Scientific Research Publishing

Open Journal of Statistics, 2018, 8, 187-219
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Simulated Minimum Hellinger Distance Inference Methods for Count Data

Andrew Luong, Claire Bilodeau, Christopher Blier-Wong

École d'actuariat, Université Laval, Québec, Canada
Email: Andrew.Luong@act.ulaval.ca, Claire.Bilodeau@act.ulaval.ca, Christopher.Blier-Wong.1@act.ulaval.ca

How to cite this paper: Luong, A., Bilodeau, C. and Blier-Wong, C. (2018) Simulated Minimum Hellinger Distance Inference Methods for Count Data. *Open Journal of Statistics*, **8**, 187-219.
https://doi.org/10.4236/ojs.2018.81012

Received: January 22, 2018
Accepted: February 25, 2018
Published: February 28, 2018

Abstract

In this paper, we consider simulated minimum Hellinger distance (SMHD) inferences for count data. We consider grouped and ungrouped data and emphasize SMHD methods. The approaches extend the methods based on the deterministic version of Hellinger distance for count data. The methods are general, it only requires that random samples from the discrete parametric family can be drawn and can be used as alternative methods to estimation using probability generating function (pgf) or methods based matching moments. Whereas this paper focuses on count data, goodness of fit tests based on simulated Hellinger distance can also be applied for testing goodness of fit for continuous distributions when continuous observations are grouped into intervals like in the case of the traditional Pearson's statistics. Asymptotic properties of the SMHD methods are studied and the methods appear to preserve the properties of having good efficiency and robustness of the deterministic version.

Keywords

Break Down Points, Robustness, Power Mixture, Esscher Transform, Mixture Discrete Distributions, Chi-Square Tests Statistics

1. Introduction

1.1. New Distribution Created Using Probability Generating Functions

Nonnegative discrete parametric families of distributions are useful for modeling count data. Many of these families do not have closed form probability mass functions nor closed form formulas to express the probability mass function (pmf) recursively. Their pmfs can only be expressed using an infinite series representation but their corresponding Laplace transforms have a closed form and,

in many situations, they are relatively simple. Probability generating functions are often used for discrete distributions but Laplace transforms are equivalent and can also be used. In this paper, we use Laplace transforms but they will be converted to probability generating functions (pgfs) whenever the need arises to link with results which already appear in the literature. We begin with a few examples to illustrate the situation often encountered when new distributions are created.

Example 1 (Discrete stable distributions) The random variable $X \geq 0$ follows a positive stable law if the probability generating function and Laplace transform are given respectively as

$$P_{\boldsymbol{\beta}}(s) = E\left(s^X\right) = \mathrm{e}^{-\lambda(1-s)^{\alpha}},\ 0 < \alpha \leq 1,\ \lambda > 0,\ \boldsymbol{\beta} = (\lambda, \alpha)',\ |s| \leq 1$$

and

$$\varphi_{\boldsymbol{\beta}}(s) = E\left(\mathrm{e}^{-sX}\right) = \mathrm{e}^{-\lambda\left(1-\mathrm{e}^{-s}\right)^{\alpha}},\ 0 < \alpha \leq 1,\ \lambda > 0,\ \boldsymbol{\beta} = (\lambda, \alpha)',\ s \geq 0.$$

The distribution was introduced by Christoph and Schreiber [1].

It is easy to see that $\varphi_{\boldsymbol{\beta}}(s) = P_{\boldsymbol{\beta}}\left(\mathrm{e}^{-s}\right)$.

The Poisson distribution can be obtained by fixing $\alpha = 1$. The distribution is infinitely divisible and displays long tail behavior. The recursive formula for its mass function has been obtained; see expression (8) given by Christoph and Schreiber [1].

Now if we allow λ to be a random variable with an inverse Gaussian distribution whose Laplace transform is given by $h(s) = \mathrm{e}^{\mu\left(1-\sqrt{1+\frac{2}{\mu}s}\right)}, s \geq -\frac{\mu}{2}$, a mixed nonnegative discrete stable distribution can be created with Laplace transform given by

$$\varphi_{\boldsymbol{\beta}}(s) = \int_0^{\infty} \left(g(s)\right)^{\lambda} \mathrm{d}H(\lambda),$$

where $g(s) = \mathrm{e}^{-(1-s)^{\alpha}}$ and $H(\lambda)$ is the distribution with Laplace transform $h(s)$. The resulting Laplace transform,

$$\varphi_{\boldsymbol{\beta}}(s) = \exp\left(\mu\left(1-\sqrt{1+\frac{2}{\mu}\left(1-\mathrm{e}^{-s}\right)^{\alpha}}\right)\right),$$

is the Laplace transform of a nonnegative infinitely divisible (ID) distribution.

We can see that it is not always straightforward to find the recursive formula for the pmf for a nonnegative count distribution. Even if it is available, it might still complicated to be used numerically for inferences meanwhile the Laplace transform or pgf can have a relatively simple representation.

We can observe that the new distribution is obtained by using the inverse Gaussian distribution as a mixing distribution. This is also an example of the use of a power mixture (PM) operator to obtain a new distribution. The PM operator will be further discussed in Section 1.2.

From a statistical point of view, when neither a closed form pmf nor a recur-

sive formula for the pmf exists, maximum likelihood estimation can be difficult to implement.

The power mixture operator was introduced by Abate and Whitt [2] (1996) as a way to create new distributions from an infinitely divisible (ID) distribution together with a mixing distribution using Laplace transforms (LT). We shall review it here in the next section, after a definition of an ID distribution.

Definition 1.1.3. A nonnegative random variable X is infinitely divisible if its Laplace transform can be written as

$$\psi(s) = \left(k_n(s)\right)^n, n = 1, 2, \cdots,$$

where $k_n(s)$ also is the Laplace transform of a random variable. In many situations, $k_n(s)$ and $\psi(s)$ belong to the same parametric family. See Panjer and Willmott [3] (1992, p42) for this definition.

Abate and Whitt [2] (1996) introduced the power mixture (PM) operator for ID distributions and also some other operators. To the operators already developed by them, we add the Esscher transform operator and the shift operator. All operators considered are discussed below.

1.2. Operational Calculus on Laplace Transforms

1.2.1. Power Mixture (PM) Operator

Suppose that X_t is an infinitely divisible nonnegative discrete random variable such that the Laplace transform can be expressed as $\left(\kappa(s)\right)^t, t \geq 0$, where $\kappa(s)$ is the Laplace transform of X, which is nonnegative and infinitely divisible as well. The power mixture (PM) with mixing distribution function $H(y)$ and Laplace transform $\kappa_H(s)$ of a nonnegative random variable Y is defined as the Laplace transform

$$\eta(s) = \mathrm{PM}(\kappa, H) = \int_0^\infty \left(\kappa(s)\right)^t \mathrm{d}H(t) = \kappa_H\left(-\log\left(\kappa(s)\right)\right).$$

Furthermore, if $H(y)$ is infinitely divisible, then the distribution with Laplace transform $\eta(s)$ is also infinitely divisible. The random variable $Y \geq 0$ with distribution $H(y)$ can be discrete or continuous but needs to be ID. This is the PM method for creating new parametric families, *i.e.*, using the PM operator. The PM method can be viewed as a form of continuous compounding method. The ID property can be dropped but as a result the new distribution created using the PM operator needs not be ID. For the traditional compounding methods, see Klugman *et al.* [4] (p141-148). Abate and Whitt [2] also mentioned other methods.

Example 2 (Generalized negative binomial) The generalized negative binomial (GNB) distribution introduced by Gerber [5] can be viewed as a power variance function distribution mixture of a Poisson distribution. The power variance function distribution introduced by Hougaard [6] is obtained by tilting the positive stable distribution using a parameter θ. It is a three-parameter continuous nonnegative distribution with Laplace transform given by

$$\kappa_H(s) = \exp\left\{-\lambda\left(\theta+s\right)^{\alpha} - \theta^{\alpha}\right\}, \lambda, \theta > 0, 0 < \alpha < 1.$$

Gerber [5] used a different parameterization and named this distribution generalized gamma. It is also called positive tempered stable distribution in finance.

Let $\kappa(s) = \mathrm{e}^{\left(\mathrm{e}^{-s}-1\right)}$ be the Laplace transform of a Poisson distribution with rate $\mu = 1$. The Laplace transform of the GNB distribution can be represented as

$$\eta(s) = \exp\left(-\lambda\left(\theta - \mathrm{e}^{-s} + 1\right)^{\alpha} - \theta^{\alpha}\right).$$

The corresponding pgf can be expressed as

$$P(s) = \exp\left(-\lambda\left(\theta - s + 1\right)^{\alpha} - \theta^{\alpha}\right).$$

The pgf is given by expression (21) in the paper by Gerber [5]. The GNB distribution is infinitely divisible. If stochastic processes are used instead of distributions, the distribution can also be derived from a stochastic process point of view by considering a Poisson process subordinated to a generalized gamma process and obtain the new distribution as the distribution of increments of the new process created. See section 6 of Abate and Whitt [2] (p92-93). See Zhu and Joe [7] for other distributions which are related to the GNB distribution.

Note that, if $H(y)$ is discrete, $\eta(s)$ is the Laplace transform of a random variable expressible as a random sum. A random sum is also called stopped sum in the literature, see chapter 9 by Johnson *et al.* [8] (p343-403). The Neymann-Type A distribution given below is an example of a distribution of a random sum.

Example 3 Let $X = \sum_{i=1}^{Y} U_i$,the U_i's conditioning on Y are independent and identically distributed and follows a Poisson distribution with rate ϕ and Y is distributed with a Poisson distribution with rate λ. Using the Power mixture operator we conclude that the LT for X is

$$\eta(s) = \exp\left(\lambda\left(\mathrm{e}^{\phi\left(\mathrm{e}^{-s}-1\right)} - 1\right)\right),$$

and the pgf is

$$P(s) = \exp\left(\lambda\left(\mathrm{e}^{\phi(s-1)} - 1\right)\right).$$

Properties and applications of the Neymann type A distribution have been studied by Johnson *et al.* [8] (p368-378). The mean and variance of X are given respectively by $E(X) = \lambda\phi$ and $V(X) = \lambda\phi(1+\phi)$. From these expressions, moment estimators (MM) have closed form expressions, see section (4.1) for comparisons between MM estimators and SMHD estimators in a numerical study. For applications often the parameter λ is smaller than the parameter ϕ.

1.2.2. Esscher Transform Operator

By tilting the density function using the Esscher transform, the Esscher trans-

form operator can be defined and, provided the tilting parameter τ introduced is identifiable, new distributions can be created from existing ones.

Let X be the original random variable with Laplace transform $\kappa(s)$. The Esscher transform operator which can be viewed as a tilting operator is defined as

$$\eta(s) = \text{Esscher}(\kappa, \tau) = \frac{\kappa(s+\tau)}{\kappa(\tau)}.$$

1.2.3. Shift Operator

Let $\kappa(s)$ be the Laplace transform of a positive continuous random variable X. The Laplace transform of $Y = X - \tau, Y \geq \tau \geq 0$ is given by $e^{\tau}\kappa(s)$. So, we can define the shift operator as

$$\eta(s) = \text{Shift}(\kappa, \tau) = e^{\tau}\kappa(s).$$

In some cases, even the pmf of Y has a closed form but the maximum likelihood (ML) estimators might be attained at the boundaries, the ML estimators might not have the regular optimum properties.

Note that parallel to the closed form pgf expressions for these new discrete distributions, it is often simple to simulate from the new distributions if we can simulate from the original distribution before the operators are applied. For example, let us consider the new distribution obtained by using the Esscher operator. It suffices to simulate from the distribution before applying the operator and apply the acceptance-rejection method to obtain a sample from the Esscher transformed distribution. The situation is similar for new distributions created by the PM operator. If we can simulate one observation from the mixing distribution of Y which gives a realized value t and if it is not difficult to draw one observation from the distribution with LT $\kappa(s)^t$ then combining these two steps, we would be able to obtain one observation from the new distribution created by the PM operator. Consequently, simulated methods of inferences offer alternative methods to inferences methods based on matching selected points of the empirical pgf with its model counterpart or other related methods, see Doray *et al.* [9] for regression methods using selected points of the pgfs. For these methods there is some arbitrariness on the choice of points which make it difficult to apply. The techniques of using a continuum number of points to match are more involved numerically, see Carrasco and Florens [10]. The new methods also avoid the arbitrariness of the choice of points which is needed for the regression methods and the *k-L* procedures as proposed by Feurverger and McDunnough [11] if characteristic functions are used instead of probability generating functions and they are more robust than methods based on matching moments (MM) in general. We can reach the same conclusions for another class of distributions namely mixture distributions created by other mixing mechanisms, see Klugman *et al.* [4], Nadarajah and Kotz [12], Nadarajah and Kotz [13]. These distributions might not display closed form pmf or the pmf are only ex-

pressible only using special functions such as the confluent hypergeometric functions. For these models, likelihood methods might also be difficult to implement.

This leads us to look for alternative methods such as the simulated minimum Hellinger distance (SMHD) methods for count data. We shall consider grouped count data and ungrouped count data. With grouped data, it leads to simulated chi-square type statistics which can be used for model testing for discrete or continuous models. These statistics are similar to the traditional Pearson statistics. For model testing with continuous distributions, continuous observations when grouped into intervals are reduced to count data and we do not need to integrate the model density functions on intervals using SMHD methods, it suffices to simulate from the continuous model and construct sample distribution functions to obtain estimate interval probabilities. Therefore, the scopes of applications of simulated methods are widened due to these features.

We briefly describe the classical minimum Hellinger distance methods introduced by Simpson [14], Simpson [15] for estimation for count data in the next section and we shall develop inference methods based on a simulated version of this HD distance following Pakes and Pollard [16] (1989), who have developed an elegant asymptotic theory for estimators obtained by minimizing a simulated objective function expressible as the Euclidean norm of a random vector of functions. As an example, they have shown that the simulated minimum chi-square estimators without weight satisfy the regularity conditions for being consistent and have an asymptotic normal distribution, see Pakes and Pollard [16] (p1048). They work with properties of some special classes of sets to check the regularity conditions of their Theorem 3.3. Meanwhile, Newey and McFadden [17] (p2187) work with properties of random functions and introduce a stochastic version of the classical equicontinuity property of real analysis. In this paper, we shall also extend the notion of continuity of real analysis to a version which only holds in probability for random functions which we call continuity in probability for a sequence of random functions which is similar to the notion of continuity with probability one as discussed by Newey and McFadden [17] (p2132) in their Theorem 2.6. We also use the property of the compact domains under considerations shrink as the sample size $n \to \infty$ to verify conditions of Theorem 3.3 given by Pakes and Pollard [16] (1989) for SMHD methods using grouped data and conditions of Theorem 7.1 of Newey and McFadden [17] (p2185) for ungrouped data. This approach appears to be new and simpler that other approaches which have been used in the literature to establish asymptotic normality for estimators using simulations; previous approaches are very general but they are also more complicated to apply. A similar notion of continuity in probability has been introduced in the literature of stochastic processes.

It is worth to mention that simulated methods of inferences are relatively recent. In advanced econometrics textbook such as the book by Davidson and McKinnon [18], only section 9.6 is devoted to simulated methods of inferences

where the authors mention simulated methods of moments (MSM). The simulated version for HD methods will be referred to as version S and the original version which is deterministic will be referred to as version D in this paper. We briefly review the Hellinger distance and chi-square distance below and subsequently develop simulated inference methods for grouped and ungrouped count data using HD distance.

1.3. Hellinger and Chi-Square Distance Estimation

Assume that we have a random sample of n independent and identically distributed

(iid) nonnegative observations $X_1,\cdots,X_n$ from a pmf $p_{\boldsymbol{\theta}}(x)$ with $x=0,1,\cdots$ and

$$\boldsymbol{\theta}=(\theta_1,\cdots,\theta_m)'$$

is the vector of parameters of interest, $\boldsymbol{\theta}_0$ is the vector of the true parameters. If the data are grouped into $r=k+1$ disjoint intervals $I_j, j=0,1,\cdots,k$ so that they form a partition of the nonnegative real line, the unweighted chi-square distance is defined to be

$$CS_n(\boldsymbol{\theta})=\sum_{j=0}^{k}\left(p_n(I_j)-p_{\boldsymbol{\theta}}(I_j)\right)^2,$$

where $p_n(I_j)$ is the proportion of the sample which fall into the interval I_j and $p_{\boldsymbol{\theta}}(I_j)$ is the probability of an observation which fall into I_j under the pmf $p_{\boldsymbol{\theta}}(x)$. If $p_{\boldsymbol{\theta}}(x)$ has no closed form expression but we can draw a sample of size $U=\tau n$ from this distribution then clearly $p_{\theta}(I_j)$ can be estimated by $p_{\theta}^{S}(I_j)$ using the simulated sample of size U which is the proportion of observations of the simulated sample which has taken a value in I_j. To illustrate their theory Pake and Pollard [16] (p1047-1048) considered simulated estimators obtained by minimizing with respect to $\boldsymbol{\theta}$ the objective function

$$Q_n(\boldsymbol{\theta})=\sum_{j=0}^{k}\left(p_n(I_j)-p_{\boldsymbol{\theta}}^{S}(I_j)\right)^2$$

and show that the estimators satisfy the regularity conditions of their Theorem 3.1 and 3.3 which lead to conclude that the simulated estimators are consistent and have an asymptotic normal distribution. As we already know, a weighted version can be more efficient, if we attempt a version S for the Pearson's chi square distance,

$$P(\boldsymbol{\theta})=\sum_{j=0}^{k}\frac{\left(p_n(I_j)-p_{\boldsymbol{\theta}}(I_j)\right)^2}{p_{\boldsymbol{\theta}}(I_j)},$$

and since the denominator of the summand involves $p_{\theta}(I_j)$, it is numerically not easy to introduce a version S. Clearly, if $p_{\boldsymbol{\theta}}^{S}(I_j)=0$, the version S of this distance will run into numerical difficulties. The traditional and deterministic

version of the Hellinger distance as given by

$$Q_n(\boldsymbol{\theta}) = \sum_{j=0}^{k}\left(\left[p_n(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}}(I_j)\right]^{\frac{1}{2}}\right)^2 \tag{1}$$

is more appropriate for a version S and it is already known that it generates minimum HD estimators which are as efficient as the minimum chi-square estimators or maximum likelihood (ML) estimators for grouped data, see Cressie-Read divergence measure with $\lambda = -\frac{1}{2}$ given by Cressie and Read [19] (p457) for version D.

Note that $\mathrm{HD}(\boldsymbol{\theta}) = 2 - 2\sum_{j=0}^{k}\left[p_n(I_j)\right]^{\frac{1}{2}}\left[p_{\boldsymbol{\theta}}(I_j)\right]^{\frac{1}{2}}$ and by using Cauchy-Schwartz inequality, we have

$$0 \le \sum_{j=0}^{k}\left[p_n(I_j)\right]^{\frac{1}{2}}\left[p_{\boldsymbol{\theta}}(I_j)\right]^{\frac{1}{2}} \le 1,$$

so that $0 \le Q_n(\boldsymbol{\theta}) \le 2$ and $Q_n(\boldsymbol{\theta})$ remains always bounded. Therefore the objective function for version S can be defined as

$$Q_n(\boldsymbol{\theta}) = \sum_{j=0}^{k}\left(\left[p_n(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}}^{S}(I_j)\right]^{\frac{1}{2}}\right)^2. \tag{2}$$

Since the objective function remains bounded and this property continues to hold for the ungrouped data case, this suggests that SMHD methods could preserve some of the nice robustness properties of version D.

For ungrouped data, it is equivalent to have grouped data but using intervals with unit length $I_j = [j, j+1), j = 0,1,\cdots$ and the number of classes is infinite, we shall develop SMHD estimation which is based on the objective function

$$Q_n(\boldsymbol{\theta}) = \sum_{j=0}^{\infty}\left(\left[p_n(j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}}^{S}(j)\right]^{\frac{1}{2}}\right)^2 = 2 - 2\sum_{i=0}^{\infty}\left[p_n(j)\right]^{\frac{1}{2}}\left[p_{\boldsymbol{\theta}}^{S}(j)\right]^{\frac{1}{2}}. \tag{3}$$

Note that for a data set the sum given by the RHS of the above expression only has a finite number of terms as $p_n(j) = 0$ when j is large.

The version D with

$$Q_n(\theta) = \sum_{j=0}^{\infty}\left(\left[p_n(j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}}(j)\right]^{\frac{1}{2}}\right)^2 = 2 - 2\sum_{i=0}^{\infty}\left[p_n(j)\right]^{\frac{1}{2}}\left[p_{\boldsymbol{\theta}}(j)\right]^{\frac{1}{2}} \tag{4}$$

has been investigated by Simpson [14], Simpson [15] who also shows that the MHD estimators have a high breakdown point of at least 50% and first order as efficient as the ML estimators. For the Poisson case, the ML estimator is the sample mean which has a zero breakdown point and consequently far less robust than the HD estimators, yet the HD estimators are first order as efficient as the ML estimators. This feature makes HD estimators attractive. For the notion of finite sample break down point as a measure of robustness, see Hogg *et al.* [20] (p594-595), Kloke and McKean [21] (p29) and for the notion of asymptotic

breakdown point for large samples, see Maronna *et al.* [22] (p58).

Simpson [14], Simpson [15] extended the works of Beran [23] for continuous distributions to discrete distributions. Beran [23] appears to be the first to introduce a weaker form of robustness not based on bounded influence function and shows that efficiency can be achieved for robust estimators not based on influence functions. Also, see Lindsay [24] for discussions on robustness of Hellinger distance estimators. Simulated versions extending some of the seminal works of Simpson will be introduced in this paper.

SMHD methods appear to be useful for actuarial studies when there is a need for fitting discrete risk models, see chapter 9 of Panjer and Willmott [3] (p292-238) for fitting discrete risk models using ML methods. The SMHD methods appear to be useful for other fields as well especially when there is a need to analyze count data with efficiency and robustness but the pmfs of the models do not have closed form expressions. For minimizing the objective functions to obtain SMHD estimators, simplex derivative free algorithm can be used and the R package already has built in functions to implement these minimization procedures.

1.4. Outlines of the Paper

In this paper, we develop unified simulated methods of inferences for grouped and ungrouped count data using HD distances and it is organized as follows. Asymptotic properties for SMHD methods are developed in Section 2 where consistency and asymptotic normality are shown in Section 2.2. Based on asymptotic properties, consistency of the SMHD estimators hold in general but high efficiencies of SMHD estimators can only be guaranteed if the Fisher information matrix of the parametric exists, a situation which is similar to likelihood estimation. One can also viewed the estimators are fully efficient within the class of simulated estimators obtained with the model pmf being replaced by a simulated version. Chi-square goodness of fit test statistics are constructed in Section 2.3. For the ungrouped case, it can be seen as having grouped data but the number of intervals with unit length and the number of intervals is infinite, it is given in Section 3 where the ungrouped SMHD estimators are shown to have good efficiencies. The breakdown point for the SMHD estimators remains at least $\frac{1}{2}$ just as for the deterministic version. A limited simulation study is included in Section 4. First, we consider the Neymann type A distribution and compare the efficiencies of the SMHD estimators versus moment (MM) estimators, simulations results appear to confirm the theoretical results showing that the SMHD estimators are more efficient than the MM estimators based on matching the first two empirical moments with their model counterparts for a selected range of parameters. The Poisson distribution is considered next and the study shows that despite being less efficient than the ML estimator, the efficiency of the SMHD estimators remain high and the estimators are far more ro-

bust than the ML estimator in the presence of outliers just as in the deterministic case as shown by Simpson [14] (p805). More works are needed in this direction in general and for assessing the performance SMHD estimators and comparisons with the performances of other traditional estimators in various parametric models in finite samples.

2. SMHD Methods for Grouped Data

2.1. Introduction

Pakes and Pollard [16] have developed a very elegant and general theory for establishing consistency and asymptotic normality of estimators obtained by minimizing the length of a random function taking values in an Euclidean space, *i.e.*, by minimizing

$$\left\|\boldsymbol{G}_n(\boldsymbol{\theta})\right\| \tag{5}$$

or

$$\left(\left\|\boldsymbol{G}_n(\boldsymbol{\theta})\right\|\right)^2 \tag{6}$$

where $\boldsymbol{G}_n(\boldsymbol{\theta})=\left(G_{n,0}(\boldsymbol{\theta}),\cdots,G_{n,k}(\boldsymbol{\theta})\right)'$ is a vector of random functions with values in a Euclidean space and $\|\cdot\|$ is the Euclidean norm and if $\boldsymbol{A}=\left(a_{ij}\right), i=1,\cdots,a,\ j=1,\cdots,b$ is a matrix of finite dimension then $\|\boldsymbol{A}\|=\left(\sum_{i=1}^{a}\sum_{j=1}^{b}a_{ij}\right)^{\frac{1}{2}}$. Their theory is summarized by their Theorem 3.1 and Theorem 3.3 given in Pakes and Pollard [16] (p1038-1043). It is very general and it is clearly applicable for both versions D and S for Hellinger distance with grouped data. Let

$$Q_n(\boldsymbol{\theta})=\left(\left\|\boldsymbol{G}_n(\boldsymbol{\theta})\right\|\right)^2,\ Q(\boldsymbol{\theta})=\left(\left\|\boldsymbol{G}(\boldsymbol{\theta})\right\|\right)^2 \tag{7}$$

and for HD distance, version D, let

$$\boldsymbol{G}_n(\boldsymbol{\theta})=\left(\left[p_n(I_0)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}}(I_0)\right]^{\frac{1}{2}},\cdots,\left[p_n(I_k)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}}(I_k)\right]^{\frac{1}{2}}\right)', \tag{8}$$

and for version S, let

$$\boldsymbol{G}_n(\boldsymbol{\theta})=\left(\left[p_n(I_0)\right]^{\frac{1}{2}}-\left[p^S_{\boldsymbol{\theta}}(I_0)\right]^{\frac{1}{2}},\cdots,\left[p_n(I_k)\right]^{\frac{1}{2}}-\left[p^S_{\boldsymbol{\theta}}(I_k)\right]^{\frac{1}{2}}\right)' \tag{9}$$

which can be reexpressed as

$$\begin{aligned}\boldsymbol{G}_n(\boldsymbol{\theta})=\Bigg(&\left(\left[p_n(I_0)\right]^{\frac{1}{2}}-\left[p_{\theta}(I_0)\right]^{\frac{1}{2}}\right)-\left(\left[p^S_{\theta}(I_0)\right]^{\frac{1}{2}}-\left[p_{\theta}(I_0)\right]^{\frac{1}{2}}\right),\cdots,\\&\left(\left[p_n(I_k)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}}(I_k)\right]^{\frac{1}{2}}\right)-\left(\left[p^S_{\boldsymbol{\theta}}(I_k)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}}(I_k)\right]^{\frac{1}{2}}\right)\Bigg)'\end{aligned}. \tag{10}$$

In general, the intervals I_i's form a partition of the nonnegative real line R_0^+

with $R_0^+ = \bigcup_{i=0}^{k} I_i$. Only in section (2.3) where we want to test goodness of fit for continuous distribution with support of the entire real line used in financial study, we might let $R = \bigcup_{i=0}^{k} I_i$, R is the real line.

Let $\boldsymbol{G}(\boldsymbol{\theta}) = \left(\left[p_{\boldsymbol{\theta}_0}(I_0) \right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}}(I_0) \right]^{\frac{1}{2}}, \cdots, \left[p_{\boldsymbol{\theta}_0}(I_k) \right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}}(I_k) \right]^{\frac{1}{2}} \right)'$, the vector of the true parameters is denoted by $\boldsymbol{\theta}_0 \in \Omega$, the parameter space Ω is assumed to be compact. Clearly, we have point wise convergence in probability with $G_n(\boldsymbol{\theta}) \xrightarrow{p} G(\boldsymbol{\theta})$ for each $\boldsymbol{\theta}$ for both versions, $G(\boldsymbol{\theta})$ is nonrandom. Clearly the set up fits into the scopes of their Theorem 3.1 and 3.3 which we shall rearrange the results of these two theorems before applying to version D and version S of Hellinger distance inferences and verify that we can satisfy the regularity conditions of these two Theorems.

2.2. Asymptotic Properties

2.2.1. Consistency

We define MHD estimators as given by the vector $\widehat{\boldsymbol{\theta}_G}$ for version D and $\widehat{\boldsymbol{\theta}_G^S}$ for version S but emphasize version S as version D has been studied by Simpson [14]. Both versions can be treated in a unified way using the following Theorem 1 for consistency which is essentially Theorem 3.1 of Pakes and Pollard [16] (p1038) and the proof has been given by the authors.

Theorem 1 (Consistency)

Under the following conditions $\tilde{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$:

a) $\left\| \boldsymbol{G}_n(\tilde{\boldsymbol{\theta}}) \right\| \le o_p(1) + \inf_{\boldsymbol{\theta} \in \Omega} \left(\left\| \boldsymbol{G}_n(\boldsymbol{\theta}) \right\| \right)$, the parameter space Ω is compact

b) $\left\| \boldsymbol{G}_n(\boldsymbol{\theta}_0) \right\| = o_p(1)$,

c) $\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \delta} \left(\dfrac{1}{\left\| \boldsymbol{G}_n(\boldsymbol{\theta}) \right\|} \right) = O_p(1)$ for each $\delta > 0$.

Theorem 3.1 states condition b) as $\boldsymbol{G}_n(\boldsymbol{\theta}_0) = o_p(1)$ but in the proof the authors just use $\left\| \boldsymbol{G}_n(\boldsymbol{\theta}_0) \right\| = o_p(1)$ so we state condition b) as $\left\| \boldsymbol{G}_n(\boldsymbol{\theta}_0) \right\| = o_p(1)$ as it is easier to use this condition when there is a need to extend to the infinite dimensional case with the space l^2.

An expression is $o_p(1)$ if it converges to 0 in probability and $O_p(1)$ if it is bounded in probability. In version D and version S for Hellinger distance we have $\inf_{\boldsymbol{\theta} \in \Omega} \left(\left\| \boldsymbol{G}_n(\boldsymbol{\theta}) \right\| \right)$ occurs at the values of the vector values of the HD estimators, so the conditions a) and b) are satisfied for both versions and compactness of the parameter space Ω is assumed. Also, for both versions $\left\| \boldsymbol{G}_n(\boldsymbol{\theta}) \right\| \xrightarrow{p} 0$ only at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $0 < Q_n(\boldsymbol{\theta}) \le 2$ otherwise, this implies that there exist real numbers u and v with $0 < u < v < \infty$ such that

$$P\left(u \le \sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \delta} \left(\frac{1}{\left\| \boldsymbol{G}_n(\boldsymbol{\theta}) \right\|} \right) \le v \right) \to 1 \text{ as } n \to \infty.$$

Therefore, for both versions of $Q_n(\boldsymbol{\theta})$ whether deterministic or simulated,

the minimum Hellinger distance estimators (MHD) are consistent. Theorem 3.1 of Pakes and Pollard [16] is an elegant theorem, its proof is also concise using the norm concept of functional analysis and it allows many results to be unified. Essentially, the same theorem remains valid with the use of the Hilbert space l^2 and its norm instead of the Euclidean space R^m and the Euclidean norm. By using l^2 and its norm the consistency for the ungrouped SMHD estimators can also be established but further asymptotic results for the ungrouped SMHD estimators will be postponed and given in Section 3.

Asymptotic normality is more complicated in general. For the grouped case, Theorem 3.3 given by Pakes and Pollard [16] (p1040) can be used to establish asymptotic normality for both versions of Hellinger distance estimators. We shall rearrange results of Theorem 3.3 under Theorem 2 and Corollary 1 given in the next section to make it easier to apply for HD estimation using both versions.

Since the proofs have been given by the authors, we only discuss here the ideas of their proofs to make it easier to follow the results of Theorem 2 and Corollary 1 in Section (2.2.2).

For both versions, $Q_n(\boldsymbol{\theta}) = \left(\left\|\boldsymbol{G}_n(\boldsymbol{\theta})\right\|\right)^2 = \left(\boldsymbol{G}_n(\boldsymbol{\theta})\right)'\left(\boldsymbol{G}_n(\boldsymbol{\theta})\right)$ but $\boldsymbol{G}_n(\boldsymbol{\theta})$ is not differentiable for version S, the traditional Taylor expansion argument cannot be used to establish asymptotic normality of estimators obtained by minimizing $\left(\left\|\boldsymbol{G}_n(\boldsymbol{\theta})\right\|\right)^2$. If we assume $\boldsymbol{G}(\boldsymbol{\theta})$ is differentiable with derivative matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta})$, then we can define the random function $Q_n^a(\boldsymbol{\theta})$ to approximate $Q_n(\boldsymbol{\theta})$ with

$$Q_n^a(\boldsymbol{\theta}) = \left(\left\|\boldsymbol{L}_n(\boldsymbol{\theta})\right\|\right)^2,\ \boldsymbol{L}_n(\boldsymbol{\theta}) = \boldsymbol{G}_n(\boldsymbol{\theta}_0) + \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0). \tag{11}$$

$\boldsymbol{G}_n(\boldsymbol{\theta}_0)$ is based on expression (8) for version D and it is based on expressions (9-10) for version S. Note that $Q_n^a(\boldsymbol{\theta})$ is differentiable for both versions.

Let $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ be the vectors which minimize $Q_n(\boldsymbol{\theta})$ and $Q_n^a(\boldsymbol{\theta})$ respectively. If the approximation is of the right order then $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ are asymptotically equivalent. This set up will allow a unified approach for establishing asymptotic for MHD estimation for both versions. For version D, it suffices to let $\tilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}_G}$ and for version S, let $\tilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}_G^S}$.

Under these circumstances, it suffices to work with $\boldsymbol{\theta}^*$ and $Q_n^a(\boldsymbol{\theta}^*)$ for asymptotic properties of $\tilde{\boldsymbol{\theta}}$ and $Q_n^a(\boldsymbol{\theta}^*)$. A regularity condition for the approximation is of the right order which implies the condition (iii) given by their Theorem 3.3, which is the most difficult to check is given as

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \le \delta_n} \sqrt{n}\left\|\boldsymbol{G}_n(\boldsymbol{\theta}) - \boldsymbol{G}(\boldsymbol{\theta}) - \boldsymbol{G}_n(\boldsymbol{\theta}_0)\right\| = o_p(1). \tag{12}$$

This condition is used to formulate Theorem 2 below and is slightly more stringent than the condition iii) of their Theorem 3.3 but it is less technical and sufifcient for SMHD estimation. Clearly, for SMHD estimation $\boldsymbol{G}_n(\boldsymbol{\theta})$ is as given by expression (9) or expression (10). For simulated unweighted simulated

minimum chi-square estimation for this condition to hold, independent samples for each θ cannot be used, see Pakes and Pollard [16] (p1048). Otherwise, only consistency can be guaranteed for estimators using version S. For version S, the simulated samples are assumed to have size $U=\tau n$ and the same seed is used across different values of θ to draw samples of size *U*. We implicitly make these assumptions for SMHD methods. These two assumption are standard for simulated methods of inferences, see section 9.6 for method of simulated moments(MSM) given by Davidson and McKinnon [19] (p383-394). For numerical optimization to find the minimum of the objective function $Q_n(\boldsymbol{\theta})$, we rely on direct search simplex methods which are derivative free and the R package already has prewritten functions to implement direct search methods.

2.2.2. Asymptotic Normality

In this section, we shall state a Theorem namely Theorem 2 which is essentially Theorem 3.3 by Pakes and Pollard [16] (p1040-1043) with the condition (4) of Theorem 2 given by expression (9) replacing their condition (iii) in their Theorem 3.3, the condition (4) implies the condition (iii) by being more stringent. We also comment on the conditions needed to verify asymptotic normality for the HD estimators based on Theorem 2.

Theorem 2

Let $\tilde{\boldsymbol{\theta}}$ be a vector of consistent estimators for $\boldsymbol{\theta}_0$, the unique vector which satisfies $\boldsymbol{G}(\boldsymbol{\theta}_0)=\mathbf{0}$.

Under the following conditions:

1) The parameter space Ω is compact, $\tilde{\boldsymbol{\theta}}$ is an interior point of Ω.

2) $\left\|\boldsymbol{G}_n(\tilde{\boldsymbol{\theta}})\right\| \le o_p\left(n^{-\frac{1}{2}}\right)+\inf_{\boldsymbol{\theta}\in\Omega}\left\|\boldsymbol{G}_n(\boldsymbol{\theta})\right\|$.

3) $\boldsymbol{G}(.)$ is differentiable at $\boldsymbol{\theta}_0$ with a derivative matrix $\boldsymbol{\Gamma}=\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)$ of full rank.

4) $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n}\sqrt{n}\left\|\boldsymbol{G}_n(\boldsymbol{\theta})-\boldsymbol{G}(\boldsymbol{\theta})-\boldsymbol{G}_n(\boldsymbol{\theta}_0)\right\|=o_p(1)$ for every sequence $\{\delta_n\}$ of positive numbers which converge to zero.

5) $\left\|\boldsymbol{G}_n(\boldsymbol{\theta}_0)\right\|=o_p(1)$.

6) $\boldsymbol{\theta}_0$ is an interior point of the parameter space Ω.

Then, we have the following representation which will give the asymptotic distribution of $\tilde{\theta}$ in Corollary 1, *i.e.*,

$$\sqrt{n}\left(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=-\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0)+o_p(1), \tag{13}$$

or equivalently, using equality in distribution,

$$\sqrt{n}\left(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=^{d}-\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\sqrt{n}\boldsymbol{\Gamma}'\boldsymbol{G}_n(\boldsymbol{\theta}_0). \tag{14}$$

The proofs of these results follow from the results used to prove Theorem 3.3 given by Pakes and Pollard [16] (p1040-1043). For expression (13) or expression (14) to hold, in general only condition 5) of Theorem 2 is needed and there is no need to assume that $G_n(\boldsymbol{\theta}_0)$ has an asymptotic distribution. From the results of

Theorem 2, it is easy to see that we can obtain the main result of the following Corollary 1 which gives the asymptotic covariance matrix for the HD estimators for both versions.

Corollary 1.

Let $\boldsymbol{Y}_n = \sqrt{n}\boldsymbol{\Gamma}'\boldsymbol{G}_n(\boldsymbol{\theta}_0)$, if $\boldsymbol{Y}_n \xrightarrow{L} N(\boldsymbol{0},\boldsymbol{V})$ then $\sqrt{n}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \xrightarrow{L} N(\boldsymbol{0},\boldsymbol{T})$ with

$$\boldsymbol{T} = (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{V}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1},$$

The matrices $\boldsymbol{T}$ and $\boldsymbol{V}$ depend on $\boldsymbol{\theta}_0$ we also adopt the notations $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{\theta}_0)$, $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta}_0)$.

We observe that condition 4) of Theorem 2 when applies to Hellinger distance or in general involve technicalities. The condition 4) holds for version D, we only need to verify for version S. Note that to verify the condition 4, it is equivalent to verify

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n} n\left(\left\|\boldsymbol{G}_n(\boldsymbol{\theta})-\boldsymbol{G}(\boldsymbol{\theta})-\boldsymbol{G}_n(\boldsymbol{\theta}_0)\right\|\right)^2 = o_p(1)$$

and for version S of Hellinger distance estimation, let

$$g_n(\boldsymbol{\theta}) = n\left(\left\|\boldsymbol{G}_n(\boldsymbol{\theta})-\boldsymbol{G}(\boldsymbol{\theta})-\boldsymbol{G}_n(\boldsymbol{\theta}_0)\right\|\right)^2$$

and for the grouped case, it is given by

$$g_n(\boldsymbol{\theta}) = n\sum_{i=0}^{k}\left(\left(\left[p_{\boldsymbol{\theta}_0}^S(I_j)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}_0}(I_j)\right]^{\frac{1}{2}}\right)-\left(\left[p_{\boldsymbol{\theta}}^S(I_j)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}}(I_j)\right]^{\frac{1}{2}}\right)\right)^2. \quad (15)$$

We need to verify that we have the sequence of functions $\{g_n(\boldsymbol{\theta})\}$ converge uniformly to 0 in probability as $n\to\infty$ and $\boldsymbol{\theta}\to\boldsymbol{\theta}_0$ or equivalently, $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n} g_n(\boldsymbol{\theta}) \xrightarrow{p} 0$ as $n\to\infty$ and $\boldsymbol{\theta}\to\boldsymbol{\theta}_0$.

Note that

$$g_n(\boldsymbol{\theta}) = n\sum_{i=0}^{k}\left(\left(\left[p_{\boldsymbol{\theta}_0}^S(I_j)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}_0}(I_j)\right]^{\frac{1}{2}}\right)^2+\left(\left[p_{\boldsymbol{\theta}}^S(I_j)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}}(I_j)\right]^{\frac{1}{2}}\right)^2\right.$$
$$\left.-2\left(\left[p_{\boldsymbol{\theta}_0}^S(I_j)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}_0}(I_j)\right]^{\frac{1}{2}}\right)\left(\left[p_{\boldsymbol{\theta}}^S(I_j)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}}(I_j)\right]^{\frac{1}{2}}\right)\right),\ g_n(\boldsymbol{\theta})\ge 0.$$

We shall outline the approach by first defining the notion of continuity in probability and let $S(\boldsymbol{\theta}_0,\delta_n) = \{\boldsymbol{\theta}\,|\,\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n\}$ which is a compact set. The compactness of this set simplifies proofs and does not appear to be used in previous approaches in the literature. Observe that $g_n(\boldsymbol{\theta}_0)=0$, it is easy to see that $g_n(\boldsymbol{\theta}) \xrightarrow{p} g_n(\boldsymbol{\theta}_0)=0$ as $\boldsymbol{\theta}\to\boldsymbol{\theta}_0$. Subsequently we establish $g_n(\boldsymbol{\theta})$ being continuous in probability for $\boldsymbol{\theta}$ and using the property that $\{g_n(\boldsymbol{\theta})\}$ is continuous in probability $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n} g_n(\boldsymbol{\theta})$ is attained at a point $\boldsymbol{\theta}=\boldsymbol{\theta}^0$ which belongs to the compact set $S(\boldsymbol{\theta}_0,\delta_n)$ in probability. This is similar to the property of nonrandom continuous function in real analysis.

Now as $n \to \infty, \delta_n \to 0$ which implies $\boldsymbol{\theta}^0 \to \boldsymbol{\theta}_0$ and by continuity in probability $g_n(\boldsymbol{\theta}_n^0) \xrightarrow{p} g_n(\boldsymbol{\theta}_0) = 0$. Therefore, $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \le \delta_n} g_n(\boldsymbol{\theta}) \xrightarrow{p} 0$ which means that $\{g_n(\boldsymbol{\theta})\}$ converges uniformly in probability as $n \to \infty$. The technical details of these arguments are given in technical appendices **TA1.1** and **TA1.2** at the end of the paper, in the section of **Appendices**.

The notion of continuity in probability has been used in a similar context in the literature of stochastic processes, see Gusalk *et al.* [25] and will be introduced in the next paragraph and we also make a few assumptions which are summarized by Assumption 1 and Assumption 2 given below along with the notion of continuity in probability. A related continuity notion namely the notion of continuity with probability one has been mentioned by Newey and McFadden [18] in their Theorem 2.6 as mentioned earlier. They also commented that this notion can be used for establishing asymptotic properties of simulated estimators introduced by Pakes [26]. Pakes [26] also has used pseudo random numbers to estimate probability frequencies for some models. For SMHD estimation, we extend a standard result of analysis which states that a continuous function attains its supremum on a compact set to a version which holds in probability. This approach seems to be new and simpler than the use of the more general stochastic equicontinuity condition given by section 2.2 in Newey and McFadden [18] (p2136-2138) to establish uniform convergence of a sequence of random functions in probability. Our approach uses the fact that as $n \to \infty$ the set $S(\boldsymbol{\theta}_0, \delta_n)$ shrinks to $\boldsymbol{\theta}_0$, a property which did not seem to have been used previously by other approaches to establish $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \le \delta_n} g_n(\boldsymbol{\theta}) \xrightarrow{p} 0$ as $n \to \infty$ and $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$. Subsequently, we define the notion of continuity in probability which is similar to the one used in stochastic processes, see Gusak *et al.* [25] (p33) for a related notion of continuity in probability for stochastic processes.

Definition 1 (Continuity in probability)

A sequence of random functions $\{g_n(\boldsymbol{\theta})\}$ is continuous in probability at $\boldsymbol{\theta}'$ if $g_n(\boldsymbol{\theta}) \xrightarrow{p} g_n(\boldsymbol{\theta}')$ whenever $\boldsymbol{\theta} \to \boldsymbol{\theta}'$. Equivalently, for any $\epsilon > 0, \delta_1 > 0$, there exists a $\delta \ge 0$ and n_0 such that $P(|g_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta}')| \le \epsilon) \ge 1 - \delta_1$ for $n \ge n_0$, whenever $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \le \delta$. This can be viewed as an extension of the classical result of continuity in real analysis. It is also well known that the supremum of a continuous function on a compact domain is attained at a point of the compact domain, see Davidson and Donsig [27] (p81) or Rudin [28] (p89) for this classical result. The equivalent property for a random function which is only continuous in probability is the supremum of the random function is attained at a point of the compact domain in probability. The compact domain we study here is given by $S(\boldsymbol{\theta}_0, \delta_n) = \{\boldsymbol{\theta} | \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \delta_n\}$ and as $n \to \infty$, $S(\boldsymbol{\theta}_0, \delta_n) \to \boldsymbol{\theta}_0$. It might be more precise to use the term sequence of random functions rather than just random function here for the notion of continuity in probability as the ran-

dom function will depend on n.

Below are the assumptions we need to make to establish asymptotic normality for SMHD estimators and they appear to be reasonable.

Assumption 1

1) The pmf of the parametric model has the continuity property with $p_{\boldsymbol{\theta}}^{\frac{1}{2}}(i) \to p_{\boldsymbol{\theta}'}^{\frac{1}{2}}(i)$ whenever $\boldsymbol{\theta} \to \boldsymbol{\theta}'$.

2) The simulated counterpart has the continuity in probability property with $\left[p_{\boldsymbol{\theta}}^{S}\right]^{\frac{1}{2}} \xrightarrow{p} \left[p_{\boldsymbol{\theta}'}^{S}\right]^{\frac{1}{2}}$ whenever $\boldsymbol{\theta} \to \boldsymbol{\theta}'$. Convergence in probability is denote by $\xrightarrow{p}$.

3) $p_{\boldsymbol{\theta}}^{\frac{1}{2}}(i)$ is differentiable with respect to $\boldsymbol{\theta}$.

In general, the condition 2) will be satisfied if the condition 1) holds and implicitly we assume the same seed is used for obtaining the simulated samples across different values of $\boldsymbol{\theta}$. For ungrouped data, we also need the notion of differentiability in probability to facilitate the application of Theorem 7.1 given by Newey and McFadden (1994, p2185-2186). Before stating their Theorem 7.1, Newey and McFadden has mentioned the notion of approximate derivative for the use of their Theorem, the definition given below will make it clearer.

Definition 2 (Differentiability in probability)

The sequence of random functions $\left\{f_{\theta}^{(n)}\right\}$ is differentiable with respect to $\boldsymbol{\theta}$ **at** $\boldsymbol{\theta}_0$ in probability if $\lim_{\epsilon \to 0} \dfrac{f_{\boldsymbol{\theta}_0+\epsilon e_j}^{(n)} - f_{\boldsymbol{\theta}_0}^{(n)}}{\varepsilon} =^{p} v_j(\boldsymbol{\theta}_0)$, $j = 1, \cdots, m$ exists and $e_i = (0, 0, \cdots, 1, 0, \cdots, 0)'$ with 1 occurring at the ith entry. Furthermore, the vector $v(\theta) = \left(v_1(\boldsymbol{\theta}), \cdots, v_m(\boldsymbol{\theta})\right)'$ is continuous and bounded in probability for all $\boldsymbol{\theta} \in S(\boldsymbol{\theta}_0, \delta_0)$ for some $\delta_0 > 0$. This concept is similar to the notion of differentiability in real analysis for nonrandom function.

A similar notion of differentiability in probability has been used in stochastic processes literature, see Gusak *et al.* [25] (p33-34), a more stringent differentiability notion namely differentiability in quadratic mean has also been used to study local asymptotic normality (LAN) property for a parametric family, see Keener [29] (p326). The notion of differentiability in probability will be used in section 3 with Theorem 7.1 of Newey and McFadden [17] to establish asymptotic normality for the SMHD estimators for the ungrouped case. We make the following assumption for $\left[p_{\boldsymbol{\theta}}^{S}\right]^{\frac{1}{2}}$ where $p_{\boldsymbol{\theta}}^{S}$ can be viewed as a proxy model for $p_{\boldsymbol{\theta}}$,

Assumption 2

$\left[p_{\boldsymbol{\theta}}^{S}(i)\right]^{\frac{1}{2}}$ with the same seed being used across different values of $\boldsymbol{\theta}$ is differentiable in probability with the same derivative vector as $p_{\boldsymbol{\theta}}^{\frac{1}{2}}(i)$ where the

derivative vector for $p_{\boldsymbol{\theta}}^{\frac{1}{2}}(i)$ is

$$s_{\theta}(i)=\left(\frac{\partial p_{\boldsymbol{\theta}}^{\frac{1}{2}}(i)}{\partial\theta_1},\cdots,\frac{\partial p_{\boldsymbol{\theta}}^{\frac{1}{2}}(i)}{\partial\theta_m}\right).$$

This assumption appears to be reasonable, this can be checked by using limit operations as in real analysis with $\left[p_{\boldsymbol{\theta}}^{S}(i)\right]^{\frac{1}{2}}\xrightarrow{p}p_{\boldsymbol{\theta}}^{\frac{1}{2}}(i)$ and $\left[p_{\boldsymbol{\theta}}^{S}(i)\right]^{\frac{1}{2}}$ is continuous in probability.

Since regularity conditions for Theorem 2 and its corollary can be met and they are justified in **TA1.1** and **TA1.2** in the Appendices, we proceed here to find the asymptotic covariance matrix $\boldsymbol{T}$.

Since $\boldsymbol{G}_n(\boldsymbol{\theta}_0)$ for version D is based on expression (8) and for version S is based on expressions (9-10), the asymptotic covariance matrix of $\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0)$ version S is just the asymptotic covariance matrix of $\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0)$ of version D multiplied by $\left(1+\frac{1}{\tau}\right)$ as the simulated sample from $p_{\boldsymbol{\theta}}(x)$ is independent from the sample given by the data, so we can focus on version D and make the adjustment for version S. We need the asymptotic covariance matrix $\boldsymbol{\Sigma}$ of the vector $\sqrt{n}\boldsymbol{u}_n=\sqrt{n}\left(p_n(I_0),\cdots,p_n(I_k)\right)'$ first then we can find the matrix $\boldsymbol{T}$ and we let $\boldsymbol{T}=\boldsymbol{T}_D$ for version D and for version S, we shall let $\boldsymbol{T}=\boldsymbol{T}_S$.

Recall that form properties of the multinomial distribution, the covariances of $\sqrt{n}p_n(I_i)$ and $\sqrt{n}p_n(I_j)$ are

$$nCov\left(p_n(I_i),p_n(I_j)\right)=\left(p_{\boldsymbol{\theta}_0}(I_i)\right)\left(1-p_{\boldsymbol{\theta}_0}(I_i)\right),\ \text{for } i=j$$

and

$$nCov\left(p_n(I_i),p_n(I_j)\right)=-\left(p_{\boldsymbol{\theta}_0}(I_i)\right)\left(p_{\boldsymbol{\theta}_0}(I_j)\right),\ \text{for } i\neq j.$$

The covariance matrix of $\sqrt{n}\boldsymbol{u}_n=\sqrt{n}\left(p_n(I_0),\cdots,p_n(I_k)\right)'$ using matrix notations can be expressed as

$$\boldsymbol{\Sigma}=\boldsymbol{Q}^{-\frac{1}{2}}\left(\boldsymbol{I}-\boldsymbol{q}\boldsymbol{q}'\right)\boldsymbol{Q}^{-\frac{1}{2}},\tag{16}$$

$\boldsymbol{Q}^{-\frac{1}{2}}$ is a diagonal matrix with diagonal elements $\left(\left[p_{\boldsymbol{\theta}_0}(I_0)\right]^{\frac{1}{2}},\cdots,\left[p_{\boldsymbol{\theta}_0}(I_k)\right]^{\frac{1}{2}}\right)'$ and the vector $\boldsymbol{q}'=\left(\left[p_{\boldsymbol{\theta}_0}(I_0)\right]^{\frac{1}{2}},\cdots,\left[p_{\boldsymbol{\theta}_0}(I_k)\right]^{\frac{1}{2}}\right)$ is the transpose of $\boldsymbol{q}$ and $\boldsymbol{I}$ is the identity matrix of dimension $r\times r$ with $r=k+1$. Using the delta method the asymptotic covariance matrix of $\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0)$ of version D is simply the asymptotic covariance matrix of $\frac{1}{2}\boldsymbol{Q}^{\frac{1}{2}}\sqrt{n}\boldsymbol{u}_n$ which is given by

$$\boldsymbol{W}_D = \frac{1}{4}\left(\boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}'\right), \tag{17}$$

and the asymptotic covariance matrix of $\sqrt{n}\boldsymbol{G}_n\left(\boldsymbol{\theta}_0\right)$, version S is

$$\boldsymbol{W}_S = \frac{1}{4}\left(1+\frac{1}{\tau}\right)\left(\boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}'\right). \tag{18}$$

We then have the vector of HD estimators version D and S given respectively by $\widehat{\boldsymbol{\theta}_G}$ and $\hat{\boldsymbol{\theta}}_G^S$ with asymptotic distributions given by

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}_G} - \boldsymbol{\theta}_0\right) \xrightarrow{L} N\left(\boldsymbol{0}, \boldsymbol{T}_D\right),$$

$$\boldsymbol{T}_D = \frac{1}{4}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\left(\boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}'\right)\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1} = \frac{1}{4}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1} = \boldsymbol{I}_G\left(\boldsymbol{\theta}_0\right)^{-1}, \tag{19}$$

$$\boldsymbol{\Gamma} = -\frac{1}{2}\begin{bmatrix} \dfrac{\partial p_{\boldsymbol{\theta}_0}\left(I_0\right)}{\partial\theta_1} & \cdots & \dfrac{\partial p_{\boldsymbol{\theta}_0}\left(I_0\right)}{\partial\theta_m} \\ \left[p_{\boldsymbol{\theta}_0}\left(I_0\right)\right]^{\frac{1}{2}} & & \left[p_{\boldsymbol{\theta}_0}\left(I_0\right)\right]^{\frac{1}{2}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial p_{\boldsymbol{\theta}_0}\left(I_k\right)}{\partial\theta_1} & \cdots & \dfrac{\partial p_{\boldsymbol{\theta}_0}\left(I_k\right)}{\partial\theta_m} \\ \left[p_{\boldsymbol{\theta}_0}\left(I_k\right)\right]^{\frac{1}{2}} & & \left[p_{\boldsymbol{\theta}_0}\left(I_k\right)\right]^{\frac{1}{2}} \end{bmatrix},$$

$\boldsymbol{I}_G\left(\boldsymbol{\theta}_0\right)$ is the model Fisher information matrix using grouped data as $\boldsymbol{q}'\boldsymbol{\Gamma} = \boldsymbol{0}$ due to $\sum_{i=0}^{k}\frac{\partial p_{\boldsymbol{\theta}}\left(I_i\right)}{\partial\boldsymbol{\theta}} = 0$ using $\sum_{i=0}^{k} p_{\boldsymbol{\theta}}\left(I_i\right) = 1$. Let $\boldsymbol{B} = -2\boldsymbol{\Gamma}$,

$$\boldsymbol{B} = \begin{bmatrix} \dfrac{\partial p_{\boldsymbol{\theta}_0}\left(I_0\right)}{\partial\theta_1} & \cdots & \dfrac{\partial p_{\boldsymbol{\theta}_0}\left(I_0\right)}{\partial\theta_m} \\ \left[p_{\boldsymbol{\theta}_0}\left(I_0\right)\right]^{\frac{1}{2}} & & \left[p_{\boldsymbol{\theta}_0}\left(I_0\right)\right]^{\frac{1}{2}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial p_{\boldsymbol{\theta}_0}\left(I_k\right)}{\partial\theta_1} & \cdots & \dfrac{\partial p_{\boldsymbol{\theta}_0}\left(I_k\right)}{\partial\theta_m} \\ \left[p_{\boldsymbol{\theta}_0}\left(I_k\right)\right]^{\frac{1}{2}} & & \left[p_{\boldsymbol{\theta}_0}\left(I_k\right)\right]^{\frac{1}{2}} \end{bmatrix},$$

so

$$\boldsymbol{T}_D = \left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}$$

with

$$\frac{\dfrac{\partial p_{\boldsymbol{\theta}_0}\left(I_i\right)}{\partial\theta_j}}{\left[p_{\boldsymbol{\theta}_0}\left(I_i\right)\right]^{\frac{1}{2}}} = \left[p_{\boldsymbol{\theta}_0}\left(I_i\right)\right]^{\frac{1}{2}}\frac{\partial \log p_{\boldsymbol{\theta}_0}\left(I_i\right)}{\partial\theta_j}, i = 0,1,\cdots,k,\ j = 1,\cdots,m.$$

Therefore for version *S*,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_G^S - \boldsymbol{\theta}_0\right) \xrightarrow{L} N\left(0, \boldsymbol{T}_S\right),\ \boldsymbol{T}_S = \left(1+\frac{1}{\tau}\right)\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1} \tag{20}$$

the simulated sample size is $U = n\tau$.

Note that for version D, the HD estimators are as efficient as the minimum chi-square estimators or ML estimators based on grouped data. The overall asymptotic relative efficiency (ARE) between version D and S for HD estimation is simply ARE = $\frac{\tau}{\tau+1}$ and we recommend to set $\tau \geq 10$ to minimize the loss of efficiency due to simulations.

An estimate for the covariance matrix

The asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}_G^S$ can be estimated if we can estimate $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)$. Using a result given by Pakes and Pollard (1989, p1043), an estimate for $\boldsymbol{\Gamma}$ is the matrix

$$\hat{\boldsymbol{\Gamma}}_n = \left[\frac{G_n\left(\widehat{\boldsymbol{\theta}}_G^S + \epsilon_n e_1\right) - G_n\left(\widehat{\boldsymbol{\theta}}_G^S\right)}{\epsilon_n}, \cdots, \frac{G_n\left(\widehat{\boldsymbol{\theta}}_G^S + \epsilon_n e_m\right) - G_n\left(\widehat{\boldsymbol{\theta}}_G^S\right)}{\epsilon_n} \right], \quad (21)$$

$e_i = (0,0,\cdots,1,0,\cdots,0)'$ with 1 occurring at the ith entry of the vector $e_i, i = 1,\cdots,m$ and $\epsilon_n = n^{-\delta}$, $\delta \leq \frac{1}{2}$ and in general we can let $\delta = \frac{1}{2}$. Note that the columns of $\hat{\boldsymbol{\Gamma}}_n$ estimate the corresponding partial derivatives given by the columns of Γ.

For ungrouped data and for version D, it is equivalent to choose $I_j = [j, j+1)$ with unit length and let $k = \infty$. If we choose $I_j = [j, j+1)$ and let $k \to \infty$ and note that $(\boldsymbol{B}'\boldsymbol{B}) \to \boldsymbol{I}(\boldsymbol{\theta}_0)$ and $\boldsymbol{I}(\boldsymbol{\theta}_0)$ the is Fisher information matrix for ungrouped data with elements given by

$$I_{h,l} = \sum_{i=0}^{\infty} p_{\boldsymbol{\theta}}(i) \left(\frac{\partial \log p_{\boldsymbol{\theta}}(i)}{\partial \theta_h} \frac{\partial \log p_{\boldsymbol{\theta}}(i)}{\partial \theta_l} \right), l = 1,\cdots,m, h = 1,\cdots,m \quad (22)$$

and $p_{\boldsymbol{\theta}}(i) = p_{\boldsymbol{\theta}}(I_i), I_i = [i, i+1), \boldsymbol{\theta} = \boldsymbol{\theta}_0$. We can foresee that the HD estimators are as efficient as ML estimators for version D, a result which is already obtained by Simpson [14]. We postpone till section (3) for a more rigorous approach to justify the related result for version S using Theorem 7.1 given by Newey and McFadden [17]. The SMHD estimators given by $\widehat{\boldsymbol{\theta}}^S$ for ungrouped data will be shown to have the property

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^S - \boldsymbol{\theta}_0\right) \xrightarrow{L} N\left(0, \left(1 + \frac{1}{\tau}\right)\left(\boldsymbol{I}(\boldsymbol{\theta}_0)\right)^{-1}\right).$$

Section 3 may be skipped for practitioners if their main interests are only on applications of the results.

2.3. Chi-Square Goodness of Fit Test Statistics

2.3.1. Simple Hypothesis

In this section, the Hellinger distance $Q_n(\boldsymbol{\theta})$ is used to construct goodness of fit test statistics for the simple hypothesis

H_0: data comes from a specified distribution with distribution $F_{\boldsymbol{\theta}_0}$, $F_{\boldsymbol{\theta}_0}$ can

be the distribution of a discrete or continuous distribution. The chi-square test statistics and their asymptotic distributions are given below with

$$4nQ_n(\boldsymbol{\theta}_0) \xrightarrow{L} \chi^2\left(r=(k+1)-1\right) \text{ for version } D \text{ and} \tag{23}$$

$$4n\left(\frac{\tau}{\tau+1}\right)Q_n(\boldsymbol{\theta}_0) \xrightarrow{L} \chi^2\left(r=(k+1)-1\right) \text{ for version } S. \tag{24}$$

The version S is of interest since it allows testing goodness of fit for discrete or continuous distribution without closed form pmfs or density functions, all we need is to be able to simulate from the specified distribution. We shall justify the asymptotic chi-square distributions given by expression (23) and expression (24) below.

Note that

$$4nQ_n(\boldsymbol{\theta}_0) = \sqrt{n}\boldsymbol{G}_n'(\boldsymbol{\theta}_0)\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0)$$

and

$$\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) \xrightarrow{L} N\left(0, \frac{1}{4}(\boldsymbol{I}-\boldsymbol{q}\boldsymbol{q}')\right)$$

for version D. For version S,

$$\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) \xrightarrow{L} N\left(0, \frac{1}{4}\left(1+\frac{1}{\tau}\right)(\boldsymbol{I}-\boldsymbol{q}\boldsymbol{q}')\right).$$

Using standard results for distribution of quadratic forms and the property of the operator trace of a matrix with $trace(\boldsymbol{I}-\boldsymbol{q}\boldsymbol{q}') = trace(\boldsymbol{I}) - trace(\boldsymbol{q}\boldsymbol{q}') = (k+1)-1 = k$, see Luong and Thompson [30] (p247); we have the asymptotic chi-square distributions as given by expression (23) and expression (24). On how to choose the intervals, the problem is rather complex as it depends on the type of alternatives we would like to detect. We can also follow the recommendations of the Pearson's statistics, see Greenwood and Nikulin [31]; also see Lehmann [32] (p341) for more discussions and references on this issue.

2.3.2. Composite Hypothesis

Just as the chi-square distance, the Hellinger distance $Q_n(\boldsymbol{\theta})$ can also be used for construction of the test statistics for the composite hypothesis,

H_0: data comes from a parametric model $\{F_{\boldsymbol{\theta}}\}$, $\{F_{\boldsymbol{\theta}}\}$ can be a discrete or continuous parametric model. The chi-square test statistics are given by

$$4nQ_n\left(\widehat{\boldsymbol{\theta}_G}\right) \xrightarrow{L} \chi^2\left(r=k-m\right), \tag{25}$$

for version D and for version S,

$$4n\left(\frac{\tau}{\tau+1}\right)Q_n\left(\widehat{\boldsymbol{\theta}_G^S}\right) \xrightarrow{L} \chi^2\left(r=k-m\right) \tag{26}$$

where $\widehat{\boldsymbol{\theta}_G}$ and $\widehat{\boldsymbol{\theta}_G^S}$ are the vector of HD estimators which minimize $Q_n(\boldsymbol{\theta})$ version D and version S respectively and assuming $k>m$. To justify these

asymptotic chi-square distributions, note that we have for version D, $4nQ_n\left(\widehat{\boldsymbol{\theta}_G}\right)=4nQ_n^a\left(\widehat{\boldsymbol{\theta}_G}\right)+o_p(1)$. It suffices to consider the asymptotic distribution of $4nQ_n^a\left(\widehat{\boldsymbol{\theta}_G}\right)$ as we have the following equalities in distribution,

$4nQ_n\left(\widehat{\boldsymbol{\theta}_G}\right)=^d nQ_n^a\left(\widehat{\boldsymbol{\theta}_G}\right)=4n\left\|L_n\left(\widehat{\boldsymbol{\theta}_G}\right)\right\|^2$, $L_n(\boldsymbol{\theta})$ as given by expression (11). Also, using expression (11) and expression (13),

$$\sqrt{n}\boldsymbol{L}_n\left(\widehat{\boldsymbol{\theta}_G}\right)=^d \sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0)+\boldsymbol{\Gamma}\sqrt{n}\left(\widehat{\boldsymbol{\theta}_G}-\boldsymbol{\theta}_0\right)$$

which can be reexpressed as

$$\sqrt{n}\boldsymbol{L}_n\left(\widehat{\boldsymbol{\theta}_G}\right)=^d \sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0)-\sqrt{n}\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\boldsymbol{G}_n(\boldsymbol{\theta}_0)$$

or

$$\sqrt{n}\boldsymbol{L}_n\left(\widehat{\boldsymbol{\theta}_G}\right)=^d \left(\boldsymbol{I}-\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\right)\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0)$$

With

$$\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0)\xrightarrow{L}N\left(0,\frac{1}{4}(\boldsymbol{I}-\boldsymbol{q}\boldsymbol{q}')\right),$$

$\boldsymbol{G}_n(\boldsymbol{\theta}_0)$ is based on expression (8) for version D. Consequently,

$$\sqrt{n}\boldsymbol{L}_n\left(\widehat{\boldsymbol{\theta}_G}\right)\xrightarrow{L}N\left(0,\frac{1}{4}\left(\boldsymbol{I}-\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'-\boldsymbol{q}\boldsymbol{q}'\right)\right)$$

by noting

$$\frac{1}{4}\left(\boldsymbol{I}-\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\right)(\boldsymbol{I}-\boldsymbol{q}\boldsymbol{q}')\left(\boldsymbol{I}-\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\right)=\frac{1}{4}\left(\boldsymbol{I}-\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'-\boldsymbol{q}\boldsymbol{q}'\right),$$

using $\boldsymbol{q}'\boldsymbol{\Gamma}=0$ and the matrix $\boldsymbol{B}=\left(\boldsymbol{I}-\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'-\boldsymbol{q}\boldsymbol{q}'\right)$ is of rank $k+1-m-1=k-m$ with the rank of the matrix $\boldsymbol{B}$ is also equal to its trace. The argument used is very similar to the one used for the Pearson's statistics, see Luong and Thompson [30] (p249).

For version S,

$$4nQ_n\left(\widehat{\boldsymbol{\theta}_G^S}\right)=^d 4nQ_n^a\left(\widehat{\boldsymbol{\theta}_G^S}\right)=4n\left\|\boldsymbol{L}_n\left(\widehat{\boldsymbol{\theta}_G^S}\right)\right\|^2$$

and

$$\sqrt{n}\boldsymbol{L}_n\left(\widehat{\boldsymbol{\theta}_G^S}\right)=^d \left(\boldsymbol{I}-\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\right)\sqrt{n}G_n(\boldsymbol{\theta}_0)$$

with

$$\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0)\xrightarrow{L}N\left(0,\frac{1}{4}\left(1+\frac{1}{\tau}\right)(\boldsymbol{I}-\boldsymbol{q}\boldsymbol{q}')\right),$$

$\boldsymbol{G}_n(\boldsymbol{\theta}_0)$ is based on expressions (9-10) for version S. This justifies the asymptotic chi-square distribution for version S as given by expression (25) and expression (26). This version is useful for model testing for nonnegative continuous models without closed form expression densities, see Luong [33]

for some positive infinitely divisible distributions without closed form densities used in actuarial sciences. It is also suitable for testing models with support on the real line used in finance such as the jump diffusion model as given by Tsay [34] (p311-319), for example. All we need is to be able to simulate from the model.

3. SMHD Methods for Ungrouped Data

For the classical version D with ungrouped data, Simpson [14] (p806) in the proof of his Theorem 2 has shown that we have equality in probability of the following expression by letting

$$\dot{s}_{\theta}(i)=\frac{\frac{\partial p_{\boldsymbol{\theta}}(i)}{\partial \boldsymbol{\theta}}}{2\sqrt{p_{\boldsymbol{\theta}}(i)}}$$

be the vector of partial derivatives with respect to $\boldsymbol{\theta}$ of $p_{\boldsymbol{\theta}}^{\frac{1}{2}}(i), i=0,1,\cdots$ and we have

$$\sqrt{n}\sum_{i=0}^{\infty}\dot{s}_{\boldsymbol{\theta}_0}(i)\left(p_n^{\frac{1}{2}}(i)-p_{\boldsymbol{\theta}_0}^{\frac{1}{2}}\right)=^{p}\sqrt{n}\sum_{i=0}^{\infty}\frac{1}{4}\frac{\partial \log p_{\boldsymbol{\theta}_0}(i)}{\partial \theta}\left(p_n(i)-p_{\boldsymbol{\theta}_0}\right) \tag{27}$$

with $\frac{1}{4}\frac{\partial \log p_{\boldsymbol{\theta}_0}(i)}{\partial \boldsymbol{\theta}}=\frac{\dot{s}_{\boldsymbol{\theta}_0}(i)}{2\sqrt{p_{\theta}(i)}}$ and $\frac{\partial \log p_{\boldsymbol{\theta}_0}(i)}{\partial \theta}, i=1,\cdots$ is the vector of the score functions with covariance matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$ which is the Fisher information matrix.

For version D, we then have

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=^{d}\left(\boldsymbol{I}(\boldsymbol{\theta}_0)\right)^{-1}\sqrt{n}\sum_{i=0}^{\infty}\frac{\partial \log p_{\boldsymbol{\theta}_0}(i)}{\partial \theta}\left(p_n(i)-p_{\boldsymbol{\theta}_0}\right)$$

or equivalently

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=^{d}\left(\boldsymbol{I}(\boldsymbol{\theta}_0)\right)^{-1}\boldsymbol{Y} \text{ with } \boldsymbol{Y}\sim N\left(0,\boldsymbol{I}(\boldsymbol{\theta}_0)\right).$$

Therefore, we can conclude that $\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\xrightarrow{L}N\left(0,\left(\boldsymbol{I}(\boldsymbol{\theta}_0)\right)^{-1}\right)$ which is the result of Theorem 2 given by Simpson [14] (p804) which shows that the MHDE estimators are as effcient as the maximum likelihood (ML) estimators.

For version S with ungrouped data, it is more natural to use Theorem 7.1 of Newey and McFadden [17] (p2185-2186) to establish asymptotic normality for SMHD estimators. The ideas behind Theorem 7.1 can be summarized as follows. In case of the objective function $Q_n(\boldsymbol{\theta})$ is non smooth and the estimators is the vector $\tilde{\boldsymbol{\theta}}$ which is obtained by minimizing $Q_n(\boldsymbol{\theta})$, we can consider the vector $\boldsymbol{\theta}^*$ which is obtained by minimizing a smooth function $Q_n^a(\boldsymbol{\theta})$ which approximates $Q_n(\boldsymbol{\theta})$ if $Q_n(\boldsymbol{\theta})$ is differentiable in probability at $\boldsymbol{\theta}_0$ with the derivative vector given by $D_n(\boldsymbol{\theta}_0)$. For SMHD estimation,

$$Q_n(\boldsymbol{\theta})=\sum_{i=0}^{\infty}\left(\left[p_n(i)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}}^{S}(i)\right]^{\frac{1}{2}}\right)^{\frac{1}{2}} \tag{28}$$

with its equivalent expression given by expression (3).

Also, if $Q_n(\boldsymbol{\theta}) \xrightarrow{p} Q(\boldsymbol{\theta})$ and assume that $Q(\boldsymbol{\theta})$ is non random and twice differentiable with second derivative matrix $\boldsymbol{H}$ with $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{\theta}_0) = \dfrac{\partial^2 Q(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}$, $Q(\boldsymbol{\theta})$ attains its minimum at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ then we can define

$$Q_n^a(\boldsymbol{\theta}) = Q_n(\boldsymbol{\theta}_0) + \boldsymbol{D}_n(\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' H(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

The vector $\boldsymbol{\theta}^*$ which minimizes $Q_n^a(\boldsymbol{\theta})$ can be obtained explicitly as $Q_n^a(\boldsymbol{\theta})$ is a quadratic function of $\boldsymbol{\theta}$, it is given by $\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 = -\boldsymbol{H}^{-1}\boldsymbol{D}_n(\boldsymbol{\theta}_0)$ and using equality in distribution

$$\sqrt{n}\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 =\right) =^d -\boldsymbol{H}^{-1}\sqrt{n}\boldsymbol{D}_n(\boldsymbol{\theta}_0).$$

If the remainder of the approximation is small, we also have

$$\sqrt{n}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \sqrt{n}\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\right) =^d -\boldsymbol{H}^{-1}\sqrt{n}\boldsymbol{D}_n(\boldsymbol{\theta}_0).$$

Before defining the remainder term $R_n(\boldsymbol{\theta})$, note that the following approximation $Q_n^b(\boldsymbol{\theta})$ can be viewed as equivalent with

$$Q_n^b(\boldsymbol{\theta}) = Q_n(\boldsymbol{\theta}_0) + \boldsymbol{D}_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + Q(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}_0)$$

as $Q(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}_0) \approx \dfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ using $\dfrac{\partial Q(\boldsymbol{\theta}_0)}{\partial \theta} = 0$, $Q(\boldsymbol{\theta})$ is minimized at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

For the approximation to be valid, we define

$$R_n(\boldsymbol{\theta}) = \frac{\sqrt{n}\left(Q_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta}_0) + \boldsymbol{D}_n'(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + Q(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}_0)\right)}{\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|}$$

and requires $\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \delta_n} \left|R_n(\boldsymbol{\theta})\right| \xrightarrow{p} 0$ as $n \to \infty, \delta_n \to 0$ as indicated by the proofs of Theorem 7.1 given by Newey and McFadden. The following Theorem 3 is essentially Theorem 7.1 given by Newey and McFadden but restated with estimators obtained by minimizing an objective function instead of maximizing an objective function and requires $\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \delta_n} \left|R_n(\boldsymbol{\theta})\right| \xrightarrow{p} 0$ which is slightly more stringent than the original condition v) of their Theorem 7.1. We also require compactness of the parameter space Ω. Newey and McFadden do not use this assumption but with this assumption, the proofs are less technical and simplified. It is also likely to be met in practice.

Theorem 3

Suppose that $Q_n(\tilde{\boldsymbol{\theta}}) \le \inf_{\boldsymbol{\theta} \in \Omega} Q_n(\boldsymbol{\theta}) + o_p\left(\dfrac{1}{n}\right)$, $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and

1) $Q(\boldsymbol{\theta})$ is minimized at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$;
2) $\boldsymbol{\theta}_0$ is an interior point of the parameter space Ω;
3) $Q(\boldsymbol{\theta})$ is twice differentiable at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with nonsingular matrix $\boldsymbol{H}$;
4) $\sqrt{n}\boldsymbol{D}_n(\boldsymbol{\theta}_0) \xrightarrow{L} N(0, \boldsymbol{K})$;

5) $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n}\left|R_n(\boldsymbol{\theta})\right| \xrightarrow{p} 0$ as $n\to\infty, \delta_n\to 0$. Then

$$\sqrt{n}\left(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{L} N\left(0, \boldsymbol{H}^{-1}\boldsymbol{K}\boldsymbol{H}^{-1}\right).$$

The regularity conditions (1-3) of Theorem 3 can easily be checked. The condition 4 follows from expression (27) established by Simpson [14]. The condition 5 might be the most difficult to check as it involve technicalities and it is verified in **TA2** of the Appendices. By assuming all can be verified, we apply Theorem 3 for SMHD estimation with $\tilde{\boldsymbol{\theta}}=\widehat{\boldsymbol{\theta}^S}$.

The objective function $Q_n(\boldsymbol{\theta})$ is as defined by expression (3),

$$Q(\boldsymbol{\theta})=\sum_{i=0}^{\infty}\left(\left[p_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}}(i)\right]^{\frac{1}{2}}\right)^2,$$

$$Q(\boldsymbol{\theta}_0)=\sum_{i=0}^{\infty}\left(\left[p_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}}\right)^2=0,$$

the matrix of second derivative of $Q(\boldsymbol{\theta}_0)$ is

$$\boldsymbol{H}=2\sum_{i=0}^{\infty}\left(\dot{s}_{\boldsymbol{\theta}_0}(i)\right)\left(\dot{s}_{\boldsymbol{\theta}_0}(i)\right)'=\frac{1}{2}\boldsymbol{I}(\boldsymbol{\theta}_0)$$

and it can be seen that

$$\boldsymbol{D}_n(\boldsymbol{\theta}_0)=-2\sum_{i=0}^{\infty}\left(\left[p_n(i)\right]^{\frac{1}{2}}-\left[p^S_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}}\right)\dot{s}_{\boldsymbol{\theta}_0}(i),$$

by performing limit operations to find derivates as in real analysis and using Assumption 1 and Assumption 2. Therefore, we have the following equality in distribution using the condition 4) of Theorem 3 and expression (27)

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}^S}-\boldsymbol{\theta}_0\right)=^d\left(\boldsymbol{I}(\boldsymbol{\theta}_0)\right)^{-1}\sqrt{n}\sum_{i=0}^{\infty}\frac{\partial\log p_{\boldsymbol{\theta}_0}(i)}{\partial\boldsymbol{\theta}}\left(p_n(i)-p^S_{\boldsymbol{\theta}_0}(i)\right), \qquad (29)$$

which is similar to the grouped case.

Now with $\left(p_n(i)-p^S_{\boldsymbol{\theta}_0}(i)\right)=\left(p_n(i)-p_{\boldsymbol{\theta}_0}(i)\right)-\left(p^S_{\boldsymbol{\theta}_0}(i)-p_{\boldsymbol{\theta}_0}(i)\right)$ with the size of the simulated sample is $U=\tau n$ and the simulated sample is independent of the sample given by data, we can argue as for the grouped case to conclude

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}^S}-\boldsymbol{\theta}_0\right)=^d\left(\boldsymbol{I}(\boldsymbol{\theta}_0)\right)^{-1}\sqrt{n}\boldsymbol{Z} \text{ with } \boldsymbol{Z}\sim N\left(0,\left(1+\frac{1}{\tau}\right)\left(\boldsymbol{I}(\boldsymbol{\theta}_0)\right)^{-1}\right). \qquad (30)$$

One might want to define the extended Cramér-Rao lower bound for simulated method estimators to be $\frac{1}{n}\left(1+\frac{1}{\tau}\right)\left(\boldsymbol{I}(\boldsymbol{\theta}_0)\right)^{-1}$; with this definition, the asymptotic covariance matrix of SMHD estimators attains this bound just as the asymptotic covariance matrix of ML estimators attain the classical Cramér-Rao lower bound. The factor $\left(1+\frac{1}{\tau}\right)$ is a common factor which also appears in other simulated methods, it can be interpreted as the adjustment factor when estimators are obtained via minimizing a simulated version of the objective function instead of the original objective function with the model distribution

being replaced by a sample distribution using a simulated sample, see Pakes and Pollard [16] (p1048) for the simulated minimum chi-square estimators, for example. Clearly, $\boldsymbol{I}(\boldsymbol{\theta}_0)$ can also be estimated numerically as in the grouped case which is given in section (2). Results of Theorem 2 and Corollary 1 allow us to establish asymptotic normality of the MHD estimators for both versions in a unified way.

We close this section by showing the asymptotic breakdown point ϵ of SHMD estimators is the same as HMD estimators under the true model with $\epsilon \geq \frac{1}{2}$ by using the argument used by Simpson for the version D of HD estimators, see Simpson [14] (p805-806) and assuming only the original data set might be contaminated, there is no contamination coming from simulated samples. This assumption appears to be reasonable as we can control the simulation procedures. We focus only on the strict parametric model and the set up is less general than the one considered by Theorem 3 of Simpson [13] (p805) which also includes distributions near the parametric model.

Let $H^{\epsilon}_{n,\boldsymbol{\theta}_0}$ be the contaminated distribution function defined as $H^{\epsilon}_{n,\boldsymbol{\theta}_0} = (1-\epsilon)F_{\boldsymbol{\theta}_0} + \epsilon K_n$, where $F_{\boldsymbol{\theta}_0}$ is the true distribution function and the distribution K_n is introduced to contaminate the model. The pmfs of $H^{\epsilon}_{n,\boldsymbol{\theta}_0}, F_{\boldsymbol{\theta}_0}$ and K_n are given respectively by $p^{\epsilon}_{n,\boldsymbol{\theta}_0}(i), p_{\boldsymbol{\theta}_0}(i)$ and $p_{K_n}(i)$. The asymptotic breakdown point is the smallest value ϵ which makes $\left|\widehat{\boldsymbol{\theta}}^S_n\right| \to \infty$. The vector $\widehat{\boldsymbol{\theta}}^S_n$ minimizes with respect to $\boldsymbol{\theta}$ the objective function

$$\rho\left(p^{\epsilon}_{n,\boldsymbol{\theta}_0}, p^S_{\boldsymbol{\theta}}\right) = \sum_{i=0}^{\infty}\left[p^{\epsilon}_{n,\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}}\left[p^S_{\boldsymbol{\theta}}(i)\right]^{\frac{1}{2}}.$$

Now with the same seed used across $\boldsymbol{\theta}$, $p^S_{\boldsymbol{\theta}}(i)$ can be viewed as a proxy pmf for the true parametric model. We let $\left|\widehat{\boldsymbol{\theta}}^S_n\right| \to \infty$ and show that this will imply $\epsilon \geq \frac{1}{2}$ in probability. As $\widehat{\boldsymbol{\theta}}^S_n$ is the vector which minimizes SHD or maximizes $\rho\left(p^{\epsilon}_{n,\boldsymbol{\theta}_0}, p^S_{\boldsymbol{\theta}}\right)$ clearly $\rho\left(p^{\epsilon}_{n,\boldsymbol{\theta}_0}, p^S_{\widehat{\boldsymbol{\theta}}^S_n}\right) \geq \rho\left(p^{\epsilon}_{n,\boldsymbol{\theta}_0}, p^S_{\boldsymbol{\theta}_0}\right)$. Now observe that

$$\rho\left(p^{\epsilon}_{n,\boldsymbol{\theta}_0}, p^S_{\boldsymbol{\theta}_0}\right) = \sum_{i=0}^{\infty}\left[(1-\epsilon)\,p_{\boldsymbol{\theta}_0}(i) + \epsilon p_{K_n}(i)\right]^{\frac{1}{2}}\left[p^S_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}}$$

but

$$\sum_{i=0}^{\infty}\left[(1-\epsilon)\,p_{\boldsymbol{\theta}_0}(i) + \epsilon p_{K_n}(i)\right]^{\frac{1}{2}}\left[p^S_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}}$$

$$\geq (1-\epsilon)^{\frac{1}{2}}\sum_{i=0}^{\infty}\left[p_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}}\left[p^S_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}} \xrightarrow{p} (1-\epsilon)^{\frac{1}{2}}$$

as $\left[p^S_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}} \xrightarrow{p} p_{\boldsymbol{\theta}_0}(i)$ which implies $\sum_{i=0}^{\infty}\left[p_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}}\left[p^S_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}} \xrightarrow{p} 1$. So, in probability, we have the lower bound

$$(1-\epsilon)^{\frac{1}{2}} \leq \rho\left(p^{\epsilon}_{n,\boldsymbol{\theta}_0}, p^S_{\widehat{\boldsymbol{\theta}}^S_n}\right).$$

With $\rho\left(p_{n,\boldsymbol{\theta}_0}^{\epsilon}, p_{\widehat{\boldsymbol{\theta}}_n^S}^S\right) = \sum_{i=0}^{\infty}\left[(1-\epsilon)\,p_{\boldsymbol{\theta}_0}(i) + \epsilon p_{K_n}(i)\right]^{\frac{1}{2}}\left[p_{\widehat{\boldsymbol{\theta}}_n^S}^S\right]^{\frac{1}{2}}$,

using the inequality $(a+b)^{\frac{1}{2}} \le a^{\frac{1}{2}} + b^{\frac{1}{2}}$ with $a \ge 0, b \ge 0$, we have the upper bound inequality

$$\rho\left(p_{n,\boldsymbol{\theta}_0}^{\epsilon}, p_{\widehat{\boldsymbol{\theta}}_n^S}^S\right) \le \sum_{i=0}^{\infty}\left((1-\epsilon)^{\frac{1}{2}}\left[p_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}} + \epsilon^{\frac{1}{2}}\left[p_{K_n}(i)\right]^{\frac{1}{2}}\right)\left[p_{\widehat{\boldsymbol{\theta}}_n^S}^S(i)\right]^{\frac{1}{2}}$$

$$\le (1-\epsilon)^{\frac{1}{2}}\sum_{i=0}^{\infty}\left[p_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}}\left[p_{\widehat{\boldsymbol{\theta}}_n^S}^S(i)\right]^{\frac{1}{2}}$$

$$+\epsilon^{\frac{1}{2}}\sum_{i=0}^{\infty}\left[p_{K_n}(i)\right]^{\frac{1}{2}}\left[p_{\widehat{\boldsymbol{\theta}}_n^S}^S(i)\right]^{\frac{1}{2}} \le \epsilon^{\frac{1}{2}}.$$

The last inequality follows from the assumption that

$\sum_{i=0}^{\infty}\left[p_{\boldsymbol{\theta}_0}(i)\right]^{\frac{1}{2}}\left[p_{\widehat{\boldsymbol{\theta}}_n^S}^S(i)\right]^{\frac{1}{2}} \xrightarrow{p} 0$ since $\left|\widehat{\boldsymbol{\theta}}_n\right| \to \infty$ which implies the two pmfs $p_{\boldsymbol{\theta}_0}(i)$ and $p_{\widehat{\boldsymbol{\theta}}_n^S}^S(i)$ are not close according to the discrepancy measure using SHD as $n \to \infty$, an argument also used by Simpson [14] to justify his expression $\rho^* = 0$, see Simpson [14] (p805-806).

Using $\sum_{i=0}^{\infty}\left[p_{K_n}(i)\right]^{\frac{1}{2}}\left[p_{\widehat{\boldsymbol{\theta}}_n^S}^S(i)\right]^{\frac{1}{2}} \le 1$, we might conclude in probability we have the inequalities $(1-\epsilon)^{\frac{1}{2}} \le \rho\left(p_{n,\boldsymbol{\theta}_0}^{\epsilon}, p_{\widehat{\boldsymbol{\theta}}_n^S}^S\right) \le \epsilon^{\frac{1}{2}}$ which implies $\epsilon \ge \frac{1}{2}$ in probability under the true model which is similar to version D. The only difference is here we have an inequality in probability. From this result, we might conclude that the SMHD estimators preserve the robustness properties of version D and the loss of asymptotic efficiency comparing to version D can be minimized if $\tau \ge 10$.

4. Numerical Issues

4.1. Methods to Approximate Probabilities

Once the parameters are estimated, probabilities can be estimated. For situations where recursive formulas exist then Panjer's method can be used, see Chapter 9 of the book by Klugman *et al.* [4]. Otherwise, we might need to approximate probabilities by simulations or by analytic methods.

In this section, we discuss some methods for approximating probabilities $p_h, h = 0,1,\cdots$ for a discrete nonnegative random variable X with pgf $P(s)$ which can be used if a recursion formula for p_h is not available. The saddlepoint method and the method based on inverting the characteristic function can be used.

See Butler [35] (p8-9) for details of the saddlepoint approximation. It can be described as using p_h^a to approximate p_h, with

$$p_h^a = \frac{1}{\sqrt{2\pi K''(\bar{s})}}\exp\left(K(\bar{s}) - \bar{s}h\right).$$

The saddlepoint $\bar{s}$ is defined implicitly, using the pgf, as the solution of $\log\left(P\left(\mathrm{e}^{s}\right)\right)=K(s)$, and $K'(\bar{s})=h$ with $K'(s)=\dfrac{\partial K(s)}{\partial s}$ and $K''(s)=\dfrac{\partial^2 K(s)}{\partial s^2}$. The function $K(s)$ is the cumulant function.

If the cumulant function does not exist, an alternative method which is based on the characteristic function, as described by Abate and Whitt [36] (p32), can be used.

4.2. A Limited Simulation Study

4.2.1. Neymann Type A distribution

As an example for illustration we choose the Neymann Type A distribution with the method of moments (MM) estimators for λ and ϕ which have been given by Johnson *et al.* [8]. The MM estimators are given by $\tilde{\lambda}=\dfrac{\bar{X}}{\tilde{\phi}}$ and $\tilde{\phi}=\dfrac{s^2}{\bar{X}}-1$ with the sample mean and variance given respectively by $\bar{X}$ and s^2. The MM estimators are classical moment estimators. We perform a limited simulation study to compare the performance of the SMHD estimator which is given by $\widehat{\boldsymbol{\theta}}^{S}=\left(\widehat{\lambda}^{S},\widehat{\phi}^{S}\right)'$ vs the MM estimators given by $\tilde{\boldsymbol{\theta}}=\left(\tilde{\lambda},\tilde{\phi}\right)'$.

For the range of parameter values, we let $\lambda=0.25,0.5,1,2,\cdots,6$, $\phi=30,40,\cdots,80,100$ are used in the study. For applications often the parameter λ for the mixing distribution much smaller than the parameter ϕ. The SMHD estimators seem to perform much better than the MM estimators, in general. The results are displayed in **Table A**. The criterion for overall relative efficiency used is the ratio $ARE=\dfrac{MSE\left(\widehat{\lambda}^{S}\right)+MSE\left(\widehat{\phi}^{S}\right)}{MSE\left(\tilde{\lambda}\right)+MSE\left(\tilde{\phi}\right)}$ with $MSE(.)$ denotes the mean square error of the estimator inside the parenthesis. The mean square error of an estimator $\hat{\pi}$ for π_0 is defined as

$$MSE\left(\hat{\pi}\right)=E\left(\hat{\pi}-\pi_0\right)^2.$$

The ratio ARE can be estimated using simulated data and they are displayed in **Table A**. Due to limited computing facilities, we only draw $M=50$ samples of size $n=1000$ and the simulated sample is fixed at $U=12000$, $\tau=12$ and the results are summarized using **Table A**. It takes around one minute using a laptop computer for obtaining the SMHD estimators for one simulated sample. The MM estimators appear to perform reasonably well for some samples but display erratic results for some other samples which account for the loss of efficiency of the MM estimators. Also, the parameter ϕ is not well estimated by the moment method but it gives reasonably good estimates for the parameter λ in general. The MM estimators are based on the sample mean and variance and these statistics are known to be nonrobust. If outliers are present, the MM estimators again might become erratic. The mean square errors (MSE) for estimat-

ing the parameters and the corresponding ratios ARE are estimated using the simulated samples and the AREs are displayed in Table A.

4.2.2. Poisson Distribution

For the Poisson model with parameter λ we compare the performance of $\widehat{\lambda_{ML}}$ the MLE for λ which is the sample mean vs the SMHD estimator $\widehat{\lambda^S}$ using the ratio $ARE = \dfrac{MSE\left(\widehat{\lambda_{ML}}\right)}{MSE\left(\widehat{\lambda^S}\right)}$ for $\lambda = 5,10,12,\cdots,20,100$. For the Poisson model, the information matrix exists and we can check the efficiency and robustness of the SHD estimator and compare it with the ML estimator which is the sample mean. Since there is only on parameter estimate we are able to fix

Table A. Asymptotic relative efficiencies between MM estimators and SMHD estimators $ARE = \dfrac{MSE\left(\widehat{\lambda^S}\right) + MSE\left(\widehat{\phi^S}\right)}{MSE\left(\hat{\lambda}\right) + MSE\left(\hat{\phi}\right)}$.

$\lambda \cdot \cdot . \phi$	30	40	50	60	80	100
0.25	0.0032	0.0082	0.0238	0.0173	0.0074	0.0063
0.5	0.0523	0.0024	0.0148	0.2053	0.0115	0.0429
1	0.0337	0.0256	0.1253	0.1502	0.0892	0.0481
2	0.0073	0.0197	0.0393	0.0536	0.2986	0.0147
3	0.0038	0.0046	0.0020	0.3167	0.0229	0.0057
4	0.0098	0.0103	0.0117	0.0156	0.0102	0.0020
5	0.0481	0.1431	0.0062	0.0073	0.0100	0.0009
6	1.0330	0.0632	0.0145	0.0236	0.0126	0.0062

Asymptotic relative efficiency between MLE $\widehat{\lambda_{ML}}$ and SHD $\hat{\lambda}$, $ARE = \dfrac{MSE(\widehat{\lambda_{ML}})}{MSE\left(\widehat{\lambda^S}\right)}$ for the Poisson model with parameter λ.

λ	5	10	12	14	16	18	20	100
$\dfrac{MSE(\widehat{\lambda_{ML}})}{MSE\left(\widehat{\lambda^S}\right)}$	0.7000	0.7864	0.9639	0.7649	0.8256	0.9480	0.8102	1.097

Asymptotic relative efficiency between MLE $\widehat{\lambda_{ML}}$ and SHD $\widehat{\lambda^S}$ for the Poisson model (λ) when 10% of data coming from the discrete positive distribution with parameter λ and $\alpha = 0.9$, *i.e.*, $0.9Poisson(\lambda) + 0.1DPS(\lambda, \alpha = 0.9)$.

λ	5	10	12	14	16	18	20	100
$\dfrac{MSE(\widehat{\lambda_{ML}})}{MSE\left(\widehat{\lambda^S}\right)}$	87.736	87.5592	43.6890	102.8376	85.9624	62.8738	51.2473	75.8619

U = 10000 for the simulated sample size from the Poisson model without slowing down the computations. It appears overall the SHD estimators performs very well for the range of parameters often encountered in actuarial studies, here we observe that the asymptotic efficiencies range from 0.7 to 1.1. We also study a contaminated Poisson model (λ) with $p = 90\%$ observations coming from the Poisson model (λ) and $q = 1 - p = 10\%$ of observations coming from a discrete positive stable (DPS) distribution with the parameter for $\alpha = 0.9$ and λ has the same value of the Poisson model. We compare the performance of the sample mean for λ which is the ML estimator vs the SMHD estimator $\widehat{\lambda^S}$ using the contaminated model Poisson model as described and estimate the $ARE = \dfrac{MSE\left(\widehat{\lambda_{ML}}\right)}{MSE\left(\widehat{\lambda^S}\right)}$ for $\lambda = 5,10,12,\cdots,20,100$ to compare the robustness of the SMHD estimator vs ML estimator in presence of contamination. The sample mean looses its efficiency and becomes very biased. The results are given at the bottom of **Table A** which shows that the $\widehat{\lambda^S}$ performs much better than the sample mean which is the ML estimator. For drawing simulated samples from the DPS distribution, the algorithm given by Devroye [37] is used.

5. Conclusion

More simulation experiments to further study the performance of the SMHD estimators vs commonly used estimators across various parametric models are needed and we do not have the computing facilities to carry out such large scale studies. Most of the computing works were carried out using only a laptop computer. So far, the simulation results confirm the theoretical asymptotic results which show that SMHD estimators have the potential of having high efficiencies for parametric models with finite Fisher information matrices and they are robust if data is contaminated; the last feature might not be shared by ML estimators.

Acknowledgements

The helps received from the Editorial staffs of OJS which lead to an improvement of the presentation of the paper are gratefully acknowledged.

References

[1] Christoph, G. and Schreiber, K. (1998) Discrete Stable Random Variables. *Statistics and Probability Letters*, **37**, 243-247. https://doi.org/10.1016/S0167-7152(97)00123-5

[2] Abate, J. and Whitt, W. (1996) An Operational Calculus for Probability Distributions via Laplace Transforms. *Advances in Applied Probability*, **28**, 75-113. https://doi.org/10.2307/1427914

[3] Panjer, H. and Willmot, G.E. (1992) Insurance Risk Models. Society of Actuaries, Chicago, IL.

[4] Klugman, H., Panjer, H.H. and Willmot, G.E. (2012) Loss Models: From Data to

Decisions. Wiley, New York.

[5] Gerber, H.U. (1991) From the Generalized Gamma to the Generalized Negative Binomial Distribution. *Insurance: Mathematics and Economics*, **10**, 303-309. https://doi.org/10.1016/0167-6687(92)90061-F

[6] Hougaard, P. (1986) Survival Models for Heterogeneous Populations Derived from Stable Distributions. *Biometrika*, **73**, 387-396. https://doi.org/10.1093/biomet/73.2.387

[7] Zhu, R. and Joe, H. (2009) Modelling Heavy Tailed Count Data Using a Generalized Poisson Inverse Gaussian Family. *Statistics and Probability Letters*, **79**, 1695-1703. https://doi.org/10.1016/j.spl.2009.04.011

[8] Johnson, N.L., Kotz, S. and Kemp, A.W. (1992) Univariate Discrete Distributions. 2nd Edition, Wiley, New York.

[9] Doray, L.G., Jiang, S.M. and Luong, A. (2009) Some Simple Methods of Estimation for the Parameters of the Discrete Stable Distributions with Probability Generating Functions. *Communications in Statistics*, *Simulation and Computation*, **38**, 2004-2007. https://doi.org/10.1080/03610910903202089

[10] Carrasco, M. and Florens, J.-P. (2000) Generalization of GMM to a Continuum Moments Conditions. *Econometric Theory*, **16**, 797-834. https://doi.org/10.1017/S0266466600166010

[11] Feuerverger, A. and McDunnough, P. (1981) On the Efficiency of Empirical Characteristic Function Procedure. *Journal of the Royal Statistical Society*, *Series B*, **43**, 20-27.

[12] Nadarajah, S and Kotz, S. (2006a) Compound Mixed Poisson Distribution I. *Scandinavian Actuarial Journal*, **2006**, 141-162. https://doi.org/10.1080/03461230600783384

[13] Nadarajah, S and Kotz, S. (2006b) Compound Mixed Poisson Distribution II. *Scandinavian Actuarial Journal*, **2006**, 163-181. https://doi.org/10.1080/03461230600715253

[14] Simpson, D.G. (1987) Minimum Hellinger Distance Estimation for the Analysis of Count Data. *Journal of the American Statistical Association*, **82**, 802-807. https://doi.org/10.1080/01621459.1987.10478501

[15] Simpson, D.G. (1989) Hellinger Deviance Tests: Efficiency, Breakdown Points and Examples. *Journal of the American Statistical Association*, **84**, 107-113. https://doi.org/10.1080/01621459.1989.10478744

[16] Pakes, A. and Pollard, D. (1989) Simulation Asymptotic of Optimization Estimators. *Econometrica*, **57**, 1027-1057. https://doi.org/10.2307/1913622

[17] Newey, W.K. and McFadden, D. (1994) Large Sample Estimation and Hypothesis Testing. In: Engle, R.F. and McFadden, D., Eds., *Handbook of Econometrics*, Vol. 4, North Holland, Amsterdam.

[18] Davidson, R. and MacKinnon, J.G. (2004) Econometric Theory and Methods. Oxford University Press, Oxford.

[19] Cressie, N. and Read, T. (1984) Multinomial Goodness of Fit Tests. *Journal of the Royal Statistical Society*, *Series B*, **46**, 440-464.

[20] Hogg, R.V., Mc Kean, J.W. and Craig, A. (2013) Introduction to Mathematical Statistics. 7th Edition, Pearson, Boston, MA.

[21] Kloke, J. and McKean, J.W. (2015) Nonparametric Statistical Using R. CRC Press, New York.

[22] Maronna, R.A., Martin, R.D. and Yohai, V.G. (2006) Robust Statistics: Theory and Methods. Wiley, New York. https://doi.org/10.1002/0470010940

[23] Beran, R. (1977) Minimum Hellinger Distance Estimates for Parametric Models. *Annals of Statistics*, **5**, 445-463. https://doi.org/10.1214/aos/1176343842

[24] Lindsay, B.G. (1994) Efficiency versus Robustness: The Case for Minimum Hellinger Distance and Related Methods. *Annals of Statistics*, **22**, 1081-1114. https://doi.org/10.1214/aos/1176325512

[25] Gusak, D., Kukush, A., Kulik, A., Mishura, Y. and Pilipenko, A. (2010) Theory of Stochastic Processes with Applications to Financial Mathematics and Risk Theory. Springer, New York.

[26] Pakes, A. (1986) Patents as Options: Some Estimates of the Value of Holding European Patent Stocks. *Econometrica*, **54**, 755-784. https://doi.org/10.2307/1912835

[27] Keener, R.W. (2016) Theoretical Statistics. Springer, New York.

[28] Davidson, K.R. and Donsig, A.P. (2009) Real Analysis and Applications. Springer, New York.

[29] Rudin, W. (1976) Principles of Mathematical Analysis. McGraw-Hill, New York.

[30] Luong, A. and Thompson, M.E. (1987) Minimum Distance Methods Based on Quadratic Distance for Transforms. *Canadian Journal of Statistics*, **15**, 239-251. https://doi.org/10.2307/3314914

[31] Greenwood, P.E. and Nikulin, M.S. (1996) A Guide to Chi-Squared Testing. Wiley, New York.

[32] Lehmann, E.L. (1999) Element of Large Sample Theory. Wiley, New York. https://doi.org/10.1007/b98855

[33] Luong, A. (2016) Cramér-Von Mises Distance Estimation for Some Positive Infinitely Divisible Parametric Families with Actuarial Applications. *Scandinavian Actuarial Journal*, **2016**, 530-549. https://doi.org/10.1080/03461238.2014.977817

[34] Tsay, R.S. (2010) The Analysis of Financial Time Series. 3rd Edition, Wiley, New York. https://doi.org/10.1002/9780470644560

[35] Butler, R.W. (2007) Saddlepoint Approximations with Applications. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511619083

[36] Abate, J. and Whitt, W. (1992) The Fourier Methods for Inverting Transforms of Probability Distributions. *Queueing Systems*, **10**, 5-87. https://doi.org/10.1007/BF01158520

[37] Devroye, L. (1993) A Triptych of Discrete Distributions Related to the Stable Laws. *Statistics and Probability Letters*, **18**, 349-351. https://doi.org/10.1016/0167-7152(93)90027-G

Appendices

Technical Appendix 1 (TA1)

TA 1.1

In this technical appendix, we shall show that a sequence of random functions $\{g_n(\boldsymbol{\theta})\}$ which is continuous in probability and bounded in probability on a compact set $\boldsymbol{\theta}$ will attain its supremum on a point of $\boldsymbol{\theta}$ in probability. Pick a sequence $\{\boldsymbol{\theta}_j\}$ in $\boldsymbol{\theta}$ with the property $g_n(\boldsymbol{\theta}_j) \xrightarrow{p} \sup_{\theta\in\ominus} g_n(\boldsymbol{\theta})$. Since $\boldsymbol{\theta}$ is compact we can extract a subsequence $\{\boldsymbol{\theta}_{j_k}\}$ from $\{\boldsymbol{\theta}_j\}$ with the property $\boldsymbol{\theta}_{j_k} \to \boldsymbol{\theta}^0$ which belongs to $\boldsymbol{\theta}$. This property in real analysis is also known under the name Bolzano-Weirstrass theorem. We then have $g_n(\boldsymbol{\theta}_{j_k}) \xrightarrow{p} g_n(\boldsymbol{\theta}^0)$ and $\sup_{\theta\in\ominus} g_n(\boldsymbol{\theta}) =^p g_n(\boldsymbol{\theta}^0)$.

TA 1.2

In this technical appendix, we shall show that the sequence of function $\{g_n(\boldsymbol{\theta})\}$ is continuous in probability and for the grouped case of Section (2.2.2), $g_n(\boldsymbol{\theta})$ for the grouped data case can also be expressed as

$$g_n(\boldsymbol{\theta}) = n\sum_{i=0}^{k}\left(\left[p_{\boldsymbol{\theta}_0}^S(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}_0}(I_j)\right]^{\frac{1}{2}}\right)^2 + n\sum_{i=0}^{k}\left(\left[p_{\boldsymbol{\theta}}^S(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}}(I_j)\right]^{\frac{1}{2}}\right)^2$$
$$- 2n\sum_{i=0}^{k}\left(\left[p_{\boldsymbol{\theta}_0}^S(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}_0}(I_j)\right]^{\frac{1}{2}}\right)\left(\left[p_{\boldsymbol{\theta}}^S(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}}(I_j)\right]^{\frac{1}{2}}\right).$$

The first two terms of the RHS of the above equation are bounded in probability as they have a limiting distributions and this implies the third term is also bounded in probability by using Cauchy-Schwartz inequality. Now using the conditions of Assumption 1 of Section (2.2.2) and implicitly the assumption of the same seed is used across different values of $\boldsymbol{\theta}$, we then have as $\boldsymbol{\theta} \to \boldsymbol{\theta}'$,

$$n\sum_{i=0}^{k}\left(\left[p_{\boldsymbol{\theta}}^S(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}}(I_j)\right]^{\frac{1}{2}}\right)^2 \xrightarrow{p} n\sum_{i=0}^{k}\left(\left[p_{\boldsymbol{\theta}'}^S(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}'}(I_j)\right]^{\frac{1}{2}}\right)^2$$

and

$$2n\sum_{i=0}^{k}\left(\left[p_{\boldsymbol{\theta}_0}^S(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}_0}(I_j)\right]^{\frac{1}{2}}\right)\left(\left[p_{\boldsymbol{\theta}}^S(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}}(I_j)\right]^{\frac{1}{2}}\right)$$
$$\xrightarrow{p} 2n\sum_{i=0}^{k}\left(\left[p_{\boldsymbol{\theta}_0}^S(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}_0}(I_j)\right]^{\frac{1}{2}}\right)\left(\left[p_{\boldsymbol{\theta}'}^S(I_j)\right]^{\frac{1}{2}} - \left[p_{\boldsymbol{\theta}'}(I_j)\right]^{\frac{1}{2}}\right).$$

From the above property, it is clear that $g_n(\boldsymbol{\theta}_0) = 0$, $g_n(\boldsymbol{\theta})$ is continuous in probability, and using **TA1.1** we conclude that there exists $\boldsymbol{\theta}^0$ which belongs to $S(\boldsymbol{\theta}_0, \delta_n) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \delta_n\}$ and $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n} g_n(\boldsymbol{\theta}) =^p g_n(\boldsymbol{\theta}^0)$ but as $n \to \infty, \boldsymbol{\theta}^0 \to \boldsymbol{\theta}_0$, $g_n(\boldsymbol{\theta}^0) \xrightarrow{p} g_n(\boldsymbol{\theta}_0) = 0$. Therefore, $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n} g_n(\boldsymbol{\theta}) \xrightarrow{p} 0$, as $n \to \infty, \boldsymbol{\theta}^0 \to \boldsymbol{\theta}_0$.

The justifications for the ungrouped case are similar using the same type of arguments but with the use of Theorem 7.1 given by Newey and McFadden [17] and will be given in **TA2**.

Technical Appendix 2 (TA2)

In this technical appendix we shall verify the condition $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n}\left|R_n(\boldsymbol{\theta})\right| \xrightarrow{p} 0$ as $n\to\infty, \delta_n\to 0$ for SMHD estimation using ungrouped data. Despite $Q(\boldsymbol{\theta}_0)=0$, we will keep $Q(\boldsymbol{\theta}_0)$ and define the sequence of functions $u_n(\boldsymbol{\theta})=\sqrt{n}\left(\left(Q_n(\boldsymbol{\theta})+Q(\boldsymbol{\theta})\right)-\left(Q_n(\boldsymbol{\theta}_0)+Q(\boldsymbol{\theta}_0)\right)\right)$ then it will allow us to express $R_n(\boldsymbol{\theta})=\dfrac{u_n(\boldsymbol{\theta})-\sqrt{n}\mathbf{D}_n'(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|}$. Now with $\dfrac{\partial Q(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}=0$, $u_n(\boldsymbol{\theta})$ is differentiable in probability at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$. The derivative vector for $u_n(\boldsymbol{\theta})$ at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ is simply $\sqrt{n}\mathbf{D}_n(\boldsymbol{\theta}_0)$ as it can be seen by performing limit operations as in real analysis and using Assumption 1 and Assumption 2 in Section 3. Therefore, we have $R_n(\boldsymbol{\theta})\xrightarrow{p}0$ as $\boldsymbol{\theta}\to\boldsymbol{\theta}_0$ by using definition of the derivative. Since $u_n(\boldsymbol{\theta})$ is approximable by $\sqrt{n}\mathbf{D}_n'(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ which is bounded in probability in a neighborhood including $\boldsymbol{\theta}_0$, $S(\boldsymbol{\theta}_0,\delta_{n_0})=\left\{\boldsymbol{\theta}\middle|\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_{n_0}\right\}$ using expression (27), we might assume $u_n(\boldsymbol{\theta})$ is bounded in probability for $\boldsymbol{\theta}\in S(\boldsymbol{\theta}_0,\delta_{n_0})$. We might also assume that $u_n(\boldsymbol{\theta}')\xrightarrow{p}u_n(\boldsymbol{\theta}'')$ as $\boldsymbol{\theta}'\to\boldsymbol{\theta}''$, using Dominated Convergence Theorem (DCT) with $Q_n(\boldsymbol{\theta}')\xrightarrow{p}Q_n(\boldsymbol{\theta}'')$ when $\boldsymbol{\theta}'\to\boldsymbol{\theta}''$, $Q_n(\boldsymbol{\theta})$ as defined by expression (28) of Section 3, the summand of $Q_n(\boldsymbol{\theta})$ satisfies the following inequalities

$$\left(\left[p_n(i)\right]^{\frac{1}{2}}-\left[p_{\boldsymbol{\theta}}^S(i)\right]^{\frac{1}{2}}\right)^2\le\left|p_n(i)-p_{\boldsymbol{\theta}}(i)\right|\le p_n(i)+p_{\boldsymbol{\theta}}(i),\ i=0,1,\cdots,$$

hence the use of DCT is justified. Therefore, $u_n(\boldsymbol{\theta})$ is continuous in probability for all $\boldsymbol{\theta}\in S(\boldsymbol{\theta}_0,\delta_{n_0})$.

Now if we define $R_n(\boldsymbol{\theta}_0)=0$, $R_n(\boldsymbol{\theta})$ is continuous in probability for all $\boldsymbol{\theta}$ which belongs to $S(\boldsymbol{\theta}_0,\delta_{n_0})$. Consequently, $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n}\left|R_n(\boldsymbol{\theta})\right|=^p\left|R_n(\boldsymbol{\theta}^0)\right|$ with $\boldsymbol{\theta}^0\in S(\boldsymbol{\theta}_0,\delta_n)$ as the set $S(\boldsymbol{\theta}_0,\delta_n)$ is compact. As $n\to\infty$, $\delta_n\to 0$, $\boldsymbol{\theta}^0\to\boldsymbol{\theta}_0$, $\left|R_n(\boldsymbol{\theta}^0)\right|\xrightarrow{p}\left|R_n(\boldsymbol{\theta}_0)\right|=0$, this establishes the result and the argument used is similar to the one used in **TA1.2** for the grouped data case.