Robust Estimation and Inference in Categorical Data*

Max Welz
welz@ese.eur.nl

Econometric Institute
Erasmus School of Economics

February 29, 2024
Job Market Paper for 2024-25 (first version)

Abstract

In empirical science, many variables of interest are categorical. Like any model, models for categorical responses can be misspecified, leading to possibly large biases in estimation. One particularly troublesome source of misspecification is inattentive responding in questionnaires, which is well-known to jeopardize the validity of structural equation models (SEMs) and other survey-based analyses. I propose a general estimator that is designed to be robust to misspecification of models for categorical responses. Unlike hitherto approaches, the estimator makes no assumption whatsoever on the degree, magnitude, or type of misspecification. The proposed estimator generalizes maximum likelihood estimation, is strongly consistent, asymptotically Gaussian, has the same time complexity as maximum likelihood, and can be applied to any model for categorical responses. In addition, I develop a novel test that tests whether a given response can be fitted well by the assumed model, which allows one to trace back possible sources of misspecification. I verify the attractive theoretical properties of the proposed methodology in Monte Carlo experiments, and demonstrate its practical usefulness in an empirical application on a SEM of personality traits, where I find compelling evidence for the presence of inattentive responding whose adverse effects the proposed estimator can withstand, unlike maximum likelihood.

*I thank Alberto Abadie, Andreas Alfons, Isaiah Andrews, Marco Avella Medina, Patrick Groenen, Nick Koning, Patrick Mair, Anna Mikusheva, Whitney Newey, Alberto Quaini, Elvezio Ronchetti, Stefanie Stantcheva, and Ming Yuan, as well as the participants of the MIT Econometrics lunch seminar and the Econometric Institute internal seminar for helpful comments and suggestions. Most of this work was carried out when I was visiting the economics department at MIT; I thank them for their hospitality. This work was supported by a grant from the Dutch Research Council (NWO), research program Vidi (Project No. VI.Vidi.195.141).

KEYWORDS: Model misspecification, robust estimation, categorical variables, inattentive responding
JEL CODES: C13, C18, C35, C83, D91.

1 Introduction

In the social, economic, psychological, and behavioral sciences, many quantities of interest are measured by means of categorical variables, often through surveys or questionnaires. After collecting such data, researchers typically model the quantities of interest by employing models specifically designed for the categorical measurements thereof, for example structural equation models for latent character traits or item response models for item difficulty. However, such models can be misspecified, and it is well-known that model misspecification leads to inconstant estimation of model parameters, which may ultimately result in incorrect research findings. In survey-collected data, a particularly relevant source of model misspecification is inattentive or careless responding by survey participants (e.g. Stantcheva, 2022; Meade & Craig, 2012; Huang et al., 2015b). While there are many estimation methods that are designed to be robust to model misspecification (see Maronna et al., 2018, for a recent overview), these methods almost exclusively designed for continuous data, and may fail to be effective or even computable when the data are categorical.

Motivated by the empirical relevance of this gap in the literature due to the increasing popularity of surveys in economics (Stantcheva, 2022) and their long-standing importance in behavioral research (e.g. Rust et al., 2020), I propose a novel estimation framework for models of categorical data that is designed to be robust to misspecification of that model. The estimator is very general in the sense that it can be applied to nearly any model of categorical data, and, crucially, it makes *no assumption whatsoever* on the type, magnitude, or location of potential misspecification. Hence, the proposed estimator is robust to an unlimited and unspecified variety of possible sources of model misspecification. I show that the estimator generalizes commonly employed maximum likelihood estimation and possesses attractive theoretical and computational properties. For instance, the estimator is strongly consistent and asymptotically normally distributed, while it is of the same time complexity as maximum likelihood, meaning that it comes at no additional computational cost. Furthermore, I develop a novel statistical test that test if a given categorical data point can be fitted well by the presumed model. The test rejects this null hypothesis if the data point in question cannot be fitted sufficiently well, and thereby helps pinpoint potential sources of model misspecification.

As such, the proposed methodology is particularly attractive to fit models for survey responses such as structural equation models or item response models because it is likely that a certain proportion of survey participants responds inattentively or carelessly (Ward & Meade, 2023). The estimator can not only withstand the harmful effects of inattentive respondents, but also identify them through the diagnostic test.

I verify the attractive theoretical and robustness properties of the proposed estimator by means of extensive simulation studies and demonstrate its practical usefulness in an empirical application on a structural equation model on a measurement of the Big-5 personality traits (Goldberg, 1992). I find compelling evidence for the presence of inattentive respondents that, if unaccounted for, have a sizable effect on parameter estimates. For instance, the correlation coefficient between two mutually contradic·tory items in a *neuroticism* scale is estimated as -0.62 by hitherto estimation methods, whereas my estimator yields a substantially stronger correlational estimate of -0.93, which is, unlike the former estimate, in line with literature on this scale. Likewise, a structural equation model fitted by my robust estimator yields stronger and internally more consistent factor loadings that can explain substantially more variation than a fit based on commonly employed methods. In addition, I demonstrate how the estimator can be used to robustly estimate scale reliability coefficients like Cronbach's α. As such, the entire process of building, testing, and verifying scales can be robustified against inattentive responding without making any assumptions on the nature of inattentive responding, a departure from previous literature.

With its focus on models for questionnaire responses, this paper ties into a growing literature concerned with the validity of research findings when employed models are misspecified due to responses that do not follow the assumed model and subsequently cannot be fitted well by that model. An ensuing poor model fit may lead a researcher to doubt or even reject the theoretical model (Lai & Green, 2016). Poor model fit can occur for two reasons. First, the theory behind the model may simply be wrong, in case of which the model is also wrong and the theory is correctly rejected by the data. Second, theory and model may in fact be correct, yet a misfit occurs due to the presence of a limited number of observations from a different population, such as inattentive respondents or heterogeneous subgroups, which may lead one to incorrectly reject a theory (Arias et al., 2020). Indeed, already a small proportion of inattentive responses can substantially deteriorate model fit (Arias et al., 2020; Huang et al., 2015a; Woods, 2006), and ultimately lead a researcher to reject a correct hypothesis or sustain an incorrect hypothesis (Arias et al., 2020; Huang et al., 2015b; Maniaci & Rogge, 2014; McGrath et al., 2010; Woods, 2006; Schmitt & Stults, 1985). Inattentive responding itself is widely prevalent (Ward & Meade, 2023; Bowling et al., 2016; Meade & Craig, 2012), with most estimates on its prevalence ranging from 10–15% of study participants (Curran, 2016; Huang et al., 2015b, 2012; Meade & Craig, 2012), while already a prevalence 5–10% can jeopardize the validity of research findings (Arias et al., 2020; Credé, 2010; Woods, 2006). In fact, Ward & Meade (2023) conjecture that inattentive responding is likely present in all survey data. Due to the damaging effects of inattentive responses, a large number of methods for their detection has emerged, ranging from consistency indicators such as psychometric antonyms/synonyms (Meade & Craig, 2012) over response times (e.g. Bowling et al., 2023) to model-based techniques, such as person-fit statistics (e.g. Drasgow et al., 1985), structural equation models (e.g. Kim et al., 2018), mixture models (e.g. Arias et al., 2020), or attention check items (e.g., Section 3 in Stantcheva, 2022). More recently, machine learning techniques have been proposed (Welz & Alfons, 2023; Schroeders et al., 2022). I refer to Alfons & Welz (2024) for a recent review of inattention in survey data. I deviate from hitherto methods

by not making any assumptions on the nature of potential misspecification and deriving statistical guarantees on the performance of my approach.

In econometrics, this paper ties into a recent stream of methodological literature on misspecification-robust estimation (e.g. Andrews et al., 2017; Armstrong & Kolesár, 2021; Bugni & Ura, 2019; Bonhomme & Weidner, 2022; Chernozhukov et al., 2018, 2022; Kitamura et al., 2013), and develops a categorical analogue to robust M-estimation (e.g. Van de Geer, 2000), which is primarily intended for models of continuous data.

This paper is organizes as follows. Section 2 describes the modeling framework and Section 3 the proposed methodology, while Sections 4 and 5 contain Monte Carlo experiments and an empirical application, respectively. Section 6 concludes.

I employ the following notation throughout. A vector $\boldsymbol{a} \in \mathbb{R}^d$ is understood as a $(d \times 1)$ column vector, and $\boldsymbol{a}^\top$ denotes its $(1 \times d)$ transpose. Further, denote by $\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}^\top \boldsymbol{a}}$ its Euclidean norm, and, for a square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$, denote by $|||\boldsymbol{A}||| = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})}$ its operator norm, where $\lambda_{\max}(\boldsymbol{A})$ is the largest eigenvalue of a matrix $\boldsymbol{A}$. Assuming that $\boldsymbol{A}$ is symmetric positive-definite, let $\boldsymbol{A}^{1/2}$ denote a lower triangular $(d \times d)$ matrix that satisfies the Cholesky factorization $\boldsymbol{A} = \boldsymbol{A}^{1/2} \left(\boldsymbol{A}^{1/2}\right)^\top$. Furthermore, let Φ_d denote the distribution function of the d-variate standard normal distribution, and let $\mathbb{P}$ and $\mathbb{E}$ be a generic probability measure and a generic expectation operator, respectively. Throughout this paper, I make the implicit assumption that all stochastic objects are well-defined on an appropriately chosen measure space. Furthermore, let the symbols "$\xrightarrow{\text{d}}$" and "$\xrightarrow{\text{a.s.}}$" denote convergence in distribution and almost sure convergence, respectively.

2 Setup and modeling framework

This section introduces the modeling framework that will be employed throughout the paper. The considered framework is very general and many well-known models emerge as special cases.

2.1 Notation and examples

Suppose one observes a k-dimensional categorical random variable $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_k)^\top$ that takes values in a finite sample space $\boldsymbol{\mathcal{Z}} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m\}$ with m possible vector-valued outcomes of dimension k. Hence, realizations of this random variable can be represented in a k-way contingency table that cross-tabulates empirical frequencies for each outcome. The individual categories in $\boldsymbol{\mathcal{Z}}$ can either be ordered (so that $\boldsymbol{Z}$ is ordinal/quantitative, like educational attainment) or unordered (so that $\boldsymbol{Z}$ is nominal/qualitative, like marital status).

Models for categorical outcomes typically parametrize the probability mass function (henceforth density) of each of the outcomes, and the primary statistical problem is to estimate the model's parameters from observed data. Formally, a model $\left\{\boldsymbol{p}(\boldsymbol{\theta}) = (p_{\boldsymbol{z}}(\boldsymbol{\theta}))_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\right\}$ subject to d parameters and parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^d$ assigns to each categorical

outcome $\boldsymbol{z} = (z_1, z_2, \ldots, z_k)^\top \in \boldsymbol{\mathcal{Z}}$ a nonnegative probability

$$p_{\boldsymbol{z}}(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{Z} = \boldsymbol{z}] = \mathbb{P}_{\boldsymbol{\theta}}[Z_1 = z_1, \ldots, Z_k = z_k] \quad \text{such that} \quad \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} p_{\boldsymbol{z}}(\boldsymbol{\theta}) = 1, \tag{1}$$

which depends on a constant but unobserved parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that one wishes to estimate by using realizations of categorical $\boldsymbol{Z}$. Such models can be used for predicting test results (like the Rasch model in item response theory), modeling the association between latent personality traits, categorical regression, or predicting how often a certain event occurs, just to name a few. I present a selection of such models in Appendix A. Textbook treatments can be found in Agresti (2010) for ordinal variables and Agresti (2012) for nominal ones. To fix ideas, the following section focuses in more detail on a particularly relevant type of categorical model, namely structural equation modeling of questionnaire responses, the harmful effects that misspecification of such models can entail, and how issues stemming from misspecification can be alleviated with the methodology proposed in this paper.

2.2 Lead example: Structural Equation Modeling

Structural equation models (SEMs) are a fundamental tool to analyze questionnaire responses in the social, behavioral, and business sciences. SEMs are typically based on factor analyses. Factor models are designed to explain the correlation between a set of modeled variables in terms of a small number of latent variables called *factors* (Mardia et al., 1979, Ch. 9). Such models are attractive for questionnaire data because questionnaires are typically designed to measure latent constructs such as attitudes or personality traits by means of multiple rating items for each construct (repeated elicitation). Multiple items measuring the same construct are called a *scale*. Formally, a factor model for r factors (like constructs) modeling q variables (like rating responses) is in its most basic form given by

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}, \tag{2}$$

where $\boldsymbol{\Sigma}$ is the population correlation matrix of the q modeled variables,[1] $\boldsymbol{\Lambda}$ is a $q \times r$ matrix of coefficients (factor loadings) that summarize the variation explained by the factors, and $\boldsymbol{\Psi}$ is a diagonal covariance matrix capturing unexplained variation. Factor models like (2) are primarily used to develop, test, and validate behavioral theory (e.g. Rust et al., 2020), are frequently utilized in economics (e.g. Almlund et al., 2011; Borghans et al., 2008; Heckman et al., 2006; Osborne-Groves, 2005), and offer many econometric extensions (see Ch. 3 in Schennach, 2022, for an overview).

In practice, one first estimates population correlation matrix $\boldsymbol{\Sigma}$ from some sample of the q modeled variables and then fits factor model (2) to the estimated correlation matrix

[1] One can also use covariance matrices for factor models, but it is generally recommended to use correlation matrices to account for scale differences in the measurement of the modeled variables (e.g. Mardia et al., 1979, Ch. 9).

by estimating $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, subject to some identification constraints. However, hitherto estimation methods like maximum likelihood are highly susceptible to misspecification of factor model (2), resulting in inconsistent estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, which in turn can lead to incorrect research findings. In the context of factor models for questionnaire responses, a prominent source of misspecification are inattentive or careless responding, which have been shown to be a serious threat for the validity of SEM analyses due to a variety of psychometric issues, such as, for example but not limited to, reduced scale reliability (Arias et al., 2020), reduced construct validity (Kam & Meyer, 2015), improper factor structure (Arias et al., 2020; Huang et al., 2015b; Woods, 2006), as well as errors in hypothesis testing (Maniaci & Rogge, 2014; McGrath et al., 2010; Woods, 2006), and there is evidence that already a small proportion of inattentive respondents of 5–10% are problematic (Arias et al., 2020; Credé, 2010; Schmitt & Stults, 1985; Woods, 2006).

Since factor analyses are fundamentally a model of the correlation structure of the modeled variables (see eq. 2), one approach to robustify factor analyses against inattentive responding (or other sources of misspecification) is to robustify the estimation of the population correlation matrix $\mathbf{\Sigma}$ that is being modeled.[2] The commonly employed Pearson sample correlation matrix of rating items (where each response option is encoded by an integer) is strongly affected by inattentive responses (Credé, 2010; Raymaekers & Rousseeuw, 2021), so a different correlation estimator is needed. An often recommended alternative to sample correlation between rating items is *polychoric* correlation (e.g. Foldnes & Grønneberg, 2022; Garrido et al., 2013; Holgado-Tello et al., 2010), but polychoric correlation estimated by maximum likelihood is also sensitive to inattentive responding (cf. Foldnes & Grønneberg, 2022) and generally yields similar estimates as sample correlation when there are five or more rating options (Rhemtulla et al., 2012). However, it turns out that estimating the correlation matrix $\mathbf{\Sigma}$ by means of polychoric correlation coefficients is robust to inattentive responding if the latter are estimated with the robust estimator proposed in this paper. It follows that one can make factor analyses on questionnaire responses robust to inattentive responding by fitting the factor model to a polychoric correlation matrix estimated by my robust estimator. Consequently, subsequent analyses that rely on estimated correlation or factor structure such as scale reliability (e.g. coefficients α (Cronbach, 1951) or ω (McDonald, 1970)) can also be robustified against inattention with this approach.

In the pairwise polychoric correlation model (Pearson & Pearson, 1922; Pearson, 1901), one observes a bivariate variable $\boldsymbol{Z} = (X, Y)^\top$ of ratings that takes values in the set $\boldsymbol{\mathcal{Z}} = \mathcal{X} \times \mathcal{Y} = \{1, 2, \ldots, J_x\} \times \{1, 2, \ldots, J_y\}$, that is, the items associated with X and Y have J_x and J_y response categories, respectively. The polychoric model assumes that the probability

[2]For factor models of continuous observed variables, Pison et al. (2003) propose to robustly estimate the correlation matrix by means of the outlier-robust *Minimum Covariance Determinant* estimator (Rousseeuw, 1985). However, this estimator crucially relies on continuity and Gaussianity of the modeled variables, and is therefore not applicable to categorical data (and may not even be computable for such).

of observing responses $\boldsymbol{z} = (x, y) \in \mathcal{X} \times \mathcal{Y}$ is given by

$$p_{\boldsymbol{z}}(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}[X = x, Y = y] = \int_{a_{x-1}}^{a_x} \int_{b_{y-1}}^{b_y} \phi_2(t, s; \varrho) \, \mathrm{d}s \, \mathrm{d}t, \tag{3}$$

where $\phi_2(\cdot, \cdot; \varrho)$ is the bivariate standard normal density with correlation parameter $\varrho \in (-1, 1)$, and real-valued threshold parameters $a_1 < \cdots < a_{J_x - 1}$, $b_1 < \cdots < b_{J_y - 1}$ with the convention $a_0 = b_0 = -\infty, a_{J_x} = b_{J_y} = +\infty$. This model is parametrized by a vector $\boldsymbol{\theta} = (\varrho, a_1, \ldots, a_{J_x - 1}, b_a, \ldots, b_{J_y - 1})^\top$ of dimension $d = J_x + J_y - 1$, where the object of primary interest is the *polychoric correlation coefficient* ϱ, which is the correlation of the two latent variables that X and Y measure. For parameter estimation in the polychoric model, maximum likelihood is typically employed (Olsson, 1979). If a given dataset comprises rating responses to q items, the associated $q \times q$ *polychoric correlation matrix* of all items holds the $q(q-1)/2$ unique pairwise polychoric correlation coefficients ϱ_{ij} for distinct items $i, j = 1, \ldots, q$. Guidelines for polychoric correlation matrices can be found in Gadermann et al. (2012).

2.3 Conceptualization of model misspecification

In the lead example of SEMs, the primarily discussed source of model misspecification are inattentive respondents in questionnaire studies. Yet, misspecifiation of models for categorical outcomes (not necessarily restricted to questionnaire responses) may manifest in numerous ways. In this section, I conceptualize general model misspecification by adopting the seminal work of Huber (1964).

Suppose that there exists a "true" but unknown parameter $\boldsymbol{\theta}_* \in \boldsymbol{\Theta}$ whose associated density a categorical random variable of interest $\boldsymbol{Z}$ follows, that is, $\boldsymbol{Z} \sim \boldsymbol{p}(\boldsymbol{\theta}) = (p_{\boldsymbol{z}}(\boldsymbol{\theta}_*))_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}}$. Subsequently, the primary statistical problem is to estimate the true $\boldsymbol{\theta}_*$ based on a sample of realizations of $\boldsymbol{Z}$. If all data points in the sample are indeed generated by the true density $\boldsymbol{p}(\boldsymbol{\theta}_*)$, then one can generally construct a consistent estimator of $\boldsymbol{\theta}_*$, for instance through maximum likelihood estimation. However, if some observed data points are *not* generated by $\boldsymbol{p}(\boldsymbol{\theta}_*)$ but some other distribution outside of model $\{\boldsymbol{p}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, then one says that the model is *misspecified*.

Without loss of generality, potential model misspecification can be conceptualized by means of a mixture density

$$\boldsymbol{\mathcal{Z}} \ni \boldsymbol{z} \mapsto f_\varepsilon(\boldsymbol{z}) = (1 - \varepsilon) p_{\boldsymbol{z}}(\boldsymbol{\theta}_*) + \varepsilon h(\boldsymbol{z}) \tag{4}$$

for *misspecification degree* $\varepsilon \in [0, 1]$ and *misspecification type* $h(\cdot)$, which is some density on $\boldsymbol{\mathcal{Z}}$ that is not contained in model $\{\boldsymbol{p}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Hence, one samples from the model distribution $p_{\boldsymbol{z}}(\boldsymbol{\theta}_*)$ with probability $1 - \varepsilon$ and from some other density $h(\boldsymbol{z})$ with probability ε. Neither ε nor $h(\cdot)$ are known and are left completely unspecified so that one makes *no assumption whatsoever* on the degree, magnitude, or type of misspecification, which might be absent altogether. Indeed, if $\varepsilon = 0$, the model is correctly specified because

the true model probability and population probability are equal to one another, that is, $f_0(\boldsymbol{z}) = p_{\boldsymbol{z}}(\boldsymbol{\theta}_*)$ for all $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$.

In the context of models for questionnaire responses (like SEMs), leaving misspecification degree ε and type $h(\cdot)$ in (4) unspecified means that the model can be misspecified due to an unlimited variety of reasons, for instance but not limited to inattentive or careless responding (e.g., straightlining, pattern responding, random-like responding), misresponses, item misunderstanding, or accurate responses that are simply not generated by the assumed model (sampling errors).

Conceptualizing model misspecification in similar fashion as (4) is commonly done in the robust statistics literature (see Maronna et al., 2018, for a recent overview) and econometric literature on misspecified moment conditions (e.g. Andrews et al., 2017; Armstrong & Kolesár, 2021; Bugni & Ura, 2019; Bonhomme & Weidner, 2022; Chernozhukov et al., 2018, 2022; Kitamura et al., 2013; Newey, 1985).

2.4 Why model in the first place?

It is worthwhile to briefly address the need for modeling in categorical data. First, minimizing the Kullback-Leibler divergence between the empirical relative frequency of the individual outcomes and some non-parametrized theoretical outcome probabilities with respect to the theoretical outcome probabilities yields as estimator the empirical relative frequencies of the outcomes, which is uninformative. Assuming a model that imposes a parametrization based on theory is standard in many fields; for instance, factor models are commonly employed in the analysis of questionnaire responses with repeated elicitation, which aids in validating, rejecting, or developing novel theories on human behavior (e.g. Ch. 9 in Mardia et al., 1979). Second, in the context of questionnaire responses, recent work by Bond & Lang (2019) has pointed out that without assumptions on the latent trait that is being measured by the questionnaire, possible location differences between two groups of respondents are not identified. Yet, Bond & Lang (2019) consider the necessary conditions for parametric and nonparametric identification to be unlikely to be satisfied in practice. In other words, a possible model on a latent trait is likely to be misspecified. However, this is exactly the type of situation where robust estimation could be useful: Robust estimators, like the one presented in this paper, implicitly allow for model misspecification while preserving estimation accuracy. Hence, if one is worried about violations of modeling assumptions like those needed for identification, it might be attractive to employ a robust estimator that is designed to work reasonably well despite violations of said assumptions.

3 Methodology

Throughout this section, suppose one observes a sample $\{\boldsymbol{Z}_i\}_{i=1}^N$ of N independent categorical draws from population density f_ε in (4), that is, the assumed model $\{\boldsymbol{p}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is

possibly misspecified. In the following, I briefly review the commonly employed maximum likelihood estimation and then proceed by describing the proposed robust estimator.

3.1 Maximum likelihood estimation

Denote by

$$N_{\boldsymbol{z}} = \sum_{i=1}^{N} \mathbb{1}\left\{\boldsymbol{Z}_i = \boldsymbol{z}\right\}$$

the empirical frequency of an outcome $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$. The maximum likelihood estimator (MLE) of the true parameter $\boldsymbol{\theta}_*$ can be expressed as

$$\widehat{\boldsymbol{\theta}}_N^{\text{MLE}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} N_{\boldsymbol{z}} \log\left(p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \right\}. \tag{5}$$

Many models for categorical data are commonly estimated via maximum likelihood (and extensions thereof), including the famous Rasch model in item response theory (Ch. 4 in Mair, 2018), the polychoric correlation model in eq. 3 (see Olsson, 1979), and even sample correlation like the Pearson correlation coefficient can be expressed as MLE.

However, while the MLE is consistent and efficient for a correctly specified model, it is well-known to be inconsistent for $\boldsymbol{\theta}_*$ when the model is misspecified and is generally highly susceptible to already small degrees of misspecification (Huber & Ronchetti, 2009; Hampel et al., 1986; Huber, 1964).

3.2 Proposed estimator

In the following, I propose an estimator of $\boldsymbol{\theta}_*$ that is asymptotically equivalent to the MLE in the absence of misspecification, and substantially more accurate than the MLE in the presence of misspecification. The estimator is based on the following observation. The empirical relative frequency of a fixed outcome $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$, denoted

$$\widehat{f}_N(\boldsymbol{z}) = N_{\boldsymbol{z}}/N = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left\{\boldsymbol{Z}_i = \boldsymbol{z}\right\},$$

is a pointwise strongly consistent nonparameteric estimator of its corresponding population cell probability

$$f_\varepsilon(\boldsymbol{z}) = (1-\varepsilon)p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_*\right) + \varepsilon h(\boldsymbol{z})$$

as $N \to \infty$ (see e.g., Chapter 19.2 in Van der Vaart, 1998). Cosequently, if there is no misspecification ($\varepsilon = 0$), then model $\{\boldsymbol{p}\left(\boldsymbol{\theta}\right) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is correctly specified because

$$f_0(\boldsymbol{z}) = p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_*\right),$$

whereas if $\varepsilon > 0$, then misspecification arises from the model's cell probabilities not being equal to the population cell probabilities, that is,

$$f_\varepsilon(\boldsymbol{z}) \neq p_{\boldsymbol{z}}(\boldsymbol{\theta}_*). \tag{6}$$

Consequently, if the model is correctly specified (and $\boldsymbol{\theta}_*$ is identified), then there exists a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ whose associated probability mass function $p_{\boldsymbol{z}}(\boldsymbol{\theta})$ the nonparametric estimator $\widehat{f}_N(\boldsymbol{z})$ will converge to, namely the true $\boldsymbol{\theta} = \boldsymbol{\theta}_*$. Conversely, if there is misspecification, then there is no $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ whose model probability $p_{\boldsymbol{z}}(\boldsymbol{\theta})$ the estimate $\widehat{f}_N(\boldsymbol{z})$ will converge to. This fact can be exploited in the estimation of the model's parameters by minimizing a certain divergence between the empirical relative outcome frequencies, $\widehat{f}_N(\boldsymbol{z})$, and theoretical outcome probabilities $p_{\boldsymbol{z}}(\boldsymbol{\theta})$ returned by the assumed model (1) at parameter value $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ to find the most accurate fit that the assumed model can offer for the observed data. Specifically, the estimator minimizes with respect to $\boldsymbol{\theta}$ the loss

$$L\left(\boldsymbol{\theta}, \widehat{f}_N\right) = \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \rho\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})}\right) p_{\boldsymbol{z}}(\boldsymbol{\theta}), \tag{7}$$

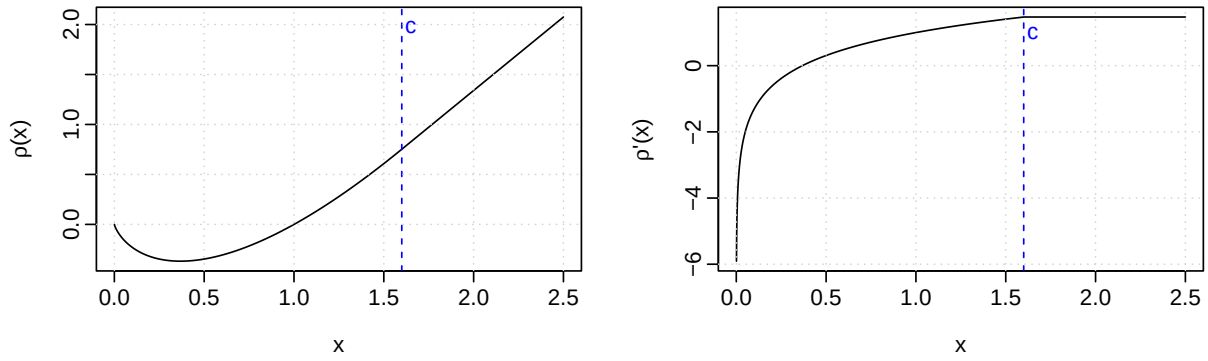
where $\rho : [0, \infty) \to \mathbb{R}$ is a predefined function that will be defined momentarily. The proposed estimator $\widehat{\boldsymbol{\theta}}_N$ is given by the value minimizing the objective loss over $\boldsymbol{\Theta}$,

$$\widehat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L\left(\boldsymbol{\theta}, \widehat{f}_N\right). \tag{8}$$

Estimators that minimize a loss function of the type in (7) are called *minimum disparity estimators* (Lindsay, 1994) because they minimize a certain disparity between empirical probabilities ($\widehat{f}_N(\boldsymbol{z})$ here) and theoretical probabilities ($p_{\boldsymbol{z}}(\boldsymbol{\theta})$ here). A small disparity indicates that the assumed model is able to fit observed data well. The disparity is governed by the choice of the function $\rho(\cdot)$. Many estimators can be written as minimum disparity estimators, including the MLE and various estimators for grouped data (Victoria-Feser & Ronchetti, 1997; Cressie & Read, 1984) through the choice of $\rho(\cdot)$ (Victoria-Feser & Ronchetti, 1997; Lindsay, 1994). In the following, I motivate a specific choice of $\rho(\cdot)$ that makes the estimator $\widehat{\boldsymbol{\theta}}_N$ less susceptible to misspecification of the polychoric model.

The fraction $\widehat{f}_N(\boldsymbol{z})/p_{\boldsymbol{z}}(\boldsymbol{\theta})$ in (7) is called a *Pearson residual*[3] (Lindsay, 1994) and can be interpreted as a goodness-of-fit measure of outcome $\boldsymbol{z}$. Values close to 1 indicate a good fit between data and assumed model at $\boldsymbol{\theta}$, whereas values toward 0 or $+\infty$ indicate a poor fit. Indeed, outcomes whose observations are primarily generated by a different distribution than that of the assumed model will generally have a Pearson residual away from 1 (see eq. 6). Hence, to achieve robustness to misspecifcation, outcome frequencies that cannot be modeled well by the assumed model, as indicated by their Pearson residual being away from 1, should

[3]Technically, Lindsay (1994) defines Pearson residuals as $\widehat{f}_N(\boldsymbol{z})/p_{\boldsymbol{z}}(\boldsymbol{\theta}) - 1$. I renounce on subtracting the value 1 because it makes notation simpler in this paper.



(a) $\rho(x)$ **(b)** $\rho'(x)$

Figure 1: Visualization of the function $\rho(x)$ in (9) (left panel) and its derivative (right panel), for $c = 1.6$ (vertical dashed blue line).

be downweighted in the estimation procedure such that they do not over-proportionally affect the fit. Such robustness can be achieved by choosing an appropriate function for $\rho(\cdot)$ in loss function (7). I propose to choose the following specification, suggested by Ruckstuhl & Welsh (2001) for robustly fitting the binomial model,

$$\rho(x) = \begin{cases} x \log(x) & \text{if } x \in [0, c], \\ x(\log(c) + 1) & \text{if } x > c, \end{cases} \tag{9}$$

where $c \in [1, \infty]$ is a prespecified tuning constant.[4] Figure 1 visualizes this function for the example choice $c = 1.6$. Note that the function $\rho(\cdot)$ is convex, but whether the estimator's optimization problem in (8) is also convex depends on the assumed model.

It is easy to see that for the choice $c = +\infty$ in function $\rho(\cdot)$, minimizing the loss (7) is equivalent to maximizing the log-likelihood objective in (5), meaning that the estimator $\widehat{\boldsymbol{\theta}}_N$ is equal to $\widehat{\boldsymbol{\theta}}_N^{\text{MLE}}$ for this choice of c. More specifically, if a Pearson residual $x = \frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})}$ of an outcome $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ is such that $x \in [0, c]$ for fixed $c \geq 1$, then the estimation procedure behaves at this cell like in maximum likelihood estimation. If this Pearson residual x equals 1, then its associated outcome can be fitted perfectly with the presumed model, so at this outcome the estimation procedure will behave like maximum likelihood *regardless* of the choice of $c \geq 1$. In the absence of misspecification ($\varepsilon = 0$), $\widehat{f}_N(\boldsymbol{z}) \xrightarrow{\text{a.s.}} p_{\boldsymbol{z}}(\boldsymbol{\theta}_*)$ as $N \to \infty$ for all $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$, meaning that all Pearson residuals are asymptotically equal to 1. In other words, if there is no misspecification, $\widehat{\boldsymbol{\theta}}_N$ is asymptotically equivalent to $\widehat{\boldsymbol{\theta}}_N^{\text{MLE}}$ no matter the choice of tuning constant $c \geq 1$. On the other hand, if an outcome's Pearson residual is far away from 1, it cannot be fitted well with the presumed model, which is typically indicative of the polychoric model being misspecified. In this case, this particular outcome should not be treated like in maximum likelihood estimation because maximum likelihood is

[4] Equation (9) is actually just a special case of a more general formulation in Ruckstuhl & Welsh (2001).

not consistent under misspecification. Instead, the outcome's influence on the final estimate should be downweighted to avoid that outcomes that cannot be fitted well dominate the fit, which happens in maximum likelihood. Such downweighting is employed by function $\rho(x)$ in (9) whenever $x > c \geq 1$, that is, the Pearson residual is sufficiently far away from the ideal value 1, where the choice of c governs what is deemed "sufficiently far". Indeed, for values $x > c$, the function $\rho(x)$ increases only linearly with x, as opposed to the non-linear exponential increase when $x \leq c$. The notion of requiring nonlinear effects for the bulk of the data and linear effects in its tails is similar to classic robust estimation as in Huber (1964).

It is shown in Figure 1 how $\rho(x)$ transitions from exponential growth to linear growth at $x = c$, as well as the boundedness of its first derivative when c is finite. Hence, if c is finite, any Pearson residual can only have a bounded effect on the final estimator, as opposed to unbounded effects in maximum likelihood estimation where $c = +\infty$. Thus, I achieve robustness against misspecification through the choice of c. The closer to 1 one chooses c, the more robust the estimator becomes. However, there is a well-known tradeoff between robustness and efficiency for robust estimators: the more robust an estimator, the more estimation variance is introduced (e.g. Huber & Ronchetti, 2009). Therefore, by choosing c, one is effectively choosing between robustness and efficiency. A characterization of the this tradeoff is work in progress.

With the proposed choice of $\rho(\cdot)$, I stress that estimator $\widehat{\boldsymbol{\theta}}_N$ in (8) has the same time complexity as maximum likelihood, that is, $O\big(\#\boldsymbol{\mathcal{Z}}\big)$, since one needs to calculate the Pearson residual of all possible outcomes for any given candidate parameter value. Consequently, the proposed estimator does not incur any additional computational cost compared to maximum likelihood, and therefore robustness can be achieved without having to pay a computational price.

3.3 Estimand

Considering the potential presence of model misspecification, it is worth studying what quantity the proposed estimator $\widehat{\boldsymbol{\theta}}_N$ in (8) estimates. In population, its estimand is given by

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L\Big(\boldsymbol{\theta}, f_\varepsilon\Big). \tag{10}$$

This minimization problem is simply the population analogue to the minimization problem in (7) that the sample-based estimator $\widehat{\boldsymbol{\theta}}_N$ solves, since the $\widehat{f}_N(\boldsymbol{z})$ are the population analogues to the $f_\varepsilon(\boldsymbol{z})$. In the absence of misspecification ($\varepsilon = 0$), estimand $\boldsymbol{\theta}_0$ equals the true parameter $\boldsymbol{\theta}_*$. In the presence of misspecification ($\varepsilon > 0$) it is generally different from $\boldsymbol{\theta}_*$. How much different it is depends on the degree and type of of misspecification as well as the choice of tuning constant c in $\rho(\cdot)$. In general, the larger ε (more severe misspecification) and c (less downweighting of hard-to-fit cells), the further $\boldsymbol{\theta}_0$ is away from $\boldsymbol{\theta}_*$. Hence, for fixed misspecification degree ε, the MLE ($c = +\infty$) will estimate a parameter that is farther or equally far away from the true $\boldsymbol{\theta}_*$ than for finite choices of c. Correspondingly, finite

choices of c lead to an estimator that is at least as accurate as the MLE, and more accurate under misspecification of the presumed model.

3.4 Assumptions

In the following, I list a set of assumptions that will be entertained in the asymptotic analysis of the proposed estimator $\widehat{\boldsymbol{\theta}}_N$ computed on random sample $\{\boldsymbol{Z}_i\}_{i=1}^N$.

Assumption Set A. *Suppose that the following assumptions hold true.*

A.1 $c \in [1, +\infty]$

A.2 $\boldsymbol{\Theta} \subset \mathbb{R}^d$ *is compact,*

A.3 $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}, f_\varepsilon)$ *is a unique global minimum, and* $\boldsymbol{\theta}_0$ *is an interior point of* $\boldsymbol{\Theta}$*,*

A.4 $p_{\boldsymbol{z}}(\boldsymbol{\theta})$ *is continuously differentiable with respect to* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and twice differentiable at* $\boldsymbol{\theta}_0$ *for all cells* $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$*,*

A.5 $\left\| \frac{\partial p_{\boldsymbol{z}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| < \infty$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$*,*

A.6 $p_{\boldsymbol{z}}(\boldsymbol{\theta}) > 0$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$*.*

A.7 $\#\{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}} : f_\varepsilon(\boldsymbol{z}) > 0\} > d$*,*

A.8 $L(\boldsymbol{\theta}, f_\varepsilon)$ *is convex in a neighborhood of* $\boldsymbol{\theta}_0$*,*

A.9 $\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_0)} \neq c$ *for any* $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$*.*

Assumption A.1 ensures that function $\rho(\cdot)$ exhibits meaningful behavior when evaluated at Pearson residuals, such as the ideal residual value 1 being included in the interval $[0, c]$. Compactness of the parameter space (Assumption A.2) is primitive and required for a technicality when proving consistency of $\widehat{\boldsymbol{\theta}}_N$.[5] Uniqueness of a global minimum in the parameter space's interior (Assumption A.3) is a common assumption in the literature on M-type estimators (e.g., Chapter 5.2 in Van der Vaart, 1998) like the one presented in this paper. Assumptions A.4 and A.5 pertain to the presumed model and require smoothness and first-order boundedness, as well as second-order differentiability at estimand $\boldsymbol{\theta}_0$. These assumptions are standard in the literature on minimum-disparity-type estimators (e.g., Cressie & Read, 1984; Ruckstuhl & Welsh, 2001; Victoria-Feser & Ronchetti, 1997). The assumption of strictly positive model probabilities (A.6) is also standard in this literature, and rules out that one divides by zero when calculating Pearson residuals. Related is Assumption A.7

[5] This assumption can possibly be modified to $\boldsymbol{\Theta}$ being open by equipping it with a specific topological structure.

imposes that the number of positive cell probabilities in population is strictly larger than the number of model parameters. In other words, there must be more sources of variation (populated cells) than parameters. One may view this assumption as a rank condition that ensures invertibility of the Hessian matrix of the minimization problem (7), and is required to prevent rank deficiency of the asymptotic covariance matrix of estimator $\widehat{\boldsymbol{\theta}}_N$. Assumption A.8 is a local convexity assumption that only becomes relevant when the population minimization problem (10) is not convex, which can happen for some models. Specifically, this assumption rules out that the gradient of population loss $L(\boldsymbol{\theta}, f_\varepsilon)$ is flat at value zero in a region around $\boldsymbol{\theta}_0$, because such a situation would precipitate an identification problem: If the gradient is zero-valued in a neighborhood for multiple adjacent points, then estimand $\boldsymbol{\theta}_0$ is not uniquely determined. As such, this assumption refines the assumption of $\boldsymbol{\theta}_0$ being a global minimum of the population loss (Assumption A.3) by requiring well-separatedness of the global minimum, which is a common assumption in the asymptotic analysis of M-type estimators (see Ch. 5.2 in Van der Vaart, 1998, for a discussion). Finally, Assumption A.9 imposes that the Pearson residual at the global minimum is not equal to tuning constant c. This is a primitive condition that is required for the loss to be twice differentiable at $\boldsymbol{\theta}_0$ (in combination with Assumption A.4), which is a requirement for the existence of the estimator's asymptotic covariance matrix.

I emphasize that no assumption in Assumption Set A restricts the type, source, or magnitude of potential misspecification of the considered model. In fact, Assumptions A.2–A.8 are also required for consistency and asymptotic normality of the MLE. Only assumptions A.1 and A.9 are specific to the proposed estimator because they pertain to tuning constant c.

3.5 Asymptotic analysis

The following theorem establishes strong consistency of $\widehat{\boldsymbol{\theta}}_N$ for $\boldsymbol{\theta}_0$. The theorem's proof and that of all subsequent mathematical statements are given in Appendix D.

Theorem 1 (Consistency)**.** *Under Assumptions A.1–A.6, it holds true that*

$$\widehat{\boldsymbol{\theta}}_N \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0,$$

as $N \to \infty$.

I now study the limit distribution of the estimator. Doing so necessitates additional notation. For fixed tuning constant $c \geq 1$, let

$$w(x) = \mathbb{1}\left\{x \in [0, c]\right\} + c\mathbb{1}\left\{x > c\right\} / z \quad \text{for } x \geq 0,$$

with first derivative

$$w'(x) = 0\mathbb{1}\left\{x \in [0, c]\right\} - c\mathbb{1}\left\{x > c\right\} / x^2,$$

and further define the d-dimensional gradient of $\log\left(p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)$ for cell $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ at parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as

$$\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{1}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \, p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \right),$$

as well as the $d \times d$ Hessian matrix of $\log(p_{\boldsymbol{z}}(\boldsymbol{\theta}))$ as

$$\boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{1}{p_{\boldsymbol{z}}(\boldsymbol{\theta})} \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} p_{\boldsymbol{z}}(\boldsymbol{\theta}) \right) - \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}) \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})^\top.$$

In addition, define d-dimensional vectors

$$\boldsymbol{w}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mathbb{1} \left\{ \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})} \in [0, c] \right\},$$

and a $d \times m$ matrix

$$\boldsymbol{W}(\boldsymbol{\theta}) = \left(\boldsymbol{w}_{\boldsymbol{z}_1}(\boldsymbol{\theta}), \boldsymbol{w}_{\boldsymbol{z}_2}(\boldsymbol{\theta}) \cdots, \boldsymbol{w}_{\boldsymbol{z}_m}(\boldsymbol{\theta}) \right)$$

that row-binds all m vectors of $\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})$ multiplied by an indicator that takes value 1 when associated population Pearson residual is in the MLE-part of the function $\rho(\cdot)$ in (9) and 0 otherwise. In similar fashion, define the m-dimensional vector

$$\boldsymbol{f}_\varepsilon = \Big(f_\varepsilon(\boldsymbol{z}_1), f_\varepsilon(\boldsymbol{z}_2), \ldots, f_\varepsilon(\boldsymbol{z}_m) \Big)^\top$$

that holds all m evaluations of the population density f_ε, and put

$$\boldsymbol{\Omega} = \operatorname{diag}(\boldsymbol{f}_\varepsilon) - \boldsymbol{f}_\varepsilon \boldsymbol{f}_\varepsilon^\top.$$

With this notation, I can establish root-N consistency and asymptotic normality of estimator $\widehat{\boldsymbol{\theta}}_N$ in the following theorem.

Theorem 2 (Asymptotic normality)**.** *Grant the assumptions of Assumption Set A. Then*

$$\sqrt{N} \left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0 \right) \xrightarrow{\mathrm{d}} \mathrm{N}_d\Big(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \Big),$$

as $N \to \infty$, where, for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{M}(\boldsymbol{\theta})^{-1} \boldsymbol{U}(\boldsymbol{\theta}) \boldsymbol{M}(\boldsymbol{\theta})^{-1},$$

with $d \times d$ symmetric matrices

$$\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{W}(\boldsymbol{\theta}) \boldsymbol{\Omega} \boldsymbol{W}(\boldsymbol{\theta})^\top \qquad \textit{and}$$

$$\boldsymbol{M}(\boldsymbol{\theta}) = \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} f_\varepsilon(\boldsymbol{z}) \left(w' \left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})} \right) \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}) \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})^\top - w \left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})} \right) \boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{\theta}) \right).$$

A strongly consistent estimator of the unobserved asymptotic covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ can be constructed as follows. Replace all population class probabilities $f_\varepsilon(\boldsymbol{z})$ by their corresponding empirical counterparts $\widehat{f}_N(\boldsymbol{z})$ in matrices $\boldsymbol{W}(\boldsymbol{\theta})$, $\boldsymbol{M}(\boldsymbol{\theta})$, and $\boldsymbol{\Omega}$. Then exploit the plug-in principle and evaluate $\boldsymbol{U}(\boldsymbol{\theta})$ and $\boldsymbol{M}(\boldsymbol{\theta})$ at the point estimate $\widehat{\boldsymbol{\theta}}_N$. Denote the ensuing plug-in estimator by $\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}_N\right)$, which is strongly consistent for $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ by Theorem 1 and the continuous mapping theorem.

3.6 Goodness-of-fit test

Suppose one wishes to test the null hypothesis that an individual cell in a k-way contingency table can be fitted well by the presumed model. Rejecting this null hypothesis is indicative of the model's misspecification, at least for that cell. This notion can be conceptualized by means of Pearson residuals. Recall that a Pearson residual of value 1 indicates that the corresponding cell can be fitted well, whereas a Pearson residual significantly larger than 1 indicates poor fit. For a given cell $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$, this translates into the natural null hypothesis with one-sided alternative

$$\mathrm{H}_0 : \frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)} = 1 \qquad \text{vs.} \qquad \mathrm{H}_1 : \frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)} > 1,$$

which is equivalent to

$$\mathrm{H}_0 : p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right) = \widehat{f}_N(\boldsymbol{z}) \qquad \text{vs.} \qquad \mathrm{H}_1 : p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right) < \widehat{f}_N(\boldsymbol{z}). \tag{11}$$

Ideally, a test for such a hypothesis will reject H_0 if the presumed model is misspecified for that cell, and sustain H_0 if it is correctly specified for that cell. It turns out that a test statistic that satisfies these two desirable properties is given by

$$T_N(\boldsymbol{z}) = \frac{p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right) - \widehat{f}_N(\boldsymbol{z})}{\sqrt{\sigma_{\boldsymbol{z}}^2\left(\boldsymbol{\theta}_0\right)/N}}, \tag{12}$$

where

$$\sigma_{\boldsymbol{z}}^2\left(\boldsymbol{\theta}\right) = \boldsymbol{g}_{\boldsymbol{z}}(\boldsymbol{\theta})^\top \boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right) \boldsymbol{g}_{\boldsymbol{z}}(\boldsymbol{\theta})$$

for gradient

$$\boldsymbol{g}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{\partial p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}.$$

The intuition behind using test statistic $T_N(\boldsymbol{z})$ to test $\mathrm{H}_0 : p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right) = \widehat{f}_N(\boldsymbol{z})$ is as follows. If the presumed model is correctly specified ($\varepsilon = 0$), it holds true that $\widehat{f}_N(\boldsymbol{z}) \xrightarrow{\text{a.s.}} p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_*\right)$ as well as $p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right) \xrightarrow{\text{a.s.}} p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_*\right)$ (by Theorem 1 and the continuous mapping theorem). It follows that the difference between $p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right)$ and $\widehat{f}_N(\boldsymbol{z})$ vanishes as N grows to infinity. Indeed, the following corollary (Corollary 1) shows that this difference, when scaled appropriately to equal $T_N(\boldsymbol{z})$, converges in distribution to a zero-mean Gaussian distribution, provided that the model is correctly specified. However, if the model is misspecified ($\varepsilon > 0$), then $\widehat{f}_N(\boldsymbol{z}) \xrightarrow{\text{a.s.}} f_\varepsilon(\boldsymbol{z})$ and $p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right) \xrightarrow{\text{a.s.}} p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)$, but $f_\varepsilon(\boldsymbol{z}) \neq p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)$ because of mmisspecification. Consequently, the difference between $p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right)$ and $\widehat{f}_N(\boldsymbol{z})$ does not converge to zero (and need not converge at all). Hence, if $T_N(\boldsymbol{z})$ is statistically significantly different from a zero-mean Gaussian sequence, there is evidence that the presumed model is misspecified, at least at cell $\boldsymbol{z}$. The following corollary of Theorem 2 formally establishes the validity of test statistic $T_N(\boldsymbol{z})$ for testing the null hypothesis $\mathrm{H}_0 : \widehat{f}_N(\boldsymbol{z}) = p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)$.

Corollary 1 (Limit distribution of test statistic)**.** *Grant the assumptions of Assumption Set A and consider a given cell* $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$*. Then, under the null hypothesis in* (11)*, the test statistic* $T_N(\boldsymbol{z})$ *in* (12) *possesses the following limit distribution:*

$$T_N(\boldsymbol{z}) \xrightarrow{\text{d}} \text{N}(0, 1),$$

as $N \to \infty$*.*

In practice, the variance term $\sigma_{\boldsymbol{z}}^2(\boldsymbol{\theta}_0)$ in test statistic $T_N(\boldsymbol{z})$ is unobserved, but it can be strongly consistently estimated by $\sigma_{\boldsymbol{z}}^2\left(\widehat{\boldsymbol{\theta}}_N\right)$, which follows from Theorem 1 and the continuous mapping theorem. Hence, in practice, one uses the approximate test statistic

$$\widehat{T}_N(\boldsymbol{z}) = \frac{p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right) - \widehat{f}_N(\boldsymbol{z})}{\sqrt{\sigma_{\boldsymbol{z}}^2\left(\widehat{\boldsymbol{\theta}}_N\right)/N}}$$

for hypothesis testing. In addition, it should be noted that using this test to test the null hypothesis in (11) for all m outcomes in sample space $\boldsymbol{\mathcal{Z}}$ for being outlying creates a multiple comparisons problem. I therefore recommend to adjust the ensuing p-values for multiple comparisons, for instance through the procedure of Benjamini & Hochberg (1995).

As additional theoretical result, I derive the influence function of estimator $\widehat{\boldsymbol{\theta}}_N$ in Appendix B. It turns out that for tuning constant choices $c > 1$, the influence function equals that of the non-robust MLE, but remains bounded at categorical data. Similar results have been derived for minimum power divergence estimators (Victoria-Feser & Ronchetti, 1997), minimum Hellinger distance estimators (Simpson, 1987; He & Simpson, 1993; Lindsay, 1994), and minimum disparity estimators at the binomial model (Ruckstuhl & Welsh, 2001). However, I argue that the influence function is not a very informative measure when working with categorical data because the influence of a single categorical observation is bounded by construction due to the bounded nature of categorical data.

3.7 Computational and practical aspects

The minimization problem of the estimator in (7) can be solved with standard gradient descent methods, which allows one to impose constraints on the model parameters. In the spirit of open and reproducible science and to enhance accessibility and adoption by empirical researchers, an R implementation of the proposed methodology is freely available in the package `robcat` (Welz, 2024, for "ROBust CATegorical data analysis"). To maximize speed and performance, the package is predominantly developed in C++ and integrated to R via `Rcpp` (Eddelbuettel, 2013).

As for the tuning constant $c \in [1, +\infty]$ in function $\rho(\cdot)$ in (9), the choice $c = 1.6$ yielded a good compromise between robustness and efficiency in numerous simulation experiments. I therefore recommend this choice for practical use, but a detailed investigation on the optimal choice of c is work in progress.

4 Monte Carlo experiments

In order to verify the theoretical guarantees of the proposed estimator and demonstrate its performance in practice, I employ a series of Monte Carlo simulation experiments. In these experiments, I emulate a SEM on questionnaire responses where some respondents only give inattentive responses. As described in Section 2.2, polychoric correlation coefficients estimated with my proposed robust estimator is the workhorse behind robustifying SEMs against inattentive responding (among other sources of misspecification). Therefore, the simulation design is twofold: In the first design, I study estimation of the pairwise polychoric model in (3), while the second design focuses on fitting SEMs.

4.1 Polychoric correlation

Consider a pairwise polychoric correlation model for two rating items (eq. 3). Let there be $J_x = J_y = 5$ response categories for each of the two rating variables $\boldsymbol{Z} = (X, Y)^\top$ and define the true thresholds in the polychoric model as

$$a_{*,1} = b_{*,1} = -1.5, \quad a_{*,2} = b_{*,2} = -0.5, \quad a_{*,3} = b_{*,3} = 0.5, \quad a_{*,4} = b_{*,4} = 1.5, \tag{13}$$

and let the true polychoric correlation coefficient be $\varrho_* = 0.5$. Then, for a bivariate standard normally distributed latent random variable with correlation coefficient ϱ_*, denoted (ξ, η), observations of rating responses (X, Y) are generated by

$$X = \begin{cases} 1 & \text{if } \xi < a_1, \\ 2 & \text{if } a_1 \leq \xi < a_2, \\ 3 & \text{if } a_2 \leq \xi < a_3, \\ 4 & \text{if } a_3 \leq \xi < a_4, \\ 5 & \text{if } a_4 \leq \xi, \end{cases} \quad \text{and} \quad Y = \begin{cases} 1 & \text{if } \eta < b_1, \\ 2 & \text{if } b_1 \leq \eta < b_2, \\ 3 & \text{if } b_2 \leq \eta < b_3, \\ 4 & \text{if } b_3 \leq \eta < b_4, \\ 5 & \text{if } b_4 \leq \eta. \end{cases} \tag{14}$$

To emulate misspecification of the polychoric model, let a fraction ε of the latent (ξ, η) be generated by a bivariate normal distribution with mean $(2, -2)^\top$, variances $(0.2, 0.2)^\top$, and zero covariance (and therefore zero correlation) that causes the polychoric model (3) to be misspecified. Hence, in simulated data with nonzero misspecification ε, the empirical frequency of cells $(x, y) \in \{(5, 1), (4, 3), (5, 2)\}$ will be inflated in the sense that they have a higher realization probability than under the true polychoric model distribution. The data points causing these three cells to be inflated are instances of *negative leverage points*. Here, such leverage points drag correlational estimates away from a positive value towards zero or, if there are sufficiently many of them, even a negative value.

For misspecification degrees $\varepsilon \in \{0, 0.1, 0.2\}$, I sample $N = 1,000$ rating responses from this data generating process and estimate the true parameter $\boldsymbol{\theta}_* = (\varrho_*, a_{*,1}, \ldots, a_{*,4}, b_{*,1}, \ldots, b_{*,4})^\top$ by means of the Pearson sample correlation coefficient

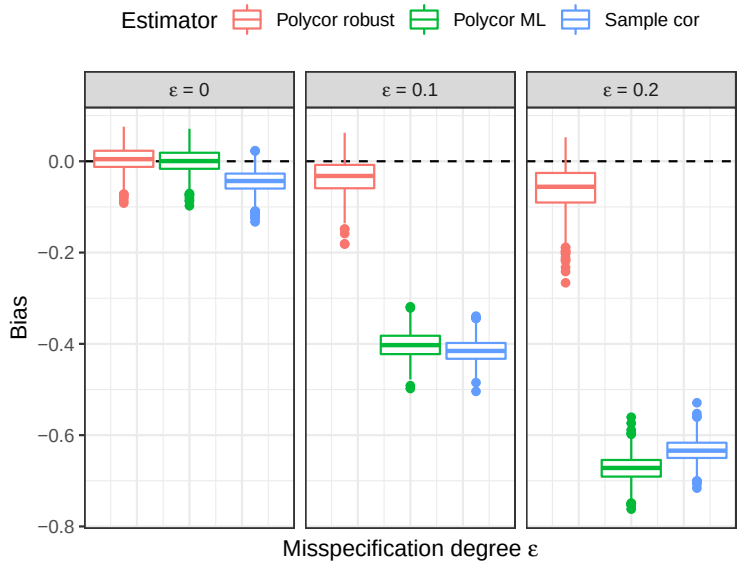


Figure 2: Boxplot visualization of the bias $(\widehat{\varrho}_N - \varrho_*)$ of the three estimators, Pearson sample correlation, the MLE, and the proposed robust estimator with $c = 1.6$, for various degrees of misspecification across 1,000 simulated datasets.

(for ϱ_*), the MLE, as well as the proposed estimator with tuning constant set to $c = 1.6$, since this choice yielded a good compromise between robustness and efficiency in further simulation studies. This procedure is repeated for 1,000 simulated datasets. As performance measures, I calculate the average bias, standard deviation across repetitions, coverage, and length of confidence intervals at significance level $\alpha = 0.05$. The coverage is defined as proportion of $(1 - \alpha)$-th confidence intervals $[\widehat{\varrho}_N \mp q_{1-\alpha/2} \cdot \text{SE}(\widehat{\varrho}_N)]$ that contain the true ϱ_*, where $q_{1-\alpha/2}$ is the $(1 - \alpha/2)$-th quantile of the standard normal distribution and $\text{SE}(\widehat{\varrho}_N)$ is the standard error of $\widehat{\varrho}_N$, which is constructed using the limit theory developed in Theorem 2. The length of a confidence interval is given by $2 \cdot q_{1-\alpha/2} \cdot \text{SE}(\widehat{\varrho}_N)$.

Figure 2 visualizes the bias of each estimator with respect to the true polychoric correlation ϱ_* across the 1,000 simulated datasets. Analogous plots for the whole parameter $\boldsymbol{\theta}_*$ can be found in Appendix C; the results are similar to those of ϱ_*. For all considered misspecification degrees, the estimates of the MLE and sample correlation are similar, which is expected because these two estimators are known to yield similar similar results when there are five or more rating options (cf. Rhemtulla et al., 2012). In the absence of misspecification, both MLE and the robust estimator yield accurate estimates. Both estimates are nearly equivalent to one another in the sense that their point estimates, standard deviation, and coverage at significance level $\alpha = 0.05$ are very similar (Table 1). However, when misspecification is introduced, MLE, sample correlation, and robust estimator yield noticeably different results. At misspecification degree $\varepsilon = 0.1$, MLE and sample correlation are substantially biased with average estimates of 0.097 and 0.084, corresponding to biases

Misspecification	Estimator	$\widehat{\varrho}_N$	Bias	StDev	Coverage	CI length
$\varepsilon = 0$	Polycor robust	0.504	0.004	0.027	0.930	0.104
	Polycor ML	0.500	0.000	0.026	0.943	0.102
	Sample cor	0.457	−0.043	0.025	0.702	0.110
$\varepsilon = 0.1$	Polycor robust	0.466	−0.034	0.038	0.911	0.152
	Polycor ML	0.097	−0.403	0.029	0.000	0.134
	Sample cor	0.084	−0.416	0.026	0.000	0.124
$\varepsilon = 0.2$	Polycor robust	0.439	−0.061	0.051	0.951	0.220
	Polycor ML	−0.172	−0.672	0.028	0.000	0.133
	Sample cor	−0.133	−0.633	0.026	0.000	0.123

Table 1: Performance measures of of the three estimators, Pearson sample correlation, the MLE, and the proposed robust estimator with $c = 1.6$, for various degrees of misspecification across 1,000 simulated datasets. The true polychoric correlation coefficient is $\rho_* = 0.5$. The performance measures are the average point estimate of the polychoric correlation coefficient, $\widehat{\varrho}_N$, average bias $(\widehat{\varrho}_N - \varrho_*)$, the standard deviation of the $\widehat{\varrho}_N$ ("StDev"), the estimator's coverage with respect to the true ϱ_* at significance level $\alpha = 0.05$, as well as the length of the estimator's confidence interval, again at level $\alpha = 0.05$.

of −0.403 and −0.416, respectively, as well as zero coverage. In contrast, the robust estimator maintains accuracy with an average estimate of 0.466, which corresponds to only a minor bias of −0.034 and a good coverage of 0.911 (Table 1). When the misspecification is increased to $\varepsilon = 0.2$, the contrast between the estimators becomes even stronger. While the robust estimator is still remarkably close to the truth with a small bias of −0.061, MLE and sample correlation produce estimates that are not only severely biased (biases of −0.672 and −0.633), but also sign-flipped: While the true correlation is strongly positive (0.5), both estimates are considerably negative (−0.172 and −0.133). It is worth noting that in the presence of misspecification, the confidence intervals of the robust estimator are wider than those of the MLE (see Table 1). This is expected because of the well-known trade-off between robustness and efficiency: An estimator that is designed to reduce bias, like a robust estimator, will inevitably have a larger estimation variance (e.g. Huber & Ronchetti, 2009). These wider confidence intervals furthermore explain why the robust estimator improves its coverage in Table 1 when the degree of misspecification is increased from 0.1 to 0.2.

This first simulation study demonstrated that already a small degree of misspecificaion of the polychoric model can render the commonly employed MLE and sample correlation unreliable, while the proposed robust estimator retains good accuracy even in the presence of considerable misspecificaion. On the other hand, when the model is correctly specified, MLE and robust estimator produce equivalent results. With these affirmative results in mind, I proceed to the second simulation study, where a structural equation model is fitted to a robustly estimated polychoric correlation matrix that was computed on data for which the model is misspecified due to inattentive responding.

4.2 Structural Equation Modeling

This second simulation design emulates a popular type of SEM, namely a confirmatory factor analysis (CFA). In a CFA, a scale compprising multiple rating items is tested for validity and reliability. That is, does the scale indeed measure the latent construct it is supposed to measure, and does it so reliably?

Suppose there are $q = 4$ rating items that jointly measure the same unidimensional ($r = 1$) latent construct in a reliable manner by having a $q \times r$ factor loadings matrix

$$\mathbf{\Lambda} = (0.75, 0.75, 0.75, 0.75)^\top$$

and diagonal $q \times q$ noise covariance matrix $\mathbf{\Psi}$ with diagonal elements all being equal to 0.4375. Then, by factor model (2), the ensuing $q \times q$ population correlation matrix $\mathbf{\Sigma}$ is characterized by having the same pairwise correlation of 0.5625 between all distinct items. It follows that this scale is indeed reliable for the latent construct with a population Cronbach-α of 0.84.[6]

I then generate $N = 1,000$ responses to the $q = 4$ rating items with $J = 5$ response options each by drawing from a zero-mean normal distribution with correlation matrix $\mathbf{\Sigma}$ and use these draws to generate rating responses according to discretization process (14), where the same thresholds as in (13) are used for each variable. To emulate inattentive responding, I replace a fraction $\varepsilon \in \{0, 0.1, 0.2\}$ of the generated rating observations with the negative leverage point $(1, 5, 1, 5)^\top$. Such a response vector corresponds to a respondent who alternates between the most extreme response points, regardless of item content.[7] Based on the resulting $N \times q$ data matrix of ratings, I estimate the correlation matrix $\mathbf{\Sigma}$ by means of the standard Pearson sample correlation matrix as well as pairwise polychoric correlation coefficients estimated robustly (with choice $c = 1.6$) and via MLE. Then, based on the estimated correlation matrix $\mathbf{\Sigma}$, the factor model (2) is fitted with $r = 1$ factor by means of maximum likelihood[8] to obtain estimates of loadings matrix $\mathbf{\Lambda}$ and error covariance matrix $\mathbf{\Psi}$. In addition, I estimate the scale's reliability by calculating an estimate of Cronbach's-α based on the estimated correlation matrix. This procedure is repeated 1,000 times. As performance measures, I calculate the root mean squared error (RMSE) between estimate and true value of each element in matrices $\mathbf{\Lambda}$ and $\mathbf{\Sigma}$ as well as Cronbach's α.

Figure 3 visualizes the simulation results by means of boxplots. Like in the simulation design for pairwise polychoric correlations (Figure 2), MLE-based estimation of the polychoric correlation matrix yields similar results as estimation based on the sample correlation matrix, which is in line with expectations (cf. Rhemtulla et al., 2012). In the absence of inattentive responding ($\varepsilon = 0$), both polychoric estimates (MLE and robust) are equivalent and accurate with respect to all three performance measures. However, once inattentive

[6] Cronbach's α (Cronbach, 1951) is a lower bound on the reliability of a scale, and is computed from the scale's correlation matrix. In general, a scale with Cronbach-α of 0.75 or higher is considered reliable (Robinson, 2018).

[7] This response vector may alternatively be interpreted as the responses of a straightliner at the first option, after recoding negatively worded items (which would here be the second and fourth item).

[8] Alternatively, one could also fit a factor model via principal components or least-squares approaches.

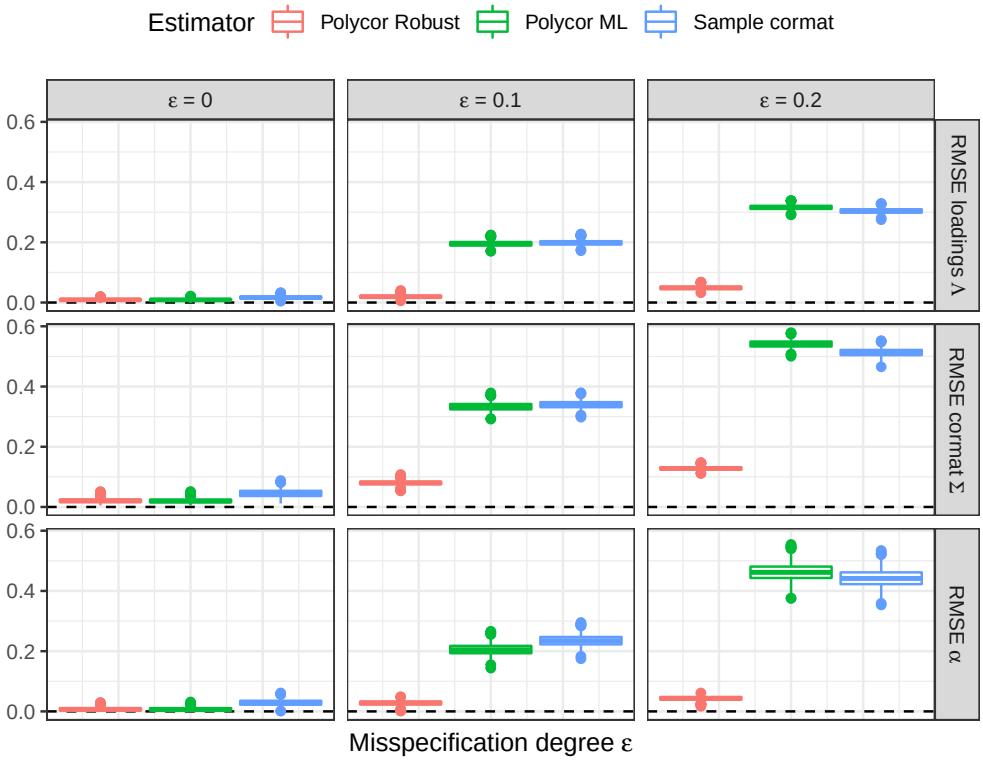


Figure 3: Boxplot visualization of the RMSE of loadings matrix $\mathbf{\Lambda}$, correlation matrix $\mathbf{\Sigma}$, and Cronbach's α for the three estimators for various degrees of misspecification across 1,000 simulated datasets.

responding is introduced, the MLE and sample correlation-based estimates display a large bias: At misspecification degree $\varepsilon = 0.1$, both have an RMSE of about 0.2 and 0.35 for the loadings matrix and correlation matrix, respectively, as well as an inaccurate estimate of Cronbach's α: While the true value of Cronbach's α of 0.84 indicates a reliable scale, the two non-robust estimators would suggest an rather unreliable scale with an estimated reliability of only about 0.65. In contrast, the robust estimator remains accurate with respect to all three performance measures. In addition, while the performance of the two non-robust estimators further deteriorates with increasing prevalence of inattentive responding ($\varepsilon = 0.2$), that of the robust estimator remains stable and almost unaffected.

To conclude this section, my Monte Carlo experiments demonstrated the performance and practical usefulness of the proposed robust estimator when applied to a SEM: While hitherto estimation approaches are highly sensitive to the adverse effects of inattentive responding, the robust estimator remains almost unaffected by their presence although it makes no assumption about the inattention's magnitude or type. On the other hand, in the absence of inattentive responding, the robust estimator yields equivalent results to hitherto estima-

tors. These simulation results suggest that a SEM can be robustified against inattentive responding by using the novel robust estimator proposed in this paper.

5 Empirical application

5.1 Background and study design

In this section, I demonstrate the proposed methodology on empirical data by using a subset of the 100 unipolar markers of the Big-5 personality traits (Goldberg, 1992).[9] Each marker is a a questionnaire item comprising a single English adjective (such as "bold" or "timid") asking respondents to indicate how accurately the adjective describes their personality using a 5-point Likert-type rating scale (*very inaccurate, moderately inaccurate, neither accurate nor inaccurate, moderately accurate*, and *very accurate*). Here, each Big-5 personality trait is measured with six pairs of adjectives that are polar opposites to one another (such as "talkative" vs. "silent"), that is, twelve items in total for each trait. It seems implausible that an attentive respondent would choose to agree (or disagree) to *both* items in a pair of polar opposite adjectives. Consequently, one expects a strongly negative correlation between polar adjectives if all respondents respond attentively (Arias et al., 2020). Hence, this dataset is well-suited to test if my robust estimator can correctly identify and account for inattentive respondents, which are presumably those who gave inconsistent responses to polar opposite adjective items.

Arias et al. (2020) collect measurements of three Big-5 traits in this way, namely *extroversion, neuroticism*, and *conscientiousness*.[10] The sample that I shall use, Sample 1 in Arias et al. (2020), consists of $N = 725$ online respondents who are all U.S. citizens, native English speakers, and tend to have relatively high levels of reported education (about 90% report to hold an undergraduate or higher degree). Concerned about respondent inattention in their data, Arias et al. (2020) construct a factor mixture model for detecting inattentive participants. Their model crucially relies on response inconsistencies to polar opposite adjectives and is designed to primarily detect inattentive straightlining responding. They find that inattentive responding is a sizable problem in their data. Their model estimates that the proportion of inattentive participants amounts to 4.7% in the *conscientiousness*, 6% in the *neuroticism*, and 7.3% in the *extroversion* scale. After some further analyses, the authors conclude that if unaccounted for, inattentive responses can substantially deteriorate the fit of theoretical models, produce spurious variance, and overall jeopardize the validity of research results.

[9] The Big-5 factor model is a fundamental model in personality psychology. It assumes that human personality can be described by five latent variables (traits), namely openness, conscientiousness, extraversion, agreeableness, and neuroticism.

[10] Arias et al. (2020) synonymously refer to *neuroticism* as *emotional stability*. Furthermore, in addition to the three listed traits, Arias et al. (2020) collect measurements of the trait *dispositional optimism* by using a different instrument, and another scale that is designed to not measure any construct. I do not consider these scales in this empirical demonstration.

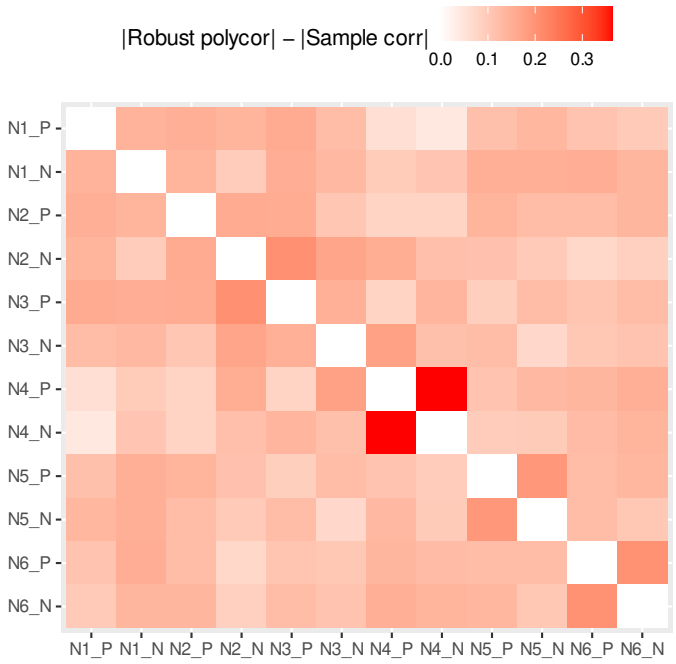


Figure 4: Difference between absolute estimates for the robustly estimated polychoric correlation matrix and the Pearson sample correlation matrix of the *neuroticism* scale, using the data of Arias et al. (2020). The items are "calm" (N1_P), "angry" (N1_N), "relaxed" (N2_P), "tense" (N2_N), "at ease" (N3_P), "nervous" (N3_N), "not envious" (N4_P), "envious" (N4_N), "stable" (N5_P), "unstable" (N5_N), "contented" (N6_P), and "discontented" (N6_N). For the item naming given in parentheses, items with identical identifier (the integer after the first "N") are polar opposites, where a last character "P" refers to the positive opposite and "N" to the negative opposite. The individual estimates of each method are provided in Table C.1 in Appendix C.

Due to the suspected presence of inattentive respodents, I employ my robust estimator with tuning constant choice $c = 1.6$[11] to estimate the polychoric correlation matrix of the twelve items in the *neuroticism* scale. The results of the remaining two scales are qualitatively similar. I then conduct a confirmatory factor analysis (which is a special case of SEM) on this estimated correlation matrix. I repeat this exercise for a non-robustly estimated correlation matrix, namely the Pearson sample correlation matrix, which is the standard way of conducting factor analyses with questionnaire data. Conducting the analysis with a polychoric correlation matrix estimated via MLE yields similar results as with a sample correlation matrix; I therefore omit the former from the reported results.

	Sample cor		MLE		Robust	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
ϱ	-0.562	0.031	-0.618	0.025	-0.925	0.062
a_1			-1.370	0.061	-1.570	0.276
a_2			-0.476	0.043	-0.560	0.203
a_3			0.121	0.042	0.109	0.187
a_4			1.060	0.054	1.080	0.105
b_1			-0.857	0.049	-0.905	0.073
b_2			-0.004	0.041	-0.040	0.091
b_3			0.608	0.045	0.640	0.364
b_4			1.580	0.071	1.171	0.811

Table 2: Parameter estimates with standard errors (SEs) for the correlation between the *neuroticism* adjective pair "envious" and "not envious" in the data of Arias et al. (2020), using sample correlation MLE, and my robust estimator with tuning constant $c = 1.6$. Each adjective item has five ordinal answer categories. The sample correlation coefficient does not model thresholds, hence no estimates for them can be reported.

5.2 Results of correlation matrix estimation

Figure 4 visualizes the absolute differences between the robustly estimated polychoric correlation matrix and sample correlation matrix of the *neuroticism* scale. For all individual unique item pairs, the robust method estimates a stronger correlation coefficient than sample correlation. The differences in absolute estimates on average amount to 0.13, ranging from only marginally larger than zero to a substantive 0.36. For correlations between polar opposite adjectives, the average absolute difference between the robust method and sample correlation is 0.20. The fact that a robust method consistently yields stronger correlation estimates than a non-robust method, particularly between polar opposite adjectives, is indicative of the presence of negative leverage points, which drag negative correlational estimates towards zero, that is, they attenuate the estimated strength of correlation. Here, such negative leverage points could be the responses of inattentive participants who report agreement or disagreement to both items in item pairs that are designed to be negatively correlated. For instance, recall that it is implausible that an attentive respondent would choose to agree (or disagree) to *both* adjectives in the pair "envious" and "not envious" (cf. Arias et al., 2020). If sufficiently many such respondents exist, then the presumably strongly negative correlation between these two opposite adjectives will be estimated to be weaker than it actually is.

To further investigate the presence of inattentive respondents who attenuate correlational estimates, I study in detail the adjective pair "not envious" and "envious", which featured the largest discrepancy between the non-robust and robust estimates in Figure 4, with an

[11] The results remain qualitatively similar for different finite choices of c.

absolute difference of 0.36. The results are summarized in Table 2. The maximum likelihood estimate of -0.618 and sample correlation estimate of -0.562 for the correlation between these two items seem remarkably weak considering that the two adjectives in question are polar opposites. In contrast, its robust estimate estimate is given by -0.925, which seems much more in line with what one would expect if all participants responded accurately and attentively (cf. Arias et al., 2020).

To study the potential presence of inattentive responses in each response cell $\boldsymbol{z} = (x, y) \in \{1, \ldots, 5\}^2$ of the item pair "envious" and "not envious", Figure 5 visualizes the empirical relative frequencies, $\widehat{f}_N(\boldsymbol{z})$, through dot size, as well as the associated Pearson residual at the robust estimate, $\widehat{f}_N(\boldsymbol{z}) \Big/ p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right)$, through dot color (the darker the blue shade, the larger). Importantly, the color of cells whose Pearson residual exceed 5.48 has been fixed to red.[12] This truncation value is equal to the value of the smallest Pearson residual that is substantially larger than the ideal value 1. I consider cells whose Pearson residual exceeds this truncation value to have a poor fit at the polychoric model. This applies to a total of 12 cells, some of which have enormous Pearson residuals. The Pearson residuals of the remaining 13 cells are reasonably close to ideal value 1, ranging from 0.65 to 1.42 with average 0.89. The Pearson residuals as well as relative empirical frequencies of all cells can be found in Table C.2 in the Appendix. It stands out that all poorly fitted cells are those whose responses might be viewed as inconsistent. Indeed, response cells $(x, y) = (1, 1), (1, 2), (2, 1), (1, 2)$ indicate that a participant reports that *neither* "envious" nor "not envious" characterizes them accurately, which are mutually contradicting responses, while for response cells $(x, y) = (4, 4), (4, 5), (5, 4), (5, 5)$ *both* adjectives characterize them accurately, which is again contradicting. As discussed previously, such responses are likely due to inattentiveness. The robust estimator suggests that such responses cannot be fitted well by the polychoric model and subsequently downweighs their influence in the estimation procedure by mapping their Pearson residual with the linear part of the $\rho(\cdot)$ function in (9). Notably, also cells $(x, y) = (1, 3), (3, 1), (3, 5), (5, 3)$ are classified as poorly fitted. These responses report (dis)agreement to one opposite adjective, while being neutral about the other opposite. It is beyond the scope of this paper to assess whether such response patterns are indicative of inattentive responding, but the robust estimator suggests that such responses at least cannot be fitted well by the polychoric model with the data of Arias et al. (2020).

Next, I perform the goodness-of-fit test derived in Corollary 1 for each response cell to assess for which cells the polychoric model achieves a statistically significantly poor fit in the "not envious"–"envious" item pair. Table 3 presents the p-values for the hypothesis test in (11), adjusted for multiple comparisons via the procedure of Benjamini & Hochberg (1995). Values for which the null hypothesis is rejected at significance level $\alpha = 0.001$ are in boldface. This choice of significance level is deliberately extremely conservative because the literature on inattentive responding recommends overwhelming evidence in fa-

[12] The truncation of the color gradient in Figure 5 prevents that the color gradient is dominated by single cells with extreme Pearson residuals, which would blur the distinction between well fitted and poorly fitted cells.

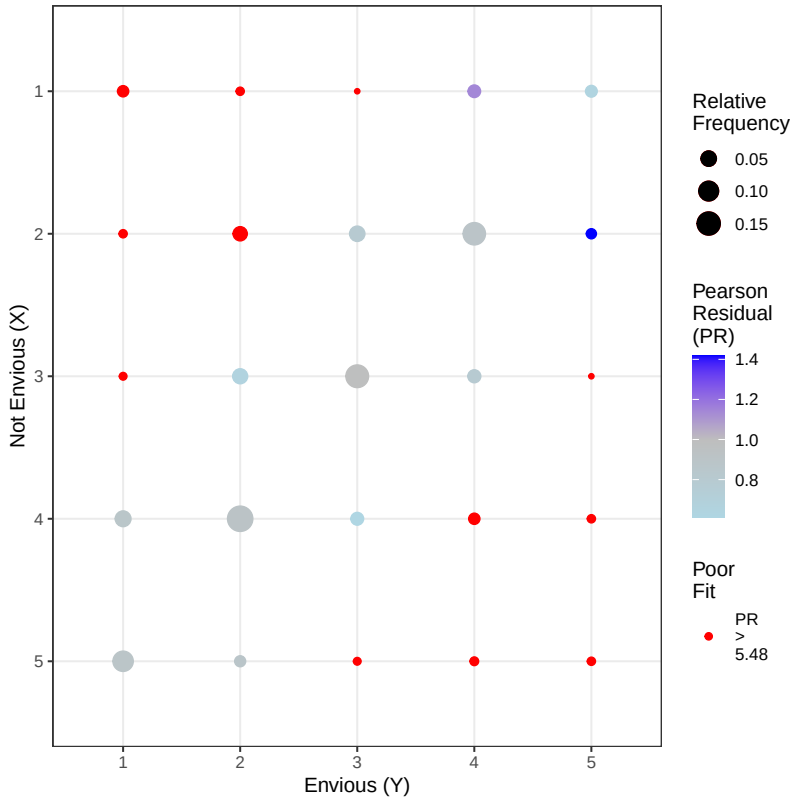


Figure 5: Dot plot of cells for the *neuroticism* item adjective pair "envious" and "not envious" in the data of Arias et al. (2020), where each item has five Likert-type response options, anchored by "very inaccurate" (= 1) and "very accurate" (= 5). Each dot's size is proportional to the relative empirical frequency of its associated cell, $\widehat{f}_N(\boldsymbol{z})$, whereas its color varies by the value of the cell's Pearson residual, $\widehat{f}_N(\boldsymbol{z})/p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right)$, at robust parameter estimate with tuning constant $c = 1.6$: The darker in blue a dot, the larger the value of the Pearson residual of its associated cell. The color of cells that could not be fitted well is fixed to red, where I deem a fit poor if the Pearson residual exceeds value 5.48 (which is the value of the smallest Pearson residual that is substantially larger than ideal value 1). The data visualized here can be found in Table C.2 in Appendix C.

		"Envious"				
	$X \setminus Y$	1	2	3	4	5
	1	$\mathbf{< 0.0001}$	**0.0002**	0.8429	0.8274	0.8429
	2	$\mathbf{< 0.0001}$	0.0033	0.8429	0.9146	0.8274
"Not Envious"	3	0.8274	0.9896	0.8429	0.8429	0.8429
	4	0.8429	0.9079	0.8863	0.5251	**0.0003**
	5	0.8429	0.8429	0.5251	$\mathbf{< 0.0001}$	$\mathbf{< 0.0001}$

Table 3: p-values, adjusted for multiple comparisons by the procedure of Benjamini & Hochberg (1995), of the cellwise goodness-of-fit test in Corollary 1 for the *neuroticism* adjective pair "envious" and "not envious" in the data of Arias et al. (2020), where each item has five Likert-type response options, anchored by "1 = very inaccurate" and "5 = very accurate". The test statistics were computed with robust estimates using tuning constant $c = 1.6$. Cells in boldface are those for which the null hypothesis of unit Pearson residual is rejected at significance level $\alpha = 0.001$ in favor of the alternative of it being larger than one.

vor of inattention before one should label responses as such (cf. Huang et al., 2012). At this significance level, the null hypothesis of a good fit is rejected for six cells, namely $(x, y) = (1, 1), (2, 1), (1, 2), (5, 4), (4, 5), (5, 5)$. These six cells comprise 5.52% of the entire sample. As discussed in the previous section, it seems likely that these responses are due to inattention because of inconsistent and contradictory responding. Either way, my test offers strong empirical evidence that these cells are outlying in the sense that they cannot be fitted well by the polychoric model and therefore lead to deteriorated model fit. This is consistent with Arias et al. (2020), who find that even a relatively small proportion of inconsistent responses can drastically reduce a model's fit. In their analyses, they estimate that 6% of all respondents in the *neuroticism* scale have been inattentive. Yet, albeit similar, I emphasize that the estimate of 5.52% can be, if at all, understood as a lower bound for the proportion of inattentive responding because of the extremely conservative significance level I chose for my analyses. For instance, the null hypothesis of good model fit was *not* rejected for the seemingly inconsistent response cell $(x, y) = (2, 2)$ (relative empirical frequency of about 4%) with a p-value of approximately 0.003, but would have been rejected at a slightly more liberal significance level. In addition, it is worth noting that the null hypothesis was also not rejected for one more seemingly inconsistent response cell, namely $(x, y) = (4, 4)$, despite a relatively large Pearson residual of 12.66. This non-rejection is likely due to low statistical power stemming from a small empirical frequency of this cell, since it only counted 14 responses (out of 725). Similar reasoning applies to the remaining four cells that were highlighted in red in Figure 4 but for which the null hypothesis of good fit was not rejected, namely those who indicate (dis)agreement to one adjective, while being neutral about its opposite. These four cells, $(x, y) = (1, 3), (3, 1), (3, 5), (5, 3)$, only count empirical frequencies of 2, 4, 2, and 4, respectively.

Overall, leveraging the proposed robust estimator, I find strong evidence for the presence of inattentive respondents in the data of Arias et al. (2020). While they substantially affect

Item	Sample corr	Robust polycor
N1_P	0.70	0.80
N1_N	0.56	0.66
N2_P	0.76	0.86
N2_N	0.68	0.78
N3_P	0.77	0.88
N3_N	0.66	0.74
N4_P	0.35	0.46
N4_N	0.46	0.54
N5_P	0.69	0.77
N5_N	0.67	0.73
N6_P	0.57	0.66
N6_N	0.64	0.71
Proportion variance	0.40	0.53
Cronbach's α	0.89	0.93

Table 4: Factor loadings estimates in the unidimensional confirmatory factor analysis of the *neuroticism* scale, using the data of Arias et al. (2020).

the correlational estimate of Pearson sample correlation, amounting to about -0.56, which is much weaker than one would expect for polar opposite items, my robust estimator can withstand their influence with an estimate of about -0.93 and also identify them by means of the proposed test. Similar conclusions follow by repeating this analysis for different item pairs.

5.3 Results of structural equation modeling

In this section, I perform confirmatory factor analyses on the data of Arias et al. (2020) in the *neuroticism* scale. That is, I fit a factor model as in (2) with a single factor to each of the two estimated correlation matrices. Table 4 presents the estimates factor loadings for each estimator. The estimated loadings of the robust estimator are consistently higher than those of the non-robust sample correlation estimator, indicating greater internal consistency. Indeed, the *neuroticism* factor of the robust analysis can explain 53% of the variation in the data, whereas the factor of the non-robust analysis can only explain 40%. In addition, the robust analysis yields higher reliability than the non-robust analysis, as can be seen by Cronbach-α estimates of 0.89 and 0.93, respectively.

The fact that a robust estimator yields a more internally consistent factor structure than a non-robust estimator is indicative of the presence of irregular data points, namely the inconsistent (and presumably inattentive) responses the robust estimator identified in the previous section and whose adverse influence it can withstand by down-weighting them in the estimation procedure.

This empirical application demonstrated how the proposed estimator can be used to robustify SEMs against inattentive respondents. I find compelling evidence for the presence

of inattentive responding, and the proposed robust estimator can not only identify them, but also account for their presence to obtain a good model fit.

6 Conclusion

In this paper, I develop a novel estimator for models of categorical variables that is designed to be robust to misspecification of such models, which is the first of its kind and can be thought of as an analogue to robust M-estimation for non-continuous variables. The estimator is shown to be consistent and asymptotically normally distributed, and possesses attractive properties with respect to robustness and computation. Crucially, the estimator makes no assumption whatsoever on the degree, magnitude, or type of misspecification. If misspecification is absent, the estimator is asymptotically equivalent to maximum likelihood estimation (MLE), but more robust than MLE in the presence of misspecification. In addition, I develop a novel diagnostic test that can test if a given categorical observation can be fitted well by the presumed model, allowing one to trace back potential sources of model misspecification. The methodology proposed in this paper is implemented in the free open source package `robcat` (Welz, 2024) in the statistical programming environment R, although it is primarily developed in C++ to maximize speed and computational performance.

I verify the enhanced robustness and theoretical properties of the novel estimator in simulation studies and demonstrate its practical usefulness in an empirical application on structural equation modeling of questionnaire responses to a Big-5 administration. I find compelling evidence for the presence of inattentive respondents. For instance, in a rating item pair with polar opposite content where a strong negative correlation is expected, the robust estimator yields a correlational estimate of -0.93, whereas non-robust estimators yield only -0.56 to -0.62; it follows that the robust estimate is more in line with the literature on the corresponding scale. Utilizing the proposed diagnostic test, I argue that the lower-than-expected estimates of the non-robust method are likely due to a few possibly inattentive participants who gave mutually contradictory responses, while the robust estimator can resist their influence. In addition, conducting a confirmatory factor analyses with the robust estimator yielded more core consistent factor loadings, higher explained variance, and greater scale reliability than commonly employed estimation methods.

Although the proposed estimator is particularly useful in models for questionnaire responses, I stress that it can be applied to any model of categorical responses. Examples include models from item response theory, counting processes, or categorical regression, thereby potentially giving rise to a new research line. I leave these exciting avenues to further research.

References

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). John Wiley & Sons.

Agresti, A. (2012). *Categorical data analysis* (3rd ed.). John Wiley & Sons.

Alfons, A. & Welz, M. (2024). Open science perspectives on machine learning for the identification of careless responding: A new hope or phantom menace? *Social and Personality Psychology Compass*. https://doi.org/10.1111/spc3.12941. In press.

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. (2011). Personality psychology and economics. *Handbook of The Economics of Education*, volume 4 of *Handbook of the Economics of Education*, 1–181. Elsevier. https://doi.org/10.1016/B978-0-444-53444-6.00001-8

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1), 42–54. https://doi.org/10.1111/j.2517-6161.1972.tb00887.x

Andrews, I., Gentzkow, M., & Shapiro, J. M. (2017). Measuring the Sensitivity of Parameter Estimates to Estimation Moments*. *Quarterly Journal of Economics*, 132(4), 1553–1592. https://doi.org/10.1093/qje/qjx023

Arias, V. B., Garrido, L., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, 52(6), 2489–2505. https://doi.org/https://doi.org/10.3758/s13428-020-01401-8

Armstrong, T. B. & Kolesár, M. (2021). Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1), 77–108. https://doi.org/https://doi.org/10.3982/QE1609

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. https://doi.org/https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bond, T. N. & Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy*, 127(4), 1629–1640. https://doi.org/10.1086/701679

Bonhomme, S. & Weidner, M. (2022). Minimizing sensitivity to model misspecification. *Quantitative Economics*, 13(3), 907–954. https://doi.org/https://doi.org/10.3982/QE1930

Borghans, L., Duckworth, A. L., Heckman, J., & ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, 43(4), 972–1059. https://doi.org/10.3368/jhr.43.4.972

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, 111(2), 218–229. https://psycnet.apa.org/doi/10.1037/pspp0000085

Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2023). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*, 26(2), 323–352. https://doi.org/10.1177/10944281211056520

Bugni, F. A. & Ura, T. (2019). Inference in dynamic discrete choice problems under local misspecification. *Quantitative Economics*, 10(1), 67–103. https://doi.org/https://doi.org/10.3982/QE917

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., & Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, 90(4), 1501–1535. https://doi.org/https://doi.org/10.3982/ECTA16294

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, 70(4), 596–612. https://doi.org/10.1177/0013164410366686

Cressie, N. & Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3), 440–464.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. https://doi.org/10.1007/BF02310555

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Daley, D. J. & Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes* (2nd ed.). Probability and Its Applications. Springer. https://doi.org/10.1007/b97277

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer. https://doi.org/10.1007/978-1-4614-6868-4. ISBN 978-1-4614-6867-7

Foldnes, N. & Grønneberg, S. (2022). The sensitivity of structural equation modeling with ordinal data to underlying non-normality and observed distributional forms. *Psychological Methods*, 27(4), 541–567. https://doi.org/10.1037/met0000385

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for likert-type and ordinal item response data: A conceptual, empirical, and practical guid. *Practical Assessment, Research, and Evaluation*, 17(1), 1–13. https://doi.org/10.7275/n560-j767

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, 18(4), 454–474. https://doi.org/10.1037/a0030005

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1), 26–42. https://doi.org/https://doi.org/10.1037/1040-3590.4.1.26

Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Annals of Probability*, 19(2), 724–739. https://doi.org/10.1214/aop/1176990448

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346), 383–393. https://doi.org/10.1080/01621459.1974.10482962

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. Wiley series in probability and mathematical statistics. Wiley.

He, X. & Simpson, D. G. (1993). Lower bounds for contamination bias: Globally minimax versus locally linear estimation. *Annals of Statistics*, 21(1), 314–337. https://doi.org/10.1214/aos/1176349028

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3), 411–482. https://doi.org/10.1086/504455

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, 44, 153–166. https://doi.org/10.1007/s11135-008-9190-y

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015a). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, 30(2), 299–311. https://doi.org/10.1007/s10869-014-9357-6

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015b). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, 100(3), 828–845. https://doi.org/10.1037/a0038510

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Huber, P. J. & Ronchetti, E. M. (2009). *Robust Statistics*. Wiley. https://doi.org/10.1002/9780470434697

Kam, C. C. S. & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, 18(3), 512–541. https://doi.org/10.1177/1094428115571894

Kim, D. S., Reise, S. P., & Bentler, P. M. (2018). Identifying aberrant data in structural equation models with IRLS-ADF. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3), 343–358. https://doi.org/10.1080/10705511.2017.1379881

Kitamura, Y., Otsu, T., & Evdokimov, K. (2013). Robustness, infinitesimal neighborhoods, and moment restrictions. *Econometrica*, 81(3), 1185–1201. https://doi.org/https://doi.org/10.3982/ECTA8617

Lai, K. & Green, S. B. (2016). The problem with having two watches: Assessment of fit when rmsea and cfi disagree. *Multivariate Behavioral Research*, 51(2-3), 220–239. https://doi.org/10.1080/00273171.2015.1134306

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics*, 22(2), 1081–1114. https://doi.org/10.1214/aos/1176325512

Mair, P. (2018). *Modern Psychometrics with R*. Springer. https://doi.org/10.1007/978-3-319-93177-7

Maniaci, M. R. & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis* (1st ed.). Series in Probability and Mathematical Statistics. Academic Press.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2018). *Robust Statistics: Theory and Methods* (2nd ed.). John Wiley & Sons.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23(1), 1–21. https://doi.org/10.1111/j.2044-8317.1970.tb00432.x

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, 136(3), 450–470. https://doi.org/10.1037/a0019216

Meade, A. W. & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455. https://psycnet.apa.org/doi/10.1037/a0028085

Newey, W. K. (1985). Generalized method of moments specification testing. *Journal of Econometrics*, 29(3), 229–256. https://doi.org/https://doi.org/10.1016/0304-4076(85)90154-X

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59(4), 1161–1167. https://doi.org/10.2307/2938179

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1), 1–32. https://doi.org/10.2307/1914288

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4), 443–460. https://doi.org/10.1007/BF02296207

Osborne-Groves, M. (2005). How important is your personality? Labor market returns to personality for women in the US and UK. *Journal of Economic Psychology*, 26(6), 827–841. https://doi.org/10.1016/j.joep.2005.03.001

Pearson, K. (1901). I. mathematical contributions to the theory of evolution, vii: On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, Series A*, 195(262-273), 1–47. https://doi.org/10.1098/rsta.1900.0022

Pearson, K. & Pearson, E. S. (1922). On polychoric coefficients of correlation. *Biometrika*, 14(1–2), 127–156. https://doi.org/10.1093/biomet/14.1-2.127

Pison, G., Rousseeuw, P. J., Filzmoser, P., & Croux, C. (2003). Robust factor analysis. *Journal of Multivariate Analysis*, 84(1), 145–172. https://doi.org/10.1016/S0047-259X(02)00007-6

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.*

Raymaekers, J. & Rousseeuw, P. J. (2021). Fast robust correlation for high-dimensional data. *Technometrics*, 63(2), 184–198. https://doi.org/10.1080/00401706.2019.1677270

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? a comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. https://doi.org/10.1037/a0029315

Robinson, M. A. (2018). Using multi-item psychometric scales for research and practice in human resource management. *Human Resource Management*, 57(3), 739–750. https://doi.org/10.1002/hrm.21852

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, 283–297. Reidel.

Ruckstuhl, A. F. & Welsh, A. H. (2001). Robust fitting of the binomial model. *Annals of Statistics*, 29(4), 1117–1136. https://doi.org/10.1214/aos/1013699996

Rudin, W. (1976). *Principles of Mathematical Analysis* (3rd ed.). McGraw-Hill.

Rust, J., Golombok, S., & Stillwell, D. (2020). *Modern psychometrics: The science of psychological assessment* (4th ed.). Routledge.

Schennach, S. (2022). Measurement systems. *Journal of Economic Literature*, 60(4), 1223–1263. https://doi.org/10.1257/jel.20211355

Schmitt, N. & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, 9, 367–373. https://doi.org/10.1177/014662168500900405

Schroeders, U., Schmidt, C., & Gnambs, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, 82(1), 29–56. https://doi.org/10.1177/00131644211004708

Simpson, D. G. (1987). Minimum hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association*, 82(399), 802–807. https://doi.org/10.1080/01621459.1987.10478501

Stantcheva, S. (2022). How to run surveys: A guide to creating your own identifying variation and revealing the invisible. Working Paper 30527, National Bureau of Economic Research. https://doi.org/10.3386/w30527

Van de Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press. https://doi.org/10.1017/9781108231596

Victoria-Feser, M.-P. & Ronchetti, E. (1997). Robust estimation for grouped data. *Journal of the American Statistical Association*, 92(437), 333–340. https://doi.org/10.1080/01621459.1997.10473631

Ward, M. & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, 74(1), 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Welz, M. (2024). ***robcat**: Robust Categorical Data Analysis*. https://github.com/mwelz/robord. R package version 0.0.1

Welz, M. & Alfons, A. (2023). *I don't care anymore: Identifying the onset of careless responding*. https://doi.org/10.48550/arXiv.2303.07167. arXiv:2303.07167

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28(3), 186–191. https://psycnet.apa.org/doi/10.1007/s10862-005-9004-7

A Popular categorical models

In addition SEM presented in Section 2.2, this section describes two more commonly used models for categorical models, namely the Rasch model from item response theory and a Poisson counting process. I stress that these three models (SEM, Rasch, Poisson) are designed for three different types of categorical variables, namely ordinal ones, nominal ones, and counting variables. The proposed robust estimator can be applied to fit all three models.

Example 1 (Rasch model)**.** The Rasch model (Rasch, 1960) is a fundamental model to model a test taker's probability to answer correctly to a test's questions. Let the test comprise k questions and let $\boldsymbol{X} = (X_1, \dots, X_k)^\top$ be a binary vector whose j-th element equals 1 if the j-th question was answered correctly, and zero otherwise. It follows that the sample space is given by $\boldsymbol{\mathcal{X}} = \{0, 1\}^k$. Further, denote by $S = \sum_{j=1}^k X_j$ the number of correctly answered questions, known as the *score*. Since the unconditional Rasch model suffers from the incidental parameter problem,[i] one instead (Andersen, 1972) works with the conditional probability for $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ given a score $s \in \{0, \dots, k\}$, being

$$p_{\boldsymbol{z}}(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{X} = \boldsymbol{x} \mid S = s] = \frac{\exp(-\boldsymbol{x}^\top \boldsymbol{\theta})}{\sum_{\boldsymbol{y} \in \boldsymbol{\mathcal{D}}(s)} \exp(-\boldsymbol{y}^\top \boldsymbol{\theta})},$$

where $\boldsymbol{z} = (\boldsymbol{x}^\top, s)^\top$, $\boldsymbol{\theta} \in \mathbb{R}^k$ is a k-vector of question difficulties, and $\boldsymbol{\mathcal{D}}(s) = \left\{\boldsymbol{x} \in \boldsymbol{\mathcal{X}} : \sum_{j=1}^k x_j = s\right\}$ denotes the set of response vector with score s.

Example 2 (Poisson process)**.** Counting processes model how often a certain event occurs in a given time period, which is a discrete outcome. Assume that one has access to counting observations in k periods, where the j-th period is given by $(a_j, b_j]$ with known finite boundaries $a_j < b_j \leq a_{j+1}$, and random variable $Z_j \in \mathbb{N}_0 = \{0, 1, 2, \dots\}$ counts the number of event in the j-th period, $j = 1, \dots, k$. Consequently, the k-dimensional random variable $\boldsymbol{Z} = (Z_1, \dots, Z_k)^\top$ holds the counts per period and takes values in sample space $\boldsymbol{\mathcal{Z}} = \mathbb{N}_0^k$.[ii] One of the most popular counting processes is the *stationary Poisson point process* (e.g. Daley & Vere-Jones, 2003, equation 2.1.1), which defines the probability of

[i] The unconditional Rasch model is defined by the individual-specific probability

$$\mathbb{P}[\boldsymbol{X}_i = \boldsymbol{x}] = \frac{\exp\left(\sum_{j=1}^k x_j(\alpha_i - \theta_j)\right)}{\prod_{j=1}^k (1 + \exp(\alpha_i - \theta_j))},$$

where θ_j parametrizes the difficulty of question j, whereas α_i parametrizes the ability of test taker $i = 1, \dots, N$, through fixed effects. Joint maximum likelihood estimates of the ability and difficulty parameters will not be consistent since the number of parameters grows with the sample size N, which is an instance of the incidental parameter problem (Neyman & Scott, 1948).

[ii] In practice, the empirical support is bounded, with the largest observed number of counts being the upper bound.

$\boldsymbol{z} = (z_1, \ldots, z_k)^\top \in \boldsymbol{\mathcal{Z}}$ as

$$p_{\boldsymbol{z}}(\lambda) = \mathbb{P}_\lambda[\boldsymbol{Z} = \boldsymbol{z}] = \prod_{j=1}^{k} \frac{\big(\lambda(b_j - a_j)\big)^{z_j}}{z_j!} \exp\big(-\lambda(b_j - a_j)\big),$$

where $\lambda > 0$ is an *intensity* parameter.

B Additional theoretical results

The *influence function* (Hampel, 1974) measures how an infinitesimal amount of contamination from a point mass distribution affects an estimator. Essential theory on M-estimation (e.g. Huber & Ronchetti, 2009, Section 3.2) yields that the MLE (eq. 5), reveals the following influence function at point $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ at density $\boldsymbol{p}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}$:

$$\text{IF}\left(\boldsymbol{z}, \widehat{\boldsymbol{\theta}}_N^{\text{MLE}}, \boldsymbol{p}(\boldsymbol{\theta})\right) = \boldsymbol{J}(\boldsymbol{\theta})^{-1} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}),$$

where

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{\theta}) p_{\boldsymbol{z}}(\boldsymbol{\theta})$$

denotes the Fisher information matrix at $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The following proposition derives the influence function of the proposed estimator $\widehat{\boldsymbol{\theta}}_N$ in (8).

Proposition B.1 (Influence function)**.** *Grant Assumption Set A. Then, the influence function of estimator $\widehat{\boldsymbol{\theta}}_N$ at cell $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ and density $\boldsymbol{p}(\boldsymbol{\theta})$ is given by*

$$\text{IF}\left(\boldsymbol{z}, \widehat{\boldsymbol{\theta}}_N, \boldsymbol{p}(\boldsymbol{\theta})\right) = \begin{cases} \text{IF}\left(x, \widehat{\boldsymbol{\theta}}_N^{\text{MLE}}, \boldsymbol{p}(\boldsymbol{\theta})\right) & \text{if } c \neq 1, \\ \left[\boldsymbol{J}(\boldsymbol{\theta}) - p_{\boldsymbol{z}}(\boldsymbol{\theta}) \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}) \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})^\top\right]^{-1} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}) p_{\boldsymbol{z}}(\boldsymbol{\theta}) & \text{if } c = 1. \end{cases}$$

I will add the proof of this proposition in a future version of this working paper.

This result is remarkable because it suggests that there exist robust estimators (governed by the choice of c) that have the same influence function as the efficient but non-robust MLE, which is due to the bounded nature of categorical variables. Similar results have been derived for minimum power divergence estimators (Victoria-Feser & Ronchetti, 1997), minimum Hellinger distance estimators (Simpson, 1987; He & Simpson, 1993; Lindsay, 1994), and minimum disparity estimators at the binomial model (Ruckstuhl & Welsh, 2001).

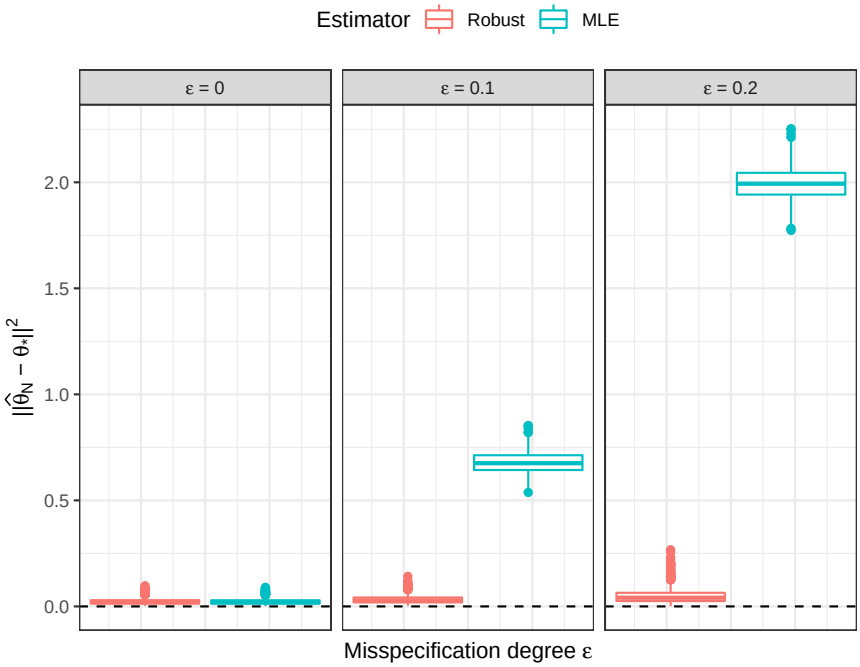


Figure C.1: Mean squared error of all estimated parameters in the polychoric model, $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_*\|^2$, in the first simulation in Section 4.

C Additional empirical results

C.1 Additional simulation results

Figure C.1 visualizes the mean squared error of all estimated parameters in the polychoric model in the first simulation in Section 4. I do not plot the results of the Pearson sample correlation coefficient because it does not estimate threshold parameters. The results for that coefficient are visualized in Figure 2 in the main text.

C.2 Additional results from the empirical application

Table C.1 lists the three estimated correlation matrices in the *neuroticism* scale in the data of Arias et al. (2020), where negatively worded items were reverse-coded, which is standard in factor analyses on questionnaire data. Table C.2 contains the data that are visualized in Figure 5.

	N1_P	N1_N	N2_P	N2_N	N3_P	N3_N	N4_P	N4_N	N5_P	N5_N	N6_P	N6_N
N1_P	1.00	-0.32	0.65	-0.44	0.63	-0.44	0.24	-0.22	0.51	-0.41	0.37	-0.29
N1_N	-0.32	1.00	-0.34	0.48	-0.34	0.40	-0.16	0.34	-0.32	0.53	-0.28	0.49
N2_P	0.65	-0.34	1.00	-0.50	0.70	-0.50	0.24	-0.24	0.50	-0.38	0.48	-0.42
N2_N	-0.44	0.48	-0.50	1.00	-0.49	0.59	-0.22	0.37	-0.37	0.50	-0.28	0.46
N3_P	0.63	-0.34	0.70	-0.49	1.00	-0.47	0.26	-0.25	0.57	-0.39	0.48	-0.44
N3_N	-0.44	0.40	-0.50	0.59	-0.47	1.00	-0.25	0.37	-0.39	0.50	-0.27	0.41
N4_P	0.24	-0.16	0.24	-0.22	0.26	-0.25	1.00	-0.56	0.24	-0.17	0.17	-0.18
N4_N	-0.22	0.34	-0.24	0.37	-0.25	0.37	-0.56	1.00	-0.30	0.40	-0.20	0.39
N5_P	0.51	-0.32	0.50	-0.37	0.57	-0.39	0.24	-0.30	1.00	-0.62	0.47	-0.41
N5_N	-0.41	0.53	-0.38	0.50	-0.39	0.50	-0.17	0.40	-0.62	1.00	-0.32	0.50
N6_P	0.37	-0.28	0.48	-0.28	0.48	-0.27	0.17	-0.20	0.47	-0.32	1.00	-0.54
N6_N	-0.29	0.49	-0.42	0.46	-0.44	0.41	-0.18	0.39	-0.41	0.50	-0.54	1.00

(a) Pearson sample correlation

	N1_P	N1_N	N2_P	N2_N	N3_P	N3_N	N4_P	N4_N	N5_P	N5_N	N6_P	N6_N
N1_P	1.00	-0.37	0.71	-0.50	0.69	-0.49	0.27	-0.24	0.58	-0.47	0.42	-0.32
N1_N	-0.37	1.00	-0.40	0.55	-0.39	0.47	-0.19	0.40	-0.39	0.60	-0.32	0.56
N2_P	0.71	-0.40	1.00	-0.55	0.75	-0.54	0.26	-0.26	0.55	-0.41	0.53	-0.47
N2_N	-0.50	0.55	-0.55	1.00	-0.54	0.65	-0.24	0.42	-0.41	0.57	-0.31	0.52
N3_P	0.69	-0.39	0.75	-0.54	1.00	-0.53	0.29	-0.28	0.63	-0.44	0.52	-0.48
N3_N	-0.49	0.47	-0.54	0.65	-0.53	1.00	-0.28	0.43	-0.44	0.58	-0.29	0.47
N4_P	0.27	-0.19	0.26	-0.24	0.29	-0.28	1.00	-0.61	0.26	-0.20	0.18	-0.20
N4_N	-0.24	0.40	-0.26	0.42	-0.28	0.43	-0.61	1.00	-0.33	0.46	-0.22	0.44
N5_P	0.58	-0.39	0.55	-0.41	0.63	-0.44	0.26	-0.33	1.00	-0.69	0.53	-0.46
N5_N	-0.47	0.60	-0.41	0.57	-0.44	0.58	-0.20	0.46	-0.69	1.00	-0.35	0.57
N6_P	0.42	-0.32	0.53	-0.31	0.52	-0.29	0.18	-0.22	0.53	-0.35	1.00	-0.58
N6_N	-0.32	0.56	-0.47	0.52	-0.48	0.47	-0.20	0.44	-0.46	0.57	-0.58	1.00

(b) Maximum likelihood estimates

	N1_P	N1_N	N2_P	N2_N	N3_P	N3_N	N4_P	N4_N	N5_P	N5_N	N6_P	N6_N
N1_P	1.00	-0.47	0.80	-0.58	0.79	-0.56	0.30	-0.26	0.63	-0.54	0.49	-0.39
N1_N	-0.47	1.00	-0.48	0.58	-0.49	0.54	-0.26	0.45	-0.47	0.68	-0.43	0.63
N2_P	0.80	-0.48	1.00	-0.66	0.85	-0.60	0.32	-0.32	0.64	-0.50	0.60	-0.56
N2_N	-0.58	0.58	-0.66	1.00	-0.70	0.76	-0.37	0.49	-0.48	0.60	-0.35	0.55
N3_P	0.79	-0.49	0.85	-0.70	1.00	-0.62	0.35	-0.39	0.66	-0.52	0.59	-0.57
N3_N	-0.56	0.54	-0.60	0.76	-0.62	1.00	-0.42	0.49	-0.52	0.58	-0.37	0.53
N4_P	0.30	-0.26	0.32	-0.37	0.35	-0.42	1.00	-0.92	0.35	-0.30	0.30	-0.33
N4_N	-0.26	0.45	-0.32	0.49	-0.39	0.49	-0.92	1.00	-0.39	0.50	-0.33	0.53
N5_P	0.63	-0.47	0.64	-0.48	0.66	-0.52	0.35	-0.39	1.00	-0.82	0.59	-0.55
N5_N	-0.54	0.68	-0.50	0.60	-0.52	0.58	-0.30	0.50	-0.82	1.00	-0.44	0.61
N6_P	0.49	-0.43	0.60	-0.35	0.59	-0.37	0.30	-0.33	0.59	-0.44	1.00	-0.75
N6_N	-0.39	0.63	-0.56	0.55	-0.57	0.53	-0.33	0.53	-0.55	0.61	-0.75	1.00

(c) Robust estimates

Table C.1: Estimated correlation matrices of the items in the *neuroticism* scale from the data in Arias et al. (2020, Sample 1; $N = 725$) using three estimators. The items are "calm" (N1_P), "angry" (N1_N), "relaxed" (N2_P), "tense" (N2_N), "at ease" (N3_P), "nervous" (N3_N), "not envious" (N4_P), "envious" (N4_N), "stable" (N5_P), "unstable" (N5_N), "contented" (N6_P), and "discontented" (N6_N). For the item naming given in parentheses, items with identical identifier (the integer after the first "N") are polar opposites, where a last character "P" refers to the positive opposite and "N" to the negative opposite.

$X \backslash Y$	1	2	3	4	5
1	9,814,457,557.73	16,011.33	11.82	1.14	0.65
2	2,424.07	10.07	0.80	0.90	1.42
3	15.48	0.65	0.99	0.80	77.14
4	0.88	0.92	0.61	12.66	222,528.08
5	0.89	0.88	36.01	55,420.33	995,017,243,197.60

(a) Pearson residuals $\widehat{f}_N(\boldsymbol{z})/p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right)$

$X \backslash Y$	1	2	3	4	5
1	0.019	0.007	0.003	0.028	0.022
2	0.007	0.040	0.050	0.138	0.014
3	0.006	0.047	0.143	0.030	0.003
4	0.054	0.189	0.029	0.019	0.007
5	0.108	0.018	0.006	0.008	0.007

(b) Empirical relative frequencies $\widehat{f}_N(\boldsymbol{z})$

$X \backslash Y$	1	2	3	4	5
1	< 0.001	< 0.001	< 0.001	0.024	0.034
2	< 0.001	0.004	0.062	0.153	0.010
3	0.001	0.072	0.145	0.038	< 0.001
4	0.061	0.205	0.047	0.002	< 0.001
5	0.120	0.020	< 0.001	< 0.001	< 0.001

(c) Estimated cell probabilities $p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right)$

Table C.2: Pearson residual (top), empirical relative frequency (center), and estimated cell probability (bottom) of each cell for the "not envious" (X)–"envious" (Y) item pair in the measurements of Arias et al. (2020) of the *neuroticism* scale. Estimate $\widehat{\boldsymbol{\theta}}_N$ was computed with tuning constant $c = 1.6$.

D Proofs

The following lemmas shall be useful in proving the mathematical statements in this paper.

D.1 Lemmas

Lemma D.1. *Under the Assumptions of Theorem 1, the sequence* $\left\{L\left(\boldsymbol{\theta}, \widehat{f}_N\right)\right\}_N$ *is equicontinuous on the parameter space* $\boldsymbol{\Theta} \ni \boldsymbol{\theta}$.

Proof. I verify equicontinuity by its definition (e.g. Definition 7.22 in Rudin, 1976). Put $L_N(\boldsymbol{\theta}) = L\left(\boldsymbol{\theta}, \widehat{f}_N\right)$ and let

$$\boldsymbol{g}_N(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} L_N(\boldsymbol{\theta}) = -\sum_{k=1}^{K} \boldsymbol{s}_k(\boldsymbol{\theta}) f_\varepsilon(\boldsymbol{z}) w\left(\frac{\widehat{f}_N(k)}{p_{\boldsymbol{z}}(\boldsymbol{\theta})}\right)$$

denote its gradient at $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. By Assumption A.5, there exists a universal constant $C > 0$ such that $\|\boldsymbol{g}_N(\boldsymbol{\theta})\| < C$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Fix $\varepsilon > 0$ and put $\delta = \varepsilon / C$. Let $\boldsymbol{\theta}_a, \boldsymbol{\theta}_b \in \boldsymbol{\Theta}$ be vectors such that $\|\boldsymbol{\theta}_a - \boldsymbol{\theta}_b\| < \delta$ and the entire line segment between these two vectors is contained in $\boldsymbol{\Theta}$. Then, by the multivariate mean value theorem, there exists a $\boldsymbol{\theta}_c$ in this line segment such that

$$L_N(\boldsymbol{\theta}_b) - L_N(\boldsymbol{\theta}_a) = \boldsymbol{g}_N(\boldsymbol{\theta})^\top (\boldsymbol{\theta}_b - \boldsymbol{\theta}_a).$$

Thus,

$$\begin{aligned} |L_N(\boldsymbol{\theta}_b) - L_N(\boldsymbol{\theta}_a)| &= |\boldsymbol{g}_N(\boldsymbol{\theta})^\top (\boldsymbol{\theta}_b - \boldsymbol{\theta}_a)| \\ &\leq \|\boldsymbol{g}_N(\boldsymbol{\theta})\| \|\boldsymbol{\theta}_b - \boldsymbol{\theta}_a\| \\ &\leq C \|\boldsymbol{\theta}_b - \boldsymbol{\theta}_a\| \\ &< C\delta \\ &= \varepsilon, \end{aligned}$$

where I have applied the Cauchy-Schwarz inequality in the second line. It follows from the definition of equicontinuity that the sequence $\{L_N(\boldsymbol{\theta})\}_N$ is equicontinuous on $\boldsymbol{\Theta}$. ■

The following lemma can be seen as a multivariate special case of the classic Berry-Esseen theorem (e.g. Theorem 2.1.3 in Vershynin, 2018).

Lemma D.2. *For* $i = 1, \ldots, N$, *put* $\boldsymbol{J}_i = \left(J_{i,\boldsymbol{z}_1}, J_{i,\boldsymbol{z}_2}, \ldots, J_{i,\boldsymbol{z}_m}\right)^\top = \left(\mathbb{1}\{\boldsymbol{Z}_i = \boldsymbol{z}_1\}, \mathbb{1}\{\boldsymbol{Z}_i = \boldsymbol{z}_2\}, \ldots, \mathbb{1}\{\boldsymbol{Z}_i = \boldsymbol{z}_m\}\right)^\top$ *and* $\boldsymbol{X}_i = -\boldsymbol{U}(\boldsymbol{\theta}_0)^{-1/2} \boldsymbol{W}(\boldsymbol{\theta}_N(\boldsymbol{t}))(\boldsymbol{J}_i - \boldsymbol{f}_\varepsilon)$, *where* $\boldsymbol{\theta}_N(\boldsymbol{t}) = \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{V}\boldsymbol{t}$ *for a fixed vector* $\boldsymbol{t} \in \mathbb{R}^d$ *and symmetric invertible matrix* $\boldsymbol{V} \in \mathbb{R}^{d \times d}$. *Then, under the assumptions of Theorem 2, the moment* $\mathbb{E}\|\boldsymbol{X}_i\|^3$ *is finite, and, in addition, there exists a constant* $C_d > 0$ *only depending on* d *such that*

$$\sup_{\boldsymbol{x} \in \mathbb{R}^d} \left| \Phi_d(\boldsymbol{x}) - \mathbb{P}\left[N^{-1/2} \sum_{i=1}^{N} \boldsymbol{X}_i \leq \boldsymbol{x}\right] \right| < C_d \mathbb{E}\|\boldsymbol{X}_i\|^3 / \sqrt{N} + o\left(N^{-1/2}\right).$$

Before I turn to its proof, note that this lemma implies that

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d}\left|\Phi_d(\boldsymbol{x})-\mathbb{P}\left[N^{-1/2}\sum_{i=1}^{N}\boldsymbol{X}_i\leq\boldsymbol{x}\right]\right|=o\left(N^{-1/2}\right)+o\left(N^{-1/2}\right).$$

Proof. For the first assertion, note that $J_{i,\boldsymbol{z}}=\mathbb{1}\{\boldsymbol{Z}_i=\boldsymbol{z}\}$ follows a Bernoulli distribution with probability parameter $f_\varepsilon(\boldsymbol{z})$ for any $\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}$. An easy calculation reveals that

$$\mathbb{E}\big|J_{i,\boldsymbol{z}}-f_\varepsilon(\boldsymbol{z})\big|^3=f_\varepsilon(\boldsymbol{z})(1-f_\varepsilon(\boldsymbol{z}))\big((1-f_\varepsilon(\boldsymbol{z}))^2+f_\varepsilon(\boldsymbol{z})^2\big). \tag{D.1}$$

Furthermore, I have that

$$\begin{aligned}
\mathbb{E}\left\|\boldsymbol{W}(\boldsymbol{\theta}_N(\boldsymbol{t}))\big(\boldsymbol{J}_i-\boldsymbol{f}_\varepsilon\big)\right\|^3&=\mathbb{E}\left\|\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}}\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))\mathbb{1}\left\{\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\in[0,c]\right\}\big(J_{i,\boldsymbol{z}}-f_\varepsilon(\boldsymbol{z})\big)\right\|^3\\
&\leq\mathbb{E}\left\|\sum_{\boldsymbol{z}\in\boldsymbol{z}}\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))\big(J_{i,\boldsymbol{z}}-f_\varepsilon(\boldsymbol{z})\big)\right\|^3\\
(\textit{Minkowski's inequality})\quad&\leq\left[\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}}\Big(\|\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))\|^3\mathbb{E}\big|J_{i,\boldsymbol{z}}-f_\varepsilon(\boldsymbol{z})\big|^3\Big)^{1/3}\right]^3\\
(\textit{equation (D.1)})\quad&=\left[\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}}\Big(\|\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))\|^3f_\varepsilon(\boldsymbol{z})(1-f_\varepsilon(\boldsymbol{z}))\big((1-f_\varepsilon(\boldsymbol{z}))^2+f_\varepsilon(\boldsymbol{z})^2\big)\Big)^{1/3}\right]^3\\
&=:\widetilde{\mu}^3,
\end{aligned}$$

where the first line follows by definition of $\boldsymbol{\theta}\mapsto\boldsymbol{W}(\boldsymbol{\theta})$. By Assumption A.5, $\|\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))\|^3<\infty$, and, clearly, $f_\varepsilon(\boldsymbol{z})(1-f_\varepsilon(\boldsymbol{z}))\big((1-f_\varepsilon(\boldsymbol{z}))^2+f_\varepsilon(\boldsymbol{z})^2\big)<\infty$. Subsequently, $\widetilde{\mu}^3$ is finite. It follows from the definition of $\boldsymbol{X}_i$ that

$$\begin{aligned}
\mathbb{E}\left\|\boldsymbol{X}_i\right\|^3&\leq\left|\left|\left|\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\right|\right|\right|^3\mathbb{E}\left\|\boldsymbol{W}(\boldsymbol{\theta}_N(\boldsymbol{t}))\big(\boldsymbol{J}_i-\boldsymbol{f}_\varepsilon\big)\right\|^3\\
&\leq\left|\left|\left|\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\right|\right|\right|^3\widetilde{\mu}^3\\
&<\infty,
\end{aligned}\tag{D.2}$$

where the second line follows from the previous display, and the third line from $\widetilde{\mu}^3<\infty$ and the fact that $\left|\left|\left|\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\right|\right|\right|<\infty$, which is implied by Assumption A.5. This proves the first assertion.

For the second assertion, note that $\boldsymbol{J}_i$ obeys a multinomial distribution with 1 trial, m events, and event probabilities gathered in vector $\boldsymbol{f}_\varepsilon$. Hence, $\mathbb{E}\boldsymbol{J}_i=\boldsymbol{f}_\varepsilon$ and $\mathbb{V}\mathrm{ar}\left[\boldsymbol{J}_i\right]=$

$\operatorname{diag}(\boldsymbol{f}_\varepsilon) - \boldsymbol{f}_\varepsilon \boldsymbol{f}_\varepsilon^\top = \boldsymbol{\Omega}$. Therefore, $\mathbb{E}\boldsymbol{X}_i = \boldsymbol{0}$, as well as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\boldsymbol{X}_i\right] &= \boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2} \boldsymbol{W}(\boldsymbol{\theta}_N(\boldsymbol{t}))\boldsymbol{\Lambda}\boldsymbol{W}(\boldsymbol{\theta}_N(\boldsymbol{t}))^\top \left(\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\right)^\top \\
&= \boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2} \boldsymbol{U}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right) \left(\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\right)^\top \\
&= \boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2} \boldsymbol{U}\left(\boldsymbol{\theta}_0\right) \left(\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\right)^\top + o\left(N^{-1/2}\right) \\
&= \boldsymbol{I}_d + o\left(N^{-1/2}\right),
\end{aligned}$$

where $\boldsymbol{I}_d$ denotes the the $d \times d$ identity matrix. The third line follows from the fact that $\boldsymbol{\theta}_N(\boldsymbol{t}) = \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{V}\boldsymbol{t} = \boldsymbol{\theta}_0 + o\left(N^{-1/2}\right)$ and the continuity of $\boldsymbol{\theta} \mapsto \boldsymbol{U}(\boldsymbol{\theta})$, where the latter is implied by Assumption A.4. It follows that $\boldsymbol{X}_i$ has mean zero and identity covariance matrix when N is large enough. This fact together with equation (D.2) allows me to apply Theorem 1.3 in Götze (1991, equation 1.5), and the second assertion follows. This completes the proof. ■

Lemma D.3. *Grant the assumptions of Theorem 2 and put $\boldsymbol{\theta}_N(\boldsymbol{t}) = \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{V}\boldsymbol{t}$ for a fixed vector $\boldsymbol{t} \in \mathbb{R}^d$ and symmetric invertible matrix $\boldsymbol{V} \in \mathbb{R}^{d\times d}$. Then the equality*

$$\begin{aligned}
&-\sqrt{N}\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} f_\varepsilon(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))\left(w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right) - w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\right) \\
&\quad = \sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} w'\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)} f_\varepsilon(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)^\top\boldsymbol{V}\boldsymbol{t} + o\left(N^{-1/2}\right) + o\left(N^{-1/2}\right) + o\left(N^{-1/2}\right)
\end{aligned}$$

holds true.

Proof. I can write

$$\begin{aligned}
&-\sqrt{N}\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} f_\varepsilon(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))\left(w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right) - w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\right) \\
&\quad = -\sqrt{N}\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} \underbrace{\frac{w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))}\right) - w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_0)}\right)}{\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))} - \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_0)}}}_{=\text{Ⓐ}} \underbrace{\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)} - \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)}_{=\text{Ⓑ}} \underbrace{f_\varepsilon(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))}_{=\text{Ⓒ}}
\end{aligned}$$

For term Ⓐ, put $h_N = N^{-1/2}\boldsymbol{V}\boldsymbol{t}$ such that $\boldsymbol{\theta}_N(\boldsymbol{t}) = \boldsymbol{\theta}_0 + h_N$ and note that by the definition of the derivative, it holds true that

$$\lim_{h_N\to 0} \frac{w\left(\frac{f_\varepsilon(k)}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_0+h_N)}\right) - w\left(\frac{f_\varepsilon(k)}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_0)}\right)}{\frac{f_\varepsilon(k)}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_0+h_N)} - \frac{f_\varepsilon(k)}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_0)}} = w'\left(\frac{f_\varepsilon(k)}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right),$$

where the right hand side exists by Assumption A.9. By definition, $h_N = o\left(N^{-1/2}\right)$. Combined with the previous display, one obtains

$$Ⓐ = \frac{w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))}\right) - w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_0)}\right)}{\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))} - \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_0)}} = w'\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right) + o\left(N^{-1/2}\right).$$

For term Ⓑ, put $g(\boldsymbol{\theta}) = \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})}$ and note that

$$\frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{\theta}) = -\boldsymbol{s}_k(\boldsymbol{\theta}) \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}.$$

Performing the Taylor expansion

$$\begin{aligned} g(\boldsymbol{\theta}_N(\boldsymbol{t})) = g\big(\boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{V}\boldsymbol{t}\big) &= g(\boldsymbol{\theta}_0) + N^{-1/2}\left(\frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{\theta}_0)\right)^\top \boldsymbol{V}\boldsymbol{t} + o\left(N^{-1/2}\right) \\ &= \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)} - N^{-1/2}\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)^\top \boldsymbol{V}\boldsymbol{t} + o\left(N^{-1/2}\right), \end{aligned}$$

it follows that

$$Ⓑ = \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)} - \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)} = -N^{-1/2}\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)^\top \boldsymbol{V}\boldsymbol{t} + o\left(N^{-1/2}\right).$$

For term Ⓒ, the continuity of $\boldsymbol{\theta} \mapsto \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})$ (Assumption A.4) implies that

$$Ⓒ = f_\varepsilon(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t})) = f_\varepsilon(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0) + o\left(N^{-1/2}\right)$$

Combining the derived expressions for Ⓐ, Ⓑ, Ⓒ completes the proof. ■

Lemma D.4. *For fixed $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$, the matrix $\boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{\theta})$ is equal to the Hessian matrix of $\log\left(p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)$, at $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. In addition, the Fisher information matrix $\boldsymbol{J}(\boldsymbol{\theta})$ of a density $\boldsymbol{p}\left(\boldsymbol{\theta}\right)$ can be expressed as*

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{\theta}) p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right).$$

Proof. Note that $\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log\left(p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) = \frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})$. Then, applying the product rule and then the chain rule,

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}\frac{\partial}{\partial\boldsymbol{\theta}}p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \\
&= \left(\frac{\partial}{\partial\boldsymbol{\theta}}\frac{1}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^\top + \frac{1}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \\
&= -\frac{1}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)^2}\left(\frac{\partial}{\partial\boldsymbol{\theta}}p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^\top + \frac{1}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \\
&= \frac{1}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}p_k(\boldsymbol{\theta})\right) - \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})^\top,
\end{aligned}$$

which is equal to the definition of $\boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{\theta})$. The second assertion follows immediately by the definition of the Fisher information matrix. ■

D.2 Proof of Theorem 1

For any $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$, the pointwise convergence $\widehat{f}_N(\boldsymbol{z}) \xrightarrow{\text{a.s.}} f_\varepsilon(\boldsymbol{z})$ holds true when $N \to \infty$ (see e.g., Chapter 19.2 in Van der Vaart, 1998). Since loss $L(\boldsymbol{\theta}, f)$ is continuous in any density f on $\boldsymbol{\mathcal{Z}}$, I know by the continuous mapping theorem that for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$L\left(\boldsymbol{\theta}, \widehat{f}_N\right) \xrightarrow{\text{a.s.}} L\left(\boldsymbol{\theta}, f_\varepsilon\right), \tag{D.3}$$

as $N \to \infty$.

Put $L_N\left(\boldsymbol{\theta}\right) = L\left(\boldsymbol{\theta}, \widehat{f}_N\right)$ and $L\left(\boldsymbol{\theta}\right) = L\left(\boldsymbol{\theta}, f_\varepsilon\right)$. I proceed to show that the convergence in (D.3) is uniform on $\boldsymbol{\Theta}$. One way of doing so (Newey, 1991) is to show that the sequence $\{L_N\left(\boldsymbol{\theta}\right)\}_N$ is equicontinuous on the parameter space $\boldsymbol{\Theta}$ and then apply a well-known result in Rudin (1976). Since $\{L_N\left(\boldsymbol{\theta}\right)\}_N$ is indeed equicontinuous on $\boldsymbol{\Theta}$ by Lemma D.1, $\boldsymbol{\Theta}$ is compact (Assumption A.2), and $\{L_N\left(\boldsymbol{\theta}\right)\}_N$ converges almost surely pointwise to $L(\boldsymbol{\theta})$ on $\boldsymbol{\Theta}$ by equation (D.3), it follows from Exercise 7.16 in Rudin (1976) that the convergence in equation (D.3) is uniform on $\boldsymbol{\Theta}$.

Therefore, when $N \to \infty$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|L(\boldsymbol{\theta}) - L_N(\boldsymbol{\theta})\right| \xrightarrow{\text{a.s.}} 0, \tag{D.4}$$

and

$$\begin{aligned}
\left| L\left(\boldsymbol{\theta}_0\right) - L\left(\widehat{\boldsymbol{\theta}}_N\right) \right| &= \left| L\left(\boldsymbol{\theta}_0\right) - L_N\left(\widehat{\boldsymbol{\theta}}_N\right) + L_N\left(\widehat{\boldsymbol{\theta}}_N\right) - L\left(\widehat{\boldsymbol{\theta}}_N\right) \right| \\
(\textit{triangle inequality}) \quad &\leq \left| L\left(\boldsymbol{\theta}_0\right) - L_N\left(\widehat{\boldsymbol{\theta}}_N\right) \right| + \left| L\left(\widehat{\boldsymbol{\theta}}_N\right) - L_N\left(\widehat{\boldsymbol{\theta}}_N\right) \right| \\
(\textit{Assumption A.3}) \quad &\leq \left| L\left(\widehat{\boldsymbol{\theta}}_N\right) - L_N\left(\widehat{\boldsymbol{\theta}}_N\right) \right| + \left| L\left(\widehat{\boldsymbol{\theta}}_N\right) - L_N\left(\widehat{\boldsymbol{\theta}}_N\right) \right| \\
(\textit{equation (D.4)}) \quad &\xrightarrow{\text{a.s.}} 0.
\end{aligned}$$

It follows that $\widehat{\boldsymbol{\theta}}_N \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0$ as $N \to \infty$, which concludes the proof. ■

D.3 Proof of Theorem 2

The proof strategy employed in this proof is similar to Ruckstuhl & Welsh (2001) and Victoria-Feser & Ronchetti (1997).

For $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and a density f on $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$, denote by

$$\boldsymbol{\eta}\left(\boldsymbol{\theta}, f\right) = \frac{\partial}{\partial \boldsymbol{\theta}} L\left(\boldsymbol{\theta}, f\right) = -\sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}) f(\boldsymbol{z}) w\left(\frac{f(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}\right)$$

the objective's gradient at $\boldsymbol{\theta}$. By Assumption A.8, $L(\boldsymbol{\theta}, f_\varepsilon)$ is convex in a neighborhood of $\boldsymbol{\theta}_0$, therefore its gradient $\boldsymbol{\eta}\left(\boldsymbol{\theta}, f_\varepsilon\right)$ is non-negative for $\boldsymbol{\theta}$ in this neighborhood. Define this neighborhood as follows. For an arbitrary vector $\boldsymbol{t} \in \mathbb{R}^d$ and an arbitrary symmetric invertible matrix $\boldsymbol{V} \in \mathbb{R}^{d \times d}$, let $\boldsymbol{\theta}_N(\boldsymbol{t}) = \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{V}\boldsymbol{t}$ define the neighborhood's boundary such that its radius about $\boldsymbol{\theta}_0$ is of length $\|\boldsymbol{\theta}_N(\boldsymbol{t})\|$. Hence, $\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), \widehat{f}_N\right) \geq 0$ and, by definition, $\sqrt{N}\boldsymbol{V}^{-1}(\boldsymbol{\theta}_N(\boldsymbol{t}) - \boldsymbol{\theta}_0) = \boldsymbol{t}$. It follows that for every $\boldsymbol{\theta}$ in the aforementioned neighborhood, the equivalence

$$\sqrt{N}\boldsymbol{V}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \leq \boldsymbol{t} \iff \boldsymbol{\eta}\left(\boldsymbol{\theta}, \widehat{f}_N\right) \geq \boldsymbol{0}, \tag{D.5}$$

holds true, where an event $\{\boldsymbol{T} \leq \boldsymbol{t}\}$ is to be understood as the event $\{T_1 \leq t_1, \ldots, T_d \leq t_d\}$ for a d-variate random variable $\boldsymbol{T} = (T_1, \ldots, T_d)^\top$. By construction of the estimator in (8), it holds true that $\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}_N, \widehat{f}_N) = \boldsymbol{0}$, and, by Theorem 1, $\widehat{\boldsymbol{\theta}}_N \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0$, as $N \to \infty$. Therefore, for N large enough, $\widehat{\boldsymbol{\theta}}_N$ is in the neighborhood of $\boldsymbol{\theta}_0$ with probability one. Hence, by (D.5),

$$\sqrt{N}\boldsymbol{V}^{-1}\left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\right) \leq \boldsymbol{t},$$

when N is large enough.

Before I continue with the proof, I first need some additional notation. For $i = 1, \ldots, N$, put

$$\boldsymbol{J}_i = \left(J_{i,\boldsymbol{z}_1}, J_{i,\boldsymbol{z}_2}, \ldots, J_{i,\boldsymbol{z}_m}\right)^\top = \left(\mathbb{1}\left\{Y_i = \boldsymbol{z}_1\right\}, \mathbb{1}\left\{Y_i = \boldsymbol{z}_2\right\}, \ldots, \mathbb{1}\left\{Y_i = \boldsymbol{z}_m\right\}\right)^\top,$$

and note that each $\boldsymbol{J}_i$ is distributed according to a multinomial distribution with one trial, m events, and event probabilities gathered in vector $\boldsymbol{f}_\varepsilon$. Hence, $\mathbb{E}[\boldsymbol{J}_i] = \boldsymbol{f}_\varepsilon$ and $\mathbb{V}\mathrm{ar}[\boldsymbol{J}_i] = \mathrm{diag}(\boldsymbol{f}_\varepsilon) - \boldsymbol{f}_\varepsilon \boldsymbol{f}_\varepsilon^\top = \boldsymbol{\Omega}$. Next, denote by

$$Z_N(\boldsymbol{z}) = N^{-1/2} \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \left(J_{i,\boldsymbol{z}} - f_\varepsilon(\boldsymbol{z}) \right) \tag{D.6}$$

the difference between $\widehat{f}_N(\boldsymbol{z})$ and $f_\varepsilon(\boldsymbol{z})$, scaled by $\sqrt{N}$. Then, put

$$\boldsymbol{Z}_N = (Z_N(\boldsymbol{z}_1), Z_N(\boldsymbol{z}_2), \ldots, Z_N(\boldsymbol{z}_m))^\top = N^{-1/2} \sum_{i=1}^{N} (\boldsymbol{J}_i - \boldsymbol{f}_\varepsilon).$$

Furthermore, it is useful to define the d-variate function

$$\begin{aligned}
&\boldsymbol{\psi}(\boldsymbol{\theta}, \boldsymbol{Z}_N) \\
&\quad = -N^{-1/2} \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}) Z_N(\boldsymbol{z}) \mathbb{1}\left\{ \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})} \in [0, c] \right\} \\
&\quad = -\sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}) Z_N(\boldsymbol{z}) \mathbb{1}\left\{ \frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})} \in [0, c] \right\} \left(\frac{1}{N} \sum_{i=1}^{N} J_{i,\boldsymbol{z}} - f_\varepsilon(\boldsymbol{z}) \right) \\
&\quad = -N^{-3/2} \left(\boldsymbol{s}_{\boldsymbol{z}_1}(\boldsymbol{\theta}) \mathbb{1}\left\{ \frac{f_\varepsilon(\boldsymbol{z}_1)}{p_{\boldsymbol{z}_1}(\boldsymbol{\theta})} \in [0, c] \right\}, \cdots, \boldsymbol{s}_{\boldsymbol{z}_m}(\boldsymbol{\theta}) \mathbb{1}\left\{ \frac{f_\varepsilon(\boldsymbol{z}_m)}{p_{\boldsymbol{z}_m}(\boldsymbol{\theta})} \in [0, c] \right\} \right) \begin{pmatrix} \sum_{i=1}^{N} J_{i,\boldsymbol{z}_1} - f_\varepsilon(\boldsymbol{z}_1) \\ \vdots \\ \sum_{i=1}^{N} J_{i,\boldsymbol{z}_m} - f_\varepsilon(\boldsymbol{z}_m) \end{pmatrix} \\
&\quad = -N^{-3/2} \sum_{i=1}^{N} \boldsymbol{W}(\boldsymbol{\theta}) (\boldsymbol{J}_i - \boldsymbol{f}_\varepsilon),
\end{aligned} \tag{D.7}$$

for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

With these definitions, I am now ready to proceed with the proof. I can write (D.6) as $\sqrt{N}\widehat{f}_N(\boldsymbol{z}) = Z_N(\boldsymbol{z}) + \sqrt{N} f_\varepsilon(\boldsymbol{z})$. Therefore, when evaluating the gradient at $(\boldsymbol{\theta}_N(\boldsymbol{t}), \widehat{f}_N)$ and scaling by $\sqrt{N}$, one obtains

$$\begin{aligned}
&\sqrt{N} \boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), \widehat{f}_N \right) \\
&\quad = -\sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t})) \widehat{f}_N(k) w\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))} \right) \\
&\quad = -\sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t})) w\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))} \right) Z_N(\boldsymbol{z}) - \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t})) w\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))} \right) f_\varepsilon(\boldsymbol{z}).
\end{aligned} \tag{D.8}$$

I now wish to Taylor-expand the function $w\left(\widehat{f}_N(\boldsymbol{z}) / p_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t})) \right)$ about the point $w\left(f_\varepsilon(\boldsymbol{z}) / p_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t})) \right)$, for which I need to check differentiability at this point. The

function $w(\cdot)$ is not differentiable at point c. By Assumption A.9, $w\Big(f_\varepsilon(\boldsymbol{z})/p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)\Big) \neq c$, which implies that there exists a neighborhood of $f_\varepsilon(\boldsymbol{z})/p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)$ that excludes c. Since $\boldsymbol{\theta}_N(\boldsymbol{t}) = \boldsymbol{\theta}_0 + o\left(N^{-1/2}\right)$, the point $\widehat{f}_N(\boldsymbol{z})/p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)$ is in this neighborhood when N is sufficiently large. Hence, I can perform the expansion

$$w\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right) = w\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right) + w'\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right)\frac{\widehat{f}_N(\boldsymbol{z}) - f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)} + o\left(N^{-1/2}\right),$$

which allows me to write (D.8) as

$$\begin{aligned}
\sqrt{N}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), \widehat{f}_N\right) &= -\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right)Z_N(\boldsymbol{z}) + \sqrt{N}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), f_\varepsilon\right) \\
&\quad - \sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))w'\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right)\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}Z_N(\boldsymbol{z}) + o\left(N^{-1/2}\right) \\
&= -\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right)Z_N(\boldsymbol{z}) + \sqrt{N}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), f_\varepsilon\right) \\
&\quad - \sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})\left(-w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right) + \mathbb{1}\left\{\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)} \in [0, c]\right\}\right)Z_N(\boldsymbol{z}) \\
&\quad + o\left(N^{-1/2}\right) \\
&= -\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))Z_N(\boldsymbol{z})\mathbb{1}\left\{\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)} \in [0, c]\right\} + \sqrt{N}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), f_\varepsilon\right) \\
&\quad + o\left(N^{-1/2}\right), \\
&= \sqrt{N}\boldsymbol{\psi}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), \boldsymbol{Z}_N\right) + \sqrt{N}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), f_\varepsilon\right) + o\left(N^{-1/2}\right),
\end{aligned} \tag{D.9}$$

where I have applied in the the second equality the identity $x \mapsto w'(x)x = -w(x) + \mathbb{1}\{x \in [0, c]\}$ and in the last equation the definition given in equation (D.7). Next, I derive a bound for the difference between the probability $\Phi_d(\boldsymbol{t})$ and the probability that the

left hand side in the previous display is nonnegative. I have that

$$
\begin{aligned}
&\left|\mathbb{P}\left[\sqrt{N}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),\widehat{f}_N\right)\geq\mathbf{0}\right]-\Phi_d(\boldsymbol{t})\right| \\
&\quad=\left|1-\mathbb{P}\left[\sqrt{N}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),\widehat{f}_N\right)\leq\mathbf{0}\right]-\Phi_d(\boldsymbol{t})\right| \\
&\quad=\left|1-\mathbb{P}\Big[\sqrt{N}\boldsymbol{\psi}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),\boldsymbol{Z}_N\right)\leq-\sqrt{N}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)\Big]-\Phi_d(\boldsymbol{t})\right|+o\left(N^{-1/2}\right) \\
&\quad=\left|1-\mathbb{P}\Big[\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\psi}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),\boldsymbol{Z}_N\right)\leq-\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)\Big]-\Phi_d(\boldsymbol{t})\right|+o\left(N^{-1/2}\right) \\
&\quad=\Bigg|\overbrace{\Phi_d\left(-\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)\right)+\Phi_d\left(\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)\right)}^{=1,\ \textit{by symmetry of standard normal density}} \\
&\qquad-\mathbb{P}\Big[\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\psi}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),\boldsymbol{Z}_N\right)\leq-\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)\Big]-\Phi_d(\boldsymbol{t})\Bigg|+o\left(N^{-1/2}\right) \\
&\quad\leq\Bigg|\Phi_d\Big(\overbrace{-\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)}^{=:\boldsymbol{x}}\Big) \\
&\qquad-\mathbb{P}\Big[\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\psi}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),\boldsymbol{Z}_N\right)\leq\overbrace{-\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)}^{=:\boldsymbol{x}}\Big]\Bigg| \\
&\qquad+\left|\Phi_d\left(\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)\right)-\Phi_d(\boldsymbol{t})\right|+o\left(N^{-1/2}\right) \\
&\quad\leq\sup_{\boldsymbol{x}\in\mathbb{R}^d}\left|\Phi_d(\boldsymbol{x})-\mathbb{P}\Big[\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\psi}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),\boldsymbol{Z}_N\right)\leq\boldsymbol{x}\Big]\right| \\
&\qquad+\left|\Phi_d\left(\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)\right)-\Phi_d(\boldsymbol{t})\right|+o\left(N^{-1/2}\right) \\
&\quad=\sup_{\boldsymbol{x}\in\mathbb{R}^d}\left|\Phi_d(\boldsymbol{x})-\mathbb{P}\Big[-N^{-1/2}\sum_{i=1}^{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{W}(\boldsymbol{\theta}_N(\boldsymbol{t}))\big(\boldsymbol{J}_i-\boldsymbol{f}_\varepsilon\big)\leq\boldsymbol{x}\Big]\right| \\
&\qquad+\left|\Phi_d\left(\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)\right)-\Phi_d(\boldsymbol{t})\right|+o\left(N^{-1/2}\right)
\end{aligned}
$$

where I have used in the second (in)equality equation (D.9), in the fifth (in)equality the triangle inequality, and in the last (in)equality the definition in (D.7). A direct application

of Lemma D.2 now yields the bound

$$\left|\mathbb{P}\left[\sqrt{N}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),\widehat{f}_N\right)\geq\mathbf{0}\right]-\Phi_d(\boldsymbol{t})\right|\leq\left|\Phi_d\left(\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)\right)-\Phi_d(\boldsymbol{t})\right| \tag{D.10}$$
$$+o\left(N^{-1/2}\right)+o\left(N^{-1/2}\right)+o\left(N^{-1/2}\right).$$

I remains to be shown that first term on the right hand side vanishes as $N\to\infty$.

Write

$$\begin{aligned}
&\sqrt{N}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)\\
&\quad=\sqrt{N}\Big(\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}),f_\varepsilon\right)-\underbrace{\boldsymbol{\eta}\left(\boldsymbol{\theta}_0,f_\varepsilon\right)}_{=\mathbf{0}}\Big)\\
&\quad=-\sqrt{N}\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}}f_\varepsilon(\boldsymbol{z})\left(\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right)-\boldsymbol{s}_k(\boldsymbol{\theta}_0)w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\right)\\
&\quad=-\sqrt{N}\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}}f_\varepsilon(\boldsymbol{z})\left(\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right)-\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\right)\\
&\qquad\underbrace{+\sqrt{N}\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}}f_\varepsilon(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))\left(w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)-w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\right)}_{=\mathbf{0}}\\
&\quad=-\sqrt{N}\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}}f_\varepsilon(k)\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))\left(w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_N(\boldsymbol{t})\right)}\right)-w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\right)\\
&\qquad-\sqrt{N}\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}}f_\varepsilon(\boldsymbol{z})w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\left(\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))-\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)\right),
\end{aligned} \tag{D.11}$$

where the first equation follows by Assumption A.3, the second equation by simply writing out the expression for gradient $\boldsymbol{\eta}\left(\boldsymbol{\theta},f_\varepsilon\right)$, and the third equation by adding an intelligent zero.

I can rewrite the second summand in equation (D.11) as

$$\sqrt{N}\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}}f_\varepsilon(\boldsymbol{z})w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\left(\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))-\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)\right)=\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}}f_\varepsilon(\boldsymbol{z})w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)\boldsymbol{V}\boldsymbol{t}+o\left(N^{-1/2}\right),$$

which follows from the Taylor expansion

$$\begin{aligned}
\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_N(\boldsymbol{t}))&=\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0+N^{-1/2}\boldsymbol{V}\boldsymbol{t})\\
&=\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)+N^{-1/2}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)\boldsymbol{V}\boldsymbol{t}+o\left(N^{-1/2}\right)\\
(\textit{by Lemma D.4})\quad&=\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)+N^{-1/2}\boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)\boldsymbol{V}\boldsymbol{t}+o\left(N^{-1/2}\right).
\end{aligned}$$

The first summand in equation (D.11) possesses a characterization derived in Lemma D.3. Using this characterization and pre-multiplying equation (D.11) by $\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}$ (which exists by Assumption A.7), I obtain

$$\begin{aligned}
&\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), f_\varepsilon\right) \\
&\quad = \boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\left[\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} f_\varepsilon(k)\left(w'\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)^\top - w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{\theta}_0)\right)\right]\boldsymbol{V}\boldsymbol{t} \\
&\qquad + o\left(N^{-1/2}\right),
\end{aligned}$$

where the term $o\left(N^{-1/2}\right)$ collects four individual terms that all vanish at rate $N^{-1/2}$.

So far, the choice of the $d \times d$ matrix $\boldsymbol{V}$ has been arbitrary, except for symmetry and invertibility. For the choice

$$\boldsymbol{V} = \boldsymbol{V}\left(\boldsymbol{\theta}_0\right) = \boldsymbol{M}\left(\boldsymbol{\theta}_0\right)^{-1}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{1/2},$$

where

$$\boldsymbol{M}\left(\boldsymbol{\theta}\right)^{-1} = \left[\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} f_\varepsilon(\boldsymbol{z})\left(w'\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}\right)\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})\boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta})^\top - w\left(\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}\right)\boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)\right]^{-1},$$

I obtain that

$$\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), f_\varepsilon\right) = \boldsymbol{t} + o\left(N^{-1/2}\right),$$

and it follows by (D.10) that

$$\begin{aligned}
&\left|\mathbb{P}\left[\sqrt{N}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), \widehat{f}_N\right) \geq \boldsymbol{0}\right] - \Phi_d(\boldsymbol{t})\right| \\
&\quad \leq \left|\Phi_d\left(\sqrt{N}\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^{-1/2}\boldsymbol{\eta}\left(\boldsymbol{\theta}_N(\boldsymbol{t}), f_\varepsilon\right)\right) - \Phi_d(\boldsymbol{t})\right| + o\left(N^{-1/2}\right) + o\left(N^{-1/2}\right) + o\left(N^{-1/2}\right) \\
&\quad = o\left(N^{-1/2}\right) + o\left(N^{-1/2}\right) + o\left(N^{-1/2}\right) + o\left(N^{-1/2}\right).
\end{aligned}$$

Using the equivalence in (D.5), it follows that

$$\mathbb{P}\left[\sqrt{N}\boldsymbol{V}\left(\boldsymbol{\theta}_0\right)^{-1}\left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\right) \leq \boldsymbol{t}\right] \longrightarrow \Phi_d\Big(\boldsymbol{t}\Big), \quad \text{as } N \to \infty.$$

Since the choice of $\boldsymbol{t} \in \mathbb{R}^d$ is arbitrary, this is sufficient to conclude that

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\right) \xrightarrow{\text{d}} \text{N}_d\Big(\boldsymbol{0}, \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\Big),$$

as $N \to \infty$, where

$$\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right) = \boldsymbol{M}\left(\boldsymbol{\theta}\right)^{-1}\boldsymbol{U}\left(\boldsymbol{\theta}\right)\boldsymbol{M}\left(\boldsymbol{\theta}\right)^{-1}.$$

This concludes the proof. ■

D.4 Proof of Corollary 1

By Theorem 2, it holds true that

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\right) \xrightarrow{\mathrm{d}} \mathrm{N}_d\Big(\mathbf{0}, \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\Big).$$

Applying the Delta method (e.g. Theorem 3.1 in Van der Vaart, 1998) on this result yields

$$\sqrt{N}\left(p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right) - p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)\right) \xrightarrow{\mathrm{d}} \mathrm{N}\Big(0,\ \boldsymbol{g}_{\boldsymbol{z}}(\boldsymbol{\theta})^\top \boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right) \boldsymbol{g}_{\boldsymbol{z}}(\boldsymbol{\theta})\Big).$$

It follows that

$$\widetilde{T}_N(\boldsymbol{z}) := \frac{p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right) - p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)}{\sqrt{\sigma^2_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)/N}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1), \tag{D.12}$$

where $\sigma^2_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = \boldsymbol{g}_{\boldsymbol{z}}(\boldsymbol{\theta})^\top \boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right) \boldsymbol{g}_{\boldsymbol{z}}(\boldsymbol{\theta})$. Suppose one wishes to test a null hypothesis of the form

$$\mathrm{H}_0 : p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right) = p,$$

for an arbitrary but fixed $p \in [0,1]$. Conditional on this null hypothesis, it follows from (D.12) that

$$\widetilde{T}_N(\boldsymbol{z}) \;\Big|\; \mathrm{H}_0 = \frac{p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right) - p}{\sqrt{\sigma^2_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right)/N}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

Choosing $p = \widehat{f}_N(\boldsymbol{z})$ yields the result. ■