Subset Selection in Linear Regression using Sequentially Normalized Least Squares: Asymptotic Theory

Jussi Määttä* Daniel F. Schmidt† Teemu Roos*

Running headline:

SNLS: Asymptotic Theory

Abstract

This article examines the recently proposed sequentially normalized least squares criterion for the linear regression subset selection problem. A simplified formula for computation of the criterion is presented, and an expression for its asymptotic form is derived without the assumption of normally distributed errors. Asymptotic consistency is proved in two senses: (i) in the usual sense, where the sample size tends to infinity, and (ii) in a non-standard sense, where the sample size is fixed and the noise variance tends to zero.

Keywords: asymptotics, consistency, linear regression, minimum description length principle, subset selection.

*Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki

†Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne

1 Introduction

Consider a design matrix $Z_n = (z_1^{\mathrm{T}}, \ldots, z_n^{\mathrm{T}})^{\mathrm{T}}$, where $z_i \in \mathbb{R}^q$ are row vectors, and a corresponding vector of observed responses $y \in \mathbb{R}^n$. The linear regression model assumes that the mean of the responses is given by a linear combination of the covariates, resulting in the generating model

$$y = Z_n \beta_* + \sigma_* \varepsilon, \tag{1}$$

where $\beta_* \in \mathbb{R}^q$ are the regression coefficients, $\varepsilon = \varepsilon^{(n)} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is a vector of independent and identically distributed random variates with $E(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = 1$, and $\sigma_* > 0$ determines the standard deviation of the errors. It is common to assume that the distribution of the errors ε is normal, but none of the results presented in this paper require this assumption.

When some of the components of β_* are exactly zero, the associated covariates are unrelated to the response, and the problem is to identify which of the components are non-zero. More formally, let $\gamma = \{\gamma_1, \ldots, \gamma_k\} \subseteq \{1, \ldots, q\}$ denote an index vector determining which $1 \leq k \leq q$ of the q covariates comprise the design submatrix

$$X_n(\gamma) = (x_1^{\mathrm{T}}, \ldots, x_n^{\mathrm{T}})^{\mathrm{T}}, \quad x_i = (z_{i,\gamma_1}, \ldots, z_{i,\gamma_k}),$$

and let Γ denote the set of all candidate subsets under consideration. If the subset includes all covariates that correspond to non-zero components in β_*, i.e., if no covariates that are related to the response are left out, the model indexed by γ is said to be *correct*. Otherwise the model is said to be *incorrect*. The correct model with the fewest coefficients is called the *true* model. In this paper, we only consider the case where the true model has at least one non-zero coefficient.

Motivated by earlier work on sequential normalized maximum likelihood, Rissanen *et al.* (2010) proposed the sequentially normalized least squares (SNLS) criterion for subset selection in linear regression problems and studied its behavior under the assumption that the model γ is correct and the errors are normally distributed. They also presented a simulation experiment in which the sequentially normalized least squares criterion was found to identify the true model with as high, or higher, probability than certain other commonly used criteria such as the Akaike information criterion (AIC) (Akaike, 1974), the Bayesian information criterion (BIC) (Schwarz, 1978), and the predictive least squares (PLS) method (Rissanen, 1986).

However, despite these promising numerical results, the theoretical properties of SNLS have been poorly understood. Given that Rissanen *et al.* (2010) only studied the limiting behavior under correct models and white Gaussian noise, there has been no guarantee that SNLS remains a valid criterion in more general situations. Such guarantees have been available for many other model selection criteria, such as the PLS criterion (Wei, 1992) and the BIC; in this paper we provide several asymptotic model selection guarantees for SNLS under a relaxed model of random noise. Specifically, we assume only that the error distribution has a mean of zero and a finite fourth moment.

Given that SNLS is derived under the assumption of Gaussian white noise, it is clearly not possible to provide any guarantees on model selection consistency in the sense of identifying the true probabilistic model if the errors ε are non-Gaussian. Instead, we focus on the consistency of the estimate of the model structure, γ, obtained by minimising the SNLS score over a *finite* set of candidate model structures, i.e., we study the probability of SNLS selecting only those covariates that are truly associated with the responses y.

These results do not immediatetely imply anything to about the more stringent notions of consistency for countably infinite sets of models, such as those studied, for example, by Barron & Cover (1991) and Csiszár & Shields (2000).

Our main results are as follows: (i) we provide a large-sample asymptotic expression which is seen to be similar to that of the predictive least squares criterion (Wei, 1992, Theorem 4.1.1), but different from the BIC formula under incorrect models (Section 3); (ii) we show that the criterion is consistent as the sample size tends to infinity (Section 4); (iii) we provide conditions under which the criterion is also consistent as the noise variance σ_*^2 tends to zero (Section 5). The proofs of all lemmas are presented in a supplementary document (Appendix S1: Proofs of Lemmas).

2 Sequentially Normalized Least Squares

The sequentially normalized least squares criterion is closely related to the (predictive) minimum description length principle and sequential coding (Rissanen, 1989; Grünwald, 2007; Roos & Rissanen, 2008). The basic idea is to sequentially define a predictive distribution $q(y_t \mid y_{1:t-1}, X_t(\gamma))$ for response y_t conditional on all the previously observed data $y_{1:t-1} = (y_1, \ldots, y_{t-1})^{\mathrm{T}}$ and $X_t(\gamma)$, using the subset of covariates specified by γ. The criterion is then defined as the accumulated negative logarithm of the predictive density for the data $y_{m+1}, \ldots, y_n$

$$\textsc{snls}(n, \gamma, y) = -\sum_{t=m+1}^{n} \log q(y_t \mid y_{1:t-1}, X_t(\gamma)),$$

where the predictive density is defined in a specific way. In Bayesian terms, the criterion corresponds to a negative logarithm of a marginal likelihood, and the above identity corresponds to its decomposition into a product of predictive densitities by the chain rule. The index m determines the number of initial samples that are ignored in the beginning of the sequence so as to guarantee that the required predictive densities are well-defined, can be viewed as the number of samples used to "kick-start" the sequential coding procedure. In practice, there are two choices for m that result in different theoretical properties; these are presented later in this article and summarized in Section 6. The SNLS criterion is inspired by the ideas of sequential communication of data, and is closely related to the sequentially normalized maximum likelihood principle (Roos & Rissanen, 2008). The resulting score can be interpreted as the length of a compressed message, in natural digits, required to communicate the data $\mathbf{y}$ using the model γ. In this sense the SNLS naturally balances the goodness-of-fit of a model against its inherent "complexity".

Given the observed data, (Z_n, y), the preferred subset of covariates is selected by choosing the subset from a set of possible choices $\Gamma \subseteq 2^{\{1,\ldots,q\}}$ that yields the smallest SNLS score. The full details of the derivation of the sequentially normalized least squares are given in Rissanen *et al.* (2010). For convenience, we summarize the results here and present a simplification of the formula given in Rissanen *et al.* (2010).

For notational simplicity, we often simply write $\text{SNLS}(n, \gamma)$. We also drop the explicit dependence on terms such as $X_n(\gamma)$ on the subset γ when it is clear from the context. Let $X_t = (x_1^{\mathrm{T}}, \ldots, x_t^{\mathrm{T}})^{\mathrm{T}}$ denote the matrix comprised of the first $t \leq n$ rows of $X_n = X_n(\gamma)$. The SNLS criterion is then defined as

$$\text{SNLS}(n, \gamma) = \left(\frac{n-m}{2}\right) \log(2\pi e \hat{\tau}_n) + \sum_{t=m+1}^{n} \log(1 + c_t) + \frac{1}{2} \log n, \quad (2)$$

where

$$\hat{\tau}_n = \hat{\tau}_n(\gamma) = \left(\frac{1}{n-m}\right) \sum_{t=m+1}^{n} \hat{e}_t^2, \quad (3)$$

and

$$
\begin{aligned}
J_t &= X_t^{\mathrm{T}} X_t, \\
\hat{\beta}_t &= J_t^{-1} X_t^{\mathrm{T}} y_{1:t}, \\
c_t &= x_t J_{t-1}^{-1} x_t^{\mathrm{T}}, \\
d_t &= x_t J_t^{-1} x_t^{\mathrm{T}}, \\
e_t &= y_t - x_t \hat{\beta}_{t-1}, \\
\hat{e}_t &= y_t - x_t \hat{\beta}_t = (1 - d_t) e_t.
\end{aligned}
$$

Using the Sherman–Morrison–Woodbury formula (Hager, 1989) and the fact that $J_t = J_{t-1} + x_t^{\mathrm{T}} x_t$, one also obtains the important identity

$$1 - d_t = \frac{1}{1 + c_t}.$$

Before Rissanen *et al.* (2010), the above quantities were already considered by Wei (1992) in his analysis of the PLS criterion.

We note that equation (2) is actually an approximation of the exact criterion. It is obtained by applying Stirling's formula to the term $\log \Gamma((n - m)/2)$, so the approximation is accurate even for small values of n.

The required least-squares estimates $\hat{\beta}_t$ based on the first t data points can be computed efficiently for all $t = m + 1, \ldots, n$ using the complete recurrence relations given in, for instance, Plackett (1950). As $\hat{\beta}_t$ is not unique for $t < |\gamma|$, it is required that m satisfies $m \geq |\gamma|$ for all $\gamma \in \Gamma$.

2.1 A Simplified Form of SNLS

The criterion as given by (2) can be simplified by noting (Wei, 1992, eq. 2.3) that

$$1 - d_t = \frac{|X_{t-1}^{\mathrm{T}} X_{t-1}|}{|X_t^{\mathrm{T}} X_t|} = \frac{|J_{t-1}|}{|J_t|}.$$

where $|\cdot|$ denotes the determinant of a matrix. Then, it is clear that $1 + c_t = |J_t|/|J_{t-1}|$, and the second term on the right hand side of (2) can be written as

$$\sum_{t=m+1}^{n} \log(1 + c_t) = \log\left(\frac{|J_{m+1}|}{|J_m|} \cdot \frac{|J_{m+2}|}{|J_{m+1}|} \cdots \frac{|J_n|}{|J_{n-1}|}\right).$$

By noting that the terms in the product telescope we arrive at the following simplification.

Proposition 1. *The* SNLS *criterion* (2) *admits the simplified formula*

$$\text{SNLS}(n, \gamma) = \left(\frac{n-m}{2}\right) \log(2\pi e \hat{\tau}_n) + \log\frac{|J_n|}{|J_m|} + \frac{1}{2}\log n. \tag{4}$$

We now introduce the following regularity assumption:

$$\lim_{n\to\infty} \left\{\frac{1}{n} Z_n^{\mathrm{T}} Z_n\right\} = \Lambda, \qquad \text{with } \Lambda \text{ positive definite.} \tag{5}$$

This implies that the same holds when Z_n is replaced by any $X_n(\gamma)$. Using Proposition 1, the proof of the large sample behaviour of the SNLS criterion is greatly simplified.

Theorem 1 (Theorem 1 in Rissanen *et al.* (2010))**.** *Under assumption* (5)*, we have*

$$\text{SNLS}(n, \gamma) = \left(\frac{n-m}{2}\right) \log(2\pi e \hat{\tau}_n) + \left(\frac{2|\gamma| + 1}{2}\right) \log n + o(\log n). \tag{6}$$

Proof. By assumption (5), we have $|J_n| = n^{|\gamma|}\, |(1/n) X_n^{\mathrm{T}} X_n|$, where the determinant term tends to a strictly positive constant due to Sylvester's criterion. Hence $\log |J_n| = |\gamma| \log n + O(1)$. Moreover $|J_m| = O(1)$. Applying Proposition 1 completes the proof. □

Proposition 1 also allows us to verify that SNLS possesses two desirable invariance properties. The first is the invariance of SNLS under all transformations of the design matrix of the form $X_n A$, with $A \in \mathbb{R}^{|\gamma| \times |\gamma|}$ positive definite. This is easily seen by: (i) noting that the quantity $\hat{\tau}_n$ depends on the specific form of the design matrix only through the least squares estimator, which is itself invariant under this class of transformations, and (ii) by using the properties of determinants to verify that the ratio in the second term of (4) is also invariant under these types of transformations of the design matrix.

An immediate and practical consequence of this result is that the *scale* of measurement chosen for the covariates has no effect on the criterion, which is intuitively pleasing as we do not expect our results to change whether we measure our covariates in centimetres or inches. In a similar fashion, it is desirable for the criterion to be invariant under scale transformations of the responses, i.e., that the SNLS criterion is invariant, up to a constant independent of γ, under all transformations of the responses of the form κy, with $\kappa \neq 0$.

From Proposition 1, and the properties of the least-squares estimator and the definition of $\hat{\tau}_n$, it is immediately obvious that the difference between SNLS scores for y and the scaled responses κy is given by

$$\text{SNLS}(n, \gamma, y) - \text{SNLS}(n, \gamma, \kappa y) = \frac{1}{2}(m - n) \log \kappa^2,$$

which is the same for all γ (assuming that m is independent of γ; see Section 5 for a discussion on several different choices of m). The important implication of this result is that scaling the responses y will have no effect on the relative ordering of models by their SNLS score, and thus have no effect on the model selected by SNLS.

2.2 SNLS and Bayesian Model Selection

The simplified SNLS formula given by (4) also reveals some interesting similarities and differences with a number of model selection criteria based on Bayesian inference. It has been shown, under suitable regularity conditions involving the likelihood function and the choice of prior distribution (Barndorff-Nielsen & Cox, 1989), that linear regression model selection criteria based on Bayesian marginal probabilities can be written in the form

$$-\log \int\int p(y|\beta,\sigma^2)p(\beta,\sigma^2)\,\mathrm{d}\beta\mathrm{d}\sigma^2 = \frac{n}{2}\log 2\pi\hat{\sigma}_n^2 + \frac{n}{2} + \frac{1}{2}\log|J_n| + O(1), \quad (7)$$

where

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\hat{\beta}_n)^2$$

is the empirical residual variance obtained by the least squares estimates $\hat{\beta}_n$ for the model γ, and $p(\beta,\sigma^2)$ is an appropriate prior distribution. Under suitable regularity conditions, a large number of model selection procedures, through asymptotic equivalency with Bayesian marginal probabilities, such as conditional normalized maximum likelihood, the minimum message length principle (Wallace, 2005) and earlier forms of the minimum description length principle (Rissanen, 1989) may also be written in the form (7), and indeed this expression forms the basis for the Bayesian information criteria (Schwarz, 1978). More recently, Hedayati & Bartlett (2012) have demonstrated that (7) holds, under suitable regularity conditions, even in the case of improper Jeffreys' priors, as long as a suitable number of the observations y are used as start-up samples, in a similar manner to the SNLS.

When interpreting (7), it is common to view the first term as measuring the goodness-of-fit of the model γ to the data y, which under suitable conditions on the design matrix Z, decreases with increasing number of covariates, while the second term, which generally increases with increasing number of covariates, may be viewed as a model complexity "penalty" term. Comparing (7) with SNLS (4), it can be seen that while the two show intriguing similarities, the SNLS criterion as given by (4) cannot be immediately written in a form equivalent to (7). The log-determinant of the Fisher information matrix term present in SNLS is twice as large as the corresponding term in the asymptotic Bayesian formula (7), which would imply that SNLS should be significantly more conservative than regular Bayesian procedures. However, comparing the quantity $\hat{\tau}_n$, given by (3), which measures the goodness-of-fit of the model in the SNLS criterion, to the usual estimate of residual variance, $\hat{\sigma}_n^2$, shows that $\hat{\tau}_n \leq \hat{\sigma}_n^2$, with equivalence only in the specific case that the responses are all zero. This reveals a crucial difference between the two approaches, and in this paper we demonstrate the interesting result that the fact that $\hat{\tau}_n$ is smaller than $\hat{\sigma}_n^2$ is directly compensated for by the larger "penalty" term in SNLS.

3 Large-Sample Behaviour

In the following, we will assume that the regressors are bounded:

$$\sup_n z_n z_n' = \alpha < \infty. \tag{8}$$

We also require the assumption that the limits

$$\lim_{n \to \infty} \left\{ \frac{1}{n} \sum_{t=1}^{n} Z_{ti} Z_{tj} Z_{tk} Z_{t\ell} \right\} \tag{9}$$

exist for all $i, j, k, \ell \in \{1, 2, \dots, q\}$. All in all, our assumptions are that the regressors are bounded (8), linearly independent [positive definiteness in (5)], and well-behaved enough to produce a covariance matrix (5) and four-fold products (9).

We let $R = R(\gamma)$ be the $(q \times q)$ idempotent matrix with entries $r_{i,i} = 1$ if and only if $i \in \gamma$, and zero otherwise. Again, assume that the limit (5) exists; recalling that the data y is generated by the linear model (1), with $\beta_* \in \mathbb{R}^q$ denoting the true, underlying regression coefficients, this assumption implies that

$$\lim_{n\to\infty}\left\{\frac{1}{n}(Z_nR)^{\mathrm{T}}(Z_nR)\right\} = R\Lambda R \quad \text{and}$$
$$\lim_{n\to\infty}\left\{\frac{1}{n}(Z_nR)^{\mathrm{T}}(Z_n\beta_*)\right\} = R\Lambda\beta_*.$$

Note that $R^- = R$, where $(\cdot)^-$ denotes the Moore–Penrose pseudoinverse; using this we may define the quantities

$$\delta = Z_n\beta_* - Z_n(R\Lambda R)^-\Lambda\beta_* \quad \text{and}$$
$$\tilde{G} = \lim_{n\to\infty}\left\{\frac{1}{n}(Z_nR)^{\mathrm{T}}\Delta(Z_nR)\right\},$$

where Δ is an $(n \times n)$ diagonal matrix with entries $\Delta_{i,i} = \delta_i^2$. If $R = R(\gamma)$ has $R_{ii} = 1$ for all i such that $(\beta_*)_i \neq 0$, then the vector δ is the zero vector and forces $\tilde{G}$ to be the zero matrix. More generally, δ can written as $\delta = Z_nL(R)\beta_*$, where $L(R) = I_q - (R\Lambda R)^-\Lambda$ is a linear transformation operating on the coefficient vector. The matrix $\tilde{G}$ is well-defined due to assumption (9).

We now give a general expression for the large sample behaviour of the sequentially normalized least squares criterion that applies to both correct and incorrect models, i.e., even when γ omits covariates which are related to the response (with corresponding non-zero entries in β_*). The result will involve

$$\hat{\sigma}_n^2 = \hat{\sigma}_n^2(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - z_i\tilde{\beta}_n\right)^2$$

which is an estimate of σ^2 based on the restricted least squares estimate $\tilde{\beta}_n =$

$((Z_nR)^{\mathrm{T}}(Z_nR))^{-}(Z_nR)^{\mathrm{T}}y_{1:n}$. First, we have the following important lemma.

Lemma 1. *Under assumptions* (5) *and* (8)*, we have*

$$\log \hat{\tau}_n = \log \hat{\sigma}_n^2(\gamma) - \left(\frac{\log n}{n}\right)\left[\frac{|\gamma|\sigma_*^2 + \mathrm{tr}((R\Lambda R)^{-}\tilde{G})}{\hat{\sigma}_n^2(\gamma)}\right] + o\left(\frac{\log n}{n}\right) \quad \textit{a.s.} \tag{10}$$

Combining the above lemma with Theorem 1, we immediately obtain the following asymptotic form of the criterion.

Corollary 1. *Under assumptions* (5) *and* (8)*, the sequentially normalized least squares criterion satisfies*

$$\begin{aligned} \textsc{snls}(n,\gamma) = {} & \left(\frac{n-m}{2}\right)\log(2\pi e\hat{\sigma}_n^2(\gamma)) \\ & + \left(2|\gamma| - \frac{|\gamma|\sigma_*^2 + \mathrm{tr}((R\Lambda R)^{-}\tilde{G})}{\hat{\sigma}_n^2(\gamma)} + 1\right)\frac{\log n}{2} \\ & + o(\log n) \quad \textit{a.s.} \end{aligned} \tag{11}$$

The above asymptotic formula resembles the asymptotic form of the predictive least squares criterion (Wei, 1992, Theorem 4.1.1). Notice that (11) agrees with the BIC formula, in which the penalty term is given by $(|\gamma|/2)\log n$, if the trace term vanishes and if $\hat{\sigma}_n^2(\gamma) \to \sigma_*^2$. These conditions are guaranteed to hold for correct models; see Lemmas 2 and 3 below.

4 Consistency as $n \to \infty$

In terms of model selection, consistency (the guarantee that a criterion will identify the true model as the sample size grows) is an important property. The BIC criterion has been shown to be consistent under a wide range of situations (Haughton, 1988). However, the asymptotic form (11) differs from the BIC formula, and it remains to be shown that the sequentially normalized least squares criterion is also consistent.

For the SNLS criterion to be consistent, we require a mild additional assumption:

$$E[\varepsilon_i^4] < \infty. \tag{12}$$

Let γ_* denote the index of the components of β_* that are non-zero. We define the sequentially normalized least squares estimator as

$$\hat{\gamma}_n = \arg\min_{\gamma \in \Gamma} \left\{ \text{SNLS}(n, \gamma) \right\}.$$

That is, we define the SNLS estimate of γ_* to be the subset γ that minimizes the SNLS criterion. In this section, we will prove the consistency of this estimate as $n \to \infty$. In a manner analogous to the proof of consistency of predictive least squares presented in Wei (1992) we treat the underfitting and overfitting cases separately.

Lemma 2. *Under assumptions* (5) *and* (8)*, the estimate of σ_*^2 based on the restricted least squares estimates satisfies*

$$\hat{\sigma}_n^2 = \sigma_*^2 + \xi + o(1) \quad \textit{a.s.}$$

for some constant $\xi \geq 0$ that depends on γ. Moreover, $\xi = 0$ if and only if γ is correct.

Lemma 2 may be combined with Lemma 1 to show that

$$\begin{aligned} \log \hat{\tau}_n &= \log \hat{\sigma}_n^2 + O\left(\frac{\log n}{n}\right) \quad \text{a.s.} \\ &= \log\left(\sigma_*^2 + \xi\right) + o(1) \quad \text{a.s.} \end{aligned} \tag{13}$$

The underfitting case is covered by the following theorem, which by the use of the above results is relatively straightforward to prove.

Theorem 2. *Let γ_2 be correct and assume without loss of generality that $1 \in \gamma_2$. Further assume* (5), (8), *and $\beta_*(1) \neq 0$ and let $\gamma_1 \subseteq \gamma_2 \setminus \{1\}$. Then*

$$\lim_{n\to\infty} \left\{ \frac{\text{SNLS}(n,\gamma_1) - \text{SNLS}(n,\gamma_2)}{n} \right\} = \rho/2 \quad \textit{a.s.},$$

where

$$\lim_{n\to\infty} \{\log \hat{\tau}_n(\gamma_1) - \log \hat{\tau}_n(\gamma_2)\} = \rho > 0 \quad \textit{a.s.} \tag{14}$$

Proof. By Theorem 1, we almost surely have

$$\begin{aligned}\text{SNLS}(n,\gamma_1) - \text{SNLS}(n,\gamma_2) = {} & \frac{n}{2}\left(\log \hat{\tau}_n(\gamma_1) - \log \hat{\tau}_n(\gamma_2)\right) \\ & + \frac{m}{2}\log\left(2\pi e \hat{\tau}_n(\gamma_2)\right) - \frac{m}{2}\log\left(2\pi e \hat{\tau}_n(\gamma_1)\right) \\ & + (|\gamma_1| - |\gamma_2|)\log n + o(\log n).\end{aligned}$$

Using (13), the above may be simplified to

$$\begin{aligned}\text{SNLS}(n,\gamma_1) - \text{SNLS}(n,\gamma_2) = {} & \frac{n}{2}\left(\log \hat{\tau}_n(\gamma_1) - \log \hat{\tau}_n(\gamma_2)\right) \\ & + O(\log n) \quad \text{a.s.}\end{aligned} \tag{15}$$

Since $\gamma_1 \subset \gamma_2$ and $\beta_*(1) \neq 0$, Lemma 2 gives us

$$\lim_{n\to\infty} \{\hat{\sigma}_n^2(\gamma_1) - \hat{\sigma}_n^2(\gamma_2)\} = \xi > 0 \quad \text{a.s.}$$

This and (13) imply (14), which can be plugged into (15) to obtain the claim. □

The somewhat more involved overfitting case is covered by the following theorem.

Theorem 3. *Let γ_2 be correct and assume without loss of generality that $1 \in \gamma_2$.*

Further assume (5), (8), (12), *and* $\beta_*(1) = 0$ *and let* $\gamma_1 = \gamma_2 \setminus \{1\}$. *Then*

$$\mathrm{pr}\left\{\textsc{snls}(n,\gamma_2) - \textsc{snls}(n,\gamma_1) > 0\right\} \geq 1 - O\left(\frac{1}{(\log n)^2}\right).$$

To prove Theorem 3, we first introduce some useful lemmas.

Lemma 3. *If* γ *is correct, then* $\mathrm{tr}((R\Lambda R)^{-}\tilde{G}) = 0$.

Lemma 4. *Let* γ_2 *be a model with* $1 \in \gamma_2$ *and assume* (5). *Let* $\gamma_1 = \gamma_2 \setminus \{1\}$. *Then there exist constants* $C \in \mathbb{R}$ *and* $N \in \mathbb{N}$ *such that*

$$\log|J_n(\gamma_2)| - \log|J_n(\gamma_1)| \geq \log n + C$$

for all $n \geq N$.

Lemma 5. *Let* γ_2 *be correct and assume* (5), (8), $1 \in \gamma_2$ *and* $\beta_*(1) = 0$. *Let* $\gamma_1 = \gamma_2 \setminus \{1\}$. *Then*

$$\log\hat{\tau}_n(\gamma_2) - \log\hat{\tau}_n(\gamma_1) = \left(\hat{\sigma}_n^2(\gamma_2) - \hat{\sigma}_n^2(\gamma_1)\right)(1 + o(1)) - \frac{\log n}{n} + o\left(\frac{\log n}{n}\right)$$

almost surely.

Lemma 6. *Under the same assumptions as in Lemma 5, we have*

$$\hat{\sigma}_n^2(\gamma_2) - \hat{\sigma}_n^2(\gamma_1) = \frac{\sigma_*^2}{n}(\varepsilon^{(n)})^{\mathrm{T}} M_n \varepsilon^{(n)} \quad \textit{a.s.}$$

where

$$\begin{aligned} M_n &= \frac{1}{n} Z_n \left(Q_n(\gamma_1) - Q_n(\gamma_2)\right) Z_n^{\mathrm{T}}, \\ Q_n(\gamma_i) &= \left(\frac{1}{n}(Z_n R(\gamma_i))^{\mathrm{T}}(Z_n R(\gamma_i))\right)^{-}. \end{aligned}$$

Lemma 7. *Under the same assumptions as in Lemma 5, we have*

$$\limsup_{n\to\infty}\left\{E\left[\frac{n}{2}\left(\hat{\sigma}_n^2(\gamma_2)-\hat{\sigma}_n^2(\gamma_1)\right)\right]\right\}<\infty$$

and if we further assume (12)*, then*

$$\limsup_{n\to\infty}\left\{\operatorname{var}\left[\frac{n}{2}\left(\hat{\sigma}_n^2(\gamma_2)-\hat{\sigma}_n^2(\gamma_1)\right)\right]\right\}<\infty$$

Now we have all the pieces we need to prove Theorem 3.

Proof of Theorem 3. All the statements that follow until the end of the proof hold with probability one (almost surely). Using Proposition 1 and equation (13), we have

$$\begin{aligned}\textsc{snls}(n,\gamma_2)-\textsc{snls}(n,\gamma_1) &= \frac{n}{2}\left(\log\hat{\tau}_n(\gamma_2)-\log\hat{\tau}_n(\gamma_1)\right)\\ &\quad+\frac{m}{2}\log(2\pi e\hat{\tau}_n(\gamma_1))-\frac{m}{2}\log(2\pi e\hat{\tau}_n(\gamma_2))\\ &\quad+\log|J_n(\gamma_2)|-\log|J_n(\gamma_1)|\\ &\quad+\log|J_m(\gamma_1)|-\log|J_m(\gamma_2)|\\ &= \frac{n}{2}\left(\log\hat{\tau}_n(\gamma_2)-\log\hat{\tau}_n(\gamma_1)\right)\\ &\quad+\log|J_n(\gamma_2)|-\log|J_n(\gamma_1)|+O(1),\end{aligned}$$

and applying Lemma 4 gives that

$$\textsc{snls}(n,\gamma_2)-\textsc{snls}(n,\gamma_1)\geq\frac{n}{2}\left(\log\hat{\tau}_n(\gamma_2)-\log\hat{\tau}_n(\gamma_1)\right)+\log n+O(1).$$

By Lemma 5, this is equivalent to

$$\textsc{snls}(n,\gamma_2)-\textsc{snls}(n,\gamma_1)\geq\frac{n}{2}\left(\hat{\sigma}_n^2(\gamma_2)-\hat{\sigma}_n^2(\gamma_1)\right)(1+o(1))+\frac{\log n}{2}+o(\log n)$$

Notice that by Lemma 7, the above lower bound consists essentially of a random variable with a bounded mean and variance and an increasing term. The claim follows by Chebyshev's inequality: denoting $U_n = (n/2)\left(\hat{\sigma}_n^2(\gamma_2) - \hat{\sigma}_n^2(\gamma_1)\right)$, we have

$$\begin{aligned}
&\mathrm{pr}\left\{\textsc{snls}(n,\gamma_2) - \textsc{snls}(n,\gamma_1) > 0\right\} \\
\geq\ &\mathrm{pr}\left\{U_n(1+o(1)) + \frac{\log n}{2} + o(\log n) > 0\right\} \\
=\ &1 - \mathrm{pr}\left\{E[U_n] - U_n \geq E[U_n] + \frac{\frac{1}{2}\log n + o(\log n)}{1+o(1)}\right\} \\
\geq\ &1 - \mathrm{pr}\left\{|E[U_n] - U_n| \geq E[U_n] + \frac{\frac{1}{2}\log n + o(\log n)}{1+o(1)}\right\} \\
\geq\ &1 - \frac{\mathrm{var}[U_n]}{\left(E[U_n] + \frac{\frac{1}{2}\log n + o(\log n)}{1+o(1)}\right)^2} = 1 - O\left(\frac{1}{(\log n)^2}\right)
\end{aligned}$$

where we used Lemma 7 in the final step. □

The proof of consistency now follows.

Theorem 4. *Assume* (5), (8) *and* (12). *Let* $\gamma_* \subseteq \{1,2,\dots,q\}$ *be the unique correct model with the least variables, and let* $\hat{\gamma}_n$ *denote the model that minimises the* SNLS *criterion using the* n *first data vectors. Then*

$$\mathrm{pr}\left\{\hat{\gamma}_n = \gamma_*\right\} \geq 1 - O\left(\frac{1}{(\log n)^2}\right).$$

Proof. Let $\gamma \subseteq \{1,2,\dots,q\}$ be an incorrect model. Then $\gamma \cup \gamma_*$ is a correct model. Choose some $j \in \gamma_* \setminus \gamma$. The regression coefficient related to j, which we may assume without loss of generality to be $\beta_*(1)$, must be nonzero. By Theorem 2,

$$\mathrm{pr}\left\{\textsc{snls}(n,\{1,2,\dots,q\}) < \textsc{snls}(n,\gamma) \text{ eventually}\right\} = 1.$$

On the other hand, consider the case where $\gamma \neq \gamma_*$ is a correct model. Then there exists some positive integer ℓ and models γ_i, $1 \leq i \leq \ell$, such that $\gamma_1 = \gamma_*$ (with possibly reordered indices) and $\gamma_\ell = \gamma$ and for $1 < i \leq \ell$, the model $\gamma_i \setminus \gamma_{i-1}$ is a singleton. We may now apply Theorem 3 $\ell - 1$ times and use the union bound to obtain

$$\text{pr}\left\{\textsc{snls}(n, \gamma_*) < \textsc{snls}(n, \gamma)\right\} \geq 1 - O\left(\frac{1}{(\log n)^2}\right)$$

which completes the proof. □

5 Consistency as $\sigma_* \to 0$

Recently, a number of papers have examined an alternative notion of consistency in which the sample size is fixed, and the noise variance goes to zero. In particular, Ding & Kay (2011) demonstrated that the BIC does not satisfy this notion of consistency, while Schmidt & Makalic (2012) showed that minimum description length criteria based on normalized maximum likelihood (Rissanen, 2000) and Bayes mixture codes (Hansen & Yu, 2001) are both consistent in this sense. This is a seemingly non-standard notion of consistency; to put it into context, consider the signal-to-noise ratio of a regression model with coefficients β_* and error variance σ_*^2 for a given design matrix Z_n,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{E[y_i^2]}{E[\varepsilon_i^2]} = \frac{\beta_*' Z_n' Z_n \beta}{n\sigma_*^2}. \tag{16}$$

From (16), it is clear that consistency as $\sigma_* \to 0$ is essentially equivalent to guaranteeing that a model selection procedure will select the true model, if it is among the set of candidate models, with probability tending to one, as the signal-to-noise ratio tends to infinity. In the context of a domain such as signal processing, in which the imprecision of signal measurement is usually modelled through the presence of random errors, this is essentially equivalent to a guarantee on the performance of a criterion with increasingly accurate measurement.

While this notion of consistency appears on the surface to be quite different to the usual notion based on increasing sample size, it is actually closely related. This can be seen recalling that the determinant of the full Fisher information matrix for β_* is $|J_n/\sigma_*^2|$. Under standard assumptions on the design matrix, such as (5), the usual notion of consistency as $n \to \infty$ implies that the probability of selecting the true model tends to one as the Fisher information tends to infinity. It is clear that the alternative notion of consistency as $\sigma_* \to 0$ implies exactly the same concept, and in this sense they are simply guarantees on model selection performance as the amount of information increases. In the standard notion of consistency, the information about β_* increases as we obtain more samples, while in the alternative notion of consistency the information increases as the samples are measured more accurately.

In order to establish the consistency of the sequentially normalized least squares criterion under this notion of consistency, we need to abandon the assumption that the number of initial observations, m, that are omitted is the same for all subsets γ. In fact, in the previous sections of this article, this assumption is only required in In fact, in the previous sections of this article, this assumption is only required if we require SNLS to be invariant under a scaling of the responses y_t (cf. Section 2.1); the other results do not depend on m being constant. In the following section, we let $m(\gamma)$ be a function of the subset γ such that

$$|\gamma| \leq m(\gamma) < n.$$

We also make the following assumptions:

$$\begin{aligned} |X_n^{\mathrm{T}}(\gamma)X_n(\gamma)| &> 0, \text{ for all } \gamma \in \Gamma, && (17) \\ \sum_{t=m(\gamma)+1}^{n} x_{t,\gamma_i}^2 &> 0, \text{ for all } i = 1, \ldots, |\gamma|, \ \gamma \in \Gamma. && (18) \end{aligned}$$

For large enough n, assumption (5) implies (17) and (18).

We then have the following result in which the choice of the function $m(\cdot)$ plays a crucial role.

Theorem 5. *Let $\hat{\gamma} \in \Gamma$ be an estimate of $\gamma_* \in \Gamma$ found by minimising a criterion of the form*

$$I(y, \gamma) = \left(\frac{n - m(\gamma)}{2} \right) \log T_n(\gamma) + c, \tag{19}$$

with

$$T_n(\gamma) = \sum_{t=m(\gamma)+1}^{n} a_t e_t^2(\gamma),$$

where $a_t > 0$ and c are constants independent of σ_*^2. Then, under assumptions (17) and (18), $\hat{\gamma}$ is a consistent estimate of γ_* as $\sigma_*^2 \to 0$ if and only if for any $|\gamma_1| > |\gamma_2|$ we have $m(\gamma_1) > m(\gamma_2)$.

Proof. To prove consistency, we show that both the probability of overfitting and the probability of underfitting approach zero as $\sigma_*^2 \to 0$. In the following discussion the symbol 'c' is used as a generic placeholder for any constant terms that are independent of σ_*^2. We first show that the estimate $\hat{\gamma}$ will not overfit as $\sigma_*^2 \to 0$. A model γ is considered to overfit if it is any correct model other than the true model, i.e., it references all the non-zero entries of β_* as well as some additional zero entries. We have

$$\begin{aligned}
e_t(\gamma) &= y_t - x_t(\gamma)\hat{\beta}_{t-1}(\gamma) \\
&= (z_t\beta_* + \sigma_*\varepsilon_t) \\
&\quad - z_t \left((Z_{t-1}R)^{\mathrm{T}}(Z_{t-1}R)\right)^{-} (Z_{t-1}R)^{\mathrm{T}}(Z_{t-1}\beta_* + \sigma_*\varepsilon^{(t-1)}) \\
&= z_t \left(\beta_* - \left((Z_{t-1}R)^{\mathrm{T}}(Z_{t-1}R)\right)^{-} (Z_{t-1}R)^{\mathrm{T}} Z_{t-1}\beta_*\right) \\
&\quad + \sigma_*(z_t \left((Z_{t-1}R)^{\mathrm{T}}(Z_{t-1}R)\right)^{-} (Z_{t-1}R)^{\mathrm{T}}\varepsilon^{(t-1)} + \varepsilon_t). \qquad (20)
\end{aligned}$$

If the model γ under consideration is correct, the first term on the right hand side of (20) vanishes; it therefore follows that the random variable

$$S_{n,\gamma} = \sum_{t=m(\gamma)+1}^{n} a_t \left(\frac{e_t^2(\gamma)}{\sigma_*^2}\right)$$

is independent of σ_*^2 if γ is correct, and we may write $T_n(\gamma) = \sigma_*^2 \cdot S_{n,\gamma}$. To show that $\hat{\gamma}$ will not overfit as $\sigma_*^2 \to 0$, it suffices to show that

$$\mathrm{pr}\left\{\mathrm{I}(y,\gamma) - \mathrm{I}(y,\gamma_*) < 0\right\}$$

tends to zero as $\sigma_*^2 \to 0$, where γ is any model such that $\gamma_* \subseteq \gamma$ and $|\gamma| > |\gamma_*|$. From (19), the probability of overfitting may be written as

$$\mathrm{pr}\left\{\left(\frac{n-m(\gamma)}{2}\right)\log\left(\sigma_*^2 \cdot S_{n,\gamma}\right) - \left(\frac{n-m(\gamma_*)}{2}\right)\log\left(\sigma_*^2 \cdot S_{n,\gamma_*}\right) + c < 0\right\}.$$

After simplification this becomes

$$\mathrm{pr}\left\{(m(\gamma_*) - m(\gamma))\log\sigma_*^2 < (n - m(\gamma_*))\log S_{n,\gamma_*} - (n - m(\gamma))\log S_{n,\gamma} + c\right\}.$$

Exponentiating both sides yields

$$\mathrm{pr}\left\{\left(\frac{1}{\sigma_*^2}\right)^{m(\gamma)-m(\gamma_*)} < c\left(\frac{S_{n,\gamma_*}^{n-m(\gamma_*)}}{S_{n,\gamma}^{n-m(\gamma)}}\right)\right\}. \tag{21}$$

The right-hand-side of (21) is a random variable independent of σ_*^2, which implies that for (21) to tend to zero as $\sigma_*^2 \to 0$, the left-hand-side of (21) must be unbounded from above as $\sigma_*^2 \to 0$, which requires that $m(\gamma) - m(\gamma_*) > 0$. Due to the fact that γ is overfitting, we have $|\gamma| > |\gamma_*|$; therefore, the only way in which $m(\gamma) - m(\gamma_*) > 0$ may be satisfied for all $\gamma_* \in \Gamma$ is if $m(\cdot)$ is chosen such that $m(\gamma_1) > m(\gamma_2)$ if $|\gamma_1| > |\gamma_2|$.

We now show that the probability of preferring an incorrect model to the true model tends to zero as $\sigma_*^2 \to 0$. The probability that an incorrect model γ is preferred to the true model γ_* is

$$\mathrm{pr}\left\{\left(\frac{n-m(\gamma)}{2}\right)\log T_n(\gamma) - \left(\frac{n-m(\gamma_*)}{2}\right)\log\left(\sigma_*^2 \cdot S_{n,\gamma_*}\right) + c < 0\right\}.$$

After simplification, this becomes

$$\mathrm{pr}\left\{\left(\frac{1}{\sigma_*^2}\right)^{n-m(\gamma_*)} < c\left(\frac{S_{n,\gamma_*}^{n-m(\gamma_*)}}{(T_n(\gamma))^{n-m(\gamma)}}\right)\right\}. \tag{22}$$

If a model γ is incorrect, then under assumptions (17) and (18), the first term on the right-hand-side of (20) will be a non-zero constant independent of σ_*, while the second term will be of order $O(\sigma_*)$ as $\sigma_* \to 0$, and it is straightforward to show that

$$E\{T_n(\gamma)\} \to c, \quad \text{and} \quad \mathrm{var}(T_n(\gamma)) \to 0,$$

as $\sigma_*^2 \to 0$, for all incorrect models. Therefore, the right-hand-side of (22) converges in distribution to $c \cdot S_{n,\gamma_*}^{n-m(\gamma_*)}$ as $\sigma_*^2 \to 0$, which is independent of σ_*^2. As

$\sup_{\gamma \in \Gamma}\{m(\gamma)\} < n$ we have $n - m(\gamma) > 0$ and the left-hand-side of (22) is unbounded from above as $\sigma_*^2 \to 0$, implying that the probability of preferring an incorrect model to the true model is vanishingly small as $\sigma_*^2 \to 0$. □

It is straightforward to see that Theorem 5 determines the conditions under which the sequentially normalized least squares criterion is consistent as $\sigma_*^2 \to 0$. Specifically, the SNLS criterion is equivalent to the criterion (19) with $a_t = (1-d_t)^2$. The result also shows that the predictive least squares criterion, which is obtained by letting $a_t = 1$, has the same consistency property.

6 Discussion

When applying the predictive least squares and sequentially normalized least squares criteria to subset selection problems, there are two natural choices of $m(\gamma)$. While consistency of the SNLS criterion as $n \to \infty$ is unaffected by the specific choice of m, as long as it is independent of n, consistency as $\sigma_* \to 0$ requires that $m(\gamma)$ satisfies a particular property (see Section 5).

The first choice of $m(\gamma)$ is to take $m(\gamma) = |\gamma|$, which from Theorem 5, ensures consistency as $\sigma_*^2 \to 0$. However, this choice has two disadvantages: (i) models with different number of parameters are assessed on a different number of samples, and (ii) the resulting criterion is not invariant under scale transformations of the data (cf. Section 2.1). The second choice is to set m to the minimum number of samples required to ensure that the least-squares estimates are unique for all models under consideration, i.e.,

$$m(\gamma) = m = \sup_{\gamma' \in \Gamma} \left\{ |\gamma'| \right\}. \tag{23}$$

This choice has the advantage that all candidate models are assessed on the same number of samples, and that the resulting criterion is invariant under scale transformations of the data. However, from Theorem 5 it is clear that the resulting criterion is inconsistent as $\sigma_*^2 \to 0$.

In our view, the choice (23) is recommended in practice. Sensitivity to an arbitrary choice such as the scale in which the data is measured seriously compromises the validity of a criterion. In contrast, consistency as $\sigma_*^2 \to 0$ is a notion that only applies in potentially unrealistic cases. Furthermore, if prediction is the goal, the lack of consistency as $\sigma_*^2 \to 0$ is moderated by the fact that the prediction error for all overfitting models in Γ will tend to zero as $\sigma_*^2 \to 0$ by virtue of the fact that the estimation error will also tend to zero.

Acknowledgments

The authors thank Jorma Rissanen and Enes Makalic for useful discussions. J.M. and T.R. are supported by the Academy of Finland under the Finnish Centre of Excellence in Computational Inference Research (COIN).

Supporting Information

Additional information for this article is available online, including Appendix S1: Proofs of Lemmas.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19**(6), 716–723.

Barndorff-Nielsen, O. E. & Cox, D. R. (1989). *Asymptotic techniques for use in statistics*. Chapman & Hall, New York, NY.

Barron, A. R. & Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37**(4), 1034–1054.

Csiszár, I. & Shields, P. C. (2000). The consistency of the BIC Markov order estimator. *Ann. Statist.* **28**(6), 1601–1619.

Ding, Q. & Kay, S. (2011). Inconsistency of the MDL: On the performance of model order selection criteria with increasing signal-to-noise ratio. *IEEE Trans. Signal Process.* **59**(5), 1959–1969.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT Press, Cambridge, MA.

Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM Rev.* **31**(2), 221–239.

Hansen, M. H. & Yu, B. (2001). Model selection and the principle of minimum description length. *J. Amer. Statist. Assoc.* **96**(454), 746–774.

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.* **16**(1), 342–355.

Hedayati, F. & Bartlett, P. L. (2012). Exchangeability characterizes optimality of sequential normalized maximum likelihood and Bayesian prediction with Jeffreys prior. In *Proc. 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, 504–510.

Plackett, R. L. (1950). Some theorems in least squares. *Biometrika* **37**(1/2), 149–157.

Rissanen, J. (1986). A predictive least squares principle. *IMA J. Math. Control Inform.* **3**(2-3), 211–222.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. World Scientific Publishing, Singapore.

Rissanen, J. (2000). MDL denoising. *IEEE Trans. Inform. Theory* **46**(7), 2537–2543.

Rissanen, J., Roos, T. & Myllymäki, P. (2010). Model selection by sequentially normalized least squares. *J. Multivariate Anal.* **101**(4), 839–849.

Roos, T. & Rissanen, J. (2008). On sequentially normalized maximum likelihood models. In *Proc. 1st Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08).*

Schmidt, D. F. & Makalic, E. (2012). The consistency of MDL for linear regression models with increasing signal-to-noise ratio. *IEEE Trans. Signal Process.* **60**(3), 1508–1510.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Ann. Statist.* **2**, 461–464.

Wallace, C. S. (2005). *Statistical and inductive inference by minimum message length.* Information Science and Statistics. Springer, Hoboken, NJ.

Wei, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20**(1), 1–42.

Jussi Määttä, Helsinki Institute for Information Technology HIIT, Department of Computer Science, PO Box 68 (Gustaf Hällströmin katu 2b), FI-00014 University of Helsinki, Finland.
E-mail: jussi.maatta@helsinki.fi