Standardizing the Empirical Distribution Function Yields the Chi-Square Statistic

Andrew Barron
Department of Statistics, Yale University
Mirta Benšić
Department of Mathematics, University of Osijek
and
Kristian Sabo
Department of Mathematics, University of Osijek

July 1, 2016

Abstract

Standardizing the empirical distribution function yields a statistic with norm square that matches the chi-square test statistic. To show this one may use the covariance matrix of the empirical distribution which, at any finite set of points, is shown to have an inverse which is tridiagonal. Moreover, a representation of the inverse is given which is a product of bidiagonal matrices corresponding to a representation of the standardization of the empirical distribution via a linear combination of values at two consecutive points. These properties are discussed also in the context of minimum distance estimation.

Keywords: Generalized least squares, minimum distance estimation, covariance of empirical distribution, covariance of quantiles, bidiagonal Cholesky factors, Helmert matrix.

1 Introduction

Let $X_1, X_2, \ldots, X_n$ be independent real–valued random variables with distribution function F. Let F_n be the empirical distribution function $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \le t\}}$ and let

$$\sqrt{n}(F_n(t) - F(t)), \quad t \in \mathcal{T},$$

be the centered empirical process evaluated at a set of points $\mathcal{T} \subset \mathbb{R}$. It is familiar that when F is an hypothesized distribution and $\mathcal{T} = \mathbb{R}$ the maximum of the absolute value of this empirical process corresponds to the Kolmogorov–Sminov test statistic, the average square corresponds to the Cramer-Von Mises test statistic and the average square with marginal standardization using the variance equal to $F(t)(1 - F(t))$ produces the Anderson-Darling statistics (average with the distribution F) (see Anderson (1952)). The covariance of the empirical process takes the form $\frac{1}{n}F(t)(1 - F(s))$ for $t \le s$. 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 $\sqrt{n}(F_n(t) - F(t))$, $t \in \mathcal{T}$. 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 $\boldsymbol{V}$. Complete standardization of the empirical distribution restricted to $\mathcal{T}$ has been put forward in Benšić (2014) leading to the distance

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

which is there analysed as a generalized least squares criterion for minimum distance parameter estimation. It fits also in the framework of the generalized method of moments (Benšić (2015)). The motivation, familiar from regression, is that the complete standardization produces more efficient estimators.

The purpose of the present work is to show statistical simplifications in the generalized least squares criterion. In particular, we show that the expression (1) is precisely equal to the chi-square test 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 formed by consecutive values $\mathcal{T}=\{t_1,\ldots,t_k\}$, 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 $F(\infty)=F_n(\infty)=1$.

Moreover, we show that $\boldsymbol{V}^{-1}$ takes a tridiagonal form with $1/P(A_j)+1/P(A_{j+1})$ for the (j,j) entries on the diagonal; $-1/P(A_j)$, for the $(j-1,j)$ entries and $-1/P(A_{j+1})$ for the $(j+1,j)$ entries and 0 otherwise.

We find an explicit standardization

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

These random variables have mean 0 and variance 1 (with $c_{n,j}=\sqrt{\frac{F(t_j)F(t_{j+1})P(A_{j+1})}{n}}$) and they are uncorrelated for $j=1,\ldots,k$. Moreover, the sum of squares

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

is precisely equal to the statistic given in expressions (1) and (2). It corresponds to a bidiagonal Cholesky decomposition of $\boldsymbol{V}^{-1}$ as $\boldsymbol{B}^{\tau}\boldsymbol{B}$ 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 $\boldsymbol{Z}=\boldsymbol{B}(\boldsymbol{F}_n-\boldsymbol{F})$, where $\boldsymbol{F}=(F(t_1),\ldots,F(t_k))^{\tau}$, as a full standardization of the vector $\boldsymbol{F}_n=(F_n(t_1),\ldots,F_n(t_k))^{\tau}$. The Z_j may also be written as

$$Z_j=\frac{P_n(A_{j+1})F(t_j)-F_n(t_j)P(A_{j+1})}{c_{n,j}} \tag{4}$$

so its marginal distribution (with an hypothesized F) comes from the trinomial distribution of $(nF_n(t_{j-1}),nP_n(A_j))$. These uncorrelated Z_j, though not independent, suggest 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 (4)) that the joint distribution of $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^\tau$ is asymptotically Normal$(0, \boldsymbol{I})$, providing a direct path to the asymptotic chi-square(k) distribution of the statistic given equivalently in (1), (2), (3).

There are natural fixed and random choices for the points $t_1, \ldots, t_k$. A natural fixed choice is to use $k-$quantiles of a reference distribution. If F is an hypothesized continuous distribution , such quantiles can be chosen such that $P(A_j) = F(t_j) - F(t_{j-1}) = 1/(k+1)$.

A natural random choice is to use empirical quantiles $t_j = X_{(n_j)}$ with $1 \le n_1 < n_2 < \ldots < n_k \le n$. If $k+1$ divides n the n_j may equal $jn/(k+1)$, for $j = 1, \ldots, k$. With empirical quantiles it is the $F(t_j) = F(X_{(n_j)})$ that are random, having the same distribution as uniform order statistic with mean $\mathcal{R}_j = n_j/(n+1)$ and covariance $V_{j,l}/(n+1)$ where $V_{j,l} = \mathcal{R}_{t_j}(1 - \mathcal{R}_{t_l})$ for $j \le l$. Once again we find that $(\boldsymbol{F} - \boldsymbol{\mathcal{R}})^\tau \boldsymbol{V}^{-1}(\boldsymbol{F} - \boldsymbol{\mathcal{R}})$, where $\boldsymbol{\mathcal{R}} = (\mathcal{R}_1, \ldots, \mathcal{R}_k)^\tau$, takes the form of a chi-square test statistic and the story is much the same. The Z_j are now multiples of $\mathcal{R}_{j-1}F(X_{(n_j)}) - \mathcal{R}_j F(X_{(n_j-1)})$ which are again mean 0, variance 1 and uncorrelated. The main difference in this case is that their exact distribution comes from the Dirichlet distribution of $(F(X_{(n_j-1)}), (F(X_{(n_j)}) - F(X_{(n_j-1)}))$ rather than from the multinomial.

The form of the $\boldsymbol{V}^{-1}$ with bidiagonal decomposition $\boldsymbol{B}^\tau \boldsymbol{B}$ and the representation of the norm square of the standardized empirical distribution (1) as the chi-square test statistic (2) provides simplified interpretation, simplified computation and simplified statistical analysis. For interpretation, we see that when choosing between tests based on the cumulative distribution (like the Anderson–Darling test) and tests based on counts in disjoint cells, the choice largely depends on whether one wants the benefit of the more complete standardization leading to the chi-square test. For computation, we see that generalized least squares simplifies due to the tridiagonal form.

As for simplification of statistical analysis, consider the case of a parametric family of distribution function F_θ with a real parameter θ. The generalized least squares procedure picks $\hat{\theta} = \hat{\theta}_{k,n}$ to minimize $(\boldsymbol{F}_n - \boldsymbol{F}_\theta)^\tau \boldsymbol{V}^{-1}(\boldsymbol{F}_n - \boldsymbol{F}_\theta)$, where $\boldsymbol{V}$ is the covariance matrix evaluated at a consistent estimator of the true θ_0.

For a fixed set of points $t_1, \ldots, t_k$ it is known (Benšić (2015)) that $\lim_n [n\mathrm{Var}(\hat{\theta}_{k,n})]$ has reciprocal $\boldsymbol{G}^\tau \boldsymbol{V}_{\theta_0} \boldsymbol{G}$ where $\boldsymbol{G}$ is the vector $-\frac{\partial}{\partial\theta}\boldsymbol{F}_\theta$ evaluated at the true parameter value $\theta = \theta_0$. Using the tridiagonal inverse we show that this $\boldsymbol{G}^\tau \boldsymbol{V}_{\theta_0}^{-1} \boldsymbol{G}$ simplifies to

$$\sum_{A\in\pi} P_{\theta_0}(A)(E[S(X)|A]))^2, \tag{5}$$

where $S(X) = \frac{\partial}{\partial\theta}\log f(X|\theta)$ is the score function evaluated at $\theta = \theta_0$ which we interpret as a Riemann–Stieltjes discretization of the Fisher information $E[S^2(X)]$. This Fisher information arises in the limit of large k.

2 Common Framework

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 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

$$\mathrm{Cov}(R_j, R_l) = \frac{1}{c}\mathcal{R}_j(1-\mathcal{R}_l) = \frac{1}{c}V_{j,l} \tag{6}$$

for $j \leq l$. Let $\boldsymbol{R} = (R_1, \dots, R_k)^\tau$ and $\boldsymbol{\mathcal{R}} = (\mathcal{R}_1, \dots, \mathcal{R}_k)^\tau$. We explore the relationship between $(\boldsymbol{R} - \boldsymbol{\mathcal{R}})^\tau \boldsymbol{V}^{-1}(\boldsymbol{R} - \boldsymbol{\mathcal{R}})$ and $\sum_{j=1}^{k+1} \frac{(r_j - \rho_j)^2}{\rho_j}$ and the structure of the inverse $\boldsymbol{V}^{-1}$ as well as construction of a version of $\boldsymbol{B}(\boldsymbol{R} - \boldsymbol{\mathcal{R}})$ with uncorrelated entries, and $\boldsymbol{V}^{-1} = \boldsymbol{B}^\tau \boldsymbol{B}$.

We have the following cases for $X_1, \dots, X_n$ 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 have increments $r_j = P_n(A_j) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A_j\}}$ and $\rho_j = P(A_j) = F(t_j) - F(t_{j-1})$ with intervals $A_j = (t_{j-1}, t_j]$. Now the covariance is $1/n$ times the covariance in a single draw, so (6) holds with $c = n$.

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

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

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 (6) holds for $c = n + 2$.

Note that in both cases we examine distribution properties of $R_j - \mathcal{R}_j$ which 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 $t_1, \dots, t_k$ and in Case 2 it is the centering of the hypothesized distribution function evaluated at the quantiles $X_{(n_1)}, X_{(n_2)}, \dots, X_{(n_k)}$.

3 Tridiagonal V^{-1} and its bidiagonal decomposition

We have two approaches to appreciating the relationship between the standardized cumulative distribution and the chi-square statistic. In this section, we use elementary matrix calculations to show that the inverse of the matrix $\boldsymbol{V}$ has a special tridiagonal structure, to derive its bidiagonal decomposition and to obtain the following identity:

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

Whereas in the next section we revisit the matter from the geometrical perspective of orthogonal projection.

Lemma 1. *If*

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

then

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

Moreover, $\boldsymbol{V}^{-1} = \boldsymbol{B}^\tau \boldsymbol{B}$, *where*

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

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 $1 \leq j < l \leq k$, we have

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

It remains to show $\boldsymbol{B}^\tau\boldsymbol{B} = \boldsymbol{V}^{-1}$. For $j = 1, \ldots, k$ it follows

$$(\boldsymbol{B}^\tau\boldsymbol{B})_{j,j} = \frac{\mathcal{R}_{j-1}^2}{\mathcal{R}_{j-1}\mathcal{R}_j\rho_j} + \frac{\mathcal{R}_{j+1}^2}{\mathcal{R}_j\mathcal{R}_{j+1}\rho_{j+1}} = \frac{1}{\rho_j} + \frac{1}{\rho_{j+1}}.$$

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

$$(\boldsymbol{B}^\tau\boldsymbol{B})_{j,j+1} = -\frac{\mathcal{R}_{j+1}\mathcal{R}_j}{\sqrt{\mathcal{R}_j\mathcal{R}_{j+1}\rho_{j+1}}\sqrt{\mathcal{R}_j\mathcal{R}_{j+1}\rho_{j+1}}} = -\frac{1}{\rho_{j+1}}.$$

Finally, for $l \geq j+2$, it follows $(\boldsymbol{B}^\tau \boldsymbol{B})_{j,l} = 0$. Since both of matrices $\boldsymbol{B}^\tau \boldsymbol{B}$ and $\boldsymbol{V}^{-1}$ are symmetric, the identity $\boldsymbol{B}^\tau \boldsymbol{B} = \boldsymbol{V}^{-1}$ holds.

$\square$

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

Proof. Note that $\mathbf{V}^{-1}$ can be written in the following way:

$$\mathbf{V}^{-1} = \begin{bmatrix} \frac{1}{\mathcal{R}_1} + \frac{1}{\mathcal{R}_2 - \mathcal{R}_1} & -\frac{1}{\mathcal{R}_2 - \mathcal{R}_1} & 0 & \cdots & 0 & 0 \\ -\frac{1}{\mathcal{R}_2 - \mathcal{R}_1} & \frac{1}{\mathcal{R}_2 - \mathcal{R}_1} + \frac{1}{\mathcal{R}_3 - \mathcal{R}_2} & -\frac{1}{\mathcal{R}_3 - \mathcal{R}_2} & \cdots & 0 & 0 \\ 0 & -\frac{1}{\mathcal{R}_3 - \mathcal{R}_2} & \frac{1}{\mathcal{R}_3 - \mathcal{R}_2} + \frac{1}{\mathcal{R}_4 - \mathcal{R}_3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\mathcal{R}_{k-1} - \mathcal{R}_{k-2}} + \frac{1}{\mathcal{R}_k - \mathcal{R}_{k-1}} & -\frac{1}{\mathcal{R}_k - \mathcal{R}_{k-1}} \\ 0 & 0 & 0 & \cdots & -\frac{1}{\mathcal{R}_k - \mathcal{R}_{k-1}} & \frac{1}{1 - \mathcal{R}_k} + \frac{1}{\mathcal{R}_k - \mathcal{R}_{k-1}} \end{bmatrix}$$

Consequently, we obtain

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

$\square$

4 Projection properties

There is, of course, 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

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

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 $\boldsymbol{C}_{i,j} = \rho_i \mathbf{1}_{\{i=j\}} - \rho_i\rho_j$. These forms use the vectors of length k, because the value $r_{k+1} = 1 - \sum_{j=1}^{k} r_j$ is linearly determined from the others. It is known (and easily checked) that the matrix $\boldsymbol{C}^{-1}$ has entries $(\boldsymbol{C}^{-1})_{i,j} = \frac{1}{\rho_i}\mathbf{1}_{\{i=j\}} - \frac{1}{\rho_{k+1}}$ for $i, j = 1, 2, \ldots, k$ (matching the Fisher information of the multinomial) and one finds from this form that $(\boldsymbol{r}-\boldsymbol{\rho})^\tau \boldsymbol{C}^{-1}(\boldsymbol{r}-\boldsymbol{\rho})$ 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 the equivalence of expressions (1) and (2). Furthermore, using suitable orthogonal vectors one can see how the chi-square statistic (2) arises as the norm square of the fully standardised cumulative distributions (3).

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 this unit vector has length $\boldsymbol{\xi}^\tau \boldsymbol{u} = \sum_{j=1}^{k+1} \left(\frac{r_j}{\sqrt{\rho_j}}\right)\sqrt{\rho_j}$ equal to 1. Accordingly, if $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_k$ and $\boldsymbol{q}_{k+1} = \boldsymbol{u}$ are orthonormal vectors, then the chi-square value 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 $Z_j = \boldsymbol{\xi}^\tau \boldsymbol{q}_j$, $j = 1, 2, \ldots, k$, or equivalently $Z_j = (\boldsymbol{\xi}-\boldsymbol{u})^\tau \boldsymbol{q}_j$.

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 $E(\boldsymbol{\xi}-\boldsymbol{u})(\boldsymbol{\xi}-\boldsymbol{u})^\tau$ reduces to $\frac{1}{c_n}[\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\tau]$ 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

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

Accordingly, $EZ_jZ_l = E\boldsymbol{q}_j^\tau(\boldsymbol{\xi} - \boldsymbol{u})(\boldsymbol{\xi} - \boldsymbol{u})^\tau \boldsymbol{q}_l = \frac{1}{c_n}\boldsymbol{q}_j^\tau(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\tau)\boldsymbol{q}_l$ is $\frac{1}{c_n}\boldsymbol{q}_j^\tau \boldsymbol{q}_l$ 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 variables (c.f. Cramer (1946), pages 416-420).

5 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 $\sum_{j=2}^{n}(Y_j - \bar{Y}_{j-1})^2 \frac{j-1}{j}$ of the independent $\mathcal{N}(0, \sigma^2)$ innovations (also known as standardized prediction errors) $Z_j = \frac{Y_j - \bar{Y}_{j-1}}{\sqrt{1+\frac{1}{j-1}}}$ and, accordingly, this sum of squares is explicitly σ^2 times a chi-square 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 $Z_j = \boldsymbol{\xi}^\tau \boldsymbol{q}_j$, where the $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k, \boldsymbol{q}_{k+1}$ are the normalization of the following orthogonal vectors in $\mathbb{R}^{k+1}$

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

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 $j = 1, \ldots, k$,

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

and

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

becomes

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

which 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, orthogonal, and of constant variance $1/c_n$. Each of these facts can be checked directly using $ER_j = \mathcal{R}_j$ and using the specified form of the covariance $\mathrm{Cov}(R_j, R_l) = \frac{1}{c_n}[\min(\mathcal{R}_j, \mathcal{R}_l) - \mathcal{R}_j\mathcal{R}_l]$.

To summarize this section we have presented uncorrelated components Z_j of the chi-square statistic. Moreover, we have provided interpretation of these components as innovations standardizing the cumulative distribution values. It provides specific proof of the equivalence of expressions (1), (2) and (3).

6 Large sample estimation properties

The results of previous sections will be used to discuss asymptotic efficiency of certain minimum distance estimators related to Case 1 and Case 2 of Section 2.

Let us 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 $P_{\boldsymbol{\theta}_0}$, we may use the generalized least squares procedure for parameter estimation so that we minimize $Q_n(\boldsymbol{\theta}) = (\boldsymbol{R}_n - \boldsymbol{\mathcal{R}}_n)^\tau(\boldsymbol{\theta})\boldsymbol{V}^{-1}(\boldsymbol{R}_n - \boldsymbol{\mathcal{R}}_n)(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$. Here $\boldsymbol{V}$ is the covariance matrix of $\boldsymbol{R}_n - \boldsymbol{\mathcal{R}}_n$ evaluated at a true value $\boldsymbol{\theta}_0$ or at a consistent estimator of the true value.

Both cases from Section 2, i.e. fixed and random $t_1, \ldots, t_k$ considered in the estimation context, fulfill this requirement. Indeed, for Case 1 (fixed $t_1 < \cdots < t_k$) we have

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

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

Here only the expectation $\boldsymbol{\mathcal{R}}_n(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$. For Case 2 (random $t_1 < \cdots < t_k$, $t_j = X_{(n_j)}$) we have

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

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

Now only the $\boldsymbol{R}_n(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$. Here $F_{\boldsymbol{\theta}_0}(X_{(n_j)})$ has a Beta$(n_j, n+1-n_j)$ distribution, $E_{\boldsymbol{\theta}_0}[F_{\boldsymbol{\theta}_0}(X_{(n_j)})] = \frac{n_j}{n+1} = F_n(X_{(n_j)})\frac{n}{n+1}$ so that

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

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.

However, there is a difference in the analysis of estimation properties for the two mentioned cases. Let us discuss them separately.

Case 1. For the fixed $t_1 < \cdots < t_k$. We can express the functional $Q_{n,\boldsymbol{V}}(\boldsymbol{\theta})$ as

$$Q_{n,\boldsymbol{V}}(\boldsymbol{\theta}) = [F_n(t_1) - F_{\boldsymbol{\theta}}(t_1), \ldots, F_n(t_k) - F_{\boldsymbol{\theta}}(t_k)]^\tau \boldsymbol{V}^{-1}[F_n(t_1) - F_{\boldsymbol{\theta}}(t_1), \ldots, F_n(t_k) - F_{\boldsymbol{\theta}}(t_k)].$$

For the complete standardization we should use the matrix

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \mathrm{Var}_{\boldsymbol{\theta}}[F_n(t_1) - F_{\boldsymbol{\theta}}(t_1), \ldots, F_n(t_k) - F_{\boldsymbol{\theta}}(t_k)]$$

that depends on the parameter. Results derived in Sections 3 and 4 then guaranty that minimizing the functional $Q_n(\boldsymbol{\theta}) = Q_{n,\boldsymbol{V}_{\boldsymbol{\theta}}}(\boldsymbol{\theta})$ 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). However, this estimation procedure can also be set in the framework of the generalized method of moments (GMM). 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 functional $Q_{n,\boldsymbol{V}}(\boldsymbol{\theta})$, 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 functional $Q_{n,\boldsymbol{V}_{\boldsymbol{\theta}^\star}}(\boldsymbol{\theta})$. Refining the notation from Section 2:

$A_i = (t_{i-1}, t_i]$, $i = 1, \ldots, k$, $A_{k+1} = (t_k, \infty)$,

$P_n(A_i) = F_n(t_i) - F_n(t_{i-1})$,

$P^\star(A_i) = F_{\boldsymbol{\theta}^\star}(t_i) - F_{\boldsymbol{\theta}^\star}(t_{i-1})$,

$P(A_i; \boldsymbol{\theta}) = F_{\boldsymbol{\theta}}(t_i) - F_{\boldsymbol{\theta}}(t_{i-1})$

$P_0(A_i) = F_{\boldsymbol{\theta}_0}(t_i) - F_{\boldsymbol{\theta}_0}(t_{i-1})$,

analogous to the results of the previous sections, we see that

$$\hat{\boldsymbol{\theta}}_{k,n} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{k+1} \frac{(P_n(A_i) - P(A_i; \boldsymbol{\theta}))^2}{P^\star(A_i)}. \tag{7}$$

If classical 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 $\boldsymbol{G}_0^\tau \boldsymbol{V}_0^{-1} \boldsymbol{G}_0$ where $\boldsymbol{G}_0$ and $\boldsymbol{V}_0$ are, respectively, the matrices

$\frac{\partial}{\partial \boldsymbol{\theta}^\tau}[F_{\boldsymbol{\theta}}(t_1), \ldots, F_{\boldsymbol{\theta}}(t_k)]^\tau$ and the covariance matrix of $[F_n(t_1), \ldots, F_n(t_k)]^\tau$ evaluated at the true parameter value $\boldsymbol{\theta}_0$.

Using the tridiagonal form for the inverse of $\boldsymbol{V}_0$ we can simplify this limit. Indeed, if the model is regular, 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

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

and

$$\begin{aligned} \boldsymbol{G}_0^\tau \boldsymbol{V}_0^{-1} \boldsymbol{G}_0 &= \sum_{i=1}^{k+1} \frac{1}{P_0(A_{i-1})} \int_{t_{i-1}}^{t_i} \boldsymbol{S}(x) f(x;\boldsymbol{\theta}_0)\, dx \int_{t_{i-1}}^{t_i} \boldsymbol{S}^\tau(x) f(x;\boldsymbol{\theta}_0)\, dx \\ &= \sum_{i=1}^{k+1} P_0(A_{i-1}) \frac{\int_{t_{i-1}}^{t_i} \boldsymbol{S}(x) f(x;\boldsymbol{\theta}_0)\, dx}{P_0(A_{i-1})} \frac{\int_{t_{i-1}}^{t_i} \boldsymbol{S}^\tau(x) f(x;\boldsymbol{\theta}_0)\, dx}{P_0(A_{i-1})} \\ &= \sum_{i=1}^{k+1} P_0(A_{i-1}) E_{\boldsymbol{\theta}_0}[\boldsymbol{S}(X)|A_{i-1}] E_{\boldsymbol{\theta}_0}[\boldsymbol{S}(X)|A_{i-1}]^\tau. \end{aligned}$$

This can be interpreted as a Riemann-Stieltjes discretization of the Fisher information which arises in the limit of large k.

Let us note the similarity of $\hat{\boldsymbol{\theta}}_{k,n}$ and the minimum chi-square estimator. From (7) 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. What was not recognized is that minimizing the norm squared of the fully standardized empirical distribution is in the same as minimizing chi-square.

Case 2. In this case we have $t_j = X_{(n_j)}$ so that the value $F_n(t_j) = n_j/n$ is predetermined. The random part $F_{\boldsymbol{\theta}}(X_{n_j})$ of $Q_n(\boldsymbol{\theta})$ depends on the parameter. But the matrix $\boldsymbol{V}$ (which should be used for complete standardization) does not depend on $\boldsymbol{\theta}$ and can be computed from uniform order statistics covariances. Now, the results from Sections 3 and 4 enable us to represent the minimizer of the functional $Q_k(\boldsymbol{\theta})$ as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{k+1} \frac{((F_{\boldsymbol{\theta}}(X_{(n_i)}) - F_{\boldsymbol{\theta}}(X_{(n_{i-1})})) - \frac{n_i - n_{i-1}}{n+1})^2}{\frac{n_i - n_{i-1}}{n+1}}. \tag{8}$$

To discuss asymptotic properties of this estimator let us suppose that all data are different and $k = n$ so that the estimator can be easily recognized as a generalized spacing estimator (GSE) (see Ghosh and Rao Jammalamadaka (2001), Cheng and Amin (1983), Ranneby (1984)). 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,

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n+1} (F_{\boldsymbol{\theta}}(X_{(n_i)}) - F_{\boldsymbol{\theta}}(X_{(n_{i-1})}))^2 = \sum_{i=1}^{n+1} h(F_{\boldsymbol{\theta}}(X_{(n_i)}) - F_{\boldsymbol{\theta}}(X_{(n_{i-1})})), \tag{9}$$

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 $h(t) = h^2$ 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 GSE with minimum variance for a given class of functions which includes $h(t) = t^2$. It is stated there that 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), Theorem 3.1, from Ghosh and Rao Jammalamadaka (2001) equals 2, which 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 Fisher information. So, for these cases $\hat{\boldsymbol{\theta}}_n$ from (9) 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 (8) in each step:

1. Let
$$Q_n(\theta, \theta') = \sum_{i=1}^{n+1} \frac{(F_\theta(X_{(i)}) - F_\theta(X_{(i-1)}) - \frac{1}{n+1})^2}{F_{\theta'}(X_{(i)}) - F_{\theta'}(X_{(i-1)})}.$$

2. Let $\theta^\star$ be a consistent estimator for real θ.

3.
$$\theta_1 = \theta^\star$$
$$\theta_{j+1} = \operatorname*{argmin}_\theta Q_n(\theta, \theta_j), \quad j = 1, 2, \ldots$$

To show this let us denote $\boldsymbol{F}_\theta = [F_{\boldsymbol{\theta}}(X_{(1)}), \ldots, F_{\boldsymbol{\theta}}(X_{(n)})]^\tau$, $\boldsymbol{e} = [1, \ldots, 1]^\tau \in \mathbb{R}^n$, $\boldsymbol{G}_\theta = \frac{\partial}{\partial \boldsymbol{\theta}}[F_{\boldsymbol{\theta}}(X_{(1)}), \ldots, F_{\boldsymbol{\theta}}(X_{(n)})]^\tau$, and

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

As in Gauss-Newton's method for nonlinear least squares, here we consider the following quadratic approximation

$$\theta \mapsto \hat{Q}_n(\theta,\theta_j) = \left(\boldsymbol{F}_{\theta_j} + \boldsymbol{G}_{\theta_j}(\theta-\theta_j) - \frac{1}{n+1}\mathbf{e}\right)^{\tau} \boldsymbol{V}_{\theta_j}^{-1} \left(\boldsymbol{F}_{\theta_j} + \boldsymbol{G}_{\theta_j}(\theta-\theta_j) - \frac{1}{n+1}\mathbf{e}\right)$$

of the function $\theta \mapsto Q_n(\theta,\theta_j) = \left(\boldsymbol{F}_{\theta} - \frac{1}{n+1}\mathbf{e}\right)^{\tau} \boldsymbol{V}_{\theta_j}^{-1} \left(\boldsymbol{F}_{\theta} - \frac{1}{n+1}\mathbf{e}\right)$ about the point θ_j.

Instead of nonlinear optimization problem $\min\limits_{\theta} Q_n(\theta,\theta_j)$, in every iteration we solve simple problem $\min\limits_{\theta} \hat{Q}_n(\theta,\theta_j)$, that has explicit solution. Then the corresponding iterative procedure is given by

$$\theta_{j+1} = \operatorname*{argmin}_{\theta} \left(\boldsymbol{F}_{\theta_j} + \boldsymbol{G}_{\theta_j}(\theta-\theta_j) - \frac{1}{n+1}\mathbf{e}\right)^{\tau} \boldsymbol{V}_{\theta_j}^{-1} \left(\boldsymbol{F}_{\theta_j} + \boldsymbol{G}_{\theta_j}(\theta-\theta_j) - \frac{1}{n+1}\mathbf{e}\right),$$

or explicitly

$$\theta_{j+1} = \theta_j + \left(\boldsymbol{G}_{\theta_j}^{\tau} \boldsymbol{V}_{\theta_j}^{-1} \boldsymbol{G}_{\theta_j}\right)^{-1} \boldsymbol{G}_{\theta_j}^{\tau} \boldsymbol{V}_{\theta_j}^{-1} \left(\frac{1}{n+1}\boldsymbol{e} - \boldsymbol{F}_{\theta_j}\right),\ j = 1,2,\ldots. \tag{10}$$

If we suppose that the sequence (θ_j) is convergent, then

$$\boldsymbol{G}_{\theta_j}^{\tau} \boldsymbol{V}_{\theta_j}^{-1} \left(\frac{1}{n+1}\boldsymbol{e} - \boldsymbol{F}_{\theta_j}\right) \to 0$$

i.e. the limit of the sequence (θ_j) is the solution of the equation

$$\boldsymbol{G}_{\theta}^{\tau} \boldsymbol{V}_{\theta}^{-1} \left(\frac{1}{n+1}\boldsymbol{e} - \boldsymbol{F}_{\theta}\right) = 0. \tag{11}$$

Let us consider the function

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

where $h(t) = \log t$. Note that

$$\mathcal{S}'(\theta) = \boldsymbol{G}_{\theta}^{\tau} \boldsymbol{V}_{\theta}^{-1} \left(\frac{1}{n+1}\boldsymbol{e} - \boldsymbol{F}_{\theta}\right),$$

i.e. the condition $\mathcal{S}'(\theta) = 0$ is exactly the same as equation (11). Finally, we showed the following: if the sequence (θ_j) given by (10) is convergent, then it converges to the 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 $h(t) = \log t$.

Here, $Q_n(\theta, \theta^\star)$ is algebraically the same functional as the one described in Case 1 if we intentionally chose fixed t_j to be the same as $x_{(n_j)}$ and behave as we are in Case 1.

7 Conclusion

In previous work Benšić (2015), has been shown 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 present work we understand that this means advocacy of minimum chi-square estimators as superior to estimators based on minimum distance between (unstandardized) cumulative distributions.

We have seen here that for both fixed $t_1, \ldots, t_k$ and quantiles $t_i = X_{(n_i)}$ the form of the covariance of $(F_n(t_i) - F(t_i),\ i = 1, \ldots, k)$ permits simple standardization and a simple relationship to chi-square statistic. For fixed $t_1, \ldots, t_k$ we also see clearly the chi-square(k) asymptotic distribution. However, we caution that 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. A modified chi-square like formulation is given for the empirical quantiles that is fully efficient.

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.

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.

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

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.

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

Fisher, R. A. (1924). The Conditions Under Which χ^2 Measures the Discrepancey 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.

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.

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

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 Subdividsion of χ^2 into Components, *Biometrika 36*, 130-134.

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.*

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

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