Eur. J. Stat. 2 (2022) 11
doi: 10.28924/ada/stat.2.11

Evaluating Pairwise Variable Selection Methods

Erhard Reschenhofer

Department of Statistics and Operations Research, University of Vienna
Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
Correspondence: erhard.reschenhofer@univie.ac.at

ABSTRACT. This paper discusses novel methods for the pairwise selection of explanatory variables from a large set of candidate pairs. These methods are applied to monthly time series of surface temperature and their performance is compared with that of conventional selection criteria such as AIC and BIC. In our frequency-domain analysis of the temperature datasets, the pairs are defined in a natural way as cosine and sine vectors of the same frequency. The results show that the new criteria are the only ones which are able to correctly identify seasonal patterns.

1. Introduction

Reschenhofer (2015a) derived conditions for the consistent selection of a subset from a large set of potential explanatory variables and proposed automatic subset selection criteria, which satisfy these conditions. For situations where the explanatory variables can only be selected pairwise, Reschenhofer (2015b) proposed analogous criteria. From a technical point of view, the latter setting is easier because it involves quantities that are χ^2-distributed with two degrees of freedom rather than only one degree of freedom, which allows the use of Rényi's representation (Rényi, 1953) for the calculation of the mean and the variance of sums of order statistics (note that a standard exponentially distributed variable can easily be obtained by dividing a $\chi^2(2)$-distributed variable by 2). The need for new subset selection criteria is due to the fact that the widely used criteria AIC (Akaike, 1973) and BIC (Schwarz, 1978) are particularly suitable for nested models but do not take the

Received: 23 Jan 2022.
Key words and phrases. subset selection; frequency-domain analysis; seasonal patterns; surface temperature.

extent of possible data snooping into account, which is particularly problematic when the set of candidate variables is large.

Criteria of the AIC-type can for the classical regression model in the most general form be written as

$$-2\log\big(f(y;X\hat{\beta},\hat{\sigma}^2 I)\big)+2(k+1)\hat{Q}+2\hat{Q}^2+\frac{14\hat{Q}^2+2k^2\hat{Q}^2-8k\hat{Q}^3+24\hat{Q}^2-32\hat{Q}^3+12\hat{Q}^4}{n-k-2} \qquad (1)$$

(Reschenhofer, 1999), where the first term is two times the negative maximum log likelihood and the other terms serve to penalize overparametrization. There are $k+1$ model parameters, namely $\beta_1,\ldots,\beta_k,\sigma^2$. The number of model parameters is used as a measure of model complexity. Misspecificaton of a model is meaasued by $\hat{Q}=\hat{\sigma}_0^2/\hat{\sigma}^2$, where $\hat{\sigma}_0^2$ is an unbiased estimator of the true error variance obtained from a large model (which is assumed to be correctly specified) and $\hat{\sigma}^2$ is the least squares (LS) estimator obtained from the possibly misspecified model $y=X\beta+u$. The criterion (1) reduces to Sawa's (1978) criterion for $n\to\infty$, the AIC for $n\to\infty$ and $\hat{Q}=1$, and the corrected AIC (Sugiura, 1978; Hurvich and Tsai, 1989) for $\hat{Q}=1$. It is clear that the criteria of the AIC-type are inconsistent because their penalty terms do not grow as $n\to\infty$. In contrast, the penalty term of the BIC, which is given by $(k+1)\log(n)$, increases as n increases but it still does not factor in the number K of candidate variables. Because of this shortcoming of the BIC, several modifications have been proposed (e.g. Zheng & Loh, 1997; Chen & Chen, 2008; Wang et al., 2009; Kim et al., 2012), which are typically not fully specified but depend on tuning parameters to be chosen by the user.

Soon after the criteria for pairwise variable selection (Reschenhofer, 2015b) had been published online in the SOP Transactions on Statistics and Analysis in 2015, this journal was apparently discontinued. For convenience of the reader, we will therefore review some of the theoretical results in the next section. In addition, we will also introduce new criteria. In Section 3, the different criteria are evaluated using real data. The design of this empirical investigation is tailored for the problem of consistently selecting pairs of variables from a large set of pairs. Firstly, the time series are extremely long. They start in January 1880 and extend to January 2021. Secondly, the number of available explanatory variables is of the same size as the length of the time series. Thirdly, the pairs of explanatory variables are formed in a natural way. Fourthly, the explanatory variables are orthogonal, which is essential for carrying out the required computations in an efficient way. Finally, there is indeed a genuine pattern in these time series, which allows a clear distinction between true and false models. Section 4 concludes.

2. Methods

2.1 Individual explanatory variables

In the following, we prefer to use criteria with multiplicative penalty terms instead of additive penalty terms. However, it is easily possible to switch back at any time from the latter to the former. For example, in the case of the final prediction error (FPE)

$$\frac{n+k}{n-k}S, \tag{2}$$

(Rothman, 1968; Akaike, 1969), where $S = \sum_{t=1}^{n} \hat{u}_t^2 = n\hat{\sigma}^2$ is the residual sum of squares, we obtain

$$n\log\left(\frac{n+k}{n-k}S\right) = n\log(S) + \log\left(1+\frac{2k}{n-k}\right)^n \approx n\log(S) + \log\big(\exp(2k)\big) = n\log(S) + 2k$$

$$= n\log(\hat{\sigma}^2) + 2k + n\log(n), \tag{3}$$

which is, apart from an additive constant, equal to the AIC

$$-2\log\big(f(y; X\hat{\beta}, \hat{\sigma}^2 I\big) + 2(k+1). \tag{4}$$

For this approximation, it is of course required that k is small compared to n.

The term k occurring in (2) coincides with the expected value of the sum of k i.i.d. $\chi^2(2)$-variables and has just the right size to ensure the unbiasedness of (2) as an estimator for the mean squared prediction error

$$E\left\|z - X\hat{\beta}\right\|^2, \tag{5}$$

where z is a new sample which is independent from the original sample y that has been used for the estimation of β. A conceptual shortcoming of model selection by minimizing some unbiased estimator of the squared prediction error is that it is implicitly assumed that each competing model is correctly specified, i.e., $EX\hat{\beta} = Ey$. However, taking a possible misspecification into account (e.g., by introducing the term $\hat{Q}$) has only a small effect on the performance of the model selection criterion because in the case of a serious misspecification the term S is decisive anyway (see Reschenhofer, 1999). We will therefore turn a blind eye to the possibility of a misspecification and focus solely on the danger of overfitting. The simplest way to do that is to assume that $Ey = 0$. To preserve unbiasedness (under the additional assumption of orthogonality of the explanatory variables) in the case of subset selection, where the residual sum of squares is minimized also with respect to the $n \times k$ matrix X whose columns can be chosen from a total of K columns, the term k in (2) has to be replaced by the expected value $\zeta_1(k, K)$ of the sum of the k largest of K independent and $\chi^2(1)$-distributed variables (Reschenhofer, 2004; for a related criterion see George and Foster, 2000). The criterion

$$FPE_{sub}(k) = \frac{n+\zeta_1(k,K)}{n-\zeta_1(k,K)} S, \tag{6}$$

is clearly an improvement over (3) because it takes the number K of available explanatory variables into account and therefore allows a fairer penalization of data snooping. Reschenhofer (2015a) argued that in the presence of some dominant explanatory variables, which are certain to be included, it is more appropriate to use $\zeta_1(1, K-k)$ instead of $\zeta_1(k+1, K)$ for the decision to increase the model dimension from k to $k+1$. A technical advantage of this stepwise procedure is that only expected values of the form $\zeta_1(1, K)$ are needed, which were approximated by $2\log(K)$ by Foster and George (1994) and used for the construction of their risk inflation criterion (RIC). A more accurate approximation is given by

$$\hat{\zeta}_1(1, K) = 2\log(K) - \log(\log(K)) \tag{7}$$

(Reschenhofer, 2004). For an alternative stepwise procedure see Reschenhofer (2010).

2.2 Pairs of explanatory variables

In contrast to time-domain models, frequency-domain models typically include only pairs of explanatory variables rather than individual explanatory variables because both a cosine vector and a sine vector of the same frequency are required to model an oscillation with an arbitrary phase. Suppose that the mean μ of an n-dimensional normal distribution with covariance matrix $\sigma^2 I$ can be written as a linear combination

$$\mu = \sum_{j=1}^{K} A_j \underbrace{\begin{pmatrix} \cos(\omega_j \cdot 1) \\ \vdots \\ \cos(\omega_j \cdot n) \end{pmatrix}}_{=:c_j} + \sum_{j=1}^{K} B_j \underbrace{\begin{pmatrix} \sin(\omega_j \cdot 1) \\ \vdots \\ \sin(\omega_j \cdot n) \end{pmatrix}}_{=:s_j}, \tag{8}$$

where $K = [(n-1)/2]$ and $0 < \omega_j = 2\pi j/n < \pi$ is the j-th Fourier frequency. Because of the orthogonality of the vectors $C_1, \dots, C_K, S_1, \dots, S_K$, the LS estimates of the coefficients A_j and B_j based on a sample y are given by

$$\hat{A}_j = \frac{\sum_{t=1}^{n} \cos(\omega_j t) y_t}{\sum_{t=1}^{n} \cos^2(\omega_j t)} = \frac{2}{n} \sum_{t=1}^{n} \cos(\omega_j t) y_t, \tag{9}$$

$$\hat{B}_j = \frac{\sum_{t=1}^{n} \sin(\omega_j t) y_t}{\sum_{t=1}^{n} \sin^2(\omega_j t)} = \frac{2}{n} \sum_{t=1}^{n} \sin(\omega_j t) y_t, \tag{10}$$

and the residual sum of squares by

$$\begin{aligned} S(K) &= \left(y - \sum \hat{A}_j C_j - \sum \hat{B}_j S_j\right)^T \left(y - \sum \hat{A}_j C_j - \sum \hat{B}_j S_j\right) \\ &= y^T y - \sum \hat{A}_j^2 C_j^T C_j \; - \sum \hat{B}_j^2 S_j^T S_j \\ &= \sum y_t^2 - \frac{n}{2} \sum \underbrace{(\hat{A}_j^2 + \hat{B}_j^2)}_{=:\hat{R}_j^2}. \end{aligned} \tag{11}$$

Under the additional assumption that $\mu = 0$, we have

$$\hat{A}_j, \dots, \hat{A}_j, \hat{B}_j, \dots, \hat{B}_j \sim N\left(0, \frac{2\sigma^2}{n} I\right) \tag{12}$$

and

$$\frac{n}{2\sigma^2} \hat{R}_1^2, \dots, \frac{n}{2\sigma^2} \hat{R}_K^2 \;\; \text{i.i.d. } \chi^2(2), \tag{13}$$

hence the expected value of the residual sum of squares of the best model with $k \le K$ pairs C_j, S_j of explanatory variables is given by

$$ES(k) = E\left(\sum_{t=1}^{n} y_t^2 - \frac{n}{2} \sum_{j=1}^{k} \hat{R}_{q(j)}^2\right) = \sigma^2 E\left(\sum_{t=1}^{n} \frac{y_t^2}{\sigma^2} - \frac{n}{2\sigma^2} \sum_{j=1}^{k} \hat{R}_{q(j)}^2\right) = \sigma^2 (n - \zeta_2(k, K)), \tag{14}$$

where q is a permutation which rearranges $\hat{R}_1^2, \dots, \hat{R}_K^2$ into descending order, i.e.,

$$\hat{R}_{q(1)}^2 \ge \cdots \ge \hat{R}_{q(K)}^2, \tag{15}$$

and

$$\zeta_2(k, K) = 2k\left(1 + \sum_{j=k+1}^{K} \frac{1}{j}\right) \tag{16}$$

is the expected value of the sum of the k largest of K independent and $\chi^2(2)$-distributed random variables. Finally, it follows from

$$\begin{aligned} ES^*(k) &= E\left(z - \sum_{j=1}^{k} \hat{A}_{q(j)} C_{q(j)} - \sum_{j=1}^{k} \hat{B}_{q(j)} S_{q(j)}\right)^T \left(z - \sum_{j=1}^{k} \hat{A}_{q(j)} C_{q(j)} - \sum_{j=1}^{k} \hat{B}_{q(j)} S_{q(j)}\right) \\ &= n\sigma^2 + E \frac{n}{2} \sum_{j=1}^{k} (\hat{A}_{q(j)}^2 + \hat{B}_{q(j)}^2) = \sigma^2 (n + \zeta_2(k, K)) \end{aligned} \tag{17}$$

that

$$FPE_{SUB}(k) = \frac{n+\zeta_2(k,K)}{n-\zeta_2(k,K)} S(k) \tag{18}$$

is an unbiased estimator for the mean squared prediction error (17).

2.3 Consistency

We study the probability of overfitting in the simplest case, where the residual sums of squares $S(0)$ and $S(1)$, respectively, of the two smallest models are compared under the assumption that $\mu = 0$. The probability that a criterion with a multiplicative penalty term h incorrectly prefers one pair over zero pairs is given by

$$P(S(1)h < S(0)) = P\left(\frac{1}{2\sigma^2}\left(S(0) - \frac{n}{2}\hat{R}^2_{q(K)}\right) h < \frac{1}{2\sigma^2} S(0)\right) = P\left(Z > \frac{S(0)}{\sigma^2}\frac{h-1}{2}\right) \approx P\left(Z > \underbrace{\frac{n(h-1)}{2}}_{=:H}\right), \tag{19}$$

where Z is the maximum of K independent and standard exponentially distributed random variables. For consistency it is required that this probability is very small, hence H must be much greater than EZ. It is therefore safe to assume that $H - EZ > 0$, which implies that

$$P(Z > H) = P(Z - EZ > H - EZ) \leq P(|Z - EZ| > H - EZ) \leq \frac{Var(Z)}{(H-EZ)^2}\,. \tag{20}$$

Using Rényi's representation (Rényi, 1953) to write Z as a linear combination

$$Z = \frac{e_1}{1} + \frac{e_2}{2} \ldots + \frac{e_K}{K} \tag{21}$$

of a random sample $e_1, \ldots, e_K$ from a standard exponential distribution, we obtain

$$Var(z) = 1 + \frac{1}{4} + \cdots + \frac{1}{K^2} < \frac{\pi}{6}\,, \tag{22}$$

hence the propability (20) vanishes if $H - EZ \to \infty$ or, equivalently, if

$$nh - n - 2EZ \to \infty. \tag{23}$$

A suitable choice for the multiplicative penalty term h is

$$h_\lambda = 1 + \frac{1}{n}\lambda\left(\log(K) + C + \frac{1}{2K}\right),\ \lambda > 2, \tag{24}$$

because

$$nh_\lambda - n - 2EZ = nh_\lambda - n - 2\left(1 + \frac{1}{2} + \cdots + \frac{1}{K}\right) \geq (\lambda - 2)\left(\log(K) + C + \frac{1}{2K}\right) \to \infty, \tag{25}$$

where $C = 0.5772156649\ldots$ is Euler's constant. The associated additive penalty term is given by

$$p_\lambda = \lambda\left(\log(K) + C + \frac{1}{2K}\right),\ \lambda > 2, \tag{26}$$

because

$$n\log(S(1)h_\lambda) = n\log\big(S(1)\big) + \log(1+\tfrac{p_\lambda}{n})^n \approx n\log\big(S(1)\big) + p_\lambda. \tag{27}$$

In the next section, the performance of this penalty term will be examined for various values of λ and compared with conventional penalty terms.

3. Empirical Results

In our empirical study, we use only stations with continuous monthly temperature measurements from January 1880 to January 2021. The homogenized station data (see Lenssen, 2019; GISTEMP Team, 2022) were downloaded from the website https://data.giss.nasa.gov/gistemp/ of the NASA Goddard Institute for Space Studies (GISS). Figure 1.a shows the annual means from 1880 to 2020 for the stations (i) Vardo, Norway (70.3670N, 31.1000E, elevation: 15m), (ii) Freudenstadt, Germany (48.4544N, 8.4100E, elevation: 797m), (iii) Kremsmünster, Austria (48.0500N, 14.1331E, elevation: 383m), (iv) Basel Binningen (47.5500N, 7.5831E, elevation: 316m), and (v) Valentia Observatory, Ireland (51.9394N, 10.2219W, elevation: 9m). Each time series exhibits an upward trend which is an indication of global warming (see, e.g., also Fomby and Vogelsang, 2002; Lai and Yoon, 2018; Chang et al., 2020; Mangat and Reschenhofer, 2020). To make this trend more visible, we plotted first the cumulative measurements in Figure 1.b and then the difference between the cumulative measurements $y_1, y_1+y_2, y_1+y_2+y_3, \ldots$ and a suitable linear trend $b, 2b, 3b, \ldots$ in Figure 1.c. The accumulation has a strong smoothing effect and the subtraction of the linear trend amplifies any possible deviation of the difference from linearity. The parameter b was obtained by averaging over the first 20 years. Analogous difference plots were produced separately for each calendar month. In each case, a faster than linear growth can be seen which is again strong evidence of an upward trend. Finally, Figure 3 shows for each station the average seasonal pattern, which must later in our evaluation study be correctly identified by the different criteria discussed in the previous section.

In the periodogram

$$I(\omega_j) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} y_t e^{-i\omega_j t}\right|^2, \tag{28}$$

obtained from the measurements over a period of 141 full years (January 1880 to December 2010), a seasonal pattern becomes apparent in the form of isolated peaks at the seasonal frequencies, which are in the case of monthly data given by $2\pi j/12, j = 1, \dots, 6$, where $2\pi j/12 = \omega_{141} = 141 \cdot 2\pi/n$ and $n = 12 \cdot 141 = 1692$. However, there is also a possible pole at frequency zero (see the first column of Figure 4), which is a further indication of the upward trend in the temperature. To get rid of this distracting feature, we take first differences. In order to retain full years, we add the January 2021 to the observation period. The observation period for the differenced series then starts in February 1880 and extends to January 2021. The second column of Figure 4 shows that the pole has indeed disappeared but the seasonal pattern is still there. The transformation has served its purpose. We are now ready to start the evaluation of the competing criteria.

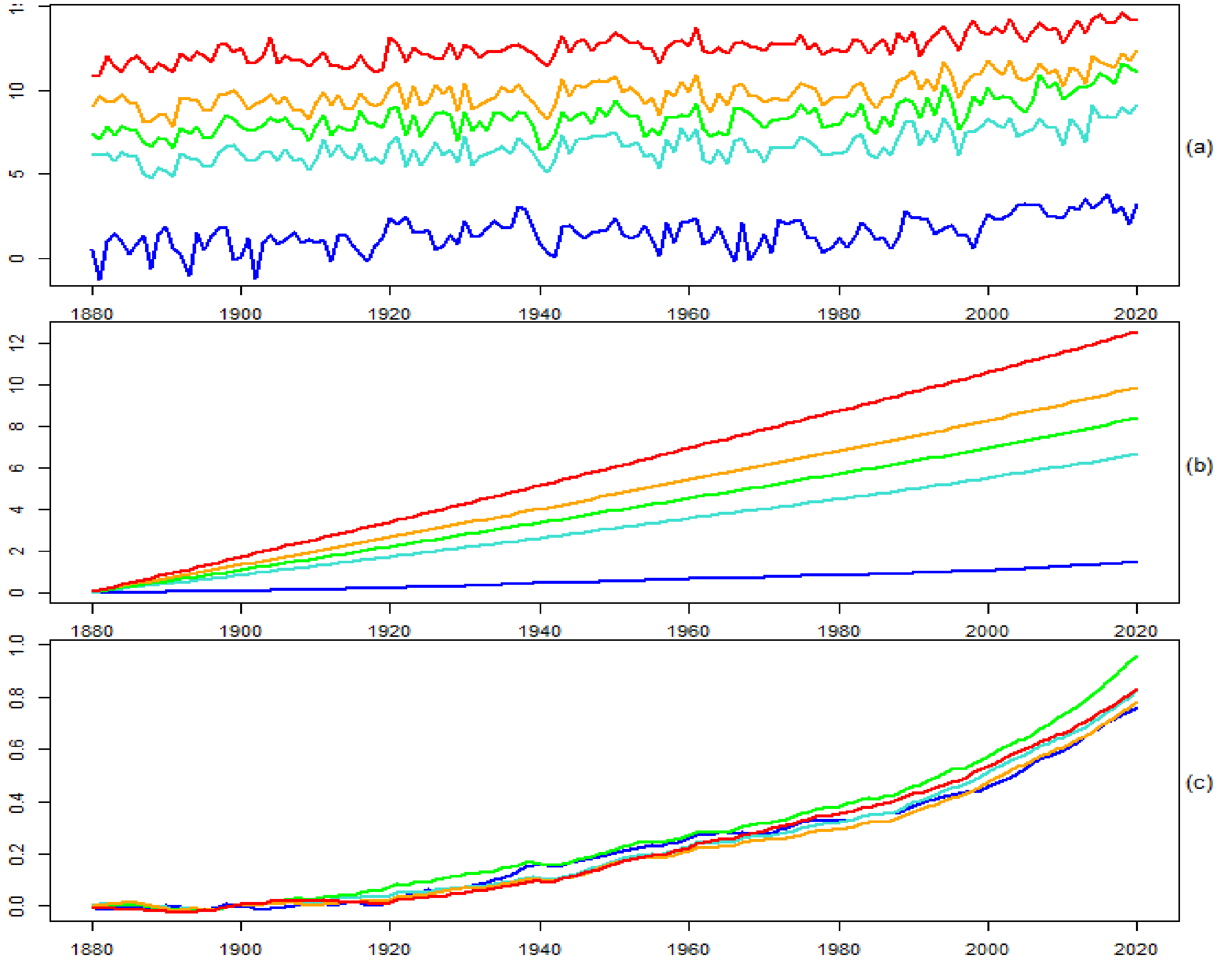

Figure 1.a Annual means from 1880 to 2020 for the stations (i) Vardo, Norway (blue) (ii) Freudenstadt, Germany (turquoise), (iii) Kremsmünster, Austria (green), (iv) Basel Binningen, Switzerland (orange), and (v) Valentia Observatory, Ireland (red). 1.b Cumulative time series. 1.c Linear trend subtracted

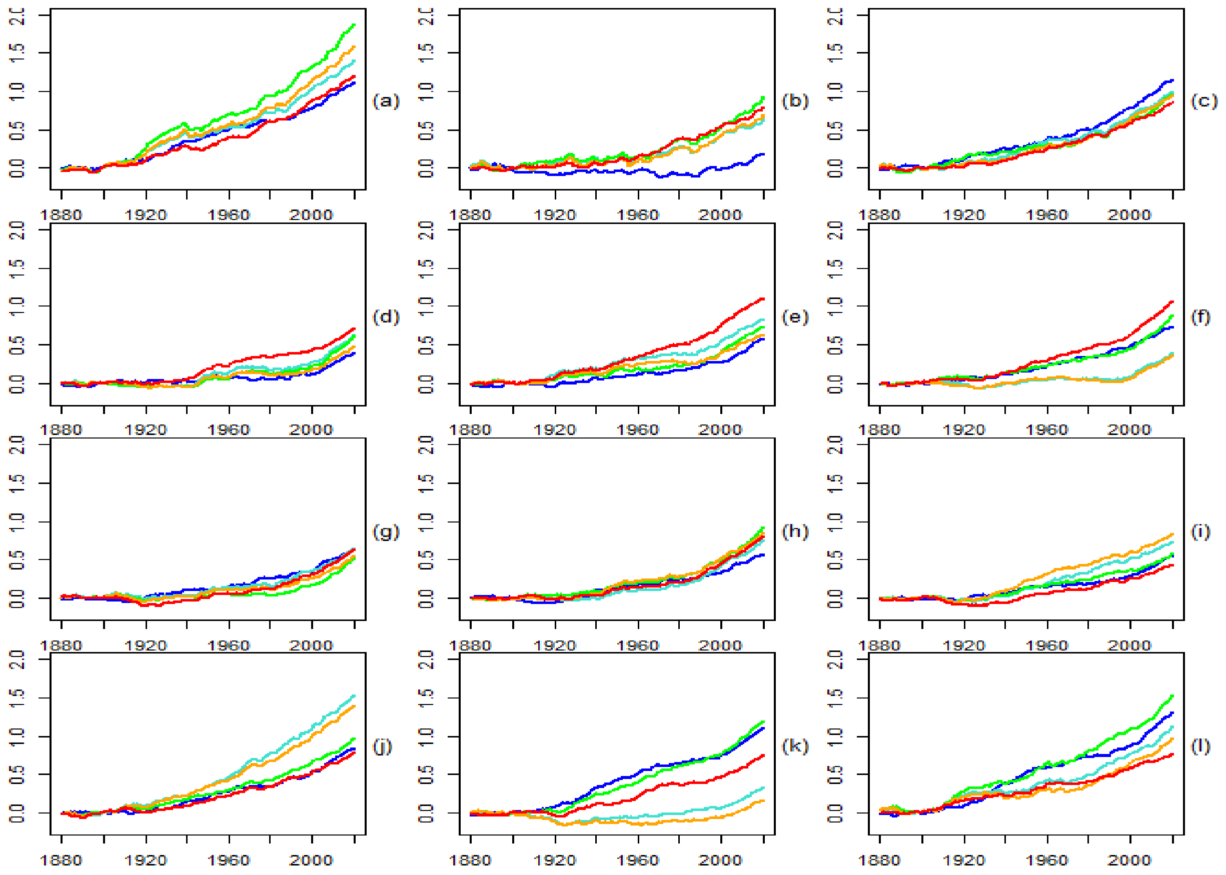

Figure 2: After the subtraction of a linear trend, the cumulative monthly temperature at Vardo (blue), Freudenstadt (turquoise), Kremsmünster (green), Basel Binningen (orange), and Valentia Observatory (red) is plotted separately for each calendar month (a: January, b: February, c: March, ...).

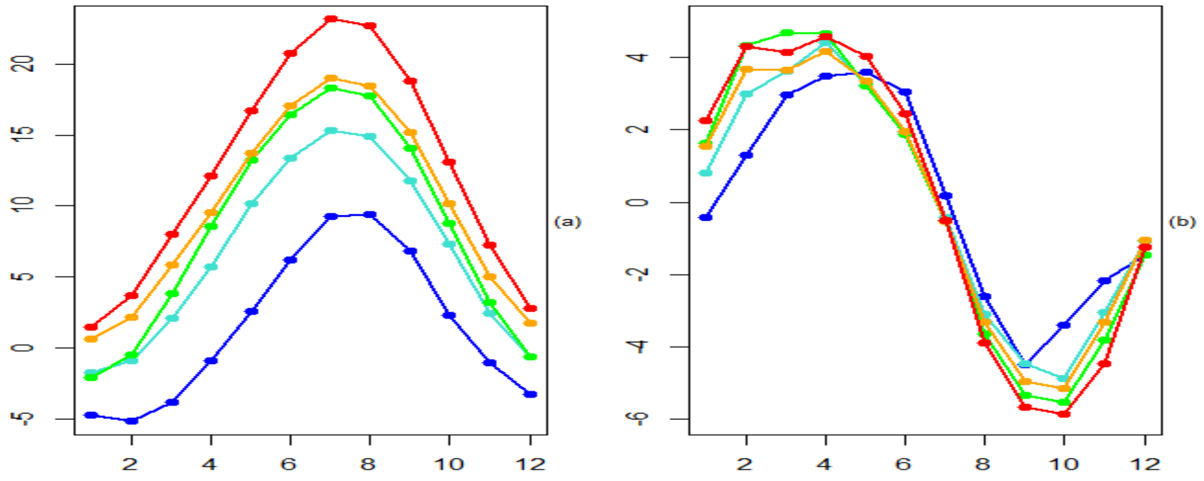

Figure 3: Mean seasonal pattern in (a) the monthly temperature and (b) the differenced monthly temperature at Vardo (blue), Freudenstadt (turquoise), Kremsmünster (green), Basel Binningen (orange), and Valentia Observatory (red).

In Subsection 2.2, we used $[(n-1)/2]$ instead of $[n/2]$ for the definition of the number K of Fourier frequencies, which excludes the seasonal frequency π. This decision was based on the fact that the distribution of the periodogram ordinate at this frequency has only one instead of two degrees of freedom because $\sin(\pi)=0$. Nevertheless, the frequency π has exceptionally been included in Figures 4. Obviously, this seasonal frequency does not explain an important part of the seasonal pattern anyway. In general, only the first two or three seasonal frequencies are really needed for the description of the seasonal pattern, which is due to the fact that this pattern does not deviate much from a simple sinusoidal shape (see Figure 3). An ideal criterion should therefore select only the pairs (C_j, S_j) corresponding to the first two or three seasonal frequencies and none of the other pairs. Because of the orthogonality of the cosine and sine vectors, we choose a stepwise approach, where the k-th pair will be selected if

$$n\log\big(S(k)\big)+p(k)<n\,log(S(k-1)). \tag{29}$$

The additive penalty term $p(k)$ can either be constant or depend on the sample size n, the total number K of available pairs or the number $K-k+1$ of the remaining pairs, which have not been chosen yet.

The criteria AIC and the BIC penalize the inclusion of an additional pair with the constant terms $p_{AIC}(k)=2\cdot 2$ and $p_{BIC}(k)=2\log(n)$, respectively. The additional factor 2 in these terms is due to the trivial fact that each pair consists of two explanatory variables. Anyhow, these two criteria penalize any pair in exactly the same way, regardless whether it is the only choice or it has been selected as the best from a large set of candidate pairs. In contrast, the criterion based on (25) explicitly penalizes overparametrization because its penalty term p_λ, which is given by (26), depends on the total number K of available pairs. However, this term may still be considered constant in the sense that it penalizes the inclusion of the first pair in the same way as the inclusion of the second or third pair. It does not take into account that in the second step there are only $K-1$ pairs left and in the third step only $K-2$. The criterion with penalty term

$$p_\zeta(k)=\ 2\,\zeta_2(1,K-k+1)=4\left(1+\sum_{j=1}^{K-k+1}\frac{1}{j}\right) \tag{30}$$

(Reschenhofer, 2015b) does just that. It looks at the number of the remaining pairs. The large-sample properties of the criteria utilizing K depend of course on the size of K, in particular on whether K increases as n increases and at which rate. In our application, $K \approx n/2$.

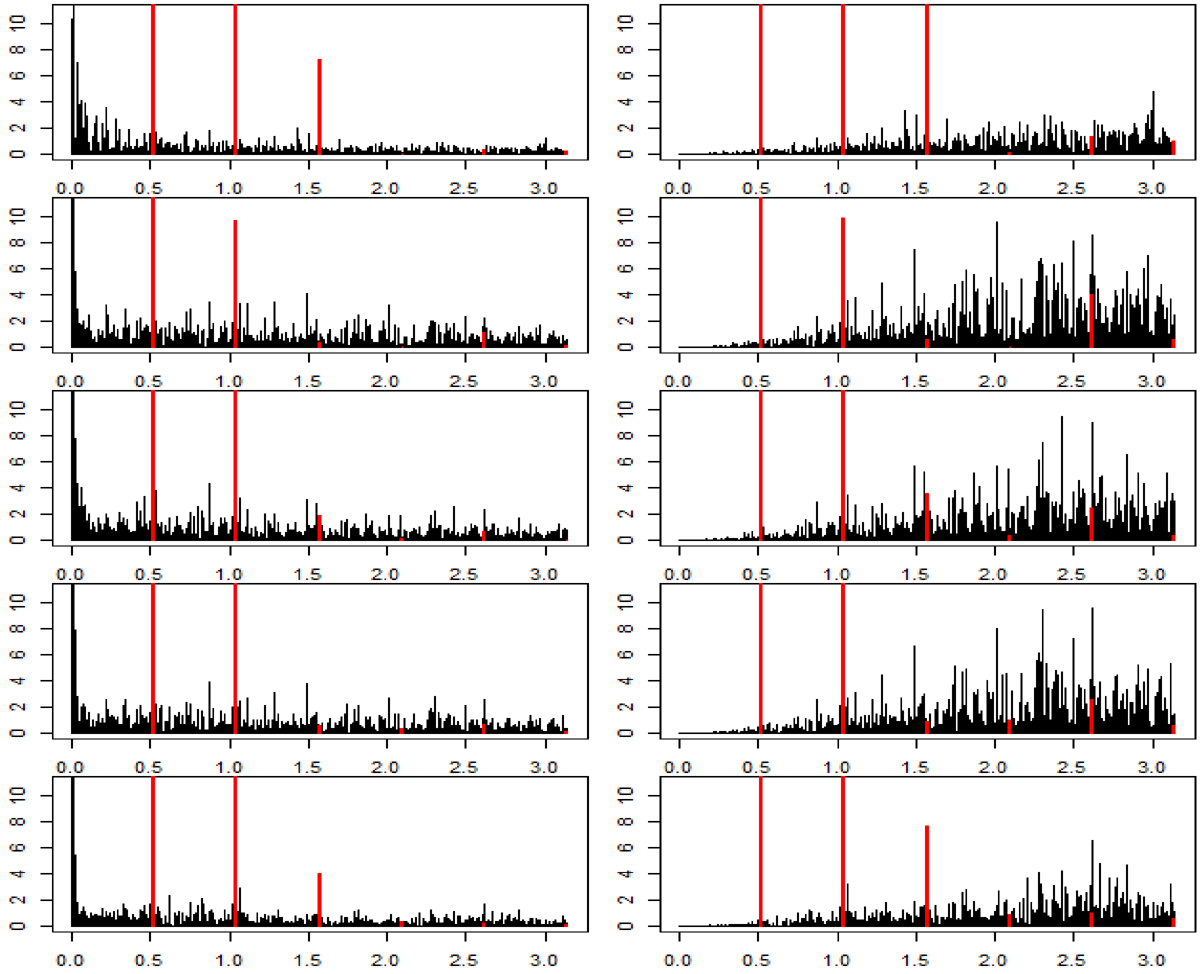

Figure 4: Periodogram of the monthly temperature from 1880-01 to 2020-12 (1st column) and periodogram of the first differences from 1880-02 to 2021-01 (2nd column) at the stations Vardo (1st row), Freudenstadt (2nd row), Kremsmünster (3rd row), Basel Binningen (4th row), and Valentia Observatory (5th row). The periodogram ordinates at the seasonal frequencies are marked red.

Table 1 shows for each of the five temperature datasets which of the seasonal frequencies (excl. π) and how many of the nonseasonal frequencies are selected by the different criteria when the whole observation period is used. Not surprisingly, the AIC selects far too many frequencies. It always selects 50 frequencies, which is the maximum allowed in our study. Since at most three frequencies are required for the description of the seasonal patterns, AIC makes at least 47 wrong decisions. BIC selects much fewer frequencies than AIC but still more than the other competing criteria which perform quite similarly. The harshest criterion is that with the penalty term (30). It is the only one which selects exactly those (green) frequencies which are absolutely necessary. When a milder criterion is desired, which selects also the optional (yellow) frequencies, the criterion with penalty term (26) and $\lambda = 2.5$ might be a good compromise.

Table 1: Selection of up to 50 pairs of cosine and sine vectors of the same frequency by seven automatic selection methods for the description of the seasonal patterns in five differenced temperature series (stations: Vardo, Freudenstadt, Kremsmünster, Basel Binningen, and Valentia Observatory). The colors green and yellow indicate which of the first five seasonal frequencies are absolutely required (green) or optional (yellow). Nonseasonal frequencies (orange) should not be selected ever.

AIC	BIC	$2\,\zeta_2$	$\lambda = 2.25$	$\lambda = 2.50$	$\lambda = 2.75$	$\lambda = 3.00$
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
0	0	0	0	0	0	0
0	0	0	0	0	0	0
47	3	0	1	1	1	0
1	1	1	1	1	1	1
1	1	0	1	1	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
1	0	0	0	0	0	0
47	5	0	3	1	0	0
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
47	5	0	3	2	2	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
48	5	0	4	3	2	0
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	0	1	1	1	1
0	0	0	0	0	0	0
0	0	0	0	0	0	0
47	8	0	4	3	1	1

4. Conclusion

In this paper, conditions for the pairwise selection of explanatory variables from a large set of pairs are derived and selection methods are proposed that satisfy these conditions. In an empirical study with long monthly temperature time series, these methods are evaluated and compared with conventional methods. In this study, the number of available pairs is equal to half of the length of the time series. The pairs are defined as cosine and sine vectors of the same frequency. Apart from a possible seasonal component, the spectrum of the time series is continuous. When the temperature time series are described by a linear combination of the cosine and sine vectors, the individual coefficients vanish at all frequencies as the length of the time series increases, but with the exception of the seasonal frequencies. It is therefore straightforward to assess the performance of the competing selection methods. Ideally, all seasonal frequencies required for the description of the seasonal pattern should be selected but none of the nonseasonal frequencies. The conventional selection criteria AIC and BIC fail to identify the seasonal pattern correctly. They select too many of the nonseasonal frequencies. Not surprisingly, this is particularly true for the AIC. But also the BIC does not sufficiently penalize overparametrization because it does not take the total number of available pairs into account. Only the new selection criteria manage to identify the seasonal pattern correctly.

References

[1] H. Akaike, Fitting autoregressive models for prediction. Ann. Inst. Stat. Math. 21 (1969) 243-247. https://doi.org/10.1007/BF02532251.

[2] H. Akaike, Information theory and an extension of the maximum likelihood principle, in B.N. Petrov, F. Csaki, Second international symposium on information theory, Akademia Kiado, Budapest (1973) 267-281. https://doi.org/10.1007/978-1-4612-1694-0_15.

[3] Y. Chang, R.K. Kaufmann, C.S. Kim, J.I. Miller, J.Y. Park, S. Park, Evaluating trends in time series of distributions: A spatial fingerprint of human effects on climate, J. Econ. 214 (2020) 274–94. https://doi.org/10.1016/j.jeconom.2019.05.014.

[4] J. Chen, Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, Biometrika 95 (2008) 759–771. https://doi.org/10.1093/biomet/asn034.

[5] T.B. Fomby, T. J. Vogelsang, The application of size-robust trend statistics to global-warming temperature series, J. Climate 15 (2002) 117–23. https://doi.org/10.1175/1520-0442(2002)015<0117:TAOSRT>2.0.CO;2.

[6] C.M. Hurvich, C.-L. Tsai, Regression and time series model selection in small samples, Biometrika 76 (1989) 297-307. https://doi.org/10.1093/biomet/76.2.297.

[7] D.P. Foster, E.I. George, The risk inflation criterion for multiple regression, Ann. Stat. 22 (1994) 1947-1975. http://www.jstor.org/stable/2242493.

[8] E.I. George, D.P. Foster, Calibration and empirical Bayes variable selection, Biometrika 87 (2000) 731-747. https://doi.org/10.1093/biomet/87.4.731.

[9] GISTEMP Team, GISS surface temperature analysis (GISTEMP), version 4, NASA Goddard Institute for Space Studies (2022). https://data.giss.nasa.gov/gistemp/.

[10] [10] Y. Kim, S. Kwon, H. Choi, Consistent model selection criteria on high dimensions, J. Mach. Learn. Res. 13 (2012) 1037-1057. https://www.jmlr.org/papers/v13/.

[11] K.S. Lai, M. Yoon, Nonlinear trend stationarity in global and hemispheric temperatures. Appl. Econ. Lett. 25 (20118) 15–18. https://doi.org/10.1080/13504851.2017.1290768.

[12] N. Lenssen, G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, D. Zyss, Improvements in the GISTEMP uncertainty model, J. Geophys. Res.: Atmosph. 124 (2019) 6307-6326. https://doi.org/10.1029/2018JD029522.

[13] M.K. Mangat, E. Reschenhofer, Frequency-domain evidence for climate change. Econometrics 8 (2020) 1-15. https://doi.org/10.3390/econometrics8030028.

[14] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (2013). https://www.R-project.org/.

[15] A. Rényi, On the theory of order statistics, Acta Math. Acad. Sci. Hung. 4 (1953) 191-231. https://doi.org/10.1007/BF02127580.

[16] E. Reschenhofer, Improved estimation of the expected Kullback-Leibler discrepancy in case of misspecification, Econ. Theory 15 (1999) 377-387. https://doi.org/10.1017/S0266466699153052.

[17] E. Reschenhofer, On subset selection and beyond. Adv. Appl. Stat. 4 (2004) 265–286. https://www.pphmj.com/abstract/436.htm.

[18] E. Reschenhofer, Discriminating between nonnested models. Far East J. Theor. Stat. 31 (2010) 117 -133. http://www.pphmj.com/abstract/4910.htm.

[19] E. Reschenhofer, Consistent variable selection in large regression models, J. Stat.: Adv. Theory Appl. 14 (2015a) 49-67. http://scientificadvances.co.in/abstract/4/174/977.

[20] E. Reschenhofer, Criteria for pairwise variable selection. SOP Transactions on Statistics and Analysis (2015b).

[21] D. Rothman, Letter to the editor, Technometrics 10 (1968) 432. https://doi.org/10.1080/00401706.1968.10490590.

[22] T. Sawa, Information criteria for discriminating among alternative regression models, Econometrica 46 (1978) 1273-1291. https://doi.org/10.2307/1913828.

[23] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461-464. https://doi.org/10.1214/aos/1176344136.

[24] N. Sugiura, Further analysts of the data by Akaike's information criterion and the finite corrections, Commun. Stat.-Theory Methods. 7 (1978) 13-26. https://doi.org/10.1080/03610927808827599.

[25] H. Wang, B. Li, C., Leng, Shrinkage tuning parameter selection with a diverging number of parameters, J. R. Stat. Soc. Ser. B. 71 (2009) 671-683. https://doi.org/10.1111/j.1467-9868.2008.00693.x.

[26] X. Zheng, W.-Y. Loh, A consistent variable selection criterion for linear models with high-dimensional covariates, Stat. Sin. 7 (1997) 311-325. http://www3.stat.sinica.edu.tw/statistica/j7n2/j7n24/j7n24.htm.