*entropy*

MDPI

*Article*

# Distance-Based Estimation Methods for Models for Discrete and Mixed-Scale Data

**Elisavet M. Sofikitou [1], Ray Liu [2], Huipei Wang [1] and Marianthi Markatou [1,]***

[1] Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, USA; esofikit@buffalo.edu (E.M.S.); huipeiwa@buffalo.edu (H.W.)

[2] Head of Oncology Data Science, AstraZeneca PLC, Gaithersburg, MD 20878, USA; ray.liu1@astrazeneca.com

* Correspondence: markatou@buffalo.edu

**Abstract:** Pearson residuals aid the task of identifying model misspecification because they compare the estimated, using data, model with the model assumed under the null hypothesis. We present different formulations of the Pearson residual system that account for the measurement scale of the data and study their properties. We further concentrate on the case of mixed-scale data, that is, data measured in both categorical and interval scale. We study the asymptotic properties and the robustness of minimum disparity estimators obtained in the case of mixed-scale data and exemplify the performance of the methods via simulation.

**Keywords:** contingency tables; disparity; mixed-scale data; pearson residuals; residual adjustment function; robustness; statistical distances

check for updates

**Citation:** Sofikitou, E.M.; Liu, R.; Wang, H.; Markatou, M. Distance-Based Estimation Methods for Models for Discrete and Mixed-Scale Data. *Entropy* **2021**, *23*, 107. https://doi.org/10.3390/e23010107

Received: 5 December 2020
Accepted: 12 January 2021
Published: 14 January 2021

## 1. Introduction

Minimum disparity estimation has been studied extensively in models where the scale of the data is either interval or ratio (Beran [1], Basu and Lindsay [2]). It has also been studied in the discrete outcomes case. Specifically, when the response variable is discrete and the explanatory variables are continuous, Pardo et al. [3] introduced a general class of distance estimators based on $\phi$-divergence measures, the minimum $\phi$-divergence estimators, and they studied their asymptotic properties. The estimators can be viewed as an extension/generalization of the Maximum Likelihood Estimator (MLE). Pardo et al. [4] used the minimum $\phi$-divergence estimator in a $\phi$-divergence statistic to perform goodness-of-fit tests in logistic regression models, while Pardo and Pardo [5] extended the previous works to address solving problems for testing in generalized linear models with binary scale data.

The case where data are measured on discrete scale (either on ordinal or generally categorical scale) has also attracted the interest of other researchers. For instance, Simpson [6] demonstrated that minimum Hellinger distance estimators fulfill desirable robustness properties and for this reason can be effective in the analysis of count data prone to outliers. Simpson [7] also suggested tests based on the minimum Hellinger distance for parametric inference which are robust as the density of the (parametric) model can be nonparametrically estimated. In contrast, Markatou et al. [8] used weighted likelihood equations to obtain efficient and robust estimators in discrete probability models and applied their methods to logistic regression, whereas Basu and Basu [9] considered robust penalized minimum disparity estimators for multinomial models with good small sample efficiency.

Moreover, Gupta et al. [10], Martín and Pardo [11] and Castilla et al. [12] used the minimum $\phi$-divergence estimator to provide solution to testing problems in polytomous regression models. Working in a similar fashion, Martín and Pardo [13] studied the properties of the family of $\phi$-divergence estimators for log-linear models with linear constraints under multinomial sampling in order to identify potential associations between various variables in multi-way contingency tables. Pardo and Martín [14] presented an overview of works associated with contigency tables of symmetric structure on the basis of minimum $\phi$-divergence estimators and minimum $\phi$-divergence test statistics. Additional works include Pardo and Pardo [15] and Pardo et al. [16]. Alternative power divergence measures have been introduced by Basu et al. [17].

The class of $f$ or $\phi-$divergences was originally introduced by Csiszár [18]. The structural characteristics of this class and their relationship to the concepts of efficiency and robustness were studied, for the case of discrete probability models, by Lindsay [19]. Basu and Lindsay [2] studied the properties of estimators derived by minimizing $f-$divergences between continuous models and presented examples showing the robustness results of these estimates. We also note that Tamura and Boos [20] studied the minimum Hellinger distance estimation for multivariate location and covariance. Additionally, formal robustness results were presented in Markatou et al. [8,21] in connection with the introduction of weighted likelihood estimation.

If $G$ is a real valued, convex function, defined on $[0, \infty)$ and such that $G(u)$ converges to 0 as $u \to \infty$, $0G(0/0) = 0$, $0G(u/0) = uG_\infty$, $G_\infty = \lim_{u\to\infty}(G(u)/u)$, the class of $\phi-$divergences is defined as

$$\rho(\tau, m_{\beta_0}) = \sum G\Big(\frac{\tau(t)}{m_{\beta_0}(t)}\Big) m_{\beta_0}(t),$$

where $\tau(\cdot)$, $m_{\beta_0}(\cdot)$ are two probability models. Notice that we define $\rho(\tau, m_{\beta_0})$ on discrete probability models first, where $\mathscr{T} = \{0, 1, 2, \ldots, T\}$ is a discrete sample space, $T$ possibly infinite, and $m_{\beta_0}(t) \in \mathscr{M} = \{m_{\boldsymbol{\beta}}(t) : \boldsymbol{\beta} \in \mathscr{B}\}$, $\mathscr{B}$ is the parameter space $\mathscr{B} \subseteq \mathbb{R}^d$. Furthermore, different forms of the function $G(u)$ provide different statistical distances or divergences.

We can change the argument of the function $G$ from $\frac{\tau(t)}{m_{\beta_0}(t)}$ to $\frac{\tau(t)}{m_{\beta_0}(t)} - 1$. Then, $G$ is a function of the Pearson residual which is defined as $\delta(t) = \frac{\tau(t)}{m_{\beta_0}(t)} - 1$, and takes values in $[-1, \infty)$. If the measurement scale is interval/ratio, then the Pearson residuals are modified to reflect and adjust for the discrepancy of scale between data, that are always discrete, and the assumed continuous probability model (see Basu and Lindsay [2]).

The Pearson residual is used by Lindsay [19], Basu and Lindsay [2] and Markatou et al. [8,21] in investigating the robustness of the minimum disparity and weighted likelihood estimators, respectively. This residual system allows one to identify distributional errors. If, in the equation of Pearson residual, we replace $\tau(t)$ with its best nonparametric representative $d(t)$, the proportion of observations in a sample with value $t$, then $\delta(t) = \frac{d(t)}{m_{\beta_0}(t)} - 1$. We note that the Pearson residuals are called so because $n \sum \delta^2(t) m(t)$ is Pearson's chi-squared distance. Furthermore, these residuals are not symmetric since they take values in $[-1, \infty]$ and are not standardized to have identical variances.

How does robustness fit into this picture? In the robustness literature, there is a denial of the model's truth. Following this logic, the framework based on disparities starts with goodness-of-fit by identifying a measure that assesses whether the model fits the data adequately. Then, we examine whether this measure of adequacy is robust and in what sense. A fundamental tool that assists in measuring the degree of robustness is the Pearson residual, because it measures model misspecification. That is, Pearson residuals provide information about the degree to which the specified model $m_\beta$ fits the data. In this context, outliers are defined as those data points that have a low probability of occurrence under the hypothesized model. Such probabilistic outliers are called *surprising observations* (Lindsay [19]). Furthermore, the robustness of estimators obtained via minimization of the divergence measures we discuss here is indicated by the shape of the associated Residual Adjustment Function (RAF), a concept that is reviewed in Section 2. Of note is that in contingency table analysis, the generalized residual system is used for examination of sources of error in models for contingency tables, see, for example, Haberman [22], Haberman and Sinharay [23]. The concept of generalized residuals in the case of generalized linear models is discussed, for example, in Pierce and Schafer [24].

Data sets are comprised of data measured on both categorical (ordinal or nominal) scale and interval/ratio scale. We can think of these data as realizations of discrete and continuous random variables respectively. Examples of data sets that include mixed-scale data are electronic health records containing diagnostic codes (discrete) and laboratory measurements (e.g., blood pressure, alanine amino transferase (ALT) measurements on interval/ratio scale) and marketing data (customer records include income and gender information). Additional examples include data from developmental toxicology (Aerts et al. [25]), where fetal data from laboratory animals include binary, categorical and continuous outcomes. In this context, the joint density of the discrete and continuous random variables is given as $m_{\boldsymbol{\beta}}(x, y) = f_{\boldsymbol{\beta}_1}(y|x) g_{\boldsymbol{\beta}_2}(x)$, where $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$ are parameter vectors indexing the joint, conditional on $x$ and probability density function of $x$.

Work on the analysis of mixed-scale data is complicated by the fact that is difficult to identify suitable joint probability distributions to describe both measurement scales of the data, although a number of ad hoc methods to the analysis of mixed-scale data have been used in applications. Olkin and Tate [26] proposed multivariate correlation models for mixed-scale data. Copulas also provide an attractive approach to modeling the joint distribution of mixed-scale data, though copulas are less straightforward to implement, and there are subtle identifiability issues that complicate the specification of a model (Genest and Nešlehová [27]).

To formulate the joint distribution in the mixed-scale variables case one can either specify the marginal distribution of the discrete variables and the conditional distribution of the continuous variables. Alternatively, one can specify the marginal distribution of the continuous variables and the conditional distribution of the discrete variables given the continuous variables. Of note here is that the direction of factorization generally yields distinct model interpretations and results. The first approach has received much attention in the literature, in the context of the analysis of data with mixtures of categorical and continuous variables. Here, the continuous variables follow different multivariate normal distributions for each possible setting of the categorical variable values; the categorical variables then follow an arbitrary marginal multinomial distribution. This model is known in the literature as the conditional Gaussian distribution model and is central in the discussion of graphical association models with mixed-scale variables (Lauritzen and Wermuth [28]). A very special case of this model is used in our simulations.

In this paper, we develop robust methods for mixed-scale data. Specifically, Section 2 reviews basic concepts in minimum disparity estimation, Section 3 defines Pearson residuals for data measured in discrete, interval/ratio and mixed-scale, and studies their properties. Section 4 establishes the optimization problem for obtaining estimators of the model parameters, while Sections 5 and 6 establish the robustness and asymptotic properties of these estimators. Finally, Section 7 presents simulations showing the performance of these methods and Section 8 offers discussions. The Appendix A includes proofs of the theoretical results.

## 2. Concepts in Minimum Disparity Estimation

Beran [1] introduced a robust method to estimate the parameters of a statistical model, called minimum Hellinger distance estimation. The parameter estimator is obtained by minimizing the Hellinger distance between a parametric model density and a nonparametric density estimator. Lindsay [19] extended the aforementioned method to incorporate many other distances, and introduced the concept of the residual adjustment function in the context of minimum disparity estimation. The Minimum Distance Estimators (MDE) of a parameter vector $\boldsymbol{\beta}$ are obtained by minimizing over $\boldsymbol{\beta}$, the distance (or disparity)

$$\rho(d, m_{\boldsymbol{\beta}}) = \sum_x G(\delta(x)) m_{\boldsymbol{\beta}}(x), \tag{1}$$

where the assumed model $m_{\boldsymbol{\beta}}$ is a probability mass function. When the model $m_{\boldsymbol{\beta}}$ is continuous, the MDE of the parameter vector $\boldsymbol{\beta}$ is obtained by minimizing over $\boldsymbol{\beta}$ the quantity

$$\rho(f^*, m^*_{\boldsymbol{\beta}}) = \int G(\delta(x)) m^*_{\boldsymbol{\beta}}(x)\, dx, \tag{2}$$

where $f^*(x) = \int k(x;t,h) d\hat{F}(t)$, $m^*_{\boldsymbol{\beta}}(x) = \int k(x;t,h) m_{\boldsymbol{\beta}}(t)\, dt$, $\hat{F}$ is the empirical distribution function obtained from the data and $k$ is a smooth family of kernel functions. One example is the normal density with mean $t$ and standard deviation $h$. Furthermore, $\delta(x)$ is the Pearson residual defined as $\delta(x) = f^*(x)/m^*(x) - 1$. Lindsay [19] and Basu and Lindsay [2] discuss the efficiency and robustness properties of these estimators.

If $G(\delta) = \frac{1}{\lambda(1+\lambda)}\left\{(1+\delta)^{(\lambda+1)} - 1\right\}$ we obtain the class of power divergence measures. Notice that we have $G(0) = 0$. Different values of $\lambda$ offer different measures; for example, when $\lambda = -2$ we obtain Neyman's chi-squared divided by 2 measure, while $\lambda = -1, -1/2$ return the Kullback-Leibler and Hellinger distances, respectively.

Under appropriate conditions, (1) and (2) can be written as

$$\sum A(\delta(x)) m_{\boldsymbol{\beta}}(x) = 0,$$

or

$$\int A(\delta(x)) \nabla m^*_{\boldsymbol{\beta}}(x)\, dx = 0,$$

where $A(\delta) = (\delta+1)G'(\delta) - G(\delta)$ and the prime denotes differentiation with respect to $\delta$.

Lindsay [19] has shown that the structural characteristics of the function $A(\delta)$ play an important role in the robustness and efficiency properties of these methods. Furthermore, without loss of generality, we can center and rescale $A(\delta)$, and define the RAF as follows.

**Definition 1** (Lindsay [19])**.** *Let $A(\delta)$ be an increasing and twice differentiable function on $[-1,\infty)$ defined as*

$$\begin{aligned} A(\delta) &= (\delta+1)G'(\delta) - G(\delta), \\ A(0) &= 0, \\ A'(0) &= 1, \end{aligned}$$

*where $G$ is strictly convex and twice differentiable with respect to $\delta$ on $[-1,\infty)$ with $G(0) = 0$. Then, $A(\delta)$ is called residual adjustment function.*

**Remark 1.** *Since $A'(\delta) = (1+\delta)G''(\delta)$, the second order differentiability of $G$, in addition to its strict convexity, implies that $A(\delta)$ is strictly increasing function of $\delta$ on $[-1,\infty)$. Thus, we can define $A(\delta)$ as above without changing the solutions of the aforementioned estimating equations in the discrete case (see Lindsay [19], p. 1089). In the continuous case, such standardization does not change the estimating properties of the associated disparities (see Basu and Lindsay [2], p. 687).*

Two fundamental and at the same time conflicting goals in robust statistics are the goals of robustness and efficiency. In the traditional literature on robustness, first order efficiency is sacrificed and, instead, safety of the estimation or testing method against outliers is guaranteed. Here, one adheres to the notion that information about robustness of a method is carried by the influence function. In our setting, using the influence function to characterize the robustness properties of the associated estimation procedures is misleading. Instead, the shape of the RAF, $A(\cdot)$, provides information to the extent of which our procedures can be characterized as robust. The interested reader is directed to Lindsay [19] for further discussion on this topic.

## 3. Pearson Residual Systems

In this section, we define various Pearson residuals, appropriate for the measurement scale of the data. We introduce our notation first.

Let $(y_i, \boldsymbol{x}_i),\ i = 1, 2, \ldots, n$ be realizations from $n$ independent and identically distributed random variables that follow a distribution with density $m_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$. Recall that we use the word density to denote a general probability function, independently of whether the random variables $\boldsymbol{X}, Y$ are discrete, continuous or mixed. In what follows, we define different Pearson residual systems that account for the measurement scale of the data and study their properties.

**Case 1:** *Both X and Y are discrete.*
In this case, the pairs $(y_i, \boldsymbol{x}_i)$ follow a discrete probability mass function $m_{\boldsymbol{\beta}}(\boldsymbol{x}_i, y_i)$. Define the Pearson residual as

$$\delta(x, y) = \frac{\frac{n_{x,y}}{n}}{m_{\boldsymbol{\beta}}(y|x)\boldsymbol{\pi}_x} - 1,$$

where $\boldsymbol{\pi}_x = P(X = x) = g(x)$, and $n_{x,y}$ is the number of observations in the cell with $Y = y$ and $X = x$.

Note that this definition of the Pearson residual is nonparametric on the discrete support of $X$. In the case of regression, one can carry out a semiparametric argument to obtain the estimators of the vector $\boldsymbol{\beta}$ and $\boldsymbol{\pi}_x$.

We now establish that, under correct model specification, the residual $\delta(x, y)$ converges, almost surely, to zero.

**Proposition 1.** *When the model is correctly specified and as $n \to \infty$,*

$$\delta(x, y) \xrightarrow{a.s.} 0.$$

**Proof.** Write

$$\begin{aligned} \delta(x, y) &= \frac{\frac{n_{x,y}}{n}}{m_{\boldsymbol{\beta}}(y|x)\boldsymbol{\pi}_x} - 1 \\ &= \frac{\frac{n_{x,y}}{n_x} \cdot \frac{n_x}{n}}{m_{\boldsymbol{\beta}}(y|x)\boldsymbol{\pi}_x} - 1. \end{aligned}$$

Then

$$\begin{aligned} \frac{n_x}{n} &= \frac{\text{(\# of observations in the sample equal to x)}}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} I(x_i = x), \end{aligned}$$

where $I(\cdot)$ is the indicator function. Furthermore,

$$E\left[\frac{1}{n} I(X_i = x)\right] = P(X = x) < \infty,$$

and by the strong law of large numbers

$$\frac{n_x}{n} \xrightarrow[n\to\infty]{a.s.} E[I(X = x)] = P(X = x) = \boldsymbol{\pi}_x.$$

Similarly,

$$\frac{n_{x,y}}{n_x} \xrightarrow{a.s.} m_{\boldsymbol{\beta}}(y|x),$$

therefore

$$\delta(x, y) \xrightarrow[n\to\infty]{a.s.} 0$$

under correct model specification. □

**Case 2:** *Y is continuous and X is discrete.*
This is the case in some *ANOVA* models. We can still define the Pearson residual in this setting as

$$\delta(x,y) = \frac{f_n(y,x)}{m_{\boldsymbol{\beta}}(y,x)} - 1,$$

where

$$\begin{aligned} f_n(y,x) &= f_n^*(y|x)g(x) \\ &= \Big\{ \int k(y,t,h)\; d\hat{F}_n(t|x) \Big\} \frac{n_x}{n} \end{aligned}$$

and

$$\begin{aligned} m_{\boldsymbol{\beta}}(y,x) &= m_{\boldsymbol{\beta}}^*(y|x)g(x) \\ &= \Big\{ \int k(y,t,h)\; dM_{\boldsymbol{\beta}}(t|x) \Big\} \boldsymbol{\pi}_x. \end{aligned}$$

Then,

$$\delta(x,y) = \frac{f_n^*(y|X=x)\frac{n_x}{n}}{m_{\boldsymbol{\beta}}^*(y|X=x)\boldsymbol{\pi}_x} - 1.$$

**Proposition 2.** *Assume the model is correctly specified and $k(y,t,h)$ is a continuous function. Then,*

$$\delta(x,y) \xrightarrow[n\to\infty]{a.s.} 0.$$

**Proof.** Under the strong law of large numbers

$$\frac{n_x}{n} \xrightarrow[n\to\infty]{a.s.} \boldsymbol{\pi}_x.$$

Under the correct model specification, continuity of the kernel function and the fact that $\hat{F}_n$ converges completely to $F$ (implication of Glivenko-Cantelli theorem),

$$\lim_{n\to\infty} \int k(y;t,h)\; d\hat{F}_n(t|x) \to \int k(y;t,h)\; dF(t|x) = \int k(y;t,h)\; dM_{\boldsymbol{\beta}}(t|x) = m_{\boldsymbol{\beta}}^*(y|x)$$

(extension of Helly-Bray lemma). Therefore,

$$\frac{\frac{n_x}{n}\, f_n^*(y|x)}{\boldsymbol{\pi}_x\, m_{\boldsymbol{\beta}}^*(y|x)} \xrightarrow{a.s.} \frac{\boldsymbol{\pi}_x}{\boldsymbol{\pi}_x} \cdot \frac{m_{\boldsymbol{\beta}}^*(y|x)}{m_{\boldsymbol{\beta}}^*(y|x)} = 1$$

and hence

$$\delta(x,y) = \frac{\frac{n_x}{n}\, f_n^*(y|x)}{\boldsymbol{\pi}_x\, m_{\boldsymbol{\beta}}^*(y|x)} - 1 \xrightarrow{a.s.} 1 - 1 = 0.$$

□

**Case 3:** *Y is continuous and X is continuous.*
In this case, the pairs $(y_i, x_i)$ follow a continuous probability distribution. The Pearson residual is then defined as

$$\delta(x,y) = \frac{f_n^*(y,x)}{m_{\boldsymbol{\beta}}^*(y,x)} - 1,$$

where

$$f_n^*(x,y) = \int k(x,y;t_1,t_2)\, d\hat{F}_n(t_1,t_2),$$
$$m_{\boldsymbol{\beta}}^*(x,y) = \int k(x,y;t_1,t_2) m_{\boldsymbol{\beta}}(t_1,t_2)\, dt_1 dt_2.$$

As an example, we take the linear regression model with random carriers $X$, and $\epsilon_i \sim N(0,1)$. Furthermore, assume that the random carriers follow a normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In this case, $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ and the quantities $z_i = (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})/\sigma$ are independent, identically distributed random variables when $\boldsymbol{\beta}$ represents the vector of true parameters. Hence, the $z_i$'s represent realizations of a random variable $Z$ that has a completely known density $f(z)$. Thus,

$$m_{\boldsymbol{\beta}}(x,y) = m_{\boldsymbol{\beta}}(z|x) \cdot g(x), \qquad z = (\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\beta})/\sigma$$

and hence

$$m_{\boldsymbol{\beta}}^*(x,y) = m_{\boldsymbol{\beta}}^*(y - x^T\boldsymbol{\beta}|X = x) g^*(x),$$
$$m_{\boldsymbol{\beta}}^*(y - x^T\boldsymbol{\beta}|X = x) = m_{\boldsymbol{\beta}}^*(z|x) = \int k(z,t,h)\, dM_{\boldsymbol{\beta}}(t|x),$$
$$g^*(x) = \int k'(x,t',h') g(t')\, dt'.$$

The kernel $k(z,t,h)$ is selected so that it facilitates easy computation. Kernels that do not entail loss of information when they are used to smooth the assumed parametric model are called transparent kernels (Basu and Lindsay [2]). Basu and Lindsay [2] provide a formal definition of transparent kernels and an insightful discussion on the point of why transparent kernels do not exhibit information loss when convoluted with the hypothesized model (see Section 3.1 of Basu and Lindsay [2]).

## 4. Estimating Equations

In this section, we concentrate on cases 1, 2 presented in the previous section. We carefully outline the optimization problems and discuss the associated estimating equations for these two cases. The case where both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are continuous has been discussed in the literature, see, for example, Markatou et al. [21].

**Case 1:** *Both* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are discrete.*
In this case, the minimum distance estimators of the parameter vector $\boldsymbol{\beta}$ and $\pi_x$ are obtained by solving the following optimization problem

$$\min_{\boldsymbol{\beta},\pi_x} \rho(d, m_{\boldsymbol{\beta}}) \tag{3}$$

subject to

$$\sum_x \pi_x = 1.$$

Optimization problem (3) is equivalent to the problem

$$\min \sum_{x,y} G(\delta(x,y)) m_{\boldsymbol{\beta}}(x,y)$$

subject to

$$\sum_x \pi_x = 1.$$

The class of $G$ functions that we use creates distances that belong in the family of $\phi$-divergences.

**Proposition 3.** *The estimating equations for $\boldsymbol{\beta}$ and $\pi_x$ are given as:*

$$\sum_{x,y} w(\delta(x,y))\, n_{x,y}\, u(y|x;\boldsymbol{\beta}) = 0,$$
$$\sum_{x,y} w(\delta(x,y))\, n_{x,y} \left\{ \frac{I(X=x)}{\pi_x} - 1 \right\} = 0. \tag{4}$$

*The function $w(\delta(x,y))$ is a weight function, such that $0 \le w(\delta(x,y)) \le 1$, and it is defined as*

$$w(\delta(x,y)) = \min\left\{ \frac{[A(\delta(x,y))+1]^+}{\delta(x,y)+1}, 1 \right\}$$

*with $[\cdot]^+$ indicating the positive part of the function $A(\delta(x,y))+1$.*

**Proof.** The main steps of the proof are provided in the Appendix A.1. □

**Remark 2.**

1. *The above two estimating equations can be solved with respect to $\boldsymbol{\beta}$ and $\pi_x$. In an iterative algorithm, we can solve the second equation (4) explicitly for $\pi_x$ to obtain*

$$\pi_x = \frac{\sum_y w(\delta(x,y))n_{x,y}}{\sum_{x,y} w(\delta(x,y))n_{x,y}}.$$

   *This means that if the model does not fit any of the $y$, observed at a particular $x$ well, the weight for this $x$ will drop as well.*
2. *When $A(\delta(x,y)) = \delta(x,y)$ the corresponding estimating equation for $\boldsymbol{\beta}$ becomes $\sum_{x,y} n_{x,y} u(y|x;\boldsymbol{\beta}) = 0$ and the MLE is obtained. This is because the corresponding weight function $w(\delta(x,y)) = 1$. In this case, the estimating equations for the $\pi_x$s become $\sum n_{x,y}\left[\frac{I(X=x)}{\pi_x} - 1\right] = 0$, the estimating equations for the MLEs of $\pi_x$.*
3. *The Fisher consistency property of the function that introduces the estimates guarantees that the expectation of the corresponding estimating function is 0, under the correct model specification.*

**Case 2:** *$\boldsymbol{Y}$ is continuous and $\boldsymbol{X}$ is discrete.*
In this case, the estimates of the parameters $\boldsymbol{\beta}$ and $\pi_x$ are obtained by solving the following optimization problem

$$\min_{\beta,\pi_x} \sum_x \int G(\delta(x,y)) m^*_{\boldsymbol{\beta}}(y,x)dy$$

subject to

$$\sum_x \pi_x = 1.$$

In general $m^*_{\boldsymbol{\beta}}(y,x) = m^*_{\boldsymbol{\beta}}(y|x)\pi_x$; in the case where $y, x$ are independent $m^*_{\boldsymbol{\beta}}(y,x) = m^*_{\boldsymbol{\beta}}(y)\pi_x$, and the optimization problem stated above is equivalent to

$$\min_{\beta,\pi_x} \sum_x \pi_x \int G(\delta(x,y)) m^*_{\boldsymbol{\beta}}(y)dy \tag{5}$$

subject to

$$\sum_x \pi_x = 1.$$

**Proposition 4.** *The estimating equations for $\boldsymbol{\beta}$ and $\pi_x$ in the case of independence of $y, x$ are given as follows:*

$$\sum_x \pi_x \int A(\delta(x,y)) \nabla_{\boldsymbol{\beta}}\, m^*_{\boldsymbol{\beta}}(y)dy = 0,$$
$$\sum_x \pi_x \int A(\delta(x,y)) \Big[\frac{I(X=x)}{\pi_x} - 1\Big]\, m^*_{\boldsymbol{\beta}}(y)dy = 0, \tag{6}$$

*where $A(\delta)$ is the residual adjustment function (RAF) that corresponds to the function $G$, and $G'(\delta)$ is the derivative of $G$ with respect to $\delta$.*

**Proof.** Straightforward, after differentiating the Lagrangian with respect to $\boldsymbol{\beta}$ and $\pi_x$. □

**Case 3:** $\boldsymbol{Y}$ *is continuous and* $\boldsymbol{X}$ *is continuous.*
In this case, we refer the reader to Basu and Lindsay [2].

## 5. Robustness Properties

Hampel et al. [29] and Hampel [30,31] define robust statistics as the "statistics of approximate parametric models", and introduce one of the fundamental tools of robust statistics, the concept of the influence function, in order to investigate the behavior of a statistic $T_n$ expressed as a functional $T(G)$. The influence function is a heuristic tool with the intuitive interpretation of measuring the bias caused by an infinitesimal contamination at a point $x$ on the estimate standardized by the mass of contamination. Its formal definition is as follows:

**Definition 2.** *The influence function of a functional $T$ at the distribution $F$ is given as*

$$IF(x;T,F) = \lim_{t \to 0} \frac{T((1-t)F + t\Delta_x) - T(F)}{t},$$

*in those $x \in \mathcal{X}$ where the limit exists, $0 \le t \le 1$ and $\Delta_x$ is the Dirac measure defined as*

$$\Delta_x(u) = \begin{cases} 1, & u = x, \\ 0, & u \neq x. \end{cases} \tag{7}$$

If an estimator has a bounded influence function, the estimator is considered to be robust to outliers, that is data which is away from the pattern set by the majority of the data. The effect of bounding the influence function is the sacrifice of efficiency; estimators with bounded influence function, while are not affected by outlying points, are not fully efficient under the correct model specification.

Our goal in calculating the influence function is to show the full efficiency of the proposed estimators. That is, the influence function of the proposed estimators, under correct model specification, equals the influence function of the corresponding maximum likelihood estimators. In our context, robustness of the estimators is quantified by the associated RAFs (see Lindsay [19] and Basu and Lindsay [2]).

In what follows, we will derive the influence function of the estimators for the parameter vector $\boldsymbol{\beta}$ in the case where both $y, x$ are discrete. Similar calculations provide the influence functions of estimators obtained under the remaining scenarios. To do so, we need to resort to the estimators' functional form, denoted by $\boldsymbol{\beta}_\epsilon$, with corresponding estimating equations

$$\sum_{s,t} w(\delta_\epsilon(s,t)) u(t|s;\boldsymbol{\beta}_\epsilon) d_\epsilon(s,t) = 0,$$

where $d_\epsilon(s,t) = (1-\epsilon)d(s,t) + \epsilon\Delta_{x,y}(s,t)$. The influence function is then obtained by differentiating the aforementioned estimating equations with respect to $\epsilon$ and then evaluating the derivative at $\epsilon = 0$.

**Proposition 5.** *The influence function of the $\boldsymbol{\beta}$ estimator is given by*

$$\boldsymbol{\beta}_0' = [A(d)]^{-1}B(x,y;d),$$

*where*

$$\begin{aligned} A(d) =& \sum_{s,t}[\delta_0(t)+1]w'(\delta_0(s,t))u(t|s;\boldsymbol{\beta}_0)u^T(t|s;\boldsymbol{\beta}_0)d(s,t) \\ & - \sum_{s,t} w(\delta_0(s,t))\nabla u(t|s;\boldsymbol{\beta}_0)d(s,t), \end{aligned}$$

$$\begin{aligned} B(x,y;d) =& \sum_{s,t}\Big[\frac{I(s=x,t=y)}{m_{\boldsymbol{\beta}_0}(t|s)\pi_s} - \frac{d(s,t)}{m_{\boldsymbol{\beta}_0}(t|s)\pi_s}w'(\delta_0(s,t))\Big]u(t|s;\boldsymbol{\beta}_0)d(s,t) \\ & - \sum_{s,t} w(\delta_0(s,t))u(t|s;\boldsymbol{\beta}_0)d(s,t) + w(\delta_0(x,y))u(t|s;\boldsymbol{\beta}_0), \end{aligned}$$

*with $u(t|s;\boldsymbol{\beta}) = \nabla \ln m_{\boldsymbol{\beta}}(t|s)$, and the subscript 0 indicates evaluation at a parametric model.*

**Proof.** The proof is obtained via straightforward differentiation and its main steps are provided in the Appendix A.2. □

**Proposition 6.** *Under the assumption that the model is correct, the influence function derived, reduces to the influence function of the MLE of $\boldsymbol{\beta}$.*

**Proof.** Under the assumption that the adopted model is the correct model, the density $d(s,t)$ is $m_{\boldsymbol{\beta}_0}(s,t)$, so that $\delta(s,t) = 0$. Now recall that $w(0) = 1$ and $w'(0) = 0$, so the expression $A(d)$ reduces to

$$\begin{aligned} A(d) &= -\sum_{s,t} \nabla u(t|s;\boldsymbol{\beta}_0)m_{\boldsymbol{\beta}_0}(s,t) \\ &= i(\boldsymbol{\beta},x,y). \end{aligned} \tag{8}$$

Furthermore, the expression $B(x,y;d)$ reduces to $u(y|x;\boldsymbol{\beta}_0)$, where we assume exchangeability of differentiation and integration and use the fact that $u(t|s;\boldsymbol{\beta}_0) = u(s,t;\boldsymbol{\beta}_0)$. Hence, the influence function is given as

$$i^{-1}(\boldsymbol{\beta};x,y)u(y|x;\boldsymbol{\beta}_0),$$

which is exactly the influence function of the MLE. Therefore, full efficiency is preserved under the model. □

## 6. Asymptotic Properties

In what follows, we establish asymptotic normality of the estimators in the case of discrete variables. The techniques for obtaining asymptotic normality in the mixed-scale case are similar and not presented here.

**Case 1:** *Both $\mathbf{X}$ and $\mathbf{Y}$ are discrete.*
Recall that the $k-$th estimating equation is given as $\sum_{x,y} w(\delta_{\boldsymbol{\beta}}(x,y))n_{x,y}u_k(y|x;\boldsymbol{\beta}) = 0$, which can be expanded in Taylor series in the neighborhood of the true parameter $\boldsymbol{\beta}_0$ to obtain:

$$\frac{1}{n}\sum_{x,y} w(\delta_{\boldsymbol{\beta}}(x,y))n_{x,y}u_k(y|x;\boldsymbol{\beta}) \cong A_n + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T B_n + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T C_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0), \tag{9}$$

where

$$A_n = \frac{1}{n}\sum_{x,y} w(\delta_{\boldsymbol{\beta}}(x,y))n_{x,y}u_k(y|x;\boldsymbol{\beta}_0),$$
$$B_n = \nabla_{\boldsymbol{\beta}}\left\{\frac{1}{n}\sum_{x,y} w(\delta_{\boldsymbol{\beta}}(x,y))n_{x,y}u_k(y|x;\boldsymbol{\beta})\right\}\Big|_{\boldsymbol{\beta}_0}, \tag{10}$$

$C_n$ is a $p \times p$ Hessian matrix whose $(t,e)-$th element is given as

$$\frac{\partial^2}{\partial\boldsymbol{\beta}_t\partial\boldsymbol{\beta}_e}\left\{\frac{1}{n}\sum_{x,y} w(\delta_{\boldsymbol{\beta}}(x,y))n_{x,y}u_k(y|x;\boldsymbol{\beta})\right\}\Big|_{\boldsymbol{\beta}_0}.$$

Under assumptions 1–8, listed in the Appendix A.3, we have the following theorem.

**Theorem 1.** *The minimum disparity estimators of the parameter vector $\boldsymbol{\beta}$ are asymptotically normal with asymptotic variance $I^{-1}(\boldsymbol{\beta}_0)$, where $I(\cdot)$ indicates the Fisher information matrix.*

## 7. Simulations

The simulation study presented below has two aims. The first one, is to indicate the versatility of the disparity methods for different data measurement scales. The second aim is to exemplify and study the robustness of these methods under different contamination scenarios.

**Case 1:** *Both X and Y are discrete.*
The Cressie-Read family of power divergence is given by

$$PWD(\boldsymbol{d},\boldsymbol{m}_{\boldsymbol{\beta}}) = \sum m_{\boldsymbol{\beta}}(x,y)\cdot\frac{[1+\delta(x,y)]^{\lambda+1}-1}{\lambda(\lambda+1)} = \sum d(x,y)\cdot\frac{[d(x,y)/m_{\boldsymbol{\beta}}(x,y)]^{\lambda}-1}{\lambda(\lambda+1)},$$

where $d(x,y) = n_{x,y}/n$ is the proportion of observations with value $x,y$ and $m_{\boldsymbol{\beta}}(x,y) = m_{\boldsymbol{\beta}}(y|x)\pi_x$ is the density function of the model of interest.

To evaluate the performance of our algorithmic procedure, we use the following disparity measures, that is,

$$\begin{aligned}
&\textit{Likelihood disparity } (\lambda = 0):\\
&LD(\boldsymbol{d},\boldsymbol{m}_{\boldsymbol{\beta}}) = \sum d(x,y)\cdot\{\log[d(x,y)/m_{\boldsymbol{\beta}}(x,y)]\},\\
&\textit{Twice-squared Hellinger's } (\lambda = -1/2):\\
&HD(\boldsymbol{d},\boldsymbol{m}_{\boldsymbol{\beta}}) = 2\cdot\sum\left[\sqrt{d(x,y)}-\sqrt{m_{\boldsymbol{\beta}}(x,y)}\right]^2,\\
&\textit{Pearson's chi-squared divided by 2 } (\lambda = 1):\\
&PCS(\boldsymbol{d},\boldsymbol{m}_{\boldsymbol{\beta}}) = \sum\frac{[d(x,y)-m_{\boldsymbol{\beta}}(x,y)]^2}{2\cdot m_{\boldsymbol{\beta}}(x,y)},\\
&\textit{Symmetric chi-squared } \left(G(\delta(x,y)) = \frac{2[\delta(x,y)]^2}{\delta(x,y)+2}\right):\\
&SCS(\boldsymbol{d},\boldsymbol{m}_{\boldsymbol{\beta}}) = 2\cdot\sum\frac{[m_{\boldsymbol{\beta}}(x,y)-d(x,y)]^2}{[m_{\boldsymbol{\beta}}(x,y)+d(x,y)]}.
\end{aligned}$$

The data are generated in four different ways using three different sample sizes $N$, say $N = 100$; $N = 1000$ and $N = 10{,}000$. The data format used can be represented in a $5 \times 5$ contingency table, with $n_{i,j}$, $i = 1,2,\ldots,5$; $j = 1,2,\ldots,5$ denoting the counts in the $ij$-th cell, $n_{i\bullet}$ and $n_{\bullet j}$ representing the row and column totals, respectively. Furthermore, the variable $x$ indicates columns, while $y$ indicates the rows. In each of the aforementioned cases/scenarios, 10,000 tables were generated and that corresponds to the number of Monte Carlo (MC) replications. Our purpose is to get the mean values of the estimates of the parameters $m_{\beta}(y|x)$'s and $\pi_x$'s along with their corresponding standard deviations (SDs). Notice that, in this setting, the estimation of $\pi_x$ and $m_{\beta}(y|x)$ is completely nonparametric, that is, no model is assumed for estimating the marginal probabilities of $X$ and $Y$.

The table was generated by using either a fixed total sample size $N$ or fixed marginal probabilities. These two data generating schemes imply two different sampling schemes that could have generated the data with consequences for the probability model one would use. For example, with fixed total sample size the distribution of the counts is multinomial, or if the row margin is fixed in advance the distribution of the counts is a product binomial distribution. In the former case of fixed $N$, we explored two different scenarios: a balanced and an imbalanced one. The imbalanced scenario allows for the presence of one zero cell in the contingency table, whereas the balanced scenario does not. In the latter case of fixed marginal probabilities, the row marginal probabilities ($m_{\beta}(y|x)$'s) were fixed, while the column marginals ($\pi_x$'s) were randomly chosen and these values were used to obtain the contingency table. In this case, we also explored a balanced and an imbalanced scenario based on whether the row marginal probabilities were chosen so that to be equal to each other or not, respectively.

Specifically, under Scenario Ia, where the total sample size $N$ was fixed and the balanced design was exploited, none of the $n_{ij}$'s ($n_{ij} \neq 0$, $\forall\, i, j = 1, 2, 3, 4, 5$) was set equal to zero, with equal row and column marginal probabilities. Table 1 presents the mean of 10,000 estimates and the corresponding SDs for all four distances ($PCS$, $HD$, $SCS$, $LD$) when $N$ is fixed under the balanced scenario. Table 1 clearly shows that all distances provide estimates approximately equal to 0.200 regardless of the sample size used. Furthermore, as the sample size increases, the SDs decrease noticeably.

In Scenario IIa, where the total sample size $N$ was fixed and the contingency table was structured using the imbalanced design, the presence of a zero cell ($n_{11} = 0$) was allowed. The results of this scenario are presented in Table 2, where the estimates were calculated exploiting all disparity measures. For the $LD$, $n_{11}$ was set equal to $10^{-8}$. The presence of zero cells in contingency tables has a large history in the relevant literature on contingency tables analysis, where several options are provided for the analysis of these tables (Fienberg [32], Agresti [33], Johnson and May [34], Poon et al. [35]). From Table 2, one could infer that the different distances handle differently the zero cell. This difference is reflected in the estimate of $\hat{m}_{\beta(y_1|x)} = \hat{m}_{\beta_1}$, because it is affected by the zero value of $n_{11}$. The strongest control is provided by the Hellinger and symmetric chi-squared distances. All distances estimate the parameters $\pi_{x_i}$ similarly, with the bias in their estimation been between 2.7% and 5.2%. The SDs are almost the same for all distances per estimate and their values are ameliorated for $N = 10{,}000$.

A referee suggested that in certain cases interest may be centered on smaller samples. We generated $2 \times 3$ tables with fixed total sample size of 50 and 70 observations. Tables 3 and 4 describe the results when the contingency tables were generated under a balanced and an imbalanced design with associated respective Scenarios Ib and IIb. More precisely, Table 3 presents the estimators of the marginal row and column probabilities obtained when $PC$, $HD$, $SCS$ and $LD$ distances are used. We notice that the increase in the sample size provides for a decrease in the overall absolute bias in estimation, defined as $\sum_{\ell=1}^{L} |\hat{\theta}_\ell - \theta_{0,\ell}|$, where $\hat{\theta}_\ell$ is the estimate of the $\ell$-th component of an $L \times 1$ vector $\boldsymbol{\theta}$ and $\theta_{0,\ell}$ is the corresponding true value. In our case, $\boldsymbol{\theta}^T = (m_{\beta_1}, m_{\beta_2}, \pi_{x_1}, \pi_{x_2}, \pi_{x_3})$. This observation applies to all distances used in our calculations. Table 4 presents results associated with the imbalanced case. The generated $2 \times 3$ tables contain two empty cells ($n_{12} = n_{21} = 0$). Once again, for calculating the $LD$, cells $n_{12} = n_{21} = 10^{-8}$. We notice that the bias associated with the estimates is rather large for all the distances, and an increased sample size does not alleviate the observed bias. Basu and Basu [9] have proposed an empty cell penalty for the minimum power-divergence estimators. This penalty leads to estimators with improved small sample properties. See also Alin and Kurt [36] for a discussion of the need of penalization in small samples.

Table 5 provides the results obtained under Scenario III. In this case, the parameter estimates were calculated using the $PCS$, $HD$, $SCS$ and $LD$ distances when the $5 \times 5$ contingency table was constructed by fixing the row marginal probabilities so that they were all set at 0.20, that is, $(0.20, 0.20, 0.20, 0.20, 0.20)$. The column marginals were randomly chosen in the interval $[0, 1]$ and summed to 1. In this case, the produced column marginal probabilities were $(0.1472, 0.2365, 0.3196, 0.2370, 0.0597)$. The simulation study reveals that the estimates of the parameters $m_{\boldsymbol{\beta}}(y|x)$'s and $\pi_x$'s do not differ substantially from the respective row and column marginal probabilities for any of the four distances utilized. The SDs are approximately the same and they get lower values for larger $N$.

Finally, in Table 6 the data generation was done by exploiting Scenario IV, that is, by having fixed the row marginal probabilities, which were not equal to each other; while, the column marginals were randomly chosen in the interval $[0, 1]$ so that they sum to 1. In particular, the row marginal probabilities were fixed at values $(0.04, 0.20, 0.20, 0.20, 0.36)$, while the column marginals used were $(0.2171, 0.1676, 0.2347, 0.1178, 0.2628)$. When $N = 100$, the value of $\hat{m}_{\beta}(y_1|x) = \hat{m}_{\beta_1}$ is not approximately 0.07 and not equal to 0.04 for all distances. However, when $N = 1000$ or $N = 10{,}000$, we get better estimates irrespectively of the disparity measure choice. The SDs are approximately the same and they become smaller as the sample size increases.

We also notice from Tables 1, 5 and 6 that in all cases the standard deviation associated with the estimates obtained when we use other than likelihood distances, is approximately the same with the standard deviation that corresponds to the likelihood estimates, thereby showing the asymptotic efficiency of the disparity estimators.

All calculations were performed using the $R$ language. Given that the problem described in this section can be viewed as a general non-linear optimization problem, the `solnp` function of the `Rsolnp` package (Ye [37]) was used to obtain the aforementioned estimates. For our calculations, we tried using a variety of different initial values ($\hat{\pi}_x^{(0)}$'s and $\hat{m}_{\boldsymbol{\beta}}^{(0)}(y|x)$'s); we notice that no matter how the initial values were chosen, the estimates were always pretty similar and very close to the observed values ($n_{i\bullet}/N$ and $n_{\bullet j}/N$ for $i, j = 1, 2, 3, 4, 5$). Only the number of iterations needed for convergence is slightly affected. Consequently, random numbers from a Uniform distribution in the interval $[0, 1]$ were set as initial values (which were not necessarily summing to 1). The `solnp` function has a built-in stopping rule and there was no need to set our own stopping rule. We only set the boundary constraints to be in the interval $[0, 1]$ for all estimates which were also subject to $\sum \pi_x = \sum m_{\boldsymbol{\beta}}(y|x) = 1$.

Other functions may also be used to obtain the estimates. For example, we used the `auglag` function of the `nloptr` package with local solvers "`lbfgs`" or "`SLSQP`" (Conn et al. [38], Birgin and Martínez [39]) which emulates Augmented Lagrangian multipliers. However, the convergence using the `solnp` function (the number of iterations was on average 2) was extremely faster than using the `auglag` function (the average number of iterations was approximately 100). For this reason, the results presented in Tables 1–6 were based only on the function `solnp`.

**Table 1.** Scenario Ia: Means and standard deviations (SDs) of 4 distances ($PCS, HD, SCS, LD$). A 5 × 5 contingency table was generated having fixed the total sample size $N$ under a balanced design with $n_{ij} \neq 0$, $\forall\, i, j = 1, 2, 3, 4, 5$. The number of Monte Carlo (MC) replications used is 10,000.

| $N$ | Statistical Distance | Summary | Estimates<br>Means and SDs over 10,000 Replications | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\hat{m}_{\beta_1}$ | $\hat{m}_{\beta_2}$ | $\hat{m}_{\beta_3}$ | $\hat{m}_{\beta_4}$ | $\hat{m}_{\beta_5}$ | $\hat{\pi}_{x_1}$ | $\hat{\pi}_{x_2}$ | $\hat{\pi}_{x_3}$ | $\hat{\pi}_{x_4}$ | $\hat{\pi}_{x_5}$ |
| 100 | PCS | Mean | 0.199 | 0.199 | 0.201 | 0.201 | 0.200 | 0.201 | 0.200 | 0.199 | 0.200 | 0.201 |
| | | SD | 0.038 | 0.041 | 0.039 | 0.039 | 0.039 | 0.038 | 0.038 | 0.037 | 0.038 | 0.038 |
| | HD | Mean | 0.199 | 0.200 | 0.200 | 0.200 | 0.201 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 |
| | | SD | 0.037 | 0.041 | 0.037 | 0.037 | 0.037 | 0.037 | 0.037 | 0.035 | 0.036 | 0.037 |
| | SCS | Mean | 0.199 | 0.201 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.199 | 0.200 | 0.201 |
| | | SD | 0.037 | 0.041 | 0.038 | 0.038 | 0.038 | 0.032 | 0.033 | 0.030 | 0.031 | 0.032 |
| | LD | Mean | 0.199 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.002 | 0.200 | 0.200 | 0.200 |
| | | SD | 0.035 | 0.039 | 0.036 | 0.036 | 0.036 | 0.035 | 0.036 | 0.036 | 0.034 | 0.035 |
| 1000 | PCS | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 |
| | | SD | 0.014 | 0.015 | 0.016 | 0.016 | 0.014 | 0.017 | 0.015 | 0.015 | 0.013 | 0.016 |
| | HD | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 |
| | | SD | 0.013 | 0.015 | 0.013 | 0.013 | 0.013 | 0.013 | 0.012 | 0.012 | 0.012 | 0.013 |
| | SCS | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 |
| | | SD | 0.014 | 0.015 | 0.013 | 0.013 | 0.013 | 0.008 | 0.009 | 0.011 | 0.012 | 0.008 |
| | LD | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 |
| | | SD | 0.013 | 0.015 | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 | 0.012 | 0.012 | 0.013 |
| 10,000 | PCS | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 |
| | | SD | 0.008 | 0.007 | 0.006 | 0.006 | 0.009 | 0.010 | 0.010 | 0.007 | 0.008 | 0.006 |
| | HD | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 |
| | | SD | 0.004 | 0.005 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 |
| | SCS | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 |
| | | SD | 0.004 | 0.005 | 0.004 | 0.004 | 0.004 | 0.007 | 0.005 | 0.008 | 0.008 | 0.004 |
| | LD | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 |
| | | SD | 0.004 | 0.005 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 |

**Table 2.** Scenario IIa Means and SDs of 4 distances ($PCS, HD, SCS, LD$). A 5 × 5 contingency table was generated having fixed the total sample size $N$ under an imbalanced design with $n_{11} = 0$. The number of MC replications used is 10,000.

| $N$ | Statistical Distance | Summary | Estimates<br>Means and SDs over 10,000 Replications | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\hat{m}_{\beta_1}$ | $\hat{m}_{\beta_2}$ | $\hat{m}_{\beta_3}$ | $\hat{m}_{\beta_4}$ | $\hat{m}_{\beta_5}$ | $\hat{\pi}_{x_1}$ | $\hat{\pi}_{x_2}$ | $\hat{\pi}_{x_3}$ | $\hat{\pi}_{x_4}$ | $\hat{\pi}_{x_5}$ |
| 100 | PCS | Mean | 0.052 | 0.197 | 0.198 | 0.198 | 0.355 | 0.165 | 0.173 | 0.172 | 0.245 | 0.245 |
| | | SD | 0.028 | 0.045 | 0.044 | 0.044 | 0.053 | 0.041 | 0.039 | 0.044 | 0.044 | 0.047 |
| | HD | Mean | 0.026 | 0.202 | 0.202 | 0.202 | 0.368 | 0.156 | 0.168 | 0.168 | 0.254 | 0.254 |
| | | SD | 0.019 | 0.049 | 0.045 | 0.045 | 0.054 | 0.041 | 0.042 | 0.041 | 0.046 | 0.049 |
| | SCS | Mean | 0.033 | 0.209 | 0.209 | 0.209 | 0.340 | 0.166 | 0.172 | 0.171 | 0.245 | 0.246 |
| | | SD | 0.022 | 0.047 | 0.045 | 0.045 | 0.051 | 0.036 | 0.036 | 0.033 | 0.038 | 0.040 |
| | LD | Mean | 0.040 | 0.200 | 0.200 | 0.200 | 0.360 | 0.160 | 0.170 | 0.170 | 0.250 | 0.250 |
| | | SD | 0.020 | 0.043 | 0.040 | 0.040 | 0.048 | 0.037 | 0.038 | 0.036 | 0.042 | 0.044 |
| 1000 | PCS | Mean | 0.044 | 0.197 | 0.197 | 0.197 | 0.365 | 0.164 | 0.170 | 0.170 | 0.248 | 0.248 |
| | | SD | 0.011 | 0.017 | 0.014 | 0.014 | 0.018 | 0.013 | 0.014 | 0.013 | 0.015 | 0.015 |
| | HD | Mean | 0.034 | 0.203 | 0.202 | 0.202 | 0.359 | 0.156 | 0.170 | 0.170 | 0.252 | 0.252 |
| | | SD | 0.005 | 0.015 | 0.013 | 0.013 | 0.016 | 0.011 | 0.012 | 0.012 | 0.013 | 0.014 |
| | SCS | Mean | 0.038 | 0.210 | 0.210 | 0.210 | 0.332 | 0.166 | 0.169 | 0.169 | 0.248 | 0.248 |
| | | SD | 0.006 | 0.015 | 0.014 | 0.014 | 0.016 | 0.014 | 0.013 | 0.011 | 0.013 | 0.014 |
| | LD | Mean | 0.040 | 0.200 | 0.200 | 0.200 | 0.360 | 0.160 | 0.170 | 0.170 | 0.250 | 0.250 |
| | | SD | 0.006 | 0.015 | 0.013 | 0.013 | 0.016 | 0.012 | 0.012 | 0.011 | 0.013 | 0.014 |
| 10,000 | PCS | Mean | 0.044 | 0.197 | 0.196 | 0.196 | 0.367 | 0.164 | 0.170 | 0.170 | 0.248 | 0.248 |
| | | SD | 0.002 | 0.006 | 0.007 | 0.007 | 0.010 | 0.007 | 0.006 | 0.005 | 0.007 | 0.008 |
| | HD | Mean | 0.034 | 0.203 | 0.202 | 0.202 | 0.359 | 0.156 | 0.171 | 0.171 | 0.252 | 0.252 |
| | | SD | 0.002 | 0.005 | 0.004 | 0.004 | 0.005 | 0.004 | 0.004 | 0.004 | 0.004 | 0.005 |
| | SCS | Mean | 0.038 | 0.210 | 0.210 | 0.210 | 0.332 | 0.166 | 0.169 | 0.169 | 0.248 | 0.248 |
| | | SD | 0.002 | 0.005 | 0.004 | 0.004 | 0.005 | 0.007 | 0.006 | 0.004 | 0.006 | 0.006 |
| | LD | Mean | 0.040 | 0.200 | 0.200 | 0.200 | 0.360 | 0.160 | 0.170 | 0.170 | 0.250 | 0.250 |
| | | SD | 0.002 | 0.005 | 0.004 | 0.004 | 0.005 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 |

**Table 3.** Scenario Ib: Means and Biases of 4 distances $(PCS, HD, SCS, LD)$. A 2 × 3 contingency table was generated having fixed the total sample size $N$ under a balanced design with $n_{ij} \neq 0,\ \forall\ i = 1, 2,\ j = 1, 2, 3$. The number of MC replications used is 10,000.

| ***N*** | **Statistical Distance** | **Summary** | **Estimates Means and Biases over 10,000 Replications** | | | | |
|---|---|---|---|---|---|---|---|
| | | | $\hat{m}_{\beta_1}$ | $\hat{m}_{\beta_2}$ | $\hat{\pi}_{x_1}$ | $\hat{\pi}_{x_2}$ | $\hat{\pi}_{x_3}$ |
| 50 | PCS | Mean | 0.5008 | 0.4992 | 0.3339 | 0.3336 | 0.3325 |
| | | Abs.Biases | 0.0008 | 0.0008 | 0.0006 | 0.0003 | 0.0009 |
| | | Overall Bias | | | 0.0034 | | |
| | HD | Mean | 0.5008 | 0.4992 | 0.3339 | 0.3335 | 0.3326 |
| | | Abs.Biases | 0.0008 | 0.0008 | 0.0006 | 0.0002 | 0.0007 |
| | | Overall Bias | | | 0.0031 | | |
| | SCS | Mean | 0.5007 | 0.4993 | 0.3338 | 0.3335 | 0.3326 |
| | | Abs.Biases | 0.0007 | 0.0007 | 0.0005 | 0.0002 | 0.0007 |
| | | Overall Bias | | | 0.0028 | | |
| | LD | Mean | 0.5008 | 0.4992 | 0.3339 | 0.3335 | 0.3326 |
| | | Abs.Biases | 0.0008 | 0.0008 | 0.0006 | 0.0002 | 0.0008 |
| | | Overall Bias | | | 0.0032 | | |
| 70 | PCS | Mean | 0.4998 | 0.5002 | 0.3333 | 0.3331 | 0.3337 |
| | | Abs.Biases | 0.0002 | 0.0002 | 0.0001 | 0.0003 | 0.0003 |
| | | Overall Bias | | | 0.0011 | | |
| | HD | Mean | 0.4998 | 0.5002 | 0.3333 | 0.3330 | 0.3336 |
| | | Abs.Biases | 0.0002 | 0.0002 | 0.0000 | 0.0003 | 0.0003 |
| | | Overall Bias | | | 0.0009 | | |
| | SCS | Mean | 0.4998 | 0.5002 | 0.3334 | 0.3331 | 0.3335 |
| | | Abs.Biases | 0.0002 | 0.0002 | 0.0000 | 0.0002 | 0.0002 |
| | | Overall Bias | | | 0.0008 | | |
| | LD | Mean | 0.4999 | 0.5001 | 0.3333 | 0.3330 | 0.3336 |
| | | Abs.Biases | 0.0001 | 0.0001 | 0.0000 | 0.0003 | 0.0003 |
| | | Overall Bias | | | 0.0009 | | |

**Table 4.** Scenario IIb: Means and Biases of 4 distances $(PCS, HD, SCS, LD)$. A 2 × 3 contingency table was generated having fixed the total sample size $N$ under an imbalanced design with $n_{12} = n_{21} = 0$. The number of MC replications used is 10,000.

| ***N*** | **Statistical Distance** | **Summary** | **Estimates Means and Biases over 10,000 Replications** | | | | |
|---|---|---|---|---|---|---|---|
| | | | $\hat{m}_{\beta_1}$ | $\hat{m}_{\beta_2}$ | $\hat{\pi}_{x_1}$ | $\hat{\pi}_{x_2}$ | $\hat{\pi}_{x_3}$ |
| 50 | PCS | Mean | 0.6391 | 0.3609 | 0.3489 | 0.2278 | 0.4234 |
| | | Abs.Biases | 0.0276 | 0.0276 | 0.0155 | 0.0611 | 0.0766 |
| | | Overall Bias | | | 0.2084 | | |
| | HD | Mean | 0.7815 | 0.2185 | 0.3346 | 0.0497 | 0.6157 |
| | | Abs.Biases | 0.1149 | 0.1149 | 0.0013 | 0.1170 | 0.1157 |
| | | Overall Bias | | | 0.4638 | | |
| | SCS | Mean | 0.6420 | 0.3580 | 0.3510 | 0.2726 | 0.3765 |
| | | Abs.Biases | 0.0247 | 0.0247 | 0.0176 | 0.1059 | 0.1235 |
| | | Overall Bias | | | 0.2964 | | |
| | LD | Mean | 0.6677 | 0.3323 | 0.3342 | 0.1660 | 0.4998 |
| | | Abs.Biases | 0.0010 | 0.0010 | 0.0009 | 0.0007 | 0.0002 |
| | | Overall Bias | | | 0.0038 | | |
| 70 | PCS | Mean | 0.6377 | 0.3623 | 0.3483 | 0.2297 | 0.4220 |
| | | Abs.Biases | 0.0290 | 0.0290 | 0.0150 | 0.0631 | 0.0780 |
| | | Overall Bias | | | 0.2141 | | |
| | HD | Mean | 0.7812 | 0.2188 | 0.3328 | 0.0491 | 0.6180 |
| | | Abs.Biases | 0.1145 | 0.1145 | 0.0005 | 0.1175 | 0.1180 |
| | | Overall Bias | | | 0.4650 | | |
| | SCS | Mean | 0.6395 | 0.3605 | 0.3505 | 0.2739 | 0.3756 |
| | | Abs.Biases | 0.0271 | 0.0271 | 0.0172 | 0.1072 | 0.1244 |
| | | Overall Bias | | | 0.3030 | | |
| | LD | Mean | 0.6657 | 0.3343 | 0.3331 | 0.1671 | 0.4998 |
| | | Abs.Biases | 0.0010 | 0.0010 | 0.0002 | 0.0004 | 0.0002 |
| | | Overall Bias | | | 0.0028 | | |

**Table 5.** Scenario III: Means and SDs of 4 distances ($PCS, HD, SCS, LD$). A 5 × 5 contingency table was generated having fixed the row marginal probabilities at (0.20, 0.20, 0.20, 0.20, 0.20). The number of MC replications used is 10,000.

| $N$ | Statistical Distance | Summary | Estimates<br>Means and SDs over 10,000 Replications | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\hat{m}_{\beta_1}$ | $\hat{m}_{\beta_2}$ | $\hat{m}_{\beta_3}$ | $\hat{m}_{\beta_4}$ | $\hat{m}_{\beta_5}$ | $\hat{\pi}_{x_1}$ | $\hat{\pi}_{x_2}$ | $\hat{\pi}_{x_3}$ | $\hat{\pi}_{x_4}$ | $\hat{\pi}_{x_5}$ |
| 100 | PCS | Mean | 0.199 | 0.200 | 0.200 | 0.200 | 0.201 | 0.153 | 0.230 | 0.302 | 0.229 | 0.086 |
| | | SD | 0.037 | 0.037 | 0.037 | 0.037 | 0.037 | 0.034 | 0.039 | 0.043 | 0.039 | 0.023 |
| | HD | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.147 | 0.230 | 0.311 | 0.230 | 0.082 |
| | | SD | 0.039 | 0.040 | 0.039 | 0.039 | 0.040 | 0.033 | 0.043 | 0.037 | 0.042 | 0.019 |
| | SCS | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.153 | 0.230 | 0.302 | 0.230 | 0.085 |
| | | SD | 0.039 | 0.085 | 0.038 | 0.038 | 0.038 | 0.033 | 0.039 | 0.043 | 0.039 | 0.022 |
| | LD | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.150 | 0.230 | 0.307 | 0.230 | 0.083 |
| | | SD | 0.038 | 0.038 | 0.038 | 0.038 | 0.038 | 0.033 | 0.041 | 0.045 | 0.040 | 0.019 |
| 1000 | PCS | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.148 | 0.236 | 0.319 | 0.236 | 0.061 |
| | | SD | 0.013 | 0.013 | 0.013 | 0.013 | 0.014 | 0.012 | 0.014 | 0.017 | 0.015 | 0.011 |
| | HD | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.147 | 0.237 | 0.320 | 0.237 | 0.059 |
| | | SD | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 | 0.011 | 0.014 | 0.015 | 0.014 | 0.008 |
| | SCS | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.148 | 0.236 | 0.319 | 0.237 | 0.060 |
| | | SD | 0.015 | 0.015 | 0.015 | 0.015 | 0.015 | 0.011 | 0.014 | 0.016 | 0.014 | 0.013 |
| | LD | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.147 | 0.237 | 0.320 | 0.237 | 0.059 |
| | | SD | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 | 0.011 | 0.014 | 0.015 | 0.013 | 0.008 |
| 10,000 | PCS | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.147 | 0.236 | 0.320 | 0.237 | 0.060 |
| | | SD | 0.006 | 0.006 | 0.006 | 0.006 | 0.006 | 0.008 | 0.006 | 0.011 | 0.006 | 0.008 |
| | HD | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.147 | 0.236 | 0.320 | 0.237 | 0.060 |
| | | SD | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.005 | 0.004 | 0.002 |
| | SCS | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.147 | 0.236 | 0.320 | 0.237 | 0.060 |
| | | SD | 0.005 | 0.005 | 0.005 | 0.005 | 0.005 | 0.004 | 0.006 | 0.008 | 0.006 | 0.008 |
| | LD | Mean | 0.200 | 0.200 | 0.200 | 0.200 | 0.200 | 0.147 | 0.236 | 0.320 | 0.237 | 0.060 |
| | | SD | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.005 | 0.005 | 0.005 | 0.002 |

**Table 6.** Scenario IV: Means and SDs of 4 distances ($PCS, HD, SCS, LD$). A 5 × 5 contingency table was generated having fixed the row marginal probabilities at (0.04, 0.20, 0.20, 0.20, 0.36). The number of MC replications used is 10,000.

| $N$ | Statistical Distance | Summary | Estimates<br>Means and SDs over 10,000 Replications | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\hat{m}_{\beta_1}$ | $\hat{m}_{\beta_2}$ | $\hat{m}_{\beta_3}$ | $\hat{m}_{\beta_4}$ | $\hat{m}_{\beta_5}$ | $\hat{\pi}_{x_1}$ | $\hat{\pi}_{x_2}$ | $\hat{\pi}_{x_3}$ | $\hat{\pi}_{x_4}$ | $\hat{\pi}_{x_5}$ |
| 100 | PCS | Mean | 0.074 | 0.197 | 0.197 | 0.197 | 0.335 | 0.214 | 0.173 | 0.228 | 0.132 | 0.253 |
| | | SD | 0.022 | 0.037 | 0.038 | 0.038 | 0.045 | 0.038 | 0.035 | 0.039 | 0.031 | 0.041 |
| | HD | Mean | 0.070 | 0.194 | 0.195 | 0.195 | 0.346 | 0.215 | 0.170 | 0.231 | 0.126 | 0.258 |
| | | SD | 0.015 | 0.039 | 0.039 | 0.039 | 0.048 | 0.041 | 0.037 | 0.042 | 0.030 | 0.044 |
| | SCS | Mean | 0.074 | 0.194 | 0.195 | 0.195 | 0.342 | 0.214 | 0.173 | 0.229 | 0.131 | 0.253 |
| | | SD | 0.015 | 0.039 | 0.039 | 0.039 | 0.048 | 0.038 | 0.035 | 0.040 | 0.030 | 0.041 |
| | LD | Mean | 0.071 | 0.195 | 0.196 | 0.196 | 0.342 | 0.214 | 0.172 | 0.230 | 0.128 | 0.256 |
| | | SD | 0.015 | 0.037 | 0.038 | 0.038 | 0.046 | 0.040 | 0.036 | 0.041 | 0.030 | 0.042 |
| 1000 | PCS | Mean | 0.042 | 0.200 | 0.200 | 0.200 | 0.358 | 0.217 | 0.168 | 0.234 | 0.119 | 0.262 |
| | | SD | 0.011 | 0.014 | 0.013 | 0.013 | 0.017 | 0.014 | 0.013 | 0.014 | 0.014 | 0.015 |
| | HD | Mean | 0.039 | 0.200 | 0.200 | 0.200 | 0.361 | 0.217 | 0.167 | 0.235 | 0.118 | 0.263 |
| | | SD | 0.006 | 0.013 | 0.013 | 0.013 | 0.015 | 0.013 | 0.012 | 0.013 | 0.010 | 0.014 |
| | SCS | Mean | 0.039 | 0.200 | 0.200 | 0.200 | 0.361 | 0.217 | 0.168 | 0.234 | 0.118 | 0.263 |
| | | SD | 0.007 | 0.013 | 0.013 | 0.013 | 0.016 | 0.016 | 0.013 | 0.014 | 0.010 | 0.015 |
| | LD | Mean | 0.040 | 0.200 | 0.200 | 0.200 | 0.360 | 0.217 | 0.167 | 0.235 | 0.118 | 0.263 |
| | | SD | 0.006 | 0.013 | 0.013 | 0.013 | 0.015 | 0.013 | 0.012 | 0.013 | 0.010 | 0.014 |
| 10,000 | PCS | Mean | 0.040 | 0.200 | 0.200 | 0.200 | 0.360 | 0.217 | 0.167 | 0.235 | 0.118 | 0.263 |
| | | SD | 0.008 | 0.005 | 0.007 | 0.007 | 0.009 | 0.006 | 0.005 | 0.005 | 0.007 | 0.006 |
| | HD | Mean | 0.040 | 0.200 | 0.200 | 0.200 | 0.360 | 0.217 | 0.167 | 0.235 | 0.118 | 0.263 |
| | | SD | 0.002 | 0.004 | 0.004 | 0.004 | 0.005 | 0.004 | 0.004 | 0.004 | 0.003 | 0.004 |
| | SCS | Mean | 0.040 | 0.200 | 0.200 | 0.200 | 0.360 | 0.217 | 0.167 | 0.235 | 0.118 | 0.263 |
| | | SD | 0.002 | 0.004 | 0.004 | 0.004 | 0.005 | 0.006 | 0.005 | 0.007 | 0.003 | 0.008 |
| | LD | Mean | 0.040 | 0.200 | 0.200 | 0.200 | 0.360 | 0.217 | 0.167 | 0.235 | 0.118 | 0.263 |
| | | SD | 0.002 | 0.004 | 0.004 | 0.004 | 0.005 | 0.004 | 0.004 | 0.005 | 0.003 | 0.005 |

**Case 2:** *X is discrete and* $\boldsymbol{Y}$ *is continuous*

In this section, we are interested in solving the optimization problem (5) when $X$ is discrete, $\boldsymbol{Y}$ is continuous and $X, \boldsymbol{Y}$ are independent of each other. To evaluate the performance of our procedure, we used Hellinger's distance, which in this case takes on the following form:

$$HD(\boldsymbol{f}^*, \boldsymbol{m}_{\boldsymbol{\beta}}^*) = \int \sum_x \left[ \sqrt{f_N^*(x, \boldsymbol{y})} - \sqrt{m_{\boldsymbol{\beta}}^*(x, \boldsymbol{y})} \right]^2 d\boldsymbol{y} = \int \sum_x \left[ \sqrt{f_{\boldsymbol{Y}}^*(\boldsymbol{y}) \cdot \frac{n_X}{N}} - \sqrt{m_X(x) \cdot m_{\boldsymbol{Y}}^*(\boldsymbol{y})} \right]^2 d\boldsymbol{y}.$$

The aim of this simulation is to obtain the minimum Hellinger distance estimators of $\pi_x$ and $\boldsymbol{\mu}$ assuming (without loss of generality) that $\sigma^2$ is known to be equal to 1. All calculations were performed in *R* language.

For this purpose, we generated mixed-type data of size $N$ using the package `OrdNor` (Amatya and Demirtas [40]). More precisely, the data are comprised of one categorical variable $X$ with three levels and probability vector $(1/3, 1/3, 1/3)$, while the continuous part is coming from a trivariate normal distribution; symbolic $\boldsymbol{Y} = (Y_1, Y_2, Y_3) \sim MVN_3(\boldsymbol{\mu}, \mathbf{I}_3)$, where $\boldsymbol{\mu}^T = (\mu_1, \mu_2, \mu_3)$. We used two different mean vectors: $\boldsymbol{\mu}^T = (0, 0, 0)$ and $\boldsymbol{\mu}^T = (0, 3, 6)$. The set of ordinal and normal variables were generated concurrently using an overall correlation matrix $\Sigma$, which consists of three components/sub-matrices: $\Sigma_{OO}$, $\Sigma_{ON}$ and $\Sigma_{NN}$, with $O$ and $N$ corresponding to "Ordinal" and "Normal" variables, respectively. More precisely, the overall correlation matrix $\Sigma$ used is the following

$$\Sigma = \begin{pmatrix} 1 & \rho_{ON} & \rho_{ON} & \rho_{ON} \\ \rho_{ON} & 1 & 0 & 0 \\ \rho_{ON} & 0 & 1 & 0 \\ \rho_{ON} & 0 & 0 & 1 \end{pmatrix},$$

where $\Sigma_{OO} = 1$, $\Sigma_{NN} = \mathbf{I}_3$, $\Sigma_{ON} = \begin{pmatrix} \rho_{ON} & \rho_{ON} & \rho_{ON} \end{pmatrix}$ and $\rho_{ON}$ represents the polyserial correlations for the $ON$ combinations (for more information on polyserial correlations refer to Olsson et al. [41]). Since $X, \boldsymbol{Y}$ were assumed to be independent, we set $\rho_{ON} = 0.0$. However, we also used weak correlations, say $\rho_{ON} = 0.1$ and 0.2, to investigate whether the estimates we receive in these cases remain reasonable.

The kernel function was the multivariate normal density $MVN_3(\mathbf{0}, \mathrm{H})$ with H being estimated by the data using the `kde` function of the `ks` package (Duong [42]), $m_{\boldsymbol{Y}}^*(\boldsymbol{y})$ represented the multivariate normal density $MVN_3(\boldsymbol{\mu}, \Sigma + \mathrm{H})$ and $m_X(x)$ was the multinomial mass function. This choice of smoothing parameter, stemmed from the fact that we were interested in evaluating the performance, in terms of robustness, of standard bandwidth selection.

To solve the optimization problem, the `solnp` function of the `Rsolnp` package (Ye [37]) was used. Specifically, the initial values set for the probabilities $\pi_{x_1}, \pi_{x_2}, \pi_{x_3}$ associated with the $X$ variable were random uniform numbers in the interval $[0, 1]$, while the initial values for the means $\mu_{y_1}, \mu_{y_2}, \mu_{y_3}$ were random numbers in the interval $[Q1(Y_i), Q3(Y_i)]$ for $i = 1, 2, 3$, where $Q1$ and $Q3$ stand for the respective 25th and the 75th quantile per component of the continuous part. Following the same procedure with the one of Basu and Lindsay [2] in the univariate continuous case, here (in the mixed-case) the numerical evaluation of the integrals was also done on the basis of the Simpson's 1/3rd rule using the `sintegral` function of the `Bolstad2` package (Bolstad [43]). Moreover, we calculated the mean values, the SDs, as well as the percentages of bias of the mean and the probability vectors for three different sample sizes: $N = 100$; $N = 1000$ and $N = 1500$ over 1000 MC replications. The bias is defined as the difference of the estimates from their "true" values, that is, $bias(\mu_{y_i}) = \hat{\mu}_{y_i} - \mu_i$ and $bias(\pi_{x_i}) = \hat{\pi}_{x_i} - 1/3$ for $i = 1, 2, 3$. The results are shown in Tables 7 and 8.

In particular, Table 7 illustrates the mean values, the SDs and the bias percentages of the corresponding minimum Hellinger distance estimators, over 1000 MC replications, for the three different sample sizes and polyserial correlations, when $\boldsymbol{\mu} = (0, 0, 0)^T$. The estimates for the $\pi_{x_i}$ are approximately equal to $1/3 = 0.333$, while the $\mu_{y_i}$ estimates are almost zero, even in the cases of weak correlations. When $\rho_{ON} = 0.0$, the sample size choice does not seem to affect the values of the estimates either overall or per component of $X, \boldsymbol{Y}$ variables. Specifically, we observe that the total absolute bias, computed as the sum of the individual component-wise absolute biases of the vectors $\boldsymbol{\pi}^T = (\pi_1, \pi_2, \pi_3)$ and $\boldsymbol{\mu}^T = (\mu_1, \mu_2, \mu_3)$ are approximately the same, with larger samples providing slightly less biases at the expense of a higher computational cost.

**Table 7.** Means, Absolute Biases and Overall Absolute Bias of the Hellinger's distance ($HD$). The data were concurrently generated with a given correlation structure (an overall correlation matrix $\Sigma$) and consist of a discrete variable $X$ with marginal probability vector $(1/3, 1/3, 1/3)$ and a continuous vector $\boldsymbol{Y} = (Y_1, Y_2, Y_3) \sim MVN_3(\boldsymbol{\mu}, \mathbf{I}_3)$, where $\boldsymbol{\mu}^T = (0, 0, 0)$ and $\mathbf{I}_3$ is a $(3 \times 3)$ identity matrix. The number of MC replications used is 1000.

| $\rho_{ON}$ | $N$ | Summary | **Estimates**<br>**Means, Biases over 1000 Replications** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $\hat{\pi}_{x_1}$ | $\hat{\pi}_{x_2}$ | $\hat{\pi}_{x_3}$ | $\hat{\mu}_{y_1}$ | $\hat{\mu}_{y_2}$ | $\hat{\mu}_{y_3}$ |
| 0.0 | 50 | Mean | 0.332 | 0.340 | 0.329 | 0.016 | 0.011 | −0.011 |
| | | Abs. Biases | 0.001 | 0.007 | 0.004 | 0.016 | 0.011 | 0.011 |
| | | Overall Bias | 0.050 | | | | | |
| | 100 | Mean | 0.330 | 0.350 | 0.320 | 0.017 | −0.018 | −0.010 |
| | | Abs. Biases | 0.003 | 0.017 | 0.013 | 0.017 | 0.018 | 0.010 |
| | | Overall Bias | 0.078 | | | | | |
| | 1000 | Mean | 0.324 | 0.337 | 0.339 | 0.001 | −0.008 | 0.007 |
| | | Abs. Biases | 0.009 | 0.004 | 0.006 | 0.001 | 0.008 | 0.007 |
| | | Overall Bias | 0.035 | | | | | |
| 0.1 | 50 | Mean | 0.351 | 0.320 | 0.329 | −0.006 | 0.003 | 0.005 |
| | | Abs. Biases | 0.018 | 0.013 | 0.004 | 0.006 | 0.003 | 0.005 |
| | | Overall Bias | 0.049 | | | | | |
| | 100 | Mean | 0.330 | 0.323 | 0.347 | 0.001 | 0.005 | −0.004 |
| | | Abs. Biases | 0.003 | 0.010 | 0.014 | 0.001 | 0.005 | 0.004 |
| | | Overall Bias | 0.037 | | | | | |
| | 1000 | Mean | 0.327 | 0.343 | 0.330 | −0.021 | 0.008 | 0.003 |
| | | Abs. Biases | 0.006 | 0.010 | 0.003 | 0.021 | 0.008 | 0.003 |
| | | Overall Bias | 0.051 | | | | | |

In Table 8, analogous results are presented with the difference that the mean vector used was $\boldsymbol{\mu} = (0, 3, 6)^T$. The $\pi_{x_i}$ estimates are very close to $1/3$ $(= 0.333)$ for all $X$ components, no matter which sample size or correlation is used. On the contrary, the interpretation of the $\mu_i$ estimates slightly differs in this case. We also calculated the overall absolute bias as well as the individual, per parameter, absolute biases. In this case, larger samples clearly provide estimates with smaller bias for both parameter vectors $\boldsymbol{\pi}, \boldsymbol{\mu}$ and for both cases, the case of independence as well as the case of weak correlations. However, the computational time increases.

In what follows, we also present -for illustration purposes- a small simulation example using a mixed-type, contaminated data set of size $N = 1000$, which was generated using `OrdNor` package setting $\rho_{ON} = 0.0$ . Once again, the data were comprised of one categorical variable $X$ with three levels and probability vector $(1/3, 1/3, 1/3)$, and a trivariate continuous vector $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$. The contamination is happening only in the continuous part on the basis of $\alpha \in \{1.00, 0.95, 0.90, 0.85, 0.80\}$, as follows: $\boldsymbol{Y} \sim \alpha \times MVN_3(\mathbf{0}, \mathbf{I}_3) + (1 - \alpha) \times MVN_3(\boldsymbol{\mu}, \mathbf{I}_3)$, where $\boldsymbol{\mu}^T = (3, 3, 3)$. This means that, $N_1 = \alpha \times N$ data were generated with $\boldsymbol{Y}$ coming from multivaraiate standard normal and the remaining $N_2 = N - N_1$ subset of the data followed a multivaraiate normal distribution with mean vector $\boldsymbol{\mu}^T = (3, 3, 3)$. It goes without saying that when $\alpha = 1.00$, there is no contamination. Here, we are still considering the same optimization problem with the one described above and, consequently, we are interested in evaluating the minimum Hellinger distance estimators over 1000 MC replications by examining/studying to what extend the contamination level affects these estimates.

**Table 8.** Means, Absolute Biases and Overall Absolute Bias of the Hellinger's distance ($HD$). The data were concurrently generated with a given correlation structure (an overall correlation matrix $\Sigma$) and consist of a discrete variable $X$ with marginal probability vector $(1/3, 1/3, 1/3)$ and a continuous vector $\boldsymbol{Y} = (Y_1, Y_2, Y_3) \sim MVN_3(\boldsymbol{\mu}, \mathbf{I}_3)$, where $\boldsymbol{\mu}^T = (0, 3, 6)$ and $\mathbf{I}_3$ is a $(3 \times 3)$ identity matrix. The number of MC replications used is 1000.

| $\rho_{ON}$ | $N$ | Summary | Estimates<br>Means, Biases over 1000 Replications | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $\hat{\pi}_{x_1}$ | $\hat{\pi}_{x_2}$ | $\hat{\pi}_{x_3}$ | $\hat{\mu}_{y_1}$ | $\hat{\mu}_{y_2}$ | $\hat{\mu}_{y_3}$ |
| 0.0 | 50 | Mean | 0.340 | 0.328 | 0.332 | −0.004 | 2.606 | 5.227 |
| | | Abs. Biases | 0.007 | 0.005 | 0.001 | 0.004 | 0.394 | 0.773 |
| | | Overall Bias | 1.184 | | | | | |
| | 100 | Mean | 0.313 | 0.350 | 0.337 | −0.004 | 2.777 | 5.593 |
| | | Abs. Biases | 0.020 | 0.017 | 0.004 | 0.004 | 0.223 | 0.407 |
| | | Overall Bias | 0.675 | | | | | |
| | 1000 | Mean | 0.338 | 0.334 | 0.328 | 0.012 | 2.972 | 5.958 |
| | | Abs. Biases | 0.005 | 0.001 | 0.005 | 0.012 | 0.028 | 0.042 |
| | | Overall Bias | 0.093 | | | | | |
| 0.1 | 50 | Mean | 0.347 | 0.323 | 0.330 | −0.021 | 2.628 | 5.249 |
| | | Abs. Biases | 0.014 | 0.010 | 0.003 | 0.021 | 0.372 | 0.751 |
| | | Overall Bias | 1.171 | | | | | |
| | 100 | Mean | 0.317 | 0.343 | 0.340 | 0.017 | 2.817 | 5.615 |
| | | Abs. Biases | 0.016 | 0.010 | 0.007 | 0.017 | 0.183 | 0.385 |
| | | Overall Bias | 0.618 | | | | | |
| | 1000 | Mean | 0.334 | 0.320 | 0.346 | −0.013 | 2.988 | 5.956 |
| | | Abs. Biases | 0.001 | 0.013 | 0.013 | 0.013 | 0.012 | 0.044 |
| | | Overall Bias | 0.096 | | | | | |
| 0.2 | 50 | Mean | 0.324 | 0.333 | 0.343 | −0.004 | 2.589 | 5.240 |
| | | Abs. Biases | 0.009 | 0.000 | 0.010 | 0.004 | 0.411 | 0.760 |
| | | Overall Bias | 1.194 | | | | | |
| | 100 | Mean | 0.329 | 0.350 | 0.321 | 0.024 | 2.763 | 5.549 |
| | | Abs. Biases | 0.004 | 0.017 | 0.012 | 0.024 | 0.237 | 0.451 |
| | | Overall Bias | 0.745 | | | | | |
| | 1000 | Mean | 0.337 | 0.344 | 0.319 | −0.011 | 2.971 | 5.951 |
| | | Abs. Biases | 0.004 | 0.011 | 0.014 | 0.019 | 0.029 | 0.049 |
| | | Overall Bias | 0.118 | | | | | |

As indicated from Table 9, when there is no contamination in the data ($\alpha = 1.00$), the estimates for the $\pi_{x_i}$s are almost equal to $1/3$, while the $\mu_y$'s estimates are almost equal to zero. As the data become more contaminated (i.e., the value of $\alpha$ decreases), the minimum disparity estimators corresponding to $X$ variable remain pretty consistent with their true values. However, this is not the case with the estimates for the $\mu_{y_i}$s, which deteriorate as the value of the contamination level $\alpha$ shifts from the target/null value, that is 1.00.

The mean parameters are estimated with reasonable bias (maximum bias is 9% for the second component of the mean) when $\alpha = 0.95$, that is the contamination is 5%. When the contamination is 10%, the bias of the mean components is relatively high but still below 19%. With higher contamination, the percentage of bias in the mean components is in the interval [28.3%, 47%]. This is the result of using standard density estimation to obtain the smoothing parameters for the different mean components. Smaller values of these component smoothing parameters result in substantial bias reduction.

**Table 9.** Means and SDs of the Hellinger's distance ($HD$). The data were concurrently generated with a given correlation structure (an overall correlation matrix $\Sigma$) and consist of a discrete variable $X$ with marginal probability vector $(1/3, 1/3, 1/3)$ and a continuous trivariate vector $\boldsymbol{Y} = (Y_1, Y_2, Y_3) \sim \alpha \times MVN_3(\mathbf{0}, \mathbf{I}_3) + (1 - \alpha) \times MVN_3(\boldsymbol{\mu}, \mathbf{I}_3)$, where $\boldsymbol{\mu}^T = (3, 3, 3)$, $\mathbf{I}_3$ is a $(3 \times 3)$ identity matrix and $\alpha = 1.00(0.05)0.80$ indicates the contamination level. The number of MC replications used is 1000.

| $\rho_{ON}$ | $N$ | $\alpha$ | **Summary** | **Estimates**<br>**Means and SDs over 1000 Replications** | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $\hat{\pi}_{x_1}$ | $\hat{\pi}_{x_2}$ | $\hat{\pi}_{x_3}$ | $\hat{\mu}_{y_1}$ | $\hat{\mu}_{y_2}$ | $\hat{\mu}_{y_3}$ |
| 0.0 | 1000 | 1.00 | Mean | 0.324 | 0.337 | 0.339 | 0.001 | −0.008 | 0.007 |
| | | | SD | 0.293 | 0.293 | 0.298 | 0.378 | 0.378 | 0.386 |
| | | 0.95 | Mean | 0.327 | 0.326 | 0.347 | 0.068 | 0.090 | 0.079 |
| | | | SD | 0.304 | 0.299 | 0.309 | 0.413 | 0.413 | 0.413 |
| | | 0.90 | Mean | 0.318 | 0.331 | 0.351 | 0.188 | 0.170 | 0.189 |
| | | | SD | 0.300 | 0.305 | 0.306 | 0.443 | 0.450 | 0.436 |
| | | 0.85 | Mean | 0.324 | 0.337 | 0.339 | 0.292 | 0.283 | 0.312 |
| | | | SD | 0.293 | 0.293 | 0.297 | 0.484 | 0.487 | 0.491 |
| | | 0.80 | Mean | 0.324 | 0.337 | 0.338 | 0.447 | 0.436 | 0.470 |
| | | | SD | 0.293 | 0.293 | 0.297 | 0.552 | 0.547 | 0.559 |

We also looked at the case where the continuous model was contaminated by a trivariate normal with mean $\boldsymbol{\mu}^T = (1.5, 1.5, 1.5)$ and covariance matrix $\mathbf{I}$. In this case (results not shown), when the contamination is 5% the maximum bias of the mean components is 6.6%, while when the contamination is 10% the maximum bias of the mean components is 13.5%. Again, in this case the bandwidth parameters were obtained by fitting a unimodal density to the data.

The above results are not surprising. A judicious selection of the smoothing parameter decreases the bias of the component estimates of the mean. Agostinelli and Markatou [44] provide suggestions of how to select the smoothing parameter that can be extended and applied in this context.

## 8. Discussion and Conclusions

In this paper, we discuss Pearson residual systems that conform to the measurement scale of the data. We place emphasis on the mixed-scale measurements scenario, which is equivalent to having both discrete (categorical or nominal) and continuous type random variables, and obtain robust estimators of the parameters of the joint probability distribution that describes those variables. We show that, disparity methods can be used to actually control against model misspecification and the presence of outliers, and these methods provide reasonable results.

The scale and nature of measurement of the data imposes additional challenges, both computationally and statistically. Detecting outliers in this multidimensional space is an open research question (Eiras-Franco et al. [45]). The concept of outliers has a long history in the field of statistics and outlier detection methods have broad applications in many scientific fields such as security (Diehl and Hampshire [46], Portnoy et al. [47]), health care (Tran et al. [48]) and insurance (Konijn and Kowalczyk [49]) to mention just a few.

Classical outlier detection methods are largely designed for single measurement scale data. Handling mixed measurement scale is a challenge with few works coming from both, the field of statistics (Fraley and Wilkinson [50], Wilkinson [51]) and the fields of engineering and computer science (Do et al. [52], Koufakou et al. [53]). All these works use some version of a probabilistic outlier, either looking for regions in the space of data that have low density (Do et al. [52], Koufakou et al. [53]) or by attaching a probability, under a model, to the suspicious data point (Fraley and Wilkinson [50], Wilkinson [51]).

Our concept of a probabilistic outlier discussed here and expressed via the construction of appropriate Pearson residuals can unify the different measurement scales, and the class of disparity functions discussed above can provide estimators for the model parameters that are not influenced unduly by potential outliers.

One of the important parameters that controls the robustness of these methods is the smoothing parameter(s) used to compute the density estimator of the continuous part of the model. In our computations, we use standard smoothing parameters obtained from utilizing appropriate *R* functions for density estimation. The results show that, depending on the level of contamination and the type of contaminating probability model, the performance of the methods is satisfactory. Specifically, a small simulation study using the model reported in the caption of Table 9 shows that the overall bias associated with the mean components of the standard multivariate normal model is low when contamination with a multivariate normal model with mean components equal to 3 is less than or equal to 10%. But even in this case, when the percentage of contamination is greater than 10%, the bias increases when the smoothing parameter used is the one obtained from the *R* density function. Here, smaller values of the smoothing parameter guarantee reduction of the bias.

Devising rules for selecting the smoothing parameter(s) in the context of mixed-scale measurements that can guarantee robustness for larger than 5% levels of contamination may be possible. However, it is the opinion of the authors that greater levels of data inhomogeneity may indicate model failure, a case where assessing model goodness of fit is of importance.

**Author Contributions:** The authors of this paper have contributed as follows. *Conceptualization*: M.M.; *Methodology*: M.M., E.M.S., R.L.; *Software*: E.M.S., H.W.; *Writing-original draft presentation*: M.M., E.M.S., R.L., H.W.; *Supervision, funding acquisition and project administration*: M.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Troup Fund, KALEIDA Health Foundation, under award number 82114, to Markatou who supported the work of the first and the third author of the paper.

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| ALT | Alanine Aminotransferase |
| HD | Twice-Squared Hellinger's Disparity |
| LD | Likelihood Disparity |
| MC | Monte Carlo Replications |
| MDE | Minimum Distance Estimators |
| MLE | Maximum Likelihood Estimator |
| PCS | Pearson's Chi-Squared Disparity Divided by 2 |
| PWD | Power Divergence Disparity |
| RAF | Residual Adjustment Function |
| SCS | Symmetric Chi-Squared Disparity |
| SD | Standard Deviation |

## Appendix A

### *Appendix A.1. Proof of Proposition 3*

**Proof.** The equations (4) are obtained from solving optimization problem (3). To solve this problem we need to form the corresponding Langrangian, which is

$$\sum_{x,y} G(\delta(x,y)) m_{\beta}(y|x)\pi_x - \lambda(\sum \pi_x - 1).$$

(i) Let $\nabla_{\boldsymbol{\beta}}$ denote gradient with respect to $\boldsymbol{\beta}$. The estimators of $\boldsymbol{\beta}$ are obtained as solutions of the set of equations:

$$\nabla_{\boldsymbol{\beta}}\Big\{\sum_{x,y} G(\delta(x,y))m_{\boldsymbol{\beta}}(y|x)\pi_x - \lambda(\sum \pi_x - 1)\Big\} = 0,$$

which can be equivalently expressed as follows,

$$\sum_{x,y} \pi_x[\nabla_{\boldsymbol{\beta}} G(\delta(x,y))]m_{\boldsymbol{\beta}}(y|x) + \sum_{x,y} \pi_x G(\delta(x,y))\nabla_{\boldsymbol{\beta}}(y|x) = 0.$$

Notice that the $\nabla_{\boldsymbol{\beta}}$ of $G(\delta(x,y))$ is given by

$$\nabla_{\boldsymbol{\beta}} G(\delta(x,y)) = -G'(\delta(x,y))(\delta(x,y)+1)\, u(y|x;\boldsymbol{\beta}),$$

where the superscript "′" denote derivative with respect to $\delta$, $\delta(x,y)$ is the Pearson residual and

$$u(y|x;\boldsymbol{\beta}) = \frac{\nabla_{\boldsymbol{\beta}} m_{\boldsymbol{\beta}}(y|x)}{m_{\boldsymbol{\beta}}(y|x)} = \nabla_{\boldsymbol{\beta}} \ln[m_{\boldsymbol{\beta}}(y|x)]$$

is the score for $\boldsymbol{\beta}$ in the conditional distribution of y given x. Therefore,

$$\sum_{x,y} A(\delta(x,y))\pi_x u(y|x;\boldsymbol{\beta})m_{\boldsymbol{\beta}}(y|x) = 0,$$

where

$$A(\delta(x,y)) = G'(\delta(x,y))[\delta(x,y)+1] - G(\delta(x,y)).$$

By making use of the fact that $\sum_x \pi_x \nabla_{\boldsymbol{\beta}} m_{\boldsymbol{\beta}}(y|x) = 0$, the resulting equations can represented as

$$\sum_{x,y} \frac{A(\delta(x,y))+1}{\delta(x,y)+1} n_{x,y} u(y|x;\boldsymbol{\beta}) = 0,$$

or equivalently,

$$\sum_{x,y} w(\delta(x,y)) n_{x,y} u(y|x;\boldsymbol{\beta}) = 0.$$

Without loss of generality, we can take,

$$w(\delta(x,y)) = \min\Big\{\frac{[A(\delta(x,y))+1]^+}{\delta(x,y)+1}, 1\Big\}, w(\delta(x,y)) \le 1.$$

(ii) We now need to obtain $\hat{\pi}_x$, which can be obtained by setting the gradient of formula with respect to $\pi_z$ equal to zero, that is, by the following equations:

$$\sum_y G'(\delta(z,y))[\nabla \pi_z \delta(z,y)]m_{\boldsymbol{\beta}}(y|z)\pi_z + \sum_y G(\delta(z,y))m_{\boldsymbol{\beta}}(y|z) - \lambda = 0.$$

Recording $A(\delta(z,y)) = G'(\delta(z,y))[\delta(z,y)+1] - G(\delta(z,y))$ and $\delta(z,y)+1 = \frac{n_{z,y}/n}{m_{\boldsymbol{\beta}}(y|z)\pi_z}$, the above equations are reduced to,

$$\sum_y A(\delta(z,y))m_{\boldsymbol{\beta}}(z,y)\frac{1}{\pi_z} + \lambda = 0$$

and we readily conclude that,

$$\pi_z = -\frac{1}{\lambda}\sum_y A(\delta(z,y))m(z,y), \forall z.$$

Furthermore, to satisfy the constraint $\sum_x \pi_x = 1$, we obtain

$$\lambda = -\sum_{x,y} A(\delta(x,y)) m_{\boldsymbol{\beta}}(x,y).$$

Therefore, we get

$$\sum_{x,y} A(\delta(x,y)) m_{\boldsymbol{\beta}}(y,x) \Big[ \frac{I(X=z)}{\pi_x} - 1 \Big] = 0$$

and by making use of the fact that $\sum_{x,y} m_{\boldsymbol{\beta}}(x,y) \Big[ \frac{I(X=z)}{\pi_x} - 1 \Big] = 0$, the above equation can be represented as

$$\sum_{x,y} w(\delta(x,y)) n_{x,y} \Big[ \frac{I(X=x)}{\pi_x} - 1 \Big] = 0$$

for any $x$ where $I(X=x)$ is the indicator function of the event $\{X=x\}$. □

*Appendix A.2. Proof of Proposition 5*

Recall that $\boldsymbol{\beta}_\epsilon$ is a solution of the set of estimating equation

$$\sum_{s,t} w(\delta_\epsilon(s,t)) u(t|s;\boldsymbol{\beta}_\epsilon) d_\epsilon(s,t) = 0, \tag{A1}$$

where $d_\epsilon(s,t) = (1-\epsilon)d(s,t) + \epsilon \nabla_{x,y}(s,t)$ and $u(t|s;\boldsymbol{\beta}) = \frac{\nabla_{\boldsymbol{\beta}} m_{\boldsymbol{\beta}}(s,t)}{m_{\boldsymbol{\beta}}(s,t)} = \nabla_{\boldsymbol{\beta}} \ln[m_{\boldsymbol{\beta}}(s,t)]$ is a $p$-dimensional vector.

The influence function of $\boldsymbol{\beta}$ is calculated by differentiating, with respect to $\epsilon$, the quantity (A1), and evaluating the derivative at $\epsilon = 0$. Thus, we need

$$\begin{aligned} \frac{d}{d\epsilon} \Big\{ & \sum_{s,t} w(\delta_\epsilon(s,t)) u(t|s;\boldsymbol{\beta}_\epsilon) d(s,t) \\ & - \epsilon \sum_{s,t} w(\delta_\epsilon(s,t)) u(t|s;\boldsymbol{\beta}_\epsilon) d(s,t) \\ & + \epsilon \sum_{s,t} w(\delta_\epsilon(s,t)) u(t|s;\boldsymbol{\beta}_\epsilon) \nabla_{(x,y)}(s,t) \Big\} \Big|_{\epsilon=0} = 0. \end{aligned} \tag{A2}$$

Taking into account that $\delta_\epsilon(s,t) = \frac{d_\epsilon(s,t)}{m_{\boldsymbol{\beta}}(s,t)} - 1 = \frac{d_\epsilon(s,t)}{m_{\boldsymbol{\beta}}(t|s)\pi_s} - 1$, the aforementioned evaluation implies

$$\begin{aligned} & \Big\{ \sum_{s,t} (\delta_0(t)+1) w_0'(\delta_0(s,t)) u(t|s;\boldsymbol{\beta}_0) u^T(t|s;\boldsymbol{\beta}_0) d(s,t) \\ & - \sum_{s,t} w(\delta_0(s,t)) \nabla u(t|s;\boldsymbol{\beta}_0) d(s,t) \Big\} \boldsymbol{\beta}_0' \\ = & \sum_{s,t} \Big\{ \frac{I(s=x,y=t)}{m_{\boldsymbol{\beta}_0}(t|s)\pi_s} - \frac{d(s,t)}{m_{\boldsymbol{\beta}_0}(t|s)\pi_s} w'(\delta_0(s,t)) \Big\} u(t|s;\boldsymbol{\beta}_0) d(s,t) \\ & - \sum_{s,t} w(\delta_0(s,t)) u(t|s;\boldsymbol{\beta}_0) d(s,t) + w(\delta_0(x,y)) u(y|x;\boldsymbol{\beta}_0), \end{aligned} \tag{A3}$$

which implies that

$$\boldsymbol{\beta}_0' = IF(\boldsymbol{\beta};F) = [A(d)]^{-1} B(x,y;d).$$

*Appendix A.3. Assumptions of Theorem 1*

The following assumptions are needed to be able to establish asymptotic normality of the estimators.

1. The weight functions are nonnegative, bounded and differentiable with respect to $\delta$.

2. The weight function is regular, that is, $w'(\delta)(\delta+1)$ is bounded, where $w'(\delta)$ is the derivative of $w$ with respect to $\delta$.

3. $\sum_{x,y} m^{\frac{1}{2}}(x,y)E[u_k^2(y|x;\boldsymbol{\beta}_0)] < \infty$.

4. The elements of the Fisher information matrix are finite and the Fisher information matrix is nonsingular.

5. $\sum_{x,y} m^{\frac{1}{2}}(x,y)E[u_i^2(y|x;\boldsymbol{\beta}_0)u_j^2(y|x;\boldsymbol{\beta}_0)] < \infty \;\; \forall i,j = 1,2,\cdots,p$.

6. If $\boldsymbol{\beta}_0$ denotes the true value of $\boldsymbol{\beta}$, there exist functions $M_{ijk}(x)$ such that $|u_{ijk}(y|x;\boldsymbol{\beta}_0)| \leq M_{ijk}(x)$, $\forall\boldsymbol{\beta}$ with $\parallel \boldsymbol{\beta}-\boldsymbol{\beta}_0 \parallel^2 < r(\boldsymbol{\beta}_0)$, $r(\boldsymbol{\beta}_0) < 0$ and $E_{\boldsymbol{\beta}_0}|M_{ijk}(y|x)| < \infty$, $\;\forall i,j,k$.

7. If $\boldsymbol{\beta}_0$ denotes the true value of $\boldsymbol{\beta}$, there is a neighborhood $N(\boldsymbol{\beta}_0)$ such that for $\boldsymbol{\beta} \in N(\boldsymbol{\beta}_0)$ the quantity $|u_t(y|x;\boldsymbol{\beta}_0)u_i(y|x;\boldsymbol{\beta}_0)u_e(y|x;\boldsymbol{\beta}_0)|$ are bounded by $M_1(y|x)$ and $M_2(y|x)$ respectively, such that their corresponding expectations are finite.

8. $A''(\delta+1)(\delta+1)$ is bounded, where $A''$ denotes the second derivative of $A$ with respect to $\delta$.

## References

1. Beran, R. Minimum Hellinger Distance Estimates for Parametric Models. *Ann. Stat.* **1977**, *5*, 445–463. [CrossRef]
2. Basu, A.; Lindsay, B.G. Minimum Disparity Estimation for Continuous Models: Efficiency, Distributions and Robustness. *Ann. Inst. Stat. Math.* **1994**, *46*, 683–705. [CrossRef]
3. Pardo, J.A.; Pardo, L.; Pardo, M.C. Minimum $\phi$-Divergence Estimator in Logistic Regression Models. *Stat. Pap.* **2005**, *47*, 91–108. [CrossRef]
4. Pardo, J.A.; Pardo, L.; Pardo, M.C. Testing In Logistic Regression Models on $\phi$-Divergences Measures. *J. Stat. Plan. Inference* **2006**, *136*, 982–1006. [CrossRef]
5. Pardo, J.A.; Pardo, M.C. Minimum $\phi$-Divergence Estimator and $\phi$-Divergence Statistics in Generalized Linear Models with Binary Data. *Methodol. Comput. Appl. Probab.* **2008**, *10*, 357–379. [CrossRef]
6. Simpson, D.G. Minimum Hellinger Distance Estimation for the Analysis of Count Data. *J. Am. Stat. Assoc.* **1987**, *82*, 802–807. [CrossRef]
7. Simpson, D.G. Hellinger Deviance Tests: Efficiency, Breakdown Points, and Examples. *J. Am. Stat. Assoc.* **1989**, *84*, 104–113. [CrossRef]
8. Markatou, M.; Basu, A.; Lindsay, B.G. Weighted Likelihood Estimating Equations: The Discrete Case with Applications to Logistic Regression. *J. Stat. Plan. Inference* **1997**, *57*, 215–232. [CrossRef]
9. Basu, A.; Basu, S. Penalized Minimum Disparity Methods for Multinomial Models. *Stat. Sin.* **1998**, *8*, 841–860.
10. Gupta, A.K.; Nguyen, T.; Pardo, L. Inference Procedures for Polytomous Logistic Regression Models Based on $\phi$-Divergence Measures. *Math. Methods Stat.* **2006**, *15*, 269–288.
11. Martín, N.; Pardo, L. New Influence Measures in Polytomous Logistic Regression Models Based on Phi-Divergence Measures. *Commun. Stat. Theory Methods* **2014**, *43*, 2311–2321. [CrossRef]
12. Castilla, E.; Ghosh, A.; Martín, N.; Pardo, L. New Robust Statistical Procedures for Polytomous Logistic Regression Models. *Biometrics* **2018**, *74*, 1282–1291. [CrossRef] [PubMed]
13. Martín, N.; Pardo, L. Minimum Phi-Divergence Estimators for Loglinear Models with Linear Constraints and Multinomial Sampling. *Stat. Pap.* **2008**, *49*, 2311–2321. [CrossRef]
14. Pardo, L.; Martín, N. Minimum Phi-Divergence Estimators and Phi-Divergence Test for Statistics in Contingency Tables with Symmetric Structure: An Overview. *Symmetry* **2010**, *2*, 1108–1120. [CrossRef]
15. Pardo, L.; Pardo, M.C. Minimum Power-Divergence Estimator in Three-Way Contingency Tables. *J. Stat. Comput. Simul.* **2003**, *73*, 819–831. [CrossRef]
16. Pardo, L.; Pardo, M.C.; Zografos, K. Minimum $\phi$-Divergence Estimator for Homogeneity in Multinomial Populations. *Sankhyā: Indian J. Stat. Ser. A (1961–2002)* **2001**, *63*, 72–92.
17. Basu, A.; Harris, I.A.; Hjort, N.L.; Jones, M.C. Robust and Efficient Estimation by Minimising a Density Power Divergence. *Biometrika* **1998**, *85*, 549–559. [CrossRef]
18. Csiszár, I. Information-Type Measures of Difference of Probability Distributions and Indirect Observations. *Stud. Sci. Math. Hung.* **1967**, *25*, 299–318.
19. Lindsay, B.G. Efficiency Versus Robustness: The Case for Minimum Hellinger Distance and Related Methods. *Ann. Stat.* **1994**, *22*, 1081–1114. [CrossRef]

20. Tamura, R.N.; Boos, D.D. Minimum Hellinger Distance Estimation for Multivariate Location and Covariance. *J. Am. Stat. Assoc.* **1986**, *81*, 223–229. [CrossRef]
21. Markatou, M.; Basu, A.; Lindsay, B.G. Weighted Likelihood Equations with Bootstrap Root Search. *J. Am. Stat. Assoc.* **1998**, *93*, 740–750. [CrossRef]
22. Haberman, S.J. Generalized Residuals for Log-Linear Models. In Proceedings of the 9th International Biometrics Conference, Boston, MA, USA, 22–27 August 1976; pp. 104–122.
23. Haberman, S.J.; Sinharay, S. Generalized Residuals for General Models for Contingency Tables with Application to Item Response Theory. *J. Am. Stat. Assoc.* **2013**, *108*, 1435–1444. [CrossRef]
24. Pierce, D.A.; Schafer, D.W. Residuals in Generalized Linear Models. *J. Am. Stat. Assoc.* **1986**, *81*, 977–986. [CrossRef]
25. Aerts, M.; Molenberghs, G.; Geys, H.; Ryan, L. *Topics in Modelling of Clustered Data*; Monographs on Statistics and Applied Probability; Chapman & Hall/CRC Press: New York, NY, USA, 1986; Volume 96.
26. Olkin, I.; Tate, R.F. Multivariate Correlation Models with Mixed Discrete and Continuous Variables. *Ann. Math. Stat.* **1961**, *32*, 448–465; With correction in **1961**, *36*, 343–344. [CrossRef]
27. Genest, C.; Nešlehová, J. A Primer on Copulas for Count Data. *ASTIN Bull.* **2007**, *37*, 475–515. [CrossRef]
28. Lauritzen, S.; Wermuth, N. Graphical Models for Associations between Variables, some of which are Qualitative and some Quantitative. *Ann. Stat.* **1989**, *17*, 31–57. [CrossRef]
29. Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.J.; Stahel, W.A. *Robust Statistics: The Approach Based on Influence Functions*; Wiley Series in Probability and Mathematical Statistics. Probability and Mathematical Statistics; Wiley: New York, NY, USA, 1986.
30. Hampel, F.R. Contributions to the Theory of Robust Estimation. Ph.D. Thesis, Department of Statistics, University of California, Berkeley, Berkeley, CA, USA, 1968. Unpublished.
31. Hampel, F.R. The Influence Curve and its Role in Robust Estimation. *J. Am. Stat. Assoc.* **1974**, *69*, 383–393. [CrossRef]
32. Fienberg, S.E. The Analysis of Incomplete Multi-Way Contingency Tables. *Biometrics* **1972**, *28*, 177–202. [CrossRef]
33. Agresti, A. *Categorical Data Analysis*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2013.
34. Johnson, W.D.; May, W.L. Combining 2 × 2 Tables That Contain Structural Zeros. *Biometrics* **1972**, *14*, 1901–1911. [CrossRef]
35. Poon, W.Y.; Tang, M.L.; Wang, S.J. Influence Measures in Contingency Tables with Application in Sampling Zeros. *Sociol. Methods Res.* **2003**, *31*, 439–452. [CrossRef]
36. Alin, A.; Kurt, S. Ordinary and Penalized Minimum Power-Divergence Estimators in Two-Way Contingency Tables. *Comput. Stat.* **2008**, *23*, 455–468. [CrossRef]
37. Ye, Y. Interior Algorithms for Linear, Quadratic, and Linearly Constrained Convex Programming. Ph.D. Thesis, Department of Engineering-Economic Systems, Stanford University, Stanford, CA, USA, 1987. Unpublished.
38. Conn, A.R.; Gould, N.I.M.; Toint, P. A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds. *SIAM J. Numer. Anal.* **1991**, *28*, 545–572. [CrossRef]
39. Birgin, E.G.; Martínez, J.M. Improving Ultimate Convergence of an Augmented Lagrangian Method. *Optim. Methods Softw.* **2008**, *23*, 177–195. [CrossRef]
40. Amatya, A.; Demirtas, H. OrdNor: An R Package for Concurrent Generation of Correlated Ordinal and Normal Data. *J. Stat. Softw.* **2015**, *68*, 1–14. [CrossRef]
41. Olsson, U.; Drasgow, F.; Dorans, N.J. The Polyserial Correlation Coefficient. *Psychmetrika* **1982**, *47*, 337–347. [CrossRef]
42. Duong, T. ks: Kernel Density Estimation and Kernel Discriminant Analysis for Multivariate Data in R. *J. Stat. Softw.* **2007**, *21*, 1–16. [CrossRef]
43. Bolstad, W.M. *Understanding Computational Bayesian Statistics*; John Wiley & Sons: Hoboken, NJ, USA, 2010.
44. Agostinelli, C.; Markatou, M. Test of Hypotheses Based on the Weighted Likelihood Methodology. *Stat. Sin.* **2001**, *11*, 499–514.
45. Eiras-Franco, C.; Martínez-Rego, D.; Guijarro-Berdiñas, B.; Alonso-Betanzos, A.; Bahamonde, A. Large Scale Anomaly Detection in Mixed Numerical and Categorical Input Spaces. *Inf. Sci.* **2019**, *487*, 115–127. [CrossRef]
46. Diehl, C.; Hampshire, J. Real-Time Object Classification and Novelty Detection for Collaborative Video Surveillance. In Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290), Honolulu, HI, USA, 12–17 May 2002; Volume 3, pp. 2620–2625.
47. Portnoy, L.; Eskin, E.; Stolfo, S. Intrusion Detection with Unlabeled Data Using Clustering. In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001), Philadelphia, PA, USA, 5–8 November 2001; pp. 5–8.
48. Tran, T.; Phung, D.; Luo, W.; Harvey, R.; Berk, M.; Venkatesh, S. An Integrated Framework for Suicide Risk Prediction. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; ACM: New York, NY, USA, 2013; pp. 1410–1418.
49. Konijn, R.M.; Kowalczyk, W. Finding Fraud in Health Insurance Data with Two-Layer Outlier Detection Approach. In Data Warehousing and Knowledge Discovery, DaWak 2011; Cuzzocrea, A., Dayal, U., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 394–405.
50. Fraley, C.; Wilkinson, L. Package 'HDoutliers'. R Package, 2020. Available online: https://cran.r-project.org/web/packages/HDoutliers/index.html (accessed on 31 December 2020).
51. Wilkinson, L. Visualizing Outliers. 2016. Available online: https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf (accessed on 31 December 2020).

52. Do, K.; Tran, T.; Phung, D.; Venkatesh, S. Outlier Detection on Mixed-Type Data: An Energy-Based Approach. In *Advanced Data Mining and Applications*; Li, J., Li, X., Wang, S., Li, J., Sheng, Q.Z., Eds.; Springer: Cham, Switzerland, 2016; pp. 111–125.
53. Koufakou, A.; Georgiopoulos, M.; Anagnostopoulos, G.C. Detecting Outliers in High-Dimensional Datasets with Mixed Attributes. In Proceedings of the 2008 International Conference on Data Mining, DMIN, Las Vegas, NV, USA, 14–17 July 2008; pp. 427–433.