Model Selection by Sequentially Normalized Least Squares

Jorma Rissanen, Teemu Roos, Petri Myllymäki

Helsinki Institute for Information Technology HIIT

Abstract

Model selection by the predictive least squares (PLS) principle has been thoroughly studied in the context of regression model selection and autoregressive (AR) model order estimation. We introduce a new criterion based on sequentially minimized squared deviations, which are smaller than both the usual least squares and the squared prediction errors used in PLS. We also prove that our criterion has a probabilistic interpretation as a model which is asymptotically optimal within the given class of distributions by reaching the lower bound on the logarithmic prediction errors, given by the so called stochastic complexity, and approximated by BIC. This holds both when the regressor (design) matrix is non-random or determined by the observed data as in AR models. The advantages of the criterion include the fact that it can be evaluated efficiently and exactly, without asymptotic approximations, and importantly, there are no adjustable hyper-parameters, which makes it applicable to both small and large amounts of data.

Key words: linear regression, time series, model selection, order estimation, predictive least squares

1 Introduction

In this paper we are concerned with deriving a model selection criterion for a class of normal models $f(y^n \mid X_n; \sigma^2, \beta) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_1^n (y_t - \beta'\bar{x}_t)^2\right)$, induced by the regression equations

$$y_t = \beta'\bar{x}_t + \epsilon_t, \tag{1}$$

where the prime indicates transposition, $\beta' = (\beta(1), \ldots, \beta(k))$, with $k \in \mathbb{N}$. The deviations $(\epsilon_t)_{t=1}^n$ are taken as an i.i.d. sequence generated by a normal distribution of zero-mean and variance σ^2. The columns $\bar{x}_t = (x_{t,1}, \ldots, x_{t,k})'$

of real valued elements, defining the regressor matrices X_t, are either non-random, or $\bar{x}_t = (y_{t-1}, \ldots, y_{t-k})'$ as in AR models.

For each $t = 1, 2, \ldots n$, let $k(t)$ be the largest integer such that the least squares estimate $b_t = (b_{t,1}, \ldots, b_{t,k(t)})'$ can be uniquely solved. Hence, typically $k(t) = \min\{t, k\}$ except for AR models, where $k(t) = \min\{t-1, k\}$. We denote by m the smallest integer t such that $k(t) = k$; note that m defined this way depends on k. In the following we omit the dependency on k for notational convenience.

Central to this work are the following three representations of data for $t = 1, 2, \ldots n$, and $k \geq k(t)$:

$$y_t = b'_{t-1}\bar{x}_t + e_t = \sum_{i=1}^{k(t)} b_{t-1,i}x_{t,i} + e_t, \tag{2}$$

$$y_t = b'_n\bar{x}_t + \hat{\epsilon}_t(n) = \sum_{i=1}^{k(t)} b_{n,i}x_{t,i} + \hat{\epsilon}_t(n), \tag{3}$$

$$y_t = b'_t\bar{x}_t + \hat{e}_t = \sum_{i=1}^{k(t)} b_{t,i}x_{t,i} + \hat{e}_t. \tag{4}$$

All the representations split y_t into a sum of two terms, the first of which can be interpreted as a predicted value, and the second as an error or a residual. The representation differ only in terms of the way we define the parameter estimates. In the first one, the parameters are estimated based on the first $t-1$ observations; in the second one, all the n observations are used; and in the third one, the t first observations are used. We describe each representation in more detail below.

In fact, the second and the third representations can be defined even for $t \leq m$, in which case $k(t) < k$, since the orthogonal projection of y^n to the linear space spanned by the columns of X_t is always unique, even though the least squares estimate may not be; for any $1 \leq t \leq n$, the fitted value $b'_n\bar{x}_t$ is obtained as the t'th element of the projected vector. In practice, the solution can be obtained from

$$b'_n\bar{x}_t = \bar{x}'_t(X_nX'_n)^{-}X_ny^n,$$

where $(\cdot)^{-}$ denotes the pseudo-inverse. However, for the first representation (2), the prediction is not unique for $t \leq m$, and we will only apply it for $t = m+1, \ldots, n$.

The predictor $b'_{t-1}\bar{x}_t$ of y_t in the first case is called the 'plug-in' predictor, in which the parameters are calculated from the data available up to $t-1$. The plug-in model defines a conditional normal density function for $t > m$,

$$f(y_t \mid y^{t-1}, X_t\,;\, b_{t-1}, \hat{\sigma}^2_{t-1}) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{t-1}}} \exp\left(-\frac{e_t^2}{2\hat{\sigma}^2_{t-1}}\right),$$

where $\hat{\sigma}^2_{t-1} = \frac{1}{t-1}\sum_{i=1}^{t-1}\hat{\epsilon}_i^2(t-1)$, and $y^{t-1} = y_1, \ldots, y_{t-1}$. The resulting joint density function obtained by multiplying the conditional densities of y_{m+1}, $\ldots$, y_n, and ignoring constant terms, defines (by its negative logarithm) the so-called *predictive minimum description length* (PMDL) criterion, studied in [4], [6], [8], and [21]. Its special case for constant variance $\hat{\sigma}^2_{t-1} = \sigma^2$ is the *predictive least squares* (PLS) criterion,

$$\mathrm{PLS}(n,k) = \sum_{t=m+1}^{n} (y_t - b'_{t-1}\bar{x}_t)^2,$$

studied in [15] and [21].

The second representation (3) is the traditional least squares formulation. The predictions are the ones that minimize the sum of squared residuals, $\sum_{t=1}^{n}\hat{\epsilon}_t(n)^2$ over all predictions of the form $b'_n\bar{x}_t$ where the parameter vector b_n is the same for all $t = 1, \ldots, n$. Model selection criteria associated with (3) include AIC [1], and BIC [19],

$$\mathrm{BIC}(n,k) = \frac{n}{2}\log\hat{\sigma}^2_n + \frac{k+1}{2}\log n,$$

where $k+1$ is the number of parameters (including the variance). The BIC criterion is obtained by an approximation of a joint density function of the data where the negative logarithm of the maximized likelihood $f(y^n \mid X_n\,;\, b_n, \hat{\sigma}^2_n)$ determines the first term. In the AIC criterion the second term is $k+1$, the number of parameters. Both criteria are often multiplied by $2/n$, so that the first term is simply the logarithm of the residual sum of squares.

Also involving the second representation, the *normalized maximum likelihood* (NML) criterion is obtained directly as the normalized version of the maximized likelihood, where the normalizing term is given by $C_{n,k} = \int_{y^n \in \mathcal{Y}} f(y^n \mid X_n\,;\, b_n, \hat{\sigma}^2_n)\, dy^n$ [2], [16], [20]. In order to make the integral finite, the range of integration $\mathcal{Y}$ has to be restricted, which requires hyper-parameters. A solution which eliminates the effect of the hyper-parameters to model selection by a second normalization is presented in [17], see also [12, 18]. The corresponding parameter-free criterion is

$$\mathrm{NML}(n,k) = \frac{n-k}{2}\log\frac{\hat{\sigma}_n^2}{n-k} + \frac{k}{2}\log\frac{\hat{R}}{k} + \frac{1}{2}\log(k(n-k)),$$

where $\hat{R} = b_n' X_n X_n' b_n / n$. We also mention that a very similar construct as a Bayesian mixture, which also requires hyper-parameters for the prior, exists [7].

The third representation, which we are interested in, is new. In it, the prediction $b_t'\bar{x}_t$ is obtained by minimizing the sum of squared deviations $\sum_{i=1}^t \hat{e}_i^2$. The difference between this and the first represenation is that here we also include the information in the t'th data point. We show that the sum of squared deviations is smaller than either the sum of the traditional least squares $\sum_{i=1}^t \hat{\epsilon}_i^2(t)$, or the sum of the squared prediction errors $\sum_{i=1}^t e_i^2$. However, since the parameters of the corresponding conditional density function $f(y_t \mid y^{t-1}, X_t\,;\, b_t, \hat{\sigma}_t^2)$ involve at each step $t > m$ the response variable y_t, the density needs to be normalized in order to obtain a proper density function.

We study the asymptotic behavior of the resulting *sequentially normalized least squares* (SNLS) criterion for both fixed designs and random ones appearing in AR models. The criterion involves no approximations and is free of any hyper-parameters which tend to affect the outcome especially for small samples.

2 Sequentially normalized least squares

We start by showing that the new representation (4) achieves a better fit, in terms of the residuals, than traditional least squares representation (3). To see this, assume that the statement holds for sequences of length t, for some $t \geq 1$:

$$\hat{s}_t = \sum_{i=1}^t \hat{e}_i^2 \leq \sum_{i=1}^t \hat{\epsilon}_i^2(t) = t\hat{\sigma}_t^2. \tag{5}$$

For $t = 1$, the two representations are identical and the assumption is trivially satisfied. By induction, the claim holds for all $t \geq 1$:

$$\begin{aligned}\hat{s}_{t+1} &= \sum_{i=1}^t \hat{e}_i^2 + \hat{e}_{t+1}^2 \leq \sum_{i=1}^t \hat{\epsilon}_i^2(t) + \hat{e}_{t+1}^2 = \sum_{i=1}^t \hat{\epsilon}_i^2(t) + \hat{\epsilon}_{t+1}^2(t+1) \\ &\leq \sum_{i=1}^t \hat{\epsilon}_i^2(t+1) + \hat{\epsilon}_{t+1}^2(t+1) = (t+1)\hat{\sigma}_{t+1}^2,\end{aligned}$$

where the first inequality follows from the assumption (5), the second equality follows from the definitions (3) and (4) via $\hat{e}_{t+1} = y_{t+1} - b'_{t+1}\bar{x}_{t+1} = \hat{\epsilon}_{t+1}(t+1)$, and the second inequality follows from the definition of the least squares.

In order to obtain a meaningful model selection criterion with a capability to find a balance between goodness of fit and complexity, we convert the squared deviations into a density model.

Consider first the simple case where the variance σ^2 is fixed. The non-normalized conditionals

$$f(y_t \mid y^{t-1}, X_t; \sigma^2, b_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}\right), \tag{6}$$

are obtained by replacing the parameter vector β in the conditional normal density function $f(y_t \mid y^{t-1}, X_t; \sigma^2, \beta)$ by the least squares estimate b_t.

For each fixed k, for $t > m$, where m is the smallest value for t for which $k(t) = k$, the well known recursions exist, see for instance [10],

$$\begin{aligned} b_t &= V_t \sum_{j=1}^{t} \bar{x}_j y_j = b_{t-1} + V_{t-1}\bar{x}_t(y_t - \bar{x}'_t b_{t-1})/(1 + c_t) && (7)\\ V_t &= (X_t X'_t)^{-1} = V_{t-1} - V_{t-1}\bar{x}_t\bar{x}'_t V_{t-1}/(1 + c_t) && (8)\\ c_t &= \bar{x}'_t V_{t-1}\bar{x}_t \\ d_t &= \bar{x}'_t V_t \bar{x}_t \\ 1 - d_t &= 1/(1 + c_t). && (9) \end{aligned}$$

The last equality was shown in [8] and [21] with the interpretation that the quantity $1 - d_t$ is the ratio of the (Fisher) information in the first $t-1$ observations relative to all the t observations, [21]. This also implies that $0 \le d_t \le 1$.

By (7) we obtain

$$\begin{aligned} \hat{y}_t &= \bar{x}'_t \left[V_{t-1}\bar{x}_t(y_t - \bar{x}'_t b_{t-1})/(1 + c_t) + b_{t-1}\right] \\ &= c_t/(1 + c_t)(y_t - \bar{x}'_t b_{t-1}) + \bar{x}'_t b_{t-1} \\ &= (1 - d_t)\bar{x}'_t b_{t-1} + d_t y_t. && (10) \end{aligned}$$

which is a weighted average of the plug-in prediction $\bar{x}'_t b_{t-1}$ and the true value y_t. This gives the remaining error as

$$\hat{e}_t = y_t - \hat{y}_t = (1 - d_t)(y_t - \bar{x}'_t b_{t-1}) = (1 - d_t)e_t, \tag{11}$$

which is seen to be smaller than the plug-in prediction error by a constant factor. The normalization of (6) is straightforward, and the result is a normal density function, the mean given by the plug-in predictor and the variance by $\tau = (1 + c_t)^2\sigma^2$.

If we in (6) replace the variance by the minimized variance $\hat{s}_t/t$ and try to normalize the result the normalizing integral will be infinite. To make it finite would require hyper-parameters. Consider instead the maximization problem

$$\max_{\sigma^2} \prod_{t=m+1}^{n} f(y_t \mid y^{t-1}, X_t; \sigma^2, b_t). \tag{12}$$

The maximizing σ^2 is

$$\hat{\tau}_n = \frac{\hat{s}_n - \hat{s}_m}{n - m} = \frac{1}{n - m} \sum_{t=m+1}^{n} \hat{e}_t^2,$$

which gives the maximized product $(2\pi e\hat{\tau}_n)^{-(n-m)/2}$. By dropping the constants we get the non-normalized and normalized conditional density functions

$$\begin{aligned} f(y_t \mid y^{t-1}, X_t) &= \frac{\hat{\tau}_t^{-(t-m)/2}}{\hat{\tau}_{t-1}^{-(t-m-1)/2}} = \hat{\tau}_{t-1}^{-1/2}\left(1 + \frac{(y_t - \hat{y}_t)^2}{\hat{\tau}_{t-1}}\right)^{-(t-m)/2} \\ \hat{f}(y_t \mid y^{t-1}, X_t) &= K^{-1}(y^{t-1})\hat{\tau}_{t-1}^{-1/2}\left(1 + \frac{(y_t - \hat{y}_t)^2}{\hat{\tau}_{t-1}}\right)^{-(t-m)/2} \\ K(y^{t-1}) &= \hat{\tau}_{t-1}^{-1/2} \int_{-\infty}^{\infty} \left(1 + \frac{(y_t - \hat{y}_t)^2}{\hat{\tau}_{t-1}}\right)^{-(t-m)/2} dy_t. \end{aligned}$$

To get the normalizing integral we substitute (11), which gives

$$K(y^{t-1}) = \hat{\tau}_{t-1}^{-1/2} \int_{-\infty}^{\infty} \left[1 + \frac{(1 - d_t)^2}{\hat{\tau}_{t-1}}(y - \bar{x}_t' b_{t-1})^2\right]^{-(t-m)/2} dy. \tag{13}$$

By change of variables

$$z = \frac{1 - d_t}{\sqrt{\hat{\tau}_{t-1}}}\,(y - \bar{x}_t' b_{t-1})$$

we get

$$K(y^{t-1}) = K_{t-1} = \frac{1}{1-d_t} \int\limits_{-\infty}^{\infty} (1+z^2)^{-(t-m)/2} dz$$
$$= \frac{\sqrt{\pi}}{1-d_t} \Gamma\left(\frac{t-m-1}{2}\right) / \Gamma\left(\frac{t-m}{2}\right),$$

the last equality by the fact that z is seen to have Student's z-distribution. We need $t > m+1$ to make the normalizer non-zero.

For $t > m+1$, the conditional density function is then given by

$$\hat{f}(y_t \mid y^{t-1}, X_t) = \frac{\hat{\tau}_{t-1}^{-1/2}}{K_{t-1}} \left(1 + \frac{(1-d_t)^2}{\hat{\tau}_{t-1}} (y_t - \bar{x}_t' b_{t-1})^2\right)^{-(t-m)/2} \tag{14}$$
$$= K_{t-1}^{-1} \frac{\hat{\tau}_t^{-(t-m)/2}}{\hat{\tau}_{t-1}^{-(t-m-1)/2}}.$$

We see that again the predictor that maximizes the conditional density function is the plug-in predictor $\bar{x}_t' b_{t-1}$.

By putting the initial density function as some prespecified function $q(y^{m+1} \mid X_{m+1})$, which will not play a role in comparison of different models, we get the desired parameter-free density function

$$\hat{f}(y^n \mid X_n) = q(y^{m+1} \mid X_{m+1}) \prod_{t=m+2}^{n} \hat{f}(y_t \mid y^{t-1}, X_t).$$

Ignoring the initial density, the negative logarithm of the remaining part is given by

$$-\ln \prod_{t=m+2}^{n} \left(\frac{\sqrt{\pi}}{1-d_t} \Gamma\left(\frac{t-m-1}{2}\right) / \Gamma\left(\frac{t-m}{2}\right)\right)^{-1} \frac{\hat{\tau}_t^{-(t-m)/2}}{\hat{\tau}_{t-1}^{-(t-m-1)/2}},$$

where we note that both the Gamma functions and the $\hat{\tau}$'s telescope. Thus, we get the following simplified expression, which we call the *sequentially normalized least squares* (SNLS) criterion:

$$\mathrm{SNLS}(n,k) = \frac{n-m}{2} \ln \hat{\tau}_n - \frac{1}{2} \ln \hat{e}_{m+1} - \ln \frac{\Gamma\left(\frac{n-m}{2}\right)}{\Gamma(1/2)} + \ln \prod_{t=m+2}^{n} \frac{\sqrt{\pi}}{1-d_t}$$
$$= \frac{n-m}{2} \ln(2\pi e \hat{\tau}_n) + \sum_{t=m+1}^{n} \ln(1+c_t) + \frac{1}{2} \ln n + O(1), \tag{15}$$

where Stirling's formula was applied on the second row to the Gamma function, and terms independent of n are implicit in the $O(1)$ term. The SNLS criterion can be used for subset selection and order estimation for both small and large data sets. One of its distinguished properties is the fact that unlike the regular NML universal model it has no hyper-parameters.

We conclude this section by a large data set behavior of the SNLS model.

Theorem 1 *If the regressor variables $\bar{x}_t$ satisfy*

$$\frac{1}{n} X_n X_n' = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i \bar{x}_i' \to \Sigma \tag{16}$$

with Σ non-singular, then

$$\mathrm{SNLS}(n,k) = \frac{n-m}{2} \ln(2\pi e \hat{\tau}_n) + \left(\frac{2k+1}{2}\right) \ln n + o(\ln n). \tag{17}$$

Proof. Use (16) to get $V_t \to t^{-1}\Sigma^{-1}$, so that $c_t = O(1/t)$, and $\ln(1+c_t) = c_t + O(1/t^2)$. By the first of the following results, derived in [15] and [21],

$$\sum_{t=m+1}^{n} c_t = k \ln n + o(\ln n) \tag{18}$$

$$\sum_{t=m+1}^{n} d_t = k \ln n + o(\ln n). \tag{19}$$

we deduce (17). □

3 Fixed regression matrix

The first theorem shows the mean square deviations in the three representations of data (2), (3), and (4), which are of some interest, and which we will need later on. Since we need the recursive formulas (7), (8), (9) we give the results for $t > m$.

Theorem 2 *If the regressor variables are non-random satisfying (16) and the data generated by (1), then*

$$\frac{1}{n-m} \sum_{t=m+1}^{n} \mathbb{E} e_t^2 = \sigma^2 \left(1 + \frac{1}{n-m} \sum_{t=m+1}^{n} c_t\right) \tag{20}$$

$$\frac{1}{n-m}\sum_{t=m+1}^{n}\mathbb{E}\hat{e}_t^2 = \sigma^2\left(1 - \frac{1}{n-m}\sum_{t=m+1}^{n} d_t\right) \tag{21}$$

$$\frac{1}{n-m}\left(\sum_{t=1}^{n}\mathbb{E}\hat{\epsilon}_t^2(n) - \sum_{t=1}^{m}\mathbb{E}\hat{\epsilon}_t^2(m)\right) = \sigma^2, \tag{22}$$

where the expectation is with the parameters β and σ.

Proof. To obtain (20) we start with $y_i = \bar{x}_i'\beta + \epsilon_i$. Since $(\epsilon_i)_{i=1}^n$ is a zero-mean i.i.d. sequence, the estimate b_{t-1} is independent of ϵ_t. Thus, we get

$$\begin{aligned}\mathbb{E}e_t^2 = \mathbb{E}[(y_t - b_{t-1}'\bar{x}_t)^2] &= \mathbb{E}[((y_t - \beta'\bar{x}_t) + (\beta'\bar{x}_t - b_{t-1}'\bar{x}_t))^2] \\ &= \mathbb{E}[(y_t - \beta'\bar{x}_t)^2] + \mathbb{E}[(\beta'\bar{x}_t - b_{t-1}'\bar{x}_t)^2]. \qquad (23)\end{aligned}$$

Using the well known result on the covariance matrix of the least squares estimates, the latter expectation in (23) becomes

$$\mathbb{E}[((\beta - b_{t-1})'\bar{x}_t)^2] = \bar{x}_t'\mathbb{E}[(\beta - b_{t-1})(\beta - b_{t-1})']\bar{x}_t = \bar{x}_t'(\sigma^2 V_{t-1})\bar{x}_t = c_t\sigma^2.$$

The former expectation in (23) clearly equals σ^2, and thus, $\mathbb{E}e_t^2 = (1 + c_t)\sigma^2$ for all $t > m$, and Eq. (20) follows.

Equation (21) follows from this by (11), and

$$\mathbb{E}\hat{e}_t^2 = (1 - d_t)^2\mathbb{E}e_t^2 = \frac{c_t\sigma^2}{(1 - d_t)^2} = (1 - d_t)\sigma^2.$$

To prove the remaining statement, Eq. (22), we use the important equality (2.6) in [21]:

$$\sum_{t=m+1}^{n} e_t^2 = \sum_{t=1}^{n}\hat{\epsilon}_t^2(n) - \sum_{t=1}^{m}\hat{\epsilon}_t^2(m) + \sum_{t=m+1}^{n} d_t e_t^2, \tag{24}$$

which implies that the expected difference in the least squares is given by

$$\sum_{t=m+1}^{n}(1 - d_t)\mathbb{E}e_t^2 = \sum_{t=m+1}^{n}(1 - d_t)(1 + c_t)\sigma^2 = (n - m)\sigma^2,$$

which implies the claim. □

The next theorem shows the asymptotic optimality of the SNLS model in terms of logarithmic prediction errors, see [14], both in the mean and almost surely, in the case where the regressor matrix is fixed.

Theorem 3 *Let the assumption (16) hold, and let the data be generated by (1). Then*

$$\mathbb{E}\,\mathrm{SNLS}(n,k) = \frac{n-m}{2}\ln(2\pi e\sigma^2) + \frac{k+1}{2}\ln n + o(\ln n), \tag{25}$$

for almost all parameters b and σ. Also,

$$\mathrm{SNLS}(n,k) = \frac{n-m}{2}\ln(2\pi e\sigma^2) + \frac{k+1}{2}\ln n + o(\ln n) \tag{26}$$

almost surely.

Before giving the proof of the theorem, we elaborate on the the meaning of *optimality* in the sense of the theorem. The idea is that for any criterion which is a negative logarithm of a probability density, say $-\ln f(y^t \mid X_t)$, we can extract a sequence of probabilistic predictions as the conditionals $f(y_t \mid y^{t-1}, X_t) = f(y^t \mid X_t)/f(y^{t-1} \mid X_t)$, $t = m+1, \ldots, n$, where f is the corresponding density function. This includes criteria such as PLS and NML. The BIC criterion is included asymptotically since it is an approximation of a (mixture) probability density, see [19], but AIC is not. Since our SNLS criterion is derived via a probability density, it is included in this class. We further note that both PLS and SNLS are applicable in the so called *on-line* prediction setting where the sample size need not be determined in advance, while NML is only applicable in the so called *batch* scenario, where the sample size is fixed in advance.

For any criterion corresponding to a density function (BIC, PLS, NML, SNLS), Theorem 1 in [14] gives a lower bound on the expectation of the sum of logarithmic errors, $\mathbb{E}\sum_{t=m+1}^{n} -\ln f(y_t \mid y^{t-1}, X_t)$, where f is again the density function used for prediction; note that the sum is completely determined by the criterion and vice versa:

$$\mathbb{E}\sum_{m+1}^{n} -\ln f(y_t \mid y^{t-1}, X_t) = \mathbb{E}\; -\ln f(y^t \mid X_t).$$

The bound cannot be beaten except for some data generating parameters which belong to a set of Lebesgue measure zero. It has a fundamental role in universal prediction, see [9], and also in statistical inference by means of the Minimum Description Length (MDL) principle [13, 18].

In the Gaussian case, the bound is given by

$$\mathbb{E}\sum_{t=m+1}^{n} -\ln f(y_t \mid y^{t-1}, X_t) \geq nH(\sigma^2) + \frac{k+1}{2}\ln n + o(\ln n),$$

where $H(\sigma^2) = \frac{1}{2}\ln(2\pi e\sigma^2)$ is the differential entropy of the Gaussian density, see [3]. Whenever the logarithmic prediction errors of a given model match this bound, the model is called *optimal.* By Eq. (25), SNLS is optimal — the difference between $n-m$ and n in the first term is insignificant compared to the $o(\ln n)$ remainder term.

Proof (of Theorem 3). To prove (25) take the mean in (17) and exchange the mean and the logarithm on the right hand side. We get by Jensen's inequality

$$\mathbb{E}\,\mathrm{SNLS}(n,k) \le \frac{n-m}{2}\ln(2\pi e\mathbb{E}\hat{\tau}_n) + \left(\frac{2k+1}{2}\right)\ln n + o(\ln n).$$

Substituting (21) and applying (19) we then conclude that

$$\mathbb{E}\,\mathrm{SNLS}(n,k) \le \frac{n-m}{2}\ln(2\pi e\sigma^2) + \frac{k+1}{2}\ln n + o(\ln n). \tag{27}$$

By Theorem 1 in [14] the opposite inequality holds for all data generating parameters except some in a set of Lebesgue measure zero, see Remark 1, and (25) holds.

The proof of the a.s. result (26) is an exercise in martingales; in fact, Problem 15, page 165, in [11]. Define $\xi_t = \hat{e}_t^2 - (1-d_t)\sigma^2$, and $s_n = \sum_{t=m+1}^{n} \xi_t/t$, which is a martingale. We have

$$\mathbb{E}s_n^2 = \sum_{t=m+1}^{n} \mathbb{E}\xi_t^2/t^2 + 2\sum_{i,j\,:\,m<i<j} \frac{\mathbb{E}\xi_i\xi_j}{ij}.$$

Since $\mathbb{E}\xi_i\xi_j = \mathbb{E}\left[\mathbb{E}_{j|i}[\xi_i\xi_j]\right] = \mathbb{E}\xi_i\mathbb{E}_{j|i}[\xi_j]$, where $\mathbb{E}_{j|i}$ denotes the conditional expectation, we have $\mathbb{E}\xi_i\xi_j = \mathbb{E}\xi_i\cdot 0 = 0$, and the second term equals zero. Since $\mathbb{E}\xi_t^2$ is uniformly bounded so are both $\mathbb{E}s_n^2$ and $\mathbb{E}|s_n|$. By Doob's martingale convergence theorem s_n converges a.s. to a finite limit. By Kronecker's lemma, this implies that

$$S_n = \frac{1}{n-m}\sum_{t=m+1}^{n}\xi_t = \frac{1}{n-m}\sum_{t=m+1}^{n}\hat{e}_t^2 - \frac{\sigma^2}{n-m}\sum_{t=m+1}^{n}(1-d_t) \to 0 \quad \text{a.s.},$$

which we write as

$$\hat{\tau}_n = \frac{\sigma^2}{n-m}\sum_{t=m+1}^{n}(1-d_t) + o(1) \quad \text{a.s.}$$

Further

$$\ln \hat{\tau}_n = \ln \sigma^2 + \ln \left[1 - \sum_{t=m+1}^{n} d_t/(n-m) \right] + o(1) \quad \text{a.s.},$$

so that by Taylor expansion and (19)

$$\ln \hat{\tau}_n = \ln \sigma^2 - \frac{k}{n-m} \ln n + o(1) \quad \text{a.s.},$$

which, together with Thm. 1, concludes the proof. □

4 AR models

We then consider the case where the data are generated by an AR model,

$$y_t = \sum_{i=1}^{k} \beta(i) y_{t-i} + \epsilon_t, \ t \geq 1, \tag{28}$$

in which the regressor matrix is random, determined by the the data y^n, and where the coefficients are again given by the parameter vector $\beta = (\beta(1), \ldots, \beta(k))$.

The following theorem shows the almost sure asymptotic optimality of the SNLS model, in the sense explained in Remark 1 above, also in this case.

Theorem 4 *Let the data be generated by an AR model (28), where the roots of the polynomial $1 - \sum_{i=1}^{k} \beta(i) z^i$ are outside the unit circle, and ϵ_t is an i.i.d. zero-mean Gaussian process with variance σ^2. The process is also assumed to be ergodic and stationary with $\mathbb{E}\bar{x}_t \bar{x}_t' = \Sigma$ nonsingular. Then for $\hat{\sigma}_n^2 = (1/n) \sum_{i=1}^{n} \hat{\epsilon}_i^2(n)$, we have*

$$\ln \hat{\tau}_n = \ln \hat{\sigma}_n^2 - \left(\frac{k}{n-m} \ln n \right) (1 + o(1)) \quad \textit{a.s.}, \tag{29}$$

and

$$\mathrm{SNLS}(k,n) = \frac{n-m}{2} \ln(2\pi e \hat{\sigma}_n^2) + \frac{k+1}{2} \ln n + o(\ln n) \quad \textit{a.s.}$$

Proof. The proof takes advantage of the proof of the asymptotic optimality of the predictive model (2) in [21]. The beginning point is the equality (24). It gives

$$\ln \frac{1}{n} \sum_{t=m+1}^{n} e_t^2 = \ln \hat{\sigma}_n^2 + \ln \left[1 + \frac{\sum_{t=m+1}^{n} d_t e_t^2}{n\hat{\sigma}_n^2} - \frac{\sum_{t=1}^{m} \hat{\epsilon}_t^2(m)}{n\hat{\sigma}_n^2} \right]. \tag{30}$$

On the other hand, by Corollary 4.2.1 in [21]

$$\ln \frac{1}{n} \sum_{t=m+1}^{n} e_t^2 = \ln \hat{\sigma}_n^2 + \left(\frac{k}{n} \ln n \right) (1 + o(1)) \quad \text{a.s.} \tag{31}$$

Hence

$$\ln \left[1 + \frac{\sum_{t=m+1}^{n} d_t e_t^2 - m\hat{\sigma}_m^2}{n\hat{\sigma}_n^2} \right] = \left(\frac{k}{n} \ln n \right) (1 + o(1)) \quad \text{a.s.,} \tag{32}$$

where $m\hat{\sigma}_m^2 = \sum_{t=1}^{m} \hat{\epsilon}_t^2(m)$.

Since the right hand side of (32) vanishes, we have

$$\frac{\sum_{t=m+1}^{n} d_t e_t^2 - m\hat{\sigma}_m^2}{n\hat{\sigma}_n^2} \to 0 \quad \text{a.s.}$$

Thus,

$$\ln \left[1 + \frac{\sum_{t=m+1}^{n} d_t e_t^2 - m\hat{\sigma}_m^2}{n\hat{\sigma}_n^2} \right] = \frac{\sum_{t=m+1}^{n} d_t e_t^2 - m\hat{\sigma}_m^2}{n\hat{\sigma}_n^2} (1 + o(1)) \quad \text{a.s.,}$$

which by (31) gives

$$\sum_{t=m+1}^{n} d_t e_t^2 = \hat{\sigma}_n^2 \left(k \ln n \right) (1 + o(1)) + m\hat{\sigma}_m^2 \quad \text{a.s.} \tag{33}$$

From (11),

$$\sum_{t=m+1}^{n} \hat{e}_t^2 = \sum_{t=m+1}^{n} e_t^2 - 2 \sum_{t=m+1}^{n} d_t e_t^2 + \sum_{t=m+1}^{n} d_t^2 e_t^2$$

which with (24) and (33) gives

$$\sum_{t=m+1}^{n} \hat{e}_t^2 = n\hat{\sigma}_n^2 - m\hat{\sigma}_m^2 - \sum_{t=m+1}^{n} d_t e_t^2 + \sum_{t=m+1}^{n} d_t^2 e_t^2$$

$$= n\hat{\sigma}_n^2 - \hat{\sigma}_n^2(k\ln n)(1+o(1)) + \sum_{t=m+1}^{n} d_t^2 e_t^2 - 2m\hat{\sigma}_m^2 \quad \text{a.s.}$$

After we show that

$$\sum_{t=m+1}^{n} d_t^2 e_t^2 = o(\ln n) \quad \text{a.s.}, \tag{34}$$

we finally get

$$\frac{1}{n}\sum_{t=m+1}^{n} \hat{e}_t^2 = \hat{\sigma}_n^2 \left[1 - \left(\frac{k}{n}\ln n\right)(1+o(1))\right] \quad \text{a.s.},$$

and, since $n/(n-m) = 1 + o(1/n)$, also

$$\hat{\tau}_n = \frac{1}{n-m}\sum_{t=m+1}^{n} \hat{e}_t^2 = \hat{\sigma}_n^2 \left[1 - \left(\frac{k}{n-m}\ln n\right)(1+o(1))\right] \quad \text{a.s.},$$

which implies the first claim, (29), by Taylor expansion. By ergodicity, (16) holds, so that the second claim in the theorem follows from (17) and (29).

It now remains to prove (34). We first show that $\bar{x}_t'\bar{x}_t \leq \alpha \ln t$ almost surely for all but finitely many t, with α large enough.

The density function for $\bar{x}_t$ is Gaussian

$$f(\bar{x}_t) = \frac{|\Sigma|^{1/2}}{(2\pi)^{k/2}} e^{-\frac{1}{2}\bar{x}_t'\Sigma^{-1}\bar{x}_t},$$

where by stationarity $\Sigma = \mathbb{E}_a \bar{x}_t\bar{x}_t'$. For $\lambda_{\min}$ the least eigenvalue of Σ,

$$f(\bar{x}_t) \leq \frac{|\Sigma|^{1/2}}{(2\pi)^{k/2}} e^{-\frac{\lambda_{\min}}{2}\bar{x}_t'\bar{x}_t}.$$

Let $A_t = \{\bar{x}_t : \bar{x}_t'\bar{x}_t \geq \alpha \ln t\}$. Then

$$P(A_t) \leq \frac{|\Sigma|^{1/2}}{(2\pi)^{k/2}} \sum_{i \geq t} \int_{B_i} e^{-\frac{\lambda_{\min}}{2}\bar{x}_i'\bar{x}_i} d\bar{x}_i,$$

where $d\bar{x}_i$ is the differential volume and $B_i = \{\bar{x}_t : \alpha \ln i \leq \bar{x}_t'\bar{x}_t \leq \alpha \ln(i+1)\}$. The integrand is upper bounded by $i^{-\gamma}$ for $\gamma = \alpha\lambda_{\min}/2$, which remains constant on the surface of the k-dimensional sphere of radius $\alpha \ln i$. Hence, the integration of the surface area over the radius difference $\alpha \ln(1+1/i) = O(\alpha/i)$ gives $i^{-\gamma}$ times the volume of B_i, or

$$i^{-\gamma} \int\limits_{\alpha \ln i \leq r \leq \alpha \ln(i+1)} dr \leq O(i^{-(1+\gamma)}(\ln i)^{k-1}).$$

The sum of this from $i = t$ to $i = \infty$ is upper bounded by $O(\int_t^\infty y^{-\gamma} dy) = O(t^{-(\gamma-1)})$ for $\gamma > 2$, which can be satisfied by making α sufficiently large, which implies $\sum_{t=1}^\infty P(A_t) < \infty$. The claim follows by Borel-Cantelli lemma, namely, that the probability of the event that $\bar{x}_t' \bar{x}_t \geq \alpha \ln t$ infinitely often is zero.

By the ergodic theorem, (16) holds, and $V_t \to t^{-1}\Sigma^{-1}$ almost surely, giving

$$d_t \leq \frac{\lambda_{\min} \bar{x}_t' \bar{x}_t}{t} \leq \frac{\lambda_{\min} \alpha \ln t}{t} = \frac{O(\ln t)}{t} \quad \text{a.s.,} \tag{35}$$

where $\lambda_{\min}$ is again the least eigenvalue of Σ. Since by (33), for any $s > m$, $\sum_{t=s}^n d_t e_t^2 = O(\ln(n/s))$, we have

$$\sum_{t=m+1}^{n} d_t^2 e_t^2 = \sum_{t=m+1}^{s-1} d_t^2 e_t^2 + \sum_{t=s}^{n} d_t^2 e_t^2 \leq O(\ln s) + \frac{O(\ln s)}{s} O\left(\ln \frac{n}{s}\right) \quad \text{a.s.,}$$

where the inequality holds by $d_t \leq 1$ and (35). Take $s = \ln n$, which implies (34) and the proof of the theorem follows. □

5 Simulation study

We study the behavior of the proposed SNLS model selection criterion in a simulation study where the AIC, BIC, PLS, NML, and SNLS (Eq. (15)) methods are used to estimate the order of an AR model. The scripts, in R language, needed to reproduce all the experiments in this paper are available for download[1].

The true order was varied between $k^* = 1, \ldots, 10$, and the sample sizes were $n = 100, 200, 400, 800, 1600$. The parameters of the AR models are generated by sampling parameter vectors uniformly at random from the range $[-1, 1]^{k^*}$ and rejecting combinations that result in unstable processes, until 1000 accepted (stable) models were produced per each (n, k^*) pair. The criteria were

[1] http://www.cs.helsinki.fi/teemu.roos/snls/snls2.R

evaluated for orders up to $k = 15$, and the order minimizing each criterion was chosen as the estimate.

Figures 1 and 2 (left panels) show the percentage of correctly estimated orders for each true order k^* and sample size n. For the lowest orders, $k^* = 1, 2$, the BIC criterion is clearly the most accurate one and wins for almost all sample sizes; this was expected since BIC is known to have a tendency to underestimate rather than overestimate the order. Likewise, it is not too surprising that AIC, which a priori favors more complex models than the other criteria, wins for the smallest sample size whenever $k^* \geq 5$. For the orders $k^* = 3$ and $k^* = 4$, BIC, PLS, NML, and SNLS share the first place with small margin. For orders $k^* \geq 5$, SNLS is usually the best method with again a very small margin.

As pointed out by an anonymous referee, the goal of model selection is not always to pick the “correct” model, but to find one that minimizes future prediction errors. To this end, we also carried out a predictive experiment where the model learned by each of the criteria was used for predicting a new outcome from the same process. The right panels of Figs. 1 and 2 show the the predictive accuracy corresponding to the chosen model order under the different criteria. Predictive accuracy is measured in terms of the squared error. The boxplots show the first to third quartile range as a box, with the median (bar) and the mean (triangle) superimposed. Overall, the differences in predictive accuracy are extremely small.

References

[1] Akaike, H. (1974), ‘A new look at the statistical model identification’, *IEEE Trans. Automat. Control*, Vol. **19**, No. 6, pp. 716–723.

[2] Barron, A.R., Rissanen, J., and Yu, B. (1998), ‘The minimum description length principle in coding and modeling’, *IEEE Trans. Inform. Theory*, Vol. **44**, No. 6, pp. 2743–2760.

[3] Cover, T.M. and Thomas J.A. (1991), *Elements of Information Theory*, John Wiley and Sons, New York, 542 pages.

[4] Davis, M.H.A. and Hemerly, E.M. (1990), ‘Order determination and adaptive control of ARX models using the PLS criterion’, *Proc. Fourth Bad Honnef Conf. on Stochastic Differential Systems. Lecture Notes in Control and Information Sci.* (N. Christopeit, Ed.), Springer, New York.

[5] Dawid, A.P. (1984), ‘Present position and potential developments: some personal views. Statistical theory. The prequential approach’, *J. Royal Statist. Soc. Ser. A*, Vol. **147**, Part 2, pp. 278–292.

[6] Hannan, E.J., Mcdougall, A.J., and Poskit, D.S. (1989), ‘Recursive estimation of autoregressions’. *J. Royal Statist. Soc. Ser. B*, Vol. **51**, No. 2, pp. 217–233.

[7] Hansen, M.H. and Yu, B. (2001), ‘Model selection and the principle of minimum description length’, *J. Am. Statist. Ass.*, Vol. **96**, No. 454, pp. 746–774.

[8] Lai, T.L. and Wei, C.Z. (1982), ‘Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems’, *Ann. Statist.*, Vol. **10**, No. 1, pp. 154–166.

[9] Merhav, N., Feder, M., and Gutman, M. (1992), ‘Universal prediction of individual sequences’, *IEEE Trans. Inform. Theory*, Vol. **38**, No. 4, pp. 1258–1270.

[10] Plackett, R.L. (1950), ‘Some theorems in least squares’, *Biometrika*, Vol. **37**, No. 1–2, pp. 149–157.

[11] Pollard, D. (2002), *A User’s Guide to Measure Theoretic Probability*, Cambridge University Press, 351 pages.

[12] Roos, T., Myllymäki, P., and Rissanen, J. (2009), ‘MDL denoising revisited’, *IEEE Trans. Signal Proc.*, Vol. **57**, No. 9, pp. 3347–3360.

[13] Rissanen, J. (1978), ‘Modeling by shortest data description’, *Automatica*, Vol. **14**, No. 5, pp. 465–471.

[14] Rissanen, J. (1986), ‘Stochastic complexity and modeling’, *Ann. Statist.*, Vol. **14**, No. 3, pp. 1080–1100.

[15] Rissanen, J. (1986), ‘A predictive least squares principle’, *IMA J. Math. Control Inform.*, Vol. **3**, No. 2–3, pp. 211–222.

[16] Rissanen, J. (1996), ‘Fisher information and stochastic complexity’, *IEEE Trans. Inform. Theory*, Vol. **42**, No. 1, pp. 40–47.

[17] Rissanen, J. (2000), ‘MDL denoising’, *IEEE Trans. Inform. Theory*, Vol. **46**, No. 7, pp. 2537–2543.

[18] Rissanen, J. (2007), *Information and Complexity in Statistical Modeling*, Springer Verlag, 142 pages.

[19] Schwarz, G. (1978). ‘Estimating the dimension of a model’. *Ann. Statist.*, Vol. **6**, No. 2, pp. 416–464.

[20] Shtarkov, Yu.M. (1987), ‘Universal sequential coding of single messages’, *Probl. Inform. Transm.*, Vol. **23**, No. 3, pp. 3–17.

[21] Wei, C.Z. (1992), ‘On predictive least squares principles’, *Ann. Statist.*, Vol. **20**, No. 1, pp. 1–42.

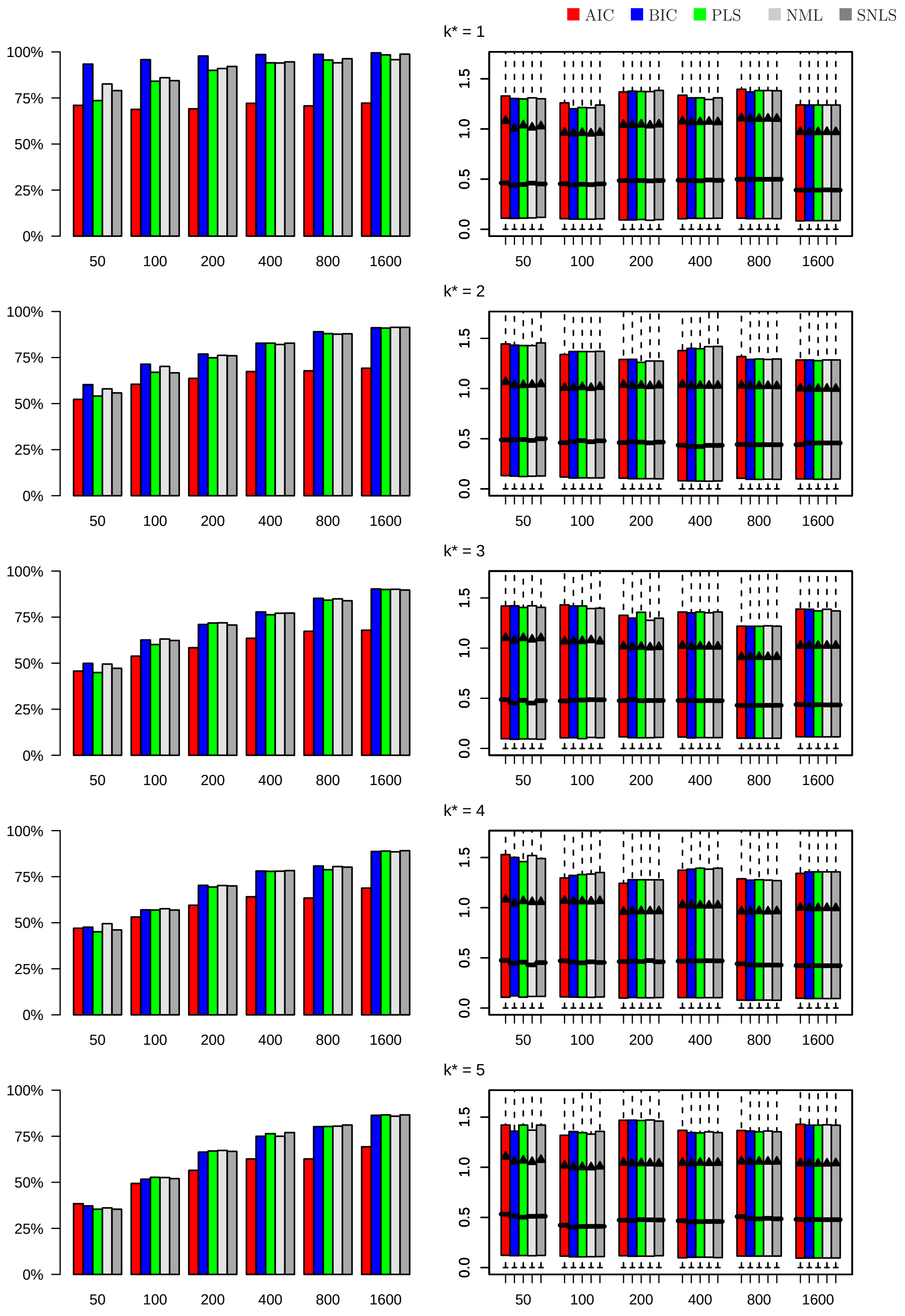

Fig. 1. Experimental results. *Left:* Percentages of correctly estimates orders for true model order $k^* = 1, \ldots, 5$ (to be continued in Fig. 2); sample sizes $n = 50, 100, 200, 400, 800, 1600$. *Right:* Boxplots of the corresponding squared prediction errors, black bars indicate median, triangles indicate mean.

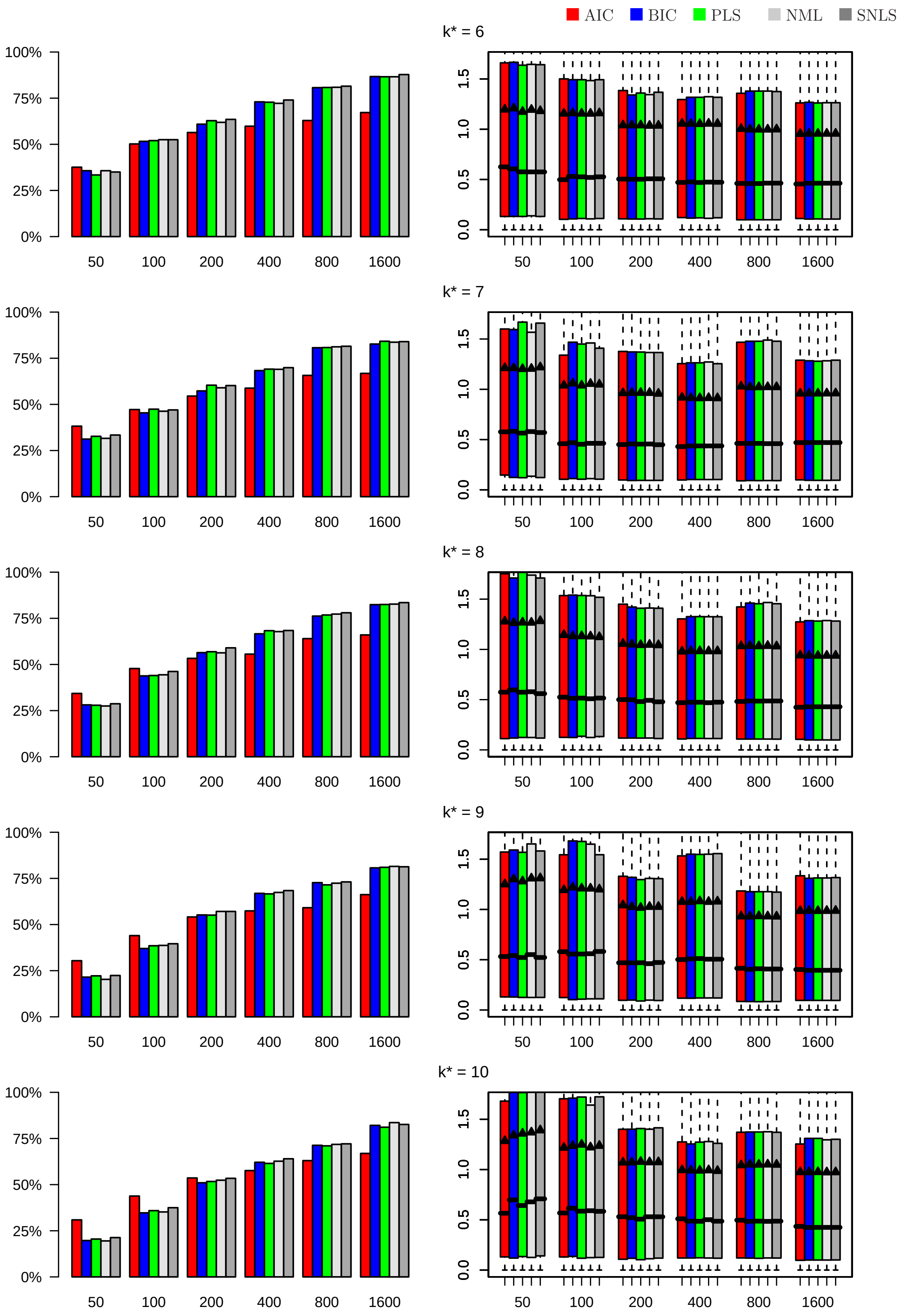

Fig. 2. Experimental results. *Left:* Percentages of correctly estimates orders for true model order $k^* = 6, \ldots, 10$ (continued from Fig. 1); sample sizes $n = 50, 100, 200, 400, 800, 1600$. *Right:* Boxplots of the corresponding squared prediction errors, black bars indicate median, triangles indicate mean.