–200. [\[CrossRef\]](#)
29. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **2005**, *67*, 301–320. [\[CrossRef\]](#)
30. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [\[CrossRef\]](#)
31. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1. [\[CrossRef\]](#) [\[PubMed\]](#)

32. Dheeru, D.; Karra Taniskidou, E. UCI Machine Learning Repository, 2017. Available online: <https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014> (accessed on 20 August 2018).
33. Sioshansi, F.P.; Pfaffenberger, W. *Electricity Market Reform: An International Perspective*; Elsevier: Oxford, UK, 2006.
34. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **1974**, *19*, 716–723. [[CrossRef](#)]
35. Wang, H.; Li, R.; Tsai, C.L. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **2007**, *94*, 553–568. [[CrossRef](#)] [[PubMed](#)]
36. Wang, H.; Li, B.; Leng, C. Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Series B Stat. Methodol.* **2009**, *71*, 671–683. [[CrossRef](#)]
37. Friedman, A. *Stochastic Differential Equations and Applications*; Dover Publications: New York, NY, USA, 2006.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

MDPI  
St. Alban-Anlage 66  
4052 Basel  
Switzerland  
Tel. +41 61 683 77 34  
Fax +41 61 302 89 18  
[www.mdpi.com](http://www.mdpi.com)

*Entropy* Editorial Office  
E-mail: [entropy@mdpi.com](mailto:entropy@mdpi.com)  
[www.mdpi.com/journal/entropy](http://www.mdpi.com/journal/entropy)





MDPI  
St. Alban-Anlage 66  
4052 Basel  
Switzerland

Tel: +41 61 683 77 34  
Fax: +41 61 302 89 18  
[www.mdpi.com](http://www.mdpi.com)



ISBN 978-3-03897-937-1