ON SEQUENTIALLY NORMALIZED MAXIMUM LIKELIHOOD MODELS

Teemu Roos and Jorma Rissanen

Helsinki Institute for Information Technology HIIT
P.O.Box 68 (Department of Computer Science)
FI-00014 University of Helsinki, FINLAND
teemu.roos@cs.helsinki.fi jorma.rissanen@mdl-research.org

1. INTRODUCTION

The important normalized maximum likelihood (NML) distribution is obtained via a normalization over all sequences of given length. It has two short-comings: the resulting model is usually not a random process, and in many cases, the normalizing integral or sum is hard to compute. In contrast, the recently proposed *sequentially normalized maximum likelihood* (SNML) models always comprise a random process and are often much easier to compute. We present some results on SNML type models in the Markovian and linear–Gaussian model classes.

In the linear–Gaussian case, the resulting *sequentially normalized least squares* (SNLS) model is particularly interesting. The associated sequentially minimized squared deviations are smaller than both the usual least squares and the squared prediction errors used in the so called *predictive least squares* (PLS) criterion. The SNLS model is asymptotically optimal within the given class of distributions by reaching the lower bound on the logarithmic prediction errors, given by the stochastic complexity, up to lower-order terms.

2. SOME MINMAX PROBLEMS

Consider the model class $\mathcal{M}_k = \{f(x^n;\theta)\}$, $\theta = \theta_1,\ldots,\theta_k$, and data sequences $x^n = x_1,\ldots,n$, for $n = 1,2,\ldots$. Let m be a large enough integer such that the ML estimate $\hat{\theta}_t = \hat{\theta}(x^t)$ can be computed for $t > m$. The number

$$\log 1/f(x^n;\hat{\theta}_n)$$

has been considered as the ideal target for the code length obtainable with the model class, [1], which, however, is not attainable, because $f(x^n;\hat{\theta}_n)$ is not a probability distribution. This leads to the minmax problem

$$\min_q \max_{x^n} \log \frac{f(x^n;\hat{\theta}_n)}{q(x^n)},$$

with the solution due to Shtarkov, known as the *normalized maximum likelihood* (NML) universal model, [2],

$$\begin{aligned} \hat{f}_{\mathrm{NML}}(x^n;\mathcal{M}_\gamma) &= \frac{f(x^n;\hat{\theta}(x^n))}{C_n} \qquad (1)\\ C_n &= \int f(y^n;\hat{\theta}(y^n))dy^n. \end{aligned}$$

However, the normalizing coefficient can be evaluated easily only for restricted model classes, and the model does not define a random process. This means that it cannot be used for prediction and its evaluation for data compression is difficult.

Now, consider for all $t > m$, the problem

$$\min_{q(x|x^{t-1})} \max_x \log \frac{f(x^{t-1},x;\hat{\theta}(x^{t-1},x))}{q(x \mid x^{t-1})}. \qquad (2)$$

The solution is given by the *conditional NML* model

$$\begin{aligned} \hat{f}(x_t \mid x^{t-1}) &= \frac{f(x^t;\hat{\theta}(x^t))}{K_t(x^{t-1})} \qquad (3)\\ K_t(x^{t-1}) &= \int f(x^{t-1},x;\hat{\theta}(x^{t-1},x))dx. \end{aligned}$$

This is proved the same way as the solution to Shtarkov's problem: First, replacing the numerator by the density function (3) does not change the solution, and the maximized ratio of the two density functions (3) and $q(x \mid x^{t-1})$, which is not smaller than unity, is made unity when the latter is selected equal to the former.

It is clear that the normalizing coefficient $K_t(x^{t-1})$, which in general is a function of x^{t-1}, is easier to calculate, at least numerically, than the normalizing coefficient in the NML universal model.

For another type of normalization, where the numerator $f(x^t\,;\hat{\theta}(x^t))$ is replaced by the conditional density $f(x_t \mid x^{t-1}\,;\hat{\theta}(x^t))$, see [3].

Putting together the conditional NML densities gives the *sequentially normalized maximum likelihood* (sNML) model:

$$f_{\mathrm{SNML}}(x^n) = f^m(x^m) \prod_{t=m+1}^{n} \hat{f}(x_t \mid x^{t-1}), \qquad (4)$$

where $f^m(x^m)$ is a suitably chosen initial distribution. The result is, by construction, a random process.

3. BERNOULLI MODEL

We begin with an example involving the Bernoulli class $\mathcal{B} = \{P(x;p)\}$, where the parameter $p = P(1)$. The ML estimate is given by $\hat{p}(x^n) = n_1/n$, where $n_1 = \sum_t x_t$ is

the number of 1's in x^n. If $n_0 = n - n_1$ the maximized likelihood is

$$P(x^n; n_1/n) = \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0}.$$

The conditional NML predictive probability can be written as

$$\hat{P}(1 \,|\, x^n) = \frac{(n_1 + 1)\, e(n_1)}{(n_0 + 1)\, e(n_0) + (n_1 + 1)\, e(n_1)}, \quad (5)$$

where $e(n_0) = (1 + 1/n_0)^{n_0}$ and $e(n_1) = (1 + 1/n_1)^{n_1}$; take $e(k) = 1$ for $k = 0$.

For instance, in the problem considered by Laplace, given a sequence of '1's, the successive probabilities of yet another '1' are $\frac{1}{2}, \frac{4}{5}, \frac{27}{31}, \frac{256}{283}, \ldots$. Compare this to the more conservative solution by Laplace,

$$P_{\rm Lap}(1 \,|\, x^n) = \frac{n_1 + 1}{n + 2},$$

which gives the same sequence as $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \ldots$, i.e., the certainty of 80 %, which is achieved by the Laplace probability on the fourth step, is achieved by conditional NML already on the second step.

The same conditional probability function $\hat{P}(1 \,|\, x^n)$ was found in [2], where it was shown to behave similarly to the Krichevski-Trofimov predictive probability

$$P_{\rm KT}(1 \,|\, x^n) = \frac{n_1 + 1/2}{n + 1}.$$

It was also found later in [4], in effect, as the solution to the following minmax problem

$$\min_{\theta} \max_{x} \log \frac{f(x^{t-1}, x; \hat{\theta}(x^{t-1}, x))}{f(x \,|\, x^{t-1}; \theta)}. \quad (6)$$

Neither Krichevski-Trofimov predictive probability nor the related Laplace probability has been shown to have any particular optimality property, except asymptotically. Takimoto and Warmuth [4] showed that for the Bernoulli models, the regret of the sNML model (4) satisfies for all sequences the inequality

$$\begin{aligned} R(f_{\rm SNML}, x^n) &:= \ln 1/f_{\rm SNML}(x^n) - \ln 1/f(x^n; \hat{\theta}(x^n)) \\ &\le \frac{1}{2} \ln(n + 1) + \frac{1}{2}. \end{aligned} \quad (7)$$

We conclude this section by noting that in the Bernoulli case, the alternative version of normalization mentioned in Sec. 2, where the numerator of (3) is replaced by $f(x_t \mid x^{t-1}\,; \hat{\theta}(x^t))$, agrees with the Laplace probability, see [3].

4. LINEAR–QUADRATIC MODELS

In the rest of the paper, we are concerned with deriving a model selection criterion for a class of normal models

$$\begin{aligned} &f(y^n \mid X_n\,; \sigma^2, b) \\ &\quad = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_1^n (y_t - b'\bar{x}_t)^2\right), \end{aligned}$$

induced by the regression equations

$$y_t = b'\bar{x}_t + \epsilon_t, \quad (8)$$

where the prime indicates transposition, $b' = (b(1), \ldots, b(k))$, 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 let m be the smallest integer t such that $k(t) = k$.

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

$$y_t = b'_{t-1}\bar{x}_t + e_t = \sum_{i=1}^{k(t)} b_{t-1,i} x_{t,i} + e_t, \quad (9)$$

$$y_t = b'_n\bar{x}_t + \hat{\epsilon}_t(n) = \sum_{i=1}^{k(t)} b_{n,i} x_{t,i} + \hat{\epsilon}_t(n), \quad (10)$$

$$y_t = b'_t\bar{x}_t + \hat{e}_t = \sum_{i=1}^{k(t)} b_{t,i} x_{t,i} + \hat{e}_t. \quad (11)$$

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 $t > m$,

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

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 [5], [6], [7], and [8]. 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 [9] and [8].

The second representation (10) is traditional, and it, too, has associated model selection criteria, including AIC [10], and BIC [11],

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

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 $f(y^n \mid X_n\,;\, b_n, \hat{\sigma}_n^2)$ 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 $C_{n,k} = \int_{y^n \in \mathcal{Y}} f(y^n \mid X_n\,;\, b_n, \hat{\sigma}_n^2)\, dy^n$ [1], [2], [12]. 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 [13], see also [14, 15]. 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 $\hat{R} = b_n' X_n X_n' b_n/n$.

The third representation, which we are interested in, is new. The sum of squared deviations $\hat{e}_t^2$ is smaller than either the sum of the traditional least squares $\hat{\epsilon}_t^2(n)$, or the sum of the squared prediction errors e_t^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, it too needs to be normalized in order to obtain a proper density function. We study the asymptotic behavior of the resulting sequentially normalized least squares 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.

5. SEQUENTIALLY NORMALIZED 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

$$f(y_t \mid y^{t-1}, X_t; \sigma^2, b_t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_t-\hat{y}_t)^2}{2\sigma^2}\right), \tag{12}$$

are obtained by replacing the parameter vector b in the conditional normal density function $f(y_t \mid y^{t-1}, X_t; \sigma^2, b)$ 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 [16],

$$\begin{aligned}
b_t &= V_t \sum_{j=1}^{t} \bar{x}_j y_j \\
&= b_{t-1} + \frac{V_{t-1}}{1+c_t}\bar{x}_t(y_t - \bar{x}_t' b_{t-1}) \quad (13)\\
V_t &= (X_t X_t')^{-1} \\
&= V_{t-1} - V_{t-1}\bar{x}_t\bar{x}_t' V_{t-1}/(1+c_t) \quad (14)\\
c_t &= \bar{x}_t' V_{t-1}\bar{x}_t \\
d_t &= \bar{x}_t' V_t \bar{x}_t \\
1-d_t &= 1/(1+c_t). \quad (15)
\end{aligned}$$

The last equality was shown in [7] and [8] 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, [8]. This also implies that $0 \le d_t \le 1$.

By (13) we obtain

$$\begin{aligned}
\hat{y}_t &= \bar{x}_t'\left[V_{t-1}\bar{x}_t(y_t - \bar{x}_t' b_{t-1})/(1+c_t) + b_{t-1}\right] \\
&= c_t/(1+c_t)(y_t - \bar{x}_t' b_{t-1}) + \bar{x}_t' b_{t-1} \\
&= (1-d_t)\bar{x}_t' b_{t-1} + d_t y_t. \quad (16)
\end{aligned}$$

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

$$\hat{e}_t = y_t - \hat{y}_t = (1-d_t)(y_t - \bar{x}_t' b_{t-1}) = (1-d_t)e_t, \tag{17}$$

which is seen to be smaller than the plug-in prediction error by a constant factor. The normalization of (12) 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 (12) 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

$$\max_{\sigma^2} \prod_{t=m+1}^{n} f(y_t \mid y^{t-1}, X_t; \sigma^2, b_t). \tag{18}$$

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 normalizing over y_t, we get the normalized conditional density function

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

The normalizing integral is given by

$$K(y^{t-1}) = \frac{\sqrt{\pi}}{1-d_t}\Gamma\left(\frac{t-m-1}{2}\right)/\Gamma\left(\frac{t-m}{2}\right).$$

The proof is omitted. We need $t > m+1$ to make the normalizer non-zero.

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

$$\hat{f}(y_t \mid y^{t-1}, X_t) = K_{t-1}^{-1} \frac{\hat{\tau}_t^{-(t-m)/2}}{\hat{\tau}_{t-1}^{-(t-m-1)/2}}.$$

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

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

The negative logarithm of this gives the *sequentially normalized least squares* (SNLS) criterion:

$$\begin{aligned} \mathrm{SNLS}(n,k) = {} & \frac{n-m}{2} \ln(2\pi e \hat{\tau}_n) \\ & + \sum_{t=m+1}^{n} \ln(1+c_t) + \frac{1}{2} \ln n + O(1), \end{aligned} \tag{19}$$

where Stirling's formula has been applied to the Gamma function, and constant terms 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*

$$\frac{1}{n} X_n X_n' = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i \bar{x}_i' \to \Sigma \tag{20}$$

with Σ non-singular, then

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

The proof of this and all subsequent theorems are left to the full version.

6. FIXED REGRESSION MATRIX

The first theorem shows the mean square deviations in the three representations of data (9), (10), and (11), which are of some interest, and which we will need later on. Since we need the recursive formulas (13), (14), (15) we give the results for $t > m$.

Theorem 2 *If the regressor variables are non-random satisfying (20) and the data generated by (8), then*

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

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

$$\frac{1}{n-m} \left(\sum_{t=1}^{n} \mathbb{E} \hat{\epsilon}_t^2(n) - \sum_{t=1}^{m} \mathbb{E} \hat{\epsilon}_t^2(m) \right) = \sigma^2, \tag{23}$$

where the expectation is with the parameters b and σ.

The next theorem shows the asymptotic optimality of the SNLS model in terms of logarithmic prediction errors, see [9], both in the mean and almost surely, in the case where the regressor matrix is fixed.

Theorem 3 *Let the assumption (20) hold, and let the data be generated by (8). Then*

$$\mathbb{E}\,\mathrm{SNLS}(n,k) = \frac{n-m}{2} \ln(2\pi e \sigma^2) + \frac{k+1}{2} \ln n + o(\ln n), \tag{24}$$

for almost all parameters b and σ. Also,

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

almost surely.

7. AR MODELS

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

$$y_t = \sum_{i=1}^{k} a_i y_{t-i} + \epsilon_t, \ t \geq 1, \tag{26}$$

in which the regressor matrix is random, determined by the the data y^n, and where we write the coefficients as a_i to avoid confusing them with b_i, where the subindex refers to time i.

The following theorem shows the almost sure asymptotic optimality of the SNLS model also in this case.

Theorem 4 *Let the data be generated by an AR model (26), where the roots of the polynomial $1 - \sum_{i=1}^{k} a_i z^i$ 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*

$$\ln \hat{\tau}_n = \ln \hat{\sigma}_n^2 - \left(\frac{k}{n-m} \ln n \right) (1 + o(1)), \tag{27}$$

almost surely, and

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

almost surely.

8. SIMULATION STUDY

We study the behavior of the proposed SNLS model selection criterion in a simulation study where the AIC, BIC, PLS, and SNLS (Eq. (19)) methods are used to estimate the order of an AR model. The scripts, in R language, needed to reproduce the experiments are available for download[1].

The true order was varied between $k^* = 1, \ldots, 10$, and the sample sizes were $n = 100, 200, 400, 800, 1600, 3200$. 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 3000 accepted (stable) models were produced per each (n, k^*) pair. The criteria were evaluated for orders up to $k = 15$, and the order minimizing each criterion was chosen as the estimate.

Tables 1 and 2 report 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 $k = 3, 4, 5$, BIC, PLS, NML, and SNLS share the first place, the last one somewhat more often than the others. For orders $k = 5, \ldots, 10$, Table 2, SNLS is clearly the best method, with the exception of the smallest sample sizes.

9. ACKNOWLEDGMENTS

This work was supported in part by the Academy of Finland under the project Civi and by the Finnish Funding Agency for Technology and Innovation under the projects Kukot and PMMA. In addition, this work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence.

10. REFERENCES

[1] A.R. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," vol. 44, no. 6, pp. 2743–2760, 1998.

[2] Yu.M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, no. 3, pp. 175–186, 1987.

[3] T. Roos, T. Silander, P. Kontkanen, and P. Myllymäki, "Bayesian network structure learning using factorized NML universal models," in *Proc. Information Theory and Applications Workshop (ITA-08)*. 2008, IEEE Press.

[4] E. Takimoto and M. Warmuth, "The last-step minimax algorithm," in *Proc. 11th International Conference on Algorithmic Learning Theo ry*, 2000, pp. 279–290.

[5] M.H.A. Davis and E.M. Hemerly, "Order determination and adaptive control of ARX models using the PLS criterion," 1990.

[6] E.J. Hannan, A.J. Mcdougall, , and D.S. Poskit, "Recursive estimation of autoregressions," *J. Royal Statist. Soc. Ser. B*, vol. 51, no. 2, pp. 217–233, 1989.

[7] T.L. Lai and C.Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *Annals of Statistics*, vol. 10, no. 1, pp. 154–166, 1982.

[8] C.Z. Wei, "On predictive least squares principles," *Annals of Statistics*, vol. 20, no. 1, pp. 1–42, 1992.

[9] J. Rissanen, "A predictive least squares principle," *IMA J. Math. Control Inform.*, vol. 3, pp. 211–222, 1986.

[10] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Control*, vol. 19, no. 6, pp. 716–723, 1974.

[11] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[12] J. Rissanen, "Fisher information and stochastic complexity," vol. 42, no. 1, pp. 40–47, 1996.

[13] J. Rissanen, "MDL denoising," vol. 46, no. 7, pp. 2537–2543, 2000.

[14] T. Roos, P. Myllymäki, and J. Rissanen, "MDL denoising revisited," Sept. 2006, preprint *arXiv:cs/0609138*.

[15] J. Rissanen, *Information and Complexity in Statistical Modeling*, Information Science and Statistics. 2007, 140 pages.

[16] R.L. Plackett, "Some theorems in least squares," *Biometrika*, vol. 37, no. 1–2, pp. 149–157, 1950.

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

Table 1. Percentages of correctly estimated orders, $k^* = 1, \ldots, 5$ *(to be continued in Table 2)* The score of the best method in each case is typeset in boldface.

		sample size, n						
		50	100	200	400	800	1600	3200
$k = 1$	AIC	70.5	71.3	72.0	70.0	71.4	70.8	70.9
	BIC	**93.5**	**96.9**	**97.9**	**98.0**	**99.4**	**99.5**	**99.4**
	PLS	75.8	86.3	91.1	93.5	96.7	97.8	98.1
	NML	82.5	88.3	89.7	91.5	94.3	95.9	96.6
	SNLS	78.5	87.5	92.2	93.9	97.0	98.1	98.3
$k = 2$	AIC	52.1	58.0	64.1	64.6	66.4	69.0	68.4
	BIC	**61.3**	**69.3**	**78.2**	**83.0**	**88.0**	90.0	**93.8**
	PLS	52.7	65.5	76.4	81.7	86.6	89.7	93.5
	NML	59.7	68.3	76.9	82.4	86.4	89.8	93.6
	SNLS	53.8	66.1	77.5	82.1	86.3	**90.1**	93.6
$k = 3$	AIC	47.1	55.5	59.2	63.6	66.5	68.6	69.2
	BIC	49.5	**63.5**	**72.3**	79.2	84.6	88.3	92.2
	PLS	45.3	61.8	71.7	79.1	84.8	**88.7**	**92.6**
	NML	**49.7**	63.1	72.1	**79.5**	84.5	88.3	92.4
	SNLS	46.5	63.0	71.1	79.3	**84.9**	88.6	**92.6**
$k = 4$	AIC	42.8	52.5	60.1	63.3	65.4	66.5	67.5
	BIC	**45.7**	59.6	67.8	76.5	**82.6**	88.3	91.4
	PLS	42.1	58.3	68.5	**77.0**	82.5	88.3	91.9
	NML	45.0	**60.2**	68.0	76.7	82.5	88.0	91.6
	SNLS	42.4	59.2	**69.4**	**77.0**	82.4	**88.5**	**92.0**
$k = 5$	AIC	**39.7**	49.6	56.9	60.5	65.7	67.1	66.8
	BIC	39.0	52.1	65.4	74.8	80.5	85.8	90.4
	PLS	39.1	53.4	65.8	75.0	**81.0**	**86.1**	90.4
	NML	39.2	52.1	**66.2**	74.8	**81.0**	85.8	90.4
	SNLS	39.4	**54.2**	66.1	**76.0**	**81.0**	86.0	**90.7**

Table 2. Percentages of correctly estimated orders, $k^* = 6, \ldots, 10$ *(continued from Table 1).* The score of the best method in each case is typeset in boldface.

		sample size, n						
		50	100	200	400	800	1600	3200
$k = 6$	AIC	**37.9**	51.0	56.5	59.5	64.1	68.7	68.2
	BIC	35.4	51.8	62.9	71.3	79.8	86.6	90.4
	PLS	34.7	52.3	62.9	72.0	80.1	86.7	**90.7**
	NML	35.6	52.8	63.2	71.7	79.9	86.5	90.6
	SNLS	36.3	**53.3**	**64.0**	**72.4**	**80.3**	**86.8**	90.5
$k = 7$	AIC	**33.7**	45.4	55.3	59.6	63.6	65.7	67.3
	BIC	29.2	43.4	59.1	69.5	77.9	82.8	88.6
	PLS	30.0	44.7	60.5	70.0	78.5	82.9	88.6
	NML	28.8	44.2	59.8	69.8	78.3	83.0	88.4
	SNLS	30.1	**46.5**	**61.2**	**70.6**	**79.4**	**83.2**	**88.9**
$k = 8$	AIC	**34.4**	**45.9**	55.7	59.6	64.5	66.3	67.6
	BIC	26.9	43.0	57.6	69.1	78.2	81.4	86.5
	PLS	28.8	44.6	58.8	69.1	77.9	82.0	86.4
	NML	26.9	43.6	58.2	69.7	78.8	81.8	**86.8**
	SNLS	28.8	**45.9**	**59.2**	**69.8**	**79.0**	**82.4**	**86.8**
$k = 9$	AIC	**30.0**	**44.1**	52.8	59.0	64.3	64.9	69.1
	BIC	23.1	39.1	52.8	66.1	75.7	82.2	86.3
	PLS	23.8	40.2	53.3	67.1	75.9	81.5	86.3
	NML	22.4	39.6	53.7	66.8	76.6	**82.3**	86.7
	SNLS	24.6	42.0	**55.0**	**67.8**	**76.7**	82.2	**86.9**
$k = 10$	AIC	**28.5**	**43.9**	51.5	59.3	64.2	67.1	67.7
	BIC	20.6	35.7	51.0	66.1	74.4	81.4	85.5
	PLS	20.1	35.7	50.7	65.0	73.4	80.8	84.8
	NML	20.2	37.1	51.9	**66.8**	74.6	81.4	**85.8**
	SNLS	21.4	37.9	**52.3**	66.5	**74.8**	**81.8**	85.6