CONSISTENT VARIABLE SELECTION IN LARGE REGRESSION MODELS

ERHARD RESCHENHOFER

Department of Statistics and Operations Research
University of Vienna
Austria
e-mail: erhard.reschenhofer@univie.ac.at

Abstract

In this paper, conditions for the consistent selection of a subset from a large set of potential regressors are derived. It is assumed that the number of potential regressors increases as the sample size increases and, in addition, that the regressors are orthogonal. Subset-selection criteria are proposed which satisfy these conditions. These criteria do not depend on any tuning parameters. It is also shown that some other criteria, which include AIC and BIC, violate these conditions. Simulation studies with different sample sizes and large sets of orthogonal regressors are conducted to compare the performance of the new criteria with that of conventional model-selection criteria. The results of these simulation studies corroborate the theoretical findings. In large samples, the consistent criteria always make the correct decisions. They include all genuine regressors and exclude the others. In contrast, AIC tends to select always the maximum number of regressors and BIC is also not competitive when the number of potential regressors grows too fast.

2010 Mathematics Subject Classification: 62F12, 62J05.
Keywords and phrases: model selection criteria, asymptotic properties, orthogonal regressors, order statistics.
Received September 1, 2015

© 2015 Scientific Advances Publishers
1. Introduction

The two most widely used model selection criteria, namely, AIC (Akaike [2]) and BIC (Schwarz [18]), take only the number \(k\) of explanatory variables (regressors) into account that are actually included in a regression model but not the total number \(K\) of candidate variables from which the \(k\) variables are selected. They are therefore particularly suitable for nested models but only of limited suitability in the case of non-nested models where arbitrary subsets of a given set of candidate variables can be chosen. The problem of not taking the total number \(K\) of potential regressors into account becomes more serious as \(K\) increases. The BIC can only identify the true model consistently as long as \(K\) is not too large. In cases where \(K\) increases with the sample size, it tends to select too many regressors. Because of this deficiency of the BIC, several modifications have been proposed (e.g., Zheng and Loh [21]; Chen and Chen [3]; Wang et al. [20]; and Kim et al. [8]). However, these modifications typically are not fully specified but rather depend on tuning parameters which must be specified by the user. In a particular situation, e.g., in a simulation experiment, there is always the danger of trying many different values of the tuning parameters and reporting only the nice results. In this paper, we therefore deal only with fully specified criteria, primarily with modifications of the FPE (Rothman [17]; Akaike [1]).

The FPE does also not take the total number of potential regressors into account. However, since it is designed as an unbiased estimator of the prediction error, it can easily be modified in a unique way for the case of nested models. There is no need for any tuning parameters. For example, Reschenhofer [12] proposed to modify the FPE under the assumption of orthogonal regressors by replacing the number \(k\), which can be interpreted in the case of nested models as the expected value of the sum of \(k\) independent \(\chi^2(1)\) variables, by the expected value of the sum of the \(k\) largest of \(K\) independent \(\chi^2(1)\) variables, which is clearly
much larger than $k$ and therefore penalizes over-fitting much harsher (for a related subset-selection approach, see Tibshirani and Knight [19]; George and Foster [6]; for a related hypothesis-testing approach see, e.g., Diehr and Hoflin [4]; Khursheed and Wallenius [7]).

The next section first sets the framework and then derives conditions for consistent subset selection under the assumptions that the number of potential regressors increases and that the regressors are orthogonal. In addition, new subset-selection criteria satisfying these conditions are proposed. Section 3 presents the results of simulation experiments which compare the performance of the new criteria with that of conventional criteria. In all of these experiments, $K$ increases as $n$ increases, either at the same rate as $n$ or at a lower rate. Section 4 concludes.

2. Selecting Regressors in Nested and Nonnested Models

2.1. The model-selection criteria FPE and AIC

Suppose that $y$ and $z$ are independent samples of size $n$ from $N(\mu, \sigma^2 I)$ and $\mu$ has a representation $\mu = X\beta$, where $X$ is an $n \times k$ matrix $X$ of rank $k$. The least-squares (LS) estimator

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$  

which minimizes the sum of squared errors

$$S(X) = \|y - X\hat{\beta}\|^2,$$  

is best linear unbiased with variance

$$\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$  

An unbiased estimator for the mean squared prediction error

$$E\|z - X\hat{\beta}\|^2 = \text{tr}(\text{var}(z) + \text{var}(X\hat{\beta})) = n\sigma^2 + \text{tr}(X^T X \text{var}(\hat{\beta})) = (n + k)\sigma^2,$$  

(4)
is given by the final prediction error

\[ FPE(k) = \frac{n + k}{n - k} S(X) \]  

(Rothman [17]; Akaike [1]) because

\[ E(S(X)) = (n - k)\sigma^2. \]  

Selecting the number of regressors by minimization of \( FPE(k) \) with respect to \( k = 1, \ldots, K \) is asymptotically equivalent to minimization of

\[ AIC(k) = n \log(S(X)) + 2k \]  

(Akaike [2]) because

\[
n \log\left( \frac{n + k}{n - k} S(X) \right) = n \log(S(X)) + \log\left(1 + \frac{2k}{n - k}\right)^n \sim n \log(S(X)) + 2k.
\]

Overfitting is penalized by the multiplicative term \((n + k)/(n - k)\) in the case of FPE and by the additive penalty term \(2k\) in the case of AIC.

2.2. The subset-selection criteria \( FPE_{sub} \) and \( FPE_{\Delta} \)

If the columns of the \( n \times k \) matrix \( X \) are chosen from a set of \( K \)-orthogonal vectors \( x_1, \ldots, x_K \) of length \( \sqrt{n} \), the LS estimator simplifies to

\[ \hat{\beta} = \frac{1}{n} X^T y, \]  

and the sum of squared errors can be written as

\[
S(X) = y^T y - 2y^T X \left( \frac{1}{n} X^T y \right) + \left( \frac{1}{n} y^T X \right) X^T X \left( \frac{1}{n} X^T y \right)
\]

\[
= y^T y - \frac{1}{n} y^T XX^T y = y^T \hat{y} - n\hat{\beta}^T \hat{\beta}.
\]
Let
\[ \hat{\beta}^2_{(1)} \geq \ldots \geq \hat{\beta}^2_{(K)} \]  
be the order statistics of
\[ \hat{\beta}_j^2 = \left( \frac{1}{n} x_j^T y \right)^2, \quad j = 1, \ldots, K, \]  
and \( x_{(1)}, \ldots, x_{(K)} \) the corresponding rearrangement of the vectors \( x_1, \ldots, x_K \). Under the additional assumption that \( \beta = 0 \),
\[ \hat{\beta} = (x_1, \ldots, x_K)^T y \sim N(0, \frac{\sigma^2}{n} I), \]  
and
\[ \frac{n\hat{\beta}_1^2}{\sigma^2}, \ldots, \frac{n\hat{\beta}_K^2}{\sigma^2} \text{ i.i.d. } \chi^2(1). \]  
It then follows from
\[ E(S(\hat{X}(k))) = E(y^T y - n\hat{\beta}(k)^T \hat{\beta}(k)) = \sigma^2 \left( E\left( \frac{y^T y}{\sigma^2} - E \sum_{j=1}^k \frac{n\hat{\beta}_j^2}{\sigma^2} \right) \right) \]
\[ = \sigma^2 (n - \zeta_1(k, K)), \]  
and
\[ E\left\| z - \hat{X}(k)\hat{\beta}(k) \right\|^2 = \sigma^2 (n + \zeta_1(k, K)), \]  
that
\[ FPE_{sub}(k) = \frac{n + \zeta_1(k, K)}{n - \zeta_1(k, K)} S(\hat{X}(k)), \]  
(Reschenhofer [12]) is an unbiased estimator for the mean squared prediction error (16), where
\[ \hat{X}(k) = (x_{(1)}, \ldots, x_{(k)}). \]
\[
\hat{\beta}(k) = \frac{1}{n} \hat{X}(k)^T y,
\] (19)

and \(\zeta_r(k, K)\) is the expected value of a random variable \(Z_r(k, K)\), which is distributed as the sum of the \(k\) largest of \(K\) i.i.d. \(\chi^2(r)\) random variables (for the distribution of sums of gamma order statistics see Khursheed and Wallenius [7]; for explicit expressions for moments of \(\chi^2\) order statistics, see Nadarajah [9]; for tables of \(\zeta_1(k, K)\), see Reschenhofer [13]). If \(\zeta_1(k, K)\) is small compared to \(n\),

\[
AIC_{\text{sub}}(k) = n \log \left( S(\hat{X}(k)) \right) + 2\zeta_1(k, K)
\] (20)

can be used instead of (17).

2.3. Stepwise procedures

In a case where there are \(k_0\) dominant regressors which are certain to be included, it seems more appropriate to compare the largest of the remaining \(K - k_0\) squared coefficients, i.e., \(n\hat{\beta}^2_{(k_0+1)} / \sigma^2\), with \(\zeta_1(1, K - k_0)\) rather than with \(\zeta_1(k_0 + 1, K) - \zeta_1(k_0, K)\) (see Foster and George [5]; Reschenhofer et al. [16], [15]). Following this line of argument, we arrive at the criterion

\[
\frac{n + \sum_{j=1}^{k} \zeta_1(1, K - (j - 1))}{n - \sum_{j=1}^{k} \zeta_1(1, K - (j - 1))} S(\hat{X}(k)).
\] (21)

If \(k\) is small, criterion (21) is very similar to

\[
\frac{n + k\zeta_1(1, K)}{n - k\zeta_1(1, K)} S(\hat{X}(k))
\] (22)

because the maximum of a large sample will hardly change if a small number of elements are removed. In contrast, the second or the third largest value might already be much smaller than the maximum.
Unfortunately, the sums $\zeta_1(1, K) + \ldots + \zeta_1(1, K - k + 1)$ and $\zeta_1(1, K) + \ldots + \zeta_1(1, K)$ can quickly become too large relative to $n$. However, this problem can easily be overcome by taking a stepwise approach, according to which model dimension $k$ should be preferred over model dimension $k - 1$ if

$$\frac{n + (k - 1) + \zeta_1(1, K - (k - 1))}{n - (k - 1) - \zeta_1(1, K - (k - 1))} S(\hat{X}(k)) < \frac{n + (k - 1)}{n - (k - 1)} S(\hat{X}(k - 1)).$$

(23)

A big advantage of the stepwise procedure is that it does not require tables of $\zeta_1(k, K)$. Only expected values of the largest order statistics occur, which may be approximated roughly by

$$\hat{\zeta}_1(1, K) = 2 \log(K) - \log(\log(K))$$

(24)

(Reschenhofer [12]).

2.4. Consistency

In the following, our focus will be on the danger of over fitting in the case where there are $k_0$ dominant regression coefficients and $K - k_0$ vanishing coefficients. Underfitting is not an issue in large samples because the implied increase in the sum of squared errors could never be compensated by the comparatively small decrease in the penalty. We explore conditions under which the probability that a criterion of the form

$$\frac{n + p(k, K)}{n - p(k, K)} S(\hat{X}(k)),$$

(25)

incorrectly prefers model dimension $k > k_0$ over dimension $k_0$ vanishes as $n$ goes to infinity. In the following, it is assumed that

$$K - k \rightarrow \infty.$$
Using (53) (see Appendix A), we obtain

\[
P \left( S(\hat{X}(k)) \frac{n + p(k, K)}{n - p(k, K)} < S(\hat{X}(k_0)) \frac{n + p(k_0, K)}{n - p(k_0, K)} \right)
\]

\[
= P \left( \left( S(\hat{X}(k_0)) - S(\hat{X}(k)) \right) \frac{n + p(k, K)}{n - p(k, K)} < S(\hat{X}(k_0)) \frac{n + p(k_0, K)}{n - p(k_0, K)} \right)
\]

\[
= P \left( \left( S(\hat{X}(k)_0) - S(\hat{X}(k)) \right) \frac{n + p(k, K)}{n - p(k, K)} < \left( S(\hat{X}(k)) - S(\hat{X}(k_0)) \right) \frac{n + p(k_0, K)}{n - p(k_0, K)} \right)
\]

\[
= P \left( \frac{S(\hat{X}(k_0))}{n - p(k_0, K)\sigma^2} \frac{n}{n + p(k, K)} \frac{2(p(k, K) - p(k_0, K))}{\sigma^2} < \frac{S(\hat{X}(k)) - S(\hat{X}(k_0))}{\sigma^2} \right)
\]

\[
\sim P(2(p(k, K) - p(k_0, K)) < Z_1(k - k_0, K - k_0))
\]

\[
\leq P(2(p(k, K) - p(k_0, K)) - \zeta_2(k - k_0, K - k_0)
\]

\[
< Z_2(k - k_0, K - k_0) - \zeta_2(k - k_0, K - k_0))
\]

\[
\leq P \left( |Z_2(k - k_0, K - k_0) - \zeta_2(k - k_0, K - k_0)| \right)
\]

\[
\geq \left( \frac{2(p(k, K) - p(k_0, K))}{\zeta_2(k - k_0, K - k_0)} - 1 \right) \zeta_2(k - k_0, K - k_0)
\]

\[
\leq \left( \frac{2(p(k, K) - p(k_0, K))}{\zeta_2(k - k_0, K - k_0)} - 1 \right)^2 \frac{\text{var}(Z_2(k - k_0, K - k_0))}{(\zeta_2(k - k_0, K - k_0))^2} \to 0,
\]  

(27)

as \( K \to \infty \), if

\[
0 = p(0, K) \leq \ldots \leq p(k, K) = o(n),
\]

(28)

and

\[
\lim \inf \frac{2(p(k, K) - p(k_0, K))}{\zeta_2(k - k_0, K - k_0)} = \omega > 1.
\]

(29)

For example, if

\[
k \leq \sqrt{n},
\]

(30)
condition (28) is met for penalties such as
\[ \hat{p}_1(k, K) = k \hat{\zeta}_1(1, K) = 2k \log(K) - k \log(\log(K)), \]
(31)
\[ p_2(k, K) = k \zeta_2(1, K) = 2k \sum_{j=1}^{K} \frac{1}{j}, \]
(32)
(see (54) in Appendix A), and
\[ \hat{p}_2(k, K) = k \hat{\zeta}_2(1, K) = 2k \log(K) + 2kC, \]
(33)
(see (55) in Appendix A) even in the extreme case where \( K = n \). In addition, it follows from
\[ k \zeta_2(1, K) - k_0 \zeta_2(1, K) = (k - k_0) \zeta_2(1, K) \geq \zeta_2(k - k_0, K - k_0), \]
(34)
and
\[ \zeta_2(k - k_0, K - k_0) = 2(k - k_0) \left( 1 + \sum_{j=k-k_0+1}^{K-k_0} \frac{1}{j} \right) \leq 2(k - k_0) \log(K). \]
(35)
that condition (29) is also met for these penalties.

In contrast, for
\[ p_2(k, K) = \zeta_2(k, K), \]
(36)
and
\[ \bar{p}_2(k, K) = 2k \log \left( \frac{K}{k} \right) + 2k - 1, \]
(37)
(see (56) in Appendix A), we obtain in the special case \( k_0 + 1 = k = \sqrt{K} \),
\[ \zeta_2(k - k_0, K - k_0) = 2 \sum_{j=1}^{K-k+1} \frac{1}{j} \sim 2 \log(K), \]
(38)
\[ \zeta_2(k, K) - \zeta_2(k_0, K) = 2k \left( 1 + \sum_{j=k+1}^{K} \frac{1}{j} \right) - (2k - 2) \left( 1 + \sum_{j=k}^{K} \frac{1}{j} \right) = 2 \sum_{j=k+1}^{K} \frac{1}{j} \sim 2(\log(K) - \log(k)) = \log(K), \]
(39)
Furthermore, for

\[ p_{BIC}(k, K) = \frac{k}{2} \log(n), \]

and \( k_0 = 0, k = 1, K = n \), we obtain

\[ \frac{2(p_{BIC}(k, K) - p_{BIC}(k_0, K))}{\zeta_2(k - k_0, K - k_0)} = \frac{2p_{BIC}(1, n)}{\zeta_2(1, n)} = \frac{\log(n)}{2\sum_{j=1}^{n-1} j} < 1. \]

This should serve as a warning against the use of

\[ BIC(k) = n \log \left( S(\hat{X}(k)) \right) + k \log(n), \]

(Schwarz [18]) in the case of nonnested models with many potential regressors. However, for small \( K \), e.g., \( K \leq \frac{3}{\sqrt{n}} \), we still have

\[ \frac{2(p_{BIC}(k, K) - p_{BIC}(k_0, K))}{\zeta_2(k - k_0, K - k_0)} \geq \frac{(k - k_0) \log(n)}{2(k - k_0) \log(K)} \geq \frac{3}{2}. \]

3. Simulation Study

For sample sizes \( n = 100, 250, 500, 1000, 5000, \) and \( 10000, K = 2m = 2[(n - 1)/2] = n - 2 \) orthogonal vectors \( x_1, \ldots, x_K \) of length \( \sqrt{n} \) are defined as the columns of the matrix
\[
X = \sqrt{2} \begin{pmatrix}
\cos(\omega_1) & \sin(\omega_1) & \cdots & \cos(\omega_K) & \sin(\omega_K) \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\cos(\omega_1) & \sin(\omega_1) & \cdots & \cos(\omega_K) & \sin(\omega_K)
\end{pmatrix},
\]

where \( \omega_j = 2\pi j / n, \) \( j = 1, \ldots, m, \) are the Fourier frequencies less than \( \pi. \)

For each sample size \( n, k_0 = 0, 1, [0.25k_1], [0.5k_1], [0.75k_1], k_1 = \lfloor \sqrt{n} \rfloor \) components of the \( K \)-dimensional vector \( \beta \) are randomly selected and set to 0.3. The remaining \( K - k_0 \) components are set to 0. Finally, \( r = 1000 \) pseudo-random samples from \( N(X\beta, \sigma^2 I) \) are generated with the software R (R Core Team [10]) and criteria of the form

\[
n \log(S(X)) + 2p(k, K),
\]

are used to determine which components are zero and which are non-zero. Here the penalties \( p(k, K) \) are given by \( k \) (AIC), (43) (BIC), (36), (37), (31), (32), and (33), respectively. The percentages of wrong decisions are given in Table 1. AIC tends to select always the maximum number of regressors, which is fixed at \( k_1, \) and BIC performs only slightly better. The criteria based on (36) and (37), which just miss to satisfy the consistency conditions, perform much better than AIC and BIC. However, even in large samples they occasionally make wrong decisions. In contrast, the consistent criteria, which are based on (31)-(33), include all genuine regressors and exclude all others in large samples.

To illustrate the improved large-sample performance of BIC when \( K \) grows at a lower rate than \( n, K = 2m = 2\lfloor n^{1/3} / 2 \rfloor \) is chosen rather than \( K = n - 2 \) for sample sizes \( n = 100, 250, 500, 1000, 5000, 10000, 50000, \) and \( 100000. \) Again, \( k_0 = 0, 1, [0.25k_1], [0.5k_1], [0.75k_1], k_1 = [0.75K] \) components of \( \beta \) are set to 0.3. The particular choice of the maximum model dimension \( k_1 \) makes sure that condition (26) is met. The results
are given in Table 2. Obviously, and not unexpectedly, the reduction in the number of potential regressors proves advantageous only for BIC but neither for (36) and (37) nor for AIC. However, even for the consistent criteria, convergence is quite slow because $K - k_1$ is small.

4. Discussion

Consistency conditions for the selection of orthogonal regressors are derived and selection criteria satisfying these conditions are proposed. The simulation results show that these criteria outperform conventional criteria such as AIC and BIC for various sample sizes. Their superiority still holds when the framework conditions are changed in favor of BIC, i.e., by choosing a very small rate of growth for the number of potential regressors and a maximum model dimension, which is only slightly smaller. Both the theoretical results and the simulation results serve as a warning against blindly relying on BIC in the case of large sample sizes. Particular caution is required when the number of potential regressors is large.

In a subsequent paper (Reschenhofer [14]), the consistency conditions are adapted to the case of non-nested frequency-domain models and applied to macroeconomic data.
Table 1. Proportions of wrong decisions ($K = n$)

<table>
<thead>
<tr>
<th>$n$</th>
<th>$k_0$</th>
<th>AIC</th>
<th>BIC</th>
<th>$\hat{p}_2$</th>
<th>$\hat{p}_2$</th>
<th>$\hat{p}_1$</th>
<th>$p_2$</th>
<th>$\hat{p}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0</td>
<td>1.0000</td>
<td>0.7740</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00090</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>0.9172</td>
<td>0.7537</td>
<td>0.09560</td>
<td>0.09550</td>
<td>0.08550</td>
<td>0.09560</td>
<td>0.09550</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.8418</td>
<td>0.7167</td>
<td>0.19200</td>
<td>0.19120</td>
<td>0.17270</td>
<td>0.19390</td>
<td>0.19390</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.6292</td>
<td>0.5825</td>
<td>0.49400</td>
<td>0.49350</td>
<td>0.47080</td>
<td>0.49600</td>
<td>0.49600</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>0.5360</td>
<td>0.5057</td>
<td>0.69530</td>
<td>0.69500</td>
<td>0.67500</td>
<td>0.69750</td>
<td>0.69750</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>0.4460</td>
<td>0.4406</td>
<td>0.99760</td>
<td>0.99760</td>
<td>0.98550</td>
<td>0.99920</td>
<td>0.99920</td>
</tr>
<tr>
<td>500</td>
<td>0</td>
<td>1.0000</td>
<td>0.4970</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00018</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>500</td>
<td>1</td>
<td>0.9555</td>
<td>0.5009</td>
<td>0.00405</td>
<td>0.00377</td>
<td>0.00123</td>
<td>0.00405</td>
<td>0.00405</td>
</tr>
<tr>
<td>500</td>
<td>5</td>
<td>0.7727</td>
<td>0.4930</td>
<td>0.00255</td>
<td>0.00255</td>
<td>0.00491</td>
<td>0.01891</td>
<td>0.01886</td>
</tr>
<tr>
<td>500</td>
<td>11</td>
<td>0.5000</td>
<td>0.4269</td>
<td>0.00382</td>
<td>0.00955</td>
<td>0.05359</td>
<td>0.05355</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>16</td>
<td>0.2727</td>
<td>0.2682</td>
<td>0.00700</td>
<td>0.00700</td>
<td>0.01336</td>
<td>0.14959</td>
<td>0.14877</td>
</tr>
<tr>
<td>500</td>
<td>22</td>
<td>0.0012</td>
<td>0.0012</td>
<td>0.00145</td>
<td>0.00145</td>
<td>0.01809</td>
<td>0.79186</td>
<td>0.79077</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>1.0000</td>
<td>0.4126</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>1000</td>
<td>1</td>
<td>0.9677</td>
<td>0.4181</td>
<td>0.00006</td>
<td>0.00003</td>
<td>0.00006</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>1000</td>
<td>7</td>
<td>0.7741</td>
<td>0.4236</td>
<td>0.00029</td>
<td>0.00029</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>1000</td>
<td>15</td>
<td>0.5161</td>
<td>0.4131</td>
<td>0.00168</td>
<td>0.00168</td>
<td>0.00003</td>
<td>0.00003</td>
<td>0.00003</td>
</tr>
<tr>
<td>1000</td>
<td>23</td>
<td>0.2586</td>
<td>0.2530</td>
<td>0.00452</td>
<td>0.00452</td>
<td>0.00003</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>1000</td>
<td>31</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>5000</td>
<td>0</td>
<td>1.0000</td>
<td>0.3078</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>5000</td>
<td>1</td>
<td>0.9857</td>
<td>0.3090</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>5000</td>
<td>17</td>
<td>0.7571</td>
<td>0.3152</td>
<td>0.00014</td>
<td>0.00014</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>5000</td>
<td>35</td>
<td>0.5000</td>
<td>0.3178</td>
<td>0.00051</td>
<td>0.00051</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>5000</td>
<td>52</td>
<td>0.2571</td>
<td>0.2494</td>
<td>0.00156</td>
<td>0.00156</td>
<td>0.00001</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>5000</td>
<td>70</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>10000</td>
<td>0</td>
<td>1.0000</td>
<td>0.2803</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>10000</td>
<td>1</td>
<td>0.9900</td>
<td>0.2785</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>10000</td>
<td>25</td>
<td>0.7500</td>
<td>0.2809</td>
<td>0.00012</td>
<td>0.00012</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>10000</td>
<td>50</td>
<td>0.5000</td>
<td>0.2852</td>
<td>0.00047</td>
<td>0.00047</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>10000</td>
<td>75</td>
<td>0.2500</td>
<td>0.2417</td>
<td>0.00092</td>
<td>0.00092</td>
<td>0.00001</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>10000</td>
<td>100</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>
Table 2. Proportions of wrong decisions ($K = n^{1/3}$)

<table>
<thead>
<tr>
<th>$n$</th>
<th>$k_0$</th>
<th>AIC</th>
<th>BPC</th>
<th>$\hat{p}_2$</th>
<th>$\hat{p}_2$</th>
<th>$\hat{p}_1$</th>
<th>$\hat{p}_2$</th>
<th>$\hat{p}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0</td>
<td>0.12346</td>
<td>0.12346</td>
<td>0.12346</td>
<td>0.12346</td>
<td>0.12346</td>
<td>0.12346</td>
<td>0.12346</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0.21667</td>
<td>0.04500</td>
<td>0.00900</td>
<td>0.01367</td>
<td>0.03833</td>
<td>0.00600</td>
<td>0.00800</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>0.17633</td>
<td>0.09933</td>
<td>0.18300</td>
<td>0.18000</td>
<td>0.09967</td>
<td>0.15833</td>
<td>0.14733</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.13300</td>
<td>0.15267</td>
<td>0.23233</td>
<td>0.22067</td>
<td>0.16133</td>
<td>0.31633</td>
<td>0.29500</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.08833</td>
<td>0.20267</td>
<td>0.13867</td>
<td>0.11900</td>
<td>0.21900</td>
<td>0.48067</td>
<td>0.44567</td>
</tr>
<tr>
<td>500</td>
<td>0</td>
<td>0.24350</td>
<td>0.02050</td>
<td>0.00450</td>
<td>0.00550</td>
<td>0.02225</td>
<td>0.00325</td>
<td>0.00350</td>
</tr>
<tr>
<td>500</td>
<td>1</td>
<td>0.19075</td>
<td>0.01325</td>
<td>0.02675</td>
<td>0.03275</td>
<td>0.01750</td>
<td>0.00150</td>
<td>0.00225</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>0.16400</td>
<td>0.01375</td>
<td>0.07375</td>
<td>0.09000</td>
<td>0.01550</td>
<td>0.00225</td>
<td>0.00250</td>
</tr>
<tr>
<td>500</td>
<td>3</td>
<td>0.10600</td>
<td>0.01400</td>
<td>0.08125</td>
<td>0.09625</td>
<td>0.01475</td>
<td>0.00175</td>
<td>0.00200</td>
</tr>
<tr>
<td>500</td>
<td>4</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00050</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>0.20800</td>
<td>0.00950</td>
<td>0.00133</td>
<td>0.00217</td>
<td>0.00967</td>
<td>0.00100</td>
<td>0.00133</td>
</tr>
<tr>
<td>1000</td>
<td>1</td>
<td>0.17833</td>
<td>0.01183</td>
<td>0.01433</td>
<td>0.01617</td>
<td>0.01183</td>
<td>0.00133</td>
<td>0.00150</td>
</tr>
<tr>
<td>1000</td>
<td>3</td>
<td>0.19733</td>
<td>0.01250</td>
<td>0.01667</td>
<td>0.01917</td>
<td>0.01317</td>
<td>0.00117</td>
<td>0.00167</td>
</tr>
<tr>
<td>1000</td>
<td>4</td>
<td>0.12550</td>
<td>0.00683</td>
<td>0.06417</td>
<td>0.07750</td>
<td>0.00733</td>
<td>0.00033</td>
<td>0.00050</td>
</tr>
<tr>
<td>1000</td>
<td>5</td>
<td>0.10600</td>
<td>0.00533</td>
<td>0.09183</td>
<td>0.11067</td>
<td>0.00550</td>
<td>0.00050</td>
<td>0.00050</td>
</tr>
<tr>
<td>1000</td>
<td>6</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>0.20667</td>
<td>0.00375</td>
<td>0.00017</td>
<td>0.00017</td>
<td>0.00258</td>
<td>0.00017</td>
<td>0.00017</td>
</tr>
<tr>
<td>1000</td>
<td>1</td>
<td>0.19808</td>
<td>0.00367</td>
<td>0.00317</td>
<td>0.00308</td>
<td>0.00300</td>
<td>0.00042</td>
<td>0.00042</td>
</tr>
<tr>
<td>1000</td>
<td>3</td>
<td>0.17108</td>
<td>0.00358</td>
<td>0.01492</td>
<td>0.01567</td>
<td>0.00283</td>
<td>0.00017</td>
<td>0.00025</td>
</tr>
<tr>
<td>1000</td>
<td>6</td>
<td>0.13592</td>
<td>0.00392</td>
<td>0.06575</td>
<td>0.07425</td>
<td>0.00267</td>
<td>0.00050</td>
<td>0.00050</td>
</tr>
<tr>
<td>1000</td>
<td>9</td>
<td>0.09092</td>
<td>0.00192</td>
<td>0.10367</td>
<td>0.11333</td>
<td>0.00133</td>
<td>0.00017</td>
<td>0.00025</td>
</tr>
<tr>
<td>10000</td>
<td>12</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>10000</td>
<td>0</td>
<td>0.20927</td>
<td>0.00367</td>
<td>0.00020</td>
<td>0.00027</td>
<td>0.00273</td>
<td>0.00020</td>
<td>0.00027</td>
</tr>
<tr>
<td>10000</td>
<td>1</td>
<td>0.20207</td>
<td>0.00360</td>
<td>0.00220</td>
<td>0.00220</td>
<td>0.00307</td>
<td>0.00060</td>
<td>0.00060</td>
</tr>
<tr>
<td>10000</td>
<td>3</td>
<td>0.17527</td>
<td>0.00220</td>
<td>0.00920</td>
<td>0.00960</td>
<td>0.00167</td>
<td>0.00033</td>
<td>0.00033</td>
</tr>
<tr>
<td>10000</td>
<td>7</td>
<td>0.13673</td>
<td>0.00160</td>
<td>0.05180</td>
<td>0.05773</td>
<td>0.00127</td>
<td>0.00007</td>
<td>0.00013</td>
</tr>
<tr>
<td>10000</td>
<td>11</td>
<td>0.09353</td>
<td>0.00147</td>
<td>0.10380</td>
<td>0.11287</td>
<td>0.00107</td>
<td>0.00007</td>
<td>0.00007</td>
</tr>
<tr>
<td>10000</td>
<td>15</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>50000</td>
<td>0</td>
<td>0.21081</td>
<td>0.00152</td>
<td>0.00007</td>
<td>0.00011</td>
<td>0.00085</td>
<td>0.00007</td>
<td>0.00011</td>
</tr>
<tr>
<td>50000</td>
<td>1</td>
<td>0.20415</td>
<td>0.00115</td>
<td>0.00037</td>
<td>0.00030</td>
<td>0.00067</td>
<td>0.00007</td>
<td>0.00007</td>
</tr>
</tbody>
</table>
Table 2. (Continued)

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>50000</td>
<td>6</td>
<td>0.17778</td>
<td>0.00100</td>
<td>0.00944</td>
<td>0.00959</td>
<td>0.00052</td>
</tr>
<tr>
<td>50000</td>
<td>13</td>
<td>0.13856</td>
<td>0.00070</td>
<td>0.05296</td>
<td>0.05515</td>
<td>0.00052</td>
</tr>
<tr>
<td>50000</td>
<td>20</td>
<td>0.09319</td>
<td>0.00044</td>
<td>0.10770</td>
<td>0.11322</td>
<td>0.00026</td>
</tr>
<tr>
<td>50000</td>
<td>27</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>100000</td>
<td>0</td>
<td>0.21424</td>
<td>0.00076</td>
<td>0.00003</td>
<td>0.00003</td>
<td>0.00047</td>
</tr>
<tr>
<td>100000</td>
<td>1</td>
<td>0.20994</td>
<td>0.00053</td>
<td>0.00021</td>
<td>0.00021</td>
<td>0.00032</td>
</tr>
<tr>
<td>100000</td>
<td>8</td>
<td>0.17585</td>
<td>0.00074</td>
<td>0.01109</td>
<td>0.01147</td>
<td>0.00041</td>
</tr>
<tr>
<td>100000</td>
<td>17</td>
<td>0.13459</td>
<td>0.00074</td>
<td>0.05209</td>
<td>0.05471</td>
<td>0.00035</td>
</tr>
<tr>
<td>100000</td>
<td>25</td>
<td>0.09500</td>
<td>0.00024</td>
<td>0.10221</td>
<td>0.10597</td>
<td>0.00018</td>
</tr>
<tr>
<td>100000</td>
<td>34</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

References

Appendix A

When we are dealing with samples from a $\chi^2$ distribution with two degrees of freedom instead of only one, the mean and the variance of sums of order statistics can easily be obtained by first noting that a simple division by 2 transforms a $\chi^2(2)$ distribution into a standard exponential distribution and then using Rényi’s representation (Rényi [11]). According to this representation, the order statistics
of a random sample $E_1, \ldots, E_K$ from a standard exponential distribution have the same distribution as

$$\frac{e_1}{1} + \ldots + \frac{e_K}{K} \geq \frac{e_2}{2} + \ldots + \frac{e_K}{K} \geq \ldots \geq \frac{e_{K-1}}{K-1} + \frac{e_K}{K} \geq \frac{e_K}{K},$$

(50)

where $e_1, \ldots, e_K$ are i.i.d. $\text{Exp}(1)$. Thus,

$$E(Z_2(k, K)) = 2E(E(1) + \ldots + E(K))$$

$$= 2E\left(\frac{e_1}{1} + \ldots + \frac{e_{k-1}}{k-1} (k-1) + \left(\frac{e_k}{k} + \frac{e_{k-1}}{k+1} + \ldots + \frac{e_K}{K}\right) k\right)$$

$$= 2E\left(e_1 + \ldots + e_{k-1} + e_k + \left(\frac{e_{k+1}}{k+1} + \ldots + \frac{e_K}{K}\right) k\right)$$

$$= 2\left(k + k \sum_{j=k+1}^{K} \frac{1}{j}\right),$$

(51)

and

$$\text{var}(Z_2(k, K)) = 4 \text{var}(E(1) + \ldots + E(k))$$

$$= 4 \text{var}\left(e_1 + \ldots + e_{k-1} + e_k + \left(\frac{e_{k+1}}{k+1} + \ldots + \frac{e_K}{K}\right) k\right)$$

$$= 4\left(k + k^2 \left(\frac{1}{k+1}\right)^2 + \ldots + \left(\frac{1}{K}\right)^2\right)$$

$$\leq 4\left(k + k^2 \sum_{j=k+1}^{K} \int_{j-1}^{j} \frac{1}{x^2} \, dx\right)$$

$$= 4\left(k + k^2 \left(\frac{1}{k} - \frac{1}{K}\right)\right)$$

$$\leq 8k.$$  

(52)
If $K - k \to \infty$, we have

$$\frac{\text{var}(Z_2(k, K))}{(E(Z_2(k, K)))^2} \leq \frac{2}{k \left(1 + \sum_{j=k+1}^{K} \frac{1}{j}\right)^2} \to 0. \quad (53)$$

Furthermore,

$$E(Z_2(1, K)) = 2 \sum_{j=1}^{K} \frac{1}{j},$$

$$= 2 \left(\log(K) + C + \frac{1}{2K} + O\left(\frac{1}{K^2}\right)\right)$$

$$= 2(\log(K) + C + o(1)), \quad (55)$$

where $C = 0.5772156649...$ is Euler’s constant, and

$$E(Z_2(k, K)) = 2k \left(1 + \log(K) + C + \frac{1}{2K} + O\left(\frac{1}{K^2}\right)\right) - \log(k) - C - \frac{1}{2k} + O\left(\frac{1}{k^2}\right)$$

$$= 2k \log\left(\frac{K}{k}\right) + 2k - 1 + o(1), \quad (56)$$

if $k \to \infty$ and $k = o(K)$.

For the more general quantity $Z_2^*$, which reduces to $Z_2$ if $k_0 = 0$, we obtain for $k_0 > 0$

$$E(Z_2^*(k_0 + 1, k, K)) = 2E(E(k_0 + 1) + \ldots + E(k))$$

$$= 2E\left(\frac{e_{k_0+1}}{k_0 + 1} + \ldots + \frac{e_{k-1}}{k-1} (k - k_0 - 1) + \left(\frac{e_k}{k} + \frac{e_1}{k+1} + \ldots + \frac{e_K}{K}\right) (k - k_0)\right)$$

$$\geq 2(k - k_0)E\left(\frac{e_k}{k} + \frac{e_1}{k+1} + \ldots + \frac{e_K}{K}\right)$$

$$= 2(k - k_0) \sum_{j=k}^{K} \frac{1}{j}, \quad (57)$$
and

$$\text{var}(z^*_2(k_0 + 1, k, K)) = 4 \text{var}(E_{(k_0+1)} + \ldots + E_{(k)})$$

$$\leq 4(k - k_0)^2 \sum_{j=k_0+1}^{K} \left( \frac{1}{j} \right)^2$$

$$\leq 4(k - k_0)^2 \sum_{j=k_0+1}^{K} \int_{j-1}^{j} \frac{1}{\lambda^2} d\lambda$$

$$= 4(k - k_0)^2 \left( \frac{1}{k_0} - \frac{1}{K} \right)$$

$$\leq 4(k - k_0)^2 \frac{1}{k_0},$$

(58)

hence

$$\frac{\text{var}(z^*_2(k_0 + 1, k, K))}{(E[z^*_2(k_0 + 1, k, K)])^2} \leq 6 \sum_{j=k+1}^{k_0} \frac{1}{j^2} \to 0$$

(59)

provided that $K - k \to \infty$. 

$\blacksquare$