CHAPTER 1

Literature Review

1.1. Empirical Likelihood

Empirical likelihood is a nonparametric method of inference based on data-driven likelihood ratio function. Like the bootstrap and jackknife, empirical likelihood inference does not require us to specify a family of distributions for the data. Also like parametric likelihood methods, empirical likelihood makes an automatic determination of the shape of confidence regions; it straightforwardly incorporates side information expressed through constraints or prior distributions; it extends to biased sampling and censored data, and it has very favorable asymptotic power properties.

Empirical likelihood was first proposed by Owen (1988) as follows. For i.i.d sample $X_1, \ldots, X_n$, the empirical likelihood ration function is

$$R(\mu) = \max\left\{ \prod_{i=1}^{n} nw_i \middle| \sum_{i=1}^{n} w_i X_i = \mu, w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\},$$

and the resulting asymptotic confidence region is

$$\{\mu \mid R(\mu) \geq \chi^2_{1,\alpha}\} = \left\{ \sum_{i=1}^{n} w_i X_i \middle| \prod_{i=1}^{n} nw_i \geq \chi^2_{1,\alpha}, w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\}. \tag{1.1.1}$$

The following research generalizes this method to more complex settings,

$$R(\theta) = \max\left\{ d\left(\sum_{i=1}^{n} w_i I_{[x=X_i]}, \sum_{i=1}^{n} \frac{1}{n} I_{[x=X_i]} \right) \middle| \sum_{i=1}^{n} w_i g(X_i, \theta) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\},$$

where $g(\cdot,\cdot)$ is general estimating function, and $d(\cdot,\cdot)$ is some divergence measure. For more details, see Owen (2010).

1.1.1. Different g. Owen (1991) first considered normal equations in linear regression as g. Later, almost all the facets in traditional linear regression were transplanted into empirical likelihood. Chen (1993, 1994) developed confidence regions for regression coefficients. Jing (1995), Adimari (1995) considered the problem of comparing the means of two populations. Davidian and Carroll (1987) investigated different variance structures in regression. Kolaczyk (1995) formulated empirical information criterion as an empirical likelihood version of AIC.

Besides linear regression, the authors in this area also took efforts to incorporate more complex models into empirical likelihood. Chen and Hall (1993) used kernel smoothing technique to construct g for quantiles. Whang (2006) extended this technique into empirical likelihood quantile regression. Kolaczyk (1994) further extended g into generalized linear model. Hall and Owen (1993) studied empirical likelihood confidence bands for kernel density estimates. Chen and Qin (2000)

studied empirical likelihood confidence intervals for local linear kernel smoothing. Peng (2004) investigated empirical likelihood confidence intervals for heavy-tailed distributions.

Data drawn from sophisticated designs have also been taken into consideration. Qin (1993) applied empirical likelihood to biased sampling data, whose distribution G related the distribution of interest F through a biasing function u, that is

$$G(A) = \frac{\int_A u(y)\ \mathrm{d}F(y)}{\int u(y)\ \mathrm{d}F(y)},$$

for every Borel set A. This result was further generalized by Qin and Zhang (1997), Qin et al. (1999) and Qin (1998) for biased sampling. Chen and Qin (1993) presented empirical likelihood for samples from a finite population. Chen and Sitter (1999) formulated an empirical likelihood that incorporated design weights. Wu and Sitter (2001) considered a setting where the entire population is known but only the sample are available. Loh et al. (1996) investigated Latin hypercube samples by empirical likelihood confidence region. Empirical likelihood analysis of cumulative hazard function was established by Murphy (1995). Adimari (1997) considered empirical likelihood inferences for the mean under independent right censoring and later Pan and Zhou (2002) extended it to do inference about more general functionals of the hazard function $\int q_n(x)\ \mathrm{d}\Lambda(x)$. Murphy and van der Vaart (1997) considered general double censoring and proportional hazard model.

Inference on infinite dimensional parameters, such as CDF and quantile functions are also visited. Owen (1995) built exact confidence bands for CDF based on empirical likelihood. Hollander et al. (1997) found asymptotic confidence bands for survival function for right censoring data. Zhang (1996a, 1999) constructed confidence bands when auxiliary information was available.

Finally, for independent data, I refer to Qin and Lawless (1994) as a useful general case. They analyzed the behavior of empirical likelihood when g was a smooth estimating function and number of parameters was less than number of constraint equations. They proposed maximum empirical likelihood estimators which could be viewed as an alternative to least square estimators when the data could not square with the model. Molanes Lopez et al. (2009) extended this work into non-smooth criterion function.

Dependent data can also be analyzed by empirical likelihood. Chuang and Chan (2002) applied empirical likelihood to unstable auto-regression. Kitamura et al. (1997) developed block-wise empirical likelihood for weakly dependent data and extended the result in Qin and Lawless (1994). Chen and Wong (2009) combined block-wise and smooth techniques in order to estimate quantiles for weakly dependent data. Nordman et al. (2007) applied block-wise empirical likelihood for long-range dependence processes but the limit distribution became a multiple Wiener integral instead of simple χ^2. The spectral approach to empirical likelihood was due to Monti (1997). Mykland (1995) extended empirical likelihood into martingale by dual likelihood technique. Suppose $m_n(\theta)$ is a zero mean martingale depending on data and model, for example, a sample estimating equation or martingale leading to Nelson–Aalen estimator, then the log dual likelihood $l(\mu)$ for dual parameter μ and fixed θ is defined by partial differential equations

$$m_n(\theta) = \nabla l_\theta(\mu)\big|_{\mu=0},$$

or through Doléans–Dade multiplicative martingale

$$l_\theta(\mu) = -\mu^T \Lambda_t(\theta) + \sum_{s \le t} \left(1 + \mu^T \Delta m_s(\theta)\right).$$

This definition coincides empirical likelihood for i.i.d data and discrete time. Wang and Zhu (2011) addressed quantile and the corresponding likelihood ratio test is $\mathrm{LR} = 2\sup_\mu l_\theta(\mu)$. regression with longitudinal data by empirical likelihood. A more recent study by Bandyopadhyay et al. (2015) introduced empirical likelihood into spatial data analysis.

1.1.2. Different d. There is also literature on the effect of different empirical divergence measures in empirical likelihood settings. Hall and Presnell (1999) used empirical entropy to construct robust confidence regions for non-robust statistics like the mean. Baggerly (1998) extended d to Cressie-Read family in the sample mean case and proved empirical likelihood was the only Bartlett correctable member of that family. In econometrics, Mittelhammer et al. (2000) used empirical entropy and Cressie–Read divergence measure in d and linked maximum empirical likelihood estimator to general method of moment. To overcome the drawback that empirical likelihood domain Θ was usually smaller than the full parameter space $\mathbb{R}^p$, Tsao and Wu (2014) defined a surjective composite similarity mapping

$$h^C(\theta) = \hat{\theta}_{\mathrm{MELE}} + \left(1 + \frac{R(\theta)}{2n}\right)\left(\theta - \hat{\theta}_{\mathrm{MELE}}\right),$$

from Θ to $\mathbb{R}^p$, and used its generalized inverse to adjust original empirical likelihood into $R\left(\left(h^C\right)^{-1}(\theta)\right)$.

1.1.3. Higher-Order Properties. Higher order properties have also been investigated. Bartlett correctable is one of the most favorite aspects of using empirical likelihood. This correction is usually based on an Edgeworth expansion of the empirical likelihood version, and has the following form: instead of using 1.1.1, the more accurate version is

$$\left\{\theta \mid -2\log R(\theta) \le \left(1 + \frac{a}{n}\right)\chi^2_{p,1-\alpha}\right\},$$

or

$$\left\{\theta \mid -2\log R(\theta) \le \left(1 - \frac{a}{n}\right)^{-1}\chi^2_{p,1-\alpha}\right\}.$$

The fist Bartlett correction was given in DiCiccio et al. (1991) for the sample mean. Zhang (1996b) extended it to general estimating function. However, Lazar and Mykland (1999) reported failure of Bartlett correction when there were nuisance parameters. This is where the empirical likelihood shows different behavior from parametric likelihoods. This problem was solved by Chen and Cui (2006). Other Bartlett correction usually follows the development of the empirical likelihood model. Biao Zhang (1998) showed that there was no first order asymptotic benefit from global side constraints in kernel density estimation. However, Chen (1997) proved that the second order asymptotic benefit could be expected via imposing side information. Newey and Smith (2004) considered higher order properties of maximum empirical likelihood estimators in the general method of moments setting. Kitamura (2001) gave some theorems on the large deviation properties of empirical likelihood, and his simulation suggested empirical likelihood performed very well at hypotheses father from the null.

1.1.4. High Dimension and Sparsity. High dimensional data and sparsity are pivotal topics in recent years and they also shed light on empirical likelihood. Tsao et al. (2004) pointed out the failure of empirical likelihood even if the number of parameters was moderately larger than the number of observations $p/n > 1/2$. Hjort et al. (2009) rescaled the empirical likelihood itself by number of parameters and asserted a normal limit distribution. Their work was generalized by Chen et al. (2009) with less restrictions. Chen et al. (2008) and Emerson et al. (2009) proposed adjusted empirical likelihood by adding pseudo observations. Bartolucci (2007) added Tikhonov regularization defined sample covariance matrix and succeeded when growing $p < n$. Lahiri et al. (2012) added component wise penalties and extended the result into weak and long-range dependency and $p > n$. Tang and Leng (2010) added popular SCAD penalty to traditional empirical likelihood for population mean and validated sparsity in high dimensional settings. Leng and Tang (2012) extended this work to generalized estimating equations and more general penalties. An interesting result from Chang et al. (2015) shows the usual block-wise empirical likelihood is still working if the number of constraints is growing with the number of parameters in some rate.

1.1.5. Bayesian. Compared to the fruitful frequentist research in empirical likelihood, Bayesian counterparts are just at the beginning. Lazar (2003) started the Bayesian empirical likelihood and did many simulations to prove its validation. Schennach (2005), Schennach et al. (2007) explained the Bayesian exponentially tilted empirical likelihood as the limit of some nonparametric procedure. Grendár and Judge (2009) found the equivalence between maximum empirical likelihood estimators and Bayesian maximum a posteriori probability estimators under model misspecification. Lancaster and Jae Jun (2010) used Bayesian exponentially tilted empirical likelihood for quantile regression. Fang and Mukerjee (2005, 2006) established probability matching prior for empirical likelihood for population mean case, which allowed the credible intervals to have validity in the frequentist sense. Chang and Mukerjee (2008) consolidated this result by showing empirical likelihood was the only member who enjoyed both Bayesian and frequentist validity. Vexler et al. (2013) used empirical likelihood as a non-parametric method to estimate Bayes factor in model selection. Vexler et al. (2014) showed some James-Stein phenomenon for Bayesian empirical likelihood. Rao and Wu (2010) a prior on empirical weights w_i in complex survey settings.

1.2. Option Pricing

Since the seminal works by Black and Scholes (1973) and Merton (1973), option valuation methodologies have been extensively developed. The Black-Scholes model has become one of the most well-known discoveries in finance literature, which relates the cross-sectional properties of option prices with the underlying asset return distribution. However, Rubinstein (1985), Melino and Turnbull (1990) pointed out several limitations in the Black-Scholes model due to the strong assumptions, such as non-normality of the returns, stochastic volatility (implied volatility smile), jumps and others. Both parametric and nonparametric approaches have been proposed to deal with these issues.

Scott (1987), Hull and White (1987) and Wiggins (1987) extended the Black and Scholes model and allowed the volatility to be stochastic. Heston (1993) developed a closed-form solution for option pricing with the underlying assets volatility

being stochastic. Duan (1995) proposed a GARCH option pricing model in an attempt to explain some systematic biases associated with the Black-Scholes model. Later Heston and Nandi (2000) provided a closed-form solution for option pricing with the underlying assets volatility following GARCH(p,q) process. Bates (1996), Bakshi, Cao and Chen (1997) derived an option pricing model with stochastic volatility and jumps. Kou (2002) provided a solution to pricing the option with the double exponential jumps diffusion process. Carr and Madan (1999) introduced the fast Fourier transform approach to option pricing given a specified characteristic function of the return, which provides an efficient computational algorithm to calculate the option prices. For further reference, see Duffie et al. (2000), Bakshi and Madan (2000) and Carr and Madan (2009) among others. All these methods are parametric based, which assume a parametric form of either the distribution of the underlying assets returns or the characteristic function of the underlying assets returns.

Nonparametric approaches have also been proposed to capture the underlying asset and option price data to reconstruct the structure of the diffusion process. For example, Hutchinson, Lo and Poggio (1994) applied the neural network techniques to price the derivatives. Ait-Sahalia and Lo (1998) used the kernel regression to fit the state-price density implicitly in option pricing. Ait-Sahalia (1996) proposed a nonparametric pricing estimation procedure for interest rate derivative securities under the assumption that the unknown volatility is independent of time. Stutzer (1996) adopted the canonical valuation method, which incorporats the no-arbitrary principle embodied in the formula for calculating the expectation of the discounted value of assets under the risk-neutral probability distribution.

1.3. Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) methods, also known as likelihood-free techniques, have appeared in the past ten years as the most satisfactory approach to intractable likelihood problems, first in genetics then in a broader spectrum of applications. Intractable likelihood is a common phenomenon in statistical modeling.

- The likelihood is expressed as a multidimensional integral,

$$l(\theta \mid Y) = \int l^*(\theta \mid Y, u)\, \mathrm{d}u,$$

  where Y is observation, u is latent variable and θ is the parameter of interest, for example, coalecent model in population genetics. Typically when the dimension of u is large, the convergence properties of MCMC like Gibbs sampler and Metropolis–Hastings algorithm are too poor to use in practice.
- The normalizing constant is unknown. This is typically the case of Gibbs random fields in order to model spatially correlated data such as epidemiology and image analysis.
- The likelihood function is not completely known, that is

$$l(\theta \mid Y) = l_1(\theta \mid Y)\, l_2(\theta).$$

  In the past, Laplace approximations by Tierney and Kadane (1986) or variational Bayes solutions by Jaakkola and Jordan (2000) have been advanced for such problems. However, Laplace approximations require some analytic knowledge of the posterior distribution, while variational Bayes solutions

replace the true model with another pseudo-model which is usually much simpler and thus misses some of the features of the original model.

The idea of ABC dated back to Rubin (1984) as an intuitive way to understand posterior distributions from a frequentist's perspective, because parameters from the posterior are more likely to be those that could have generated the observed data. The first ABC algorithm was born in Tavaré et al. (1997) and Pritchard et al. (1999) as For more details, see Marin et al. (2012).

Algorithm 1 Prichard's Modified ABC

1 Sample parameters θ_i from the prior distribution $\pi(\theta)$;
2 Sample data Z_i based on the model $f(z \mid \theta_i)$;
3 Accept θ_i if $\rho(S(Z_i), S(X_{\text{obs}})) \leq \varepsilon$, for some metric ρ and summary statistics S.

1.3.1. Different Sampling Schemes. In practice, if one uses non-informative prior, simulation would be very inefficient, because of high rejection rate of prior sample locating in low posterior probability regions. As an answer to this problem, Marjoram et al. (2003) applied Metropolis–Hastings algorithm in sampling from prior distributions and built MCMC-ABC algorithm. Picchini (2014) used this method to analyze data from stochastic differential equations. Lee and Łatuszyński (2014) backed up this method by several theoretical results including variance bound and geometric ergodicity. Ratmann et al. (2009) used tolerance level $\varepsilon = \rho(S_i, S_{\text{obs}})$ as an additional parameter of the model and proposed ABC_μ, as method to assess model uncertainty. Wilkinson (2013) replaced the hard accept-reject scheme by soft kernel smoothing, called noisy ABC,

$$\pi_\varepsilon(\theta, z \mid Y) \propto \pi(\theta) f(z \mid \theta) K_\varepsilon(Y - z),$$

where K_ε is a well-chosen kernel parameterized by the bandwidth ε. He also made the valuable point that noisy ABC simulated exactly from the posterior conditioning on observations with errors. Sisson et al. (2007) combine partial rejection control and ABC to solve the inefficiency when prior and posterior are dissimilar. In order to analyze hidden Markov models, Jasra et al. (2012) proposed ABC filtering incorporated sequential Monte Carlo (SMC) method. This procedure was theoretically justified in parameter estimation by Dean et al. (2014). Using SMC sampler will result in a bias in approximation to the posterior. To overcome this problem, Beaumont et al. (2009) incorporated population Monte Carlo method into ABC-PRC by modifying the importance sampling weights with an component-wise random walk estimator of likelihood and using decreasing tolerance levels. MCMC-ABC also suffers from poor mixture properties when tolerance level is small. To rectify this weakness, Baragatti et al. (2013) proposed ABC parallel tempering scheme. The basic technique is running several MCMC-ABC chains with different tolerance levels and swap some chains under certain conditions. They recommended ABC-PT when the posterior was multi-modal.

1.3.2. Calibration. McKinley et al. (2009) performed a simulation comparing ABC-MCMC and ABC-SMC. The conclusions are the choice of the distance, the summary statistics are paramount to the success of ABC, while the tolerance level does not seem to have a strong influence.

For parameter estimation, the ideal summary statistics would be the sufficient statistics. However, for most real problems, it is impossible to find them. Joyce and Marjoram (2008) considered sequential inclusion of summary statistics based on likelihood ratios. Nevertheless, their method does not address the issue of construction of summary statistics and does not take into account the sequential nature of likelihood ratios. Aeschbacher et al. (2012) advocated inclusion via boosting. Blum et al. (2013) summarized several other methods to select summary statistics, such as information criterion, partial least square regression, neural network. Fearnhead and Prangle (2012) used polynomial regression to estimate the posterior mean and used it as a summary statistic. To our knowledge, this is the first method which is able to construct automatic summary statistics. Ruli et al. (2013) suggested using score function of composite likelihood as summary statistics. Barthelmé and Chopin (2014) applied expectation propagation approximation in ABC, that is approximating the posterior by

$$\pi(\theta)\prod_{i=1}^{n}\int f(Z_i \mid Y_1,\ldots,Y_{i-1},\theta)\, I_{[|S_i(Z_i)-S_i(Y_i)|\leq\varepsilon]}\,\mathrm{d}Z_i,$$

where S_i is a local summary statistics for a lower dimensional data Z_i, typically $S_i(Z_i) = Z_i$. By replacing global summary statistics with local ones, they proved their algorithm EP-ABC was faster than usual ABC because the accept rate might be higher.

For model selection, the situation is complex. Grelaud et al. (2009) used sufficient statistics in model selection between Gibbs random fields. However, Marin et al. (2014) suggested the statistics auxiliary under all candidate models as the best summary statistics in model selection based on Bayesian factor.

Calibration of tolerance levels has also attracted many attention. Biau et al. (2012) viewed the rejection based on metric as a k-nearest neighbor procedure and then calibration on ε is equivalent to calibration of k. Their results favor $k \approx N^{(p+4)/(p+d+4)}$, where N is the number of simulations from model, p is the dimension of the parameters of interest, and $d > 4$ is the dimension of the summary statistics, and under this calibration, they derived rate of convergence of mean square error of density estimation. Ratmann et al. (2013) treated the calibration of ε as a statistical hypothesis testing problem and obtained ε as a critique value in hypothesis testing. The advantage of their method is the MAP estimate is the same under full posterior and ABC posterior and the Kullback–Leibler divergence of the two distributions is small.

1.3.3. Post-Process. Besides modifying sampling scheme and summary statistics, one could also improve inference by carefully processing the output from ABC. Viewing approximation to the posterior as a conditional density estimation problem, Beaumont et al. (2002) applied local linear regression by replacing the simulated raw θ by

$$\theta^* = \theta - (S(z) - S(Y))^T\hat{\beta},$$

where $\hat{\beta}$ is obtained by a weighted least square regression, using weights of the form $K_\varepsilon(\rho(S(z), S(Y)))$. Blum and François (2010) generalized this idea to nonlinear regression with heteroskedasticity estimated by a neural net with one hidden layer. Leuenberger and Wegmann (2010) addressed the same issue using inverse regression.

1.4. Sufficient Dimension Reduction

Sufficient dimension reduction (SDR) is an emerging topic in recent statistical area. As one of the answers to high dimension problems , SDR usually has solid theoretical background to guarantee large sample consistency and valid statistical procedures to select the dimension of results, comparing with machine learning techniques such as manifold learning. The original problem is formulated as follows. Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$, and there is a unknown lower dimensional transformation $S : \mathbb{R}^p \to \mathbb{R}^d$, where $d < p$, the one we need to estimate, which satisfies

$$P(Y \leq y \mid X) = P(Y \leq y \mid S(X)), \forall y \in \mathbb{R}.$$

Most papers in SDR focus on the case where S is a linear transformation, that is $S(X) = \beta^T X$, where $\beta \in \mathbb{R}^{p \times d}$. Note that even in linear SDR, inference differs from transitional parameter estimation, because for any non-singular matrix T, βT is still a valid dimension reduction transformation. As a result, the exact values of each entries in β is not identifiable, but the column space of β is uniquely determined. So the parameters in linear SDR should essentially be in a space called central subspace denoted by $S_{Y|X}$ in Cook (1994) and the optimization problems are in general constrained in the set of subspaces called Grassmann manifold instead of the usual Euclidean space. Sometimes, we are only interested in $E(Y \mid X)$, for example, in linear regression. Then a weak assumption can be made as

$$E(Y \mid X) = E\left(Y \mid \beta^T X\right).$$

The corresponding space of β is called central mean subspace $S_{E(Y|X)}$ in Cook and Li (2002). Yin and Cook (2002) generalized the idea into central k-th moment subspace $S^{(k)}_{Y|X}$ defined as

$$E\left(Y^j \mid X\right) = E\left(Y^j \mid \beta^T X\right), \text{for} j = 1, \ldots, k.$$

To estimate the conditional variance, Zhu and Zhu (2009a) introduced the notion of central variance subspace $S_{\mathrm{Var}(Y|X)}$, defined as

$$\mathrm{Var}(Y \mid X) = E\left(\mathrm{Var}(Y \mid X) \mid \beta^T X\right).$$

To estimate the defined space, statisticians innovated several methods, which can be roughly classified into three categories: inverse regression methods, non-parametric methods, and semi-parametric methods. For simplicity, they usually assume X has zero mean and identity variance-covariance matrix. Lee et al. (2013) formulated the ideas by general measure theory and central σ-field $\mathscr{G}_{Y|X}$ as

$$Y \perp\!\!\!\perp X \mid \mathscr{G}_{Y|X}.$$

For more details, I refer to Ma and Zhu (2013).

1.4.1. Inverse Regression. The first inverse regression method, sliced inverse regression (SIR) proposed by Li (1991), is a precursor of SDR. In that paper, he proved

$$E(X \mid Y) \in S_{Y|X},$$

and used principal component analysis (PCA) to get the main direction of several estimators $\hat{E}(X \mid Y = y_1), \ldots, \hat{E}(X \mid Y = y_s)$ by sliced mean. This paper also built an exemplary approach for other inverse regression based methods. He showed that key quantities, mostly conditional moments, belonged to the corresponding central space, then use PCA to find the main direction of these key quantities.

This paper also invents two standard conditions in linear SDR, named linearity condition

$$E\left(X \mid \beta^T X\right) = L\beta^T X,$$

and constant covariance condition

$$\mathrm{Var}\left(X \mid \beta^T X\right) = Q,$$

where Q is a non-random matrix. The two condition restrict the usage of SIR to nearly normal X. Lately, Dong and Li (2010) relaxed the linearity condition to polynomial condition, that is, $E\left(X \mid \beta^T X\right)$ is a polynomial function of $\beta^T X$. Inspired by SIR, Zhu et al. (1996) proposed kernel inverse regression using kernel technique to estimate the same key quantities in SIR. Wu (2008) generalized this to nonlinear SDR problem. Wang et al. (2014) applied probability integral transformation to SIR in order to solve nonlinear SDR. Li et al. (2011) replaced moments key quantities by a more robust quantile-like quantities defined by a sliced SVM. They proved $\psi(y) \in S_{Y|X}$, if ψ was a solution of a generalized SVM problem,

$$\left(\psi(y), t(y)\right) = \arg\min_{\psi,t} \psi^T \hat{\Sigma}_X \psi + \lambda E_{X,Y}\left(1 - \left(I_{[Y \le y]} - I_{[Y > y]}\right)\left[\psi^T\left(X - \overline{X}\right) - t\right]\right)_+.$$

Their method, called principal support vector machine, can be applied to both linear and kernelized nonlinear SDR problem.As pointed in Cook and Weisberg (1991), SIR fails when there are some symmetric patterns, , so they proposed sliced average variance estimation (SAVE) using the second conditional moments,

$$\mathrm{span}\left(I_p - \mathrm{Var}\left(X \mid Y\right)\right) \subset S_{Y|X}.$$

Zhu et al. (2007) suggested a hybrid of SIR and SAVE by a convex combination. Li and Wang (2007) proposed direction regression (DR) based on

$$\mathrm{span}\left(2I_p - E\left(\left(X - \tilde{X}\right)\left(X - \tilde{X}\right)^T \middle| Y, \tilde{Y}\right)\right) \subset S_{Y|X}.$$

To avoid tuning parameters such as the number of slices and bandwidth of kernel, Zhu et al. (2010a) proposed discretization-expectation procedure. Zhu et al. (2010c) proposed cumulative slicing estimation and Li et al. (2005) proposed contour regression. The above methods usually use non-parametric or semi-parametric method to estimate the key quantities, to take the advantage of explicit likelihood, Cook and Forzani (2009) introduced likelihood acquired directions as MLE of inverse regression.

For central mean subspace, Li and Duan (1989) proved the column space of ordinary least squares is a subspace of $S_{E(Y|X)}$. Li (1992) proposed principal Hessian directions and Cook and Li (2002) proposed iterative Hessian transformations to recover $S_{E(Y|X)}$.

Multivariate response settings are also considered. Three main stream methods are developed. The first is generalizing sliceing into hypercubes defined by different topologies, for instance, Aragon (1997), Hsing (1999), Setodji and Cook (2004).The second is recovering the joint central subspace from marginal central subspaces, for instance, Cook and Setodji (2003), Saracco (2005), Yin and Bura (2006). The last one is projecting multivariate response onto lower dimensional space, for instance, Li et al. (2008), Zhu et al. (2010c).

1.4.2. Non-Parametric Methods. Non-parametric methods do not require the linearity condition or constant covariance condition. And thus these methods are more flexible than inverse regression methods. However, they still rely on continuous X, and hence could not be applied to categorical predictors.

The first non-parametric method is the minimum average variance estimation (MAVE) by Xia et al. (2002), which estimated the central mean subspace by kernel weighted least square subject to Grassmann manifold restriction. The advantage of this method is exhaustiveness, meaning that it would recover the whole $S_{E(Y|X)}$ if d is correctly specified. This method settles the characters of almost all non-parametric methods, that is, smoothing approach to unknown link function m defined as

$$Y = m\left(\beta^T X\right) + \varepsilon.$$

Lately, Xia (2007) proposed density based MAVE to estimate central subspace, basically replacing response Y by kernel smoothing $K_b\left(Y - y\right)$, and Wang and Xia (2008) proposed sliced regression. Hernández and Velilla (2005), Yin and Cook (2005), Yin et al. (2008) replaced the weighted least square by other loss functions.

1.4.3. Semi-Parametric Methods. As far as the authors know, there is only one semi-parametric method by Ma and Zhu (2012) now. In the same way as in sufficient statistics , they decomposed the likelihood functions into two parts, one of which contained only predictors, and the other of which contained response, and the only interest in dimension reduction community was the latter part. Based on this observation, they formulated consistency estimating equations by influence function class defined as

$$\left\{f\left(Y,X\right) - E\left(f\left(Y,X\right) \mid \beta^T X, Y\right) : E\left(f\left(Y,X\right) \mid X\right) = E\left(f\left(Y,X\right) \mid \beta^T X\right), \forall f\right\},$$

particularly, the following forms were used

$$f\left(Y,X\right) = \left(g\left(Y,\beta^T X\right) - E\left(g\left(Y,\beta^T\right) \mid \beta^T X\right)\right)\left(\alpha\left(X\right) - E\left(\alpha\left(X\right) \mid \beta^T X\right)\right),$$

for any g and α. This approach does not require moment conditions like linearity condition or constant covariance condition, or continuous condition of predictors. The authors also shew in their paper several inverse regression methods could be derived by different settings of g and α, and the statistical intuitive of linearity condition and constant covariance condition could also be explained under this framework.

1.4.4. Inference about d. Statistical inference about reduction dimension is a characteristic of sufficient dimension reduction. There are sequential test methods by Schott (1994), Velilla (1998), Bura and Cook (2001), Cook and Yin (2001), Cook et al. (2004), Cook and Ni (2005). Ye and Weiss (2003) and Zhu and Zeng (2006) proposed bootstrap method to select d. BIC-type methods are considered in Zhu et al. (2006), Zhu and Zhu (2007), Luo et al. (2009). A more fashion method is proposed by Zhu et al. (2010b) as sparse eigen-decomposition strategy. They imposed adaptive LASSO penalty to spectral decomposition problem in inverse regression, and the minimization could be solved very effectively by LARS algorithm in Efron et al. (2004).

1.4.5. High Dimension. The laurel of modern statistics should be crowned to inference under increasing number of parameters. To avoid singularity of marginal covariance matrix of X when $p > n$, some authors incorporate partial least squares into inverse regression, such as in Li et al. (2007), Cook et al. (2007), Zhu and Zhu (2009b), Zhu et al. (2010c). Another strategy is to utilize the sparsity principle, such as LASSO, SCAD and Dantzig selector. This leads to sure independence ranking and screening procedure in Zhu et al. (2011). Wu et al. (2008) applied Tikhonov penalties to kernel sliced inverse regression to solve nonlinear SDR.

CHAPTER 2

Higher-Order Properties of Bayesian Empirical Likelihood: Univariate Case

2.1. Introduction

Empirical likelihood, over the years, has become a very popular topic of statistical research. The name was coined by Owen in his classic 1986 paper, although similar ideas are found even earlier in the works of Hartley and Rao (1968), Thomas and Grunkemeier (1975), Rubin et al. (1981) and others. The main advantage of empirical likelihood is that it involves fewer assumptions than a regular likelihood, and yet shares the same asymptotic properties of the latter.

Research in this area has primarily been frequentist with a long list of important theoretical developments accompanied by a large number of applications. To our knowledge, the first Bayesian work in this general area appeared in the article of Lazar (2003) followed by some related work in Schennach (2005), Schennach et al. (2007) , the latter introducing the concept of "exponentially tilted empirical likelihood". Lazar (2003) suggested using empirical likelihood as a substitute for the usual likelihood and carrying out Bayesian analysis in the usual way.

Baggerly (1998) Baggerly (1998) viewed empirical likelihood as a method of assigning probabilities to a n-cell contingency table in order to minimize a goodness-of-fit criterion. He selected Cressie–Read power divergence statistics as one such criterion for construction of confidence regions in a number of situations and pointed out also how the usual empirical likelihood, exponentially tilted empirical likelihood and others could be viewed as special cases of the Cressie–Read criterion by appropriate choice of the the power parameter. This was also discussed in Owen (2010) who pointed out that all members of the Cressie–Read family led to "empirical divergence analogues of the empirical likelihood in which asymptotic χ^2 calibration held for the mean".

The objective of this article is to provide an asymptotic expansion of the posterior distribution based on empirical likelihood and its variations under certain regularity conditions and a mean constraint. The work is inspired by the work of Fang and Mukerjee (2006) who provided a somewhat different expansion subject to a mean constraint. Unlike Fang and Mukerjee (2005, 2006), our result is based on the derivatives of the pseudo likelihood with respect to the parameter of interest evaluated at the maximum empirical likelihood estimator, and a rigorous expansion is provided with particular attention to the remainder terms. Moreover, we consider a general estimating equation which includes the mean example of Fang and Mukerjee (2006) as a special case. The need for different pseudo-likelihoods for statistical inference is felt all the more in these days, especially for the analysis of high-dimensional data, where the usual likelihood based analysis is hard to perform, These alternative likelihoods are equally valuable for approximate Bayesian

computations , a topic which has only recently surfaced in the statistics literature (see e.g. **?**)

Asymptotic expansion of the posterior based on a regular likelihood was given earlier in Johnson (1970), and later in Ghosh et al. (1982). We follow their approach with many necessary modifications in view of the fact that any meaningful prior needs to have support in a data-driven compact set which grows with number of observations. As a special case of our result, we get the celebrated Bernstein–von Mises theorem. The latter was mentioned in Lazar (2003) for the special case of empirical likelihood, but here we provide a rigorous derivation with the needed regularity conditions in a general framework. The asymptotic expansion can also be used in providing asymptotic expansions of the posterior moments, quantiles and other quantities of interest. Moreover, we utilize this asymptotic expansion to find some moment matching priors, earlier given in Ghosh and Liu (2011) based on the regular likelihood. In constrast to Ghosh and Liu (2011), the moment matching prior does not depend on the expectation of the derivatives of the log-likelihood function, but depends instead on the second and third central moments of the unbiased estimating function, say $g(X,\theta)$. In the particular case, $g(X,\theta) = X - \theta$, the prior depends only on knowledge about the second and third central moments of the distribution, and does not require specification of a full likelihood. The moment matching priors differ also from the reference priors as introduced in **?**. The latter is an analogue of Jeffreys' prior under most circumstances, with the Godambe information matrix (**?**) replacing the Fisher information matrix.

2.2. Basic Settings

Suppose $X_1, \ldots, X_n$ are independent and identically distributed random vectors satisfying $E\{g(X_1,\theta)\} = 0$, where $\theta \in \mathbb{R}$. In this context, Owen (1988), formulated empirical likelihood as a nonparametric likelihood of the form $\prod_{i=1}^{n} w_i(\theta)$, where w_i is the probability mass assigned to $X_i\,(i = 1, \ldots, n)$ satisfying the constraints

$$\begin{cases} w_i > 0, \text{foralli}; \\ \sum_{i=1}^{n} w_i = 1; \\ \sum_{i=1}^{n} w_i g(X_i,\theta) = 0. \end{cases} \tag{2.2.1}$$

The target is to maximize $\prod_{i=1}^{n} w_i$ or equivalently $\sum_{i=1}^{n} \log w_i$ with respect to $w_1, \ldots, w_n$ subject to the constraints given in Eq. (3.2.1). Applying the Lagrange multiplier method, the solution turns out to be

$$\hat{w}_i^{\mathrm{EL}}(\theta) = \frac{1}{n\{1 + \nu g(X_i,\theta)\}}, i = 1, \ldots, n, \tag{2.2.2}$$

where ν, the Lagrange multiplier satisfies

$$\sum_{i=1}^{n} \frac{g(X_i,\theta)}{1 + \nu g(X_i,\theta)} = 0. \tag{2.2.3}$$

It may be noted that in Fang and Mukerjee (2005, 2006), $g(X_i,\theta) = X_i - \theta, i = 1, \ldots, n$.

Closely related to the empirical likelihood is the exponentially tilted empirical likelihood where the objective is to maximize the Shannon entropy $-\sum_{i=1}^{n} w_i \log w_i$

with the same constraints in Eq. (3.2.1). The resulting solution is

$$\hat{w}_i^{\mathrm{ET}}(\theta) = \frac{\exp\{-\nu g(X_i,\theta)\}}{\sum_{j=1}^n \exp\{-\nu g(X_i,\theta)\}},$$

where ν, the Lagrange multiplier, satisfies

$$\sum_{i=1}^n \exp\{-\nu g(X_i,\theta)\} g(X_i,\theta) = 0. \tag{2.2.4}$$

The exponentially tilted empirical likelihood is related to Kullback-Leibler divergence between two empirical distributions, one with weights w_i assigned to the n sample points, and the other with uniform weights $1/n$ assigned to the sample points.

The general Cressie–Read divergence criterion given by

$$\mathrm{CR}(\lambda) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^n \left\{ (nw_i)^{-\lambda} - 1 \right\}.$$

We focus on the cases $\lambda \geq 0$ and $\lambda \leq -1$, because in these cases, $\mathrm{CR}(\lambda)$ is a convex function of the $w_i\,(i = 1,\ldots,n)$, and hence the minimization problem will produce a unique solution. The limiting cases $\lambda \to 0$ and $\lambda \to -1$ correspond to the usual empirical likelihood and the exponentially tilted empirical likelihood as defined earlier.

For convex $\mathrm{CR}(\lambda)$, its minimum will be attained in the compact set H_n determined by data. The Lagrange multiplication method now gives the weights

$$\hat{w}_i^{\mathrm{CR}}(\theta) = \frac{1}{n}\{\mu + \nu g(X_i,\theta)\}^{-1/(\lambda+1)},\ i = 1,\ldots,n, \tag{2.2.5}$$

where we abbreviate $\mu(\theta)$ as μ and $\nu(\theta)$ as ν, which satisfy

$$\begin{cases} \sum_{i=1}^n \{\mu + \nu g(X_i,\theta)\}^{-1/(\lambda+1)} = n, \\ \sum_{i=1}^n \{\mu + \nu g(X_i,\theta)\}^{-1/(\lambda+1)} X_i = 0. \end{cases} \tag{2.2.6}$$

We now introduce the posterior based on an empirical likelihood. The basic idea was first introduced by Lazar (2003) with several numerical examples. The intuition relies on close relationship between the empirical likelihood and the empirical distribution. Owen (2010) formulated the two concepts under the same optimization framework, that is, they shared the same objective function, but the former was solved under parametric constraints, while the latter was not. Considering this similarity, we can use the empirical likelihood as a valid distribution parameterized by some inferential target. Within the Bayesian paradigm, writing $\hat{w}_i(\theta)$ as generic notation for either $\hat{w}_i^{\mathrm{EL}}$, $\hat{w}_i^{\mathrm{ET}}$ or $\hat{w}_i^{\mathrm{CR}}$, and a prior with probability density function $\rho(\theta)$, with support in H_n, the profile (pseudo) posterior is

$$\pi(\theta \mid X_1,\ldots,X_n) = \frac{\prod_{i=1}^n \hat{w}_i(\theta)\,\rho(\theta)}{\int_{H_n} \prod_{i=1}^n \hat{w}_i(\theta)\,\rho(\theta)\,\mathrm{d}\theta}. \tag{2.2.7}$$

The main objective of this paper is to provide an asymptotic expansion of $\pi(\theta \mid X_1,\ldots,X_n)$. This will include in particular the Bernstein–von Mises theorem. Towards our main result, we develop a few necessary lemmas in the next section. Some of these lemmas are also of independent interest as they point out some interesting features pertaining to empirical likelihood.

2.3. Lemmas

We first point out the natural domain of θ in empirical likelihood settings. In practice, some values of θ will result in an empty feasible set under constraints Eq. (3.2.1). The set of θ values which guarantees a non-empty feasible set, and thus a solution of the optimization problem, constitutes a natural domain of the empirical likelihood. One may question whether the size of the natural domain is large enough to contain the true value. The following lemma alleviates this worry.

LEMMA 1. Assume $g(\cdot,\cdot)$ is a continuous function, then the natural domain defined by the constraints Eq. (3.2.1) is a compact set and is nondecreasing with respect to the sample size n.

PROOF. By the third constraint of Eq. (3.2.1), θ is a continuous function of $w_1, w_2, \ldots, w_n$, but w_i are defined on a simplex which is a compact set due to the first constraint of Eq. (3.2.1). We may recall that a continuous function maps compact sets to compact sets. Hence, θ is naturally defined on a compact set denoted by H_n.

If all the $g(X_i,\theta), i = 1,\ldots,n$ are non-positive or all are non-negative, then the constraints Eq. (3.2.1) are violated and $H_n = \emptyset$. Hence, we define the domain as

$$\begin{aligned} H_n &= \left(\left[\bigcap_{i=1}^{n}\{g(X_i,\theta)\geq 0\}\right]\bigcup\left[\bigcap_{i=1}^{n}\{g(X_i,\theta)\leq 0\}\right]\right)^c \\ &= \left[\bigcup_{i=1}^{n}\{g(X_i,\theta)\geq 0\}^c\right]\bigcap\left[\bigcup_{i=1}^{n}\{g(X_i,\theta)\leq 0\}^c\right]. \end{aligned}$$

As n increases, both $\{[\bigcup_{i=1}^{n}\{g(X_i,\theta)\geq 0\}^c]\}$ and $\{[\bigcup_{i=1}^{n}\{g(X_i,\theta)\leq 0\}^c]\}$ will increase, and so will their intersection H_n. □

Although, intuitively we expect the empirical likelihood to behave as the true likelihood, we need some theoretical support to show that the former enjoys some of the basic properties of the latter. In particular, we need to verify that ν and μ are smooth functions of θ and the (pseudo) Fisher Information based on the empirical likelihood is positive.

We first establish the positiveness of the Fisher information . We consider the three cases separately to introduce more transparency and continuity in our approach.

Our first lemma shows that the Lagrange multipliers $\nu(\theta)$ and $\mu(\theta)$ are both smooth functions of θ, under the following mild assumptions,

ASSUMPTION 1. *For any θ in natural domain H_n, and $n \geq 3$,*

$$\operatorname{pr}\{g(X_i,\theta) = 0, i = 1,\ldots,n\} = 0.$$

And

ASSUMPTION 2. *$g(x,\theta)$ is a continuous multivariate function with continuous derivatives in θ.*

LEMMA 2. Under Assumptions 1 and 2, for the empirical likelihood, exponentially tilted empirical likelihood and Cressie–Read empirical likelihood(λ), the Lagrange multipliers $\nu(\theta)$ and $\mu(\theta)$ are smooth functions of θ.

PROOF. We first consider the empirical likelihood and observe that, $\nu(\theta)$ is a implicit function of θ in view of (3.2.3) . Further

$$\frac{\partial}{\partial \nu}\sum_{i=1}^{n}\frac{g(X_i,\theta)}{1+\nu g(X_i,\theta)}=-\sum_{i=1}^{n}\frac{g^2(X_i,\theta)}{\{1+\nu g(X_i,\theta)\}^2}<0,$$

so that by the implicit function theorem, ν is differentiable in θ. Moreover, differentiating both sides of Eq. (3.2.3) with respect to θ, one gets

$$\begin{aligned}0 &= \sum_{i=1}^{n}\frac{1}{1+\nu g(X_i,\theta)}\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}-\sum_{i=1}^{n}\frac{\nu g(X_i,\theta)}{\{1+\nu g(X_i,\theta)\}^2}\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}\\ &\quad -\sum_{i=1}^{n}\frac{g^2(X_i,\theta)}{\{1+\nu g(X_i,\theta)\}^2}\frac{\mathrm{d}\nu}{\mathrm{d}\theta},\end{aligned}$$

which on simplification leads to

$$\frac{\mathrm{d}\nu}{\mathrm{d}\theta}=-\frac{\sum_{i=1}^{n}\{1+\nu g(X_i,\theta)\}^{-2}\,\mathrm{d}g(X_i,\theta)/\mathrm{d}\theta}{\sum_{i=1}^{n}\{1+\nu g(X_i,\theta)\}^{-2}g^2(X_i,\theta)},\tag{2.3.1}$$

Next, for exponentially tilted empirical likelihood, in view of Eq. (3.2.5) and the relation

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\nu}\left[\sum_{i=1}^{n}\exp\{-\nu g(X_i,\theta)\}g(X_i,\theta)\right]\\ =\;&-\sum_{i=1}^{n}\exp\{-\nu g(X_i,\theta)\}g^2(X_i,\theta)<0,\end{aligned}$$

once again, the implicit function theorem guarantees the differentiability of ν in θ. Further, differentiating both sides of Eq. (3.2.5) with respect to θ, one gets

$$\frac{\mathrm{d}\nu}{\mathrm{d}\theta}=\frac{\sum_{i=1}^{n}\exp\{-\nu(\theta)g(X_i,\theta)\}\{\mathrm{d}g(X_i,\theta)/\mathrm{d}\theta\}\{1-\nu g(X_i,\theta)\}}{\sum_{i=1}^{n}\exp\{-\nu(\theta)g(X_i,\theta)\}g^2(X_i,\theta)}.\tag{2.3.2}$$

A similar conclusion is achieved for $\nu(\theta)$ and $\mu(\theta)$ defined in Eq. (3.2.7) in connection with CR(λ). Specifically, defining

$$\begin{cases}F_1=\sum_{i=1}^{n}\{\mu+\nu g(X_i,\theta)\}^{-1/(\lambda+1)}-n,\\ F_2=\sum_{i=1}^{n}\{\mu+\nu g(X_i,\theta)\}^{-1/(\lambda+1)}g(X_i,\theta),\end{cases}$$

it follows that,

$$\frac{\partial(F_1,F_2)}{\partial(\mu,\nu)}=-\frac{1}{\lambda+1}\begin{pmatrix}\sum_{i=1}^{n}q_i & \sum_{i=1}^{n}q_ig(X_i,\theta)\\ \sum_{i=1}^{n}q_ig(X_i,\theta) & \sum_{i=1}^{n}q_ig(X_i,\theta)^2\end{pmatrix},$$

where $q_i=\{\mu+\nu g(X_i,\theta)\}^{-1/(\lambda+1)-1}$. Then the determinant of Jacobian is

$$\begin{aligned}\det\frac{\partial(F_1,F_2)}{\partial(\mu,\nu)} &= \left(\frac{1}{\lambda+1}\right)^2\left[\sum_{i=1}^{n}q_i\sum_{i=1}^{n}q_ig(X_i,\theta)^2-\left\{\sum_{i=1}^{n}q_ig(X_i,\theta)\right\}^2\right]\\ &= \left(\frac{1}{\lambda+1}\right)^2\left(\sum_{i=1}^{n}q_i\right)^2\left[\sum_{i=1}^{n}\frac{q_i}{\sum_{j=1}^{n}q_j}g(X_i,\theta)^2-\left\{\sum_{i=1}^{n}\frac{q_i}{\sum_{j=1}^{n}q_j}g(X_i,\theta)\right\}\right]^2>0.\end{aligned}$$

Again, by implicit function theorem, one gets differentiability of $\mu(\theta)$ and $\nu(\theta)$with respect to θ, and

$$\begin{aligned}
\begin{pmatrix} \mu' \\ \nu' \end{pmatrix} &= \left(\frac{\partial(F_1,F_2)}{\partial(\mu,\nu)}\right)^{-1}\begin{pmatrix} \partial F_1/\partial\theta \\ \partial F_2/\partial\theta \end{pmatrix} \\
&= \left(-\frac{1}{\lambda+1}\right)(\lambda+1)^2\frac{1}{\sum_{i=1}^n q_i\sum_{i=1}^n q_i g(X_i,\theta)^2-\{\sum_{i=1}^n q_i g(X_i,\theta)\}^2} \\
&\times\begin{pmatrix} \sum_{i=1}^n q_i g(X_i,\theta)^2 & -\sum_{i=1}^n q_i g(X_i,\theta) \\ -\sum_{i=1}^n q_i g(X_i,\theta) & \sum_{i=1}^n q_i \end{pmatrix} \\
&\times\Bigg((\lambda+1)^{-1}\sum_{i=1}^n q_i\nu\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta},-(\lambda+1)\sum_{i=1}^n q_i\nu g(X_i,\theta)\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta} \\
&+\sum_{i=1}^n\{\mu+\nu g(X_i,\theta)\}^{-1/(\lambda+1)}\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}\Bigg)^T. \qquad (2.3.3)
\end{aligned}$$

□

The next result shows that all the derivatives of the Lagrange multipliers $\nu(\theta)$ and $\mu(\theta)$ are smooth functions of $\theta \in H_n$. We provide a unified proof for all three cases where we utilize the previous lemma. with an assumption stronger than Assumption 2,

ASSUMPTION 3. *$g(x,\theta)$ is a multivariate continuous function and $(K+4)$th-order differentiable in θ.*

LEMMA 3. Under Assumptions 1 and 3, all derivatives of $\nu(\theta)$ and $\mu(\theta)$ are smooth functions of θ for $\theta \in H_n$.

PROOF. The result is proved by induction. We have seen already in Lemma 2, that the first derivatives of $\nu'(\theta)$ and $\mu'(\theta)$ are smooth functions of θ. Suppose the result holds for all kth derivatives of $\nu(\theta)$ and $\mu(\theta)$ for $k=1,\ldots,K$. Then writing

$$\frac{\mathrm{d}^k\nu}{\mathrm{d}\theta^k} = h_k\{\nu(\theta),\theta\}, k=1,\ldots,K,$$

$$\frac{\mathrm{d}^{k+1}\nu}{\mathrm{d}^{k+1}\theta} = \frac{\partial h_k}{\partial\nu}\frac{\mathrm{d}\nu}{\mathrm{d}\theta}+\frac{\partial h_k}{\partial\theta}$$

which is also a smooth function of θ by the induction hypothesis and Lemma 1. A similar proof works for $\mu(\theta)$. □

We know that when the number of constraints and dimension of the parameters are the same, the corresponding empirical likelihood is maximized at $\theta=\tilde{\theta}$, the M-estimator of θ based on $\sum_{i=1}^n g(X_i,\theta)=0$. Thus, $\nu\left(\tilde{\theta}\right)=0$ and $\mu\left(\tilde{\theta}\right)=1$. We next show that $\tilde{l}(\theta)$ has a negative second order derivative when evaluated at $\tilde{\theta}$.

LEMMA 4. Under Assumptions 1 and 2, $\mathrm{d}^2\tilde{l}\left(\tilde{\theta}\right)/\mathrm{d}\theta^2 < 0$ where $\tilde{l}(\theta) = n^{-1}\sum_{i=1}^n\log\hat{w}_i(\theta)$ where $\hat{w}_i$ is either $\hat{w}_i^{\mathrm{EL}}$, $\hat{w}_i^{\mathrm{ET}}$ or $\hat{w}_i^{\mathrm{CR}}$ $(i=1,\ldots,n)$.

PROOF. We begin with $\tilde{l}(\theta) = n^{-1}\sum_{i=1}^{n}\log\hat{w}_i^{\mathrm{EL}}(\theta) = -\sum_{i=1}^{n}\log\{1+\nu(X_i-\theta)\} - \log n$. Hence by Eq. (3.2.2), Eq. (3.2.3) and Eq. (3.3.1),

$$\begin{aligned}\frac{\mathrm{d}\tilde{l}(\theta)}{\mathrm{d}\theta} &= \frac{1}{n}\nu(\theta)\sum_{i=1}^{n}\frac{1}{1+\nu g(X_i\theta)}\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta} - \frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i\theta)}{1+\nu g(X_i\theta)}\frac{\mathrm{d}\nu}{\mathrm{d}\theta}\\ &= \frac{1}{n}\nu(\theta)\sum_{i=1}^{n}\frac{1}{1+\nu g(X_i\theta)}\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}.\end{aligned}$$

Thus

$$\left.\frac{\mathrm{d}^2\tilde{l}(\theta)}{\mathrm{d}\theta^2}\right|_{\theta=\tilde{\theta}} = -\frac{\left\{\sum_{i=1}^{n}\mathrm{d}g\left(X_i,\tilde{\theta}\right)/\mathrm{d}\theta\right\}^2}{n\sum_{i=1}^{n}g^2\left(X_i,\tilde{\theta}\right)} < 0.$$

Next we consider $\tilde{l}(\theta) = n^{-1}\sum_{i=1}^{n}\log\hat{w}_i^{\mathrm{ET}}(\theta) = -\nu n^{-1}\sum_{i=1}^{n}g(X_i,\theta) - \log\sum_{i=1}^{n}\exp\{-\nu g(X_i,\theta)\}$. Then

$$\begin{aligned}\frac{\mathrm{d}\tilde{l}(\theta)}{\mathrm{d}\theta} &= -\frac{\mathrm{d}\nu}{\mathrm{d}\theta}\frac{1}{n}\sum_{i=1}^{n}g(X_i,\theta)\\ &\quad + \frac{\sum_{i=1}^{n}\exp\{-\nu g(X_i,\theta)\}\{(\mathrm{d}\nu/\mathrm{d}\theta)g(X_i,\theta)+\nu\,\mathrm{d}g(X_i,\theta)/\mathrm{d}\theta\}}{\sum_{i=1}^{n}\exp\{-\nu g(X_i,\theta)\}}\\ &= -\frac{\mathrm{d}\nu}{\mathrm{d}\theta}\frac{1}{n}\sum_{i=1}^{n}g(X_i,\theta) + \nu\frac{\sum_{i=1}^{n}\exp\{-\nu g(X_i,\theta)\}\,\mathrm{d}g(X_i,\theta)/\mathrm{d}\theta}{\sum_{i=1}^{n}\exp\{-\nu g(X_i,\theta)\}}\\ &\quad -\nu n^{-1}\sum_{i=1}^{n}\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}.\end{aligned}$$

Thus, by Eq. (3.3.1)

$$\left.\frac{\mathrm{d}^2\tilde{l}(\theta)}{\mathrm{d}\theta^2}\right|_{\theta=\tilde{\theta}} = -\frac{\mathrm{d}\nu}{\mathrm{d}\theta}\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}g\left(X_i,\tilde{\theta}\right)}{\mathrm{d}\theta} = -\frac{\left\{\sum_{i=1}^{n}\mathrm{d}g\left(X_i,\tilde{\theta}\right)/\mathrm{d}\theta\right\}^2}{n\sum_{i=1}^{n}g^2\left(X_i,\tilde{\theta}\right)} < 0.$$

Finally, for the Cressie–Read case, $\tilde{l}(\theta) = n^{-1}\sum_{i=1}^{n}\log\hat{w}_i^{\mathrm{CR}}(\theta) = -\{n(\lambda+1)\}^{-1}\sum_{i=1}^{n}\log\{\mu+\nu g(X_i,\theta)\}$. Then by Eq. (3.3.4),

$$\left.\frac{\mathrm{d}^2\tilde{l}(\theta)}{\mathrm{d}\theta^2}\right|_{\theta=\tilde{\theta}} = -\frac{\left\{\sum_{i=1}^{n}\mathrm{d}g\left(X_i,\tilde{\theta}\right)/\mathrm{d}\theta\right\}^2}{n\sum_{i=1}^{n}g^2\left(X_i,\tilde{\theta}\right)}.$$

□

Let $b = \left[\left\{n^{-1}\sum_{i=1}^{n}\mathrm{d}g\left(X_i,\tilde{\theta}\right)/\mathrm{d}\theta\right\}^2 / \left\{n^{-1}\sum_{i=1}^{n}g\left(X_i,\tilde{\theta}\right)^2\right\}\right]^{-1/2}$. The main result is proved in the next section.

2.4. Main Result

Before stating the main theorem, we need a few notations. We assume that the prior density $\rho(\theta)$ has a Kth-order continuous derivative at $\tilde{\theta}$. Let $\rho_K(\theta) = \rho\left(\tilde{\theta}\right) + \rho'\left(\tilde{\theta}\right)\left(\theta-\tilde{\theta}\right) + \cdots + \rho^{(K)}\left(\tilde{\theta}\right)\left(\theta-\tilde{\theta}\right)$ the Kth-order Taylor approximation

of the prior density. Further denote the higher-order derivatives of (pseudo) log empirical likelihood as

$$a_{kn}(\theta)=\frac{1}{k!}\sum_{i=1}^{n}\frac{\mathrm{d}^k\tilde{l}(\theta)}{\mathrm{d}\theta^k},k=3,\ldots,K+3.$$

Define the summation index set

$$I_{i,k}=\left\{(m_{3,i},\ldots,m_{K+3,i})\in\mathbb{N}^K:\sum_{u=3}^{K+3}m_{u,i}=i,\sum\nolimits_{u=3}^{K+3}m_{u,i}(u-2)=k\right\}.$$

Let $y=\sqrt{n}b\left(\theta-\tilde{\theta}\right)$ be the normalized posterior random variable and

$$\begin{aligned}\alpha_k(y,n) &= \frac{1}{k!}\rho^{(k)}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k+\sum_{j=0}^{k-1}\frac{1}{j!}\rho^{(j)}\left(\tilde{\theta}\right)\\ &\quad\times\sum_{i=\lceil(k-j)/(K+1)\rceil}^{k-j}\frac{1}{i!}\sum_{I_{i,k-j}}\binom{i}{m_{3,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left\{a_{un}\left(\tilde{\theta}\right)\right\}^{m_{u,i}}\left(\frac{y}{b}\right)^{k+2i+j},\end{aligned}$$

where, $k=0,\ldots,K$. For special cases $k=0,1$, we have $\alpha_0(y,n)=\rho\left(\tilde{\theta}\right)$ and $\alpha_1(y,n)=\rho'\left(\tilde{\theta}\right)y/b+\rho\left(\tilde{\theta}\right)a_{3n}\left(\tilde{\theta}\right)(y/b)^3$. Now define $Y_{(1)}=\sqrt{n}b\left(h_1-\tilde{\theta}\right)$ and $Y_{(n)}=\sqrt{n}b\left(h_2-\tilde{\theta}\right)$ as the normalized lower and upper bounds of the support of the distribution. Now for any $\xi\in\left(Y_{(1)},Y_{(n)}\right)$ and $H_n=[h_1,h_2]$, let

$$P_K(\xi,n)=\sum_{k=0}^{K}\left\{\int_{Y_{(1)}}^{\xi}\alpha_k(y,n)\exp\left(-\frac{y^2}{2}\right)\mathrm{d}y\right\}n^{-k/2}.$$

To control the higher-order error terms, we need the following assumptions.

Assumption 4. *For any* $(l_1,\ldots,l_j)\subset\{2,\ldots,K+3\}$,

$$E\left\{\prod_{i=1}^{j}\frac{\mathrm{d}^{l_i}g(X_1,\theta)}{\mathrm{d}\theta^{l_i}}\right\}<\infty.$$

Moreover, we also need an assumption to guarantee the consistency of M-Estimator,

Assumption 5. $g(\cdot,\theta)$ *is either bounded or monotone in* θ.

Now we state the main theorem of this section.

Theorem 1 (Fundamental Theorem for Expansion). Let $X_1,X_2,\ldots,X_n$ be independent and identically distributed. Assume the prior density $\rho(\theta)$ has a support containing H_n and has $(K+1)$th-order continuous derivative. Under Assumptions 1 , 3 ,4 and 5, there exist constants $N_1>0$ and $M_1>0$, such that

(2.4.1)

$$\left|\int_{Y_{(1)}}^{\xi}\prod_{i=1}^{n}\hat{w}_i\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)\mathrm{d}y-P_K(\xi,n)\right|\leq M_1n^{-(K+1)/2},\text{ a.s.}$$

for any $n>N_1$ and $\xi\in\left(Y_{(1)},Y_{(n)}\right)$.

Proof. See Appendix A.4. □

This theorem can not only be used to prove asymptotic expansion of the posterior cumulative distribution function , the main result of this paper, but it can also be used to find the asymptotic expansions of the posterior mean, quantiles and many other quantities of interest, as in Johnson (1970) and Vexler et al. (2014) .

Next we write the posterior cumulative distribution function as

$$\Pi\left(\theta\leq\tilde{\theta}+\frac{\xi}{\sqrt{n}b}\middle|X_1,\ldots,X_n\right)=\frac{\int_{h_1}^{\tilde{\theta}+\xi/\sqrt{n}b}\prod_{i=1}^{n}\hat{w}_i\left(\theta\right)\rho\left(\theta\right)\,\mathrm{d}\theta}{\int_{h_1}^{h_2}\prod_{i=1}^{n}\hat{w}_i\left(\theta\right)\rho\left(\theta\right)\,\mathrm{d}\theta}.$$

Moreover, let $R_n=\left(Y_{(1)},Y_{(n)}\right)$ and

$$\Phi\left(\xi\mid R_n\right)=\frac{\int_{Y_{(1)}}^{\xi}\varphi\left(y\right)\,\mathrm{d}y}{\int_{Y_{(1)}}^{Y_{(n)}}\varphi\left(y\right)\,\mathrm{d}y},$$

where $\varphi\left(y\right)$ is standard normal density, be restricted to R_n. Define polynomial $\gamma_i\left(\xi,n\right),i=1,\ldots,n$ recursively as

$$\int_{Y_{(1)}}^{\xi}\alpha_k\left(y,n\right)\exp\left(-\frac{y^2}{2}\right)\,\mathrm{d}y=\sum_{j=0}^{k}\left\{\int_{Y_{(1)}}^{Y_{(n)}}\alpha_j\left(y,n\right)\exp\left(-\frac{y^2}{2}\right)\,\mathrm{d}y\right\}\gamma_{k-j}\left(\xi,n\right).$$

The first two terms of $\gamma_i\left(\xi,n\right)$ are

$$\gamma_0\left(\xi,n\right)=\frac{\rho\left(\tilde{\theta}\right)\int_{Y_{(1)}}^{\xi}\exp\left(-y^2/2\right)\,\mathrm{d}y}{\rho\left(\tilde{\theta}\right)\int_{Y_{(1)}}^{Y_{(n)}}\exp\left(-y^2/2\right)\,\mathrm{d}y}=\frac{\Phi\left(\xi\right)-\Phi\left(Y_{(1)}\right)}{\Phi\left(Y_{(n)}\right)-\Phi\left(Y_{(1)}\right)}-=\Phi\left(\xi\mid R_n\right),$$

and

$$\begin{aligned}\gamma_1\left(\xi,n\right) &= \frac{\int_{Y_{(1)}}^{\xi}\exp\left(-y^2/2\right)\left\{\rho'\left(\tilde{\theta}\right)y/b+\rho\left(\tilde{\theta}\right)a_{3n}\left(\tilde{\theta}\right)\left(y/b\right)^3\right\}\,\mathrm{d}y}{\rho\left(\tilde{\theta}\right)\int_{Y_{(1)}}^{Y_{(n)}}\exp\left(-y^2/2\right)\,\mathrm{d}y}\\ &\quad-\frac{\int_{Y_{(1)}}^{Y_{(n)}}\exp\left(-y^2/2\right)\left\{\rho'\left(\tilde{\theta}\right)y/b+\rho\left(\tilde{\theta}\right)a_{3n}\left(\tilde{\theta}\right)\left(y/b\right)^3\right\}\,\mathrm{d}y}{\rho\left(\tilde{\theta}\right)\int_{Y_{(1)}}^{Y_{(n)}}\exp\left(-y^2/2\right)\,\mathrm{d}y}\Phi\left(\xi\mid R_n\right)\\ &= \left\{\frac{\rho'\left(\tilde{\theta}\right)}{b\rho\left(\tilde{\theta}\right)}\right\}\frac{\varphi\left(Y_{(1)}\right)-\varphi\left(\xi\right)-\Phi\left(\xi\mid R_n\right)\left\{\varphi\left(Y_{(1)}\right)-\varphi\left(Y_{(n)}\right)\right\}}{\int_{Y_{(1)}}^{Y_{(n)}}\varphi\left(y\right)\,\mathrm{d}y}\\ &\quad+\left\{\frac{a_{3n}\left(\tilde{\theta}\right)}{b^3}\right\}\left\{\frac{Y_{(1)}^2\varphi\left(Y_{(1)}\right)+2\varphi\left(Y_{(1)}\right)-\xi^2\varphi\left(\xi\right)-2\varphi\left(\xi\right)}{\int_{Y_{(1)}}^{Y_{(n)}}\varphi\left(y\right)\,\mathrm{d}y}\right.\\ &\quad\left.-\Phi\left(\xi\mid R_n\right)\frac{Y_{(1)}^2\varphi\left(Y_{(1)}\right)+2\varphi\left(Y_{(1)}\right)-Y_{(n)}^2\varphi\left(Y_{(n)}\right)-2\varphi\left(Y_{(n)}\right)}{\int_{Y_{(1)}}^{Y_{(n)}}\varphi\left(y\right)\,\mathrm{d}y}\right\}.\end{aligned}$$

We now provide the next important result result of this section, namely the asymptotic expansion of the posterior distribution function.

THEOREM 2 (Asymptotic Expansion of the Posterior Cumulative Distribution Function). Use the same assumptions as in Theorem 1. Then there exist constants

N_2 and M_2, such that

(2.4.2)
$$\left|\Pi\left(\theta \leq \tilde{\theta} + \frac{\xi}{\sqrt{n}b}\middle| X_1, \ldots, X_n\right) - \Phi\left(\xi \mid R_n\right) - \sum_{i=1}^{K} \gamma_i\left(\xi, n\right) n^{-i/2}\right| \leq M_2 n^{-(K+1)/2}, \text{ a.s.},$$

for any $n \geq N_2$ and $\xi \in \left(Y_{(1)}, Y_{(n)}\right)$.

Proof. By Theorem 1, we have

$$\left|\int_{Y_{(1)}}^{\xi} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + \frac{y}{\sqrt{n}b}\right) \rho\left(\tilde{\theta} + \frac{y}{\sqrt{n}b}\right) \mathrm{d}y - P_K\left(\xi, n\right)\right| \leq M_1 n^{-(K+1)/2}, \tag{2.4.3}$$

and

(2.4.4)
$$\left|\int_{Y_{(1)}}^{Y_{(n)}} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + \frac{y}{\sqrt{n}b}\right) \rho\left(\tilde{\theta} + \frac{y}{\sqrt{n}b}\right) \mathrm{d}y - P_K\left(Y_{(n)}, n\right)\right| \leq M_1 n^{-(K+1)/2}.$$

By definition

$$\Pi\left(\theta \leq \tilde{\theta} + \frac{\xi}{\sqrt{n}b} \mid X_1, X_2, \ldots, X_n\right) = \frac{\int_{Y_{(1)}}^{\xi} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + y/\sqrt{n}b\right) \rho\left(\tilde{\theta} + y/\sqrt{n}b\right) \mathrm{d}y}{\int_{Y_{(1)}}^{Y_{(n)}} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + y/\sqrt{n}b\right) \rho\left(\tilde{\theta} + y/\sqrt{n}b\right) \mathrm{d}y}.$$

We know that all the terms in Eq. (3.7.2) and Eq. (3.7.3), are integrals of continuous functions over bounded closed intervals. Hence they are almost surely bounded below by some constant C_1 and bounded above by some constant C_2, for all $n > N_1$. Then

$$\begin{aligned}
&\left|\Pi\left(\theta \leq \tilde{\theta} + \frac{\xi}{\sqrt{n}b} \mid X_1, X_2, \ldots, X_n\right) - \frac{P_K\left(\xi, n\right)}{P_K\left(Y_{(n)}, n\right)}\right| \\
= &\left|\frac{\int_{Y_{(1)}}^{\xi} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + y/\sqrt{n}b\right) \rho\left(\tilde{\theta} + y/\sqrt{n}b\right) \mathrm{d}y}{\int_{Y_{(1)}}^{Y_{(n)}} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + y/\sqrt{n}b\right) \rho\left(\tilde{\theta} + y/\sqrt{n}b\right) \mathrm{d}y} - \frac{P_K\left(\xi, n\right)}{P_K\left(Y_{(n)}, n\right)}\right| \\
= &\left|\frac{\int_{Y_{(1)}}^{\xi} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + y/\sqrt{n}b\right) \rho\left(\tilde{\theta} + y/\sqrt{n}b\right) \mathrm{d}y - P_K\left(\xi, n\right)}{\int_{Y_{(1)}}^{Y_{(n)}} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + y/\sqrt{n}b\right) \rho\left(\tilde{\theta} + y/\sqrt{n}b\right) \mathrm{d}y}\right. \\
&\left. - \frac{\left\{\int_{Y_{(1)}}^{Y_{(n)}} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + y/\sqrt{n}b\right) \rho\left(\tilde{\theta} + y/\sqrt{n}b\right) \mathrm{d}y - P_K\left(Y_{(n)}, n\right)\right\} P_K\left(\xi, n\right)}{\left\{\int_{Y_{(1)}}^{Y_{(n)}} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + y/\sqrt{n}b\right) \rho\left(\tilde{\theta} + y/\sqrt{n}b\right) \mathrm{d}y\right\} P_K\left(Y_{(n)}, n\right)}\right| \\
\leq &\frac{1}{\left|\int_{Y_{(1)}}^{Y_{(n)}} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + y/\sqrt{n}b\right) \rho\left(\tilde{\theta} + y/\sqrt{n}b\right) \mathrm{d}y\right|} \qquad (2.4.5) \\
&\times \left\{\left|\int_{Y_{(1)}}^{\xi} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + \frac{y}{\sqrt{n}b}\right) \rho\left(\tilde{\theta} + \frac{y}{\sqrt{n}b}\right) \mathrm{d}y - P_K\left(\xi, n\right)\right|\right. \qquad (2.4.6) \\
&\left. + \left|\int_{Y_{(1)}}^{Y_{(n)}} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + \frac{y}{\sqrt{n}b}\right) \rho\left(\tilde{\theta} + \frac{y}{\sqrt{n}b}\right) \mathrm{d}y - P_K\left(Y_{(n)}, n\right)\right| \left|\frac{P_K\left(\xi, n\right)}{P_K\left(Y_{(n)}, n\right)}\right|\right\} \\
\leq &\frac{1}{C_1}\left\{M_1 n^{-(K+1)/2} + M_1 n^{-(K+1)/2} \frac{C_2}{C_1}\right\} = \frac{M_1}{C_1}\left(1 + \frac{C_2}{C_1}\right) n^{-(K+1)/2}. \qquad (2.4.7)
\end{aligned}$$

Now we find the quotient series of $P_K(\xi, n)/P_K\left(Y_{(n)}, n\right)$. By the definition of $\gamma_i(\xi, n)$, through simple calculation, we have

$$\frac{P_K(\xi, n)}{P_K\left(Y_{(n)}, n\right)} = \sum_{i=0}^{\infty} \gamma_i(A, n)\, n^{-i/2},$$

By the discussion following Lemma 7 in the Appendix A.2 we know that all γ_i are almost surely uniformly bounded for all large n. Thus, there exists a constant M_3, such that

$$\left| \frac{P_K(\xi, n)}{P_K\left(Y_{(n)}, n\right)} - \Phi(\xi \mid R_n) - \sum_{i=1}^{K} \gamma_i(A, n)\, n^{-i/2} \right| \leq M_3 n^{-(K+1)/2}. \tag{2.4.8}$$

We combine Eq. (3.7.4) and Eq. (3.7.5), to get Eq. (2.4.2). □

Let $K = 2$. Then we get asymptotic normality of the posterior distribution.

COROLLARY 1 (Bernstein-von Mises Theorem). Use the assumption in Theorem 1 with $K = 2$, then the posterior distribution converges in distribution to normal distribution almost surely, that is

$$\sqrt{n}b\left(\theta - \tilde{\theta}\right) \Big| X_1, \ldots, X_n \to N(0, 1), \text{ a.s.},$$

Indeed, Theorem 1 builds a strong foundation for asymptotic expansions of many other quantities based on the posterior, such as the mean and higher-order posterior moments. This follows simply by replacing the prior density $\rho(\theta)$ with an appropriate function. Here we use the posterior mean as an example. More examples can be found in Johnson (1970).

EXAMPLE 1. Replace $\rho(\theta)$ in Eq. (2.4.1) by $y\rho(\theta)$,

$$\left| \int_{Y_{(1)}}^{Y_{(n)}} \prod_{i=1}^{n} \hat{w}_i \left(\tilde{\theta} + \frac{y}{\sqrt{n}b}\right) \left\{ y\rho\left(\tilde{\theta} + \frac{y}{\sqrt{n}b}\right) \right\} \mathrm{d}y - P_K^N\left(Y_{(n)}, n\right) \right| \leq M_1 n^{-(K+1)/2},$$

where

$$P_K^N(\xi, n) = \sum_{k=0}^{K} \left\{ \int_{Y_{(1)}}^{\xi} \alpha_k(y, n) \exp\left(-\frac{y^2}{2}\right) y\, \mathrm{d}y \right\} n^{-(K+1)/2}.$$

Applying the same argument as in the proof of Theorem 2, the asymptotic expansion of the posterior mean is

$$E\left\{\sqrt{n}b\left(\theta - \tilde{\theta}\right) \mid X\right\} = \left\{ \frac{\rho'\left(\tilde{\theta}\right)}{\rho\left(\tilde{\theta}\right) b} \frac{\int_{Y_{(1)}}^{Y_{(n)}} y^2 \varphi(y)\, \mathrm{d}y}{\int_{Y_{(1)}}^{Y_{(n)}} \varphi(y)\, \mathrm{d}y} + \frac{a_{3n}}{b^3} \frac{\int_{Y_{(1)}}^{Y_{(n)}} y^4 \varphi(y)\, \mathrm{d}y}{\int_{Y_{(1)}}^{Y_{(n)}} \varphi(y)\, \mathrm{d}y} \right\} n^{-1} + O_P\left(n^{-\frac{3}{2}}\right).$$

Since $Y_{(n)} \to +\infty$ and $Y_{(1)} \to -\infty$ a.s. as $n \to \infty$,

$$\lim_{n\to\infty} \frac{\int_{Y_{(1)}}^{Y_{(n)}} y^2 \varphi(y)\, \mathrm{d}y}{\int_{Y_{(1)}}^{Y_{(n)}} \varphi(y)\, \mathrm{d}y} = \int_{\mathbb{R}} y^2 \varphi(y)\, \mathrm{d}y = 1, \lim_{n\to\infty} \frac{\int_{Y_{(1)}}^{Y_{(n)}} y^4 \varphi(y)\, \mathrm{d}y}{\int_{Y_{(1)}}^{Y_{(n)}} \varphi(y)\, \mathrm{d}y} = \int_{\mathbb{R}} y^4 \varphi(y)\, \mathrm{d}y = 3.$$

Then a moment matching prior (Ghosh and Liu (2011)) is found as the solution of

$$\frac{\rho'(\theta)}{\rho(\theta)} = -\lim_{n\to\infty} \frac{3a_{3n}}{b^2}.$$

For the empirical likelihood and the exponentially tilted empirical likelihood, some heavy algebra yields

$$\begin{aligned} a_{3n}\left(\tilde{\theta}\right) &= \frac{\left\{\sum_{i=1}^{n}\partial g\left(X_i,\tilde{\theta}\right)/\partial\theta\right\}^3\sum_{i=1}^{n}g^3\left(X_i,\tilde{\theta}\right)}{3n\left\{\sum_{i=1}^{n}g^2\left(X_i,\tilde{\theta}\right)\right\}^3} \\ &\quad -\frac{\left\{\sum_{i=1}^{n}\partial g\left(X_i,\tilde{\theta}\right)/\partial\theta\right\}^2\sum_{i=1}^{n}g\left(X_i,\tilde{\theta}\right)\partial g\left(X_i,\tilde{\theta}\right)/\partial\theta}{n\left\{\sum_{i=1}^{n}g^2\left(X_i,\tilde{\theta}\right)\right\}^2} \\ &\quad +\frac{\sum_{i=1}^{n}\partial g\left(X_i,\tilde{\theta}\right)/\partial\theta\sum_{i=1}^{n}\partial^2 g\left(X_i,\tilde{\theta}\right)/\partial\theta^2}{2n\left\{\sum_{i=1}^{n}g^2\left(X_i,\tilde{\theta}\right)\right\}}. \end{aligned}$$

Using strong law of large numbers,

$$\begin{aligned} \lim_{n\to\infty}a_{3n} &= \frac{\left[E\left\{\partial g\left(X_1,\theta\right)/\partial\theta\right\}\right]^3E\left\{g^3\left(X_1,\theta\right)\right\}}{3\left[E\left\{g^2\left(X_1,\theta\right)\right\}\right]^3}-\frac{\left[E\left\{\partial g\left(X_1,\theta\right)/\partial\theta\right\}\right]^2E\left\{g\left(X_1,\theta\right)\partial g\left(X_1,\theta\right)/\partial\theta\right\}}{\left\{Eg^2\left(X_1,\theta\right)\right\}^2} \\ &\quad +\frac{E\left\{\partial g\left(X_1,\theta\right)/\partial\theta\right\}E\left\{\partial^2 g\left(X_1,\theta\right)/\partial\theta^2\right\}}{2E\left\{g^2\left(X_1,\theta\right)\right\}}\text{ a.s. }\left(P_\theta\right). \end{aligned}$$

Hence, we have the following corollary.

COROLLARY 2. Assume the conditions in Theorem 1 are satisfied at $K=4$. Then the first order moment matching prior of Bayesian empirical likelihood and Bayesian exponentially tilted empirical likelihood is

$$\begin{aligned} \rho\left(\theta\right) &= \exp\left\{-\int_{-\infty}^{\theta}\Big(\frac{\left[E\left\{\partial g\left(X_1,s\right)/\partial s\right\}\right]^3E\left\{g^3\left(X_1,s\right)\right\}}{\left[E\left\{g^2\left(X_1,s\right)\right\}\right]^4}\right. \\ &\quad +\frac{3\left[E\left\{\partial g\left(X_1,s\right)/\partial s\right\}\right]^2E\left\{g\left(X_1,s\right)\partial g\left(X_1,s\right)/\partial s\right\}}{\left[E\left\{g^2\left(X_1,s\right)\right\}\right]^3} \\ &\quad \left.-\frac{3E\left\{\partial g\left(X_1,s\right)/\partial s\right\}E\left\{\partial^2 g\left(X_1,s\right)/\partial s^2\right\}}{2\left[E\left\{g^2\left(X_1,s\right)\right\}\right]^2}\Big)\,\mathrm{d}s\right\}. \end{aligned}$$

In the special case, $g\left(x,\theta\right)=x-\theta$, by Corollary 2, the moment matching prior is

$$\rho\left(\theta\right)=\exp\left(\int_{-\infty}^{\theta}\frac{E\left\{\left(X_1-s\right)^3\right\}}{\left[E\left\{\left(X_1-s\right)^2\right\}\right]^4}\,\mathrm{d}s\right).$$

2.5. Simulation Results

In this section, we give some simulation results. Here we take $g\left(X_i,\theta\right)=X_i-\theta, i=1,\ldots,n$. Let $K=3$, we compare the first order approximation with normal approximation and second order approximation. By heavy algebra, we get for all the three empirical likelihoods,

$$\tilde{l}^{(3)}\left(\overline{X}\right)=\frac{2n^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^3}{\left\{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right\}^3}.$$

So $\tilde{l}^{(3)}\left(\overline{X}\right)$ for the three empirical likelihoods are asymptotically equivalent up to the second order. The true cumulative distribution function is calculated by numerical integration. The normal approximation polynomial is $\Phi\left(\xi \mid R_n\right)$, and the second order approximation polynomial is

$$\Phi\left(\xi \mid R_n\right) + \frac{1}{\sqrt{n}}\left[\left\{\frac{\rho'\left(\overline{X}\right)}{\rho\left(\overline{X}\right)b}\right\}\left\{\frac{\varphi\left(Y_{(1)}\right)-\varphi\left(\xi\right)}{\int_{Y_{(1)}}^{Y_{(n)}}\varphi\left(y\right)\,\mathrm{d}y} - \frac{\varphi\left(Y_{(1)}\right)-\varphi\left(Y_{(n)}\right)}{\int_{Y_{(1)}}^{Y_{(n)}}\varphi\left(y\right)\,\mathrm{d}y}\Phi\left(\xi \mid R_n\right)\right\}\right.$$
$$+\left\{\frac{2n^{-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^3}{6b^9}\right\}\left\{\frac{Y_{(1)}^2\varphi\left(Y_{(1)}\right)+2\varphi\left(Y_{(1)}\right)-\xi^2\varphi\left(\xi\right)-2\varphi\left(\xi\right)}{\int_{Y_{(1)}}^{Y_{(n)}}\varphi\left(y\right)\,\mathrm{d}y}\right.$$
$$\left.\left.-\frac{Y_{(1)}^2\varphi\left(Y_{(1)}\right)+2\varphi\left(Y_{(1)}\right)-Y_{(n)}^2\varphi\left(Y_{(n)}\right)-2\varphi\left(Y_{(n)}\right)}{\int_{Y_{(1)}}^{Y_{(n)}}\varphi\left(y\right)\,\mathrm{d}y}\Phi\left(\xi \mid R_n\right)\right\}\right].$$

We take samples of size $n = 10$, and80 from a t distribution with degrees of freedom 100, and the Cauchy prior. Set Cressie–Read divergence parameter $\lambda = 2$. Then

$$\begin{aligned}\rho\left(\overline{X}\right) &= \frac{1}{\pi\left(1+\overline{X}^2\right)},\\ \rho'\left(\overline{X}\right) &= -\frac{2\overline{X}}{\pi\left(1+\overline{X}^2\right)^2}.\end{aligned}$$

The results are given in Figure 2.5.1 on page 25 and Figure 2.5.2 on page 26. In both plots, the red line stands for normal approximation of the posterior cumulative distribution function, the blue line stands for the first order approximation, the green line stands for the posterior based on the empirical likelihood, the purple line stands for the posterior based on the exponentially tilted empirical likelihood, and the black line stands for the Cressie–Read divergence empirical likelihood. We see that even when the sample size is 10, the three types of empirical likelihoods are quite close to each other, which supports the fact they are equivalent at least up to the second order, and the second order approximation works well. The first order approximation is closer than the normal approximation, which lends credence to our theorem. When the sample size increases to 80, all the lines almost coincide with each other, which means that the approximations are quite successful.

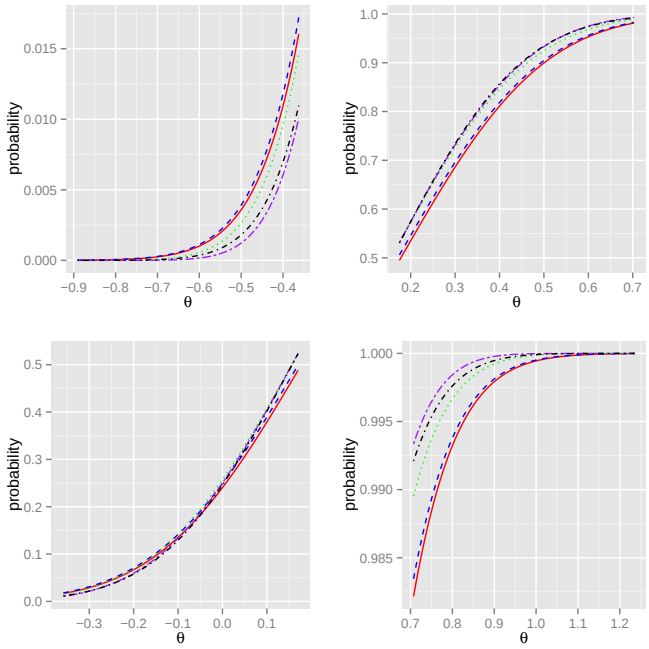


FIGURE 2.5.1. Posterior cumulative distribution function when sample size is 10

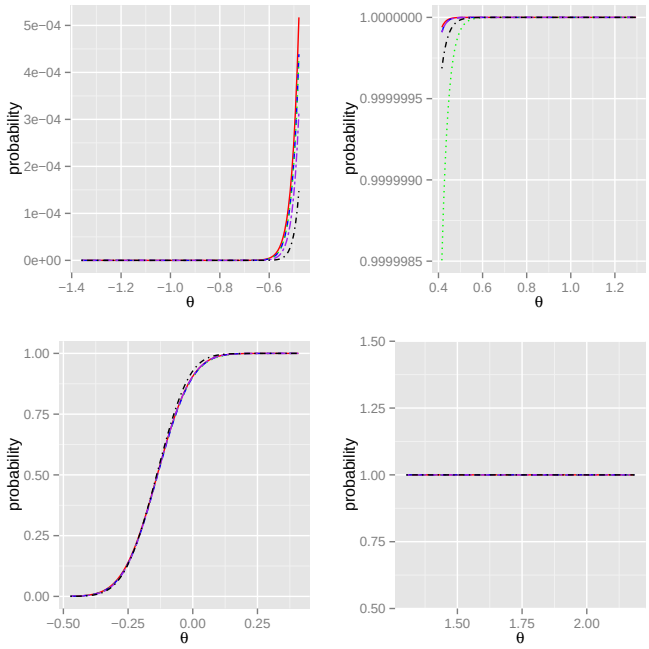


FIGURE 2.5.2. Posterior cumulative distribution function when sample size is 80

2.6. Discussion

The paper provides an asymptotic expansion of the posterior based on an empirical likelihood subject to a linear constraint. The Bernstein –von Mises theorem and asymptotic expansions of the cumulative distribution function and the posterior mean are obtained as corollaries. Future work will include an extension to the multivariate case as well as expansions subject to multiple constraints. Another potential topic of research is asymptotic expansion of posteriors under regression constraints, extending the arguments of **??**

CHAPTER 3

Higher-Order Properties of Bayesian Empirical Likelihood: Multivariate Case

3.1. Introduction

Empirical likelihood, over the years, has become a very popular topic of statistical research. The name was coined by Owen in his classic 1986 paper, although similar ideas are found even earlier in the works of Hartley and Rao (1968), Thomas and Grunkemeier (1975), Rubin et al. (1981) and others. The main advantage of empirical likelihood is that is involves fewer assumptions that a regular likelihood, and yet shares the same asymptotic properties of the latter.

Research in this area has primarily been frequentist with a long list of important theoretical developments accompanied by a large number of applications. To our knowledge, the first Bayesian work in this general area appeared in the article of Lazar (2003) followed by some related work in Schennach (2005), Schennach et al. (2007) , the latter introducing the concept of "exponentially tilted empirical likelihood". Lazar (2003) suggested using empirical likelihood as a substitute for the usual likelihood and carry out Bayesian analysis in the usual way.

Baggerly (1998) viewed empirical likelihood as a method of assigning probabilities to a n-cell contingency table in order to minimize a goodness-of-fit criterion. He selected Cressie-Read power divergence statistics as one such criterion for construction of confidence regions in a number of situations and pointed out also how the usual empirical likelihood, exponentially tilted empirical likelihood and others can be viewed as special cases of the Cressie-Read criterion by appropriate choice of the the power parameter. This was also discussed in Owen (2010) who pointed our that all members of the Cressie-Read family lead to "empirical divergence analogues of the empirical likelihood in which asymptotic χ^2 calibration holds for the mean".

The objective of this article is to provide an asymptotic expansion of the posterior distribution based on empirical likelihood and its variations under certain regularity conditions and a mean constraint. The work is inspired by the work of Fang and Mukerjee (2006) who provided a somewhat different expansion subject to a mean constraint. Unlike Fang and Mukerjee (2005, 2006), our result is based on the derivatives of the pseudo likelihood with respect to the parameters of interest evaluated at the maximum EL estimator, and a rigorous expansion is provided with particular attention in handling the remainder terms. Moreover, we consider a general estimating equation which includes the mean example of Fang and Mukerjee (2006) as a special case. The need for different pseudo-likelihoods for statistical inference is felt all the more in these days, especially for the analysis of high-dimensional data, where the usual likelihood based analysis is hard to perform,

These alternative likelihoods are equally valuable for approximate Bayesian computations (ABC), a topic which has only recently surfaced in the statistics literature (see e.g. Comuet et al. , 2008).

Asymptotic expansion of the posterior based on a regular likelihood was given earlier in Johnson (1970), and later in Ghosh et al. (1982). We follow their approach with many necessary modifications in view of the fact that any meaningful prior needs to have support in the nondecreasing compact set H. As a special case of our result, we get the celebrated Bernstein von-Mises Theorem. The latter was mentioned in Lazar (2003) for the special case of empirical likelihood, but here we provide a rigorous derivation with the needed regularity conditions.

Unlike in univariate case, where the posterior is centered at M-estimator, which guarantees the consistency under mild regular conditions, in multivariate case, the posterior is centralized at maximum generalized empirical likelihood estimator. This concept is generalized from empirical likelihood case which is fully explored in Qin and Lawless (1994) to exponentially tilted empirical likelihood and even Cressie-Read case. A similar concept has been developed in Newey and Smith (2004) under the similar name. There is slight different between their definition and ours in both intuition and mathematical definition, however, numerical study shows the performance of these two are quite similar. Furthermore, in order for the empirical likelihood sample moments to be finite, we need more restrictive conditions to guarantee not only the consistency of generalized empirical likelihood estimator, but also its law of iterative logarithm.

The organization of the remaining sections of this paper is as follows. In Section 2 of this paper, we consider the basic settings of the empirical likelihood, exponentially tilted empirical likelihood, and finally the more general Cressie-Read divergence criterion. Section 3 contains some basic lemmas pertaining to these three formulations. While both empirical and exponentially tilted empirical likelihood are indeed limiting cases of the general Cressie-Read formulation, for technical reasons as as well as for the sake of transparency, we have presented some of these lemmas separately for these three cases. Section 4 contains the main result, namely the asymptotic expansion of the posterior and presents a unified derivation. Some simulation results are presented in Section 5. Section 6 contains some concluding remarks.

3.2. Basic Settings

Suppose $X_1, \ldots, X_n$ are independent and identically distributed random vectors satisfying $Eg(X_1, \theta) = 0$, where $\theta \in \mathbb{R}^p$ and $g(x, \theta) = (g_1(x, \theta), g_2(x, \theta), \ldots, g_r(x, \theta)) \in \mathbb{R}^r$. Like in Qin and Lawless (1994), we focus on the situation $r > p$. When $r \leq p$, the posterior will still asymptotically center around M-estimator, and all the arguments are the same as univariate case. However, when $r > p$, M-estimator may not even exist, we need further generalization. In this context, Owen (1988), formulated empirical likelihood as a nonparametric likelihood of the form $\prod_{i=1}^{n} w_i(\theta)$, where w_i is the probability mass assigned to $X_i\,(i = 1, \ldots, n)$ satisfying the constraints

$$\begin{cases} w_i > 0, \text{foralli}; \\ \sum_{i=1}^{n} w_i = 1; \\ \sum_{i=1}^{n} w_i g(X_i, \theta) = 0. \end{cases} \tag{3.2.1}$$

The target is to maximize $\prod_{i=1}^{n} w_i$ or equivalently $\sum_{i=1}^{n} \log w_i$ with respect to $w_1, \ldots, w_n$ subject to the constraints given in (3.2.1). Applying the Lagrange multiplier method, the solution turns out to be

$$\hat{w}_i^{\mathrm{EL}}(\theta) = \frac{1}{n\left[1 + \nu^T g(X_i, \theta)\right]}, i = 1, 2, \ldots, n, \tag{3.2.2}$$

where $\nu \in \mathbb{R}^r$, the Lagrange multiplier satisfies

$$\sum_{i=1}^{n} \frac{g(X_i, \theta)}{1 + \nu^T g(X_i, \theta)} = 0. \tag{3.2.3}$$

It may be noted that Fang and Mukerjee (2005, 2006) $g(X_i, \theta) = X_i - \theta$.

Closely related to the empirical likelihood is the exponentially tilted empirical likelihood where the objective is to maximize the Shannon entropy $-\sum_{i=1}^{n} w_i \log w_i$ still subject to the constraints in (3.2.1). The resulting solution is given by

$$\hat{w}_i^{\mathrm{ET}}(\theta) = \frac{\exp\left(-\nu^T g(X_i, \theta)\right)}{\sum_{j=1}^{n} \exp\left(-\nu^T g(X_i, \theta)\right)}, \tag{3.2.4}$$

where ν, the Lagrange multiplier, satisfies

$$\sum_{i=1}^{n} \exp\left(-\nu^T g(X_i, \theta)\right) g(X_i, \theta) = 0. \tag{3.2.5}$$

The exponentially tilted empirical likelihood is related to Kullback-Leibler divergence between two empirical distributions, one with weights w_i assigned to the n sample points, and the other with uniform weights $1/n$ assigned to the sample points.

The general Cressie-Read divergence criterion given by

$$\mathrm{CR}(\lambda) = \frac{2}{\lambda(\lambda + 1)} \sum_{i=1}^{n} \left[(nw_i)^{-\lambda} - 1\right].$$

We focus on the cases $\lambda \geq 0$ and $\lambda \leq -1$, because in these cases, $\mathrm{CR}(\lambda)$ is a convex function of the $w_i\,(i = 1, \ldots, n)$,hence the minimization problem will produce a unique solution. The following lemma also shows within this range, the resulting empirical weights behaviour more like a likelihood. The limiting cases $\lambda \to 0$ and $\lambda \to -1$ correspond to the usual empirical likelihood and the exponentially tilted empirical likelihood as defined earlier.

For convex $\mathrm{CR}(\lambda)$, its minimum will be attained in the compact set H_n determined by data. The Lagrange multiplication method now gives the weights

$$\hat{w}_i^{\mathrm{CR}}(\theta) = \frac{1}{n}\left(\mu + \nu^T g(X_i, \theta)\right)^{-\frac{1}{\lambda+1}}, i = 1, 2, \ldots, n, \tag{3.2.6}$$

where $\mu \equiv \mu(\theta)$ and $\nu \equiv \nu(\theta)$ satisfy

$$\begin{cases} \sum_{i=1}^{n} \left(\mu + \nu^T T g(X_i, \theta)\right)^{-\frac{1}{\lambda+1}} = n, \\ \sum_{i=1}^{n} \left(\mu + \nu^T g(X_i, \theta)\right)^{-\frac{1}{\lambda+1}} X_i = 0. \end{cases} \tag{3.2.7}$$

We now introduce the posterior based on an empirical likelihood. The basic ideal was first introduced by Lazar (2003) with bunch of numerical examples. The intuition relies on close relationship between empirical likelihood and empirical distribution. Owen (2010) formulated 105 4 the two concepts under the same optimization framework, that is, they shared the same objective function, but the

former one was solved under parametric constraints, while the latter was not. Considering this similarity, we can use the empirical likelihood as a valid distribution parametrized by inferential target. To concrete this intuition in Bayesian philosophy, writing $\hat{w}_i(\theta)$ as generic notation for either $\hat{w}_i^{\mathrm{EL}}$, $\hat{w}_i^{\mathrm{ET}}$ or $\hat{w}_i^{\mathrm{CR}}$, π with a prior probability density function $\rho(\theta)$, with support in H_n, the profile (pseudo) posterior is given by

$$\pi(\theta \mid X_1, X_2, \ldots, X_n) = \frac{\prod_{i=1}^{n} \hat{w}_i(\theta)\, \rho(\theta)}{\int_{H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \rho(\theta)\, \mathrm{d}\theta}. \tag{3.2.8}$$

The main objective of this paper is to provide an asymptotic expansion of $\pi(\theta \mid X_1, X_2, \ldots, X_n)$. This will include in particular the Bernstein-von Mises theorem. Towards this end, we develop a few necessary lemmas in the next section.

3.3. Lemmas

We first give an explanation of natural domain of θ, under empirical likelihood settings. In practice, some values of θ will result in an empty feasible set in constraints (3.2.1). The which guarantees an non-empty feasible set, and thus a solution of the optimization problem constitutes 120 natural domain of empirical likelihood. One may questions whether the size of the natural domain is large enough to contain the true value. The following lemma alleviates this worry.

LEMMA 1. *Assume $g(\cdot,\cdot)$ is a continuous vector value function, then the natural domain defined by the constraints (3.2.1) is a compact set and nondecreasing with respect to sample size n.*

PROOF. By the third constraint of (3.2.1), θ is a continuous vector value function of $w_1, w_2, \ldots, w_n$, but w_i are defined on a simplex which is a compact set through the first constraint of (3.2.1). We may recall that continuous function maps compact sets to compact sets. Hence, θ is naturally defined on a compact set denoted by H.

If for any $j = 1, 2, \ldots, r$, $g_j(X_i, \theta)$, $i = 1, 2, \ldots, n$, are all non-positive or all non-negative, then the constraints (3.2.1) are violated and $H = \emptyset$. Hence ,

$$\begin{aligned} H &= \left\{ \bigcup_{j=1}^{r} \left\{ \left[\bigcap_{i=1}^{n} \left(g_j(X_i, \theta) \geq 0 \right) \right] \bigcup \left[\bigcap_{i=1}^{n} \left(g_j(X_i, \theta) \leq 0 \right) \right] \right\} \right\}^c \\ &= \bigcap_{j=1}^{r} \left\{ \left[\bigcup_{i=1}^{n} \left(g_j(X_i, \theta) \geq 0 \right)^c \right] \bigcap \left[\bigcup_{i=1}^{n} \left(g_j(X_i, \theta) \leq 0 \right)^c \right] \right\}. \end{aligned}$$

With n increases, both $\bigcup_{i=1}^{n} \left(g_j(X_i, \theta) \geq 0 \right)^c$ and $\bigcup_{i=1}^{n} \left(g_j(X_i, \theta) \leq 0 \right)^c$ will increase, so does their intersection H. □

Although, intuitively we expect the empirical likelihood to behave as the true likelihood, we need some theoretical support to show that the former enjoys some of the basic properties of the latter. In particular, we need to verify that ν and μ are smooth functions of θ and the (pseudo) Fisher Information based on the empirical likelihood is positive.

We first establish the positiveness of the Fisher information . We consider the three cases separately to introduce more transparency and continuity in our approach.

Our first lemma shows that the Lagrange multipliers $\nu(\theta)$ and $\mu(\theta)$ are all smooth functions of θ.

LEMMA 2. *For empirical likelihood, exponentially tilted empirical likelihood and Cressie-Read with parameter(λ), the Lagrange multipliers $\nu(\theta)$ and $\mu(\theta)$ are smooth functions of θ.*

PROOF. We first consider empirical likelihood and observe that, $\nu(\theta)$ is a implicit function of θ in view of (3.2.3) . Further

$$\frac{\partial}{\partial \nu}\sum_{i=1}^{n}\frac{g(X_i,\theta)}{1+\nu^T g(X_i,\theta)}=-\sum_{i=1}^{n}\frac{g(X_i,\theta)\,g^T(X_i,\theta)}{\left(1+\nu^T g(X_i,\theta)\right)^2},$$

is negative definite, so that by the implicit function theorem, ν is differentiable in θ. Moreover, differentiating both sides of (3.2.3) with respect to ν, one gets

$$\begin{aligned}0 &= \sum_{i=1}^{n}\frac{1}{1+\nu^T g(X_i,\theta)}\frac{\partial g(X_i,\theta)}{\partial\theta^T}\frac{\partial\theta}{\partial\nu}-\sum_{i=1}^{n}\frac{\nu^T g(X_i,\theta)}{\left(1+\nu^T g(X_i,\theta)\right)^2}\frac{\partial g(X_i,\theta)}{\partial\theta^T}\frac{\partial\theta}{\partial\nu}\\ &\quad -\sum_{i=1}^{n}\frac{g(X_i,\theta)\,g^T(X_i,\theta)}{\left(1+\nu^T g(X_i,\theta)\right)^2},\end{aligned}$$

which on simplification leads to

$$\frac{\partial\theta}{\partial\nu}=-\left[\sum_{i=1}^{n}\frac{1}{\left(1+\nu^T g(X_i,\theta)\right)^2}\frac{\partial g(X_i,\theta)}{\partial\theta}\right]^{-1}\sum_{i=1}^{n}\frac{g(X_i,\theta)\,g^T(X_i,\theta)}{\left(1+\nu^T g(X_i,\theta)\right)^2},\tag{3.3.1}$$

Next, for exponentially tilted empirical likelihood, in view of (3.2.5) and the relation

$$\begin{aligned}&\frac{\partial}{\partial\nu}\left(\sum_{i=1}^{n}\exp\left(-\nu^T g(X_i,\theta)\right)g(X_i,\theta)\right)\\ =\;&-\sum_{i=1}^{n}\exp\left(-\nu^T g(X_i,\theta)\right)g(X_i,\theta)\,g^T(X_i,\theta).\end{aligned}$$

Note that this matrix is negative definite. Once again, the implicit function theorem guarantees the differentiability of ν in θ. Further, differentiating both sides of (3.2.5) with respect to θ, one gets

$$\frac{\partial\nu}{\partial\theta}=\left(\sum_{i=1}^{n}\frac{g(X_i,\theta)\,g^T(X_i,\theta)}{\exp\left(-\nu^T g(X_i,\theta)\right)}\right)^{-1}\left(\sum_{i=1}^{n}\frac{1-\nu^T g(X_i,\theta)}{\exp\left(-\nu^T g(X_i,\theta)\right)}\frac{\partial g(X_i,\theta)}{\partial\theta}\right).\tag{3.3.2}$$

A similar conclusion is achieved for $\nu(\theta)$ and $\mu(\theta)$ defined in (3.2.7) in connection with CR(λ). Specifically, defining

$$\begin{cases}F_1=\sum_{i=1}^{n}\left(\mu+\nu^T g(X_i,\theta)\right)^{-\frac{1}{\lambda+1}}-n,\\ F_2=\sum_{i=1}^{n}\left(\mu+\nu^T g(X_i,\theta)\right)^{-\frac{1}{\lambda+1}}g(X_i,\theta),\end{cases}$$

it follows that,

$$\frac{\partial(F_1,F_2)}{\partial(\mu,\nu)}=-\frac{1}{\lambda+1}\begin{pmatrix}\sum_{i=1}^{n}q_i & \sum_{i=1}^{n}q_i g(X_i,\theta)\\ \sum_{i=1}^{n}q_i g^T(X_i,\theta) & \sum_{i=1}^{n}q_i g(X_i,\theta)\,g^T(X_{i,},\theta)\end{pmatrix},$$

where $q_i = \left(\mu + \nu^T g\left(X_i, \theta\right)\right)^{-\frac{1}{\lambda+1}-1}$. Then the determinant of Jacobian is

$$\begin{aligned}
\det \frac{\partial\left(F_1, F_2\right)}{\partial\left(\mu, \nu\right)} &= \left(\frac{1}{\lambda+1}\right)^2 \left(\sum_{i=1}^n q_i \sum_{i=1}^n q_i g\left(X_i, \theta\right) g^T\left(X_i, \theta\right) - \sum_{i=1}^n q_i g\left(X_i, \theta\right) \sum_{i=1}^n q_i g^T\left(X_i, \theta\right)\right) \\
&= \left(\frac{1}{\lambda+1}\right)^2 \left(\sum_{i=1}^n q_i\right)^2 \sum_{i=1}^n \frac{q_i}{\sum_{j=1}^n q_j} \left[g\left(X_i, \theta\right) - \left(\sum_{i=1}^n \frac{q_i}{\sum_{j=1}^n q_j} g\left(X_i, \theta\right)\right)\right] \\
&\quad \times \left[g\left(X_i, \theta\right) - \left(\sum_{i=1}^n \frac{q_i}{\sum_{j=1}^n q_j} g\left(X_i, \theta\right)\right)\right]^T,
\end{aligned}$$

which is positive definite. Again, by implicit function theorem, one gets differentiability of $\mu\left(\theta\right)$ and $\nu\left(\theta\right)$with respect to θ, and

$$\begin{aligned}
&\begin{pmatrix} \partial\mu/\partial\theta \\ \partial\nu/\partial\theta \end{pmatrix} \\
= &\left(\frac{\partial\left(F_1, F_2\right)}{\partial\left(\mu, \nu\right)}\right)^{-1} \begin{pmatrix} \partial F_1/\partial\theta \\ \partial F_2/\partial\theta \end{pmatrix} \qquad (3.3.3) \\
= &\left(-\frac{1}{\lambda+1}\right)\left(\lambda+1\right)^2 \left[\sum_{i=1}^n q_i \sum_{i=1}^n q_i g\left(X_i, \theta\right) g^T\left(X_i, \theta\right) - \sum_{i=1}^n q_i g\left(X_i, \theta\right) \sum_{i=1}^n q_i g^T\left(X_i, \theta\right)\right]^{-1} \\
&\times \begin{pmatrix} \sum_{i=1}^n q_i g\left(X_i, \theta\right) g^T\left(X_i, \theta\right) & -\sum_{i=1}^n q_i g\left(X_i, \theta\right) \\ -\sum_{i=1}^n q_i g^T\left(X_i, \theta\right) & \sum_{i=1}^n q_i \end{pmatrix} \\
&\times \begin{pmatrix} -\left(\lambda+1\right)^{-1} \sum_{i=1}^n q_i \nu^T \partial g\left(X_i, \theta\right)/\partial\theta \\ -\left(\lambda+1\right) \sum_{i=1}^n q_i \nu^T g\left(X_i, \theta\right) \partial g\left(X_i, \theta\right)/\partial\theta + \sum_{i=1}^n \left(\mu + \nu^T g\left(X_i, \theta\right)\right)^{\frac{1}{\lambda+1}} \partial g\left(X_i, \theta\right)/\partial\theta \end{pmatrix} \qquad (3.3.4)
\end{aligned}$$

□

The next result shows that all the derivatives of the Lagrange multipliers $\nu\left(\theta\right)$ and $\mu\left(\theta\right)$ are smooth functions of $\theta \in H$. We provide a unified proof for all three cases where we utilize the previous lemma.

ASSUMPTION 6. *Assume for any $(k_1, k_2, \ldots, k_p) \in \mathbb{N}^p$, satisfying $\sum_{i=1}^p k_i = k \leq K+4$, the higher-order mixture partial derivatives*

$$\frac{\partial^k g\left(X, \theta\right)}{\partial\theta_1^{k_1} \partial\theta_2^{k_2} \cdots \partial\theta_p^{k_p}},$$

exists and continuous in θ, almost surely in X.

LEMMA 3. *Under the Assumption 6, all partial derivatives of $\nu\left(\theta\right)$ and $\mu\left(\theta\right)$ are smooth functions of θ for $\theta \in H$.*

PROOF. The result is proved by induction. We have see already in Lemma 2, the gradient $\nabla\nu\left(\theta\right)$ and $\nabla\mu\left(\theta\right)$ are smooth functions of θ. Suppose the result holds for all kth partial derivatives of $\nu\left(\theta\right)$ and $\mu\left(\theta\right)$ for $k \leq K$. The writing

$$\nabla^k \nu\left(\theta\right) = h_k\left(\nu\left(\theta\right), \theta\right), 1 \leq k \leq K,$$

$$\nabla^{k+1} \nu\left(\theta\right) = \frac{\partial h_k}{\partial \nu^T} \nabla\nu + \frac{\partial h_k}{\partial\theta}$$

which is also a smooth function of θ by the induction hypothesis and Lemma 1. A similar proof works for $\mu\left(\theta\right)$. □

3.4. Maximum Generalized Empirical Likelihood Estimator

One important justification of maximum likelihood estimator is the good performance substantiated by large sample theory, such as consistency and asymptotic normality. It can be shown the above theoretical results can be seamlessly transplanted onto the estimator resulted from maximum empirical weights. Qin and Lawless (1994) already pointed out the validity to use empirical likelihood to define maximum empirical likelihood estimator. We extend the same idea into more general Cressie-Read family.

DEFINITION 1 (Maximum Generalized Empirical Likelihood Estimator). Let $\hat{w}_i(\theta)$ be the empirical weights on sample $X_1, X_2, \ldots, X_n$, generating from empirical likelihood, exponentially tilted empirical likelihood or Cressie-Read family. Then

$$\tilde{\theta} = \arg\min_{\theta \in H} -\sum_{i=1}^{n} \log \hat{w}_i(\theta),$$

called maximum generalized empirical likelihood estimator.

Lemma 3 establishes the continuity of empirical likelihood and Lemma 1 reveals the compactness of H, then these results guarantee the existence of $\tilde{\theta}$. So this concept is well-defined. One advantage of this estimator over the traditional M-estimator is that it can always produce a good quality estimator even when traditional one has no solution. Before exploring the asymptotic properties of this new estimator, we derive an equivalent formulation of maximum generalized empirical likelihood estimator, which may be more appropriate for numerical computation and theoretical derivation.

First, we define Ψ functions under the three cases. In empirical likelihood case, let

$$\begin{aligned}
\Psi_1(x \mid \theta, \nu) &= \frac{g(x,\theta)}{1+\nu^T g(x,\theta)}, \\
\Psi_2(x \mid \theta, \nu) &= \frac{1}{1+\nu^T g(x,\theta)} \nu^T \frac{\partial g(x,\theta)}{\partial \theta}.
\end{aligned}$$

Let $\Psi^{\mathrm{EL}}(x \mid \theta, \nu) = (\Psi_1, \Psi_2)^T$. In exponentially tilted empirical likelihood, let

$$\begin{aligned}
\Psi_1(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) &= \nu^T \frac{\partial g(x,\theta)}{\partial \theta} - \lambda_1 \exp\left(-\mu - \nu^T g(x,\theta)\right) \nu^T \frac{\partial g(x,\theta)}{\partial \theta} \\
&\quad - \exp\left(-\mu - \nu^T g(x,\theta)\right) \nu^T \frac{\partial g(x,\theta)}{\partial \theta} \lambda_2^T g(x,\theta) + \exp\left(-\mu - \nu^T g(x,\theta)\right) \lambda_2^T \frac{\partial g(x,\theta)}{\partial \theta}, \\
\Psi_2(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) &= g(x,\theta) - \lambda_1 \exp\left(-\mu - \nu^T g(x,\theta)\right) g(x,\theta) - \exp\left(-\mu - \nu^T g(x,\theta)\right) g(x,\theta) \lambda_2^T g(x,\theta), \\
\Psi_3(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) &= 1 - \lambda_1 \exp\left(-\mu - \nu^T g(x,\theta)\right) - \exp\left(-\mu - \nu^T g(x,\theta)\right) \lambda_2^T g(x,\theta), \\
\Psi_4(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) &= \exp\left(-\mu - \nu^T g(x,\theta)\right) - \frac{e}{n}, \\
\Psi_4(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) &= \exp\left(-\mu - \nu^T g(x,\theta)\right) g(x,\theta),
\end{aligned}$$

where λ_1 and λ_2 are new Lagrange multipliers which we will define in later lemma. Let $\Psi^{\mathrm{ET}}(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) = (\Psi_1, \ldots, \Psi_5)^T$. In Cressie-Read case, let

$$\begin{aligned}
\Psi_1(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) &= \frac{1}{\lambda+1}\left[\frac{\nu^T \partial g(x,\theta)/\partial\theta}{\mu+\nu^T g(x,\theta)} - \lambda_1 \frac{\nu^T \partial g(x,\theta)/\partial\theta}{(\mu+\nu^T g(x,\theta))^{1+1/(\lambda+1)}}\right. \\
&\quad \left. - \frac{\left(\nu^T \partial g(x,\theta)/\partial\theta\right)\left(\lambda_2^T g(x,\theta)\right)}{(\mu+\nu^T g(x,\theta))^{1+1/(\lambda+1)}} + \frac{\lambda_2^T \partial g(x,\theta)/\partial\theta}{(\mu+\nu^T g(x,\theta))^{1+1/(\lambda+1)}}\right], \\
\Psi_2(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) &= \frac{1}{\lambda+1}\left[\frac{g(x,\theta)}{\mu+\nu^T g(x,\theta)} - \lambda_1 \frac{g(x,\theta)}{(\mu+\nu^T g(x,\theta))^{1+1/(\lambda+1)}} - \frac{g(x,\theta)\left(\lambda_2^T g(x,\theta)\right)}{(\mu+\nu^T g(x,\theta))^{1+1/(\lambda+1)}}\right], \\
\Psi_3(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) &= \frac{1}{\lambda+1}\left[\frac{1}{\mu+\nu^T g(x,\theta)} - \lambda_1\left(\mu+\nu^T g(x,\theta)\right)^{-1-1/(\lambda+1)}\right. \\
&\quad \left. - \left(\mu+\nu^T g(x,\theta)\right)^{-1-1/(\lambda+1)} \lambda_2^T g(x,\theta)\right], \\
\Psi_4(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) &= \left(\mu+\nu^T g(x,\theta)\right)^{-1/(\lambda+1)} - 1, \\
\Psi_5(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) &= \left(\mu+\nu^T g(x,\theta)\right)^{-1/(\lambda+1)} g(x,\theta).
\end{aligned}$$

Let $\Psi^{\mathrm{CR}}(x \mid \theta, \nu, \mu, \lambda_1, \lambda_2) = (\Psi_1, \ldots, \Psi_5)^T$. Use above definitions, we elicit the following lemma.

LEMMA 4. *The maximum generalized empirical likelihood estimator is the solution of* $n^{-1}\sum_{i=1}^n \Psi(X_i \mid \theta, \nu, \mu, \lambda_1, \lambda_2) = 0$. *If we denote the solution as* $(\theta^*, \nu^*, \mu^*, \lambda_1^*, \lambda_2^*)$, *then* $\tilde{\theta} = \theta^*$, $\nu\left(\tilde{\theta}\right) = \nu^*$ *and* $\mu\left(\tilde{\theta}\right) = \mu^*$.

PROOF. For explicit, we state the proof separately under the three kinds of empirical likelihoods.

First in empirical likelihood. The relationship between θ and ν are restricted by equation constraint (3.2.3). By the Definition 1, the maximum generalized empirical likelihood estimator can also be the solution of

$$\max_{\theta,\nu} l(\theta) = -\frac{1}{n}\sum_{i=1}^n \log\left(1+\nu^T g(X_i,\theta)\right),$$

subject to,

$$\sum_{i=1}^n \frac{g(X_i,\theta)}{1+\nu^T g(X_i,\theta)} = 0.$$

Directly calculation shows

$$\frac{\partial l(\theta)}{\partial \nu} = \frac{1}{n}\sum_{i=1}^n \frac{g(X_i,\theta)}{1+\nu^T g(X_i,\theta)} = \frac{1}{n}\sum_{i=1}^n \Psi_1(X_i \mid \theta, \nu).$$

So the solution of unconstrained problem automatically satisfies the constraint (3.2.3). Moreover,

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{1}{n}\sum_{i=1}^n \Psi_2(X_i \mid \theta, \nu).$$

So the equation system $n^{-1}\sum_{i=1}^{n}\Psi\left(X_i \mid \theta,\nu\right)=0$ is the first order necessary condition of the optimization problem, hence the lemma holds in empirical likelihood case.

Next, we consider exponentially tilted empirical likelihood. Reviewing the Lagrange method to get the empirical weights (3.2.4), we find

$$\hat{w}_i\left(\theta\right)=\exp\left(-1-\mu-\nu^T g\left(X_i,\theta\right)\right),$$

where μ and ν satisfy

$$\begin{aligned}\sum_{i=1}^{n}\exp\left(-1-\mu-\nu^T g\left(X_i,\theta\right)\right) &= 1,\\ \sum_{i=1}^{n}\exp\left(-1-\mu-\nu^T g\left(X_i,\theta\right)\right)g\left(X_i,\theta\right) &= 0.\end{aligned}$$

Hence, the optimization problem in Definition 1 shares the same maximum as the problem

$$\max_{\theta,\nu,\mu} l\left(\theta\right)=\frac{1}{n}\sum_{i=1}^{n}\left(-1-\mu-\nu^T g\left(X_i,\theta\right)\right),$$

subject to

$$\begin{aligned}\sum_{i=1}^{n}\exp\left(-1-\mu-\nu^T g\left(X_i,\theta\right)\right) &= 1,\\ \sum_{i=1}^{n}\exp\left(-1-\mu-\nu^T g\left(X_i,\theta\right)\right)g\left(X_i,\theta\right) &= 0.\end{aligned}$$

Unlike in empirical likelihood case, here we do not have any shortcut. The Lagrange multiplier method is the only choice. The Lagrangian can be written as

$$\begin{aligned}L\left(\theta,\nu,\mu,\lambda_1,\lambda_2\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(-1-\mu-\nu^T g\left(X_i,\theta\right)\right)+\lambda_1\left(\sum_{i=1}^{n}\exp\left(-\mu-\nu^T g\left(X_i,\theta\right)\right)-e\right)\\ &\quad+\lambda_2^T\left(\sum_{i=1}^{n}\exp\left(-1-\mu-\nu^T g\left(X_i,\theta\right)\right)g\left(X_i,\theta\right)\right).\end{aligned}$$

Calculation shows $\nabla L=n^{-1}\sum_{i=1}^{n}\Psi\left(X_i \mid \theta,\nu,\mu,\lambda_1,\lambda_2\right)$. Hence, the same argument supporting empirical likelihood case validate the lemma in exponentially tilted empirical likelihood case.

The Cressie-Read case can follow exactly the same procedure in exponentially tilted empirical likelihood case. □

This lemma offer a better numerical scheme to get the maximum generalized empirical likelihood estimator than the original definition. In original definition, for each iteration of θ, one need to solve nonlinear equations to get Lagrange multiplier ν, which will introduce another iteration. By using the first order condition as in this lemma, we can use usual nonlinear equation solver to get the result in a single layer iteration. The benefit of Lemma 4 also enlighten a quick solution on asymptotic properties of maximum generalized empirical likelihood estimator.

Indeed, it is trivial to see,

$$\begin{aligned} E\Psi^{\mathrm{EL}}(X \mid \theta_0, 0) &= 0, \\ E\Psi^{\mathrm{ET}}(X \mid \theta_0, 0, -1, 0) &= 0, \\ E\Psi^{\mathrm{CR}}(X \mid \theta_0, 0, 1, 0, 1) &= 0. \end{aligned}$$

Hence, maximum generalized empirical likelihood estimator can also be interpreted as an ordinary M-estimator defined by Ψ function. The consistency and asymptotic normality can be easily extracted from ordinary M-estimator theory. However, in order to expand the posterior around maximum generalized empirical likelihood estimator, we need a slight stronger asymptotic property called law of iterative logarithm. Particularly, the consistency requires the following condition.

ASSUMPTION 7. *add later consistency*

The assumptions we need come from He and Wang (1995). We state them as follows.

ASSUMPTION 8. *Let* $\eta = (\theta, \nu, \mu, \lambda_1, \lambda_2)$. *Let* $\psi(\eta) = E\Psi(X \mid \eta)$, $u(x, \eta, d) = \sup_{|\tau-\eta|\leq d} |\Psi(x \mid \tau) - \Psi(x \mid \eta)|$, *where* $|\cdot|$ *takes sup-norm* $|\eta| = \max_{1\leq j\leq p} |\eta_j|$. *Let* η_n *satisfy*

$$\frac{1}{\sqrt{n \log\log n}} \sum_{i=1}^{n} \Psi(X_i \mid \eta_n) \to 0, \text{ a.s.}.$$

The following conditions guarantee law of iterative logarithm.

1 For each fixed $\eta \in H$, $\Psi(x \mid \eta)$ is square integrable in η and separable in sense of Doob: there is zero measure set N and a countable subset of $H' \subset H$, such that for every open set $U \subset H$, and every closed interval A, the sets $\{x : \Psi(x \mid \eta) \in A, \forall \eta \in U\}$ and $\{x : \Psi(x \mid \eta) \in A, \forall \eta \in U \cap H'\}$ differ by a subset of N.
   (i) There is a $\eta_0 \in H$, such that $\psi(\eta_0) = 0$, and ψ has a non-singular derivative at η_0.
   (ii) There exist positive numbers $a, b, c, d, \alpha, \beta$, and d_0 such that $\alpha \geq \beta > 2$, and
      (i) $|\psi(\eta)| \geq a|\eta - \eta_0|$, for $|\eta - \eta_0| \leq d_0$,
      (ii) $Eu(x, \eta, d) \leq bd$, for $|\eta - \eta_0| + d \leq d_0$,
      (iii) $Eu^{\alpha}(x, \eta, d) \leq cd^{\beta}$, for $|\eta - \eta_0| + d \leq d_0$.
   (iii) $|\eta_n - \eta_0| \leq d_0$ almost surely as n goes infinity.

Actually, the second condition is merely requiring we have a theoretical target, and the fourth condition is automatically satisfied if we have consistency. Immediately by the theorem in He and Wang (1995), we get the law of iterative logarithm for both maximum generalized empirical likelihood estimator and the Lagrange multipliers. Now we are ready to state another lemma to tie the empirical likelihood weights and likelihood together with following condition on the unknown distribution.

ASSUMPTION 9. *Let* $(k_{j1}, k_{j2}, \ldots, k_{jp}) \in \mathbb{N}^p$, *and* $\sum_{i=1}^{p} k_{ji} = k_j \leq K + 4$, $j = 1, \ldots, l$, *then*

$$E\left(\prod_{j=1}^{l} \frac{\partial^{k_l} g(X, \theta_0)}{\partial\theta_1^{k_{l1}} \partial\theta_2^{k_{l2}} \cdots \partial\theta_p^{k_{lp}}}\right),$$

exists and finite.

	MGELE ETEL	MDE ETEL	MGELE CR	MDE CR
N=20	0.1394	0.1320	0.2130	0.1408
N=50	0.0759	0.0759	0.1867	0.0766
N=100	0.0536	0.0536	0.2099	0.0539
N=200	0.0392	0.0401	0.2214	0.0396
N=500	0.0274	0.0298	0.1824	0.0274

LEMMA 5. *Let $\omega_i(\theta)$ be the unnormalized empirical weights from empirical likelihood, that is $\omega_i(\theta) = n\hat{w}_i(\theta)$ in empirical likelihood or Cressie-Read case, and $\omega_i(\theta) = \exp(-\nu(\theta) g(X_i,\theta))$ in exponentially tilted empirical likelihood. Use the same notation in Assumption 6. Under the Assumption 6, 7, 8 and 9, then for any $k \leq K+4$, any number $s \in \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \omega_i\left(\tilde{\theta}\right)^s \prod_{j=1}^{l} \frac{\partial^{k_l} g(X,\theta_0)}{\partial\theta_1^{k_{l1}}\partial\theta_2^{k_{l2}}\cdots\partial\theta_p^{k_{lp}}} = E\left(\prod_{j=1}^{l} \frac{\partial^{k_l} g(X,\theta_0)}{\partial\theta_1^{k_{l1}}\partial\theta_2^{k_{l2}}\cdots\partial\theta_p^{k_{lp}}}\right) \text{ a.s.}.$$

PROOF. By the similar proof in Owen (2010), we have

$$\lim_{n\to\infty} \max_{1\leq i\leq n} n^{-1/3} g(X_i,\theta_0) = 0 \text{ a.s.}.$$

By law of iterative logarithm on Lagrange multiplier ν, there exists some constant C_1

$$\overline{\lim_{n\to\infty}} \frac{\sqrt{n}\nu\left(\tilde{\theta}\right)}{\sqrt{2\log\log n}} = C_1 \text{ a.s.}.$$

Hence

$$\overline{\lim_{n\to\infty}} \nu^T\left(\tilde{\theta}\right) \max_{1\leq i\leq n} g\left(X_i,\tilde{\theta}\right) = o\left(\sqrt{\frac{2\log\log n}{n}} \times n^{1/3}\right) = o\left(\frac{\sqrt{2\log\log n}}{n^{1/6}}\right) = 0.$$

So $\nu^T\left(\tilde{\theta}\right) g\left(X_i,\tilde{\theta}\right)$ are uniformly going to zero. Hence $\omega_i\left(\tilde{\theta}\right)$ are uniformly going to 1 and the lemma holds. □

This lemma supports consistency of empirical sample moment to true moment and further justifies the intuition to use empirical likelihood as a valid likelihood.

Newey and Smith (2004) proposes a similar estimator replacing the objective function in Definition 1 with divergence measure when one calculates empirical weights. Their definition coincides with us when the divergence measure is empirical likelihood. Although, their starting point is generalized method of moment estimator and the intuition is slight different from us, the simulation hardly differs the performance. In Table **??**, we present the simulation result.

3.5. old stuff have not been modified

LEMMA 6. *$\mathrm{d}^2\tilde{l}\left(\tilde{\theta}\right)/\mathrm{d}\theta^2 < 0$ where $\tilde{l}(\theta) = n^{-1}\sum_{i=1}^{n}\log\hat{w}_i(\theta)$ where $\hat{w}_i$ is either $\hat{w}_i^{\mathrm{EL}}$, $\hat{w}_i^{\mathrm{ET}}$ or $\hat{w}_i^{\mathrm{CR}}$ $(i=1,2,\ldots,n)$.*

PROOF. We begin with $\tilde{l}(\theta) = n^{-1}\sum_{i=1}^{n}\log\hat{w}_i^{\mathrm{EL}}(\theta) = -\sum_{i=1}^{n}\log[1+\nu(X_i-\theta)] - \log n$. Hence by (3.2.2) and (3.2.3),

$$\frac{\mathrm{d}\tilde{l}(\theta)}{\mathrm{d}\theta} = \frac{1}{n}\nu\sum_{i=1}^{n}\frac{1}{1+\nu g(X_i\theta)}\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta} - \frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i\theta)}{1+\nu g(X_i\theta)}\frac{\mathrm{d}\nu}{\mathrm{d}\theta} = \frac{1}{n}\nu(\theta)\sum_{i=1}^{n}\frac{1}{1+\nu g(X_i\theta)}\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}.$$

Thus

$$\left.\frac{\mathrm{d}^2\tilde{l}(\theta)}{\mathrm{d}\theta^2}\right|_{\theta=\tilde{\theta}} = -\frac{\left(\sum_{i=1}^n \mathrm{d}g\left(X_i,\tilde{\theta}\right)/\mathrm{d}\theta\right)^2}{n\sum_{i=1}^n g\left(X_i,\tilde{\theta}\right)^2} < 0.$$

Next we consider $\tilde{l}(\theta) = n^{-1}\sum_{i=1}^n \log \hat{w}_i^{\mathrm{ET}}(\theta) = -\left(\nu n^{-1}\sum_{i=1}^n g(X_i,\theta) + \log\sum_{i=1}^n \exp\left(-\nu g(X_i,\theta)\right)\right)$. Then

$$\begin{aligned}\frac{\mathrm{d}\tilde{l}(\theta)}{\mathrm{d}\theta} &= -\frac{\mathrm{d}\nu}{\mathrm{d}\theta}\frac{1}{n}\sum_{i=1}^n g(X_i,\theta) \\ &\quad + \frac{\sum_{i=1}^n \exp\left(-\nu g(X_i,\theta)\right)\left(\mathrm{d}\nu/\mathrm{d}\theta g(X_i,\theta) + \nu\,\mathrm{d}g(X_i,\theta)/\mathrm{d}\theta\right)}{\sum_{i=1}^n \exp\left(-\nu g(X_i,\theta)\right)} \\ &= -\frac{\mathrm{d}\nu}{\mathrm{d}\theta}\frac{1}{n}\sum_{i=1}^n g(X_i,\theta) + \nu\frac{\sum_{i=1}^n \exp\left(-\nu g(X_i,\theta)\right)\mathrm{d}g(X_i,\theta)/\mathrm{d}\theta}{\sum_{i=1}^n \exp\left(-\nu g(X_i,\theta)\right)} \\ &\quad -\nu n^{-1}\sum_{i=1}^n \frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}.\end{aligned}$$

Thus

$$\left.\frac{\mathrm{d}^2\tilde{l}(\theta)}{\mathrm{d}\theta^2}\right|_{\theta=\tilde{\theta}} = -\frac{\mathrm{d}\nu}{\mathrm{d}\theta}\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{d}g\left(X_i,\tilde{\theta}\right)}{\mathrm{d}\theta} = -\frac{\left(\sum_{i=1}^n \mathrm{d}g\left(X_i,\tilde{\theta}\right)/\mathrm{d}\theta\right)^2}{n\sum_{i=1}^n g\left(X_i,\tilde{\theta}\right)^2} < 0.$$

Finally, for CR, $\tilde{l}(\theta) = n^{-1}\sum_{i=1}^n \log \hat{w}_i^{\mathrm{CR}}(\theta) = -\left[n(\lambda+1)\right]^{-1}\sum_{i=1}^n \log\left(\mu + \nu g(X_i,\theta)\right)$. Then by (3.3.4),

$$\frac{\mathrm{d}\tilde{l}(\theta)}{\mathrm{d}\theta} = .$$

Thus

$$\left.\frac{\mathrm{d}^2\tilde{l}(\theta)}{\mathrm{d}\theta^2}\right|_{\theta=\overline{X}} = -\frac{\left(\sum_{i=1}^n \mathrm{d}g\left(X_i,\tilde{\theta}\right)/\mathrm{d}\theta\right)^2}{n\sum_{i=1}^n g\left(X_i,\tilde{\theta}\right)^2}.$$

□

Let $b = \left[\left(n^{-1}\sum_{i=1}^n \mathrm{d}g\left(X_i,\tilde{\theta}\right)/\mathrm{d}\theta\right)^2 / \left(n^{-1}\sum_{i=1}^n g\left(X_i,\tilde{\theta}\right)^2\right)\right]^{-1/2}$. Now we prove the main result in the next section.

3.6. Assumptions

1 In multivariate empirical likelihood settings, matrix Z has full column rank with probability 1.

2 $E\prod_{s=1}^K (X_{ij_s} - \theta_{j_s,0}), (j_1, j_2, \ldots, j_K) \subset \{1, 2, \ldots, K\}$, where θ_0 is the true expectation of X, exists and finite.

3 The support of prior contains H_n, and $\rho(\theta) \in C^K$.

4 With probability 1 in P_X^n, $n^{-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ is a positive definite matrix.

3.7. main result

Let $\overline{X}$ be the sample mean. Let $\nu = \nu(\theta)$ be Lagrange multiplier obtained from 3.2.3 which are smooth functions of θ by in appendix. Let $\hat{l} = \hat{l}_n(\theta) = n^{-1}\sum_{i=1}^{n} \ln \hat{w}_i(\theta)$ be the logarithm of empirical likelihood. Let P_X^n be the underlying probability measure for sample $X_1, X_2, \ldots, X_n$.

For multivariate empirical likelihood, let

$$Z = \begin{pmatrix} X_{11}-X_{21} & X_{12}-X_{22} & \cdots & X_{1j}-X_{2j} & \cdots & X_{1p}-X_{2p} \\ X_{11}-X_{31} & X_{12}-X_{32} & \cdots & X_{1j}-X_{3j} & \cdots & X_{1p}-X_{3p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{i1}-X_{j1} & X_{i2}-X_{l2} & \cdots & X_{ij}-X_{lj} & \cdots & X_{ip}-X_{lp} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{p-1,1}-X_{p1} & X_{p-1,2}-X_{p2} & \cdots & X_{p-1,j}-X_{pj} & \cdots & X_{p-1,p}-X_{pp} \end{pmatrix}.$$

Let K be any positive integer and $K \geq 3$. For multivariate case in empirical likelihood, let B be the Cholesky decomposition of negative of the Hessian matrix

$$-\left.\frac{\partial^2 \hat{l}}{\partial\theta^T \partial\theta}\right|_{\theta=\overline{X}} = B^T B.$$

Let $Y = \sqrt{n}B\left(\theta - \overline{X}\right)$, define differential operators

$$\delta_i = \frac{1}{i!}\left(Y^T B^{-T}\nabla\right)^i,$$

where ∇ is the gradient operator.

We expand the prior around the sample mean $\overline{X}$ and have

$$\rho_K(\theta) = \rho\left(\overline{X}\right) + \delta_1\rho\left(\overline{X}\right)n^{-\frac{1}{2}} + \delta_2\rho\left(\overline{X}\right)n^{-1} + \cdots + \delta_K\rho\left(\overline{X}\right)n^{-\frac{K}{2}},$$

where $\delta_i\rho\left(\overline{X}\right)$ are results from applying the differential operators prior density then evaluating at $\overline{X}$. We abbreviate $\delta_i\rho\left(\overline{X}\right), i = 2, 3, \ldots K$ as $\delta_i\rho$.

Let H_n be the convex hull of samples.

Let $I_{i,h} = \left\{(m_{3,i}, m_{4,i}, \ldots, m_{K+3,i}) \in \mathbb{N}^K : \sum_{u=3}^{K+3} m_{u,i} = i, \sum_{u=3}^{K+3} m_{u,i}(u-2) = h\right\}$.

Let $P_K(A, n)$ be the expansion polynomial

$$\begin{aligned} & P_K(A,n) \\ = & \left(\rho\left(\overline{X}\right)\int_{A\cap H_n}\exp\left(-\frac{Y^TY}{2}\right)\mathrm{d}Y\right)n^{-\frac{1}{2}} + \left(\int_{A\cap H_n}\exp\left(-\frac{Y^TY}{2}\right)\left(\delta_1\rho + \rho\left(\overline{X}\right)\delta_3\hat{l}\right)\mathrm{d}Y\right)n^{-1} \\ & + \sum_{h=2}^{K}\left\{\int_{A\cap H_n}\exp\left(-\frac{Y^TY}{2}\right)\left[\delta_h\rho\right.\right. \\ & \left.\left. + \sum_{j=0}^{h-1}\delta_j\rho\sum_{\frac{h-j}{K+1}\leq i\leq h-j}\frac{1}{i!}\sum_{I_{i,h-j}}\binom{i}{m_{3,i}, m_{4,i}, \ldots, m_{K+3,i}}\prod_{u=3}^{K+3}\left(\delta_u\hat{l}\right)^{m_{u,i}}\right]\mathrm{d}Y\right\}n^{-\frac{h+1}{2}} \\ = & \int_{A\cap H_n}\exp\left(-\frac{Y^TY}{2}\right)\sum_{h=0}^{K}\alpha_h(Y,n)\,n^{-\frac{h}{2}}\,\mathrm{d}n^{-\frac{1}{2}}Y, \end{aligned}$$

where $\alpha_h(Y, n)$ are corresponding coefficients before each order of n.

THEOREM 1 (main theorem). *Under the assumptions 2, 1, 4 and 3, there exist a positive constant M_1 and a large integer N_1, such that for any Borel set $A \subset \mathbb{R}^p$, and any $n > N_1$*

$$\left| \int_{A \cap H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y - P_K(A,n) \right| \leq M_1 n^{-\frac{K+2}{2}}, \text{ a.s.}.$$

This theorem can be used to prove many asymptotic results

THEOREM 2 (Asymptotic expansion). *Under the same assumptions as theorem 1, there exist a positive constant M_2 and a large integer N_2, such that for any Borel set $A \subset \mathbb{R}^p$ and any $n > N_2$,*

(3.7.1)

$$\left| \Pi\left(B\left(\theta - \overline{X}\right) \in A | X_1, X_2, \ldots, X_n\right) - \Phi_p(A|H_n) - \sum_{i=1}^{K} \gamma_i(A,n)\, n^{-\frac{i}{2}} \right| \leq M_1 n^{-\frac{K+1}{2}}, \text{ a.s.}.$$

PROOF. By theorem 1, we have

$$\left| \int_{A \cap H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y - P_K(A,n) \right| \leq M_1 n^{-\frac{K+2}{2}}, \tag{3.7.2}$$

and

$$\left| \int_{H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y - P_K(H_n,n) \right| \leq M_1 n^{-\frac{K+2}{2}}. \tag{3.7.3}$$

By definition

$$\Pi\left(B\left(\theta - \overline{X}\right) \in A | X_1, X_2, \ldots, X_n\right) = \frac{\int_{A \cap H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y}{\int_{H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y}.$$

We know that all terms in equation (3.7.2) and equation (3.7.3), are almost surely bounded by some constant C_1 for all $n > N_1$, then there exists a constant

$$\begin{aligned}
& \left| \Pi\left(B\left(\theta - \overline{X}\right) \in A | X_1, X_2, \ldots, X_n\right) - \frac{P_K(A,n)}{P_K(H_n,n)} \right| \\
\leq\ & \left| \frac{\int_{A \cap H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y}{\int_{H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y} - \frac{P_K(A,n)}{P_K(H_n,n)} \right| \\
=\ & \left| \frac{\int_{A \cap H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y - P_K(A,n)}{\int_{H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y} \right. \\
& \left. - \frac{\left(\int_{H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y - P_K(H_n,n)\right) P_K(A,n)}{\int_{H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y P_K(H_n,n)} \right| \\
\leq\ & \frac{1}{\left| \int_{H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y \right|} \left(\left| \int_{A \cap H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y - P_K(A,n) \right| \right. \\
& \left. + \left| \int_{H_n} \prod_{i=1}^{n} \hat{w}_i(\theta)\, \pi(\theta)\, \mathrm{d}n^{-\frac{1}{2}} Y - P_K(H_n,n) \right| \left| \frac{P_K(A,n)}{P_K(H_n,n)} \right| \right) \\
\leq\ & \frac{1}{C_1} \left(M_1 n^{-\frac{K+2}{2}} + M_1 n^{-\frac{K+2}{2}} \frac{C_2}{C_1} \right) = \frac{M_1}{C_1} \left(1 + \frac{C_1}{C_2} \right) n^{-\frac{K+2}{2}}.
\end{aligned} \tag{3.7.4}$$

Now we find the quotient series of $P_K(A,n)/P_K(H_n,n)$. Let

$$\frac{P_K(A,n)}{P_K(H_n,n)} = \sum_{i=0}^{\infty} \gamma_i(A,n)\, n^{-\frac{i}{2}},$$

then by the product rule of series we have the coefficients $\gamma_i(A,n)$ are determined by

$$\int_{A\cap H_n} \exp\left(-\frac{Y^TY}{2}\right) \alpha_h(Y,n)\, \mathrm{d}Y = \sum_{j=0}^{h} \int_{H_n} \exp\left(-\frac{Y^TY}{2}\right) \alpha_j(Y,n)\, \mathrm{d}Y \gamma_{h-j}(A,n).$$

Through simple calculation, we can find first two items of $\gamma_i(A,n)$ is

$$\begin{aligned}
\gamma_0(A,n) &= \frac{\rho(\overline{X}) \int_{A\cap H_n} \exp\left(-\frac{Y^TY}{2}\right) \mathrm{d}Y}{\rho(\overline{X}) \int_{H_n} \exp\left(-\frac{Y^TY}{2}\right) \mathrm{d}Y} = \frac{\int_{A\cap H_n} \exp\left(-\frac{Y^TY}{2}\right) \mathrm{d}Y}{\int_{H_n} \exp\left(-\frac{Y^TY}{2}\right) \mathrm{d}Y} = \Phi_p(A|H_n), \\
\gamma_1(A,n) &= \frac{\int_{A\cap H_n} \exp\left(-\frac{Y^TY}{2}\right) \left(\delta_1\rho + \rho(\overline{X})\, \delta_3\hat{l}\right) \mathrm{d}Y}{\int_{H_n} \exp\left(-\frac{Y^TY}{2}\right) \mathrm{d}Y} \\
&\quad - \frac{-\int_{H_n} \exp\left(-\frac{Y^TY}{2}\right) \left(\delta_1\rho + \rho(\overline{X})\, \delta_3\hat{l}\right) \mathrm{d}Y \Phi_p(A|H_n)}{\int_{H_n} \exp\left(-\frac{Y^TY}{2}\right) \mathrm{d}Y} \\
&= \frac{\int_{A\cap H_n} \left(\delta_1\rho + \rho(\overline{X})\, \delta_3\hat{l}\right) \varphi_p(Y)\, \mathrm{d}Y}{\Phi_p(H_n)} - \frac{\int_{H_n} \left(\delta_1\rho + \rho(\overline{X})\, \delta_3\hat{l}\right) \varphi_p(Y)\, \mathrm{d}Y}{\Phi_p(H_n)} \Phi_p(A|H_n),
\end{aligned}$$

where φ_p is the density of dimension p standard normal distribution, Φ_p is the probability or conditional probability of dimension p standard normal distribution. By the discussion following lemma 10, we know that all γ_i are almost surely uniformly bounded for all large n. Then there exists a constant M_2, such that

$$\left| \frac{P_K(A,n)}{P_K(H_n,n)} - \Phi_p(A|H_n) - \sum_{i=1}^{K} \gamma_i(A,n)\, n^{-\frac{i}{2}} \right| \le M_2 n^{-\frac{K+1}{2}}. \tag{3.7.5}$$

Combine equation (3.7.4) and equation (3.7.5), we get equation (3.7.1). □

We will show detail proof of the this theorem in appendix. The proof are similar to Johnson (1970), with some modification to apply to empirical likelihood framework.

Corollary 1 (Asymptotic normality).

3.8. application

3.8.1. Simulation study. For simplicity, we simulate the situation where $p = 1$. In this case the first two terms in equation (3.7.1) will be $\Phi_1\left((-\infty,\xi] \mid \left(X_{(1)}, X_{(n)}\right)\right)$

and

$$\begin{aligned}\gamma_1\left((-\infty,\xi],n\right) &= \frac{\rho'\left(\overline{X}\right)\hat{\sigma}\int_{(-\infty,\xi]\cap\left(X_{(1)},X_{(n)}\right)} y\varphi_1(y)\,\mathrm{d}y + \rho\left(\overline{X}\right)\hat{\sigma}^3\hat{l}^{(3)}\left(\overline{X}\right)\int_{(-\infty,\xi]\cap\left(X_{(1)},X_{(n)}\right)} y^3\varphi(y)\,\mathrm{d}y}{\Phi\left(\left(X_{(1)},X_{(n)}\right)\right)} \\ &\quad - \frac{\rho'\left(\overline{X}\right)\hat{\sigma}\int_{\left(X_{(1)},X_{(n)}\right)} y\varphi_1(y)\,\mathrm{d}y + \rho\left(\overline{X}\right)\hat{\sigma}^3\hat{l}^{(3)}\left(\overline{X}\right)\int_{\left(X_{(1)},X_{(n)}\right)} y^3\varphi(y)\,\mathrm{d}y}{\Phi\left(\left(X_{(1)},X_{(n)}\right)\right)} \\ &\quad \Phi_1\left((-\infty,\xi]\,|\,\left(X_{(1)},X_{(n)}\right)\right),\end{aligned}$$

where $X_{(i)}$ are order statistics, and

$$\hat{l}^{(3)} = \frac{2n^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^3}{\hat{\sigma}^6}.$$

3.9. discussion

CHAPTER 4

On the Empirical Likelihood Option Pricing

4.1. Introduction

Since the seminal works by Black and Scholes (1973) and Merton (1973), option valuation methodologies have been extensively developed. The Black-Scholes model has become one of the most well-known discoveries in the finance literature, which relates the cross-sectional properties of option prices with the underlying asset?returns distribution. However, Rubinstein (1985), Melino and Turnbull (1990) pointed out several limitations in the Black-Scholes model due to the strong assumptions, such as non-normality of the returns, stochastic volatility (implied volatility smile), jumps and others. Both parametric and nonparametric approaches have been proposed to deal with these issues.

Scott (1987), Hull and White (1987) and Wiggins (1987) extended the Black and Scholes model and allowed the volatility to be stochastic. Heston (1993) developed a closed-form solution for option pricing with the underlying asset volatility being stochastic. Duan (1995) proposed a GARCH option pricing model in an attempt to explain some systematic biases associated with the Black-Scholes model. Later Heston and Nandi (2000) provided a closed-form solution for option pricing with the underlying asset volatility following GARCH(p,q) process. Bates (1996), Bakshi, Cao and Chen (1997) derived an option pricing model with stochastic volatility and jumps. Kou (2002) provided a solution to pricing the option with the double exponential jumps diffusion process. Carr and Madan (1999) introduced the fast Fourier transform approach to option pricing given a specified characteristic function of the return, which provides an efficient computational algorithm to calculate the option prices. For further reference, see Duffie et al. (2000), Bakshi and Madan (2000) and Carr and Madan (2009) among others. All these methods are parametric based, which assume a parametric form of either the distribution of the underlying assets returns or the characteristic function of the underlying assets?returns.

Nonparametric approaches have also been proposed to capture the underlying asset and option price data to reconstruct the structure of the diffusion process. For example, Hutchinson, Lo and Poggio (1994) applied neural network techniques to price the derivatives. Ait-Sahalia and Lo (1998) used kernel regression to fit the state-price density implicitly in option pricing. Ait-Sahalia (1996) proposed a nonparametric pricing estimation procedure for interest rate derivative securities under the assumption that the unknown volatility is independent of time. Stutzer (1996) adopted the canonical valuation method, which incorporats the no-arbitrary principle embodied in the formula for calculating the expectation of the discounted value of assets under the risk-neutral probability distribution.

One of the most important nonparametric methodologies is the empirical likelihood, which conducts likelihood-based statistical inference by profiling a nonparametric likelihood. See Owen (1988, 1990, 2001), DiCiccio and Romano (1989) and Hall and La Scala (1990) for instance. For the application of EL method to time series, see Mykland (1995), Chuang and Chan (2002) and Ling and Chan (2006) among others. Kitamura (1997) introduced a blockwise empirical likelihood method for weakly dependent time series. Nordman, Sibbertsen and Lahiri (2007) modified the blockwise methods to cope with various dependence structures and achieve better finite sample performance. Yau (2012) studied the application of EL to long-memory time series.

In this paper, we implement the EL method to price the derivative or options under risk neutral measure. We firstly construct an empirical probability constraint using the historical holding period return time series observations, without assuming the distribution family of the returns. On the other hand, we view the derivative / option price as the parameter of interest directly in the empirical likelihood optimization procedure. An empirical likelihood based estimate of the parameter (e.g. call option price) is obtained and the asymptotic properties of the EL ratio are studied. We further introduce a block-wise empirical likelihood procedure for the weakly dependent processes. Monte Carlo simulation and empirical results for S&P 500 index option are discussed.

The remaining of the paper is organized as follows. Section 2 provides a detail empirical likelihood procedure in option pricing. Asymptotic properties are discussed and a robust confidence interval is constructed. Section 3 provides some empirical performance of the empirical likelihood option pricing including both Monte Carlo simulation and S&P 500 Index options. Section 4 concludes the paper with discussions.

4.2. Empirical Likelihood on Option Pricing

Let $P(t)$ be the underlying asset price at time t, $D(t)$ the future dividend at time t, $r(s,t)$ the gross risk-free interest rate during time s and t with $r(t,t)=1$, $\mathcal{P}$ the physical probability measure, and $\mathcal{Q}$ the risk-neutral probability measure (See Huang and Litzenberger (1988)), under which the price process plus the accumulated dividends are martingales after normalization if no arbitrage exists in the pricing systems. To be specific, the latter leads to the following pricing formula.

$$\begin{aligned} P(t) &= E^{\mathcal{Q}}[\frac{P(T)+\sum_{s=t}^{T} D(s)r(s,T)}{r(t,T)}] \\ &= E^{\mathcal{P}}[\frac{P(T)+\sum_{s=t}^{T} D(s)r(s,T)}{r(t,T)}\frac{d\mathcal{Q}}{d\mathcal{P}}]. \end{aligned} \tag{4.2.1}$$

Here $\frac{d\mathcal{Q}}{d\mathcal{P}}$ is the Radon-Nykodym density of the marginal measure. One can price an option or a derivative security by evaluating the expected discounted value of it under the $\mathcal{Q}$. For example, the call option price with strike price K and expiring date T is given by

$$C(t,T) = \frac{E^{\mathcal{Q}} max[P_T - K, 0]}{r(t,T)} \tag{4.2.2}$$

The following subsection illustrates the idea of estimating $C(t,T)$ through EL coupled with the change-of-measure constraint.

4.2.1. The Estimating Procedure. Suppose historical data is available in the format of $\{(P(t), D(t)), t = -1, -2, ..., -H\}$. A nonparametric way of estimating the option price could be built on approximating $\mathcal{Q}$ by a discrete distribution supported on the observed value of option price, namely $HPR(-i-T,-i)/r(-i-T,-i)$, $1 \leq i \leq H-T$ with the corresponding probability denoted by π_i. Here $HPR(s,t)$ is the holding period return between time s and t. If there is no dividend, $HPR(-i-T,i) = P(-i)/P(-i-T)$. Then (4.2.1) can be approximated by

$$1 = \sum_{i=1}^{H-T} \frac{HPR(-i-T,-i)}{r(-i-T,-i)}\pi_i \tag{4.2.3}$$

Correspondingly, we can estimate the option price by approximating (4.2.2) by

$$\hat{C}(t,T) \quad = \quad \sum_i \frac{max[P_i(T)-K,0]}{r(t,T)}\pi_i. \tag{4.2.4}$$

Note that the choice of π_i subjecting to (4.2.3) is not unique. Stutzer (1996) use the idea of maximum entropy, namely maximizing $\sum_{i=1}^{H-T} \pi_i log\pi_i$ subject to (4.2.3). Here we adopt the Empirical likelihood idea (Owen, 1988) by changing the objective function from entropy to empirical likelihood, namely maximizing $\sum_{i=1}^{H-T} log\pi_i$. Meanwhile, By noticing that the sequence $HPR(-i-T,-i)/r(-i-T,-i)$, $1 \leq i \leq H-T$ possesses a reasonable amount of dependence, we suggest adopting the block wise version of the algorithm as follows. Group the data into Q blocks with with length M is the length of the moving block. Set L to be the step size of the moving block. We obtain block weight π_i^* by maximizing $\sum_{i=1}^{H-T} log\pi_i^*$ subject to

$$1 = \sum_{i=1}^{Q} \pi_i^* \left[\frac{1}{M} \sum_{j=1}^{M} \frac{HPR(-i*L-j-T,-i*L-j)}{r(-i*L-j-T,-i*L-j)} \right] \tag{4.2.5}$$

Then estimate the option price by

$$C = \sum_{i=1}^{Q} \frac{max[P_i(T)-K,0]}{r(t,T)}\pi_i^* \tag{4.2.6}$$

This blocking idea has been studied by Kitamura (1997). He argued that using block-wise methods has a much better empirical performance for weakly dependent processes by moving average the noise terms. The estimation procedure in the spirit of Kitamura (1997) would be slightly different, i.e.

$$max_{C,\pi_i^*} \sum_{i=1}^{Q} log\pi_i^* \tag{4.2.7}$$

subject to (4.2.5) and (4.2.6), and the maximizing C will be our estimator. Actually, Peng (2015) has shown that these two approaches yield the same asymptotic property. In our simulation below, we will adopt the second method since it is well known and there is existing package for implementation. Particularly, Qin and Lawless (1994) provided a Lagrangean with multipliers approach to solve the above mentioned optimization problem. We can either apply the numerical optimization

process or derive the solution similar to Qin and Lawless (1994). For more details about the Lagrangean optimization or the basic properties of the empirical likelihood procedure, see Owen (1990) and Qin and Lawless (1994).

4.2.2. Asymptotic Properties. In this subsection, we discuss some basic asymptotic properties of the option price with respect to the empirical likelihood process (Equation (6) / (7)), which helps us to understand the asymptotic distribution of our estimate and conduct further inference.

THEOREM 3. Consider

$$f(HPR_t, C) = (\frac{max[P_i(T) - K, 0]}{r(t,T)} - C, \frac{HPR(-t-T,t)}{r(t-T,t)} - 1)^T$$

and further assume that:
(i) the derivative price (C) is in a compact set Θ;
(ii) C_0 is unique solution of $E(f(HPR_t, C)) = 0$;
(iii) For sufficient small $\delta > 0$ and $\eta > 0$,

$$E[sup_{C^* \in O(C,\delta)} ||f(HPR, C^*)|| < \infty$$

for all $C \in \Theta$;
(iv) If a sequence of C_j, $j = 1, 2, ...$ converges to some C as $j \to \infty$, $f(HPR_t, C_j)$ converges to $f(HPR_t, C)$ for all HPR_t except on a null set, which may vary with C;
(v) C_0 is an interior point of Θ;
(vi) $Var(H^{-\frac{1}{2}} \sum_{i=1}^{H} f(HPR_i, C_0)) \to S > 0$;
(vii) For block-wise empirical likelihood approach, we further assume the weak dependent condition, $\sum_{k=1}^{\infty} \alpha_X(k)^{1-1/d} < \infty$ for some constant $d > 1$. And we require additional assumptions,

$$E||f(HPR_t, C_0)||^{2d} < \infty, \text{ for } d > 1$$

$$Esup_{C^* \in O(C_0,\delta)} ||f(HPR_t, C^*)||^{2+\epsilon} < K, \text{ for some } \epsilon > 0.$$

Then,

$$LR_0 = 2 \sum_{i=1}^{Q} log(1 + \gamma(\hat{C})^T f(HPR_i, \hat{C})) \to_{dist.} \chi_1^2$$

where K is a finite number, $\gamma(\hat{C})$ is the Lagrange multiplier vector and Q is the total number of states. Particularly for non-blockwise empirical likelihood case (i.e. Equation (6)), $Q = H - T$.

Theorem 1 provides an asymptotic distribution of the likelihood ratio LR_0, which can be further applied to inference of the estimate. We omit the detailed proof here.[1] For independent observations of HPR_i, we only require the assumptions (i)-(vi) to have the asymptotic property of the likelihood ratio; and for weak-dependent observations of HPR_i, assumption (vii) is additionally required.

4.3. Empirical Results

In this section, we will first compare our method with Black-Scholes model through simulation and then conduct an empirical analysis on the option pricing for the S&P 500 index call options.

[1] Our proof is based on Theorems 1 & 2 in Kitamura (1997).

4.3.1. Monte Carlo Simulation. Following Hutchinson et al. (1994), Ait-Sahalia and Lo (1995) and Stutzer (1996), we generate a geometric Brownian motion process with a 10 percent drift and 20 percent annualized volatility. Firstly we simulated 2 years of historical daily stock returns with $253 \times 2 = 506$ observations. We repeat 200 samples, and for each sample, three different prices are calculated. 1. the estimated price by the empirical likelihood option pricing procedure; 2. the estimated price by the Black-Scholes model with historical volatility; 3. the actual price by the Black-Scholes model with actual volatility. The performance of the first two prices are compared based on the mean absolute percentage error (MAPE) with respective to the third price, which is considered to be minicing the true price. The comparison is make at different price-to-strike price ratios (i.e. $P/X = \frac{9}{10}, 1, \frac{9}{8}$) and different expiration dates (i.e. $T = \frac{1}{13}, \frac{1}{4}, \frac{1}{2}$).

[Table 1]

Table 1 provides the simulation performance: Panel A reports the MAPE of the empirical likelihood (EL) option price, and Panel B reports the MAPE of the historical volatility based Black Scholes price. In the perfect Black-Scholes world, the Black-Scholes formula using the historical volatility outperforms the empirical likelihood option pricing methodology. This is because the Black-Scholes formula only requires the second moment information and 506 observations can provide a very good estimate of the second moment, but empirical likelihood method will automatically capture the higher order moment information, which will not benefit in pricing the options in the perfect Black-Scholes world.

We are also interested in the accuracy of the empirical likelihood option pricing for different moneyness and days-to-maturity. The empirical likelihood option pricing method provides very good performance in pricing the in-the-money options with small MAPE. However, the MAPE are very significant for out-of-the-money options. At-the-money option pricing error is in between. On the other hand, the pricing errors have different patterns for in-the-money, at-the-money and out-of-the-money options. For in-the-money and at-the-money options, the fewer days to maturity, the smaller pricing errors are. For out-of-the-money option, the fewest days to maturity case has the largest pricing error, with a possible reason that the price magnitude of the out-of-the-money options with very few days to maturity is already very small.

4.3.2. S&P 500 Index Options. We also implement the empirical likelihood option pricing method in pricing the S&P 500 index options. The daily return data is from Center for Research in Security Prices (CRSP) and the option data is from OptionMetrics. The daily return data is from Jan 2011 to Dec 2012. We use the year 2011 daily return data as the formation period, and then test its performance in the year 2012 daily index options pricing comparing with the historical volatility based Black Scholes model and the true values. We only keep the options which have the Moneyness closest to 1 and days to maturity between 15 to 50.

[Figure 1]

Figure 1 shows the time series of the option prices. The red line is true value of the market daily close price, the green line is the empirical likelihood option price and the blue line is the Black Scholes option price using the historical volatility. Due to the stock price movement, the true option prices vary from 1.5 to 3.7. However, the historical volatility based Black Scholes option prices are consistently

TABLE 1. Monte Carlo Simulation in a Black-Scholes Market

This table reports the mean absolute percentage error (MAPE) of the empirical likelihood (EL) option price to the ideal Black-Scholes price (Panel A), the historical volatility based Black Scholes price to the ideal Black-Scholes price (Panel B) for different combination of the relative exercise prices (P/K) and time to expiration date. The price dynamics follow the Geometric Brownian Motion with $\mu = 0.1$ and $\sigma = 0.2$. The relative exercise prices (P/K) are chosen as Rubinstein (1985), Stutzer (1996). The time to expiration date are $1/13, 1/4, 1/2$ years, respectively.

Panel A:

	Hist Var vs Ideal BS	Days to Maturity		
		1/13	1/4	1/2
Moneyness (P/K)	9/10	0.139	0.066	0.046
	1	0.022	0.020	0.019
	9/8	0.000	0.003	0.005

Panel B:

	EL vs Ideal BS	Days to Maturity		
		1/13	1/4	1/2
Moneyness (P/K)	9/10	0.724	0.514	0.537
	1	0.088	0.149	0.230
	9/8	0.003	0.025	0.058

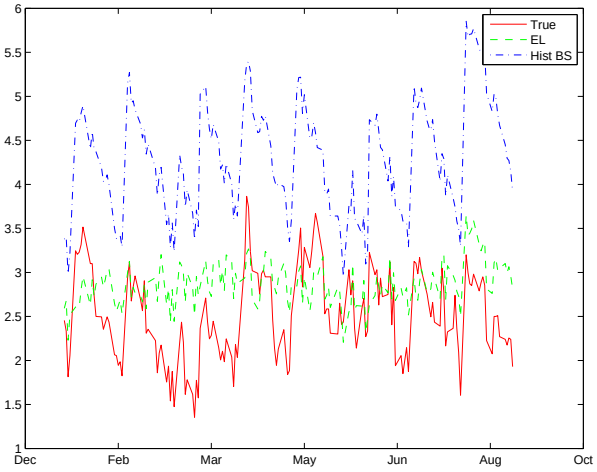
overpriced for the at-the-money call options, as is documented in Hull and White (1987). the contrast, our EL option prices are a lot closer to the true option market prices. This is because our empirical likelihood methodology also captures the high order moment information, while the historical volatility based Black Scholes option model only captures the second moment information.

4.4. Conclusion

In this paper, we introduce an empirical likelihood method to price the derivatives or options under risk-neutral measure $\mathcal{Q}$. Monte Carlo simulations show that our new pricing methodology performs reasonably well for the at-the-money and in-the-money options. We also apply our empirical likelihood option pricing method to S&P 500 index options, and the results demonstrate that our method outperforms the historical volatility based Black Scholes model. This is due to the advantage of EL method in capturing the high order moment information.

FIGURE 4.4.1. Comparison of the S&P 500 index option prices and EL option prices

This figure shows the time series of three S&P 500 index option prices. We only keep the options have the Moneyness closest to 1 and days to maturity between 15 to 50. The red line is true value of the market daily close price, the green line is the empirical likelihood option price and the blue line is the Black Scholes option price using the historical volatility.



CHAPTER 5

Approximate Bayesian Computation via Sufficient Dimension Reduction

5.1. Introduction

There are two main objectives of this article. First, we want to provide some theoretical results related to the currently emerging topic of approximate Bayesian computation (ABC). The second is to show some connectivity between ABC and another important emerging topic of research, namely, sufficient dimension reduction (SDR). While the latter has surfaced primarily in the frequentist's domain of research, it is possible to tie it with ABC as well. In particular, we want to show how ABC can be carried through nonlinear SDR.

Modern science invokes more and more Byzantine stochastic models, such as stochastic kinetic network (Wilkinson (2011)), differential equation system (Picchini (2014)) and multi-hierarchical model (Jasra et al. (2012)), whose computational complexity and intractability challenge the application of classical statistical inference. Traditional maximum likelihood methods will malfunction when the evaluation of likelihoods becomes slow and inaccurate. Lack of analytical form of the likelihood also undermines the usage of Bayesian inferential tools, such as Markov chain Monte Carlo (MCMC), Laplace approximation (Tierney and Kadane (1986)), variational Bayes (Jaakkola and Jordan (2000)) and posterior expansion (Johnson (1970), Zhong and Ghosh).

The ABC methodology stems from the observation that the interpretability of the candidate model usually leads to an applicable sampler of data given parameters, and ingeniously circumvents the evaluation of likelihood functions. The idea behind ABC can be summarized as follows:

Algorithm 2 Idea of ABC

1 Sample parameters θ_i from the prior distribution $\pi(\theta)$;
2 Sample data Z_i based on the model $f(z \mid \theta_i)$;
3 Compare the simulated data Z_i and the observed data $X_{i,\text{obs}}$, to accept or reject θ_i.

Rubin (1984) first mentioned this idea and Tavaré et al. (1997) proposed the first version of ABC, which studying population genetics. The prototype of ABC in recent research was given in Pritchard et al. (1999), where the comparison of two data sets was simplified to a comparison of summary statistics S and the accept-reject decision was made up to a certain error tolerance. We can view this algorithm as a modified version of accept-reject algorithm (Robert and Casella (2013)). The posterior is sampled by altering the frequency of the proposal distribution, that is,

Algorithm 3 Prichard's Modified ABC

1 Sample parameters θ_i from the prior distribution $\pi(\theta)$;
2 Sample data Z_i based on the model $f(z \mid \theta_i)$;
3 Accept θ_i if $\rho(S(Z_i), S(X_{\text{obs}})) \leq \varepsilon$, for some metric ρ .

the prior. Now the full posterior distribution is approximated by the following two steps (Fearnhead and Prangle (2012)):

$$\pi(\theta \mid X_{\text{obs}}) \approx \pi(\theta \mid S_{\text{obs}}) \approx \pi(\theta \mid S_{\text{sim}} \in O(S_{\text{obs}}, \varepsilon)), \tag{5.1.1}$$

where $O(S_{\text{obs}}, \varepsilon)$ means a neighborhood defined by the comparison measure ρ and tolerance level ε. We may note that the first approximation is exact when S is sufficient. Allowing the summary statistics to vary in an acceptable range sacrifices a little accuracy in exchange for a significant improvement in computational efficiency, which makes the algorithm more practical and user-friendly.

Pursuant to Algorithm 3, there are multiple generalizations in the statistical literatures. Marjoram et al. (2003) introduced MCMC-ABC algorithm to concentrate the samples in high posterior probability region, thereby increasing the accept rate. Noisy ABC, proposed by Wilkinson (2013), makes use of all the prior samples by assigning kernel weights instead of hard-threshold accept-reject mechanism and hence reduces the computational burden. This perspective is corroborated in Fearnhead and Prangle (2012) by convergence of Bayesian estimators. When the dependence structure between hierarchies is intractable, ABC filtering technique innovated by Jasra et al. (2012) comes to the rescue. Later in Dean et al. (2014), a consistency argument is established for the specific case of hidden Markov models. Moreover, many ABC algorithm above can be easily coded in a parallel way, and hence take advantages of modern CPU, GPU structures. This feature makes ABC algorithms extremely time-saving against long-established, looping-based MCMC and MLE algorithms.

Despite the fruitful results on ABC both from applied and theoretical points of view. However, there exist only a handful of papers which focus on the effect of the choice of summary statistics on the approximation quality. The quintessential case is the summary statistics are sufficient, and the resultant ABC sampler produces exact samples from the true posterior distribution when ε goes to zero. Nevertheless, in a labyrinthine model, it is difficult to extract sufficient statistics, except for some very special case, such as exponential random graph models (e.g. Grelaud et al. (2009)). Joyce and Marjoram (2008) proposed a concept called $\varepsilon-$sufficient to quantify the effect of statistics. Nonetheless, this property is also difficult to verify in complicated models. If we are interested only in model selection, Prangle et al. (2014) designs a semi-automatic algorithm to construct summary statistics via logistic regression. And laterly, Marin et al. (2014) gives sufficient conditions on summary statistics in order to choose the right model based on the Bayes factor. They advocate that the ideal summary statistics are ancillary in both model candidates. One of our contribution comes from the mathematical analysis of the consequence of conditioning the parameters of interest on consistent statistics and intrinsically inconsistent statistics, and appraises the efficiency of the posterior approximation based on the former. Generally speaking, using consistent statistics results in right concentration of the approximate posterior, while less efficiency of statistics leads to less efficiency of approximation. One byproduct is our theorem

vindicates the usage of the posterior mean as summary statistics as in Fearnhead and Prangle (2012).

In addition to the pure theoretical contribution, we also extend the two-step algorithm in Fearnhead and Prangle (2012) in a more flexible and nonparametric way for automatic constructing summary statistics. We borrow the idea from another thriving topic, namely sufficient dimension reduction (SDR). The motivation of SDR which generalizes the concept of sufficient statistics is to estimate a transformation φ, either linear or nonlinear, such that

$$Y \perp\!\!\!\perp X \mid \varphi(X). \tag{5.1.2}$$

The first SDR method titled sliced inverse regression dates back to Li (1991), followed by principle Hessian direction in Li (1992) and also by Cook and Weisberg (1991), and Cook (1998). As we step in the era of big data, this idea leads to a sea of papers on both linear and nonlinear, predictor and response. Among the more recent work, we refer to Cook and Li (2002), Xia et al. (2002), Li et al. (2005), Li and Dong (2009), Wu (2008), Yeh et al. (2009), Su and Cook (2011) and Su and Cook (2012). The association between SDR and ABC relies on the shared mathematical formulation. If we think θ as the response and X as the predictor, then an ideal summary statistics $S(X)$ will give

$$\theta \perp\!\!\!\perp X \mid S(X).$$

This simple observation offers raison d'etre to use existing SDR methods in constructing summary statistics. The employment of dimension reduction methods in our algorithm is different from that in Blum et al. (2013). In Blum et al. (2013), dimension reduction methods, such as best subset selection, projection techniques and regularization approaches, are applied to reduce the dimension of existing summary statistics, but here, we try to reduce the size of the original data. Particularly in our paper, we incorporate the principal support vector machine for nonlinear dimension reduction given in Li et al. (2011) into ABC, which uses the principal component of support vectors in reproducing kernel Hilbert space (RKHS) as a nonparametric estimator of φ.

The outline of remaining sections is as follows. Section 5.2 contains asymptotic results on the partial posterior. We gradually relax the restriction on summary statistics and investigate the relationship between the partial posterior and the full posterior. As a side result, we give a lemma building a bridge between the recent prior free inferential model (Martin and Liu (2013) and Martin and Liu (2015)) and traditional Bayesian inference. Section 5.3 elicits a new ABC algorithm which automatically produces summary statistics through nonlinear SDR. A simulation result is provided in this section as well. Section 5.4 briefly discusses the results and points out some possible future generalizations.

5.2. Asymptotic Properties of Partial Posterior

Suppose $X_1, \ldots, X_n \mid \theta$ are i.i.d. with common PDF $f(x \mid \theta)$, and there exists a true but unknown value θ_0. Without loss of generality, we assume $\theta \in \mathbb{R}$, and all probability density functions are with respect to the Lebesgue measure. For illustration purpose, we define the following terminology.

DEFINITION 2 (Partial Posterior). Let $S = S(X_1, \ldots, X_n)$ be statistics of the data. Given a prior $\pi(\theta)$, we call the distribution

$$\pi(\theta \mid S) \propto \pi(\theta)\, g(S \mid \theta)$$

the partial posterior, where $g(S \mid \theta)$ is the probability density function of statistic $S(X_1, \ldots, X_n)$ derived from the data density, and correspondingly,

$$\pi(\theta \mid X_1, \ldots, X_n) \propto \pi(\theta)\, f(X_1, \ldots, X_n \mid \theta)$$

is called the full posterior.

From equation (5.1.1), partial posterior significantly reduces the complexity of full posterior by replacing the dependence on full data by lower dimensional statistics S. If the partial posterior deviates from the full posterior too much, then no matter how delicately we sample from $\pi(\theta \mid S_{\text{sim}} \in O(S_{\text{obs}}, \varepsilon))$, how small ε we choose, the resultant samples would not behave like ones drawn from the original full posterior, which makes the subsequent Bayesian analysis fragile and unreliable. Therefore, theoretical connection between some easily verifiable properties and asymptotic behaviour of partial posterior is of relevance. In particular, we want to study consistency and asymptotic normality of our Bayesian procedures. The following theorems try to demonstrate the connection between the asymptotic behaviour of summary statistics and that of partial posterior. We start from the most popular statistics, the maximum likelihood estimator (MLE) of θ.

THEOREM 3. *Let $\hat{\theta}$, the MLE of θ, be a strongly consistent estimator, and let $\hat{I}$ be the observed Fisher information evaluated at $\hat{\theta}$, and the full posterior holds Bernstein–von Mises theorem. Then for any $\varepsilon > 0$, and any t, the partial posterior after conditioned on $\hat{\theta}$ satisfies*

$$\lim_{n\to\infty} P\left(\sqrt{n\hat{I}}\left(\theta - \hat{\theta}\right) \leq t \mid \hat{\theta} \in O(\theta_0, \varepsilon)\right) = \Phi(t), \text{ a.s.}.$$

PROOF. See C.1. □

REMARK 1. There is a slight difference between

$$\lim_{n\to\infty} P\left(\sqrt{n\hat{I}}\left(\theta - \hat{\theta}\right) \leq t \mid \hat{\theta} \in O(\theta_0, \varepsilon)\right) = \Phi(t), \text{ a.s.}.(P_{\theta_0})$$

and

$$\lim_{n\to\infty} P\left(\sqrt{n\hat{I}}\left(\theta - \hat{\theta}\right) \leq t \mid \hat{\theta}\right) = \Phi(t), \text{ a.s.}.$$

By definition,

$$P\left(\sqrt{n\hat{I}}\left(\theta - \hat{\theta}\right) \leq t \mid \hat{\theta}\right) = \lim_{\varepsilon\to 0} \frac{P\left(\sqrt{n\hat{I}}\left(\theta - \hat{\theta}\right) \leq t, \hat{\theta} \in O(s, \varepsilon)\right)}{P\left(\hat{\theta} \in O(s, \varepsilon)\right)}. \tag{5.2.1}$$

The result of Theorem 3 can only be used to prove

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty} P\left(\sqrt{n\hat{I}}\left(\theta - \hat{\theta}\right) \leq t \mid \hat{\theta} \in O(\theta_0, \varepsilon)\right) = \Phi(t), \text{ a.s.},$$

switching order of limits in equation (5.2.1).

REMARK 2. The definition of $P\left(\theta \mid \hat{\theta} \in O(\theta_0, \varepsilon)\right)$ is different from the approximation $P\left(\theta \mid \hat{\theta} \in O\left(\hat{\theta}_{\text{obs}}, \varepsilon\right)\right)$. In former case, $\hat{\theta}$ is evaluated at $X_1, \ldots, X_n \sim$

$\pi(x \mid \theta_0)$, the observed data, while the latter evaluates $\hat{\theta}$ at $Z_1, \ldots, Z_m \sim \pi(z \mid \theta)$, the simulated data.

By assumptions , the asymptotic distribution of the full posterior is still normal, and we have

$$\begin{aligned} & \sup_{t\in\mathbb{R}} \left| P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right) \le t \mid \hat{\theta} \in O(\theta_0, \varepsilon)\right) - P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right) \le t \mid X_1, \ldots, X_n\right)\right| \\ \le & \sup_{t\in\mathbb{R}} \left| P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right) \le t \mid \hat{\theta} \in O(\theta_0, \varepsilon)\right) - \Phi(t)\right| + \sup_{s\in\mathbb{R}} \left| P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right) \le s \mid X_1, \ldots, X_n\right) - \Phi(s)\right| \\ \le & 2\varepsilon \to 0, (\text{as} n \to \infty). \end{aligned}$$

Hence, we can informally say that two random variables $\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right) \mid \hat{\theta}$ and $\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right) \mid X_1, \ldots, X_n$ are close in distribution. Note that both random variables asymptotically center at consistent MLE, and hence will eventually concentrate at θ_0. Meanwhile, the scale factors in both random variables are $\sqrt{n\hat{I}}$, which ensures the same square root credible intervals. In this sense, we feel that the partial posterior conditioned on MLE has the same efficiency as the full posterior. Later theorems will tell us that if the summary statistics are not efficient, the corresponding partial likelihood will have a different scale factor, and thus lose efficiency and result in a larger credible interval.

A slightly modified proof of Theorem 3 can be used to support the posterior mean as a summary statistic in Fearnhead and Prangle (2012) and we still have a similar result, namely

$$\lim_{n\to\infty} P\left(\sqrt{n\hat{I}}\left(\theta - E(\theta \mid X_1, \ldots, X_n)\right) \le t \mid E(\theta \mid X_1, \ldots, X_n) \in O(\theta_0, \varepsilon)\right) = \Phi(t), \text{ a.s.}.$$

The key fact to support the assert above comes from Ghosh and Liu (2011), that is, the higher order closeness of the posterior mean and the MLE,

$$\lim_{n\to\infty} \sqrt{n}\left(E(\theta \mid X_1, \ldots, X_n) - \hat{\theta}\right) = 0, \text{ a.s.}. \tag{5.2.2}$$

Indeed, any estimator who has the same or higher order of closeness to MLE will work as an efficient summary statistic.

Theorem 3 can be generalized to more intricate models. The following example shows the same phenomenon in data generated from a Markov process.

EXAMPLE 2. Immigration-emigration process is a crucial model in survival analysis and can be viewed as a special case of mass-action stochastic kinetic network (Wilkinson (2011)). The model is defined by a birth procedure and a death procedure during an infinitesimal time interval, namely,

$$P(X(t+\mathrm{d}t) = x_1 \mid X(t) = x_0) = \begin{cases} \lambda\,\mathrm{d}t + o(\mathrm{d}t), & x_1 = x_0 + 1, \\ \mu x_0\,\mathrm{d}t + o(\mathrm{d}t), & x_1 = x_0 - 1, \\ 1 - \lambda\,\mathrm{d}t - \mu x_0\,\mathrm{d}t + o(\mathrm{d}t), & x_1 = x_0. \end{cases}$$

Assume that we observe full data in the time interval $[0, T]$. Let $T_i, i = 1, \ldots, n$ be the event times and $X_i = X(T_i), i = 1, \ldots, n$. Let X_0 be initial population, $T_0 = 0$, $T_{n+1} = T$. Then by Gillespie's algorithm, the likelihood is proportional to

$$\lambda^{r_1} \exp(-\lambda T)\, \mu^{r_2} \exp(-\mu A_T),$$

where r_1 and r_2 are number of events corresponding to immigration and emigration, and

$$A_T = \int_0^T X(t)\,\mathrm{d}t.$$

The MLEs are

$$\hat{\lambda} = \frac{r_1}{T}, \hat{\mu} = \frac{r_2}{A_T},$$

and they are strongly consistent estimators of λ and μ when T goes to infinity. By the computation in C.2.1, we have the partial posterior density function of $\sqrt{T}(\mu - \hat{\mu})$ conditioned on $\hat{\mu}$, r_1 and T given by

$$\lim_{T\to\infty} \pi\left(\sqrt{T}(\mu - \hat{\mu}) = t \mid \hat{\mu}, r_1, T\right) = \frac{\hat{\mu}}{\sqrt{2\pi\hat{\lambda}}} \exp\left(-\frac{\hat{\lambda}}{\hat{\mu}^2} t^2\right), \text{ a.s.}.$$

The MLE seems to be a perfect surrogate for the full data. However, in many cases, use of MLE is prohibitive due to heavy computational burden, particularly when the likelihood function is intractable. This is when the ABC comes on stage. M-estimator is a generalization of the MLE, which is also consistent and asymptotically normal under mild conditions. Many M-estimators can be easily calculated, especially some moment estimators. To give an idea of the nature of approximation, we consider the following examples.

Example 3. Gamma distribution can be used to model hazard functions in survival analysis. The shape parameter of gamma distribution determines the trend of hazard and hence is a vital parameter to estimate. Assume $X_1, \ldots, X_n \sim \text{Gamma}(\alpha, \beta)$, where we know the scale parameter β, but not the shape parameter α. The MLE of α is the solution of

$$-\log\Gamma(\alpha) - \alpha\log\beta + (\alpha - 1)\sum_{i=1}^{n}\log X_i - \frac{\sum_{i=1}^{n} X_i}{\beta} = 0,$$

which involves repeated evaluation of the gamma function in search of the root. A simple M-estimator $\tilde{\alpha} = \overline{X}/\beta$ is derived from its mean equation,

$$\sum_{i=1}^{n}(X_i - \alpha\beta) = 0.$$

Now we consider the partial posterior $\pi(\alpha \mid \tilde{\alpha})$, when the prior is $\pi(\alpha) \propto \exp(-\lambda\alpha)$. By the calculation in C.2.2, we show that the limit of cumulative probability function of $\sqrt{n}\tilde{\alpha}^{-1}(\alpha - \tilde{\alpha})$ given $\tilde{\alpha}$ is

$$\lim_{n\to\infty} P\left(\sqrt{n}\tilde{\alpha}^{-1}(\alpha - \tilde{\alpha}) \le t \mid \tilde{\alpha}\right) = \Phi(t), \text{ a.s.},$$

which means that the Bernstein-von Mises theorem holds for the partial posterior conditioned on the M-estimator $\tilde{\alpha}$. The scale factor of the partial posterior is $\sqrt{n}\tilde{\alpha}^{-1}$, which is smaller than that of the full posterior, $\sqrt{n\psi'(\alpha)}$, where $\psi(\alpha)$ is digamma function. That results in a larger credible interval based on the partial posterior.

Example 4. Another example is Laplace distribution with PDF

$$f_{\mu,\lambda}(t) = \frac{1}{2\lambda}\exp\left(-\frac{|t-\mu|}{\lambda}\right).$$

Here we want inference the location parameter μ holding λ fixed. The MLE is the the sample median and the moment estimator is the sample mean. Here we calculate the partial posterior based on sample mean. By the calculation in C.2.3, we find that the characteristic function of $\sqrt{n}\left(\mu - \overline{X}\right)$ converges to $\exp\left(-\lambda^2 t^2\right)$, which is the characteristic function of normal distribution.

Example 4 uses the following lemma which is of independent interest.

LEMMA 7. *Assume X has the same distribution as $h(Y, \theta)$, where h is a function one-to-one in θ and Y is a random variable independent of θ. Let $\theta = g(y, x)$ and $y = u(x, \theta)$ be the solutions of equation $x = h(y, \theta)$. Further assume $\partial u(x, \theta)/\partial x$ exists and is not equal to zero. Then the posterior distribution of θ conditioned on X under the uniform prior has the same distribution as $g(Y, x)$, where x is fixed.*

REMARK 3. Although not quite related to ABC, this lemma gives another interpretation of inferential model of Martin and Liu (2013) and Martin and Liu (2015). In their settings, Y is called unobserved ancillary variable, and $g(y, x)$ is $\Theta_x(u)$ in their notation. They claim that their procedure results in a distribution of θ without referring to a prior. However, by our lemma, this model is mathematically the same as a posterior given a uniform prior.

The following theorems are built upon the Theorem 2.1 in Rivoirard et al. (2012), which guarantees the asymptotic normality of linear functionals of nonparametric posterior. So we need all the assumptions in that theorem. Additionally, we need the following assumptions.

ASSUMPTION 10. *There is a neighbourhood $\theta \in O(\theta_0, \varepsilon)$ such that $\int_{\mathbb{R}} g(x, \theta_0)\, \pi(x \mid \theta)\, \mathrm{d}x$ is a continuous twice differentiable in θ and the second order derivative is bounded by some constant L.*

ASSUMPTION 11. *M-estimator $\tilde{\theta}$ and MLE $\hat{\theta}$ are both strongly consistent and asymptotically normal.*

ASSUMPTION 12. *Bernstein–von Mises theorem and posterior consistency hold for the full posterior of θ.*

ASSUMPTION 13. *For any $\theta \in \Theta$, $E_{\theta_0} \log f(X \mid \theta) \leq E_{\theta_0} \log f(X \mid \theta_0)$.*

Now we can articulate the theorem.

THEOREM 4. *Under the Assumptions 10, 11, 12, 13, and conditions of Theorem 2.1 in Rivoirard et al. (2012), for any ε and t,*

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}\left(\theta - \tilde{\theta}\right)}{\sqrt{\tilde{V}}} \leq t \middle| \tilde{\theta} \in O(\theta_0, \varepsilon)\right) = \Phi(t), \text{ a.s.},$$

where $\tilde{V} = V_0 / G_1\left(\tilde{\theta}, \tilde{\theta}\right)^2$ is Godambe information.

PROOF. See C.3. □

Using similar arguments as Theorem 3, the partial posterior $\sqrt{n\tilde{V}^{-1}}\left(\theta - \tilde{\theta}\right) \mid \tilde{\theta}$ is asymptotically close in distribution to the full posterior $\sqrt{n\hat{I}}\left(\theta - \hat{\theta}\right) \mid$

$X_1, \ldots, X_n$. Since both the M-estimator and the MLE are strongly consistent, the partial posterior still concentrates around the right θ_0, but now the asymptotic $\alpha-$level credible interval based on the partial posterior, namely

$$\left(\tilde{\theta} - Z_{\alpha/2}\sqrt{\frac{\tilde{V}}{n}}, \tilde{\theta} + Z_{\alpha/2}\sqrt{\frac{\tilde{V}}{n}}\right),$$

will be larger than that based on the full posterior,

$$\left(\hat{\theta} - Z_{\alpha/2}\sqrt{\frac{\hat{I}^{-1}}{n}}, \hat{\theta} + Z_{\alpha/2}\sqrt{\frac{\hat{I}^{-1}}{n}}\right),$$

where $Z_{\alpha/2}$ is $(1-\alpha/2)$ quantile of standard normal distribution, because the Godambe information $\tilde{V}^{-1}$ is typically no larger than Fisher information $\hat{I}$. Hence, we lose efficiency if we condition the posterior on an inefficient estimator, which coincides our intuition.

For extreme tortuous models, even finding a consistent estimator can be quite hard. There are still some simple statistics which may be consistent to some functions of θ. Unless they are ancillary statistics, they always contain some information about the parameters of interest. Moreover, in real case, we use several statistics, each of which gives independent information of the full posterior. In the remainder of this section, we will mathematically qualify what independent information means and show that using more than one statistic will improve the efficiency.

Let S_i, $i = 1, \ldots, q$ be statistics of the sample. We make the following trivial assumptions.

ASSUMPTION 14. *The joint distribution of $S_1, \ldots, S_q$ converges in distribution to a multivariate normal distribution $N\left(h\left(\theta_0\right), n^{-1/2}\Sigma\left(\theta_0\right)\right)$. And each S_i converges to $h_i\left(\theta_0\right)$ almost surely. Further, assume $\Sigma\left(\theta_0\right)$ is positive definite.*

Assumption 14 characterizes the statement independent information. Because if $\Sigma\left(\theta_0\right)$ has a lower rank, then some of S_i can be expressed as a linear combination of other S_j asymptotically. Then the partial posterior can be reduced to partial posterior based solely on S_j.

In order to prove the theorem, we need some more technical assumptions.

ASSUMPTION 15. *Let $h\left(\theta_0\right)$ be a linear functional of the distribution function, that is*

$$h\left(\theta_0\right) = \int_{\mathbb{R}} g\left(x\right) f\left(x \mid \theta_0\right) \mathrm{d}x,$$

where $g\left(x\right) \in \mathbb{R}^q$.

ASSUMPTION 16. *Let $S = \left(S_1, \ldots, S_q\right)$, assume*

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g\left(X_i\right) - S\right) = 0, \text{ a.s.}.$$

and there exists a strongly consistent estimator $\tilde{\Sigma}$ of $\Sigma\left(\theta_0\right)$

Assumption 15 is a natural consequence when we apply some version of strong law of large numbers to prove convergence of statistics. Only Assumption 16 seems quite restrictive. Based on all these assumptions, the theorem describing the partial posterior conditioned on less informative statistics can be found as following.

THEOREM 5. *Under the Assumptions 14, 15, 16 and conditions of Theorem 2.1 in Rivoirard et al. (2012), for any vector* $a \in \mathbb{R}^q$,

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left| P\left(\frac{\sqrt{n} a^T \left(h\left(\theta\right) - S\right)}{\sqrt{a^T \tilde{\Sigma} a}} \leq t \middle| S \right) - \Phi\left(t\right) \right| = 0, \text{ a.s.}.$$

PROOF. See C.4. □

Theorem 5 extends the asymptotic results about M-estimators to more general statistics, particularly the intrinsically inconsistent statistics defined as follows.

DEFINITION 3 (Intrinsic Consistency). Let S be an non-ancillary statistic and converges to $h\left(\theta_0\right)$ almost surely. If $h\left(\cdot\right)$ is an one-to-one function and has an inverse function, then we say S is intrinsically consistent. Otherwise, we say S is intrinsically inconsistent.

If S is a one dimensional intrinsic inconsistent statistic, then Theorem 5 asserts the $(1-\alpha)$ asymptotic credible set based on the partial posterior is

$$\left\{ \theta : S - Z_{\alpha/2} \sqrt{\frac{\tilde{\Sigma}}{n}} \leq h\left(\theta\right) \leq S + Z_{\alpha/2} \sqrt{\frac{\tilde{\Sigma}}{n}} \right\}.$$

In an extreme case, when sample size n is large enough, such that $Z_{\alpha/2}/\sqrt{n} \approx 0$, the asymptotic credible interval by the full posterior would be close to the singleton $\left\{\hat{\theta}\right\}$. However, the credible set based on the partial posterior would be $\{\theta : h\left(\theta\right) = S\}$. By the definition of intrinsic inconsistency, h is not a one-to-one function. Then the set $\{\theta : h\left(\theta\right) = S\}$ would possibly hold multiple elements, hence larger than that from the full posterior. Again, in this sense, we perceive loss of efficiency due to conditioning the posterior on arbitrary statistics.

Another interesting use of Theorem 5 is a more pragmatic asymptotic assessment of effectiveness of including many statistics than that in Joyce and Marjoram (2008). In their settings, the effectiveness of summary statistics is measured by the difference between log-likelihoods, thus not operable when likelihood functions are intractable. On the other hand, our approach only requires the asymptotic behaviour of statistics, and the corresponding credible set with q statistics can be developed by Cremer device as

$$\left\{ \theta : n \left(h\left(\theta\right) - S\right)^T \tilde{\Sigma} \left(h\left(\theta\right) - S\right) \leq \chi^2_{1-\alpha,q} \right\},$$

where $\chi^2_{1-\alpha,q}$ is $(1-\alpha)$ quantile of chi-square distribution with degree of freedom q. To select summary statistics, we can compare the asymptotic credible sets with and without the current statistic. If the difference is small, then we can safely throw the current statistic away.

5.3. Approximate Bayesian Computation via Nonlinear Sufficient Dimension Reduction

In principle, almost all the existing dimension reduction methods are valid in estimating the summary statistics. However, there is a slight difference between the setting of SDR and ABC. In the theory of SDR, the independent assumption 5.1.2 must hold rigorously, which implies $Y \mid X$ has exactly the same distribution

Algorithm 4 Principal Support Vector Machine

1 (Optional) Marginally standardize data $X_1, \ldots, X_n$. The purpose of this step is so that the kernel κ treats different components of X_i more or less equally.

2 Choose kernel κ and the number of basis functions k (usually around $n/3 \sim 2n/3$). Compute $K = (\kappa(X_i, X_j))_{n\times n}$. Let $Q = I_n - J_n/n$, where J_n is the $n \times n$ matrix whose entries are 1. Compute largest k eigenvalues $\lambda_1, \ldots, \lambda_k$ and corresponding eigenvectors $w_1, \ldots, w_k$ of matrix QKQ. Let $\Psi = (w_1, \ldots, w_k)$ and $P_\Psi = \Psi\left(\Psi^T\Psi\right)^{-1}\Psi^T$ be the projection matrix onto Ψ.

3 Partition the response variable space Y into h slices defined by $y_1, \ldots, y_{h-1}$. For each $y_s, s = 1, \ldots, h-1$, define a new response variable $\tilde{Y}_{si} = I_{[Y_i \le y_s]} - I_{[Y_i > y_s]}$. Then solve the modified support vector machine problem as a standard quadratic programming

$$\min_\alpha -1^T\alpha + \frac{1}{4}\alpha^T \mathrm{diag}\left(\tilde{Y}_s\right) P_\Psi \mathrm{diag}\left(\tilde{Y}_s\right)\alpha,$$

subject to constraints

$$\begin{cases} 0 \le \alpha \le \lambda, \\ \tilde{Y}_s^T\alpha = 0, \end{cases}$$

where $\mathrm{diag}\left(\tilde{Y}_s\right)$ is a diagonal matrix using $\tilde{Y}_s$ as diagonal, λ is a hyperparameter in ordinary support vector machine. The coefficients of support vectors in RKHS are

$$c_s^* = \frac{1}{2}\left(\Psi^T\Psi\right)^{-1}\Psi^T \mathrm{diag}\left(\tilde{Y}_s\right)\alpha_s.$$

4 Let d be the target dimension. Compute the eigenvectors $v_1, \ldots, v_d$ of first largest d eigenvalues of the matrix $\sum_{s=1}^{h-1} c_s^* c_s^{*T}$. Let $V = (v_1, \ldots, v_d)$.

5 Let $K(x, X) = \left(\kappa(x, X_i) - n^{-1}\sum_{j=1}^n \kappa(x, X_j)\right)$ be a n dimensional vector. Then the estimated transformation $\hat{\varphi}(x) = V^T\left(\mathrm{diag}\left(\lambda_1, \ldots \lambda_k\right)\right)^{-1}\Psi^T K(x, X)$.

as $Y \mid S(X)$. However, by our Theorem 3, 4 and 5, the two distributions are only close in large but finite sample.

In our paper, we choose principal support vector machine in Li et al. (2011). Suppose we have a regression problem (Y_i, X_i), and search a nonlinear transformation $\varphi : \mathbb{R}^p \to \mathbb{R}^d$, such that $Y \perp\!\!\!\perp X \mid \varphi(X)$. Then the main steps in principal support vector machine are given in Algorithm 4 .

By slicing the response variable space, we discretize variation of Y. The support vector machine in Step 3 recognizes the robust separate hyperplanes. We will expect the variation of Y along the directions within hyperplanes to be negligible and that along the directions perpendicular to the hyperplanes the most part of covariation between Y and X is explained. The principal component analysis on the support vectors in Step 4 estimates the principal perpendicular directions and hence creates the sufficient directions in RHKS.

Algorithm 5 ABC via PSVM

1 Sample θ_i from the prior $\pi(\theta)$ and sample $X_{i1}, \ldots, X_{in}$ from the model $f(x \mid \theta_i)$.
2 View $(\theta_i, X_{i1}, \ldots X_{in})$ as a multivariate regression problem and reduce the dimension from n to d via principal support vector machine. Denote the estimated transformation as $\hat{S}(X_1, \ldots, X_n)$.
3 Either use existent samples in Step 1 or repeat it and get new sample. Calculate the estimated summary statistics $\hat{S}_i = \hat{S}(X_{i1}, \ldots, X_{in})$ on each set $X_{i1}, \ldots, X_{in}$ corresponding to prior samples θ_i. Also calculate $\hat{S}_{\text{obs}} = \hat{S}(X_1, \ldots, X_n)$ on the observed data set.
4 Based on the metric $\rho\left(\hat{S}_i, \hat{S}_{\text{obs}}\right)$, make the decision of accept or reject of θ_i.

Based on Algorithm 4, we formulate our two-step approximate Bayesian computation algorithm in Algorithm 5.

Algorithm 5 directly generalizes the semi-automatic ABC in Fearnhead and Prangle (2012). In their algorithm, the summary statistics are fixed as posterior means and recommended the estimation method is polynomial regression. Our algorithm relaxes the restriction on summary statistics and lets the data and nonparametric algorithm together find them adaptively. One significant difference between our algorithm and the conventional ABC is in Step 1, where each prior sample θ_i produces exact n simulated data, because the nonparametric estimator of statistics should take n arguments so that it can be evaluated at both observed data and simulated data.

Next, we will show a simple simulation example to illustrate our algorithm.

EXAMPLE 5. Autoregressive model with lag one, AR(1).

$$Y_i = \beta Y_{i-1} + \varepsilon.$$

Set $Y_1 = 1$ and number of observation is 100. Assume $\varepsilon \sim N\left(0, 0.5^2\right)$, true regression coefficient 0.6. We put uniform prior in $(-1, 1)$ on β, then the true posterior distribution is $N\left(\sum_{i=1}^{99} Y_i Y_{i+1} / \left(1 + \sum_{i=1}^{99} Y_i^2\right), \left(1 + \sum_{i=1}^{99} Y_i^2\right)^{-1}\right)$. Now we apply our algorithm with the target dimension $d = 1$ and slicing pieces $h = 4$ with the slicing parameters y_k as quartiles. The sample size from the prior is 1000, with $k = 100/2 = 500$. Kernel κ is chosen as Gaussian kernel $\kappa(x_i, x_j) = \exp\left(-10^{-5} \times |x_i - x_j|^2\right)$. Then the posterior density estimated from ABC samples are plotted in Fig. 5.3.1.

The slight skewness in Fig. 5.3.1 possibly due the the small sample size of the observed data. Another interesting result of this simulation is shown in Fig. 5.3.2. There is a strong linear relationship between the estimated summary statistic and MLE

$$\hat{\beta} = \frac{\sum_{i=1}^{99} Y_i Y_{i+1}}{\sum_{i=1}^{99} Y_i^2},$$

which is one of the most efficient summary statistics based on Theorems 3 and 4. Hence, our algorithm will automatically approach the most efficient summary statistics in a nonparametric way.

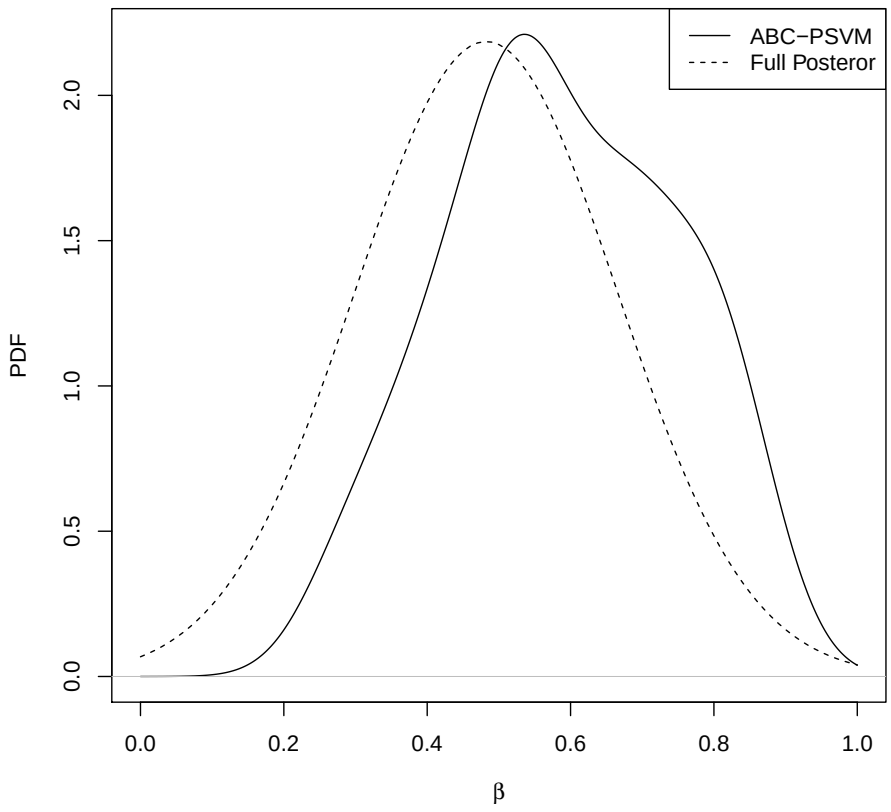


FIGURE 5.3.1. ABC Density vs True Posterior Density

5.4. Discussion

In this paper, we explore ABC both from a theoretical and computational points of view. The theory part architects the foundation of ABC by linking asymptotic properties of statistics to that of the partial posterior. The application part innovates the algorithm by virtue of bridging selection of summary statistics and SDR. However, although the theory in Li (1992) is very powerful and may be used as a theoretical guard of our algorithm, it heavily depends on the relation (5.1.2) holding rigorously. We do not know whether the result from principal support vector machine would be defunct if (5.1.2) is only valid in ε-sufficient way. Moreover, bringing in dimension reduction regression settings perhaps moderates the usage when there are multiple parameters of interest, and may need advance techniques such as envelope models of Su and Cook (2011, 2012).

Acknowledgment

Ghosh's research was partially supported by an NSF Grant.

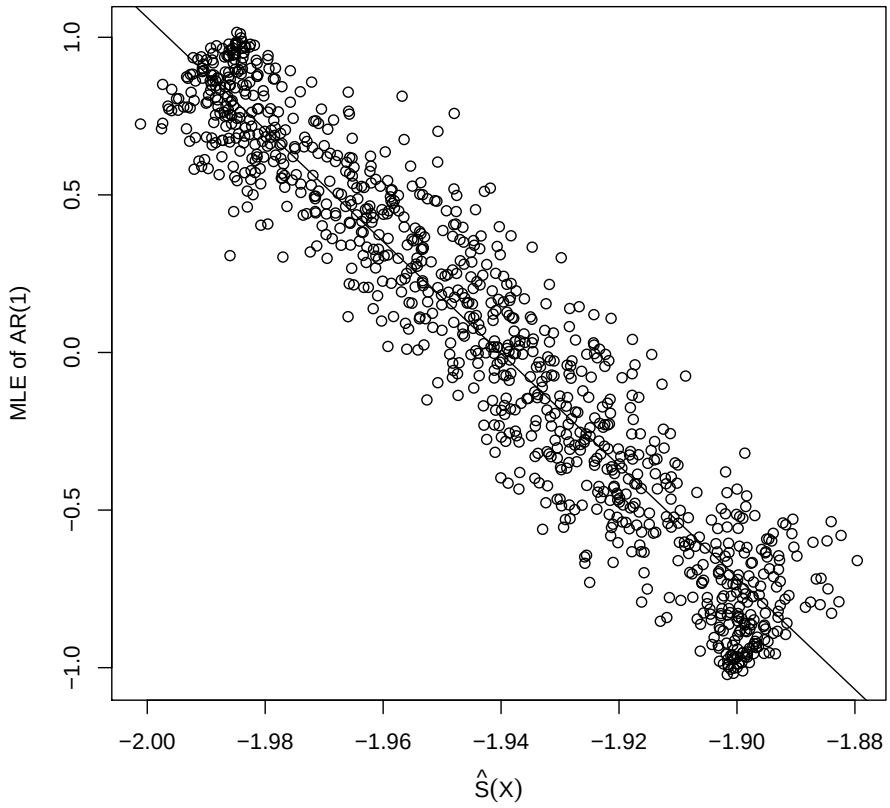


FIGURE 5.3.2. Estimated Summary Statistic vs MLE

APPENDIX A

Proof for Higher-Order Properties of Bayesian Empirical Likelihood: Univariate Case

A.1. Behavior Of Log Empirical Likelihood In The Tail

The Taylor expansion consists of expanding the log empirical likelihood and prior density around the mean and then control the tail part of the log empirical likelihood. In order to implement this idea, the tail part of $\tilde{l}(\theta)$ must vanish faster than the required polynomial order. In this section, we will show that indeed the tail part of $\tilde{l}(\theta)$ vanishes at an exponential rate.

LEMMA 5. Under Assumptions 1 and 2, for any $\delta_1 > 0$, there exist $\varepsilon_1 > 0$ and N_3, such that

$$\tilde{l}(\theta) - \tilde{l}\left(\tilde{\theta}\right) \leq -\varepsilon, \text{ a.s.},$$

for any $\left|b\left(\theta - \tilde{\theta}\right)\right| \geq \delta_1$ and $\theta \in H_n$, where $b = \left[\left\{n^{-1}\sum_{i=1}^{n} \mathrm{d}g\left(X_i, \tilde{\theta}\right)/\mathrm{d}\theta\right\}^2 / \left\{n^{-1}\sum_{i=1}^{n} g\left(X_i, \tilde{\theta}\right)^2\right\}\right]^{-1/2}$

PROOF. By Lemma 3, we know that $\tilde{\theta}$ is the unique maximizer. Therefore, for any $\theta \neq \tilde{\theta}$, $\tilde{l}(\theta) < \tilde{l}\left(\tilde{\theta}\right)$. The set

$$\left\{\theta : \left|b\left(\theta - \tilde{\theta}\right)\right| \geq \delta_1\right\} \cap H_n$$

is a compact set, and $\tilde{l}(\theta)$ is a continuous function. Hence there exists a $\theta^* \in \left\{\theta : \left|b\left(\theta - \tilde{\theta}\right)\right| \geq \delta_1\right\} \cap H_n$, such that for any $\theta \in \left\{\theta : \left|b\left(\theta - \tilde{\theta}\right)\right| \geq \delta_1\right\} \cap H_n$,

$$\tilde{l}(\theta) \leq \tilde{l}(\theta^*).$$

Therefore, $\tilde{l}(\theta) \leq \tilde{l}(\theta^*) < \tilde{l}\left(\tilde{\theta}\right)$, which is equivalent to

$$\tilde{l}(\theta) - \tilde{l}\left(\tilde{\theta}\right) \leq \tilde{l}(\theta^*) - \tilde{l}\left(\tilde{\theta}\right) < 0.$$

Let $\varepsilon_1 = \left\{\tilde{l}\left(\tilde{\theta}\right) - \tilde{l}(\theta^*)\right\}/2$, then we have

$$\tilde{l}(\theta) - \tilde{l}\left(\tilde{\theta}\right) \leq \tilde{l}(\theta^*) - \tilde{l}\left(\tilde{\theta}\right) < \varepsilon_1.$$

□

A.2. Higher-Order Derivatives

In order to expand around the mean, we need to control the remainder terms in the Taylor expansion, which involves the finiteness of higher-order derivatives of $\tilde{l}$.

LEMMA 6. Let

$$\omega_i(\theta) = \begin{cases} \{1+\nu(\theta)\, g(X_i,\theta)\}^{-1}, & \text{forempiricallikelihood,} \\ \exp\{-\nu(\theta)\, g(X_i,\theta)\}, & \text{forexponentiallytiltedempiricallikelihood,} \\ \{\mu(\theta)+\nu(\theta)\, g(X_i,\theta)\}^{-1/(\lambda+1)}, & \text{forCressie}-\text{Readempiricallikelihood.} \end{cases}$$

and

$$D = \begin{cases} n^{-1}\sum_{i=1}^n \omega_i^r g(X_i,\theta)^2, & \text{forempiricalandexponentiallytilted,} \\ n^{-1}\sum_{i=1}^n \omega_i^{\lambda+2}\left[g(X_i,\theta) - \left\{n^{-1}\sum_{i=1}^n \omega_i^{\lambda+2} g(X_i,\theta)\right\}\right]^2, & \text{forCressie}-\text{Read.} \end{cases}$$

Then under Assumptions 1 and 3, for any $k = 2, \ldots K+3$,

$$\frac{\mathrm{d}^k}{\mathrm{d}\theta^k}\tilde{l}(\theta) \quad = \quad D^{-r_k} P_k(M_1, M_2, \ldots, \nu, \mu),$$

where P_k are polynomial function , and all the $r_j < C(k)$, and $C(k)$ is some constant only depending on$k$, and M_j are the weighted average of higher order derivatives of g, i.e.

$$M_j = \frac{1}{n}\sum_{i=1}^n \omega_i^r \prod_l \frac{\mathrm{d}^l g(X_i,\theta)}{\mathrm{d}\theta^l},$$

l can be the same .

PROOF. From Lemma 2, we know that

$$\frac{\mathrm{d}\nu}{\mathrm{d}\theta} = D^{-1} P_1\left\{\frac{1}{n}\sum_{i=1}^n \omega_i^{r_1}\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}, \frac{1}{n}\sum_{i=1}^n \omega_i^{r_2} g(X_i,\theta), \ldots, \nu, \mu\right\}.$$

Moreover,

$$\begin{aligned} \frac{\mathrm{d}\tilde{l}^{\mathrm{EL}}(\theta)}{\mathrm{d}\theta} &= \sum_{i=1}^n \frac{1}{1+\nu g(X_i,\theta)}\left\{\frac{\mathrm{d}\nu}{\mathrm{d}\theta} g(X_i,\theta) + \nu\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}\right\} \\ &= \frac{\mathrm{d}\nu}{\mathrm{d}\theta}\frac{1}{n}\sum_{i=1}^n \omega_i(\theta)\, g(X_i,\theta) + \nu\frac{1}{n}\sum_{i=1}^n \omega_i(\theta)\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}, \end{aligned}$$

$$\frac{\mathrm{d}\tilde{l}^{\mathrm{ET}}(\theta)}{\mathrm{d}\theta} = -\frac{\mathrm{d}\nu}{\mathrm{d}\theta}\frac{1}{n}\sum_{i=1}^n g(X_i,\theta) - \nu\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta} + \nu\sum_{i=1}^n \hat{w}_i(\theta)\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta},$$

$$\begin{aligned} \frac{\mathrm{d}\tilde{l}^{\mathrm{CR}}(\theta)}{\mathrm{d}\theta} &= -\frac{1}{\lambda+1}\sum_{i=1}^n \omega_i^{\lambda+1}\left\{\frac{\mathrm{d}\mu}{\mathrm{d}\theta} + \frac{\mathrm{d}\nu}{\mathrm{d}\theta} g(X_i,\theta) + \nu\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}\right\} \\ &= -\frac{1}{\lambda+1}\left\{\frac{\mathrm{d}\mu}{\mathrm{d}\theta}\frac{1}{n}\sum_{i=1}^n \omega_i^{\lambda+1} + \frac{\mathrm{d}\nu}{\mathrm{d}\theta}\frac{1}{n}\sum_{i=1}^n \omega_i^{\lambda+1} g(X_i,\theta) + \nu\frac{1}{n}\sum_{i=1}^n \omega_i^{\lambda+1}\frac{\mathrm{d}g(X_i,\theta)}{\mathrm{d}\theta}\right\}. \end{aligned}$$

So for $k = 1$, the lemma holds. Assume for $k = n$, the lemma holds. Then for $k = n+1$,

$$\frac{\mathrm{d}^{n+1}\tilde{l}(\theta)}{\mathrm{d}\theta^{n+1}} \quad = \quad \frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\mathrm{d}^k\tilde{l}(\theta)}{\mathrm{d}\theta} = -r_k D^{-r_k-1}\left(\frac{\mathrm{d}D}{\mathrm{d}\theta}\sum_{i=1}^{k_n}\frac{\mathrm{d}P_k}{\mathrm{d}M_i}\frac{\mathrm{d}M_i}{\mathrm{d}\theta} + \frac{\mathrm{d}P_k}{\mathrm{d}\nu}\frac{\mathrm{d}\nu}{\mathrm{d}\theta} + \frac{\mathrm{d}P_k}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\theta}\right)$$

The partial derivative of P_k is still a polynomial. Also, D itself is a polynomial in $n^{-1}\sum_{i=1}^{n}\omega_i^r g\left(X_i,\theta\right)^2$, μ, and ν. Next

$$\begin{aligned}\frac{\mathrm{d}M_i}{\mathrm{d}\theta} &= \frac{1}{n}\sum_{i=1}^{n}\left\{r\omega_i^{r-1}\frac{\mathrm{d}\omega_i}{\mathrm{d}\theta}\prod_l\frac{\mathrm{d}^l g\left(X_i,\theta\right)}{\mathrm{d}\theta^l}+\omega_i^r\sum_{l_j}\prod_{l\neq l_j}\frac{\mathrm{d}^l g\left(X_i,\theta\right)}{\mathrm{d}\theta^l}\frac{\mathrm{d}^{l_j+1} g\left(X_i,\theta\right)}{\mathrm{d}\theta^{l_j+1}}\right\}\\ &= \frac{r}{n}\sum_{i=1}^{n}\omega_i^{r-1}\prod_l\frac{\mathrm{d}^l g\left(X_i,\theta\right)}{\mathrm{d}\theta^l}\frac{\mathrm{d}\omega_i}{\mathrm{d}\theta}+\sum_{l_j}\frac{1}{n}\sum_{i=1}^{n}\omega_i^r\prod_{l\neq l_j}\frac{\mathrm{d}^l g\left(X_i,\theta\right)}{\mathrm{d}\theta^l}\frac{\mathrm{d}^{l_j+1} g\left(X_i,\theta\right)}{\mathrm{d}\theta^{l_j+1}}.\end{aligned}$$

Similar to the calculation of the first order derivative of empirical log likelihood, we know the $\mathrm{d}\omega_i/\mathrm{d}\theta$ are polynomials involving terms like M_i. So for $k=n+1$, the higher order derivatives of empirical log likelihood are of a similar form. Hence, by mathematical induction, the lemma holds for all n. □

By Lemma 10, the higher-order derivatives of log empirical likelihood are rational functions of the sample moments of higher-order derivatives of g. We can anticipate that higher order derivatives of log empirical likelihood can be bounded in a small neighborhood of the true parameter when sample size is large , provided the population moments of higher-order derivatives of function g are finite. This we prove in the following lemma .

LEMMA 7. Under Assumptions 3 , 4 and 5 , there exist constants δ_2, C_3 and N_4 such that for any $\left|b\left(\theta-\tilde{\theta}\right)\right|\leq\delta_2$ and $n>N_4$, any $j=1,\ldots,k$,

$$\left|\frac{\mathrm{d}^j\tilde{l}\left(\theta\right)}{\mathrm{d}\theta^j}\right|\leq C_3. \tag{A.2.1}$$

PROOF. All ω_i in Lemma 10 are equal to 1 when evaluated at $\tilde{\theta}$. Under the assumption of finite moments, by strong law of large numbers, and strong consistency of the M-estimator $\tilde{\theta}$, we have

$$M_j=\frac{1}{n}\sum_{i=1}^{n}\prod_l\frac{\mathrm{d}^l g\left(X_i,\tilde{\theta}\right)}{\mathrm{d}\theta^l}\rightarrow E\left\{\prod_l\frac{\mathrm{d}^l g\left(X_1,\theta_0\right)}{\mathrm{d}\theta^l}\right\}<\infty,\text{ a.s.}.$$

By Lemma 3, the higher order derivatives are continuous functions. Hence for any small number $\varepsilon_2>0$, there exists a constant δ_2 such that whenever $\left|b\left(\theta-\tilde{\theta}\right)\right|\leq\delta_2$,

$$\left|\frac{\mathrm{d}^j\tilde{l}\left(\theta\right)}{\mathrm{d}\theta^j}-\frac{\mathrm{d}^j\tilde{l}\left(\tilde{\theta}\right)}{\mathrm{d}\theta^j}\right|<\varepsilon_2.$$

By Lemma 10, there exists a constant N_4, such that whenever $n>N_4$.

$$\left|\frac{\mathrm{d}^j\tilde{l}\left(\theta\right)}{\mathrm{d}\theta^j}-\frac{P_k\left[E\left\{\prod_l\mathrm{d}^l g\left(X_1,\theta_0\right)/\mathrm{d}\theta^l\right\},\ldots,0,1\right]}{D^{r_k}}\right|<\varepsilon_2.$$

By assumption, all the moments are bounded when $k\leq K+3$. Then there exist a constant C_3, such that

$$D^{-r_k}P_k\left[E\left\{\prod_l\frac{\mathrm{d}^l g\left(X_1,\theta_0\right)}{\mathrm{d}\theta^l}\right\},\ldots,0,1\right]\leq C_3,$$

which leads to (A.2.1). □

A.3. Expansion Near The M-Estimator

LEMMA 8. . Under Assumptions 1 and 2,there exists a $\delta_3 > 0$, such that

$$\sum_{i=1}^{n} \log \hat{w}_i(\theta) - \sum_{i=1}^{n} \log \hat{w}_i\left(\tilde{\theta}\right) \leq -\frac{1}{4}y^2,$$

for any $\theta \in \left\{\theta : \left|b\left(\theta - \tilde{\theta}\right)\right| < \delta_3\right\} \cap H$.

PROOF. By Taylor expansion,

$$\frac{1}{n}\left\{\sum_{i=1}^{n} \log \hat{w}_i(\theta) - \sum_{i=1}^{n} \log \hat{w}_i\left(\tilde{\theta}\right)\right\} = \frac{\mathrm{d}\tilde{l}\left(\tilde{\theta}\right)}{\mathrm{d}\theta}\left(\theta - \tilde{\theta}\right) + \frac{1}{2}\frac{\mathrm{d}^2\tilde{l}(\theta^*)}{\mathrm{d}\theta^2}\left(\theta - \tilde{\theta}\right)^2,$$

where $\left|\theta^* - \tilde{\theta}\right| \leq \left|\theta - \tilde{\theta}\right|$. By Lemma 4, the first term in above equation is zero. By Lemma 10, we know that $\mathrm{d}^2\tilde{l}(\theta)/\mathrm{d}\theta^2$ is a continuous function in θ. Thus there exists a δ_3, such that for any $\left|b\left(\theta^* - \tilde{\theta}\right)\right| \leq \left|b\left(\theta - \tilde{\theta}\right)\right| < \delta_3$,

$$\left|\frac{\mathrm{d}^2\hat{l}(\theta^*)}{\mathrm{d}\theta^2}\left(\theta - \tilde{\theta}\right)^2 + \left|b\left(\theta - \tilde{\theta}\right)\right|^2\right| < \frac{1}{2}\left|b\left(\theta - \tilde{\theta}\right)\right|^2.$$

Hence $\mathrm{d}^2\tilde{l}(\theta^*)/\mathrm{d}\theta^2\left(\theta - \tilde{\theta}\right)^2 < -\left|b\left(\theta - \tilde{\theta}\right)\right|^2/2$, so that,

$$\sum_{i=1}^{n} \log \hat{w}_i(\theta) - \sum_{i=1}^{n} \log \hat{w}_i\left(\tilde{\theta}\right) < \frac{1}{2} \times \frac{1}{2}\left|\sqrt{n}b\left(\theta - \tilde{\theta}\right)\right|^2 = \frac{1}{4}y^2.$$

□

The next lemma plays a key role in expanding the posterior, and can be interpreted as an empirical likelihood version of the Edgeworth expansion.

LEMMA 9. Under Assumptions 1 , 3 , 4 and 5 , then there exist δ_4, M_3 and N_5, such that

$$\left|\int_{-\sqrt{n}\delta_4}^{\sqrt{n}\delta_4} \exp\left\{-\frac{1}{2}y^2 + \sum_{k=3}^{K+3} a_{kn}\left(\frac{y}{b}\right)^k n^{-(k-2)/2}\right\} - \prod_{i=1}^{n} \frac{\hat{w}_i(\theta)}{\hat{w}_i\left(\tilde{\theta}\right)}\,\mathrm{d}y\right| \leq M_3 n^{-(K+2)/2}, \text{ a.s.}.$$

PROOF. Let $\delta_4 \leq \min(\delta_2, \delta_3)$ in Lemma 7 and Lemma 8

$$\begin{aligned}
&\int_{-\sqrt{n}\delta_4}^{\sqrt{n}\delta_4} \exp\left\{-\frac{1}{2}y^2 + \sum_{k=3}^{K+3} a_{kn}\left(\frac{y}{b}\right)^k n^{-(k-2)/2}\right\} - \prod_{i=1}^{n} \frac{\hat{w}_i(\theta)}{\hat{w}_i\left(\tilde{\theta}\right)}\,\mathrm{d}y \\
= &\int_{-\sqrt{n}\delta_4}^{\sqrt{n}\delta_4} \exp\left\{\sum_{i=1}^{n} \log \hat{w}_i(\theta) - \sum_{i=1}^{n} \log \hat{w}_i\left(\tilde{\theta}\right)\right\} \\
&\left[\exp\left\{-\frac{1}{2}y^2 + \sum_{k=3}^{K+3} a_{kn}\left(\frac{y}{b}\right)^k n^{-(k-2)/2} - \sum_{i=1}^{n} \log \hat{w}_i(\theta) + \sum_{i=1}^{n} \log \hat{w}_i\left(\tilde{\theta}\right)\right\} - 1\right] \mathrm{d}y.
\end{aligned}$$

By Lemma 8, Lemma 7 and Taylor expansion, the above equation is bounded by

$$\int_{-\sqrt{n}\delta_4}^{\sqrt{n}\delta_4} \exp\left(-\frac{y^2}{4}\right)\left|\exp\left\{-a_{K+4,n}\left(\theta^*\right)\left(\frac{y}{b}\right)^{K+4} n^{-(K+2)/2}\right\}-1\right| \mathrm{d}y$$

$$\leq \int_{-\sqrt{n}\delta_4}^{\sqrt{n}\delta_4} \exp\left(-\frac{y^2}{4}\right)\left|\exp\left\{-C_1\left(\frac{y}{b}\right)^{K+4} n^{-(K+2)/2}\right\}-1\right| \mathrm{d}y. \tag{A.3.1}$$

where $\left|\theta^*-\tilde{\theta}\right| \leq \left|\theta-\tilde{\theta}\right| < \delta_4$, and C_4 is some constant dependent on δ_4, N_5 and $a_{K+4,n}\left(\tilde{\theta}\right)$. For sufficiently large n, and sufficiently small δ_4, $a_{K+4,n}\left(\theta^*\right)$ is very close to $a_{K+4,n}\left(\tilde{\theta}\right)$, and by Lemma 7, $a_{K+4,n}\left(\tilde{\theta}\right)$ is finite. Hence, for very large n, $\exp\left\{-C_4\left(y/b\right)^{K+4} n^{-(k+2)/2}\right\}-1$ does not change sign on either $[-\sqrt{n}\delta_4, 0]$ and $[0, \sqrt{n}\delta_4]$. So without loss of generality, we assume $\exp\left\{-C_4\left(y/b\right)^{K+4} n^{-(K+2)/2}\right\}-1 \geq 0$ on $[-\sqrt{n}\delta_4, \sqrt{n}\delta_4]$. With $t=\sqrt{n}$, and $t \in \mathbb{R}^+$, the last term in (A.3.1) can be written as

$$\int_{-\delta_4 t}^{\delta_4 t} \exp\left(-\frac{y^2}{4}\right)\left\{\exp\left(-\frac{C_4}{b^{K+4}} y^{K+4} t^{-K-2}\right)-1\right\} \mathrm{d}y.$$

If we can show that

$$\lim_{t\to+\infty} \frac{\int_{-\delta_4 t}^{\delta_4 t} \exp\left(-y^2/4\right)\left\{\exp\left(-C_4 y^{K+4} t^{-K-2}/b^{K+4}\right)-1\right\} \mathrm{d}y}{t^{-K-2}} = C_5,$$

for some $C_5 < \infty$, the lemma is proved. Take the derivative with respect to t in both the numerator and the denominator. In the denominator, $\left(t^{-K-2}\right)' = -\left(K+2\right) t^{-K-3}$. In the numerator,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\delta_4 t}^{\delta_4 t} \exp\left(-\frac{y^2}{4}\right)\left\{\exp\left(-\frac{C_4}{b^{K+4}} y^{K+4} t^{-K-2}\right)-1\right\} \mathrm{d}y \\
= &\int_{-\delta_4 t}^{\delta_4 t} \exp\left(-\frac{y^2}{4}\right)\left(-\frac{C_4}{b^{K+4}} y^{K+4}\right)\left(-K-2\right) t^{-K-3} \exp\left(-\frac{C_4}{b^{K+4}} y^{K+4} t^{-K-2}\right) \mathrm{d}y \\
&+\exp\left\{-\frac{\left(\delta_4 t\right)^2}{4}\right\}\left[\exp\left\{-\frac{C_4}{b^{K+4}}\left(\delta_4 t\right)^{K+4} t^{-K-2}\right\}-1\right]\delta_4 - \exp\left\{-\frac{\left(-\delta_4 t\right)^2}{4}\right\} \\
&\left[\exp\left\{-\frac{C_4}{b^{K+4}}\left(-\delta_4 t\right)^{K+4} t^{-K-2}\right\}-1\right]\left(-\delta_4\right) \\
= &\frac{\left(K+2\right) C_4}{b^{K+4}} t^{-K-3} \int_{-\delta_4 t}^{\delta_4 t} y^{K+4} \exp\left(-\frac{y^2}{4}-\frac{C_4}{b^{K+4}} y^{K+4} t^{-K-2}\right) \mathrm{d}y \\
&+\delta_4\left[\exp\left\{-\left(\frac{\delta_4^2}{4}-\frac{C_4}{b^{K+4}}\delta_4^{K+4}\right) t^2\right\}-\exp\left(-\frac{\delta_4^2 t^2}{4}\right)\right] \\
&+\delta_4\left(\exp\left[-\left\{\frac{\delta_4^2}{4}-\frac{C_1}{b^{K+4}}\left(-\delta_4\right)^{K+4}\right\} t^2\right]-\exp\left(-\frac{\delta_4^2 t^2}{4}\right)\right).
\end{aligned} \tag{A.3.2}$$

We choose δ_4 sufficiently small such that

$$\delta_4 < \min\left(\sqrt[K+2]{\frac{b^{K+4}}{4\left|C_4\right|}}, \delta_2, \delta_3\right). \tag{A.3.3}$$

Hence,

$$0 < \frac{\delta_4^2}{4} - \frac{|C_4|}{b^{K+4}}\delta_4^{K+4} \leq \frac{\delta_4^2}{4} - \frac{C_4}{b^{K+4}}\left(\pm\delta_4\right)^{K+4},$$

and

$$\lim_{t\to+\infty} \frac{\delta_4\left[\exp\left\{-\left(\delta_4^2/4 - |C_4|\,\delta_2^{K+4}/b^{K+4}\right)t^2\right\} - \exp\left(-\delta_4^2 t^2/4\right)\right]}{-\left(K+2\right)t^{-K-3}} = 0.$$

Hence, the last two terms in (A.3.2) tend to zero when $t \to +\infty$. Now we consider the ratio

$$\frac{\left\{\left(K+2\right)C_4/b^{K+4}\right\}t^{-K-3}\int_{-\delta_4 t}^{\delta_4 t} y^{K+4}\exp\left(-y^2/4 - C_4 y^{K+4}t^{-K-2}/b^{K+4}\right)\,\mathrm{d}y}{-\left(K+2\right)t^{-K-3}}$$

$$\frac{C_4}{b^{K+4}}\int_{-\delta_4 t}^{\delta_4 t} y^{K+4}\exp\left(-\frac{y^2}{4} - \frac{C_4}{b^{K+4}}y^{K+4}t^{-K-2}\right)\,\mathrm{d}y. \tag{A.3.4}$$

Since δ_4 satisfies (A.3.3), (A.3.4) is bounded by

$$\begin{aligned}
& \frac{|C_4|}{b^{K+4}}\int_{-\delta_4 t}^{\delta_4 t}|y|^{K+4}\exp\left\{-\frac{y^2}{4} - \frac{|C_4|}{b^{K+4}}\left(\delta_4 t\right)^{K+2}y^2 t^{-K-2}\right\}\,\mathrm{d}y \\
= \quad & \frac{|C_4|}{b^{K+4}}\int_{-\delta_4 t}^{\delta_4 t}|y|^{K+4}\exp\left\{-\left(\frac{\delta_4^2}{4} - \frac{|C_4|}{b^{K+4}}\delta_4^{K+4}\right)y^2\right\}\,\mathrm{d}y \\
\to \quad & \frac{|C_4|}{b^{K+4}}\sqrt{2\pi\left\{2\left(\frac{\delta_4^2}{4} - \frac{|C_4|}{b^{K+4}}\delta_4^{K+4}\right)\right\}^{-1}}\left\{2\left(\frac{\delta_4^2}{4} - \frac{|C_4|}{b^{K+4}}\delta_4^{K+4}\right)\right\}^{-(K+4)/2} \\
& \frac{2^{(K+4)/2}\Gamma\left\{\left(K+4+1\right)/2\right\}}{\sqrt{\pi}} < \infty.
\end{aligned}$$

So by L'Hostiple's rule, we have

$$\begin{aligned}
& \lim_{t\to+\infty}\frac{\int_{-\delta_4 t}^{\delta_4 t}\exp\left(-y^2/4\right)\left\{\exp\left(-C_4 y^{K+4}t^{-K-2}/b^{K+4}\right) - 1\right\}\,\mathrm{d}y}{t^{-K-2}} \\
= \quad & \lim_{t\to+\infty}\frac{\left[\int_{-\delta_4 t}^{\delta_4 t}\exp\left(-y^2/4\right)\left\{\exp\left(-C_4 y^{K+4}t^{-K-2}/b^{K+4}\right) - 1\right\}\,\mathrm{d}y\right]'}{\left(t^{-K-2}\right)'} = C_5 < \infty.
\end{aligned}$$

□

LEMMA 10. Under Assumptions 1 , 3 , 4 and 5 , there exists $\delta_4 > 0$, and constants M_4, N_6, such that

(A.3.5)

$$\left|\int_{-\sqrt{n}\delta_4}^{\sqrt{n}\delta_4}\left[\exp\left\{-\frac{1}{2}y^2 + \sum_{k=3}^{K+3}a_{kn}\left(\frac{y}{b}\right)^k n^{-(k-2)/2}\right\}\rho_K\left(\theta\right) - \prod_{i=1}^{n}\frac{\hat{w}_i\left(\theta\right)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho\left(\theta\right)\right]\,\mathrm{d}y\right| \leq M_4 n^{-\frac{1}{2}(K+1)}, \text{ a.s.}.$$

PROOF. Use δ_4 in Lemma 9, and apply Taylor expansion of $\tilde{l}\left(\theta\right)$ around $\tilde{\theta}$. Then for any $\tilde{\theta} - \delta_4/b \leq \theta \leq \tilde{\theta} + \delta_4/b$, there exists a θ^* which satisfies $\left|b\left(\theta^* - \tilde{\theta}\right)\right| \leq$

$\left|b\left(\theta-\tilde{\theta}\right)\right|$. This leads to

$$\begin{aligned}\tilde{l}(\theta) &= \tilde{l}\left(\tilde{\theta}\right)+\frac{\mathrm{d}\tilde{l}\left(\tilde{\theta}\right)}{\mathrm{d}\theta}\left(\theta-\tilde{\theta}\right)+\frac{1}{2}\frac{\mathrm{d}^2\tilde{l}\left(\tilde{\theta}\right)}{\mathrm{d}\theta^2}\left(\theta-\tilde{\theta}\right)^2+\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\theta-\tilde{\theta}\right)^k \\ &\quad+\frac{1}{(K+4)!}\frac{\mathrm{d}^{K+4}}{\mathrm{d}\theta}\tilde{l}(\theta^*)\left(\theta-\tilde{\theta}\right)^{K+4} \\ &= \tilde{l}\left(\tilde{\theta}\right)-\frac{1}{2}y^2n^{-1}+\sum_{k=3}^{K+3}a_{kn}\left(\frac{y}{b}\right)^k n^{-k/2}+\frac{1}{(K+4)!}\frac{\mathrm{d}^{K+4}}{\mathrm{d}\theta}\tilde{l}(\theta^*)\left(\theta-\tilde{\theta}\right)^{K+4}.\end{aligned}$$

Now

$$\begin{aligned}&\left|\exp\left\{-\frac{1}{2}y^2+\sum_{k=3}^{K+3}a_{kn}\left(\frac{y}{b}\right)^k n^{-(k-2)/2}\right\}\rho_K(\theta)-\prod_{i=1}^{n}\frac{\hat{w}_i(\theta)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho(\theta)\right| \\ &\leq\left|\exp\left\{-\frac{1}{2}y^2+\sum_{k=3}^{K+3}a_{kn}\left(\frac{y}{b}\right)^k n^{-(k-2)/2}\right\}\rho_K(\theta)-\prod_{i=1}^{n}\frac{\hat{w}_i(\theta)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho_K(\theta)\right| \\ &+\left|\prod_{i=1}^{n}\frac{\hat{w}_i(\theta)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho_K(\theta)-\prod_{i=1}^{n}\frac{\hat{w}_i(\theta)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho(\theta)\right| \\ &\leq|\rho_K(\theta)|\exp\left\{n\tilde{l}(\theta)-n\tilde{l}\left(\tilde{\theta}\right)\right\}\left|\exp\left[n\left\{\tilde{l}\left(\tilde{\theta}\right)-\frac{1}{2}y^2n^{-1}+\sum_{k=3}^{K+3}a_{kn}\left(\frac{y}{b}\right)^k n^{-k/2}-\tilde{l}(\theta)\right\}\right]-1\right| \\ &+\exp\left\{n\tilde{l}(\theta)-n\tilde{l}\left(\tilde{\theta}\right)\right\}|\rho_K(\theta)-\rho(\theta)|.\end{aligned}$$

By Lemma 9, the first term in the right hand side is bounded by

$$\left\{\max_{\tilde{\theta}-\delta_4/b\leq\theta\leq\tilde{\theta}+\delta_4/b}\rho_K(\theta)\right\}M_3n^{-(K+2)/2}.$$

By Lemma 8, and Taylor expansion of $\rho(\theta)$, the second term is bounded by

$$\begin{aligned}&\int_{-\sqrt{n}\delta_4}^{\sqrt{n}\delta_4}\exp\left(-\frac{y^2}{4}\right)\frac{n^{-(K+1)/2}}{(K+1)!}\rho^{K+1}(\theta^*)\,y^{K+1}\,\mathrm{d}y \\ \leq\;&\frac{1}{(K+1)!}\left\{\max_{\tilde{\theta}-\delta_4/b\leq\theta\leq\tilde{\theta}+\delta_4/b}\rho^{K+1}(\theta^*)\right\}\left\{\int_{-\sqrt{n}\delta_4}^{\sqrt{n}\delta_4}\exp\left(-\frac{y^2}{4}\right)y^{(K+1)/2}\,\mathrm{d}y\right\}n^{-(K+1)/2} \\ =\;&\left[\frac{1}{(K+1)!}\left\{\max_{\tilde{\theta}-\delta_4/b\leq\theta\leq\tilde{\theta}+\delta_4/b}\rho^{K+1}(\theta^*)\right\}\int_0^{\infty}\exp\left(-\frac{y^2}{4}\right)y^{(K+1)/2}\,\mathrm{d}y\right]n^{-(K+1)/2}\end{aligned}$$

Hence (B.4.3) holds.

□

A.4. Proof Of The Fundamental Theorem For Expansion

We first intuitively derive $P_K(\xi, n)$. First, we expand

$$\begin{aligned}
& \exp\left\{\sum_{k=3}^{K+3} a_{kn}\left(\frac{y}{b}\right)^k n^{-(k-2)/2}\right\} \\
= & \sum_{i=0}^{K+1}\frac{1}{i!}\left\{\sum_{k=3}^{K+3} a_{kn}\left(\frac{y}{b}\right)^k n^{-(k-2)/2}\right\}^i \\
= & 1+\sum_{i=1}^{K+1}\frac{1}{i!}\sum_{\sum_{u=3}^{K+3} m_{u,i}=i}\binom{i}{m_{3,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(a_{un}\right)^{m_{u,i}}\left(\frac{y}{b}\right)^{\sum_{u=3}^{K+3} m_{u,i}u} n^{-\left\{\sum_{u=3}^{K+3} m_{u,i}(u-2)\right\}/2}.
\end{aligned}$$

Multiplying the above expression by ρ_K,

$$\begin{aligned}
& \exp\left\{\sum_{k=3}^{K+3} a_{kn}\left(\frac{y}{b}\right)^k n^{-(k-2)/2}\right\}\rho_K(\theta) \\
= & \left[1+\sum_{i=1}^{K+1}\frac{1}{i!}\sum_{\sum_{u=3}^{K+3} m_{u,i}=i}\binom{i}{m_{3,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(a_{un}\right)^{m_{u,i}}\left(\frac{y}{b}\right)^{\sum_{u=3}^{K+3} m_{u,i}u} n^{-\left\{\sum_{u=3}^{K+3} m_{u,i}(u-2)\right\}/2}\right] \\
& \times\left\{\rho\left(\tilde{\theta}\right)+\sum_{j=1}^{K}\frac{1}{j!}\rho^{(j)}\left(\frac{y}{b}\right)^j n^{-j/2}\right\} \\
= & \rho\left(\tilde{\theta}\right)+\sum_{j=1}^{K}\frac{1}{j!}\rho^{(j)}\left(\frac{y}{b}\right)^j n^{-j/2}+\rho\left(\tilde{\theta}\right)\sum_{i=1}^{K+1}\frac{1}{i!} \\
& \sum_{\sum_{u=3}^{K+3} m_{u,i}=i}\binom{i}{m_{3,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(a_{un}\right)^{m_{u,i}}\left(\frac{y}{b}\right)^{\sum_{u=3}^{K+3} m_{u,i}u} n^{-\left\{\sum_{u=3}^{K+3} m_{u,i}(u-2)\right\}/2} \\
& +\left[\sum_{i=1}^{K+1}\frac{1}{i!}\sum_{\sum_{u=3}^{K+3} m_{u,i}=i}\binom{i}{m_{3,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(a_{un}\right)^{m_{u,i}}\left(\frac{y}{b}\right)^{\sum_{u=3}^{K+3} m_{u,i}u} n^{-\left\{\sum_{u=3}^{K+3} m_{u,i}(u-2)\right\}/2}\right] \\
& \times\sum_{j=1}^{K}\frac{1}{j!}\rho^{(j)}\left(\frac{y}{b}\right)^j n^{-j/2}.
\end{aligned}$$

For the third term in the right hand side of the above equation, we change the summation index. Let $\sum_{u=3}^{K+3} m_{u,i}(u-2)=h$. For any $\sum_{u=3}^{K+3} m_{u,i}=i$, $i\leq h\leq i(K+1)$, $h/(K+1)\leq i\leq h$. Thus the third term in the summation can be rearranged as

$$\sum_{h=1}^{(K+1)^2}\left\{\rho\left(\tilde{\theta}\right)\sum_{i=\lceil h/(K+1)\rceil}^{h}\frac{1}{i!}\sum_{I_{i,h}}\binom{i}{m_{3,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(a_{un}\right)^{m_{u,i}}\left(\frac{y}{b}\right)^{\sum_{u=3}^{K+3} m_{u,i}u}\right\} n^{-h/2}.$$

Similarly for the fourth term, let $\sum_{u=3}^{K+3} m_{u,i}\,(u-2)+j=h$. Then the summation can be rearranged as

$$\sum_{h=2}^{(K+1)^2+K}\left\{\sum_{j=1}^{h-1}\frac{1}{j!}\rho^{(j)}\left(\frac{y}{b}\right)^{j}\sum_{i=\lceil (h-j)/(K+1)\rceil}^{h-j}\frac{1}{i!}\sum_{I_{i,h-j}}\binom{i}{m_{3,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(a_{un}\right)^{m_{u,i}}\left(\frac{y}{b}\right)^{\sum_{u=3}^{K+3}m_{u,i}u}\right\}n^{-h/2}.$$

We collect same order terms of n, and denote the summation of all the terms with order higher than K to be $R_K(Y)$. Then we get the product as

$$\begin{aligned}&\rho\left(\tilde{\theta}\right)+\left\{\rho'\left(\frac{y}{b}\right)+\rho\left(\tilde{\theta}\right)a_{3n}\left(\frac{y}{b}\right)^{3}\right\}n^{-1/2}\\&+\sum_{h=2}^{K}\left\{\frac{1}{h!}\rho^{(h)}\left(\frac{y}{b}\right)^{h}+\sum_{j=0}^{h-1}\frac{1}{j!}\rho^{(j)}\left(\frac{y}{b}\right)^{j}\right.\\&\left.\times\sum_{i=\lceil (h-j)/(K+1)\rceil}^{h-j}\frac{1}{i!}\sum_{I_{i,h-j}}\binom{i}{m_{3,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(a_{un}\right)^{m_{u,i}}\left(\frac{y}{b}\right)^{\sum_{u=3}^{K+3}m_{u,i}u}\right\}n^{-h/2}+R_K(y).\end{aligned}$$

Integrating any Borel set $\left(Y_{(1)},\xi\right]$, we get the polynomial $P_K(\xi,n)$. Now we prove Theorem 1.

.

PROOF. Let $A_1=\{|y|\geq\delta_4\sqrt{n}\}\cap\left(Y_{(1)},\xi\right)$ and $A_2=\{|y|<\delta_4\sqrt{n}\}\cap\left(Y_{(1)},\xi\right)$, where δ_4 is in Lemma 15. Then

$$\begin{aligned}&\left|\int_{Y_{(1)}}^{\xi}\prod_{i=1}^{n}\frac{\hat{w}_i\left(\tilde{\theta}+y/\sqrt{n}b\right)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)\mathrm{d}y-P_K(\xi,n)\right|\\=&\left|\int_{Y_{(1)}}^{\xi}\left\{\prod_{i=1}^{n}\frac{\hat{w}_i\left(\tilde{\theta}+y/\sqrt{n}b\right)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)-\exp\left(-\frac{y^2}{2}\right)\sum_{h=0}^{K}\alpha_h(y,n)\,n^{-h/2}\right\}\mathrm{d}y\right|\\\leq&\left|\int_{A_1}\left\{\prod_{i=1}^{n}\frac{\hat{w}_i\left(\tilde{\theta}+y/\sqrt{n}b\right)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)-\exp\left(-\frac{y^2}{2}\right)\sum_{h=0}^{K}\alpha_h(y,n)\,n^{-h/2}\right\}\mathrm{d}y\right|\\&+\left|\int_{A_2}\left\{\prod_{i=1}^{n}\frac{\hat{w}_i\left(\tilde{\theta}+y/\sqrt{n}b\right)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)-\exp\left(-\frac{y^2}{2}\right)\sum_{h=0}^{K}\alpha_h(y,n)\,n^{-h/2}\right\}\mathrm{d}y\right|.\end{aligned}$$

For the first term, by Lemma 5, we have

$$\begin{aligned}
&\left|\int_{A_1}\prod_{i=1}^{n}\frac{\hat{w}_i\left(\tilde{\theta}+y/\sqrt{n}b\right)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)-\exp\left(-\frac{y^2}{2}\right)\sum_{h=0}^{K}\alpha_h\left(y,n\right)n^{-h/2}\,\mathrm{d}y\right| \\
\leq\ & \int_{A_1}\exp\left\{n\hat{l}\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)-n\hat{l}\left(\tilde{\theta}\right)\right\}\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)\,\mathrm{d}y \\
&+\left|\int_{A_1}\exp\left(-\frac{y^2}{4}\right)\sum_{h=0}^{K}\alpha_h\left(y,n\right)n^{-h/2}\,\mathrm{d}y\right| \\
\leq\ & \exp\left(-n\varepsilon\right)\int_{A_1}\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)\,\mathrm{d}y \\
&+\exp\left(-\frac{\delta_4^2 n}{4}\right)\left|\int_{A_1}\exp\left(-\frac{y^2}{4}\right)\sum_{h=0}^{K}\alpha_h\left(y,n\right)n^{-h/2}\,\mathrm{d}y\right| \\
\leq\ & \exp\left(-n\varepsilon\right)\int_{\mathbb{R}}\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)\,\mathrm{d}y+\exp\left(-n\frac{\delta_4^2}{4}\right)\sum_{h=0}^{K}\left\{\int_{Y_{(1)}}^{Y_{(n)}}\exp\left(-\frac{y^2}{4}\right)\left|\alpha_h\left(y,n\right)\right|\,\mathrm{d}y\right\}n^{-h/2}.
\end{aligned}$$

The above terms are exponentially decreasing with respect to n. Hence there exist N_1, and M_5, such that for any $n \geq N_1$,

$$\left|\int_{A_1}\prod_{i=1}^{n}\frac{\hat{w}_i\left(\tilde{\theta}+y/\sqrt{n}b\right)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)-\exp\left(-\frac{y^2}{2}\right)\sum_{h=0}^{K}\alpha_h\left(y,n\right)n^{-h/2}\,\mathrm{d}y\right| \leq M_5 n^{-(K+1)/2}.$$

For the second term, by Lemma 15, we have

$$\begin{aligned}
&\left|\int_{A_2}\prod_{i=1}^{n}\frac{\hat{w}_i\left(\tilde{\theta}+y/\sqrt{n}b\right)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)-\exp\left(-\frac{y^2}{2}\right)\sum_{h=0}^{K}\alpha_h\left(y,n\right)n^{-h/2}\,\mathrm{d}y\right| \\
\leq\ & \left|\int_{A_2}\prod_{i=1}^{n}\frac{\hat{w}_i\left(\tilde{\theta}+y/\sqrt{n}b\right)}{\hat{w}_i\left(\tilde{\theta}\right)}\rho\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)-\exp\left\{-\frac{1}{2}y^2+\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right\}\rho_K\left(\theta\right)\,\mathrm{d}y\right| \\
&+\left|\int_{A_2}\exp\left\{-\frac{1}{2}y^2+\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right\}\rho_K\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)\right. \\
&\left.-\exp\left(-\frac{y^2}{2}\right)\sum_{h=0}^{K}\alpha_h\left(y,n\right)n^{-h/2}\,\mathrm{d}y\right| \\
\leq\ & M_4 n^{-(K+1)/2} \\
&+\left|\int_{A_2}\exp\left\{-\frac{1}{2}y^2+\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right\}\rho_K\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)\right. \\
&\left.-\exp\left(-\frac{y^2}{2}\right)\sum_{h=0}^{K}\alpha_h\left(y,n\right)n^{-h/2}\,\mathrm{d}y\right|.
\end{aligned}$$

For the second term in the right hand side, we add and subtract $R_K(y)$ in integrand, and by Taylor expansion,

$$\begin{aligned}
&\left|\int_{A_2}\exp\left\{-\frac{1}{2}y^2+\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right\}\rho_K\left(\tilde{\theta}+\frac{y}{\sqrt{n}b}\right)-\exp\left(-\frac{y^2}{2}\right)\sum_{h=0}^{K}\alpha_h(y,n)\,n^{-h/2}\,\mathrm{d}y\right| \\
\leq\ &\left|\int_{A_2}\exp\left(-\frac{y^2}{2}\right)\rho_K(\theta)\left[\exp\left\{\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right\}-\sum_{i=0}^{K+1}\frac{1}{i!}\left\{\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right\}^i\right]\right| \\
&+\left|\int_{A_2}\exp\left(-\frac{y^2}{2}\right)R_K(y)\,\mathrm{d}n^{-1/2}y\right| \\
=\ &\left|\int_{A_2}\exp\left(-\frac{y^2}{2}\right)\rho_K(\theta)\frac{1}{(K+2)!}\exp(L)\left\{\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right\}^{K+2}\mathrm{d}y\right| \\
&+\left|\int_{A_2}\exp\left(-\frac{y^2}{2}\right)R_K(y)\,\mathrm{d}y\right|,
\end{aligned}$$

where $|L|\leq\left|\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)(y/b)^k n^{-(K-2)/2}\right|$. We know that $R_K(y)$ is a polynomial with order $n^{-(K+1)/2}$. So there exists an M_6, such that

$$\left|\int_{A_2}\exp\left(-\frac{y^2}{2}\right)R_K(y)\,\mathrm{d}y\right|\leq M_6 n^{-(K+1)/2}.$$

For the first term,

$$\begin{aligned}
&\left|\int_{A_2}\exp\left(-\frac{y^2}{2}\right)\rho_K(\theta)\frac{1}{(K+2)!}\exp(L)\left\{\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right\}^{K+2}\mathrm{d}n^{-1/2}Y\right| \qquad \text{(A.41)} \\
\leq\ &\frac{1}{(K+2)!}\left|\int_{A_2}\exp\left(-\frac{y^2}{2}\right)\rho_K(\theta)\exp\left\{\left|\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right|\right\}\left\{\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right\}^K\right. \\
=\ &\frac{1}{(K+2)!}\left|\int_{A_2}\rho_K(\theta)\exp\left\{-\frac{y^2}{2}+\left|\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right|\right\}\right. \\
&\left.\left\{\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right\}^{K+2}\mathrm{d}y\right|.
\end{aligned}$$

We need δ_4 sufficiently small, so that there exist C_6 and C_7, such that

$$\begin{aligned}
\frac{y^2}{2}-\left|\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2}\right| &\geq C_6 y^2, \\
\sum_{k=3}^{K+3}a_{kn}\left(\tilde{\theta}\right)\left(\frac{y}{b}\right)^k n^{-(K-2)/2} &\leq C_7 y^3 n^{-1/2}.
\end{aligned}$$

Hence, (B.5.1) is bounded by

$$\begin{aligned} & \frac{n^{K+2}}{(K+2)!}\left|\int_{A_n}\exp\left(-C_6y^2\right)\left(C_7y^3n^{-1/2}\right)^{K+2}\,\mathrm{d}y\right| \\ \leq\ & \frac{C_7^{K+2}}{(K+2)!}\left|\int_{Y_{(1)}}^{Y_{(n)}}\exp\left(-C_6y^2\right)y^{3(K+2)}n^{-(K+2)/2}\,\mathrm{d}y\right| \\ \leq\ & \frac{C_7^{K+2}}{(K+2)!}\left|\int_{Y_{(1)}}^{Y_{(n)}}\exp\left(-C_6y^2\right)y^{3(K+2)}\,\mathrm{d}y\right|n^{-(K+2)/2}. \end{aligned}$$

Adding all the parts , we get the inequality in Theorem 1 . □

APPENDIX B

Proof for Higher-Order Properties of Bayesian Empirical Likelihood: Multivariate Case

In the appendix, we give the detail proof of the theorem .

B.1. differentiability with respect to θ.

In this section, we prove the fact that the empirical likelihood is a smooth function of the mean parameter θ. Thus we can take arbitrary order of derivative with respect to θ. We begin the proof by the fact that the Lagrange multipliers are smooth functions of θ.

LEMMA 8. *Under the assumption 1, $\nu(\theta) \in C^{\infty}(H_n)$ with probability 1 in P_X^n.*

PROOF. $\nu(\theta)$ is a implicit function defined by equation (3.2.3) . Let

$$G_1(\nu, \theta) = \sum_{i=1}^{n} \frac{X_i - \theta}{1 + \nu^T (X_i - \theta)},$$

By implicit function theorem, we only need to show that $\det\left(\frac{\partial G_1}{\partial \nu}\right) \neq 0$. For any $1 \leq j, l \leq p$,

$$\frac{\partial G_{1j}}{\partial \nu_l} = \frac{\partial}{\partial \nu_l} \sum_{i=1}^{n} \frac{X_{ij} - \theta_j}{1 + \sum_{k=1}^{p} \nu_k (X_{ik} - \theta_k)} = -\sum_{i=1}^{n} \frac{X_{ij} - \theta_j}{1 + \nu^T (X_i - \theta)} \frac{X_{il} - \theta_l}{1 + \nu^T (X_i - \theta)}.$$

Let

$$\Delta = \left(\frac{X_{ij} - \theta_j}{1 + \nu^T (X_i - \theta)} \right)_{n \times p}.$$

Then $\frac{\partial G_1}{\partial \nu} = \Delta^T \Delta$. If we want $\det\left(\frac{\partial G_1}{\partial \nu}\right) \neq 0$, we need Δ has a full column rank. Suppose $\operatorname{rank}(\Delta) < p$, then there is a vector $\alpha \neq 0$, such that $\Delta \alpha = 0$, that is

$$\sum_{j=1}^{p} \frac{X_{ij} - \theta_j}{1 + \nu^T (X_i - \theta)} \alpha_j = 0,$$

$$\frac{1}{1 + \nu^T (X_i - \theta)} \sum_{j=1}^{p} (\alpha_j X_{ij} - \alpha_j \theta_j) = 0.$$

Note that $\hat{w}_i = \frac{1}{n[1 + \nu^T (X_i - \theta)]} > 0$, so we have

$$\sum_{j=1}^{p} \alpha_j X_{ij} = \sum_{j=1}^{p} \alpha_j \theta_j.$$

For any $1 \leq i \neq l \leq p$,

$$\sum_{j=1}^{p} \alpha_j X_{ij} = \sum_{j=1}^{p} \alpha_j \theta_j = \sum_{j=1}^{p} \alpha_j X_{lj},$$

that is

$$\begin{aligned} \sum_{j=1}^{p} \left(X_{ij} - X_{lj}\right) \alpha_j &= 0, \\ Z\alpha &= 0. \end{aligned}$$

Contradictory with the assumption 1. Hence, we have $\det\left(\frac{\partial G_1}{\partial \nu}\right) \neq 0$, and therefore, by implicit function theorem, ν is differentiable with respect to θ. Now we prove the existence of higher order derivatives by induction. For $k = 1$, take the derivative with respect to θ_l at the both sides of equation (3.2.3),

$$\sum_{i=1}^{n} \sum_{s=1}^{p} \frac{\partial \nu_s}{\partial \theta_l} \frac{X_{is} - \theta_s}{1 + \nu^T \left(X_i - \theta\right)} \frac{X_{ij} - \theta_j}{1 + \nu^T \left(X_i - \theta\right)} = -\delta_{jl} \sum_{i=1}^{n} \frac{1}{1 + \nu^T \left(X_i - \theta\right)} + \nu_l \sum_{i=1}^{n} \frac{X_{ij} - \theta_j}{\left[1 + \nu^T \left(X_i - \theta\right)\right]^2},$$

where δ_{jl} is Kronecker's delta, or in matrix form

$$\left(\Delta^T \Delta\right) \frac{\partial \nu}{\partial \theta} = -nI_p + n\Delta^T W \nu,$$

where W is the vector of weights $\hat{w}_i\left(\theta\right)$. Hence

$$\frac{\partial \nu}{\partial \theta} = -n\left(\Delta^T \Delta\right)^{-1} + n\left(\Delta^T \Delta\right)^{-1} \Delta^T W \nu^T.$$

We know both Δ, W and ν are differentiable with respect to θ. Suppose $k \leq K$, we have

$$\nabla^k \nu = f_k\left(\nu\left(\theta\right)\right),$$

where $f_k \in C^\infty$ are a bunch of smooth functions of ν, then for $k = K + 1$,

$$\nabla^{K+1} \nu = \nabla\left(\nabla^K \nu\right).$$

We know that ν is differentiable, and f_K are smooth functions, so the gradient above is well defined, and

$$\nabla^{K+1} \nu = \nabla f_K\left(\nu\right) \nabla \nu,$$

are also smooth functions. By mathematical induction, we have $\nu \in C^\infty\left(H_n\right)$. □

The smoothness of Lagrange multipliers leads to the smoothness of empirical likelihood weights, and thus logarithm of empirical likelihood. These results justify the Taylor expansion of logarithm of empirical likelihood around the sample mean.

B.2. the behavior of the logarithm of empirical likelihood and higher order derivatives around sample mean

We expand the posterior around the sample mean which is indeed the maximum of log empirical likelihood.

LEMMA 9. *$\overline{X}$ maximizes the log empirical likelihood $\hat{l}\left(\theta\right)$, and $\nabla \hat{l}\left(\overline{X}\right) = 0$.*

PROOF. By the inequality of arithmetic mean and geometric mean, we have

$$\hat{l}(\theta)=\frac{1}{n}\sum_{i=1}^{n}\log\hat{w}_i(\theta)=\log\left(\prod_{i=1}^{n}\hat{w}_i(\theta)\right)^{\frac{1}{n}}\leq\log\frac{\sum_{i=1}^{n}\hat{w}_i(\theta)}{n}=\log\frac{1}{n}, \tag{B.2.1}$$

where the equality holds if and only if all the $\hat{w}_i(\theta)$ are equal, i.e. , $\hat{w}_i(\theta)=n^{-1}$. So $\theta=\sum_{i=1}^{n}\hat{w}_i(\theta)X_i=\sum_{i=1}^{n}n^{-1}X_i=\overline{X}$. So $\hat{w}_i\left(\overline{X}\right)=n^{-1}$ and $\nu\left(\overline{X}\right)=0$. By lemma 8, $\hat{l}(\theta)$ is a smooth function of θ, so at the maximal $\overline{X}$, $\nabla\hat{l}(\theta)=0$. □

Next, we prove that the higher order derivatives of log empirical likelihood evaluated at $\overline{X}$ are solely smooth functions of the sample central moments. Then the remainder of Taylor expansion can be bounded by the power of n^{-1}.

LEMMA 10. *For any $k=2,3,\ldots$,*

$$\nabla^k\hat{l}(\theta)=f_k\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{s=1}^{l_1}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^{t_{l_1,1}}},\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{s=1}^{l_1}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^{t_{l_1,2}}},\ldots,\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{s=1}^{l_m}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^{t_{l_m,q}}},\nu\right),$$

where f_k are rational function and the denominator are only the power of $\Delta^T\Delta$, $1\leq l_i\leq k$, and $l_i\leq t_{l_i,j}\leq C_k<\infty$.

PROOF. We prove this lemma use mathematical induction. When $u=2$, by equation (B.3.1) below, we have

$$\nabla^2\hat{l}(\theta)=f_2\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(X_{ij}-\theta_j)(X_{is}-\theta_s)}{[1+\nu^T(X_i-\theta)]^2},\nu\right).$$

Suppose for $u=k$, we have

$$\nabla^k\hat{l}(\theta)=f_k\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{s=1}^{l_1}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^{t_{l_1,1}}},\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{s=1}^{l_1}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^{t_{l_1,2}}},\ldots,\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{s=1}^{l_m}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^{t_{l_m,q}}},\nu\right).$$

When $u=k+1$, for any $v=1,2,\ldots,p$,

$$\frac{\partial\nabla^k\hat{l}(\theta)}{\partial\theta_v}=\nabla f_k\frac{\partial}{\partial\theta_v}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{s=1}^{l_1}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^{t_{l_1,1}}},\ldots,\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{s=1}^{l_m}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^{t_{l_m,q}}},\nu\right).$$

For any l and t ,

$$\begin{aligned}
&\frac{\partial}{\partial\theta_v}\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{s=1}^{l}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^t}\\
=\;&\frac{1}{n}\sum_{i=1}^{n}\frac{1}{[1+\nu^T(X_i-\theta)]^{2t}}\left\{-\sum_{s=1}^{l}\delta_{j_sv}\prod_{r\neq s}(X_{ij_r}-\theta_{j_r})\left[1+\nu^T(X_i-\theta)\right]^t-\prod_{s=1}^{l}(X_{ij_s}-\theta_{j_s})\right.\\
&\left.\times t\left[1+\nu^T(X_i-\theta)\right]^{t-1}\left[\sum_{h=1}^{p}\frac{\partial\nu_h}{\partial\theta_t}(X_{ih}-\theta_h)-\nu_v\right]\right\}\\
=\;&-\sum_{s=1}^{l}\delta_{j_sv}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{r\neq s}(X_{ij_r}-\theta_{j_r})}{[1+\nu^T(X_i-\theta)]^t}\right\}-t\sum_{h=1}^{p}\frac{\partial\nu_h}{\partial\theta_t}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{(X_{ih}-\theta_h)\prod_{s=1}^{l}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^{t+1}}\right\}\\
&+t\nu_v\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{s=1}^{l}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^{t+1}}\right\},
\end{aligned}$$

which is still a smooth function of $\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{s=1}^{l}(X_{ij_s}-\theta_{j_s})}{[1+\nu^T(X_i-\theta)]^t}$, and the denominator is only introduced by term $\frac{\partial\nu}{\partial\theta}$, which contain only $\Delta^T\Delta$. Note that the gradient of a rational function f_k is still a rational function. Let $C_{k+1}=C_k+1$. Then the statement holds when $u=k+1$, by mathematical induction, it also holds for any k. □

By lemma 10, when evaluated at $\theta=\overline{X}$, since $\nu\left(\overline{X}\right)=0$, then

$$\nabla^k\hat{l}\left(\overline{X}\right)=f_k\left(\frac{1}{n}\sum_{i=1}^{n}\prod_{s=1}^{l_1}\left(X_{ij_s}-\theta_{j_s}\right),\ldots,\frac{1}{n}\sum_{i=1}^{n}\prod_{s=1}^{l_m}\left(X_{ij_s}-\theta_{j_s}\right)\right).$$

By assumption 2 and SLLN, when $n\to\infty$,

$$\frac{1}{n}\sum_{i=1}^{n}\prod_{s=1}^{l}\left(X_{ij_s}-\overline{X}_{j_s}\right)\to E\prod_{s=1}^{l}\left(X_{ij_s}-\theta_{j_s,0}\right)<\infty,\text{ a.s.}.$$

By the smoothness of f_k , for any ε, there exist a sufficient large $N_{(k)}$, such that for any $n>N_{(k)}$,

$$\left|\nabla^k\hat{l}\left(\overline{X}\right)-f_k\left(E\prod_{s=1}^{l}\left(X_{ij_s}-\theta_{j_s,0}\right)\right)\right|<\varepsilon,\text{ a.s.}.$$

By smoothness of $\nabla^k\hat{l}\left(\theta\right)$, there exist a $\delta>0$, such that for any $\left|B\left(\theta-\overline{X}\right)\right|<\delta$,

$$\left|\nabla^k\hat{l}\left(\theta\right)-\nabla^k\hat{l}\left(\overline{X}\right)\right|<\varepsilon.$$

Hence,

$$\left|\nabla^k\hat{l}\left(\theta\right)-f_k\left(E\prod_{s=1}^{l}\left(X_{ij_s}-\theta_{j_s,0}\right)\right)\right|\leq\left|\nabla^k\hat{l}\left(\overline{X}\right)-f_k\left(E\prod_{s=1}^{l}\left(X_{ij_s}-\theta_{j_s,0}\right)\right)\right|+\left|\nabla^k\hat{l}\left(\theta\right)-\nabla^k\hat{l}\left(\overline{X}\right)\right|<2\varepsilon.$$

Note that f_k is a rational function and the denominator contains only the power of $\Delta^T\Delta$, then there exists a constant $M_{(k)}$, such that $\left|f_k\left(E\prod_{s=1}^{l}\left(X_{ij_s}-\theta_{j_s,0}\right)\right)\right|<M_{(k)}$. Therefore,

$$M_{(k)}-2\varepsilon<\nabla^k\hat{l}\left(\theta\right)<M_{(k)}+2\varepsilon.$$

Particularly, we can compute $\nabla^2\hat{l}\left(\overline{X}\right)=n\left[\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right]^{-1}$, which is a positive definite matrix with probability 1 in P_X^n by assumption 4 .

LEMMA 11. *Under the assumption 4,there exists a $\delta_1>0$, such that*

$$\sum_{i=1}^{n}\ln\hat{w}_i\left(\theta\right)-\sum_{i=1}^{n}\ln\hat{w}_i\left(\overline{X}\right)\leq-\frac{1}{4}Y^TY,$$

for any $\theta\in\left\{\left|B\left(\theta-\overline{X}\right)\right|\leq\delta_1\right\}\cap H_n$.

PROOF. By Taylor expansion,

$$\frac{1}{n}\left(\sum_{i=1}^{n}\ln\hat{w}_i\left(\theta\right)-\sum_{i=1}^{n}\ln\hat{w}_i\left(\overline{X}\right)\right)=\left(\nabla\hat{l}\left(\overline{X}\right)\right)^T\left(\theta-\overline{X}\right)+\frac{1}{2}\left(\theta-\overline{X}\right)^T\nabla^2\hat{l}\left(\theta^*\right)\left(\theta-\overline{X}\right),$$

where $\left|\theta^* - \overline{X}\right| \leq \left|\theta - \overline{X}\right|$. By lemma 9, the first term in above equation is zero. By lemma 10, we know that $\nabla^2 \hat{l}(\theta)$ is a continuous function in θ. There exists a δ_1, such that for any $\left|B\left(\theta^* - \overline{X}\right)\right| \leq \left|B\left(\theta - \overline{X}\right)\right| < \delta_1$,

$$\left|\left(\theta - \overline{X}\right)^T \nabla^2 \hat{l}(\theta^*)\left(\theta - \overline{X}\right) + \left|B\left(\theta - \overline{X}\right)\right|^2\right| < \frac{1}{2}\left|B\left(\theta - \overline{X}\right)\right|^2,$$

hence $\left(\theta - \overline{X}\right)^T \nabla^2 \hat{l}(\theta^*)\left(\theta - \overline{X}\right) < -2^{-1}\left|B\left(\theta - \overline{X}\right)\right|^2$. Therefore,

$$\sum_{i=1}^{n} \ln \hat{w}_i(\theta) - \sum_{i=1}^{n} \ln \hat{w}_i\left(\overline{X}\right) < \frac{1}{2} \times \frac{1}{2}\left|\sqrt{n}B\left(\theta - \overline{X}\right)\right|^2 = \frac{1}{4} Y^T Y.$$

□

B.3. bell shape of the logarithm of empirical likelihood

If we want the expansion of posterior is uniform on the parameter space, we do not only need the Taylor expansion around the sample mean, but also require controlling the value of posterior in the region far from sample mean. This can be achieved by showing empirical likelihood have "bell" shape around the sample mean.

LEMMA 12. *Under the assumption 1 and 2,* $\frac{\partial^2 \hat{l}(\theta)}{\partial\theta\partial\theta^T}$ *is negative definite almost surely in* P_X^n.

PROOF. For any $j = 1, 2, \ldots, p$, take the first derivative of logarithm of empirical likelihood,

$$\begin{aligned}
\frac{\partial \hat{l}_n(\theta)}{\partial \theta_j} &= \frac{1}{n}\frac{\partial}{\partial \theta_j}\sum_{i=1}^{n}\left[-\ln\left(1 + \sum_{k=1}^{p}\nu_k(\theta)\left(X_{ik} - \theta_k\right)\right)\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{p}\frac{\partial \nu_k}{\partial \theta_j}\frac{X_{ik} - \theta_k}{n\left[1 + \nu^T\left(X_i - \theta\right)\right]} + \sum_{i=1}^{n}\frac{1}{n\left[1 + \nu^T\left(X_i - \theta\right)\right]}\nu_j = \nu_j.
\end{aligned}$$

Then the Hessian of logarithm of empirical likelihood is

$$\frac{\partial^2 \hat{l}(\theta)}{\partial\theta^T\partial\theta} = \frac{\partial \nu}{\partial \theta} = -n\left(\Delta^T\Delta\right)^{-1} + n\left(\Delta^T\Delta\right)^{-1}\Delta^T W\nu^T.$$

Note that

$$\Delta\nu = \left(\frac{\nu^T\left(X_i - \theta\right)}{1 + \nu^T\left(X_i - \theta\right)}\right)_{i=1}^{n} = \left(1 - \frac{1}{1 + \lambda^T\left(X_i - \theta\right)}\right)_{i=1}^{n} = 1_{n\times 1} - nW,$$

where $1_{n\times 1}$ is a column vector with all entries are equal to 1, so

$$W = \frac{1}{n}\left(1_{n\times 1} - \Delta\nu\right).$$

Replace W in Hessian,

$$\begin{aligned}
\frac{\partial^2 \hat{l}(\theta)}{\partial\theta^T\partial\theta} &= n\left[-\left(\Delta^T\Delta\right)^{-1} + \left(\Delta^T\Delta\right)^{-1}\Delta^T\frac{1}{n}\left(1_{n\times 1} - \Delta\nu\right)\nu^T\right] \\
&= -n\left(\Delta^T\Delta\right)^{-1} + \left(\Delta^T\Delta\right)^{-1}\left(\Delta^T 1_{n\times 1}\right) - \left(\Delta^T\Delta\right)^{-1}\Delta^T\Delta\nu\nu^T.
\end{aligned}$$

Also note that

$$\Delta^T 1_{n\times 1} = \left(\sum_{i=1}^{n}\frac{X_{ij} - \theta_j}{1 + \nu^T\left(X_i - \theta\right)}\right)_{j=1}^{n} = 0.$$

Therefore,

$$\frac{\partial^2 \hat{l}(\theta)}{\partial \theta^T \partial \theta} = -n\left(\Delta^T \Delta\right)^{-1} - \nu\nu^T, \tag{B.3.1}$$

is obviously negative definite since in proof of lemma 8, we know that Δ has a full column rank. Particularly when $\theta = \overline{X}$, $\frac{\partial^2 \hat{l}(\overline{X})}{\partial \theta^T \partial \theta} = -\left[\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2\right]^{-1}$. □

Now we can control the tail part by the following lemma.

LEMMA 13. *Under the assumption 1, for any $\delta > 0$, there exists an $\varepsilon > 0$, N_2, and $D_1 \subset \mathbb{R}^p$, $P_X^n(D_1) = 1$, such that*

$$\hat{l}(\theta) - \hat{l}\left(\overline{X}\right) \leq -\varepsilon, \text{ a.s.},$$

for any $\left|B\left(\theta - \overline{X}\right)\right| \geq \delta$ and $\theta \in H_n$.

PROOF. In the proof of lemma 12, we know that $\nu\left(\overline{X}\right) = 0$, thus $\hat{l}\left(\overline{X}\right) = -\log n$ and score function $\hat{l}'\left(\overline{X}\right) = \nu\left(\overline{X}\right) = 0$. By strictly convexity of log empirical likelihood from lemma 12, we know that $\overline{X}$ is the unique maximal. Therefore, for any $\theta \neq \overline{X}$, $\hat{l}(\theta) < \hat{l}\left(\overline{X}\right)$. Note that the set

$$\left\{\theta \middle| \left|B\left(\theta - \overline{X}\right)\right| \geq \delta\right\} \cap H_n$$

is a compact set, and $\hat{l}(\theta)$ is a continuous function, then there exists a $\theta^* \in \left\{\theta \middle| \left|B\left(\theta - \overline{X}\right)\right| \geq \delta\right\} \cap H_n$, such that for any $\theta \in \left\{\theta \middle| \left|B\left(\theta - \overline{X}\right)\right| \geq \delta\right\} \cap H_n$,

$$\hat{l}(\theta) \leq \hat{l}(\theta^*).$$

Therefore, $\hat{l}(\theta) \leq \hat{l}(\theta^*) < \hat{l}\left(\overline{X}\right)$,

$$\hat{l}(\theta) - \hat{l}\left(\overline{X}\right) \leq \hat{l}(\theta^*) - \hat{l}\left(\overline{X}\right) < 0.$$

Let $\varepsilon = \frac{1}{2}\left(\hat{l}\left(\overline{X}\right) - \hat{l}(\theta^*)\right)$, then we have

$$\hat{l}(\theta) - \hat{l}\left(\overline{X}\right) \leq \hat{l}(\theta^*) - \hat{l}\left(\overline{X}\right) < \varepsilon.$$

□

B.4. expansion near the sample mean

After we control the tail of the posterior, the remaining work are solely based on smoothness of the log empirical likelihood and prior. Particularly the form of expansion polynomials are determined by the product of Taylor expansion of empirical likelihood and prior at sample mean.

LEMMA 14. *There exist a δ_2, a constant M_1 and N_1, such that*

$$\left|\int_{\left\{B\left(\theta-\overline{X}\right)\leq\delta_2\right\}\cap H_n} \exp\left(-\frac{1}{2}Y^TY + \sum_{k=3}^{K+3}\delta_k \hat{l} n^{-\frac{k-2}{2}}\right) - \prod_{i=1}^n \frac{\hat{w}_i(\theta)}{\hat{w}_i\left(\overline{X}\right)}\,\mathrm{d}n^{-\frac{1}{2}}Y\right| \leq M_1 n^{-\frac{1}{2}(K+3)}, \text{ a.s.}.$$

PROOF. First, we can choose δ_2 small enough so that $\left\{B\left(\theta-\overline{X}\right)\leq\delta_2\right\}\subset H_n$.

$$\begin{aligned}
&\int_{\left\{B\left(\theta-\overline{X}\right)\leq\delta_2\right\}\cap H_n}\exp\left(-\frac{1}{2}Y^TY+\sum_{k=3}^{K+3}\delta_k\hat{l}n^{-\frac{k-2}{2}}\right)-\prod_{i=1}^{n}\frac{\hat{w}_i\left(\theta\right)}{\hat{w}_i\left(\overline{X}\right)}\,\mathrm{d}n^{-\frac{1}{2}}Y\\
=&\int_{\left\{B\left(\theta-\overline{X}\right)\leq\delta_2\right\}}\exp\left(\sum_{i=1}^{n}\ln\hat{w}_i\left(\theta\right)-\sum_{i=1}^{n}\ln\hat{w}_i\left(\overline{X}\right)\right)\\
&\left[\exp\left(\left(-\frac{1}{2}Y^TY+\sum_{k=3}^{K+3}\delta_k\hat{l}n^{-\frac{k-2}{2}}\right)-\sum_{i=1}^{n}\left(\ln\hat{w}_i\left(\theta\right)-\ln\hat{w}_i\left(\overline{X}\right)\right)\right)-1\right]\mathrm{d}n^{-\frac{1}{2}}Y.
\end{aligned}$$

By lemma 11, and Taylor expansion, the above equation can be bounded by

$$\begin{aligned}
&\int_{\left\{\left|B\left(\theta-\overline{X}\right)\right|\leq\delta_2\right\}}\exp\left(-\frac{Y^TY}{4}\right)\left|\exp\left(-\frac{n^{-\frac{K+2}{2}}}{(K+4)!}\left[\left(\theta-\overline{X}\right)\nabla\right]^{K+4}\hat{l}\left(\theta^*\right)\right)-1\right|\mathrm{d}n^{-\frac{1}{2}}Y\\
=&\ n^{-\frac{1}{2}}\int_{\left\{\sqrt{Y^TY}\leq\delta_2\sqrt{n}\right\}}\exp\left(-\frac{Y^TY}{4}\right)\left|\exp\left(-\frac{n^{-\frac{K+2}{2}}}{(K+4)!}\left(Y^TB^{-T}\nabla\right)^{K+4}\hat{l}\left(\theta^*\right)\right)-1\right|\mathrm{d}Y.
\end{aligned}$$

By lemma 10, we know that $\nabla^{K+4}\hat{l}\left(\theta^*\right)$ are bounded by some constants. Note that for any Y, and any vector a, there exists a constant C_1, such that

$$\left(Y^Ta\right)^2\leq C_1\left(Y^TY\right), \tag{B.4.1}$$

so we can bounded the above equation by

$$n^{-\frac{1}{2}}\int_{\left\{\sqrt{Y^TY}\leq\delta_2\sqrt{n}\right\}}\exp\left(-\frac{Y^TY}{4}\right)\left(\exp\left(-n^{-\frac{K+2}{2}}C_3\left(Y^TY\right)^{\frac{K+4}{2}}\right)-1\right)\mathrm{d}Y,$$

where $C_3=C_1^{\frac{K+4}{2}}$. Denote the above integrand to be $F\left(Y,\sqrt{n}\right)$. We can the inequality in this lemma if we can prove the following

$$\lim_{n\to\infty}\frac{n^{-\frac{1}{2}}\int_{\left\{\sqrt{Y^TY}\leq\delta_2\sqrt{n}\right\}}F\left(Y,\sqrt{n}\right)\mathrm{d}Y}{n^{-\frac{1}{2}(K+3)}}=C_2, \tag{B.4.2}$$

for some constant $C_2<\infty$. Let $t=\sqrt{n}$, and relax $t\in\mathbb{R}^+$. The the above formula can be written as

$$\frac{\int_{\left\{\sqrt{Y^TY}\leq\delta_2t\right\}}F\left(Y,t\right)\mathrm{d}Y}{t^{-K-2}}.$$

Take the derivatives of both numerator and denominator with respect to t. For the denominator $\left(t^{-K-2}\right)'=-\left(K+2\right)t^{-K-3}$. For the numerator, we change to n dimensional spherical coordinate system,

$$\begin{aligned}
Y_1&=r\cos\left(\varphi_1\right),\\
Y_2&=r\sin\left(\varphi_1\right)\cos\left(\varphi_2\right),\\
&\vdots\\
Y_p&=r\sin\left(\varphi_1\right)\sin\left(\varphi_2\right)\cdots\sin\left(\varphi_{p-1}\right).
\end{aligned}$$

So the numerator can be written as

$$\begin{aligned}
&\int_{\{r\leq\delta_2 t\}} \exp\left(-\frac{r^2}{4}\right)\left(\exp\left(-C_3 t^{-K-2} r^{K+4}\right)-1\right)\\
&\times r^{p-1}\sin^{p-2}(\varphi_1)\sin^{p-3}(\varphi_2)\cdots\sin(\varphi_{p-2})\,\mathrm{d}r\,\mathrm{d}\varphi_1\cdots\mathrm{d}\varphi_{p-1}\\
=&\int_0^{2\pi}\mathrm{d}\varphi_{p-1}\int_0^{\pi}\sin^{p-2}(\varphi_1)\,\mathrm{d}\varphi_1\cdots\int_0^{\pi}\sin(\varphi_{p-2})\,\mathrm{d}\varphi_{p-2}\\
&\times\int_0^{\delta_2 t}\exp\left(-\frac{r^2}{4}\right)\left(\exp\left(-C_3 t^{-K-4} r^{K+4}\right)-1\right)r^{p-1}\,\mathrm{d}r\\
\leq&\ 2\pi(\pi)^{p-2}\int_0^{\delta_2 t}\exp\left(-\frac{r^2}{4}\right)\left(\exp\left(-C_3 t^{-K-2} r^{K+4}\right)-1\right)r^{p-1}\,\mathrm{d}r\\
=&\ 2\pi^{p-1}\int_0^{\delta_2 t}\exp\left(-\frac{r^2}{4}\right)\left(\exp\left(-C_3 t^{-K-2} r^{K+4}\right)-1\right)r^{p-1}\,\mathrm{d}r.
\end{aligned}$$

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_0^{\delta_2 t}\exp\left(-\frac{r^2}{4}\right)\left(\exp\left(-C_3 t^{-K-2} r^{K+4}\right)-1\right)r^{p-1}\,\mathrm{d}r\\
=&\int_0^{\delta_2 t}-(K+2)\,t^{-K-3}\left(-C_3 r^{K+4}\right)\exp\left(-\frac{r^2}{4}\right)\exp\left(-C_3 t^{-K-2} r^{K+4}\right)r^{p-1}\,\mathrm{d}r\\
&+\exp\left(-\frac{\delta_2^2 t^2}{4}\right)\left(\exp\left(-C_3 t^{-K-2}\left(\delta_2 t\right)^{K+4}\right)-1\right)\\
\leq&\ C_3(K+2)\,t^{-K-3}\int_0^{\delta_2 t} r^{K+p+3}\exp\int_{A\cap H_n}\exp\left(-\frac{Y^T Y}{4}\right)\left|\alpha_h(Y,n)\right|\,\mathrm{d}Y\\
&\left(-\left(\frac{1}{4}-C_3 t^{-K-2} r^{K+2}\right)r^2\right)\mathrm{d}r+\exp\left(-\left(\frac{1}{4}-C_3\delta_2^{K+2}\right)\delta_2^2 t^2\right)-\exp\left(-\frac{\delta_2^2 t^2}{4}\right).
\end{aligned}$$

If

$$\delta_2 < \sqrt[K+2]{\frac{1}{4C_3}},$$

then

$$\frac{1}{4}-C_3\delta_2^{K+2}>0,$$

hence

$$\lim_{t\to+\infty}\frac{\exp\left(-\left(\frac{1}{4}-C_3\delta_2^{K+2}\right)\delta_2^2 t^2\right)-\exp\left(-\frac{\delta_2^2 t^2}{4}\right)}{-(K+2)\,t^{-K-3}}=0.$$

For the first term in derivative of numerator, we have for any $t>0$.

$$\begin{aligned}
&\int_0^{\delta_2 t} r^{K+p+3}\exp\left(-\left(\frac{1}{4}-C_3 t^{-K-2} r^{K+2}\right)r^2\right)\mathrm{d}r\\
\leq&\int_0^{\delta_2 t} r^{K+p+3}\exp\left(-\left(\frac{1}{4}-C_3\delta_2^{K+2}\right)r^2\right)\mathrm{d}r\\
\leq&\int_{\mathbb{R}}|r|^{K+p+3}\exp\left(-\left(-\frac{1}{4}-C_3\delta_2^{K+2}\right)r^2\right)\mathrm{d}r<+\infty.
\end{aligned}$$

Therefore, there exists a constant C_2, such that

$$\begin{aligned}&\lim_{t\to+\infty}\frac{C_3\left(K+2\right)t^{-K-3}\int_0^{\delta_2 t}r^{K+p+3}\exp\left(-\left(\frac{1}{4}-C_3t^{-K-2}r^{K+2}\right)r^2\right)\,\mathrm{d}r}{-\left(K+2\right)t^{-K-3}}\\ =\;&-C_3\lim_{t\to+\infty}\int_0^{\delta_2 t}r^{K+p+3}\exp\left(-\left(\frac{1}{4}-C_3t^{-K-2}r^{K+2}\right)r^2\right)\,\mathrm{d}r=C_2.\end{aligned}$$

By L'Hostiple's rule, we have equation (B.4.2) holds, and therefore, the lemma holds. □

LEMMA 15. *Under the assumption 1, 2 and 3, there exist a $\delta_2>0$, a constant M_2 and N_2, such that*

(B.4.3)

$$\left|\int_{\left\{\left|B\left(\theta-\overline{X}\right)\right|\leq\delta\right\}\cap H_n}\exp\left(-\frac{1}{2}Y^TY+\sum_{k=3}^{K+3}\delta_k\hat{l}n^{-\frac{k-2}{2}}\right)\rho_K\left(\theta\right)-\prod_{i=1}^{n}\frac{\hat{w}_i\left(\theta\right)}{\hat{w}_i\left(\overline{X}\right)}\rho\left(\theta\right)\,\mathrm{d}n^{-\frac{1}{2}}Y\right|\leq M_2n^{-\frac{1}{2}(K+2)},\text{ a.s.}.$$

PROOF. Apply Taylor expansion to $\hat{l}\left(\theta\right)$ around $\overline{X}$, for any $\theta\in H_n$, there exists a θ^*satisfies $\left|B\left(\theta^*-\overline{X}\right)\right|\leq\left|B\left(\theta-\overline{X}\right)\right|$, such that,

$$\begin{aligned}\hat{l}\left(\theta\right)&=\hat{l}\left(\overline{X}\right)+\nabla\hat{l}\left(\overline{X}\right)\left(\theta-\overline{X}\right)+\frac{1}{2}\left(\theta-\overline{X}\right)^T\frac{\partial^2\hat{l}\left(\overline{X}\right)}{\partial\theta^T\partial\theta}\left(\theta-\overline{X}\right)+\sum_{k=3}^{K+3}\frac{1}{k!}\left[\left(\theta-\overline{X}\right)^T\nabla\right]^k\hat{l}\left(\overline{X}\right)\\&\quad+\frac{1}{(K+4)!}\left[\left(\theta-\overline{X}\right)\nabla\right]^{K+4}\hat{l}\left(\theta^*\right)\\&=\hat{l}\left(\overline{X}\right)-\frac{1}{2}Y^TYn^{-1}+\sum_{k=3}^{K+3}\delta_k\hat{l}n^{-\frac{k}{2}}++\frac{1}{(K+4)!}\left[\left(\theta-\overline{X}\right)\nabla\right]^{K+4}\hat{l}\left(\theta^*\right).\end{aligned}$$

Now we have

$$\begin{aligned}&\left|\exp\left(-\frac{1}{2}Y^TY+\sum_{k=3}^{K+3}\delta_k\hat{l}n^{-\frac{k-2}{2}}\right)\rho_K\left(\theta\right)-\prod_{i=1}^{n}\frac{\hat{w}_i\left(\theta\right)}{\hat{w}_i\left(\overline{X}\right)}\rho\left(\theta\right)\right|\\&\leq\left|\exp\left(-\frac{1}{2}Y^TY+\sum_{k=3}^{K+3}\delta_k\hat{l}n^{-\frac{k-2}{2}}\right)\rho_K\left(\theta\right)-\prod_{i=1}^{n}\frac{\hat{w}_i\left(\theta\right)}{\hat{w}_i\left(\overline{X}\right)}\rho_K\left(\theta\right)\right|\\&+\left|\prod_{i=1}^{n}\frac{\hat{w}_i\left(\theta\right)}{\hat{w}_i\left(\overline{X}\right)}\rho_K\left(\theta\right)-\prod_{i=1}^{n}\frac{\hat{w}_i\left(\theta\right)}{\hat{w}_i\left(\overline{X}\right)}\rho\left(\theta\right)\right|\\&\leq\left|\rho_K\left(\theta\right)\right|\exp\left(n\left(\hat{l}\left(\theta\right)-\hat{l}\left(\overline{X}\right)\right)\right)\left|\exp\left(n\left(\hat{l}\left(\overline{X}\right)-\frac{1}{2}Y^TYn^{-1}+\sum_{k=3}^{K+3}\delta_k\hat{l}n^{-\frac{k-2}{2}}-\hat{l}\left(\theta\right)\right)\right)-1\right|\\&+\exp\left(n\left(\hat{l}\left(\theta\right)-\hat{l}\left(\overline{X}\right)\right)\right)\left|\rho_K\left(\theta\right)-\rho\left(\theta\right)\right|.\end{aligned}$$

For the first part of the above formula, by lemma 14, we have it to be bounded by

$$\max_{\theta\in\left\{\left\{B\left(\theta-\overline{X}\right)\leq\delta_2\right\}\cap H_n\right\}}\rho_K\left(\theta\right)M_1n^{-\frac{1}{2}(K+3)}.$$

For the second part, by lemma 11, Taylor expansion in $\rho(\theta)$, and equation (B.4.1), we have the upper bound to be

$$\begin{aligned}
& \int_{\left\{Y^TY\le\delta_2^2 n\right\}} \exp\left(-\frac{Y^TY}{4}\right)\frac{n^{-\frac{K+1}{2}}}{(K+1)!}\left(Y^TB^{-T}\nabla\right)^{K+1}\rho(\theta^*)\,\mathrm{d}n^{-\frac{1}{2}}Y \\
\le\ & \frac{1}{(K+1)!}\max_{\theta\in\left\{\left\{B\left(\theta-\overline{X}\right)\le\delta_2\right\}\cap H_n\right\}}\left|\nabla^{K+1}\rho(\theta^*)\right|\int_{\left\{Y^TY\le\delta_2^2 n\right\}}\exp\left(-\frac{Y^TY}{4}\right)C_1^{\frac{K+1}{2}}\left(Y^TY\right)^{\frac{K+1}{2}}\mathrm{d}Yn^{-\frac{K+2}{2}} \\
\le\ & \frac{C_1^{\frac{K+1}{2}}}{(K+1)!}\max_{\theta\in\left\{\left\{B\left(\theta-\overline{X}\right)\le\delta_2\right\}\cap H_n\right\}}\left|\nabla^{K+1}\rho(\theta)\right|\int_{\mathbb{R}^p}\exp\left(-\frac{Y^TY}{4}\right)\left(Y^TY\right)^{\frac{K+1}{2}}\mathrm{d}Yn^{-\frac{K+2}{2}} \\
=\ & \frac{C_1^{\frac{K+1}{2}}}{(K+1)!}\max_{\theta\in\left\{\left\{B\left(\theta-\overline{X}\right)\le\delta_2\right\}\cap H_n\right\}}\left|\nabla^{K+1}\rho(\theta)\right|2\pi^{p-1}\int_0^\infty\exp\left(-\frac{r^2}{4}\right)r^{K+1}r^{p-1}\,\mathrm{d}rn^{-\frac{K+2}{2}} \\
=\ & \left[\frac{2\pi^{p-1}C_1^{\frac{K+1}{2}}}{(K+1)!}\max_{\theta\in\left\{B\left(\theta-\overline{X}\right)\le\delta_2\right\}\cap H_n}\left|\nabla^{K+1}\rho(\theta)\right|\int_0^\infty r^{p+K}\exp\left(-\frac{r^2}{4}\right)\mathrm{d}r\right]n^{-\frac{K+2}{2}}
\end{aligned}$$

Since $p\ge 1$, we have equation (B.4.3) holds. □

B.5. proof of the main theorem

We first intuitively derive . First, we expand

$$\begin{aligned}
& \exp\left(\sum_{k=3}^{K+3}\delta_k\hat{l}n^{-\frac{k-2}{2}}\right) \\
=\ & \sum_{i=0}^{K+1}\frac{1}{i!}\left(\sum_{k=3}^{K+3}\delta_k\hat{l}n^{-\frac{k-2}{2}}\right)^i \\
=\ & 1+\sum_{i=1}^{K+1}\frac{1}{i!}\sum_{\sum_{u=3}^{K+3}m_{u,i}=i}\binom{i}{m_{3,i},m_{4,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(\delta_u\hat{l}\right)^{m_{u,i}}n^{-\frac{1}{2}\sum_{u=3}^{K+3}m_{u,i}(u-2)}.
\end{aligned}$$

Then we product the above expansion by ρ_K,

$$\begin{aligned}
& \exp\left(\sum_{k=3}^{K+3}\frac{1}{k!}\delta_k\hat{l}n^{-\frac{k-2}{2}}\right)\rho_K(\theta) \\
= & \left[1+\sum_{i=1}^{K+1}\frac{1}{i!}\sum_{\sum_{u=3}^{K+3}m_{u,i}=i}\binom{i}{m_{3,i},m_{4,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(\delta_u\hat{l}\right)^{m_{u,i}}n^{-\frac{1}{2}\sum_{u=3}^{K+3}m_{u,i}(u-2)}\right] \\
& \left(\rho\left(\overline{X}\right)+\sum_{j=1}^{K}\delta_j\rho n^{-\frac{j}{2}}\right) \\
= & \rho\left(\overline{X}\right)+\sum_{j=1}^{K}\delta_j\rho n^{-\frac{j}{2}}+\rho\left(\overline{X}\right)\sum_{i=1}^{K+1}\frac{1}{i!} \\
& \sum_{\sum_{u=3}^{K+3}m_{u,i}=i}\binom{i}{m_{3,i},m_{4,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(\delta_u\hat{l}\right)^{m_{u,i}}n^{-\frac{1}{2}\sum_{u=3}^{K+3}m_{u,i}(u-2)} \\
& +\left[\sum_{i=1}^{K+1}\frac{1}{i!}\sum_{\sum_{u=3}^{K+3}m_{u,i}=i}\binom{i}{m_{3,i},m_{4,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(\delta_u\hat{l}\right)^{m_{u,i}}n^{-\frac{1}{2}\sum_{u=3}^{K+3}m_{u,i}(u-2)}\right]\sum_{j=1}^{K}\delta_j\rho n^{-\frac{j}{2}}.
\end{aligned}$$

For the third term in above equation, we change the summation index. Let $\sum_{u=3}^{K+3}m_{u,i}(u-2)=h$. Note that for any $\sum_{u=3}^{K+3}m_{u,i}=i$, $i\leq h\leq i(K+1)$, $h/(K+1)\leq i\leq h$. Thus the third term summation can be rearranged as

$$\sum_{h=1}^{(K+1)^2}\left[\rho\left(\overline{X}\right)\sum_{\frac{h}{K+1}\leq i\leq h}\frac{1}{i!}\sum_{I_{i,h}}\binom{i}{m_{3,i},m_{4,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(\delta_u\hat{l}\right)^{m_{u,i}}\right]n^{-\frac{h}{2}}.$$

Similarly for the fourth term, let $\sum_{u=3}^{K+3}m_{u,i}(u-2)+j=h$, then the summation can be rearranged as

$$\sum_{h=2}^{(K+1)^2+K}\left[\sum_{j=1}^{h-1}\delta_j\rho\sum_{\frac{h-j}{K+1}\leq i\leq h-j}\frac{1}{i!}\sum_{I_{i,h-j}}\binom{i}{m_{3,i},m_{4,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(\delta_u\hat{l}\right)^{m_{u,i}}\right]n^{-\frac{h}{2}}.$$

We collect the same order term of n, and denote the summation of all the terms with order higher than K to be $R_K(Y)$, then we get the product as

$$\begin{aligned}
& \rho\left(\overline{X}\right)+\left(\delta_1\rho+\rho\left(\overline{X}\right)\delta_3\hat{l}\right)n^{-\frac{1}{2}} \\
& +\sum_{h=2}^{K}\left[\delta_h\rho+\sum_{j=0}^{h-1}\delta_j\rho\sum_{\frac{h-j}{K+1}\leq i\leq h-j}\frac{1}{i!}\sum_{I_{i,h-j}}\binom{i}{m_{3,i},m_{4,i},\ldots,m_{K+3,i}}\prod_{u=3}^{K+3}\left(\delta_u\hat{l}\right)^{m_{u,i}}\right]n^{-\frac{h}{2}}+R_K(Y).
\end{aligned}$$

Integral over any Borel set $A\cap H_n$, we can get the polynomial $P_K(A,n)$. Now we can prove the main theorem .

PROOF. Let $A_1 = \{|Y|_2 \geq \delta_2\sqrt{n}\}$ and $A_2 = \{|Y|_2 < \delta_2\sqrt{n}\}$. Then

$$\begin{aligned}
& \left|\int_{A\cap H_n}\prod_{i=1}^{n}\frac{\hat{w}_i(\theta)}{\hat{w}_i(\overline{X})}\rho(\theta)\,\mathrm{d}n^{-\frac{1}{2}}Y - P_K(A,n)\right| \\
= & \left|\int_{A\cap H_n}\prod_{i=1}^{n}\frac{\hat{w}_i(\theta)}{\hat{w}_i(\overline{X})}\rho(\theta) - \exp\left(-\frac{Y^TY}{2}\right)\sum_{h=0}^{K}\alpha_h(Y,n)\,n^{-\frac{h}{2}}\,\mathrm{d}n^{-\frac{1}{2}}Y\right| \\
\leq & \left|\int_{A\cap H_n\cap A_1}\prod_{i=1}^{n}\frac{\hat{w}_i(\theta)}{\hat{w}_i(\overline{X})}\rho(\theta) - \exp\left(-\frac{Y^TY}{2}\right)\sum_{h=0}^{K}\alpha_h(Y,n)\,n^{-\frac{h}{2}}\,\mathrm{d}n^{-\frac{1}{2}}Y\right| \\
& + \left|\int_{A\cap H_n\cap A_2}\prod_{i=1}^{n}\frac{\hat{w}_i(\theta)}{\hat{w}_i(\overline{X})}\rho(\theta) - \exp\left(-\frac{Y^TY}{2}\right)\sum_{h=0}^{K}\alpha_h(Y,n)\,n^{-\frac{h}{2}}\,\mathrm{d}n^{-\frac{1}{2}}Y\right|.
\end{aligned}$$

For the first term, by lemma 13, we have

$$\begin{aligned}
& \left|\int_{A\cap H_n\cap A_1}\prod_{i=1}^{n}\frac{\hat{w}_i(\theta)}{\hat{w}_i(\overline{X})}\rho(\theta) - \exp\left(-\frac{Y^TY}{2}\right)\sum_{h=0}^{K}\alpha_h(Y,n)\,n^{-\frac{h}{2}}\,\mathrm{d}n^{-\frac{1}{2}}Y\right| \\
\leq & \int_{A\cap H_n\cap A_1}\exp\left(n\left(\hat{l}(\theta) - \hat{l}(\overline{X})\right)\right)\rho(\theta)\,\mathrm{d}n^{-\frac{1}{2}}Y \\
& + \left|\int_{A\cap H_n\cap A_1}\exp\left(-\frac{Y^TY}{4} - \frac{Y^TY}{4}\right)\sum_{h=0}^{K}\alpha_h(Y,n)\,n^{-\frac{h}{2}}\,\mathrm{d}n^{-\frac{1}{2}}Y\right| \\
\leq & \exp(-n\varepsilon)\int_{A\cap H_n\cap A_1}\rho(\theta)\,\mathrm{d}B(\theta-\overline{X}) \\
& + \exp\left(-\frac{\delta_2^2 n}{4}\right)\left|\int_{A\cap H_n\cap A_1}\exp\left(-\frac{Y^TY}{4}\right)\sum_{h=0}^{K}\alpha_h(Y,n)\,n^{-\frac{h}{2}}\,\mathrm{d}n^{-\frac{1}{2}}Y\right| \\
\leq & \exp(-n\varepsilon)\int_{\mathbb{R}}\rho(\theta)\,\mathrm{d}B(\theta-\overline{X}) + \exp\left(-n\frac{\delta_2^2}{4}\right)\sum_{h=0}^{K}\left(\int_{A\cap H_n}\exp\left(-\frac{Y^TY}{4}\right)|\alpha_h(Y,n)|\,\mathrm{d}Y\right)n^{-\frac{h+1}{2}}.
\end{aligned}$$

Note that the above terms are exponentially decreasing with respect to n, so there exists an N_3, and M_3, such that for any $n \geq N_3$,

$$\left|\int_{A\cap H_n\cap A_1}\prod_{i=1}^{n}\frac{\hat{w}_i(\theta)}{\hat{w}_i(\overline{X})}\rho(\theta) - \exp\left(-\frac{Y^TY}{2}\right)\sum_{h=0}^{K}\alpha_h(Y,n)\,n^{-\frac{h}{2}}\,\mathrm{d}n^{-\frac{1}{2}}Y\right| \leq M_3 n^{-\frac{K+2}{2}}.$$

For the second term, by lemma 15, we have

$$\begin{aligned}
& \left| \int_{A\cap H_n\cap A_2} \prod_{i=1}^{n} \frac{\hat{w}_i(\theta)}{\hat{w}_i(\overline{X})} \rho(\theta) - \exp\left(-\frac{Y^TY}{2}\right) \sum_{h=0}^{K} \alpha_h(Y,n)\, n^{-\frac{h}{2}}\, \mathrm{d}n^{-\frac{1}{2}Y} \right| \\
\leq & \left| \int_{A\cap H_n\cap A_2} \prod_{i=1}^{n} \frac{\hat{w}_i(\theta)}{\hat{w}_i(\overline{X})} \rho(\theta) - \exp\left(-\frac{1}{2}Y^TY + \sum_{k=3}^{K+3} \delta_k \hat{l} n^{-\frac{k-2}{2}}\right) \rho_K(\theta)\, \mathrm{d}n^{-\frac{1}{2}Y} \right| \\
& + \left| \int_{A\cap H_n\cap A_2} \exp\left(-\frac{1}{2}Y^TY + \sum_{k=3}^{K+3} \delta_k \hat{l} n^{-\frac{k-2}{2}}\right) \rho_K(\theta) - \exp\left(-\frac{Y^TY}{2}\right) \sum_{h=0}^{K} \alpha_h(Y,n)\, n^{-\frac{h}{2}}\, \mathrm{d}n^{-\frac{1}{2}Y} \right| \\
\leq & M_2 n^{-\frac{K+2}{2}} \\
& + \left| \int_{A\cap H_n\cap A_2} \exp\left(-\frac{1}{2}Y^TY + \sum_{k=3}^{K+3} \delta_k \hat{l} n^{-\frac{k-2}{2}}\right) \rho_K(\theta) - \exp\left(-\frac{Y^TY}{2}\right) \sum_{h=0}^{K} \alpha_h(Y,n)\, n^{-\frac{h}{2}}\, \mathrm{d}n^{-\frac{1}{2}Y} \right|.
\end{aligned}$$

For the second term of above, we add and subtract $R_K(Y)$ in integrand, and by Taylor expansion,

$$\begin{aligned}
& \left| \int_{A\cap H_n\cap A_2} \exp\left(-\frac{1}{2}Y^TY + \sum_{k=3}^{K+3} \delta_k \hat{l} n^{-\frac{k-2}{2}}\right) \rho_K(\theta) - \exp\left(-\frac{Y^TY}{2}\right) \sum_{h=0}^{K} \alpha_h(Y,n)\, n^{-\frac{h}{2}}\, \mathrm{d}n^{-\frac{1}{2}Y} \right| \\
\leq & \left| \int_{A\cap H_n\cap A_2} \exp\left(-\frac{Y^TY}{2}\right) \rho_K(\theta) \left[\exp\left(\sum_{k=3}^{K+3} \delta_k \hat{l} n^{-\frac{k-2}{2}}\right) - \sum_{i=0}^{K+1} \frac{1}{i!} \left(\sum_{k=3}^{K+3} \delta_k \hat{l} n^{-\frac{k-2}{2}}\right)^i \right] \mathrm{d}n^{-\frac{1}{2}Y} \right| \\
& + \left| \int_{A\cap H_n\cap A_2} \exp\left(-\frac{Y^TY}{2}\right) R_K(Y)\, \mathrm{d}n^{-\frac{1}{2}Y} \right| \\
= & \left| \int_{A\cap H_n\cap A_2} \exp\left(-\frac{Y^TY}{2}\right) \rho_K(\theta) \frac{1}{(K+2)!} \exp(L) \left(\sum_{k=3}^{K+3} \delta_k \hat{l} n^{-\frac{k-2}{2}}\right)^{K+2} \mathrm{d}n^{-\frac{1}{2}Y} \right| \\
& + \left| \int_{A\cap H_n\cap A_2} \exp\left(-\frac{Y^TY}{2}\right) R_K(Y)\, \mathrm{d}n^{-\frac{1}{2}Y} \right|,
\end{aligned}$$

where $|L| \leq \left|\sum_{k=3}^{K+3} \delta_k \hat{l} n^{-\frac{k-2}{2}}\right|$. We know that $R_K(Y)$ is a polynomial with order $n^{-\frac{1}{2}(K+1)}$, so there exists an M_3, such that

$$\left| \int_{A\cap H_n\cap A_2} \exp\left(-\frac{Y^TY}{2}\right) R_K(Y)\, \mathrm{d}n^{-\frac{1}{2}Y} \right| \leq M_3 n^{-\frac{1}{2}(K+2)}.$$

For the first term,

$$\left| \int_{A\cap H_n\cap A_2} \exp\left(-\frac{Y^TY}{2}\right) \rho_K(\theta) \frac{1}{(K+2)!} \exp(L) \left(\sum_{k=3}^{K+3} \delta_k \hat{l} n^{-\frac{k-2}{2}}\right)^{K+2} \mathrm{d}n^{-\frac{1}{2}}Y \right| \tag{B.5.1}$$

$$\begin{aligned}
&\leq \frac{1}{(K+2)!} \left| \int_{A\cap H_n\cap A_2} \exp\left(-\frac{Y^TY}{2}\right) \rho_K(\theta) \exp\left(\left|\sum_{k=3}^{K+3} \delta_k \hat{l} n^{-\frac{k-2}{2}}\right|\right) \left(\sum_{k=3}^{K+3} \delta_k \hat{l} n^{-\frac{k-2}{2}}\right)^{K+2} \mathrm{d}n^{-\frac{1}{2}}Y \right| \\
&= \frac{1}{(K+2)!} \left| \int_{A\cap H_n\cap A_2} \rho_K(\theta) \exp\left(-n\left\{\frac{(\theta-\overline{X})^T B^2 (\theta-\overline{X})}{2} - \left|\sum_{k=3}^{K+3} \frac{\left[(\theta-\overline{X})^T\nabla\right]^k \hat{l}}{k!}\right|\right\}\right) \right. \\
&\quad \left. \left\{ n\sum_{k=3}^{K+3} \frac{\left[(\theta-\overline{X})^T\nabla\right]^k \hat{l}}{k!} \right\}^{K+2} \mathrm{d}B(\theta-\overline{X}) \right|.
\end{aligned}$$

We need $\delta_2$sufficiently small, so that there exist an C_4 and C_5, such that

$$\begin{aligned}
\frac{(\theta-\overline{X})^T B^2 (\theta-\overline{X})}{2} - \left|\sum_{k=3}^{K+3} \frac{\left[(\theta-\overline{X})^T\nabla\right]^k \hat{l}}{k!}\right| &\geq C_4 (\theta-\overline{X})^T B^2 (\theta-\overline{X}), \\
\sum_{k=3}^{K+3} \frac{\left[(\theta-\overline{X})^T\nabla\right]^k \hat{l}}{k!} &\leq C_5 \left[(\theta-\overline{X})^T\nabla\right]^3 \hat{l}.
\end{aligned}$$

Hence, equation (B.5.1) can be bounded by

$$\begin{aligned}
&\frac{n^{K+2}}{(K+2)!} \left| \int_{A\cap H_n\cap A_n} \exp\left(-nC_4 (\theta-\overline{X})^T B^2 (\theta-\overline{X})\right) \left\{C_5 \left[(\theta-\overline{X})^T\nabla\right]^3 \hat{l}\right\}^{K+2} \mathrm{d}B(\theta-\overline{X}) \right| \\
\leq\ & \frac{C_5^{K+2} n^{K+2}}{(K+2)!} \left| \int_{A\cap H_n} \exp\left(-C_4 Y^TY\right) \left(\delta_3\hat{l}\right)^{K+2} n^{-\frac{3(K+2)}{2}} \mathrm{d}n^{-\frac{1}{2}}Y \right| \\
\leq\ & \frac{C_5^{K+2}}{(K+2)!} \left| \int_{A\cap H_n} \exp\left(-C_4 Y^TY\right) \left(\delta_3\hat{l}\right)^{K+2} \mathrm{d}Y \right| n^{-\frac{K+3}{2}}.
\end{aligned}$$

Add all the parts together, we get the inequality in . □

APPENDIX C

Proof of Approximate Bayesian Computation via Sufficient Dimension Reduction

C.1. Proof of Theorem 3

PROOF.

$$\begin{aligned} & P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right)\leq t \mid \hat{\theta}\in O\left(\theta_0,\varepsilon\right)\right) \\ = & \frac{P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right)\leq t, \hat{\theta}\in O\left(\theta_0,\varepsilon\right)\right)}{P\left(\hat{\theta}\in O\left(\theta_0,\varepsilon\right)\right)} \\ = & \frac{\int I_{\left[\hat{\theta}\in O(\theta_0,\varepsilon)\right]} I_{\left[\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right)\leq t\right]} \prod_{i=1}^{n}\pi\left(X_i \mid \theta\right)\pi\left(\theta\right)\,\mathrm{d}X_i\,\mathrm{d}\theta}{\int I_{\left[\hat{\theta}\in O(\theta_0,\varepsilon)\right]} \prod_{i=1}^{n}\pi\left(X_i \mid \theta\right)\pi\left(\theta\right)\,\mathrm{d}X_i\,\mathrm{d}\theta}. \end{aligned}$$

Let $P^{\infty}\left(\theta\right)$ be the probability measure on infinite independent and identically distributed sequence $X_1,\ldots,X_n,\ldots,$. Then

$$\begin{aligned} & P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right)\leq t \mid \hat{\theta}\in O\left(\theta_0,\varepsilon\right)\right) \\ = & \frac{E_{\pi(\theta)}E_{P^{\infty}(\theta_0)}I_{\left[\hat{\theta}\in O(\theta_0,\varepsilon)\right]} I_{\left[\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right)\leq t\right]} \exp\left(\sum_{i=1}^{n}\log\pi\left(X_i \mid \theta\right)-\log\pi\left(X_i \mid \theta_0\right)\right)}{E_{\pi(\theta)}E_{P^{\infty}(\theta_0)}I_{\left[\hat{\theta}\in O(\theta_0,\varepsilon)\right]} \exp\left(\sum_{i=1}^{n}\log\pi\left(X_i \mid \theta\right)-\log\pi\left(X_i \mid \theta_0\right)\right)}. \end{aligned} \tag{C.1.1}$$

By strong consistency of the maximum likelihood estimator, for any $\varepsilon>0$, there is an $N>0$, such that for any $n>N$, $P\left(\hat{\theta}\in O\left(\theta_0,\varepsilon\right)\mid\theta_0\right)=1$. Thus, we can drop the indicator $I_{\left[\hat{\theta}\in O(\theta_0,\varepsilon)\right]}$ without changing the value in (C.1.1). We can change the order of integration. So the numerator of (C.1.1) is

$$\begin{aligned} & E_{P^{\infty}(\theta_0)}\left(\prod_{i=1}^{n}\pi\left(X_i \mid \theta_0\right)\right)^{-1} E_{\pi(\theta)} I_{\left[\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right)\leq t\right]} \prod_{i=1}^{n}\pi\left(X_i \mid \theta\right) \\ = & E_{P^{\infty}(\theta_0)}\left(\prod_{i=1}^{n}\pi\left(X_i \mid \theta_0\right)\right)^{-1} \int I_{\left[\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right)\leq t\right]} \prod_{i=1}^{n}\pi\left(X_i \mid \theta\right)\pi\left(\theta\right)\,\mathrm{d}\theta \\ = & E_{P^{\infty}(\theta_0)}\left(\prod_{i=1}^{n}\pi\left(X_i \mid \theta_0\right)\right)^{-1} P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right)\leq t \mid X_1,\ldots,X_n\right) E_{\pi(\theta)}\prod_{i=1}^{n}\pi\left(X_i \mid \theta\right). \end{aligned}$$

By Bernstein–von Mises theorem,

$$\lim_{n\to\infty} P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right)\leq t \mid X_1,\ldots,X_n\right)=\Phi\left(t\right), \text{ a.s. } P^{\infty}\left(\theta_0\right).$$

Hence, the result holds. □

To prove a similar result about conditioning on posterior mean, we go through similar steps as in the proof of Theorem (3). The only needed change is to prove

$$\lim_{n\to\infty} P\left(\sqrt{n\hat{I}}\left(\theta - E\left(\theta \mid X_1,\ldots,X_n\right)\right) \le t \mid X_1,\ldots,X_n\right) = \Phi\left(t\right), \text{ a.s. } P^{\infty}\left(\theta_0\right).$$

We know from Ghosh and Liu (2011) that with probability 1, (5.2.2) holds. By conditioning on $X_1,\ldots,X_n$, both posterior mean and maximum likelihood estimator are fixed numbers and

$$\lim_{n\to\infty} \sqrt{n}\left(E\left(\theta \mid X_1,\ldots,X_n\right) - \hat{\theta}\right) = 0. \text{ a.s.}.$$

Hence, if we assume the CDF of the full posterior is continuous and asymptotically normal, then

$$\begin{aligned}
& \left|P\left(\sqrt{n\hat{I}}\left(\theta - E\left(\theta \mid X_1,\ldots,X_n\right)\right) \le t \mid X_1,\ldots,X_n\right) - \Phi\left(t\right)\right| \\
\le\ & \left|P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right) + \hat{I}^{1/2}\sqrt{n}\left(E\left(\theta \mid X_1,\ldots,X_n\right) - \hat{\theta}\right) \le t \mid X_1,\ldots,X_n\right)\right. \\
& \left. -P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right) \le t \mid X_1,\ldots,X_n\right)\right| + \left|P\left(\sqrt{n\hat{I}}\left(\theta-\hat{\theta}\right) \le t \mid X_1,\ldots,X_n\right) - \Phi\left(t\right)\right| \\
\to\ & 0, \text{as}\left(n\to\infty\right).
\end{aligned}$$

C.2. Derivation of Examples

C.2.1. Derivation of Example 2. First, we get the distribution of μ.

$$\mu^{r_2}\exp\left(-\mu\frac{r_2}{\hat{\mu}}\right)\left|-\frac{r_2}{\left(\hat{\mu}\right)^2}\right| \propto r_2\mu^{r_2}\exp\left(-\frac{\mu}{\hat{\mu}}r_2\right).$$

Now we sum out r_2. By definition $X_0 + r_1 - r_2 \ge 0$. Hence the distribution of $\hat{\mu}$ is proportional to

$$\sum_{r_2=0}^{X_0+r_1} r_2\left(\mu\exp\left(-\frac{\mu}{\hat{\mu}}\right)\right)^{r_2}.$$

Let $U = \mu\exp\left(-\mu/\hat{\mu}\right)$ and $R = X_0 + r_1 = X_0 + \hat{\lambda}T$. Let

$$\begin{aligned}
L &= \sum_{r_2=0}^{R} r_2U^{r_2} = U\frac{\mathrm{d}}{\mathrm{d}U}\sum_{r_2=0}^{R}U^{r_2} = U\frac{\mathrm{d}}{\mathrm{d}U}\left(\frac{1-U^{R+1}}{1-U}\right) \\
&= U\left(1-U\right)^{-2}\left[1-\left(R+1\right)U^R + RU^{R+1}\right].
\end{aligned}$$

For fixed t, consider $\mu = \hat{\mu} + t/\sqrt{T}$.

$$\begin{aligned}
\log U &= -\frac{\mu}{\hat{\mu}} + \log\mu = -1 - \frac{t}{\sqrt{T}\hat{\mu}} + \log\hat{\mu} + \log\left(1+\frac{t}{\sqrt{T}\hat{\mu}}\right) \\
&= \log\hat{\mu} - 1 - \frac{t^2}{2\left(\hat{\mu}\right)^2 T} + o\left(T^{-1}\right).
\end{aligned}$$

Hence

$$\lim_{T\to\infty} U = \frac{\mu_0}{e}, \text{ a.s.}.$$

the limit is a constant.

$$R\log U=\left(X_0+\hat{\lambda}T\right)\left(\log\hat{\mu}-1\right)-\frac{\hat{\lambda}t^2}{2\left(\hat{\mu}\right)^2}+o\left(1\right).$$

$$L=\frac{UR}{\left(1-U\right)^2}\left(\frac{1}{R}-\frac{R+1}{R}\exp\left(R\log U\right)+U\exp\left(R\log U\right)\right)$$

Note that the density of $t=\sqrt{T}\left(\mu-\hat{\mu}\right)$ is only proportional to L, hence only the terms containing t will affect the limit distribution, other terms can be omitted. Also recall that,

$$\lim_{T\to\infty}\frac{1}{R}=\lim_{T\to\infty}\frac{1}{X_0+\hat{\lambda}T}=0,$$

we have

$$\begin{aligned}\lim_{T\to\infty}L &\propto \lim_{T\to\infty}\left(\frac{1}{R}-\frac{R+1}{R}\exp\left(R\log U\right)+U\exp\left(R\log U\right)\right)\\ &= \lim_{T\to\infty}\left(U-1-\frac{1}{R}\right)\exp\left(\left(X_0+\hat{\lambda}T\right)\left(\log\hat{\mu}-1\right)-\frac{\hat{\lambda}t^2}{2\left(\hat{\mu}\right)^2}+o\left(1\right)\right)\\ &= \lim_{T\to\infty}\left(\frac{\mu_0}{e}-1\right)\exp\left(\left(X_0+\hat{\lambda}T\right)\left(\log\hat{\mu}-1\right)\right)\exp\left(-\frac{\hat{\lambda}t^2}{2\left(\hat{\mu}\right)^2}+o\left(1\right)\right)\\ &\propto \exp\left(-\frac{\hat{\lambda}t^2}{2\left(\hat{\mu}\right)^2}\right).\end{aligned}$$

C.2.2. Derivation of Example 3. We know $\overline{X}\sim\text{Gamma}\left(n\alpha,\beta/n\right)$, so $\tilde{\alpha}\sim\text{Gamma}\left(n\alpha,n^{-1}\right)$.

$$\begin{aligned}\pi\left(\alpha\mid\tilde{\alpha}\right) &\propto \pi\left(\tilde{\alpha}\mid\alpha\right)\pi\left(\alpha\right)=\frac{1}{\Gamma\left(n\alpha\right)n^{-n\alpha}}\left(\tilde{\alpha}\right)^{n\alpha-1}\exp\left(-n\tilde{\alpha}\right)\exp\left(-\lambda\alpha\right)\\ &\propto \frac{\left(n\tilde{\alpha}\exp\left(-\lambda/n\right)\right)^{n\alpha}}{\Gamma\left(n\alpha\right)}.\end{aligned}$$

Next we will show $\pi\left(\sqrt{n}\left(\alpha-\tilde{\alpha}\right)/b\mid\tilde{\alpha}\right)\to N\left(0,1\right)$ a.s., for some suitable b. The PDF of t is proportional to

$$\frac{\left(n\tilde{\alpha}\exp\left(-\lambda/n\right)\right)^{n\left(bt/\sqrt{n}+\tilde{\alpha}\right)}}{\Gamma\left(n\left(bt/\sqrt{n}+\tilde{\alpha}\right)\right)}\frac{b}{\sqrt{n}}.$$

Take logarithm, and drop all the terms not related to t, since those terms can be divided from both numerator and denominator,

$$\sqrt{n}bt\log\left(n\tilde{\alpha}\right)-\lambda b\frac{t}{\sqrt{n}}-\log\Gamma\left(n\tilde{\alpha}\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)\right). \tag{C.2.1}$$

Using Stirling formula to approximate gamma function,

$$\begin{aligned}&\log\Gamma\left(n\tilde{\alpha}\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)\right)\\ \approx\ &\left[n\tilde{\alpha}\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)-\frac{1}{2}\right]\log\left(n\tilde{\alpha}\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)\right)-n\tilde{\alpha}\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)+\frac{1}{2}\log 2\pi.\end{aligned}$$

So (C.2.1) can be written as

$$\begin{aligned}
&\sqrt{n}bt\log\left(n\tilde{\alpha}\right)-\lambda b\frac{t}{\sqrt{n}}-\sqrt{n}bt\log\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)\\
&-\sqrt{n}bt\log\left(n\tilde{\alpha}\right)-\left(n\tilde{\alpha}-\frac{1}{2}\right)\log\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)+\sqrt{n}bt\\
&-\sqrt{n}bt\log\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)-\left(n\tilde{\alpha}-\frac{1}{2}\right)\log\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)+\sqrt{n}bt-\lambda b\frac{t}{\sqrt{n}}. \qquad \text{(C.2.2)}
\end{aligned}$$

Now we can apply Taylor expansion for term $\log\left(1+bt/\left(\sqrt{n}\tilde{\alpha}\right)\right)$,

$$\log\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)=\frac{bt}{\sqrt{n}\tilde{\alpha}}-\frac{b^2t^2}{2n\tilde{\alpha}^2}+\frac{b^3t^3}{3n^{3/2}\tilde{\alpha}^3}+o\left(\frac{t^3}{n^{3/2}}\right).$$

Substituting the expansion into (C.2.2),

$$\begin{aligned}
&-\sqrt{n}bt\log\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)-\left(n\tilde{\alpha}-\frac{1}{2}\right)\log\left(1+\frac{bt}{\sqrt{n}\tilde{\alpha}}\right)+\sqrt{n}bt-\lambda b\frac{t}{\sqrt{n}}\\
=\;&-\sqrt{n}bt\left(\frac{bt}{\sqrt{n}\tilde{\alpha}}-\frac{b^2t^2}{2n\tilde{\alpha}^2}+\frac{b^3t^3}{3n^{3/2}\tilde{\alpha}^3}+o\left(\frac{t^3}{n^{3/2}}\right)\right)\\
&-\left(n\tilde{\alpha}-\frac{1}{2}\right)\left(\frac{bt}{\sqrt{n}\tilde{\alpha}}-\frac{b^2t^2}{2n\tilde{\alpha}^2}+\frac{b^3t^3}{3n^{3/2}\tilde{\alpha}^3}+o\left(\frac{t^3}{n^{3/2}}\right)\right)+\sqrt{n}bt-\lambda b\frac{t}{\sqrt{n}}\\
=\;&-\frac{b^2t^2}{\tilde{\alpha}}+\frac{b^3t^3}{2\sqrt{n}\tilde{\alpha}^2}-o\left(\frac{t^3}{\sqrt{n}}\right)-\sqrt{n}bt+\frac{b^2t^2}{2\tilde{\alpha}}-\frac{b^3t^3}{2\sqrt{n}\alpha^2}-o\left(\frac{t^3}{\sqrt{n}}\right)\\
&+\frac{bt}{2\sqrt{n}\tilde{\alpha}}-\frac{b^2t^2}{4n\tilde{\alpha}^2}+o\left(\frac{t^2}{n}\right)+\sqrt{n}bt-\lambda b\frac{t}{\sqrt{n}}\\
\approx\;&-\frac{b^2t^2}{2\tilde{\alpha}}.
\end{aligned}$$

If we set $b=\sqrt{\tilde{\alpha}}$, then the rescaled partial posterior convergence to standard normal.

C.2.3. Derivation of Example 4. First we prove lemma 7.

PROOF. Assume Y has a probability density function $f(y)$. Let $y=u(x,\theta)$ be the solution of equation $x=h(y,\theta)$. Then $h(Y,\theta)$ has a probability density function

$$f\left(u\left(x,\theta\right)\right)\left|\frac{\partial u\left(x,\theta\right)}{\partial x}\right|.$$

Then the posterior distribution under the uniform prior is proportional to

$$f\left(u\left(X,\theta\right)\right)\left|\frac{\partial u\left(X,\theta\right)}{\partial x}\right|.$$

Now we find the probability density function of $g(Y,X)$. By assumptions, we know $y=u(x,\theta)$ is also the solution of $\theta=g(y,x)$. Hence the probability density function is also

$$f\left(u\left(X,\theta\right)\right)\left|\frac{\partial u\left(X,\theta\right)}{\partial x}\right|.$$

□

We know if X, Y are independent exponential random variables with mean λ, then $X - Y$ is Laplace distributed with $\mu = 0$ and the same λ. So we know our sample Z has the same distribution as $X - Y + \mu$. So the sample mean $\overline{Z}$ has the same distribution as $\overline{X} - \overline{Y} + \mu$. It is easy to check $\overline{X}$ and $\overline{Y}$ have gamma distribution with location parameter n and scale parameter $n^{-1}\lambda$. Hence the posterior distribution of μon $\overline{Z}$ under the uniform prior has the same distribution as $\overline{Z} - \left(\overline{X} - \overline{Y}\right)$. Hence the posterior distribution $\sqrt{n}\left(\mu - \overline{Z}\right)$ has the same distribution as $-\sqrt{n}\left(\overline{X} - \overline{Y}\right)$ We know the characteristic function of $\sqrt{n}\overline{X}$ is

$$\left[1 - \frac{\lambda}{n} i\left(\sqrt{n}t\right)\right]^{-n},$$

So the characteristic function of $-\sqrt{n}\left(\overline{X} - \overline{Y}\right)$ is

$$\left[1 - \frac{\lambda}{n} i\left(\sqrt{n}t\right)\right]^{-n}\left[1 - \frac{\lambda}{n} i\left(-\sqrt{n}t\right)\right]^{-n} = \left(1 + \frac{\lambda^2 t^2}{n}\right)^{-n} \to \exp\left(-\lambda^2 t^2\right).$$

Hence $\sqrt{n}\left(\mu - \overline{Z}\right)$ has an asymptotic normal distribution with zero mean and variance $2\lambda^2$.

C.3. Proof of Theorem 4

LEMMA 16. *Under Assumptions 10, 11 and 12, for any ε, δ_1 and δ_2, there exists an N, such that for any $n \geq N$,*

$$P_{\theta_0}^{\infty}\left(\omega : P_{\omega}^{n}\left(\sqrt{n}\left|G\left(\theta, \tilde{\theta}\right) - G_1\left(\hat{\theta}, \tilde{\theta}\right)\left(\theta - \tilde{\theta}\right)\right| \leq 2\varepsilon\right) \geq 1 - \delta_1\right) \geq 1 - \delta_2.$$

PROOF. By Taylor expansion and Assumption 10,

$$\left|G\left(\theta, \tilde{\theta}\right) - G\left(\hat{\theta}, \tilde{\theta}\right) - G_1\left(\hat{\theta}, \tilde{\theta}\right)\left(\theta - \hat{\theta}\right)\right| \leq L\left(\theta - \hat{\theta}\right)^2, \tag{C.3.1}$$

and

$$\left|G\left(\tilde{\theta}, \tilde{\theta}\right) - G\left(\hat{\theta}, \tilde{\theta}\right) - G_1\left(\hat{\theta}, \tilde{\theta}\right)\left(\tilde{\theta} - \hat{\theta}\right)\right| \leq L\left(\tilde{\theta} - \hat{\theta}\right)^2. \tag{C.3.2}$$

By posterior consistency, there exists a $\Omega_1 \subset \Omega$, $P_{\theta_0}^{\infty}(\Omega_1) = 1$, such that for any $\omega \in \Omega_1$, random variable $(\theta \mid X_1(\omega), \ldots, X_n(\omega))$ converges in probability to θ_0. Hence $(\theta - \theta_0) \mid X_1(\omega), \ldots, X_n(\omega) = o_{P_{\omega}^n}(1)$. By Bernstein–von Mises, there exists a $\Omega_2 \subset \Omega$, $P_{\theta_0}^{\infty}(\Omega_2) = 1$, such that for any $\omega \in \Omega_2$, random variable $\left(\sqrt{n\hat{I}}\left(\theta - \hat{\theta}\right) \mid X_1(\omega), \ldots, X_n(\omega)\right)$ converges in distribution to a standard normal random variable. Hence $\left(\sqrt{n\hat{I}}\left(\theta - \hat{\theta}\right) \mid X_1(\omega), \ldots, X_n(\omega)\right) = O_{P_{\omega}^n}(1)$. For any $\omega \in \Omega_1 \cap \Omega_2$,

$$\begin{aligned}\left(\sqrt{n}L\left(\theta - \hat{\theta}\right)^2 \mid X_1(\omega), \ldots, X_n(\omega)\right) &= L\left(\sqrt{n}\left(\theta - \hat{\theta}\right) \times \left(\theta - \hat{\theta}\right) \mid X_1(\omega), \ldots, X_n(\omega)\right) \\ &= LO_{P_{\omega}^n}(1) \times o_{P_{\omega}^n}(1) = o_{P_{\omega}^n}(1),\end{aligned}$$

which means for any ε and δ_1, there exists an N_1 such that

$$P_{\omega}^{n}\left(\sqrt{n}L\left(\theta - \hat{\theta}(\omega)\right)^2 \leq \varepsilon \mid X_1(\omega), \ldots, X_n(\omega)\right) \geq 1 - \delta_1.$$

By Assumptions 11, $\tilde{\theta}-\theta_0 = o_{P^\infty_{\theta_0}}(1)$, $\sqrt{n}\left(\tilde{\theta}-\theta_0\right) = O_{P^\infty_{\theta_0}}(1)$, $\hat{\theta}-\theta_0 = o_{P^\infty_{\theta_0}}(1)$ and $\sqrt{n}\left(\hat{\theta}-\theta_0\right) = O_{P^\infty_{\theta_0}}(1)$ .Then

$$\sqrt{n}L\left(\tilde{\theta}-\hat{\theta}\right)^2 \leq 2L\left[\sqrt{n}\left(\tilde{\theta}-\theta_0\right)^2 + \sqrt{n}\left(\hat{\theta}-\theta_0\right)^2\right] = o_{P^\infty_{\theta_0}}(1),$$

which means for any ε and δ_2, there exists an N_2, such that for any $n \geq N_2$,

$$P^\infty_{\theta_0}\left(\omega : \sqrt{n}L\left(\tilde{\theta}(\omega)-\hat{\theta}(\omega)\right)^2 \leq \varepsilon\right) \geq 1-\delta_2.$$

Let $\Omega_\varepsilon = \left\{\omega : \sqrt{n}L\left(\tilde{\theta}(\omega)-\hat{\theta}(\omega)\right)^2 \leq \varepsilon\right\}$. For any $\omega \in \Omega_1 \cap \Omega_2 \cap \Omega_\varepsilon$,

$$\begin{aligned}
& \sqrt{n}\left|G\left(\theta,\tilde{\theta}\right) - G_1\left(\hat{\theta},\tilde{\theta}\right)\left(\theta-\tilde{\theta}\right)\right| \\
\leq\ & \sqrt{n}\left|G\left(\theta,\tilde{\theta}\right) - G\left(\hat{\theta},\tilde{\theta}\right) - G_1\left(\hat{\theta},\tilde{\theta}\right)\left(\theta-\hat{\theta}\right)\right| \\
& + \sqrt{n}\left|G\left(\tilde{\theta},\tilde{\theta}\right) - G\left(\hat{\theta},\tilde{\theta}\right) - G_1\left(\hat{\theta},\tilde{\theta}\right)\left(\tilde{\theta}-\hat{\theta}\right)\right| \\
\leq\ & \sqrt{n}L\left(\theta-\hat{\theta}\right)^2 + \sqrt{n}L\left(\tilde{\theta}-\hat{\theta}\right)^2 \leq 2\varepsilon,
\end{aligned}$$

with probability $1-\delta_1$ (P^n_ω). Also recall $P^\infty_{\theta_0}(\Omega_1 \cap \Omega_2 \cap \Omega_\varepsilon) \geq 1-\delta_2$. Hence for and $n \geq \max\{N_1, N_2\}$,

$$P^\infty_{\theta_0}\left(\omega : P^n_\omega\left(\sqrt{n}\left|G\left(\theta,\tilde{\theta}\right) - G_1\left(\hat{\theta},\tilde{\theta}\right)\left(\theta-\tilde{\theta}\right)\right| \leq 2\varepsilon\right) \geq 1-\delta_1\right) \geq 1-\delta_2.$$

□

Remark 4. This result is weaker than the settings in posterior consistency and Bernstein–von Mises theorem. In posterior consistency, the posterior random variable $(\theta \mid X_1, \ldots, X_n)$ is convergence in probability in P^n_ω almost surely in $P^\infty_{\theta_0}$. Similar statement can be expressed in Bernstein–von Mises. However, in this lemma, the rescaled posterior random variable is convergence in probability in P^n_ω in probability in $P^\infty_{\theta_0}$.

Remark 5. There is slight difference between the rescaled posterior random variable in this lemma and in

(C.3.3)

$$\sqrt{n}\left[\int_{\mathbb{R}} g\left(x,\tilde{\theta}\right)\pi(x \mid \theta)\,\mathrm{d}x - \int_{\mathbb{R}} g\left(x,\tilde{\theta}\right)\pi\left(x \mid \tilde{\theta}\right)\mathrm{d}x - \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{\mathbb{R}} g\left(x,\tilde{\theta}\right)\pi(x \mid \theta)\,\mathrm{d}x\Big|_{\theta=\tilde{\theta}}\right)\left(\theta-\tilde{\theta}\right)\right] \to 0, \text{ a.s.},$$

. The first order differential term is $G_1\left(\tilde{\theta},\tilde{\theta}\right)$. However, since $G_1\left(\tilde{\theta},\tilde{\theta}\right) \to G_1(\theta_0,\theta_0)$ and $G_1\left(\hat{\theta},\tilde{\theta}\right) \to G_1(\theta_0,\theta_0)$ almost surely in $P^\infty_{\theta_0}$and $\sqrt{n}$ term is absorbed by $\left(\theta-\tilde{\theta}\right)$, we have proved a weak version of (C.3.3).

Proof. Under the Assumptions 10, 11 and 12, we have the result from Lemma 16,

(C.3.4) $P^\infty_{\theta_0}\left(\omega : P^n_\omega\left(\sqrt{n}\left|G\left(\theta,\tilde{\theta}\right) - G_1\left(\hat{\theta},\tilde{\theta}\right)\left(\theta-\tilde{\theta}\right)\right| \leq 2\varepsilon\right) \geq 1-\delta_1\right) \geq 1-\delta_2.$

Let $\Omega_1 = \left\{\omega : P_\omega^n \left(\sqrt{n}\left|G\left(\theta,\tilde{\theta}\right) - G_1\left(\hat{\theta},\tilde{\theta}\right)\left(\theta-\tilde{\theta}\right)\right| \le 2\varepsilon\right) \ge 1-\delta_1\right\}$, $C_n = E_{\pi(\theta)}E_{P^\infty(\theta_0)}I_{\left[\tilde{\theta}\in O(\theta_0,\varepsilon)\right]}\exp\left(\sum_{i=1}^n\right.$
Now by the same technique used in the proof of Theorem 3, we have for sufficiently large N,

$$\begin{aligned}
& \left| P\left(\frac{\sqrt{n}\left(\theta-\tilde{\theta}\right)}{\sqrt{\tilde{V}}} \le t \middle| \tilde{\theta}\in O\left(\theta_0,\varepsilon\right)\right) - \Phi\left(t\right)\right| \\
= & \left| \frac{E_{\pi(\theta)}E_{P^\infty(\theta_0)}I_{\left[\tilde{\theta}\in O(\theta_0,\varepsilon)\right]}I_{\left[\sqrt{n\tilde{V}^{-1}}\left(\theta-\tilde{\theta}\right)\le t\right]}\prod_{i=1}^n f\left(X_i\mid\theta\right)/f\left(X_i\mid\theta_0\right)}{E_{\pi(\theta)}E_{P^\infty(\theta_0)}I_{\left[\tilde{\theta}\in O(\theta_0,\varepsilon)\right]}\prod_{i=1}^n f\left(X_i\mid\theta\right)/f\left(X_i\mid\theta_0\right)} - \Phi\left(t\right)\right| \\
\le & C_n^{-1}E_{\pi(\theta)}E_{P^\infty(\theta_0)}I_{\Omega_1}\left|P\left(\sqrt{n\tilde{V}^{-1}}\left(\theta-\tilde{\theta}\right)\le t\mid X_1,\ldots,X_n\right) - \Phi\left(t\right)\right|\prod_{i=1}^n f\left(X_i\mid\theta\right)/f\left(X_i\mid\theta_0\right) \\
& + C_n^{-1}E_{\pi(\theta)}E_{P^\infty(\theta_0)}I_{\Omega_1^c}\left|P\left(\sqrt{n\tilde{V}^{-1}}\left(\theta-\tilde{\theta}\right)\le t\mid X_1,\ldots,X_n\right) - \Phi\left(t\right)\right|\prod_{i=1}^n f\left(X_i\mid\theta\right)/f\left(X_i\mid\theta_0\right).
\end{aligned}$$

First considering the samples within Ω_1, by (C.3.4), $\sqrt{V_0^{-1}}\left\{\sqrt{n}\left[G_1\left(\tilde{\theta},\tilde{\theta}\right)\left(\theta-\tilde{\theta}\right) - G\left(\theta,\tilde{\theta}\right)\right]\right\}$ converges in probability to 0. By Theorem 2.1 in Rivoirard et al. (2012), we have,

(C.3.5)
$$\lim_{n\to\infty}\sup_{t\in\mathbb{R}}\left|P\left(\sqrt{nV_0^{-1}}\left(G\left(\theta,\tilde{\theta}\right) - \frac{1}{n}\sum_{i=1}^n g\left(X_i,\tilde{\theta}\right)\right)\le t\mid X_1,\ldots,X_n\right) - \Phi\left(t\right)\right| = 0, \text{ a.s.}\,\theta_0,$$

where

$$V_0 = \int_{\mathbb{R}}\left(g\left(x,\tilde{\theta}\right) - \int_{\mathbb{R}} g\left(y,\tilde{\theta}\right)\pi\left(y\mid\theta_0\right)\,\mathrm{d}y\right)^2\pi\left(x\mid\theta_0\right)\,\mathrm{d}x.$$

Hence, $\sqrt{nV_0^{-1}}\left(G\left(\theta,\tilde{\theta}\right) - n^{-1}\sum_{i=1}^n g\left(X_i,\tilde{\theta}\right)\right)$ converges in distribution to standard normal distribution. By the definition of M-estimator, $n^{-1}\sum_{i=1}^n g\left(X_i,\tilde{\theta}\right) = 0$. Assume that for every θ, the equation in t, $\int_{\mathbb{R}} g\left(x,t\right)\pi\left(x\mid\theta\right)\,\mathrm{d}x = 0$ has only one solution $t=\theta$, then

$$G\left(\tilde{\theta},\tilde{\theta}\right) = \int_{\mathbb{R}} g\left(x,\tilde{\theta}\right)\pi\left(x\mid\tilde{\theta}\right)\,\mathrm{d}x = 0.$$

Hence,

$$\sqrt{n\tilde{V}^{-1}}\left(\theta-\tilde{\theta}\right) = \sqrt{nV_0^{-1}}\left(G\left(\theta,\tilde{\theta}\right) - \frac{1}{n}\sum_{i=1}^n g\left(X_i,\tilde{\theta}\right)\right) + \sqrt{V_0^{-1}}\left\{\sqrt{n}\left[G_1\left(\tilde{\theta},\tilde{\theta}\right)\left(\theta-\tilde{\theta}\right) - G\left(\theta,\tilde{\theta}\right)\right]\right\},$$

by Slutsky's theorem, converges in distribution to standard normal distribution. Hence for large N,

$$\sup_{t\in\mathbb{R}}\left|P\left(\sqrt{n\tilde{V}^{-1}}\left(\theta-\tilde{\theta}\right)\le t\mid X_1,\ldots,X_n\right) - \Phi\left(t\right)\right| \le \varepsilon,$$

and

$$C_n^{-1}E_{\pi(\theta)}E_{P^\infty(\theta_0)}I_{\Omega_1}\sup_{t\in\mathbb{R}}\left|P\left(\sqrt{n\tilde{V}^{-1}}\left(\theta-\tilde{\theta}\right)\leq t\mid X_1,\ldots,X_n\right)-\Phi(t)\right|\prod_{i=1}^{n}f(X_i\mid\theta)/f(X_i\mid\theta_0)$$

$$\leq \varepsilon C_n^{-1}E_{\pi(\theta)}E_{P^\infty(\theta_0)}I_{\Omega_1}\prod_{i=1}^{n}f(X_i\mid\theta)/f(X_i\mid\theta_0)=\varepsilon. \tag{C.3.6}$$

Next, considering the samples outside Ω_1. It is trivial that

$$\sup_{t\in\mathbb{R}}\left|P\left(\sqrt{n\tilde{V}^{-1}}\left(\theta-\tilde{\theta}\right)\leq t\mid X_1,\ldots,X_n\right)-\Phi(t)\right|\leq 2.$$

By Assumption 13 and the strong law of large numbers, and the property of the Kullback-Leibler information number

$$\prod_{i=1}^{n}f(X_i\mid\theta)/f(X_i\mid\theta_0)=\exp\left(n\left(\frac{1}{n}\sum_{i=1}^{n}\log f(X_i\mid\theta)-\frac{1}{n}\sum_{i=1}^{n}\log f(X_i\mid\theta_0)\right)\right)\leq 1.\text{ a.s. }(P_{\theta_0})$$

Hence

$$C_n^{-1}E_{\pi(\theta)}E_{P^\infty(\theta_0)}I_{\Omega_1^c}\sup_{t\in\mathbb{R}}\left|P\left(\sqrt{n\tilde{V}^{-1}}\left(\theta-\tilde{\theta}\right)\leq t\mid X_1,\ldots,X_n\right)-\Phi(t)\right|\prod_{i=1}^{n}f(X_i\mid\theta)/f(X_i\mid\theta_0).$$

$$\leq 2C_n^{-1}E_{\pi(\theta)}E_{P^\infty(\theta_0)}I_{\Omega_1^c}=2P^\infty(\Omega_1^c\mid\theta_0)=2\delta_2. \tag{C.3.7}$$

Hence, combining (C.3.6) and (C.3.7),

$$\sup_{t\in\mathbb{R}}\left|P\left(\frac{\sqrt{n}\left(\theta-\tilde{\theta}\right)}{\sqrt{\tilde{V}}}\leq t\middle|\tilde{\theta}\in O(\theta_0,\varepsilon)\right)-\Phi(t)\right|\leq\varepsilon+2\delta_2.\text{ a.s.}$$

□

C.4. Proof of Theorem 5

Proof. By Theorem 2.1 in Rivoirard et al. (2012), we have

(C.4.1)

$$\lim_{n\to\infty}\sup_{t\in\mathbb{R}}\left|P\left(\sqrt{n\mathrm{Var}_{\theta_0}\left(a^Tg(X)\right)^{-1}}\left(\int a^Tg(x)f(x\mid\theta)\,\mathrm{d}x-\frac{1}{n}\sum_{i=1}^{n}a^Tg(X_i)\right)\leq t\mid X_1,\ldots,X_n\right)-\Phi(t)\right|=0,\text{ a}$$

By Assumption 16 and Slutsky's theorem, we know $n^{-1/2}\sum_{i=1}^{n}g(X_i)$ has the same asymptotic distribution as S. However, by central limit theorem, $n^{-1/2}\sum_{i=1}^{n}g(X_i)$ has an asymptotic normal distribution with variance matrix as $\mathrm{Var}_{\theta_0}(g(X))$. Hence, $\mathrm{Var}_{\theta_0}(g(X))=\Sigma(\theta_0)=\lim_{n\to\infty}\tilde{\Sigma}$, a.s.. Hence, we can replace $\mathrm{Var}_{\theta_0}(g(X))$ in (C.4.1) by its strong consistent estimator, and get

$$\lim_{n\to\infty}\sup_{t\in\mathbb{R}}\left|P\left(\sqrt{n\left(a^T\tilde{\Sigma}a\right)^{-1}}\left(\int a^Tg(x)f(x\mid\theta)\,\mathrm{d}x-n^{-1}\sum_{i=1}^{n}a^Tg(X_i)\right)\leq t\mid X_1,\ldots,X_n\right)-\Phi(t)\right|=0,\text{ a.s..}$$

By Assumption 16, we can replace $\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}a^Tg(X_i)\right)$ by $\sqrt{n}a^TS$, and finally obtain

$$\lim_{n\to\infty}\sup_{t\in\mathbb{R}}\left|P\left(\sqrt{n\left(a^T\tilde{\Sigma}a\right)^{-1}}\left(\int a^Tg(x)f(x\mid\theta)\,\mathrm{d}x-a^TS\right)\leq t\mid X_1,\ldots,X_n\right)-\Phi(t)\right|=0,\text{ a.s..}$$

The remainder of the proof uses the same argument used in the proof of Theorem 3. □

Reference

Bibliography

Adimari, G., 1995. Empirical likelihood confidence intervals for the difference between means. STATISTICA-BOLOGNA- 55, 87–94.

Adimari, G., 1997. Empirical likelihood type confidence intervals under random censorship. Annals of the Institute of Statistical Mathematics 49 (3), 447–466.

Aeschbacher, S., Beaumont, M. A., Futschik, A., 2012. A novel approach for choosing summary statistics in approximate bayesian computation. Genetics 192 (3), 1027–1047.

Aragon, Y., 1997. A gauss implementation of multivariate sliced inverse regression. Computational Statistics 12 (3), 355–372.

Baggerly, K. A., 1998. Empirical likelihood as a goodness-of-fit measure. Biometrika 85 (3), 535–547.

Bandyopadhyay, S., Lahiri, S. N., Nordman, D. J., et al., 2015. A frequency domain empirical likelihood method for irregularly spaced spatial data. The Annals of Statistics 43 (2), 519–545.

Baragatti, M., Grimaud, A., Pommeret, D., 2013. Likelihood-free parallel tempering. Statistics and Computing 23 (4), 535–549.

Barthelmé, S., Chopin, N., 2014. Expectation propagation for likelihood-free inference. Journal of the American Statistical Association 109 (505), 315–333.

Bartolucci, F., 2007. A penalized version of the empirical likelihood ratio for the population mean. Statistics & probability letters 77 (1), 104–110.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., Robert, C. P., 2009. Adaptive approximate bayesian computation. Biometrika, asp052.

Beaumont, M. A., Zhang, W., Balding, D. J., 2002. Approximate bayesian computation in population genetics. Genetics 162 (4), 2025–2035.

Biao Zhang, B. Z., 1998. A note on kernel density estimation with auxiliary information. Communications in Statistics-Theory and Methods 27 (1), 1–11.

Biau, G., Cérou, F., Guyader, A., 2012. New insights into approximate bayesian computation. arXiv preprint arXiv:1207.6461.

Blum, M. G., François, O., 2010. Non-linear regression models for approximate bayesian computation. Statistics and Computing 20 (1), 63–73.

Blum, M. G., Nunes, M. A., Prangle, D., Sisson, S. A., et al., 2013. A comparative review of dimension reduction methods in approximate bayesian computation. Statistical Science 28 (2), 189–208.

Bura, E., Cook, R. D., 2001. Extending sliced inverse regression: The weighted chi-squared test. Journal of the American Statistical Association 96 (455), 996–1003.

Chang, I. H., Mukerjee, R., 2008. Bayesian and frequentist confidence intervals arising from empirical-type likelihoods. Biometrika 95 (1), 139–147.

Chang, J., Chen, S. X., Chen, X., 2015. High dimensional generalized empirical likelihood for moment restrictions with dependent data. Journal of Econometrics

185 (1), 283–304.
Chen, J., Qin, J., 1993. Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. Biometrika 80 (1), 107–116.
Chen, J., Sitter, R., 1999. A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. Statistica Sinica 9 (2), 385–406.
Chen, J., Variyath, A. M., Abraham, B., 2008. Adjusted empirical likelihood and its properties. Journal of Computational and Graphical Statistics 17 (2), 426–443.
Chen, S. X., 1993. On the accuracy of empirical likelihood confidence regions for linear regression model. Annals of the Institute of Statistical Mathematics 45 (4), 621–637.
Chen, S. X., 1994. Empirical likelihood confidence intervals for linear regression coefficients. Journal of Multivariate Analysis 49 (1), 24–40.
Chen, S. X., 1997. Empirical likelihood-based kernel density estimation. Australian Journal of Statistics 39 (1), 47–56.
Chen, S. X., Cui, H., 2006. On bartlett correction of empirical likelihood in the presence of nuisance parameters. Biometrika 93 (1), 215–220.
Chen, S. X., Hall, P., 1993. Smoothed empirical likelihood confidence intervals for quantiles. The Annals of Statistics, 1166–1181.
Chen, S. X., Peng, L., Qin, Y.-L., 2009. Effects of data dimension on empirical likelihood. Biometrika 96 (3), 711–722.
Chen, S. X., Qin, Y. S., 2000. Empirical likelihood confidence intervals for local linear smoothers. Biometrika 87 (4), 946–953.
Chen, S. X., Wong, C. M., 2009. Smoothed block empirical likelihood for quantiles of weakly dependent processes. Statistica Sinica 19 (1), 71.
Chuang, C.-S., Chan, N. H., 2002. Empirical likelihood for autoregressive models, with applications to unstable time series. Statistica Sinica 12 (2), 387–408.
Cook, R. D., 1994. On the interpretation of regression plots. Journal of the American Statistical Association 89 (425), 177–189.
Cook, R. D., 1998. Principal hessian directions revisited. Journal of the American Statistical Association 93 (441), 84–94.
Cook, R. D., Forzani, L., 2009. Likelihood-based sufficient dimension reduction. Journal of the American Statistical Association 104 (485), 197–208.
Cook, R. D., Li, B., 2002. Dimension reduction for conditional mean in regression. Annals of Statistics, 455–474.
Cook, R. D., Li, B., Chiaromonte, F., 2007. Dimension reduction in regression without matrix inversion. Biometrika 94 (3), 569–584.
Cook, R. D., Li, B., et al., 2004. Determining the dimension of iterative hessian transformation. The Annals of Statistics 32 (6), 2501–2531.
Cook, R. D., Ni, L., 2005. Sufficient dimension reduction via inverse regression. Journal of the American Statistical Association 100 (470).
Cook, R. D., Setodji, C. M., 2003. A model-free test for reduced rank in multivariate regression. Journal of the American Statistical Association 98 (462), 340–351.
Cook, R. D., Weisberg, S., 1991. Sliced inverse regression for dimension reduction: Comment. Journal of the American Statistical Association 86 (414), pp. 328–332. URL http://www.jstor.org/stable/2290564
Cook, R. D., Yin, X., 2001. Special invited paper: Dimension reduction and visualization in discriminant analysis. Australian and New Zealand Journal of Statistics 43 (2), 147–199.

Davidian, M., Carroll, R. J., 1987. Variance function estimation. Journal of the American Statistical Association 82 (400), 1079–1091.

Dean, T. A., Singh, S. S., Jasra, A., Peters, G. W., 2014. Parameter estimation for hidden markov models with intractable likelihoods. Scandinavian Journal of Statistics 41 (4), 970–987.

DiCiccio, T., Hall, P., Romano, J., 1991. Empirical likelihood is bartlett-correctable. The Annals of Statistics, 1053–1061.

Dong, Y., Li, B., 2010. Dimension reduction for non-elliptically distributed predictors: second-order methods. Biometrika 97 (2), 279–294.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al., 2004. Least angle regression. The Annals of statistics 32 (2), 407–499.

Emerson, S. C., Owen, A. B., et al., 2009. Calibration of the empirical likelihood method for a vector mean. Electronic Journal of Statistics 3, 1161–1192.

Fang, K.-t., Mukerjee, R., 2005. Expected lengths of confidence intervals based on empirical discrepancy statistics. Biometrika 92 (2), 499–503.

Fang, K.-T., Mukerjee, R., 2006. Empirical-type likelihoods allowing posterior credible sets with frequentist validity: Higher-order asymptotics. Biometrika 93 (3), 723–733.

Fearnhead, P., Prangle, D., 2012. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74 (3), 419–474.

Ghosh, J., Sinha, B., Joshi, S., 1982. Expansions for posterior probability and integrated bayes risk. Statistical Decision Theory and Related Topics III 1, 403–456.

Ghosh, M., Liu, R., 2011. Moment matching priors. Sankhya A 73 (2), 185–201.

Grelaud, A., Robert, C. P., Marin, J.-M., Rodolphe, F., Taly, J.-F., et al., 2009. Abc likelihood-free methods for model choice in gibbs random fields. Bayesian Analysis 4 (2), 317–335.

Grendár, M., Judge, G., 2009. Asymptotic equivalence of empirical likelihood and bayesian map. The Annals of Statistics, 2445–2457.

Hall, P., Owen, A. B., 1993. Empirical likelihood confidence bands in density estimation. Journal of Computational and Graphical Statistics 2 (3), 273–289.

Hall, P., Presnell, B., 1999. Biased bootstrap methods for reducing the effects of contamination. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 661–680.

Hartley, H. O., Rao, J., 1968. A new estimation theory for sample surveys. Biometrika 55 (3), 547–557.

He, X., Wang, G., 1995. Law of the iterated logarithm and invariance principle for m-estimators. Proceedings of the American Mathematical Society 123 (2), 563–573.

Hernández, A., Velilla, S., 2005. Dimension reduction in nonparametric kernel discriminant analysis. Journal of Computational and Graphical Statistics 14 (4), 847–866.

Hjort, N. L., McKeague, I. W., Van Keilegom, I., 2009. Extending the scope of empirical likelihood. The Annals of Statistics, 1079–1111.

Hollander, M., McKeague, I. W., McKeague, I. W., 1997. Likelihood ratio-based confidence bands for survival functions. Journal of the American Statistical Association 92 (437), 215–226.

Hsing, T., 1999. Nearest neighbor inverse regression. Annals of statistics, 697–731.

Jaakkola, T. S., Jordan, M. I., 2000. Bayesian parameter estimation via variational methods. Statistics and Computing 10 (1), 25–37.

Jasra, A., Singh, S. S., Martin, J. S., McCoy, E., 2012. Filtering via approximate bayesian computation. Statistics and Computing 22 (6), 1223–1237.

Jing, B.-Y., 1995. Two-sample empirical likelihood method. Statistics & probability letters 24 (4), 315–319.

Johnson, R. A., 1970. Asymptotic expansions associated with posterior distributions. The Annals of Mathematical Statistics 41 (3), 851–864.

Joyce, P., Marjoram, P., 2008. Approximately sufficient statistics and bayesian computation. Statistical applications in genetics and molecular biology 7 (1).

Kitamura, Y., 2001. Asymptotic optimality of empirical likelihood for testing moment restrictions. Econometrica 69 (6), 1661–1672.

Kitamura, Y., et al., 1997. Empirical likelihood methods with weakly dependent processes. The Annals of Statistics 25 (5), 2084–2102.

Kolaczyk, E. D., 1994. Empirical likelihood for generalized linear models. Statistica Sinica 4 (1), 199–218.

Kolaczyk, E. D., 1995. An information criterion for empirical likelihood with general estimating equations. Unpublished manuscript, Department of Statistics, University of Chicago.

Lahiri, S. N., Mukhopadhyay, S., et al., 2012. A penalized empirical likelihood method in high dimensions. The Annals of Statistics 40 (5), 2511–2540.

Lancaster, T., Jae Jun, S., 2010. Bayesian quantile regression methods. Journal of Applied Econometrics 25 (2), 287–307.

Lazar, N. A., 2003. Bayesian empirical likelihood. Biometrika 90 (2), 319–326.

Lazar, N. A., Mykland, P. A., 1999. Empirical likelihood in the presence of nuisance parameters. Biometrika 86 (1), 203–211.

Lee, A., Łatuszyński, K., 2014. Variance bounding and geometric ergodicity of markov chain monte carlo kernels for approximate bayesian computation. Biometrika 101 (3), 655–671.

Lee, K.-Y., Li, B., Chiaromonte, F., et al., 2013. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. The Annals of Statistics 41 (1), 221–249.

Leng, C., Tang, C. Y., 2012. Penalized empirical likelihood and growing dimensional general estimating equations. Biometrika 99 (3), 703–716.

Leuenberger, C., Wegmann, D., 2010. Bayesian computation and model selection without likelihoods. Genetics 184 (1), 243–252.

Li, B., Artemiou, A., Li, L., 2011. Principal support vector machines for linear and nonlinear sufficient dimension reduction. The Annals of Statistics, 3182–3210.

Li, B., Dong, Y., 2009. Dimension reduction for nonelliptically distributed predictors. The Annals of Statistics, 1272–1298.

Li, B., Wang, S., 2007. On directional regression for dimension reduction. Journal of the American Statistical Association 102 (479), 997–1008.

Li, B., Wen, S., Zhu, L., 2008. On a projective resampling method for dimension reduction with multivariate responses. Journal of the American Statistical Association 103 (483), 1177–1186.

Li, B., Zha, H., Chiaromonte, F., 2005. Contour regression: a general approach to dimension reduction. Annals of statistics, 1580–1616.

Li, K.-C., 1991. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association 86 (414), 316–327.

Li, K.-C., 1992. On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma. Journal of the American Statistical Association 87 (420), 1025–1039.

Li, K.-C., Duan, N., 1989. Regression analysis under link violation. The Annals of Statistics, 1009–1052.

Li, L., Cook, R. D., Tsai, C.-L., 2007. Partial inverse regression. Biometrika.

Loh, W.-L., et al., 1996. On latin hypercube sampling. The annals of statistics 24 (5), 2058–2080.

Luo, R., Wang, H., Tsai, C.-L., et al., 2009. Contour projected dimension reduction. The Annals of Statistics 37 (6B), 3743–3778.

Ma, Y., Zhu, L., 2012. A semiparametric approach to dimension reduction. Journal of the American Statistical Association 107 (497), 168–179.

Ma, Y., Zhu, L., 2013. A review on dimension reduction. International Statistical Review 81 (1), 134–150.

Marin, J.-M., Pillai, N. S., Robert, C. P., Rousseau, J., 2014. Relevant statistics for bayesian model choice. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76 (5), 833–859.

Marin, J.-M., Pudlo, P., Robert, C. P., Ryder, R. J., 2012. Approximate bayesian computational methods. Statistics and Computing 22 (6), 1167–1180.

Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S., 2003. Markov chain monte carlo without likelihoods. Proceedings of the National Academy of Sciences 100 (26), 15324–15328.

Martin, R., Liu, C., 2013. Inferential models: A framework for prior-free posterior probabilistic inference. Journal of the American Statistical Association 108 (501), 301–313.

Martin, R., Liu, C., 2015. Conditional inferential models: combining information for prior-free probabilistic inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77 (1), 195–217.

McKinley, T., Cook, A. R., Deardon, R., 2009. Inference in epidemic models without likelihoods. The International Journal of Biostatistics 5 (1).

Mittelhammer, R. C., Judge, G. G., Miller, D. J., 2000. Econometric Foundations Pack with CD-ROM. Vol. 1. Cambridge University Press.

Molanes Lopez, E. M., KEILEGOM, I. V., Veraverbeke, N., 2009. Empirical likelihood for non-smooth criterion functions. Scandinavian Journal of Statistics 36 (3), 413–432.

Monti, A. C., 1997. Empirical likelihood confidence regions in time series models. Biometrika 84 (2), 395–405.

Murphy, S., 1995. Likelihood ratio-based confidence intervals in survival analysis. Journal of the American Statistical Association 90 (432), 1399–1405.

Murphy, S. A., van der Vaart, A. W., 1997. Semiparametric likelihood ratio inference. The Annals of Statistics, 1471–1509.

Mykland, P. A., 1995. Dual likelihood. The Annals of Statistics, 396–421.

Newey, W. K., Smith, R. J., 2004. Higher order properties of gmm and generalized empirical likelihood estimators. Econometrica 72 (1), 219–255.

Nordman, D. J., Sibbertsen, P., Lahiri, S. N., 2007. Empirical likelihood confidence intervals for the mean of a long-range dependent process. Journal of Time Series Analysis 28 (4), 576–599.

Owen, A., 1991. Empirical likelihood for linear models. The Annals of Statistics, 1725–1747.

Owen, A. B., 1988. Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75 (2), 237–249.

Owen, A. B., 1995. Nonparametric likelihood confidence bands for a distribution function. Journal of the American Statistical Association 90 (430), 516–521.

Owen, A. B., 2010. Empirical likelihood. CRC press.

Pan, X.-R., Zhou, M., 2002. Empirical likelihood ratio in terms of cumulative hazard function for censored data. Journal of Multivariate Analysis 80 (1), 166–188.

Peng, L., 2004. Empirical-likelihood-based confidence interval for the mean with a heavy-tailed distribution. Annals of Statistics, 1192–1214.

Picchini, U., 2014. Inference for sde models via approximate bayesian computation. Journal of Computational and Graphical Statistics 23 (4), 1080–1100.

Prangle, D., Fearnhead, P., Cox, M. P., Biggs, P. J., French, N. P., 2014. Semi-automatic selection of summary statistics for abc model choice. Statistical applications in genetics and molecular biology 13 (1), 67–82.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., Feldman, M. W., 1999. Population growth of human y chromosomes: a study of y chromosome microsatellites. Molecular biology and evolution 16 (12), 1791–1798.

Qin, J., 1993. Empirical likelihood in biased sample problems. The Annals of Statistics, 1182–1196.

Qin, J., 1998. Semiparametric likelihood based method for goodness of fit tests and estimation in upgraded mixture models. Scandinavian journal of statistics 25 (4), 681–691.

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. The Annals of Statistics, 300–325.

Qin, J., Zhang, B., 1997. A goodness-of-fit test for logistic regression models based on case-control data. Biometrika 84 (3), 609–618.

Qin, J., et al., 1999. Empirical likelihood ratio based confidence intervals for mixture proportions. The Annals of Statistics 27 (4), 1368–1384.

Rao, J., Wu, C., 2010. Bayesian pseudo-empirical-likelihood intervals for complex surveys. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (4), 533–544.

Ratmann, O., Andrieu, C., Wiuf, C., Richardson, S., 2009. Model criticism based on likelihood-free inference, with an application to protein network evolution. Proceedings of the National Academy of Sciences 106 (26), 10576–10581.

Ratmann, O., Camacho, A., Meijer, A., Donker, G., 2013. Statistical modelling of summary values leads to accurate approximate bayesian computations. arXiv preprint arXiv:1305.4283.

Rivoirard, V., Rousseau, J., et al., 2012. Bernstein–von mises theorem for linear functionals of the density. The Annals of Statistics 40 (3), 1489–1523.

Robert, C., Casella, G., 2013. Monte Carlo statistical methods. Springer Science & Business Media.

Rubin, D. B., 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics 12 (4), 1151–1172.

Rubin, D. B., et al., 1981. The bayesian bootstrap. The annals of statistics 9 (1), 130–134.

Ruli, E., Sartori, N., Ventura, L., 2013. Approximate bayesian computation with composite score functions. Statistics and Computing, 1–14.

Saracco, J., 2005. Asymptotics for pooled marginal slicing estimator based on sirα approach. Journal of multivariate Analysis 96 (1), 117–135.

Schennach, S. M., 2005. Bayesian exponentially tilted empirical likelihood. Biometrika 92 (1), 31–46.

Schennach, S. M., et al., 2007. Point estimation with exponentially tilted empirical likelihood. The Annals of Statistics 35 (2), 634–672.

Schott, J. R., 1994. Determining the dimensionality in sliced inverse regression. Journal of the American Statistical Association 89 (425), 141–148.

Setodji, C. M., Cook, R. D., 2004. K-means inverse regression. Technometrics 46 (4), 421–429.

Sisson, S. A., Fan, Y., Tanaka, M. M., 2007. Sequential monte carlo without likelihoods. Proceedings of the National Academy of Sciences 104 (6), 1760–1765.

Su, Z., Cook, R. D., 2011. Partial envelopes for efficient estimation in multivariate linear regression. Biometrika, asq063.

Su, Z., Cook, R. D., 2012. Inner envelopes: efficient estimation in multivariate linear regression. Biometrika 99 (3), 687–702.

Tang, C. Y., Leng, C., 2010. Penalized high-dimensional empirical likelihood. Biometrika 97 (4), 905–920.

Tavaré, S., Balding, D. J., Griffiths, R. C., Donnelly, P., 1997. Inferring coalescence times from dna sequence data. Genetics 145 (2), 505–518.

Thomas, D. R., Grunkemeier, G. L., 1975. Confidence interval estimation of survival probabilities for censored data. Journal of the American Statistical Association 70 (352), 865–871.

Tierney, L., Kadane, J. B., 1986. Accurate approximations for posterior moments and marginal densities. Journal of the american statistical association 81 (393), 82–86.

Tsao, M., Wu, F., 2014. Extended empirical likelihood for estimating equations. Biometrika, asu014.

Tsao, M., et al., 2004. Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. The Annals of Statistics 32 (3), 1215–1221.

Velilla, S., 1998. Assessing the number of linear components in a general regression problem. Journal of the American Statistical Association 93 (443), 1088–1098.

Vexler, A., Deng, W., Wilding, G. E., 2013. Nonparametric bayes factors based on empirical likelihood ratios. Journal of statistical planning and inference 143 (3), 611–620.

Vexler, A., Tao, G., Hutson, A., 2014. Posterior expectation based on empirical likelihoods. Biometrika 101 (3), 711–718.

Wang, H., Xia, Y., 2008. Sliced regression for dimension reduction. Journal of the American Statistical Association 103 (482), 811–821.

Wang, H. J., Zhu, Z., 2011. Empirical likelihood for quantile regression models with longitudinal data. Journal of Statistical Planning and Inference 141 (4), 1603–1615.

Wang, T., Guo, X., Zhu, L., Xu, P., 2014. Transformed sufficient dimension reduction. Biometrika, asu037.

Whang, Y.-J., 2006. Smoothed empirical likelihood methods for quantile regression models. Econometric Theory 22 (02), 173–205.

Wilkinson, D. J., 2011. Stochastic modelling for systems biology. CRC press.

Wilkinson, R. D., 2013. Approximate bayesian computation (abc) gives exact results under the assumption of model error. Statistical applications in genetics and molecular biology 12 (2), 129–141.

Wu, C., Sitter, R. R., 2001. A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association 96 (453), 185–193.

Wu, H.-M., 2008. Kernel sliced inverse regression with applications to classification. Journal of Computational and Graphical Statistics 17 (3).

Wu, Q., Liang, F., Mukherjee, S., 2008. Consistency of regularized sliced inverse regression for kernel models. Tech. rep., Technical report. Duke University and University of Illinois Urbana-Champaign.

Xia, Y., 2007. A constructive approach to the estimation of dimension reduction directions. The Annals of Statistics, 2654–2690.

Xia, Y., Tong, H., Li, W., Zhu, L.-X., 2002. An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (3), 363–410.

Ye, Z., Weiss, R. E., 2003. Using the bootstrap to select one of a new class of dimension reduction methods. Journal of the American Statistical Association 98 (464).

Yeh, Y.-R., Huang, S.-Y., Lee, Y.-J., 2009. Nonlinear dimension reduction with kernel sliced inverse regression. Knowledge and Data Engineering, IEEE Transactions on 21 (11), 1590–1603.

Yin, X., Bura, E., 2006. Moment-based dimension reduction for multivariate response regression. Journal of Statistical Planning and Inference 136 (10), 3675–3688.

Yin, X., Cook, R. D., 2002. Dimension reduction for the conditional kth moment in regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (2), 159–175.

Yin, X., Cook, R. D., 2005. Direction estimation in single-index regressions. Biometrika 92 (2), 371–384.

Yin, X., Li, B., Cook, R. D., 2008. Successive direction extraction for estimating the central subspace in a multiple-index regression. Journal of Multivariate Analysis 99 (8), 1733–1757.

Zhang, B., 1996a. Confidence intervals for a distribution function in the presence of auxiliary information. Computational statistics & data analysis 21 (3), 327–342.

Zhang, B., 1996b. On the accuracy of empirical likelihood confidence intervals for m-functionals. Journal of Nonparametric Statistics 6 (4), 311–321.

Zhang, B., 1999. Bootstrapping with auxiliary information. Canadian Journal of Statistics 27 (2), 237–249.

Zhu, L., Miao, B., Peng, H., 2006. On sliced inverse regression with high-dimensional covariates. Journal of the American Statistical Association 101 (474).

Zhu, L., Wang, T., Zhu, L., Ferré, L., 2010a. Sufficient dimension reduction through discretization-expectation estimation. Biometrika 97 (2), 295–304.

Zhu, L.-P., Li, L., Li, R., Zhu, L.-X., 2011. Model-free feature screening for ultrahigh-dimensional data. Journal of the American Statistical Association 106 (496).

Zhu, L.-P., Yu, Z., Zhu, L.-X., 2010b. A sparse eigen-decomposition estimation in semiparametric regression. Computational Statistics & Data Analysis 54 (4), 976–986.

Zhu, L.-P., Zhu, L.-X., 2007. On kernel method for sliced average variance estimation. Journal of Multivariate Analysis 98 (5), 970–991.

Zhu, L.-P., Zhu, L.-X., 2009a. Dimension reduction for conditional variance in regressions. Statistica Sinica 19 (2), 869.

Zhu, L.-P., Zhu, L.-X., 2009b. On distribution-weighted partial least squares with diverging number of highly correlated predictors. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (2), 525–548.

Zhu, L.-P., Zhu, L.-X., Feng, Z.-H., 2010c. Dimension reduction in regressions through cumulative slicing estimation. Journal of the American Statistical Association 105 (492), 1455–1466.

Zhu, L.-X., Fang, K.-T., et al., 1996. Asymptotics for kernel estimate of sliced inverse regression. The Annals of Statistics 24 (3), 1053–1068.

Zhu, L.-X., Ohtaki, M., Li, Y., 2007. On hybrid methods of inverse regression-based algorithms. Computational statistics & data analysis 51 (5), 2621–2635.

Zhu, Y., Zeng, P., 2006. Fourier methods for estimating the central subspace and the central mean subspace in regression. Journal of the American Statistical Association 101 (476), 1638–1651.

Ait-Sahalia, Y. (1996). Nonparametric Pricing of Interest Rate Derivative Securities. *Econometrica*, 64(3), 527-560.

Ait-Sahalia, Y., & Lo, A. W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *The Journal of Finance*, 53(2), 499-547.

Bakshi, G., Cao, C., & Chen, Z. (1997). Empirical performance of alternative option pricing models. *The Journal of Finance*, 52(5), 2003-2049.

Bakshi, G., & Madan, D. (2000). Spanning and derivative-security valuation. *Journal of Financial Economics*, 55(2), 205-238.

Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *Review of Financial Studies*, 9(1), 69-107.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, 637-654.

Carr, P., & Madan, D. (1999). Option valuation using the fast Fourier transform. *Journal of Computational Finance*, 2(4), 61-73.

Carr, P., & Madan, D. (2009). Saddlepoint methods for option pricing. *Journal of Computational Finance*, 13(1), 49.

Chan, N. H., & Ling, S. (2006). Empirical likelihood for GARCH models. *Econometric Theory*, 22(03), 403-428.

Chuang, C. S., & Chan, N. H. (2002). Empirical likelihood for autoregressive models, with applications to unstable time series. *Statistica Sinica*, 12(2), 387-408.

Diciccio, T. J., & Romano, J. P. (1989). On adjustments based on the signed root of the empirical likelihood ratio statistic. *Biometrika*, 76(3), 447-456.

Duan, J. C. (1995). The GARCH option pricing model. *Mathematical Finance*, 5(1), 13-32.

Duffie, D., Pan, J., & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68(6), 1343-1376.

Hall, P., & La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review/Revue Internationale de Statistique*, 109-127.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6(2), 327-343.

Heston, S. L., & Nandi, S. (2000). A closed-form GARCH option valuation model. *Review of Financial Studies*, 13(3), 585-625.

Huang, C., & Litzenberger, R. H. (1988). Foundations for financial economics. *Princeton University*.

Hull, J., & White, A. (1987). The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, 42(2), 281-300.

Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *The Journal of Finance*, 49(3), 851-889.

Kou, S. G. (2002). A jump-diffusion model for option pricing. *Management science*, 48(8), 1086-1101.

Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics*, 25(5), 2084-2102.

Melino, A., & Turnbull, S. M. (1990). Pricing foreign currency options with stochastic volatility. *Journal of Econometrics*, 45(1), 239-265.

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8(4), 323-361.

Mykland, P. A. (1995). Dual likelihood. *The Annals of Statistics*, 396-421.

Nordman, D. J., Sibbertsen, P., & Lahiri, S. N. (2007). Empirical likelihood confidence intervals for the mean of a long-range dependent process. *Journal of Time Series Analysis*, 28(4), 576-599.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2), 237-249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 90-120.

Owen, A. B. (2001). Empirical likelihood. *CRC press*.

Peng, H. (2015) On a class of maximum empirical likelihood estimators. Manuscript

Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 300-325.

Rubinstein, M. (1985). Nonparametric tests of alternative option pricing models using all reported trades and quotes on the 30 most active CBOE option classes from August 23, 1976 through August 31, 1978. *The Journal of Finance*, 40(2), 455-480.

Scott, L. O. (1987). Option pricing when the variance changes randomly: Theory, estimation, and an application. *Journal of Financial and Quantitative Analysis*, 22(04), 419-438.

Stutzer, M. (1996). A simple nonparametric approach to derivative security valuation. *The Journal of Finance*, 51(5), 1633-1652.

Wiggins, J. B. (1987). Option values under stochastic volatility: Theory and empirical estimates. *Journal of Financial Economics*, 19(2), 351-372.

Yau, C. Y. (2012). Empirical likelihood in long-memory time series models. *Journal of Time Series Analysis*, 33(2), 269-275.