Noname manuscript No.
(will be inserted by the editor)

A Note on Weighted Least Square Distribution Fitting and Full Standardization of the Empirical Distribution Function

Andrew R. Barron · Mirta Benšić · Kristian Sabo

Received: date / Accepted: date

Abstract The relationship between the norm square of the standardized cumulative distribution and the chi-square statistic is examined using the form of the covariance matrix as well as the projection perspective. This investigation enables us to give uncorrelated components of the chi-square statistic and to provide interpretation of these components as innovations standardizing the cumulative distribution values. The norm square of the standardized difference between empirical and theoretical cumulative distributions is also examined as an objective function for parameter estimation. Its relationship to the chi-square distance enables us to discuss the large sample properties of these estimators and a difference in their properties in the cases that the distribution is evaluated at fixed and random points.

Keywords weighted least squares · minimum chi-square · empirical distribution · distribution fitting

This work was supported by the Croatian Science Foundation through research grants IP-2016-06-6545 and IP-2016-06-6777.

A. R. Barron
Department of Statistics and Data Science, Yale University
P.O. Box 208290
New Haven, CT 06520, USA
E-mail: andrew.barron@yale.edu

M. Benšić
Department of Mathematics, University of Osijek
Trg Ljudevita Gaja 6, Osijek, Croatia
E-mail: mirta@mathos.hr

K. Sabo
Department of Mathematics, University of Osijek
Trg Ljudevita Gaja 6, Osijek, Croatia
E-mail: ksabo@mathos.hr

1 Introduction

Let n be a fixed integer and $X_1, X_2, \ldots, X_n$ be independent real–valued random variables with distribution function F. Let F_n be the empirical distribution function $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \le t\}}$ (here $\mathbf{1}_C$ is the indicator function of the set C) and, with F an hypothesized distribution, let $Y_n(t) = \sqrt{n}(F_n(t) - F(t)), \quad t \in \mathcal{T}$, be the empirical process evaluated at a set of points $\mathcal{T} \subset \mathbb{R}$. The covariance of the empirical process takes the form $E[Y_n(t)Y_n(s)] = F(t)(1 - F(s))$ for $t \le s$ (see eg. van der Vaart and Wellner (1996), Shorack and Wellner (2009)).

Certain functions of $F_n - F$ correspond to familiar test statistics. Indeed, the maximum of the absolute value is the Kolmogorov–Smirnov test statistic, the average square is the Cramer–Von Mises test statistic, and the average square with marginal standardization using the variance function $F(t)(1 - F(t))$ produces the Anderson-Darling statistics (average with the distribution F) (see Anderson (1952)). These statistics use the whole empirical process with $\mathcal{T} = \mathbb{R}$.

For finite $\mathcal{T}$, let $\boldsymbol{V}$ denote the corresponding symmetric covariance matrix of the column vector $\sqrt{n}(\boldsymbol{F}_n - \boldsymbol{F})$ with entries $\sqrt{n}(F_n(t) - F(t)), t \in \mathcal{T}$. Finite $\mathcal{T}$ counterparts to the Kolmogorov–Smirnov, Cramer–Von Mises, and Anderson-Darling statistics have been considered in Henze (1996) and Choulakian et al. (1994). In particular, a finite $\mathcal{T}$ counterpart to the Anderson–Darling statistic is $n(\boldsymbol{F}_n - \boldsymbol{F})^T(\mathrm{Diag}(\boldsymbol{V}))^{-1}(\boldsymbol{F}_n - \boldsymbol{F})$, which uses only the diagonal entries of $\boldsymbol{V}$.

Here we focus on the complete standardization of the empirical distribution restricted to $\mathcal{T} = \{t_1, \ldots, t_k\}$ leading to the squared distance

$$n(\boldsymbol{F}_n - \boldsymbol{F})^T \boldsymbol{V}^{-1} (\boldsymbol{F}_n - \boldsymbol{F}) \tag{1}$$

and to estimation procedures that minimize it. This quadratic form is the squared Mahalanobis distance between the vectors $\boldsymbol{F}_n$ and $\boldsymbol{F}$. The motivation, familiar from regression, is that the complete standardization produces more efficient estimators.

Such estimators are usually named "weighted least squares" (e.g. Swain et al. (1988)) or "generalized least squares" (e.g. Benšić (2014), Benšić (2015)). However, as we shall see, the tridiagonal form of the matrix $\boldsymbol{V}^{-1}$ (see e.g. Barrett (1978), Barrett (1979)) puts them in the minimum chi-square context. Indeed, the norm square of the standardized empirical distribution given in expression (1) is in fact equal to the chi-square statistic

$$n \sum_{A \in \pi} \frac{(P_n(A) - P(A))^2}{P(A)}, \tag{2}$$

where π is the partition of $\mathbb{R}$ into the $k+1$ intervals A_j, $j = 1, \ldots, k+1$, formed by consecutive values $\mathcal{T} = \{t_1, \ldots, t_k\}$ with $t_1 < t_2 < \cdots < t_k$, where $A_1 = (-\infty, t_1], A_2 = (t_1, t_2], \ldots, A_k = (t_{k-1}, t_k]$ and $A_{k+1} = (t_k, \infty)$. Here

$P_n(A_j) = F_n(t_j) - F_n(t_{j-1}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A_j\}}$ and $P(A_j) = F(t_j) - F(t_{j-1})$ with $F(-\infty) = F_n(-\infty) = 0$ and $F(\infty) = F_n(\infty) = 1$.

We provide a simple explicit standardization. Indeed (1) and (2) are shown to be equal to the sum of squares

$$\sum_{j=1}^{k} Z_j^2 \tag{3}$$

of convenient choice of uncorrelated zero mean and unit variance random variables Z_j which are proportional to

$$F_n(t_{j+1})F(t_j) - F(t_{j+1})F_n(t_j).$$

As we shall show, the equivalence of the formulas (1), (2) and (3) holds also for the case of random sets $\mathcal{T}$ with cut-points based on empirical quantiles. This extends and amplifies a result credited to Kulldorff in Hartley and Pfaffenberger (1972) showing equivalence of analogous expressions (1) and (2) in the empirical quantile case.

In this note we address the relationship between the standardized cumulative distribution and the chi-square statistic by using the tridiagonal form of the matrix $\boldsymbol{V}^{-1}$ as well as the projection perspective. It enables us to give the uncorrelated components of the chi-square statistic and to discuss asymptotic properties of the estimators that minimize (1) for random and fixed choices of points in the finite set $\mathcal{T}$.

In Section 2 we explain the framework which we use in this note. In Section 3 we address two ways in which the relationship between the standardized cumulative distribution and the chi-square statistic can be seen. In Section 4 we present uncorrelated components of the chi-square statistic and provide interpretation of these components as innovations standardizing the cumulative distribution values.

In the last section we discuss a difference in large sample properties for estimators that minimize (1) for fixed and random choices of points in $\mathcal{T}$. Some of these estimators have interpretation as regression procedures based on discrepencies between the empirical distribution function and its theoretical counterpart. Minimizing (1) is often used for estimating distributional parameters. Examples include research concerning parametrized distributions, for which the maximum likelihood estimate sometimes doesn't exist. We can find them in some textbooks (see e.g. Johnson et al. (1994) and Rinne (2009)) as well as scientific papers which discuss and compare different estimation methods especially in reliability and survival analysis (e.g. Torres (2014), Kundu (2005), Benšić (2014), Dey (2014) and Bdair (2012)). As they are mainly applied to continuous distributions, the set of points $\mathcal{T}$ in which the empirical and theoretical distributions will be evaluated is obviously very important. It is natural to set $\mathcal{T}$ to be random using empirical quantiles. That leads us to the distribution of uniform order statistics and, in case of the distance (1), to the conventional weighted least squares estimator that seeks to minimize the distance between the vector of "uniformized" order statistics and

the corresponding vector of expected values, proposed by Swain et al. (1988). However, it was mentioned in Swain et al. (1988), based on practice, that this method, based on ordered statistics, failed to achieve the quality that had been expected and they suggested a different weighting matrix in Johnson's translation system. In contrast, the fixed choice of $\mathcal{T}$ leads us directly to the classical Pearson minimum chi-square estimator for which best asymptotically normal (BAN) distribution properties are well known (see e.g. Hsiao (2006) for its BAN properties and see also Amemiya (1976), Berkson (1949), Berkson (1980), Bhapkar (1966), Fisher (1924), Taylor (1953) for more about minimum chi-square estimation). However, the fixed choice is more naturally made with discrete distributions than with continuous. At the end of the last section we give an iterative procedure which does produce a BAN estimator through the minimization of (1) and random $\mathcal{T}$, based on ordered statistics, which can be naturally applied to continuous distributions.

2 Common Framework

Fix $k \in \mathbb{N}$ and $n \in \mathbb{N}$ and let $r_1, r_2, \ldots, r_{k+1}$ be random variables with sum 1, let $\rho_1, \rho_2, \ldots, \rho_{k+1}$ be their expectations, and for $j \leq k+1$ let

$$R_j = \sum_{i=1}^{j} r_i \text{ and } \mathcal{R}_j = \sum_{i=1}^{j} \rho_i$$

be their cumulative sums. We are interested in the differences $R_j - \mathcal{R}_j$. Suppose that there is a constant $c = c_n$ such that

$$\mathrm{Cov}(R_j, R_l) = \frac{1}{c}\mathcal{R}_j(1 - \mathcal{R}_l) = \frac{1}{c}V_{jl} \tag{4}$$

for $j \leq l$. Let

$$\boldsymbol{R} = (R_1, \ldots, R_k)^T \text{ and } \boldsymbol{\mathcal{R}} = (\mathcal{R}_1, \ldots, \mathcal{R}_k)^T. \tag{5}$$

In this paper we highlight the relationship between the quadratic forms $(\boldsymbol{R}-\boldsymbol{\mathcal{R}})^T\boldsymbol{V}^{-1}(\boldsymbol{R}-\boldsymbol{\mathcal{R}})$ and $\sum_{j=1}^{k+1}\frac{(r_j-\rho_j)^2}{\rho_j}$. We show they are equal and examine properties of ingredients of these statistics by matrix decompositions and by geometrical projection properties. In particular, we confirm the tridiagonal form of $\boldsymbol{V}^{-1}$ and decompose it as $\boldsymbol{V}^{-1} = \boldsymbol{W}^T\boldsymbol{W}$ with $\boldsymbol{W}$ bidiagonal. Furthermore, we show the factor $\boldsymbol{W}(\boldsymbol{R} - \boldsymbol{\mathcal{R}})$ of the quadratic form has uncorrelated entries proportional to $R_{j+1}\mathcal{R}_j - R_j\mathcal{R}_{j+1}$.

We have the following cases for $X_1, \ldots, X_n$ i.i.d. with distribution function F.

Case 1: With fixed $t_1 < \cdots < t_k$ and $t_0 = -\infty$, $t_{k+1} = \infty$ we set

$$R_j = F_n(t_j) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t_j\}},$$

with expectations $\mathcal{R}_j = F(t_j)$. These R_j and $\mathcal{R}_j$ have increments

$$r_j = F_n(t_j) - F_n(t_{j-1}) = P_n(A_j) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A_j\}}$$

and $\rho_j = F(t_j) - F(t_{j-1}) = P(A_j)$, respectively. Now the covariance is $1/n$ times the covariance in a single draw, so the expression (4) holds with $c = n$.

Case 2: With fixed integers $1 \le n_1 < n_2 < \cdots < n_k \le n$ and ordered statistics

$$X_{(n_1)} \le X_{(n_2)} \le \cdots \le X_{(n_k)}$$

we set $t_j = X_{(n_j)}$ and

$$R_j = F(X_{(n_j)})$$

with expectation $\mathcal{R}_j = n_j/(n+1)$. These have increments $r_j = P(A_j)$ and $\rho_j = (n_j - n_{j-1})/(n+1)$. Now, when F is continuous the joint distribution of the R_j is the Dirichlet distribution of uniform quantiles and the covariance expression (4) holds for $c = n + 2$.

In both cases we are examining distribution properties of $R_j - \mathcal{R}_j$. It is $F_n(t_j) - F(t_j)$ in Case 1 and $F(t_j) - F_n(t_j)n/(n+1)$ in Case 2. Thus, the difference $\boldsymbol{R} - \boldsymbol{\mathcal{R}}$ is a vector of centered cumulative distributions. In Case 1 it is the centering of the empirical distribution at $t_1, \ldots, t_k$ and in Case 2 it is the centering of the hypothesized distribution function evaluated at the quantiles $X_{(n_1)}, X_{(n_2)}, \ldots, X_{(n_k)}$.

3 Relationship between the standardized cumulative distribution and the chi-square statistic

We have two approaches to appreciating the relationship between the standardized cumulative distribution and the chi-square statistic. Firstly, we use matrix calculations to obtain the following identity:

$$(\boldsymbol{R} - \boldsymbol{\mathcal{R}})^T \boldsymbol{V}^{-1} (\boldsymbol{R} - \boldsymbol{\mathcal{R}}) = \sum_{j=1}^{k+1} \frac{(r_j - \rho_j)^2}{\rho_j}. \tag{6}$$

After that, we revisit the matter from the geometrical perspective of orthogonal projection.

The first approach uses the form of the covariance matrix

$$\boldsymbol{V} = \begin{bmatrix} \mathcal{R}_1(1-\mathcal{R}_1) & \mathcal{R}_1(1-\mathcal{R}_2) & \cdots & \mathcal{R}_1(1-\mathcal{R}_k) \\ \mathcal{R}_1(1-\mathcal{R}_2) & \mathcal{R}_2(1-\mathcal{R}_2) & \cdots & \mathcal{R}_2(1-\mathcal{R}_k) \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{R}_1(1-\mathcal{R}_k) & \mathcal{R}_2(1-\mathcal{R}_k) & \cdots & \mathcal{R}_k(1-\mathcal{R}_k) \end{bmatrix}.$$

A matrix of this form is said to have the triangle property (Barrett (1978)). General characterization of the inverses of positive definite symmetric tridiagonal matrices (see Barrett (1978) and Barrett (1979)) enable expression of $\boldsymbol{V}^{-1}$ in the following tridiagonal form:

$$\boldsymbol{V}^{-1} = \begin{bmatrix} \frac{1}{\rho_1}+\frac{1}{\rho_2} & -\frac{1}{\rho_2} & 0 & \cdots & 0 & 0 \\ -\frac{1}{\rho_2} & \frac{1}{\rho_2}+\frac{1}{\rho_3} & -\frac{1}{\rho_3} & \cdots & 0 & 0 \\ 0 & -\frac{1}{\rho_3} & \frac{1}{\rho_3}+\frac{1}{\rho_4} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\rho_{k-1}}+\frac{1}{\rho_k} & -\frac{1}{\rho_k} \\ 0 & 0 & 0 & \cdots & -\frac{1}{\rho_k} & \frac{1}{\rho_k}+\frac{1}{\rho_{k+1}} \end{bmatrix}.$$

Also, as a consequence of the QR decomposition of a symmetric tridiagonal matrix (see e.g. Bar-On (1997)), we can see that $\boldsymbol{V}^{-1} = \boldsymbol{W}^T\boldsymbol{W}$, where

$$\boldsymbol{W} = \begin{bmatrix} -\frac{\mathcal{R}_2}{\sqrt{\mathcal{R}_1\mathcal{R}_2\rho_2}} & \frac{\mathcal{R}_1}{\sqrt{\mathcal{R}_1\mathcal{R}_2\rho_2}} & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\mathcal{R}_3}{\sqrt{\mathcal{R}_2\mathcal{R}_3\rho_3}} & \frac{\mathcal{R}_2}{\sqrt{\mathcal{R}_2\mathcal{R}_3\rho_3}} & \cdots & 0 & 0 \\ 0 & 0 & -\frac{\mathcal{R}_4}{\sqrt{\mathcal{R}_3\mathcal{R}_4\rho_4}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{\mathcal{R}_k}{\sqrt{\mathcal{R}_{k-1}\mathcal{R}_k\rho_k}} & \frac{\mathcal{R}_{k-1}}{\sqrt{\mathcal{R}_{k-1}\mathcal{R}_k\rho_k}} \\ 0 & 0 & 0 & \cdots & 0 & -\frac{1}{\sqrt{\mathcal{R}_k\rho_{k+1}}} \end{bmatrix}.$$

These forms of the matrices $\boldsymbol{V}^{-1}$ and $\boldsymbol{W}$ are verified by matrix multiplication: $\boldsymbol{V}^{-1}\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{V}^{-1} = \boldsymbol{W}^T\boldsymbol{W}$ (see Appendix A).

In order to show the equation (6), note that the non-zero elements of the matrix $\mathbf{V}^{-1}$ can be written in the following way:

$$\begin{aligned} \tfrac{1}{\rho_1}+\tfrac{1}{\rho_2} &= \tfrac{1}{\mathcal{R}_1}+\tfrac{1}{\mathcal{R}_2-\mathcal{R}_1}, & -\tfrac{1}{\rho_2} &= -\tfrac{1}{\mathcal{R}_2-\mathcal{R}_1}, \\ \tfrac{1}{\rho_2}+\tfrac{1}{\rho_3} &= \tfrac{1}{\mathcal{R}_2-\mathcal{R}_1}+\tfrac{1}{\mathcal{R}_3-\mathcal{R}_2}, & -\tfrac{1}{\rho_3} &= -\tfrac{1}{\mathcal{R}_3-\mathcal{R}_2}, \\ &\vdots & &\vdots \\ -\tfrac{1}{\rho_k} &= -\tfrac{1}{\mathcal{R}_k-\mathcal{R}_{k-1}}, & \tfrac{1}{\rho_k}+\tfrac{1}{\rho_{k+1}} &= \tfrac{1}{1-\mathcal{R}_k}+\tfrac{1}{\mathcal{R}_k-\mathcal{R}_{k-1}}. \end{aligned}$$

Consequently, observing a number of cancellations in computation of the quadratic form, we obtain

$$\begin{aligned} (\boldsymbol{R}-\boldsymbol{\mathcal{R}})^T\boldsymbol{V}^{-1}(\boldsymbol{R}-\boldsymbol{\mathcal{R}}) &= \frac{1}{\mathcal{R}_1}(R_1-\mathcal{R}_1)^2 + \frac{1}{1-\mathcal{R}_k}(R_k-\mathcal{R}_k)^2 \\ &\quad + \sum_{j=2}^{k}\frac{1}{\mathcal{R}_j-\mathcal{R}_{j-1}}(R_{j-1}-R_j-\mathcal{R}_{j-1}+\mathcal{R}_j)^2 \\ &= \sum_{j=1}^{k+1}\frac{(r_j-\rho_j)^2}{\rho_j}. \end{aligned}$$

The equation (6) can be also reached from examination of projection properties. First note that there is an invertible linear relationship between the cumulative R_j and individual r_j values via

$$R_j = \sum_{i=1}^{j} r_i \text{ and } r_j = R_j - R_{j-1}, \quad j = 1, 2, \ldots, k+1.$$

Accordingly, we will have the same norm-squares

$$c_n(\boldsymbol{R} - \boldsymbol{\mathcal{R}})^T \boldsymbol{V}^{-1} (\boldsymbol{R} - \boldsymbol{\mathcal{R}}) \text{ and } c_n(\boldsymbol{r} - \boldsymbol{\rho})^T \boldsymbol{C}^{-1} (\boldsymbol{r} - \boldsymbol{\rho})$$

for standardized version of the vectors $\boldsymbol{R}$ and $\boldsymbol{r}$ where $\boldsymbol{C}/c_n$ is the covariance matrix of the vector $\boldsymbol{r}$ with $\boldsymbol{C}_{i,j} = \rho_i \delta_{ij} - \rho_i \rho_j$. (Here $\delta_{ij} = \mathbf{1}_{\{i=j\}}$.) Per equation (5) these vectors $\boldsymbol{R}$ and $\boldsymbol{\mathcal{R}}$ are in $\mathbb{R}^k$ with the understanding that $R_{k+1} = 1$. Likewise we take $\boldsymbol{r}$ and $\boldsymbol{\rho}$ to be vectors in $\mathbb{R}^k$ because the value $r_{k+1} = 1 - \sum_{j=1}^{k} r_j$ is linearly determined from the others. Correspondingly, $\boldsymbol{V}$ and $\boldsymbol{C}$ are $k \times k$ covariance matrices. It is known (and easily checked) that the matrix $\boldsymbol{C}^{-1}$ has entries $(\boldsymbol{C}^{-1})_{i,j} = \frac{1}{\rho_i}\delta_{ij} - \frac{1}{\rho_{k+1}}$ for $i, j = 1, 2, \ldots, k$ (matching the Fisher information of the multinomial) and one finds from this form that $(\boldsymbol{r} - \boldsymbol{\rho})^T \boldsymbol{C}^{-1} (\boldsymbol{r} - \boldsymbol{\rho})$ is algebraically the same as

$$\sum_{j=1}^{k+1} \frac{(r_j - \rho_j)^2}{\rho_j}$$

as stated in Neyman (1949). So this is another way to see (6).

Furthermore, using suitable orthogonal vectors one can see how the chi-square statistic arises as the norm square of the fully standardised cumulative distributions.

The chi-square value $\sum_{j=1}^{k+1} \frac{(r_j - \rho_j)^2}{\rho_j}$ is the norm square $\| \boldsymbol{\xi} - \boldsymbol{u} \|^2$ of the difference between the vector with entries $\xi_j = \frac{r_j}{\sqrt{\rho_j}}$ and the unit vector $\boldsymbol{u}$ with entries $\sqrt{\rho_j}$, for $j = 1, \ldots, k+1$. Here we examine the geometry of the situation in $\mathbb{R}^{k+1}$. The projection of $\boldsymbol{\xi}$ in the direction of the unit vector $\boldsymbol{u}$ has length $\boldsymbol{\xi}^T \boldsymbol{u} = \sum_{j=1}^{k+1} \left(\frac{r_j}{\sqrt{\rho_j}} \right) \sqrt{\rho_j}$ equal to 1. The difference $\boldsymbol{\xi} - \boldsymbol{u}$ is the error of this projection. Work with an orthonormal basis of $\mathbb{R}^{k+1}$, in which one of the basis vectors is $\boldsymbol{u}$ (and hence the k other orthonormal vectors are orthogonal to $\boldsymbol{u}$). In particular, let $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_k$ and $\boldsymbol{q}_{k+1} = \boldsymbol{u}$ be any such orthonormal vectors in $\mathbb{R}^{k+1}$. The chi-square value $\| \boldsymbol{\xi} - \boldsymbol{u} \|^2$ is the squared length of the projection of $\boldsymbol{\xi}$ onto the space orthogonal to $\boldsymbol{u}$, spanned by $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k$. So it is given by $\sum_{j=1}^{k} Z_j^2$ where $Z_j = \boldsymbol{\xi}^T \boldsymbol{q}_j$, $j = 1, 2, \ldots, k$, or equivalently $Z_j = (\boldsymbol{\xi} - \boldsymbol{u})^T \boldsymbol{q}_j$.

This sort of analysis is familiar in linear regression theory. A difference here is that the entries of $\boldsymbol{\xi}$ are not uncorrelated. Nevertheless, the covariance $E(\boldsymbol{\xi} - \boldsymbol{u})(\boldsymbol{\xi} - \boldsymbol{u})^T$ reduces to $\frac{1}{c_n}[\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T]$ since it has entries

$$E \frac{(r_j - \rho_j)(r_l - \rho_l)}{\sqrt{\rho_j \rho_l}} = \frac{1}{c_n} \frac{\rho_j \mathbf{1}_{j=l} - \rho_j \rho_l}{\sqrt{\rho_j \rho_l}}$$

which simplifies to

$$\frac{1}{c_n}(\delta_{jl} - \sqrt{\rho_j}\sqrt{\rho_l}).$$

Accordingly, $EZ_jZ_l = E\boldsymbol{q}_j^T(\boldsymbol{\xi}-\boldsymbol{u})(\boldsymbol{\xi}-\boldsymbol{u})^T\boldsymbol{q}_l = \frac{1}{c_n}\boldsymbol{q}_j^T(\boldsymbol{I}-\boldsymbol{u}\boldsymbol{u}^T)\boldsymbol{q}_l$ is $\frac{1}{c_n}\boldsymbol{q}_j^T\boldsymbol{q}_l$ equal to 0 for $j \neq l$. Thus the Z_j are indeed uncorrelated and have constant variance $\frac{1}{c_n}$.

This is a standard way in which we know that the chi-square statistic with $k+1$ cells is a sum of k uncorrelated and standardized random variables (c.f. Cramer (1946), pages 416-420).

4 A convenient choice of orthogonal vectors

Here we wish to benefit from an explicit choice of the orthonormal vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k$ orthogonal to $\boldsymbol{q}_{k+1} = \boldsymbol{u}$. We are motivated in this by the analysis in Stigler (1984). For an i.i.d. sample $Y_1 \ldots Y_n$ from $\mathcal{N}(\mu, \sigma^2)$, the statistic $\sum_{j=1}^n (Y_j - \bar{Y}_n)^2$ is the sum of squares $\sum_{j=2}^n \frac{j-1}{j}(Y_j - \bar{Y}_{j-1})^2$ of the independent $\mathcal{N}(0, \sigma^2)$ innovations (also known as standardized prediction errors) $Z_j = \frac{Y_j - \bar{Y}_{j-1}}{\sqrt{1+1/(j-1)}}$ and, accordingly, this sum of squares is explicitly σ^2 times a chi-square distributed random variable with $n-1$ degrees of freedom. These innovations decorrelate the vector of $(Y_i - \bar{Y}_n)$ using $\boldsymbol{q}_j$ like those below, with ρ_i replaced with $\frac{1}{n}$. According to Stigler (1984) and Kruskal (1946), analysis of this type originates with Helmert (1876) (cf. Rao (1973), pp. 182–183).

The analogous choice for our setting is to let $Z_j = \boldsymbol{\xi}^T\boldsymbol{q}_j$, where the $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k, \boldsymbol{q}_{k+1}$ are the normalizations of the following orthogonal vectors in $\mathbb{R}^{k+1}$:

$$\begin{bmatrix} -\sqrt{\rho_1} & -\sqrt{\rho_1} & -\sqrt{\rho_1} & \cdots & -\sqrt{\rho_1} & \sqrt{\rho_1} \\ \frac{\mathcal{R}_1}{\sqrt{\rho_2}} & -\sqrt{\rho_2} & -\sqrt{\rho_2} & \cdots & -\sqrt{\rho_2} & \sqrt{\rho_2} \\ 0 & \frac{\mathcal{R}_2}{\sqrt{\rho_3}} & -\sqrt{\rho_3} & \cdots & -\sqrt{\rho_3} & \sqrt{\rho_3} \\ 0 & 0 & \frac{\mathcal{R}_3}{\sqrt{\rho_4}} & \cdots & -\sqrt{\rho_4} & \sqrt{\rho_4} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\sqrt{\rho_k} & \sqrt{\rho_k} \\ 0 & 0 & 0 & \cdots & \frac{\mathcal{R}_k}{\sqrt{\rho_{k+1}}} & \sqrt{\rho_{k+1}} \end{bmatrix}. \tag{7}$$

Essentially the same choices of orthogonal q_j for determination of uncorrelated components Z_j of $\boldsymbol{\xi} - \boldsymbol{u}$ are found in Irwin (1949). See also Irwin (1942), as well as Lancaster (1949) and Lancaster (1965) where the matrix from Irwin (1949) is explained as a particular member of a class of generalizations of the Helmert matrix.

The norm of the j-th such column for $j = 1, \ldots, k$ equals $\sqrt{\mathcal{R}_j + \frac{\mathcal{R}_j^2}{\rho_{j+1}}}$ which is $\sqrt{\frac{\mathcal{R}_j \mathcal{R}_{j+1}}{\rho_{j+1}}}$, so that, for $j = 1, \ldots, k$,

$$\boldsymbol{q}_j = \frac{1}{\sqrt{\frac{\mathcal{R}_j \mathcal{R}_{j+1}}{\rho_{j+1}}}} \left[-\sqrt{\rho_1}, \ldots, -\sqrt{\rho_j}, \frac{\mathcal{R}_j}{\sqrt{\rho_{j+1}}}, 0, \ldots, 0 \right]^T$$

and

$$Z_j = \boldsymbol{\xi}^T \boldsymbol{q}_j \;\text{ with }\; \xi_i = \frac{r_i}{\sqrt{\rho_i}}$$

becomes

$$Z_j = \frac{-r_1 - \cdots - r_j + \frac{r_{j+1}\mathcal{R}_j}{\rho_{j+1}}}{\sqrt{\frac{\mathcal{R}_j \mathcal{R}_{j+1}}{\rho_{j+1}}}}.$$

This is

$$Z_j = \frac{r_{j+1}\mathcal{R}_j - R_j \rho_{j+1}}{\sqrt{\mathcal{R}_j \mathcal{R}_{j+1} \rho_{j+1}}}$$

or, equivalently, for $j = 1, 2, \ldots, k$

$$Z_j = \frac{R_{j+1}\mathcal{R}_j - R_j \mathcal{R}_{j+1}}{\sqrt{\mathcal{R}_j \mathcal{R}_{j+1} \rho_{j+1}}}$$

which are the innovations of the cumulative values R_{j+1} (the standardized error of linear prediction of R_{j+1} using $R_1, \ldots, R_j$). As a consequence of the above properties of the $\boldsymbol{q}_j$, these Z_j are mean zero, uncorrelated, and of constant variance $1/c_n$. Each of these facts also can be checked directly using $ER_j = \mathcal{R}_j$ and using the specified form of the covariance $\mathrm{Cov}(R_j, R_l) = \frac{1}{c_n}[\min(\mathcal{R}_j, \mathcal{R}_l) - \mathcal{R}_j \mathcal{R}_l]$.

As we have said, any choice of orthogonal vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k$, orthogonal to the vector $\boldsymbol{u}$, may be used in showing the identity (6). The advantage of the choice (7) is the simplicity of the resulting components $Z_1, \ldots, Z_k$ and their direct relationship to the cumulative distribution. Furthermore, this choice makes these Z_j match the entries of $\boldsymbol{W}(\boldsymbol{R} - \boldsymbol{\mathcal{R}})$ when $\boldsymbol{W}$ is chosen to be the bidiagonal factor in the representation $\boldsymbol{V}^{-1} = \boldsymbol{W}^T \boldsymbol{W}$ of the tridiagonal $\boldsymbol{V}^{-1}$. We remark that the matrix inverse and orthonormal projection proofs of the equivalence of the weighted norm squares of (1), (2) and (3) may also be seen as specialization of Lemma 2 in the Appendix A concerning weighted inner products of vectors built from partial sums.

To summarize this section, specialized to Case 1, let us point out that we find an explicit standardization

$$Z_j = \frac{F_n(t_{j+1})F(t_j) - F_n(t_j)F(t_{j+1})}{c_{n,j}}, \;\; j = 1, \ldots, k, \tag{8}$$

with $c_{n,j}^2 = F(t_j)F(t_{j+1})P(A_{j+1})/n$. These random variables $Z_1, Z_2, \ldots, Z_k$ have mean 0 and variance 1 and they are uncorrelated. Moreover, the sum of squares

$$\sum_{j=1}^{k} Z_j^2$$

is precisely equal to the statistics given in expressions (1) and (2). It corresponds to a bidiagonal Cholesky decomposition of $\boldsymbol{V}^{-1}$ as $\boldsymbol{W}^T\boldsymbol{W}$ with $\boldsymbol{B}$ given by $-F(t_{j+1})/c_{n,j}$ for the (j,j) entries, $F(t_j)/c_{n,j}$ for the $(j,j+1)$ entries and 0 otherwise, yielding the vector $\boldsymbol{Z} = \boldsymbol{W}(\boldsymbol{F}_n - \boldsymbol{F})$, where $\boldsymbol{F} = (F(t_1), \ldots, F(t_k))^T$, as a full standardization of the vector $\boldsymbol{F}_n = (F_n(t_1), \ldots, F_n(t_k))^T$.

The Z_j may also be written as

$$Z_j = \frac{P_n(A_{j+1})F(t_j) - F_n(t_j)P(A_{j+1})}{c_{n,j}} \tag{9}$$

so its marginal distribution (with an hypothesized F) comes from the trinomial distribution of $(nF_n(t_j), nP_n(A_{j+1}))$. These uncorrelated Z_j, though not independent, suggest finite-sample approximation to the distribution of $\sum_j Z_j^2$ from convolution of the distributions of Z_j^2 rather than the asymptotic chi-square.

Nevertheless, when $t_1, \ldots, t_k$ are fixed, it is clear by the multivariate central limit theorem (for the standardized sum of the i.i.d. random variables comprising $P_n(A_{j+1})$ and $F_n(t_j)$ from (9)) that the joint distribution of $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^T$ is asymptotically $\mathcal{N}(0, \boldsymbol{I})$, providing a direct path to the asymptotic chi-square(k) distribution of the statistic given equivalently in (1), (2) and (3).

A reviewer has suggested to consider a limiting analogue of our decomposition into uncorrelated variables (9). By empirical process theory (van der Vaart and Wellner (1996) or Shorack and Wellner (2009)) $Y_n(t) = \sqrt{n}(F_n(t) - F(t))$ has the same means and covariances as the limiting Gaussian process $B(t) = W(F(t)) - F(t)W(1)$ which is a Brownian bridge $W(\tau) - \tau W(1)$ evaluated at $\tau = F(t)$. Accordingly, our statistics $\sqrt{n}[F_n(t_{j+1})F(t_j) - F(t_{j+1})F_n(t_j)]$, $j = 1, \ldots, k$, which equal $Y_n(t_{j+1})F(t_j) - F(t_{j+1})Y_n(t_j)$, converge in distribution to that of

$$B(t_{j+1})F(t_j) - F(t_{j+1})B(t_j),\ j = 1, \ldots, k$$

which analogously whitens the Brownian bridge.

5 Large sample estimation properties

The results of previous sections will be used to discuss asymptotic efficiency of weighted least squares estimators related to Case 1 and Case 2 of Section 2.

In the previous sections the distribution F was regarded as a fixed hypothesized distribution. This made the choice of the matrix $\boldsymbol{V}$ for standardization especially clear. Now, for estimation, the distribution will be regarded as a member of a parametric family, and the choice of $\boldsymbol{V}$ is accordingly more delicate.

We consider the case of i.i.d. random sample $X_1, \ldots, X_n$ with distribution function from a parametric family $F_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, $t_1 < \cdots < t_k$, and let $\boldsymbol{R}_n = \boldsymbol{R}$ and $\boldsymbol{\mathcal{R}}_n = \boldsymbol{\mathcal{R}}$ be as in Section 2. The vector $\boldsymbol{R}_n - \boldsymbol{\mathcal{R}}_n$, which we denote by $(\boldsymbol{R}_n - \boldsymbol{\mathcal{R}}_n)(\boldsymbol{\theta})$, can be considered as a vector depending on the data and the parameter. Let $\boldsymbol{\theta}_0$ denote the true parameter value. If $(\boldsymbol{R}_n - \boldsymbol{\mathcal{R}}_n)(\boldsymbol{\theta}_0)$ converges to zero in probability $P_{\boldsymbol{\theta}_0}$, it is natural to use the weighted least squares procedure for parameter estimation, so that we minimize the objective function

$$Q_{n,\boldsymbol{V}}(\boldsymbol{\theta}) = [(\boldsymbol{R}_n - \boldsymbol{\mathcal{R}}_n)(\boldsymbol{\theta})]^T \boldsymbol{V}^{-1} [(\boldsymbol{R}_n - \boldsymbol{\mathcal{R}}_n)(\boldsymbol{\theta})] \tag{10}$$

for $\boldsymbol{\theta} \in \Theta$. The matrix $\boldsymbol{V}$ for complete standardization is the covariance matrix of $\sqrt{c_n}(\boldsymbol{R}_n - \boldsymbol{\mathcal{R}}_n)$ with entries $V_{ij} = \mathcal{R}_i(1 - \mathcal{R}_j)$ for $i \leq j$. How we deal with possible dependence of $\boldsymbol{V}$ on the parameter $\boldsymbol{\theta}$ is discussed below. In some situations we may use a known value of $\boldsymbol{V}$ at the true $\boldsymbol{\theta}_0$ (as in Case 2, where $V_{ij} = (n_i/n)(1 - n_j/n)$) or we may either use a consistent estimate of $\boldsymbol{\theta}_0$ or use the current parameter $\boldsymbol{\theta}$. Implications of these choices for asymptotic efficiency are discussed.

Both cases from Section 2, i.e. fixed and random $t_1, \ldots, t_k$, are considered in the estimation context. Indeed, for Case 2 (random $t_1 < \cdots < t_k$, $t_j = X_{(n_j)}$) we have:

$$\boldsymbol{R}_n(\boldsymbol{\theta}) = [F_{\boldsymbol{\theta}}(X_{(n_1)}), \ldots, F_{\boldsymbol{\theta}}(X_{(n_k)})]^T$$

$$\boldsymbol{\mathcal{R}}_n = [F_n(X_{(n_1)}), \ldots, F_n(X_{(n_k)})]^T \frac{n}{n+1}.$$

The estimator in this case coincides with the estimator proposed in Swain et al. (1988). Here only the $\boldsymbol{R}_n(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$ and $F_{\boldsymbol{\theta}_0}(X_{(n_j)})$ has a Beta$(n_j, n + 1 - n_j)$ distribution and $E_{\boldsymbol{\theta}_0}[F_{\boldsymbol{\theta}_0}(X_{(n_j)})] = \frac{n_j}{n+1} = F_n(X_{(n_j)})\frac{n}{n+1}$ so that

$$\boldsymbol{\mathcal{R}}_n = \left[\frac{n_1}{n+1}, \ldots, \frac{n_k}{n+1}\right]^T.$$

For Case 1 (fixed $t_1 < \cdots < t_k$) we have

$$\boldsymbol{R}_n = [F_n(t_1), \ldots, F_n(t_k)]^T, \quad F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$$

and

$$\boldsymbol{\mathcal{R}}_n(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \boldsymbol{R}_n = [F_{\boldsymbol{\theta}}(t_1), \ldots, F_{\boldsymbol{\theta}}(t_k)]^T.$$

The estimator in this case coincides with the estimator considered in Benšić (2015). Here only the expectation $\boldsymbol{\mathcal{R}}_n(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$.

In both cases, the convergence in probability $P_{\boldsymbol{\theta}_0}$ of $(\boldsymbol{R}_n - \boldsymbol{\mathcal{R}}_n)(\boldsymbol{\theta}_0)$ to zero is a consequence of the form of the variances of the mean zero differences,

which are $(1/n)\mathcal{R}_j(1-\mathcal{R}_j)$ in Case 1 and $(1/(n+2))\mathcal{R}_j(1-\mathcal{R}_j)$ in Case 2, in accordance with expression (4).

In both cases we will assume that $\boldsymbol{F}_{\boldsymbol{\theta}}(t)$ and its gradient $\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{F}_{\boldsymbol{\theta}}(t)$ are continuous. Additional regularity assumptions from cited literature may arise in the discussion below.

There are some differences in the analysis of estimation properties for the two mentioned cases. Let us discuss them separately.

Case 1. For the fixed $t_1 < \cdots < t_k$, $\boldsymbol{F}_{\boldsymbol{\theta}} = (F_{\boldsymbol{\theta}}(t_1), \ldots, F_{\boldsymbol{\theta}}(t_k))^T$ and $\boldsymbol{F}_n = (F_n(t_1), \ldots, F_n(t_k))^T$ we can express the function (10) as

$$Q_{n,\boldsymbol{V}}(\boldsymbol{\theta}) = (\boldsymbol{F}_n - \boldsymbol{F}_{\boldsymbol{\theta}})^T \boldsymbol{V}^{-1} (\boldsymbol{F}_n - \boldsymbol{F}_{\boldsymbol{\theta}}).$$

Denote by $\boldsymbol{V}_{\boldsymbol{\theta}}$ the covariance matrix of $\sqrt{n}(\boldsymbol{F}_n - \boldsymbol{F}_{\boldsymbol{\theta}})$ which, as we know, has entries $F_{\boldsymbol{\theta}}(t_i)(1 - F_{\boldsymbol{\theta}}(t_j))$ for $i \leq j$. Note here that $\boldsymbol{V}_{\boldsymbol{\theta}}$ depends on the parameter. This leads first to the objective function $Q_n(\boldsymbol{\theta}) = Q_{n,\boldsymbol{V}_{\boldsymbol{\theta}}}(\boldsymbol{\theta})$. Equation (6) from Section 3 then guarantees that minimizing this function $Q_{n,\boldsymbol{V}_{\boldsymbol{\theta}}}(\boldsymbol{\theta})$ leads to the classical Pearson minimum chi-square estimator (see e.g. Hsiao (2006) for its best asymptotically normal (BAN) distribution properties and see also Amemiya (1976), Berkson (1949), Berkson (1980), Bhapkar (1966), Fisher (1924), Taylor (1953) for more about minimum chi-square estimation).

Estimation in this case can also be set in the framework of the generalized method of moments (GMM), with alternative choices of the covariance for standardization. Indeed, if we use a fixed $\boldsymbol{V}$ or we use $\boldsymbol{V}_{\boldsymbol{\theta}^\star}$ where $\boldsymbol{\theta}^\star$ is a consistent estimator of the true parameter value instead of $\boldsymbol{V}_{\boldsymbol{\theta}}$ in the function $Q_{n,\boldsymbol{V}}(\boldsymbol{\theta})$, then, as shown in Benšić (2015), this estimation procedure can be seen as a GMM procedure.

Let $\hat{\boldsymbol{\theta}}_{k,n}$ denote the estimator obtained by minimization of the function $Q_{n,\boldsymbol{V}_{\boldsymbol{\theta}^\star}}(\boldsymbol{\theta})$. Refining the notation from Section 2:

$A_i = (t_{i-1}, t_i]$, $i = 1, \ldots, k$, $A_{k+1} = (t_k, \infty)$,
$P_n(A_i) = F_n(t_i) - F_n(t_{i-1})$,
$P^\star(A_i) = F_{\boldsymbol{\theta}^\star}(t_i) - F_{\boldsymbol{\theta}^\star}(t_{i-1})$,
$P(A_i;\boldsymbol{\theta}) = F_{\boldsymbol{\theta}}(t_i) - F_{\boldsymbol{\theta}}(t_{i-1})$
$P_{\boldsymbol{\theta}_0}(A_i) = F_{\boldsymbol{\theta}_0}(t_i) - F_{\boldsymbol{\theta}_0}(t_{i-1})$,

from the tridiagonal form of the weighting matrix and equation (6) we see that

$$\hat{\boldsymbol{\theta}}_{k,n} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{k+1} \frac{(P_n(A_i) - P(A_i;\boldsymbol{\theta}))^2}{P^\star(A_i)}. \tag{11}$$

If classical regularity assumptions of the generalized method of moments theory are fulfilled (see e.g. Newey and McFadden (1994), Harris and Matyas (1999)) it is shown in Benšić (2015) that $\lim_n[n\mathrm{Var}(\hat{\boldsymbol{\theta}}_{k,n})]$ has inverse $\boldsymbol{G}_{\boldsymbol{\theta}_0}^T \boldsymbol{V}_{\boldsymbol{\theta}_0}^{-1} \boldsymbol{G}_{\boldsymbol{\theta}_0}$ where $\boldsymbol{G}_{\boldsymbol{\theta}_0}$ and $\boldsymbol{V}_{\boldsymbol{\theta}_0}$ are, respectively, the Jacobian matrix $\frac{\partial}{\partial\boldsymbol{\theta}}[F_{\boldsymbol{\theta}}(t_1), \ldots, F_{\boldsymbol{\theta}}(t_k)]^T$ and the covariance matrix of $\sqrt{n}[F_n(t_1), \ldots, F_n(t_k)]^T$ evaluated at the true parameter value $\boldsymbol{\theta}_0$.

It is fruitful to examine the quantity $\boldsymbol{G}_{\boldsymbol{\theta}}^T \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \boldsymbol{G}_{\boldsymbol{\theta}}$ and how it simplifies. Using Lemma 2 in the Appendix A, with $\boldsymbol{A}$ and $\boldsymbol{B}$ chosen as columns of $\boldsymbol{G}_{\boldsymbol{\theta}}$ and $\rho_i = P_{\boldsymbol{\theta}}(A_i) = F_{\boldsymbol{\theta}}(t_i) - F_{\boldsymbol{\theta}}(t_{i-1})$, the tridiagonal form of $\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}$ allows simplification of $\boldsymbol{G}_{\boldsymbol{\theta}}^T \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \boldsymbol{G}_{\boldsymbol{\theta}}$ to see that it equals

$$I^{\mathcal{T}}(\boldsymbol{\theta}) = \sum_{j=1}^{k+1} \frac{1}{P_{\boldsymbol{\theta}}(A_j)} [\frac{\partial}{\partial \boldsymbol{\theta}} P_{\boldsymbol{\theta}}(A_j)][\frac{\partial}{\partial \boldsymbol{\theta}} P_{\boldsymbol{\theta}}(A_j)]^T \tag{12}$$

which one recognize to be the Fisher information of a multinomial with probabilities $P_{\boldsymbol{\theta}}(A_j)$, $j = 1, \ldots, k+1$. The interpretation is that, when restricted to the multinomial counts in the partition formed by $\mathcal{T}$, the GMM procedure (here shown to be related to the minimum chi-square) inherits the asymptotic efficiency for that multinomial setting. The relative efficiency using a fixed partition $\mathcal{T}$ compared to the full data situation is given, in the scalar parameter case, by the ratio of $I^{\mathcal{T}}(\theta)/I(\theta)$ where $I(\theta) = \int \frac{1}{f(x,\theta)} (\frac{\partial}{\partial \theta} f(x,\theta))^2 \, dx$ is the full Fisher information.

For additional understanding of the inverse of $\lim_n [n\mathrm{Var}(\hat{\boldsymbol{\theta}}_{k,n})]$ suppose the model has a differentiable density $f(\boldsymbol{x}, \boldsymbol{\theta})$ satisfying the classical regularity, and let $\boldsymbol{S}(x) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(x, \boldsymbol{\theta})|_{\boldsymbol{\theta}_0}$ be the population score function evaluated at the true parameter value. Now we have

$$\boldsymbol{G}_{\boldsymbol{\theta}_0} = \begin{bmatrix} [E_{\boldsymbol{\theta}_0}(\boldsymbol{S}\mathbf{1}_{(-\infty,t_1]})]^T \\ \vdots \\ [E_{\boldsymbol{\theta}_0}(\boldsymbol{S}\mathbf{1}_{(-\infty,t_k]})]^T \end{bmatrix}$$

and

$$\begin{aligned} \boldsymbol{G}_{\boldsymbol{\theta}_0}^T \boldsymbol{V}_{\boldsymbol{\theta}_0}^{-1} \boldsymbol{G}_{\boldsymbol{\theta}_0} &= \sum_{i=1}^{k+1} \frac{1}{P_{\boldsymbol{\theta}_0}(A_i)} \int_{t_{i-1}}^{t_i} \boldsymbol{S}(x) f(x; \boldsymbol{\theta}_0) \, dx \int_{t_{i-1}}^{t_i} \boldsymbol{S}^T(x) f(x; \boldsymbol{\theta}_0) \, dx \\ &= \sum_{i=1}^{k+1} P_{\boldsymbol{\theta}_0}(A_i) \frac{\int_{t_{i-1}}^{t_i} \boldsymbol{S}(x) f(x; \boldsymbol{\theta}_0) \, dx}{P_{\boldsymbol{\theta}_0}(A_i)} \frac{\int_{t_{i-1}}^{t_i} \boldsymbol{S}^T(x) f(x; \boldsymbol{\theta}_0) \, dx}{P_{\boldsymbol{\theta}_0}(A_i)} \\ &= \sum_{i=1}^{k+1} P_{\boldsymbol{\theta}_0}(A_i) E_{\boldsymbol{\theta}_0}[\boldsymbol{S}(X)|A_i] E_{\boldsymbol{\theta}_0}[\boldsymbol{S}(X)|A_i]^T. \end{aligned}$$

This can be interpreted as a Riemann-Stieltjes discretization of the Fisher information which arises in the limit of large k. So the GMM in this deterministic partition procedure is fully efficient in the limit as first $n \to \infty$ and then $k \to \infty$.

Let us note the similarity of $\hat{\boldsymbol{\theta}}_{k,n}$ and the minimum chi-square estimator. From (11), we see that they differ only in the denominator, so we can interpret $\hat{\boldsymbol{\theta}}_{k,n}$ as a modified minimum chi-square. It is well known that various minimum chi-square estimators are in fact generalized least squares (see e.g. Amemiya

(1976), Harris and Kanji (1983), Hsiao (2006)) and BAN estimators. Likewise, the norm square of standardizing the empirical distribution has been known to also provide a generalized least squares estimator. Here we give a clear and simple way that summarize these findings through the complete standardization of the empirical distribution.

Case 2. In this case we have $t_j = X_{(n_j)}$ so that the value $\mathcal{R}_j = F_n(t_j) = n_j/n$ is predetermined. The random part within $Q_{n,\boldsymbol{V}}(\boldsymbol{\theta})$ is $R_j(\boldsymbol{\theta}) = F_{\boldsymbol{\theta}}(X_{(n_j)})$ which we note depends on the parameter. Nevertheless, the covariance matrix $\boldsymbol{V}$ in this case has constant entries $V_{jl} = (n_j/n)(1 - n_l/n)$ for $j \le l$. Now, the results summarized in Sections 3 (see also Swain et al. (1988)) enable us to represent the minimizer of the function $Q_{n,\boldsymbol{V}}(\boldsymbol{\theta})$ as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{k+1} \frac{((F_{\boldsymbol{\theta}}(X_{(n_i)}) - F_{\boldsymbol{\theta}}(X_{(n_{i-1})})) - \frac{n_i - n_{i-1}}{n+1})^2}{\frac{n_i - n_{i-1}}{n+1}}. \tag{13}$$

As it was mentioned in Swain et al. (1988), page 276, based on practice, there is "a weight matrix which yields fits to empirical CDFs that are usually superior in many respects" to the estimator (13). Nowadays, this can be explained in the view of the generalized spacing estimator (GSE) (see Ghosh and Rao Jammalamadaka (2001), Cheng and Amin (1983), Ranneby (1984)). To discuss this, let us suppose for simplicity that all data are different and $k = n$ so that the estimator can be easily recognized as the GSE. Namely, if $n_i - n_{i-1} = 1$ then

$$Q_n(\boldsymbol{\theta}) = (n+2)(n+1) \sum_{i=1}^{n+1} \left((F_{\boldsymbol{\theta}}(X_{(n_i)}) - F_{\boldsymbol{\theta}}(X_{(n_{i-1})})) - \frac{1}{n+1} \right)^2.$$

Obviously,

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n+1} (F_{\boldsymbol{\theta}}(X_{(n_i)}) - F_{\boldsymbol{\theta}}(X_{(n_{i-1})}))^2 = \sum_{i=1}^{n+1} h(F_{\boldsymbol{\theta}}(X_{(n_i)}) - F_{\boldsymbol{\theta}}(X_{(n_{i-1})})), \tag{14}$$

where $h(t) = t^2$. Detailed discussion about conditions for consistency and asymptotic normality for this type of estimator the interested reader can find in Ghosh and Rao Jammalamadaka (2001). If we apply these results with $h(t) = t^2$ it comes out that we face a lack of BAN distribution properties with $\hat{\boldsymbol{\theta}}_n$. To illustrate this, let us suppose, for simplicity, that $\boldsymbol{\theta} = \theta$ is a scalar. Theorem 3.2 from Ghosh and Rao Jammalamadaka (2001) gives necessary and sufficient condition on h to generate a GSE with minimum variance for a given class of functions which includes $h(t) = t^2$. It is stated there that the asymptotic variance of a GSE is minimized with $h(t) = a \log t + bt + c$ where a, b and c are constants. Based on the results formulated in Theorem 3.1 from the same paper, it is also possible to calculate the asymptotic variance of the GSE for the given function h under some regular conditions on the population density. Thus, for $h(t) = t^2$, the expression (9) in Theorem 3.1 from Ghosh

and Rao Jammalamadaka (2001) is equal to 2. This means that asymptotic variance of our estimator (under mild conditions on the population density) is $\frac{2}{I(\theta_0)}$, where $I(\theta_0)$ denotes the Fisher information. So, for these cases, $\hat{\boldsymbol{\theta}}_n$ from (14) is not BAN. It is only 50% efficient asymptotically.

However, it is possible to reach the BAN distribution property for Case 2 and $k = n$ through an iterative procedure which includes a modification of the denominator in (13) in each step. For simplicity let us discuss the one-dimensional parameter case. Also assume that the density function $f_\theta(x)$ exists and has Fisher information $I(\theta) = \int f_\theta(x)(\frac{\partial}{\partial\theta}\log f_\theta(x))^2\,dx$ that is a bounded function of θ.

1. Let
$$Q_n(\theta,\theta') = \sum_{i=1}^{n+1} \frac{(F_\theta(X_{(i)}) - F_\theta(X_{(i-1)}) - \frac{1}{n+1})^2}{F_{\theta'}(X_{(i)}) - F_{\theta'}(X_{(i-1)})}. \tag{15}$$
2. Let $\theta^\star$ be a consistent estimator for real θ.
3.
$$\theta_1 = \theta^\star$$
$$\theta_{j+1} = \operatorname*{argmin}_\theta Q_n(\theta,\theta_j), \quad j = 1,2,\ldots$$

The use of the denominator $F_{\theta'}(X_{(i)}) - F_{\theta'}(X_{(i-1)})$ in (15) rather than the expected value $(n_i - n_{i-1})/(n+1)$ at the true parameter value is a hybrid that allows to adapt to distribution variability at the most recent parameter value.

To show the desired properties, let us fix the data set $x_1,\ldots,x_n$ for a given sample size n and denote here:

$$\boldsymbol{F}_\theta = [F_\theta(x_{(1)}),\ldots,F_\theta(x_{(n)})]^T, \quad \boldsymbol{G}_\theta = \frac{\partial}{\partial\theta}[F_\theta(x_{(1)}),\ldots,F_\theta(x_{(n)})]^T,$$

and

$$\boldsymbol{V}_\theta = \begin{bmatrix} F_\theta(x_{(1)})(1-F_\theta(x_{(1)})) & F_\theta(x_{(1)})(1-F_\theta(x_{(2)})) & \cdots & F_\theta(x_{(1)})(1-F_\theta(x_{(n)})) \\ F_\theta(x_{(1)})(1-F_\theta(x_{(2)})) & F_\theta(x_{(2)})(1-F_\theta(x_{(2)})) & \cdots & F_\theta(x_{(2)})(1-F_\theta(x_{(n)})) \\ \vdots & \vdots & \vdots & \vdots \\ F_\theta(x_{(1)})(1-F_\theta(x_{(n)})) & F_\theta(x_{(2)})(1-F_\theta(x_{(n)})) & \cdots & F_\theta(x_{(n)})(1-F_\theta(x_{(n)})) \end{bmatrix}.$$

We take advantage of the fact that the $Q_n(\theta,\theta')$ can also be expressed by the tools we have developed. The $\frac{1}{n+1}$ in the definition of $Q_n(\theta,\theta')$ provides the difference in consecutive entries of the vector $\boldsymbol{\mathcal{R}}$ with entries $\mathcal{R}_j = \frac{j}{n+1}$, $j = 1,\ldots,n$. Thus, it holds that

$$Q_n(\theta,\theta') = (\boldsymbol{F}_\theta - \boldsymbol{\mathcal{R}})^T\, \boldsymbol{V}_{\theta'}^{-1}\, (\boldsymbol{F}_\theta - \boldsymbol{\mathcal{R}}).$$

As in the Gauss-Newton method for nonlinear least squares, here we consider the following quadratic approximation $\theta \mapsto \hat{Q}_n(\theta,\theta_j)$,

$$\hat{Q}_n(\theta,\theta_j) = \left(\boldsymbol{F}_{\theta_j} + \boldsymbol{G}_{\theta_j}(\theta-\theta_j) - \boldsymbol{\mathcal{R}}\right)^T \boldsymbol{V}_{\theta_j}^{-1} \left(\boldsymbol{F}_{\theta_j} + \boldsymbol{G}_{\theta_j}(\theta-\theta_j) - \boldsymbol{\mathcal{R}}\right),$$

of the function $\theta \mapsto Q_n(\theta, \theta_j) = (\boldsymbol{F}_\theta - \boldsymbol{\mathcal{R}})^T \boldsymbol{V}_{\theta_j}^{-1} (\boldsymbol{F}_\theta - \boldsymbol{\mathcal{R}})$ about the point θ_j.

Instead of solving the nonlinear optimization problem $\min_\theta Q_n(\theta, \theta_j)$, in every iteration $j = 1, 2, \ldots$ we solve the simpler quadratic minimization problem $\min_\theta \hat{Q}_n(\theta, \theta_j)$, that has explicit solution

$$\operatorname*{argmin}_\theta \hat{Q}_n(\theta, \theta_j) = \theta_j + \left(\boldsymbol{G}_{\theta_j}^T \boldsymbol{V}_{\theta_j}^{-1} \boldsymbol{G}_{\theta_j} \right)^{-1} \boldsymbol{G}_{\theta_j}^T \boldsymbol{V}_{\theta_j}^{-1} \left(\boldsymbol{\mathcal{R}} - \boldsymbol{F}_{\theta_j} \right).$$

Then the corresponding iterative procedure is given by

$$\theta_{j+1} = \theta_j + \left(\boldsymbol{G}_{\theta_j}^T \boldsymbol{V}_{\theta_j}^{-1} \boldsymbol{G}_{\theta_j} \right)^{-1} \boldsymbol{G}_{\theta_j}^T \boldsymbol{V}_{\theta_j}^{-1} \left(\boldsymbol{\mathcal{R}} - \boldsymbol{F}_{\theta_j} \right), \; j = 1, 2, \ldots. \tag{16}$$

This is an iterative algorithm for computation of the estimate.

Generally, it is not easy to obtain conditions that guarantee, for given data, the convergence of the sequence (θ_j), indexed by the iteration number j. Nevertheless, if θ_j converges, the differences $\theta_{j+1} - \theta_j$ converges to zero, and then, provided $\boldsymbol{G}_{\theta_j}^T \boldsymbol{V}_{\theta_j}^{-1} \boldsymbol{G}_{\theta_j}$ is bounded, it follows that

$$\boldsymbol{G}_{\theta_j}^T \boldsymbol{V}_{\theta_j}^{-1} \left(\boldsymbol{\mathcal{R}} - \boldsymbol{F}_{\theta_j} \right) \to 0.$$

Since the function $\theta \mapsto \boldsymbol{G}_\theta^T \boldsymbol{V}_\theta^{-1} (\boldsymbol{\mathcal{R}} - \boldsymbol{F}_\theta)$ is continuous, the limit of the sequence (θ_j) is a solution of the equation

$$\boldsymbol{G}_\theta^T \boldsymbol{V}_\theta^{-1} (\boldsymbol{\mathcal{R}} - \boldsymbol{F}_\theta) = 0. \tag{17}$$

As for the matter of the boundedness of $\boldsymbol{G}_\theta^T \boldsymbol{V}_\theta^{-1} \boldsymbol{G}_\theta$ (used in the identification of this algorithmic limit) we again find that it equals

$$\sum_{i=1}^{n+1} \frac{\left(\frac{\partial}{\partial \theta} (F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})) \right)^2}{F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})}$$

which is the Fisher information $I^{\mathcal{T}}(\theta)$ as in (12) but now it is based on the partition $\mathcal{T} = \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ formed by the data. In general, $I^{\mathcal{T}}(\theta) \leq I(\theta)$ (as confirmed in Appendix B) and this bound holds uniformly over all data $x_1, \ldots, x_n$. We assumed $I(\theta)$ to be a bounded function of θ. Thus, $\boldsymbol{G}_{\theta_j}^T \boldsymbol{V}_{\theta_j}^{-1} \boldsymbol{G}_{\theta_j}$ is bounded and hence, if (θ_j) is convergent, the limit of the algorithm satisfies (17).

On the other hand, let us consider the function

$$\mathcal{S}(\theta) = \sum_{i=1}^{n+1} h(F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})), \tag{18}$$

where $h(t) = \log t$. It has gradient

$$\mathcal{S}'(\theta) = \sum_{i=1}^{n+1} \frac{\frac{\partial}{\partial(\theta)} F_\theta(x_{(i)}) - \frac{\partial}{\partial(\theta)} F_\theta(x_{(i-1)})}{F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})},$$

which is the same as

$$\sum_{i=1}^{n+1} \frac{\frac{\partial}{\partial(\theta)} F_\theta(x_{(i)}) - \frac{\partial}{\partial(\theta)} F_\theta(x_{(i-1)})}{F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})} \left[\frac{1}{n+1} - (F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})) \right].$$

Using the form of $\boldsymbol{V}_\theta^{-1}$ we find (again using Lemma 2 in Appendix A) this is the same as

$$\mathcal{S}'(\theta) = \boldsymbol{G}_\theta^T \boldsymbol{V}_\theta^{-1} (\boldsymbol{\mathcal{R}} - \boldsymbol{F}_\theta),$$

i.e. the condition $\mathcal{S}'(\theta) = 0$ is exactly the same as equation (17).

Finally, for arbitrary data, this argument shows the following: if the sequence (θ_j) given by (16) is convergent, then it converges to a stationary point of the function $\theta \mapsto \sum_{i=1}^{n+1} h(F_\theta(x_{(n_i)}) - F_\theta(x_{(n_{i-1})}))$, where $h(t) = \log t$. In the case of a unique stationary point, this estimator is same as the generalised spacings estimator with the statistically efficient choice of h. The analysis here may be interpreted as linking the iterative algorithm (16) with the iteratively reweighed least squares interpretation of optimization of the log-probability criterion.

Here, $Q_n(\theta, \theta^\star)$ and the choice of $\boldsymbol{V}_{\boldsymbol{\theta}^\star}$ are algebraically the same functions as described in Case 1 if we intentionally chose fixed t_j to be the same as $x_{(n_j)}$ and behave as if we were in Case 1.

Thus, for both the fixed and empirical quantile choices of partition, we have estimation motivated by minimizing of the norm square of standardized empirical discrepancies between empirical and theoretical distributions, which have the same asymptotic efficiencies as optimization of log-probability criteria motivated by likelihood.

6 Conclusion

In previous work Benšić (2014) showed by simulations that fully standardizing the cumulative distribution produces estimators that are superior to those that minimize the Cramer-Von Mises and Anderson-Darling statistics. Now, as a result of the presented perspective, we make it easy to understand that this means advocacy of minimum chi-square estimators as superior to estimators based on minimum distance between (unstandardized) cumulative distributions.

We gave here the common framework in which, for both fixed $t_1, \ldots, t_k$ and quantiles $t_i = X_{(n_i)}$, the form of the covariance of $(F_n(t_i) - F(t_i),\ i = 1, \ldots, k)$ assures a simple relationship to chi-square statistics. However, we caution that, when using all the empirical quantiles ($k = n$, $n_i = i$, $t_i = X_{(i)}$), the standardized $(F(X_{(i)}) - \frac{i}{n+1}, i = 1, \ldots, n)$ is not shown to have an effective norm square for estimation, being only 50% efficient, when the standardization is based on the covariance at the true parameter value. A modified chi-square like formulation is given for the empirical quantiles that is fully efficient.

As noted in Section 4, the fully standardized cumulative distribution statistic $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$ is asymptotically $\mathcal{N}(0, \boldsymbol{I})$. Thus the asymptotic distribution of $\boldsymbol{Z}$ does not depend on the hypothesized distribution F (that is, it is asymptotically distribution free), unlike the vector of $k+1$ components $\sqrt{n}(P_n(A) - P(A))/\sqrt{P(A)}$ whose (asymptotic) distribution depends in particular on the vector of components $\sqrt{P(A)}$ to which it is orthogonal. As $\boldsymbol{Z}$ is asymptotically distribution-free, it is akin to the test statistic components studied in Khmaladze (2013). A difference is that there the objective was to provide a class of such asymptotically distribution-free statistics for discrete settings whereas our objective is to clarify understanding of the fully standardized cumulative distribution for improved efficiency of estimation.

Acknowledgments. We are very grateful to anonymous reviewers for providing many excellent comments, which enhanced the quality of this article.

Appendix A

Lemma 1 *Let k be a fixed integer and $\boldsymbol{\rho}$ be vector in $\mathbb{R}^{k+1}$, such that $\rho_i > 0$ for $1 \le i \le k+1$ and $\sum_{i=1}^{k+1} \rho_i = 1$. Let $\boldsymbol{\mathcal{R}}$ be the vector in $\mathbb{R}^k$ with entries $\mathcal{R}_j = \sum_{i=1}^{j} \rho_i$, $j = 1, \ldots, k$. Let $\boldsymbol{V}$ be the symmetric $k \times k$ matrix with entries*

$$V_{jl} = \mathcal{R}_j(1 - \mathcal{R}_l), \quad j \le l.$$

Then $\boldsymbol{V}^{-1}$ has the tridiagonal form

$$\boldsymbol{V}^{-1} = \begin{bmatrix} \frac{1}{\rho_1} + \frac{1}{\rho_2} & -\frac{1}{\rho_2} & 0 & \cdots & 0 & 0 \\ -\frac{1}{\rho_2} & \frac{1}{\rho_2} + \frac{1}{\rho_3} & -\frac{1}{\rho_3} & \cdots & 0 & 0 \\ 0 & -\frac{1}{\rho_3} & \frac{1}{\rho_3} + \frac{1}{\rho_4} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\rho_{k-1}} + \frac{1}{\rho_k} & -\frac{1}{\rho_k} \\ 0 & 0 & 0 & \cdots & -\frac{1}{\rho_k} & \frac{1}{\rho_k} + \frac{1}{\rho_{k+1}} \end{bmatrix}.$$

Proof. Firstly, let us show that $\boldsymbol{V}\boldsymbol{V}^{-1} = \boldsymbol{V}^{-1}\boldsymbol{V} = \boldsymbol{I}$. Since $\boldsymbol{V}$ is symmetric, it is enough to show $\boldsymbol{V}\boldsymbol{V}^{-1} = \boldsymbol{I}$. In order to do this, note that for $j = 1, \ldots, k$ we have

$$\begin{aligned} \left(\boldsymbol{V}\boldsymbol{V}^{-1}\right)_{j,j} &= \sum_{s=1}^{k} V_{j,s} V^{-1}_{s,j} \\ &= -\mathcal{R}_{j-1}(1 - \mathcal{R}_{j-1})\frac{1}{\rho_j} + \mathcal{R}_j(1 - \mathcal{R}_j)\left(\frac{1}{\rho_j} + \frac{1}{\rho_{j+1}}\right) - \\ &- \mathcal{R}_j(1 - \mathcal{R}_{j+1})\frac{1}{\rho_{j+1}} = 1, \end{aligned}$$

where $\mathcal{R}_0 = 0$ and $\mathcal{R}_{k+1} = 1$. Similarly, for $1 \le j < l \le k$, we have

$$\begin{aligned}\left(\boldsymbol{V}\boldsymbol{V}^{-1}\right)_{j,l} &= \sum_{s=1}^{k} V_{j,s} V^{-1}_{s,l} \\ &= -\mathcal{R}_{l-1}(1-\mathcal{R}_l)\frac{1}{\rho_l} + \mathcal{R}_{l-1}(1-\mathcal{R}_{l-1})\left(\frac{1}{\rho_l} + \frac{1}{\rho_{l+1}}\right) - \\ &- \mathcal{R}_l(1-\mathcal{R}_{l+1})\frac{1}{\rho_{l+1}} = 0.\end{aligned}$$

□

Lemma 2 *Let k be a fixed integer and $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{\rho}$ be vectors in $\mathbb{R}^{k+1}$, with entries a_j, b_j and ρ_j, $j = 1, \ldots, k+1$ respectively, such that:*

1. $\sum_{i=1}^{k+1} a_i = 0$,
2. $\sum_{i=1}^{k+1} b_i = 0$,
3. $\rho_i > 0$ *for* $1 \le i \le k+1$ *and* $\sum_{i=1}^{k+1} \rho_i = 1$.

Let $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{\mathcal{R}}$ be the vectors in $\mathbb{R}^k$ with entries

$$A_j = \sum_{i=1}^{j} a_i, \quad B_j = \sum_{i=1}^{j} b_i, \quad \mathcal{R}_j = \sum_{i=1}^{j} \rho_i, \quad j = 1, \ldots, k.$$

Let $\boldsymbol{V}$ be the symmetric $k \times k$ matrix with entries

$$V_{jl} = \mathcal{R}_j(1-\mathcal{R}_l), \quad j \le l.$$

Then

$$\boldsymbol{A}^T\boldsymbol{V}^{-1}\boldsymbol{B} = \sum_{i=1}^{k+1} \frac{a_i b_i}{\rho_i}. \tag{19}$$

Proof. Similar as in Section 3, the proof can be done by matrix manipulation as well as from the geometrical perspective of orthogonal projection. Here we show the second approach.

Let us denote

$\boldsymbol{\alpha} = [a_1/\sqrt{\rho_1}, \ldots, a_{k+1}/\sqrt{\rho_{k+1}}]^T$,
$\boldsymbol{\beta} = [b_1/\sqrt{\rho_1}, \ldots, b_{k+1}/\sqrt{\rho_{k+1}}]^T$ and
$\boldsymbol{u} = [\sqrt{\rho_1}, \ldots, \sqrt{\rho_{k+1}}]^T$.

These have $\boldsymbol{\alpha}^T\boldsymbol{u} = \sum_i (a_i/\sqrt{\rho_i})\sqrt{\rho_i} = \sum_i a_i = 0$ and likewise $\boldsymbol{\beta}^T\boldsymbol{u} = \sum_i \beta_i = 0$ so they are orthogonal to $\boldsymbol{u}$. Accordingly $\boldsymbol{\alpha}^T\boldsymbol{\beta} = \sum_{j=1}^{k} \tilde{\alpha}_j\tilde{\beta}_j$ where $\tilde{\alpha}_j = \boldsymbol{\alpha}^T\boldsymbol{q}_j$ and $\tilde{\beta}_j = \boldsymbol{\beta}^T\boldsymbol{q}_j$ where $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k$ are orthonormal vectors in $\mathbb{R}^{k+1}$, orthogonal to $\boldsymbol{u}$. Using the choice of these $\boldsymbol{q}_j$ as in Section 4 we find

$$\tilde{\alpha}_j = \frac{-A_j + \alpha_{j+1}\mathcal{R}_j/\rho_{j+1}}{\sqrt{\mathcal{R}_j\mathcal{R}_{j+1}/\rho_{j+1}}}$$

and

$$\tilde{\beta}_j = \frac{-B_j + \beta_{j+1}\mathcal{R}_j/\rho_{j+1}}{\sqrt{\mathcal{R}_j\mathcal{R}_{j+1}/\rho_{j+1}}}.$$

Now, using $\alpha_{j+1} = A_{j+1} - A_j$ for $j < K$ and $\alpha_{k+1} = 0 - A_k$, we see that $\tilde{\boldsymbol{\alpha}} = \boldsymbol{W}^T\boldsymbol{A}$ and $\tilde{\boldsymbol{\beta}} = \boldsymbol{W}^T\boldsymbol{B}$ with bidiagonal $\boldsymbol{W}$. The result then follows upon confirming $\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{V}^{-1}$.

In order to show that $\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{V}^{-1}$, let us point out that, for $j = 1, \ldots, k$,

$$\left(\boldsymbol{W}^T\boldsymbol{W}\right)_{j,j} = \frac{\mathcal{R}_{j-1}^2}{\mathcal{R}_{j-1}\mathcal{R}_j\rho_j} + \frac{\mathcal{R}_{j+1}^2}{\mathcal{R}_j\mathcal{R}_{j+1}\rho_{j+1}} = \frac{1}{\rho_j} + \frac{1}{\rho_{j+1}},$$

for $j = 1, \ldots, k-1$ we have

$$\left(\boldsymbol{W}^T\boldsymbol{W}\right)_{j,j+1} = -\frac{\mathcal{R}_{j+1}\mathcal{R}_j}{\sqrt{\mathcal{R}_j\mathcal{R}_{j+1}\rho_{j+1}}\sqrt{\mathcal{R}_j\mathcal{R}_{j+1}\rho_{j+1}}} = -\frac{1}{\rho_{j+1}},$$

and, finally, for $l \geq j+2$, $\left(\boldsymbol{W}^T\boldsymbol{W}\right)_{j,l} = 0$. Since both of matrices $\boldsymbol{W}^T\boldsymbol{W}$ and $\boldsymbol{V}^{-1}$ are symmetric, the identity $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{V}^{-1}$ holds.

□

Appendix B

Here we discuss the Fisher information result $I^{\mathcal{T}}(\boldsymbol{\theta}) \leq I(\boldsymbol{\theta})$ for the partition formed by any $\mathcal{T}$. This can be seen as a consequences of the general chain rule of the Fisher information $I_{X,Y}(\boldsymbol{\theta}) = I_Y(\boldsymbol{\theta}) + I_{X|Y}(\boldsymbol{\theta})$ by specializing to the case that $Y = g(X)$ is a function of X. Indeed, then $I_Y(\boldsymbol{\theta}) \leq I_{X,Y}(\boldsymbol{\theta}) = I_X(\boldsymbol{\theta})$. In our case, where $\mathcal{T} = \{t_1, \ldots, t_k\}$ with $t_1 < t_2 < \cdots < t_k$, the function g is given by $g(x) = j$ for $t_{j-1} < x \leq t_j$. This function provides the membership label of x in the partition formed by $\mathcal{T}$. It is recognized that, to handle this case, one needs $I_{X,Y}(\boldsymbol{\theta})$ for jointly distributed X, Y even when X is continuous and Y is discrete. The inequality $I^{\mathcal{T}}(\boldsymbol{\theta}) \leq I(\boldsymbol{\theta})$ is seen to hold for any partition $\mathcal{T}$, including the case that $\mathcal{T}$ is based on a data set (via empirical quantiles).

Concerning the chain rule of Fisher information in the twice differentiable case, it is an immediate consequence of the factorization $f_{\boldsymbol{\theta}}(x, y) = f_{\boldsymbol{\theta}}(y)f_{\boldsymbol{\theta}}(x|y)$ by taking expectation of

$$\frac{\partial^2}{\partial\theta^2}\log f_{\boldsymbol{\theta}}(x, y) = \frac{\partial^2}{\partial\theta^2}\log f_{\boldsymbol{\theta}}(y) + \frac{\partial^2}{\partial\theta^2}\log f_{\boldsymbol{\theta}}(x|y),$$

taking advantage of representation of the respective Fisher informations $I_{X,Y}(\boldsymbol{\theta})$, $I_Y(\boldsymbol{\theta})$ and $I_{X|Y}(\boldsymbol{\theta})$ as minus the expected values of the terms in this identity.

Alternatively, in the squared first derivative representation, the chain rule is seen as the Pythagorean identity associated with the L_2 projection property

$$\frac{\partial}{\partial\theta}\log f_{\boldsymbol{\theta}}(y) = E_{\boldsymbol{\theta}}[\frac{\partial}{\partial\theta}\log f_{\boldsymbol{\theta}}(X, Y)|Y = y]. \tag{20}$$

Indeed, the right side is

$$\int \frac{f_{\boldsymbol{\theta}}(x,y)}{f_{\boldsymbol{\theta}}(y)} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(x,y)\right] dx = \frac{1}{f_{\boldsymbol{\theta}}(y)} \int \frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x,y)\, dx = \frac{\frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(y)}{f_{\boldsymbol{\theta}}(y)}$$

provided the derivative can be exchanged with the integral as indicated. Then the chain rule is the expected value in expansion of the square of

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(x,y) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(y) + \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(x|y),$$

as in Zamir (1998). Thus $I_{X,Y}(\boldsymbol{\theta}) = I_Y(\boldsymbol{\theta}) + I_{X|Y}(\boldsymbol{\theta})$ and hence $I_Y(\boldsymbol{\theta}) \leq I_{X,Y}(\boldsymbol{\theta})$. When Y is a function of X we have $I_{X,Y}(\boldsymbol{\theta}) = I_X(\boldsymbol{\theta})$ and hence one has the "data processing" inequality $I_Y(\boldsymbol{\theta}) \leq I_X(\boldsymbol{\theta})$, as claimed. It is also Jensen's inequality applied to (20), as in (Ibragimov and Has'minskii, 1981, Theorem I.7.2).

References

Amemiya, T. (1976). The maximum likelihood, the minimum chi- square and the nonlinear weighted least-squares estimator in the general qualitative response model, *Journal of the American Statistical Association 71*, 347–351.

Anderson, T. W. and D. A. Darling (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes, *The Annals of Mathematical Statistics, 23,* 193–212.

Bar-On I., B. Codenotti, M. Leoncini (1997). A fast parallel Cholesky decomposition algorithm for tridiagonal symmetric matrices, *SIAM. J. Matrix Anal. and Appl., 18,* 403—418.

Barrett, W. W., P. J. Feinsilver (1978). Gaussian families and a theorem on patterned matrices,*J. Appl. Prob. 15*, 514-522.

Barrett, W. W. (1979). A theorem on inverses of tridiagonal matrices,*Linear Algebra and its Applications 27*, 211—217.

Bdair, O. M. (2012) Different methods of estimation for Marshall Olkin exponential distribution, *Journal of Applied Statistical Science 19* 13—29

Benšić, M. (2014). Fitting distribution to data by a generalized nonlinear least squares method, *Communications in Statistics — Simulation and Computation 43*, 687–705.

Benšić, M. (2015). Properties of the generalized nonlinear least squares method applied for fitting distribution to data, *Discussiones Mathematicae Probability and Statistics 35*, 75–94.

Berkson, J. (1949). Minimum χ^2 and maximum likelihood solution in terms of a linear transform, with particular reference to bio-assay, *Journal of the American Statistical Association, 44*, 273–278.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood!, *The Annals of Statistics 8*, 457–487.

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data, *Journal of the American Statistical Association 61*, 228–235.

Cheng, R. C. H. and N. A. K. Amin (1983). Estimating parameters in continuous univariate distributions with a shifted origin, *J. Roy. Statisti. Soc. Ser. B 45*, 394–403.

Choulakian, V., R. A. Lockhart, and M. A. Stephens (1994). Cramer-von Mises statistics for discrete distributions, *Canad. J. Statist. 22*, 125—137.

Cramer, H (1946). *Mathematical Methods of Statistics*, Princeton University Press.

Dey, S. (2014) Two-parameter Rayleigh distribution: Different methods of estimation, *American Journal of Mathematical and Management Sciences 33* 55—74.

Fisher, R. A. (1924). The conditions under which χ^2 measures the discrepancy between observation and hypothesis, *Journal of the Royal Statistical Society 87*, 442–450.

Ghosh, K. and S. Rao Jammalamadaka (2001). A general estimation method using spacings, *Jouranal of Statistical Planing and Inferences, 93*, 71–82.

Harris, D. and L. Matyas (1999). Introduction to the generalized method of moment estimation, *in: Matyas, L. (Ed.), Generalized Method of Moment Estimation, Cambridge University Press, Cambridge* 3–30.

Harris, R. R. and G. K. Kanji (1983). On the use of minimum chi-square estimation, *The Statistician 23*, 379–394.

Hartley, H. O., R. C. Pfaffenberger (1972). Quadratic forms in order statistics used as goodnes-of-fit criteria, *Biometrika 59*, 605—611.

Helmert, F. R. (1986). Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Beobachtungsfehlers directer Beobachtungen gleicher Genauigkeit, *Astronomische Nachrichten 88*, columns 113–132.

Henze, N. (1996). Empirical-distribution-function goodness-of-fit tests for discrete models, *Canad. J. Statist. 24*, 81–93.

Hsiao, C. (2006). Minimum chi-square, *Encyclopedia of Statistical Scineces 7*, John Wiley & Sons.

Ibragimov, I. A., R. Z. Has'minskii, (1981). *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.

Irwin, J. O. (1942). On the distribution of a weighted estimate of variance and on analysis of variance in certain cases of unequal weighting, *Journal of the Royal Statistical Society 105*, 115–118.

Irwin, J. O. (1949). A note on the subdivision of χ^2 into components, *Biometrika 36*, 130-134.

Johnson, N. L., S. Kotz, N. Balakrishnan (1994) *Continuous Univariate Distributions, Vol 1.*, J. Wiley & Sons, Inc., New York-Singapore

Kantar, Y. M. (2015) Generalized least squares and weighted least squares estimation methods for distributional parameters, *REVSTAT - Statistical Journal 13*, 263—285

Kundu D., M. Z. Raqab (2005) Generalized Rayleigh distribution: different methods of estimations, *Computational Statistics & Data Analysis 49*, 187–

200.

Khmaladze, E. (2013). Note on distribution free testing for discrete distributions, Ann. Statist. 41, 2979–2993.

Kruskal, W. H. (1946), Helmert's distribution, *American Mathematical Monthly 53,* 435–438.

Lancaster, H. O. (1949). The derivation and partition of χ^2 in certain discrete distributions, *Biometrika 36*, 117–129.

Lancaster, H. O. (1965). The Helmert matrices, *The American Mathematical Monthly 72,* 4–12.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing, *in Engle, R. and D. McFadden, eds., Handbook of Econometrics, Vol. 4, New York: North Holland.*

Neyman, J. (1949). Contribution to the theory of the χ^2 test, *In Proc. Berkeley Symp. Math. Stat. Prob.,* 239–273.

Ranneby, B. (1984). The maximum spacing method: an estimation method related to the maximum likelihood method, *Scand. J. Statist. 11,* 93–112.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications (2nd ed.)*, New York: John Wiley.

Rinne, H. (2009). *The Weibull Distribution. A Handbook*, Taylor & Francis Group.

Shorack, G. R., J. A. Wellner (2009). *Empirical Processes with Applications to Statistics*, SIAM, Philadelphija.

Stigler, S. M. (1984). Kruskal's proof of the joint distribution of $\bar{X}$ and s^2, *The American Statistician 38*, 134—135.

Swain, J., S. Venkatraman, J. Wilson (1988). Least-squares estimation of distribution function in Johnson's translation system. *J. Statist. Comput. Simulation 29*, 271—279.

Taylor, W. F. (1953). Distance functions and regular best asymptotically normal estimates, *The Annals of Mathematical Statistics 24*, 85–92.

Torres, F. J. (2014). Estimation of parameters of the shifted Gompertz distribution using least squares, maximum likelihood and moment methods. *Journal of Computational and Applied Mathematics 255*, 867—877.

van der Vaart, A. W., J. A. Wellner (1996). *Weak Convergence and Empirical Processes With Applications to Statistics*, Springer-Verlag, New York.

Zamir, R. (1998). A proof of the Fisher information inequality via a data processing argument. *IEEE Transaction on Information Theory 44*, 1246—1250.