ON THE CONSISTENCY OF SEQUENTIALLY NORMALIZED LEAST SQUARES

Daniel F. Schmidt[1] *and Teemu Roos*[2]

[1] Centre for MEGA Epidemiology, The University of Melbourne,
Melbourne, AUSTRALIA, dschmidt@unimelb.edu.au
[2]Helsinki Institute for Information Technology HIIT, University of Helsinki,
P.O.Box 68, FIN-00014 Helsinki, FINLAND, teemu.roos@cs.helsinki.fi

1. INTRODUCTION

We examine the Sequentially Normalized Least Squares (SNLS) criterion for linear regression model selection. In particular, we present: (i) a simplified formula for computing the SNLS score, (ii) an asymptotic representation of the SNLS score even in the case of model misspecification, and (iii) a proof of the consistency of SNLS as a model selection tool.

Consider a complete design matrix $\mathbf{Z}_n = (\mathbf{z}_1, \ldots, \mathbf{z}_q)$ of covariates $\mathbf{z}_i \in \mathbb{R}^n$, and a corresponding set of observed responses $\mathbf{y} \in \mathbb{R}^n$. It is common to assume that the mean of the responses is a linear combination of the covariates, resulting in the generating model

$$\mathbf{y} = \mathbf{Z}_n \boldsymbol{\beta}_* + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{\beta}_* \in \mathbb{R}^q$ are the regression coefficients and $\varepsilon_i \sim N(0, \sigma^2)$ are independently and identically distributed normal variates. It is often assumed that some (potentially all) of the components of $\boldsymbol{\beta}_*$ are exactly zero (i.e., the associated covariates are unrelated to the response), and the problem examined in this paper is the selection of covariates which are related to the response.

More formally, let $\gamma \subset \{1, \ldots, q\}$ denote an index vector determining which of the covariates comprise the design submatrix $\mathbf{X}_n(\gamma)$, and let Γ denote the set of all candidate subsets. The model indexed by γ describes the data as being generated by

$$\mathbf{y} = \mathbf{X}_n(\gamma)\boldsymbol{\beta}_\gamma + \boldsymbol{\varepsilon}$$

where $\mathbf{X}_n(\gamma)$ is the design matrix given by

$$\mathbf{X}_n(\gamma) = (\mathbf{z}_{\gamma_1}, \ldots, \mathbf{z}_{\gamma_k})$$

and $\boldsymbol{\beta}_\gamma$ is the corresponding vector of coefficients of size $k = |\gamma|$. The problem examined in this paper can then be formulated as one of selecting covariates from the full design matrix $\mathbf{Z}_n$, i.e. choosing a suitable subset γ, and constructing a predictive distribution for future data arising from this model. There are a large range of methods available for performing model selection in the linear regression setting; we choose to study the recently proposed Sequentially Normalized Least Squares criterion (SNLS) [1].

2. SEQUENTIALLY NORMALIZED LEAST SQUARES

The SNLS criterion is closely related to both the principle of Predictive Minimum Description Length (PMDL) [2] and the Sequentially Normalized Maximum Likelihood (SNML) principle [3]. The basic idea is to sequentially define a predictive distribution $p(y_t \mid \mathbf{y}_{1:t-1}, \gamma)$ for data y_t conditioned on all the previously observed data $\mathbf{y}_{1:t-1} = (y_1, \ldots, y_{t-1})$, using the subset of covariates specified by γ. The SNLS criterion is then defined as the accumulated negative-log of the predictive density for the data $y_{m+1}, \ldots, y_n$ with $m \geq |\gamma|$, i.e.

$$\sum_{t=m+1}^{n} \ln 1/p(y_t \mid \mathbf{y}_{1:t-1}, \gamma)$$

The "optimal" subset of covariates may be selected by choosing the one from Γ that yields the smallest SNLS score. The full details of the derivation of the SNLS criterion are given in [1]; for convenience we summarise the results here and present a simplification of the final formula given in [1]. For notational simplicity, we drop the explicit dependence of $\mathbf{X}_n(\gamma)$ on the subset γ. Let $\bar{\mathbf{x}}_i$ denote the i-th *row* of $\mathbf{X}_n$, and let $\mathbf{X}_t = (\bar{\mathbf{x}}_1', \ldots, \bar{\mathbf{x}}_t')'$ denote the matrix comprised of the first $t \leq n$ rows. The SNLS criterion is given by

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

where

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

and

$$\begin{aligned} e_t &= y_t - \bar{x}_t \hat{\boldsymbol{\beta}}_{t-1} \\ \hat{e}_t &= y_t - \bar{x}_t \hat{\boldsymbol{\beta}}_t = (1 - d_t) e_t \\ \mathbf{J}_t &= \mathbf{X}_t' \mathbf{X}_t \\ \hat{\boldsymbol{\beta}}_t &= \mathbf{J}_t^{-1} \mathbf{X}_t' \mathbf{y}_{1:t} \\ 1 - d_t &= 1/(1 + c_t) \\ c_t &= \bar{\mathbf{x}}_t' \mathbf{J}_{t-1}^{-1} \bar{\mathbf{x}}_t \end{aligned}$$

with the notation $\mathbf{y}_{1:t} = (y_1, \ldots, y_t)$. The $\hat{\boldsymbol{\beta}}_t$ are the least-squares estimates for the first t data points, and the complete recurrence relations given in, for instance [4], offer an efficient way of computing them. As $\hat{\boldsymbol{\beta}}_t$ is not unique for $t < k$ the sequential process must start at sample $m + 1 \geq k + 1$, where $k = |\gamma|$.

The SNLS criterion as given by (2) can be simplified by noting that [5]

$$1 - d_t = \frac{|\mathbf{X}'_{t-1}\mathbf{X}_{t-1}|}{|\mathbf{X}'_t\mathbf{X}_t|} = \frac{|\mathbf{J}_{t-1}|}{|\mathbf{J}_t|}$$

Then it is clear that $1 + c_t = |\mathbf{J}_t|/|\mathbf{J}_{t-1}|$, and the second term in (2) can be written as

$$\sum_{t=m+1}^{n} \ln(1 + c_t) = \ln \frac{|\mathbf{J}_{m+1}|}{|\mathbf{J}_m|} \cdot \frac{|\mathbf{J}_{m+2}|}{|\mathbf{J}_{m+1}|} \cdots \frac{|\mathbf{J}_n|}{|\mathbf{J}_{n-1}|}$$

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

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

$$\begin{aligned} \mathrm{SNLS}(\mathbf{y}, \gamma) \quad &= \left(\frac{n-m}{2}\right) \ln(2\pi e \hat{\tau}_n) \\ &\quad + \ln \frac{|\mathbf{J}_n|}{|\mathbf{J}_m|} + \frac{1}{2} \ln n + O(1). \end{aligned} \tag{3}$$

Using this form of the criterion, the proof of the large sample behaviour of SNLS (Theorem 1 from [1]) is trivial.

Theorem 1 (Rissanen *et al.*, [1]) *Under the assumption that*

$$\lim_{n\to\infty} \left\{ \frac{1}{n} \mathbf{Z}'_n \mathbf{Z}_n \right\} = \boldsymbol{\Lambda} \tag{4}$$

with $\boldsymbol{\Lambda}$ a positive-definite matrix, then

$$\begin{aligned} \mathrm{SNLS}(\mathbf{y}, \gamma) \quad &= \left(\frac{n-m}{2}\right) \ln(2\pi e \hat{\tau}_n) \\ &\quad + \left(\frac{2k+1}{2}\right) \ln n + O(1) \end{aligned} \tag{5}$$

Proof. Noting that by (4), $|\mathbf{J}_n| = O(n^k)$ and $|\mathbf{J}_m| = O_n(1)$, and using these in (3) completes the proof. □

3. LARGE-SAMPLE BEHAVIOR

We let $\mathbf{R}$ be the $(q \times q)$ idempotent matrix with entries $r_{i,i} = 1$ iff $i \in \gamma$, and zero otherwise. Again, assume that limit (4) exists; recalling that the data $\mathbf{y}$ is generated by the linear-normal model (1), with $\boldsymbol{\beta}_*$ denoting the true, underlying regression coefficients, this assumption implies that

$$\begin{aligned} \lim_{n\to\infty} \left\{ \frac{1}{n} (\mathbf{Z}_n\mathbf{R})'(\mathbf{Z}_n\mathbf{R}) \right\} &= \mathbf{R}\boldsymbol{\Lambda}\mathbf{R}, \\ \lim_{n\to\infty} \left\{ \frac{1}{n} (\mathbf{Z}_n\mathbf{R})'(\mathbf{Z}_n\boldsymbol{\beta}_*) \right\} &= \mathbf{R}\boldsymbol{\Lambda}\boldsymbol{\beta}_*. \end{aligned}$$

Further assume that

$$\sup_n ||\bar{\mathbf{x}}_n|| < \infty. \tag{6}$$

We now give a general expression for the large sample behaviour of the SNLS criterion, even in the case of misspecification, i.e., when γ omits covariates which are related to the response (with corresponding non-zero entries in $\boldsymbol{\beta}_*$).

Theorem 2 *Under assumptions (4) and (6), the SNLS criterion satisfies*

$$\begin{aligned} \mathrm{SNLS}(\mathbf{y}, \gamma) = &\left(\frac{n-m}{2}\right) \ln(2\pi e \hat{\sigma}_n^2) \\ &+ \left(\frac{2k - \frac{k(\sigma^2+\xi)}{\hat{\sigma}_n^2} + 1}{2}\right) \ln n + o(\ln n). \end{aligned} \tag{7}$$

where $\xi > 0$ is the error due to misspecification, and $\hat{\sigma}_n^2 = \left(\frac{1}{n}\right) \sum_{i=1}^n (y_i - \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}_n)^2$.

Proof. Note that $\mathbf{R}^- = \mathbf{R}$, where $(\cdot)^-$ denotes the pseudo-inverse; using this we may define the quantities

$$\boldsymbol{\delta} = \mathbf{X}_n\boldsymbol{\beta}_* - \mathbf{X}_n(\mathbf{R}\boldsymbol{\Lambda}\mathbf{R})^-(\boldsymbol{\Lambda}\boldsymbol{\beta}_*)$$

and

$$\tilde{\boldsymbol{G}} = \lim_{n\to\infty} \left\{ \frac{1}{n} (\mathbf{X}_n\mathbf{R})' \boldsymbol{\Delta} (\mathbf{X}_n\mathbf{R}) \right\}$$

where $\boldsymbol{\Delta}$ is an $(n \times n)$ diagonal matrix with entries $\Delta_{i,i} = \delta_i^2$. Theorem 4.1.1 in [5] gives the asymptotic (a.s.) form of the sum of squared errors for the predictive least squares method as:

$$\sum_{i=m+1}^{n} e_i^2 = n\hat{\sigma}_n^2 + (\ln n)[k(\sigma^2+\xi)](1+o(1)) \quad \text{a.s.,} \tag{8}$$

where $\xi = \mathrm{Tr}((\mathbf{R}\boldsymbol{\Lambda}\mathbf{R})^-\tilde{\mathbf{G}})/k$ is the per sample squared error due to misspecification.

We now need a similar result for the SNLS errors (instead of the PLS errors), namely for the sum

$$\sum_{i=m+1}^{n} \hat{e}_i^2 = \sum_{i=m+1}^{n} (e_i^2 - 2d_i e_i^2 + d_i^2 e_i^2). \tag{9}$$

The first term inside the parentheses, involving the PLS errors, e_i^2 is given by (8) above. The middle term appears in Thm. 2.1 in [5]:

$$\sum_{i=m+1}^{n} e_i^2 = n\hat{\sigma}_n^2 - m\hat{\sigma}_m^2 + \sum_{i=m+1}^{n} d_i e_i^2, \tag{10}$$

where the last term on the RHS is the one we need (just negated and multiplied by two).

Rissanen *et al.* present a bound on the third term in (9); [1, Eq. (34)]

$$\sum_{i=m+1}^{n} d_t^2 e_t^2 = o(\ln n),$$

showing that the third term is negligible.

By (8), (9), and (10), the sum of SNLS errors is given by

$$\sum_{i=m+1}^{n} \hat{e}_t^2 = n\hat{\sigma}_n^2 - (\ln n)[k(\sigma^2+\xi)](1+o(1)).$$

Dividing by $n\hat{\sigma}_n^2$ and taking the log gives

$$\ln \frac{1}{n\hat{\sigma}_n^2} \sum_{i=m+1}^{n} \hat{e}_i^2 = \ln\left(1 - \left(\frac{\ln n}{n}\right)\left[\frac{k(\sigma^2+\xi)}{\hat{\sigma}_n^2}\right](1+o(1))\right).$$

Applying the Taylor approximation $\ln(1+x) = x + O(x^2)$ to the right-hand side yields

$$\ln \sum_{i=m+1}^{n} \hat{e}_i^2 = \ln n\hat{\sigma}_n^2 - \left(\frac{\ln n}{n}\right)\left[\frac{k(\sigma^2+\xi)}{\hat{\sigma}_n^2}\right](1+o(1)).$$

Subtracting $\ln(n-m) = \ln n + O_n(1/n)$ from both sides, and recalling the definition

$$\hat{\tau}_n = \left(\frac{1}{n-m}\right)\sum_{i=m+1}^{n} \hat{e}_i^2$$

we arrive at the formula:

$$\ln \hat{\tau}_n = \ln \hat{\sigma}_n^2 - \left(\frac{\ln n}{n}\right)\left[\frac{k(\sigma^2+\xi)}{\hat{\sigma}_n^2}\right](1+o(1)). \quad (11)$$

Combining (11) with (5) yields the asymptotic formula (7) and completes the proof. □

4. CONSISTENCY OF SNLS

In terms of model selection, consistency, or the guarantee that a criterion will discover the truth as the sample size grows is an important property. Formally, let γ_* denote the index of the components of $\boldsymbol{\beta}_*$ that are non-zero; we shall call the associated covariates the "relevant covariates". Define the SNLS estimate of the true model index as

$$\hat{\gamma} = \arg\min_{\gamma\in\Gamma} \{\mathrm{SNLS}(\mathbf{y}, \gamma)\}$$

that is, the chosen subset is the one that minimises the SNLS criterion. The asymptotic formula (7) provides a simple basis to prove the consistency of the SNLS estimate of γ.

Corrolary 1. *Assuming (4) and (6), and assuming that the data is generated by the linear-quadratic model (1) with* $\gamma_* \in \Gamma$*, then*

$$\Pr(\hat{\gamma} = \gamma_*) \to 1 \text{ as } n \to \infty$$

Proof. To complete this proof we use the result that the maximum likelihood estimate of σ^2 using the restricted least squares (RLS) estimates of $\boldsymbol{\beta}$, restricted to the subspace defined by γ, can be expressed as

$$\hat{\sigma}_n^2 = \sigma^2 + \xi + o(1). \quad (12)$$

This is a result of the consistency of the RLS estimates in estimating the parameter vector in the restricted subspace defined by γ that is closest in a weighted quadratic sense to the true parameter vector $\boldsymbol{\beta}_*$, i.e., it minimises the error due to misspecification. Using (12) in (7) yields

$$\mathrm{SNLS}(\mathbf{y}, \gamma) = \left(\frac{n-m}{2}\right)\ln(2\pi e\hat{\sigma}_n^2) + \left(\frac{2k - \frac{k(\sigma^2+\xi)}{\sigma^2+\xi+o(1)} + 1}{2}\right)\ln n + o(\ln n). \quad (13)$$

where $k = |\gamma|$ is the number of regression parameters being estimated. It is clear that asymptotically, the second term in the right hand side of (13) reduces to $(k+1)/2 \log n$, and the entire SNLS criterion is asymptotically equivalent to BIC. Using the results from [6] regarding the consistency of BIC completes the proof.

ACKNOWLEDGMENTS

The authors thank Jorma Rissanen and Enes Makalic for useful discussions. This work was supported in part by the Academy of Finland under project Modest, and the IST Programme of the European Community, under the PASCAL Network of Excellence.

5. REFERENCES

[1] J. Rissanen, T. Roos, and P. Myllymäki, "Model selection by sequentially normalized least squares," *Journal of Multivariate Analysis*, vol. 101, no. 4, pp. 839–849, 2010.

[2] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. World Scientific, 1989.

[3] T. Roos and J. Rissanen, "On sequentially normalized maximum likelihood models," in *Proc. 1st Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)*, Tampere International Center for Signal Processing, 2008, (Invited Paper).

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

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

[6] D. M. A. Haughton, "On the choice of a model to fit data from an exponential family," *The Annals of Statistics*, vol. 16, no. 1, pp. 342–355, March 1988.