# Robust Sequential Prediction in Linear Regression with Student's *t*-distribution

**Jussi Määttä** and **Teemu Roos**
Helsinki Institute for Information Technology HIIT
Department of Computer Science
University of Helsinki, Finland

## Abstract

The Predictive Least Squares (PLS) model selection criterion is known to be consistent in the context of linear regression. For small sample sizes, however, it can exhibit erratic behavior. We show that this shortcoming can be amended by incorporating a Student's $t$-distribution into PLS. The resulting criterion is shown to be asymptotically equivalent to PLS but significantly more robust for small sample sizes. A scale parameter involved with the $t$-distribution can be used to incorporate an estimate of the scale of the noise; it is shown that the new criterion is robust with regard to the choice of this parameter and that its effect disappears asymptotically. The recently proposed Sequentially Normalized Least Squares (SNLS) criterion can be written in a form that exposes a similar interpretation with the exception that the scale parameter of the $t$-distribution is estimated sequentially from the data. Numerical experiments are presented; they indicate that using a Student's $t$-distribution enhances model selection performance and that the benefit of the scale estimator of SNLS is negligible.

## Introduction

Linear regression has recently received attention in the *sequential* or *online* setting, where work has been done in selecting a subset of the covariates (Määttä, Schmidt, and Roos 2015) and finding a predictor that minimizes the worst-case regret (Bartlett et al. 2015). The probabilistic case that we consider also fits to the *prequential* framework of Dawid (1984).

In this article, we concentrate on the subset selection problem, also called the model selection problem. We assume a fixed design matrix $\boldsymbol{Z}_n \in \mathbb{R}^{n\times q}$ that consists of the row vectors $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$. Associated with each sample $\boldsymbol{z}_t$, we have a response $y_t \in \mathbb{R}$. The goal is to select a non-empty subset of the covariates, $\gamma \subseteq \{1, 2, \ldots, q\}$, that strikes a good balance between underfitting (poor prediction of the training data) and overfitting (poor generalization for future data).

In order to assess the performance of subset selection methods, one often introduces the assumption that the data $(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n)$ comes from the linear model

$$y_t = \boldsymbol{z}_t \boldsymbol{\beta} + \varepsilon_t, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is a fixed coefficient vector and the $\varepsilon_t$'s are i.i.d. noise with $\mathrm{E}[\varepsilon_t] = 0$ and $\mathrm{E}[\varepsilon_t^2] = \sigma^2 < \infty$. In this setting, a subset selection method is said to be *consistent* if its probability of selecting the $\gamma$ that corresponds exactly to the non-zero elements of $\boldsymbol{\beta}$ will approach one when the sample size $n$ tends to infinity. This $\gamma$ is referred to as the *true* model or subset.

For the batch case, where the score of a subset $\gamma$ cannot be represented in a sequential manner, there are numerous methods (McQuarrie and Tsai 1998). Perhaps the most well-known of these is the Bayesian Information Criterion (Akaike 1978; Schwarz 1978), or BIC, which is also known as the Schwarz Information Criterion (SIC). For the model (1), the BIC criterion is

$$\mathrm{BIC}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma) := n \log\left(\widehat{\sigma}^2_{n,\gamma}\right) + |\gamma| \log n, \tag{2}$$

where $|\gamma|$ is the cardinality of $\gamma$ and

$$\widehat{\sigma}^2_{n,\gamma} := \frac{1}{n} \sum_{t=1}^{n} \left(y_i - \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}_n^{(\gamma)}\right)^2. \tag{3}$$

Here and later $\widehat{\boldsymbol{\beta}}_n^{(\gamma)} \in \mathbb{R}^q$ denotes the maximum likelihood estimate of $\boldsymbol{\beta}$ computed using the first $n$ samples and with the restriction that the entries of $\widehat{\boldsymbol{\beta}}_n^{(\gamma)}$ that are not present in $\gamma$ are forced to be zeros.

As for information criteria based on sequential prediction, we are aware of only two (besides Bayesian methods that admit a sequential interpretation). The first is the Predictive Least Squares (PLS) criterion, introduced by Rissanen (1986), which is defined as

$$\begin{aligned} \mathrm{PLS}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma) &:= \sum_{t=m+1}^{n} e^2_{t,\gamma} \\ &:= \sum_{t=m+1}^{n} \left(y_t - \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}_{t-1}^{(\gamma)}\right)^2. \end{aligned} \tag{4}$$

The starting index $m$ is typically set to $q$ in order to make $\widehat{\boldsymbol{\beta}}_m^{(\gamma)}$ uniquely defined for all $\gamma$. Note that each term of the sum depends only on the samples seen so far. PLS is known to be consistent under certain reasonable regularity assumptions (Wei 1992). The second sequential method is the Sequentially Normalized Least Squares (SNLS) criterion (Rissanen, Roos, and Myllymäki 2010). Its derivation is based upon the idea of considering the errors $\widehat{e}_{t,\gamma} := y_t - \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}_t^{(\gamma)}$ and using them to construct a predictive distribution for each sample.

In the following, we start by discussing the similarities and differences of PLS and SNLS. We will observe that PLS can be seen as using a Gaussian distribution for prediction, while SNLS relies on Student's $t$-distribution and additionally a sequential scale parameter estimator. We will attempt to isolate the effect of the $t$-distribution by constructing a "hybrid" criterion which uses the $t$-distribution but replaces the scale parameter estimator with a constant. We will show that the hybrid is asymptotically equivalent to PLS and hence consistent. We then present results from numerical simulations to demonstrate that SNLS and the hybrid are indeed more robust than PLS and the benefit of SNLS's scale parameter estimator is negligible. Finally, we discuss the results and suggest some open problems for further study.

# PLS and SNLS

## The Predictive Least Squares Criterion

For an arbitrary constant $\lambda^2 > 0$, the PLS criterion (4) has the form

$$\frac{\mathrm{PLS}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma)}{2\lambda^2} + \frac{(n-m)\log\left(2\pi\lambda^2\right)}{2} = -\log \prod_{t=m+1}^{n} f\left(y_t \mid \mu = \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}_{t-1}^{(\gamma)}, \lambda^2\right), \quad (5)$$

where $f(\,\cdot \mid \mu, \lambda^2)$ is the Gaussian density function with mean $\mu$ and variance $\lambda^2$. Note that the affine transformation of PLS on the left-hand side of (5) does not affect subset selection. Therefore we may interpret PLS as doing sequential prediction: to each new observation, it assigns a Gaussian predictive density with a mean that depends on the previously seen samples and an arbitrary fixed scale parameter.

We now introduce three regularity assumptions:

$$\sup_{n,i} |z_{n,i}| < \infty, \quad (6)$$

$$\lim_{n\to\infty} \left(\frac{1}{n}\boldsymbol{Z}_n^{\mathrm{T}} \boldsymbol{Z}_n\right) = \boldsymbol{\Lambda}, \quad \text{with } \boldsymbol{\Lambda} \text{ positive definite}, \quad (7)$$

$$\lim_{n\to\infty} \left(\frac{1}{n}\sum_{t=1}^{n} z_{t,i} z_{t,j} z_{t,k} z_{t,\ell}\right) = \omega_{ijk\ell} \in \mathbb{R} \quad \text{for all } i,j,k,\ell \in \{1,2,\ldots,q\}. \quad (8)$$

In other words, we assume that the design matrix is bounded (6) and has a well-behaving covariance structure (7). Assumption (7) is sufficient for PLS to be strongly consistent (Wei 1992, Theorem 3.4). The fourfold products in (8) are more difficult to interpret, but a clear special case is the existence of the limits $\lim_{n\to\infty}(1/n)\sum_{t=1}^{n} z_{t,j}^4$ for all $j$; in any case, assumption (8) may be considered optional since is only required in the consistency proof of SNLS and none of our results rely on it.

The following theorem describes the behavior of $\mathrm{PLS}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma)$ as the sample size grows. We will require this result later in order to show that our PLS/SNLS hybrid criterion is consistent.

**Theorem 1.** *Assume* (6) *and* (7)*. Then there exist some* $0 \le \xi_1 \le \xi_2 < \infty$*, dependent on* $\gamma$*, for which the following inequalities hold:*

$$\begin{aligned}\sigma^2 + \xi_1 &\le \operatorname*{p\text{-}lim\,inf}_{n\to\infty} \left(\frac{\mathrm{PLS}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma)}{n-m}\right) \\ &\le \operatorname*{p\text{-}lim\,sup}_{n\to\infty} \left(\frac{\mathrm{PLS}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma)}{n-m}\right) \\ &\le \sigma^2 + \xi_2.\end{aligned}$$

*If we also assume* (8)*, the above inequalities hold with almost sure convergence and* $\xi_1 = \xi_2$.

We postpone the proofs of all theorems to the appendix.

## The Sequentially Normalized Least Squares Criterion

The sequentially normalized least squares (SNLS) criterion was introduced by Rissanen et al. (2010) and its asymptotic theory was studied by Määttä et al. (2015). All results presented in this section may be found in the latter article.

SNLS is based on the idea of using not only the first $t-1$ but also the $t$'th sample to predict the $t$'th response. Where PLS uses the errors $e_{t,\gamma} = y_t - \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}_{t-1}^{(\gamma)}$, SNLS instead considers the terms

$$\widehat{e}_t := y_t - \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}_t^{(\gamma)}. \quad (9)$$

Hence, it appears to "cheat" by using $y_t$ to help predict $y_t$. However, this is not the full story. The original derivation of SNLS assigns Gaussian densities with a common fixed variance to the $\widehat{e}_{t,\gamma}$'s and optimizes the variance parameter over the product of the densities. The end result of the derivation can be written as

$$\mathrm{SNLS}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma) = -\log \prod_{t=m+2}^{n} g\Bigg(y_t \mid \nu = t-m-1, \mu = \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}_{t-1}^{(\gamma)}, \lambda^2 = \frac{\widehat{\tau}_{t-1,\gamma}}{(1-d_{t,\gamma})^2}\Bigg), \quad (10)$$

where $g(\,\cdot \mid \nu, \mu, \lambda^2)$ is the density of the non-standardized Student's $t$-distribution:

$$g(y \mid \nu, \mu, \lambda^2) := \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\lambda^2}} \left(1 + \frac{1}{\nu}\frac{(y-\mu)^2}{\lambda^2}\right)^{-\frac{\nu+1}{2}},$$

and where the building blocks of $\lambda^2$ are

$$1 - d_{t,\gamma} = \frac{\det(\boldsymbol{Z}_{t-1,\gamma}^{\mathrm{T}} \boldsymbol{Z}_{t-1,\gamma})}{\det(\boldsymbol{Z}_{t,\gamma}^{\mathrm{T}} \boldsymbol{Z}_{t,\gamma})} \quad \text{and}$$

$$\widehat{\tau}_{t-1,\gamma} = \frac{1}{t-m-1} \sum_{s=m+1}^{t-1} (1-d_{s,\gamma})^2 e_{s,\gamma}^2,$$

where $\boldsymbol{Z}_{t,\gamma}$ indicates the submatrix of $\boldsymbol{Z}_t$ with the columns indexed by $\gamma$. Note that in (10), the predictive distribution used for $y_t$ does not use $y_t$ in its parameters, so the apparent cheating in (9) disappears.

Consider the parameters to the Student distribution in (10). The integer $\nu = t-m-1$ makes the predictive distribution closer and closer to a Gaussian density as the number of samples increases. The mean parameter $\mu$ takes the same value as in PLS. The scale parameter $\lambda^2$, which asymptotically becomes the variance of the predictive distribution, consists of two terms with the following interpretations: If the $\varepsilon_t$'s follow a zero-mean normal distribution, the number $d_{t,\gamma}$ can be interpreted as the Fisher information ratio of $\boldsymbol{z}_{t,\gamma}$ with respect to $\boldsymbol{Z}_{t,\gamma}$ (Wei 1992, pp. 4–5) and will tend to zero in all reasonable situations. The variance estimator $\widehat{\tau}_{n,\gamma}$ agrees with $\widehat{\sigma}^2_{n,\gamma}$ in the limit.

Equation (10) can be written in the computationally simpler form

$$\begin{aligned}\mathrm{SNLS}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma) = {} & \left(\frac{n-m}{2}\right)\log\left(\widehat{\tau}_{n,\gamma}\right) \\ & + \log\left(\frac{\det(\boldsymbol{Z}^{\mathrm{T}}_{n,\gamma}\boldsymbol{Z}_{n,\gamma})}{\det(\boldsymbol{Z}^{\mathrm{T}}_{m,\gamma}\boldsymbol{Z}_{m,\gamma})}\right) \\ & - \frac{1}{2}\log\left(e^2_{m+1,\gamma}\right) + c(n,m),\end{aligned} \tag{11}$$

where $c(n,m)$ does not depend on the data and thus does not affect subset selection. There are also various asymptotic simplifications of SNLS. In particular, we have

$$\begin{aligned}\mathrm{SNLS}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma) = {} & \left(\frac{n-m}{2}\right)\log\left(2\pi e\widehat{\tau}_{n,\gamma}\right) \\ & + \left(\frac{2|\gamma|+1}{2}\right)\log n + o(\log n).\end{aligned} \tag{12}$$

Since $m$ is a constant and $\widehat{\tau}_{n,\gamma}$ is convergent, we may consider the further simplified form

$$\mathrm{SNLSa}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma) := n\log\left(\widehat{\tau}_{n,\gamma}\right) + 2|\gamma|\log n \tag{13}$$

which will still select the same subset as (10) when the sample size is large enough.

It is interesting to compare (13) to BIC (see eq. (2)). It can be seen that SNLS penalizes twice as much for the number of parameters. On the other hand, if $\gamma$ contains the true model, then $n\log(\widehat{\tau}_{n,\gamma}) = n\log(\widehat{\sigma}^2_{n,\gamma}) - |\gamma|\log n + o(\log n)$ almost surely and SNLS becomes asymptotically equivalent to BIC.

We also mention that SNLS was originally derived with the assumption that the noise terms $\varepsilon_t$ follow a Gaussian distribution; remarkably, however, all the asymptotic results hold for all noise distributions with a finite fourth moment.

## The PLS/SNLS Hybrid Criterion

Compare the log-product forms of PLS (5) and SNLS (10). They are similar in that they use the same location parameter $\mu_t = \boldsymbol{z}_t\widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}$ for the $t$'th sample, but they differ in the choice of the predictive distribution and whether they attempt to estimate the scale of the responses $y_t$. We may attempt to isolate the effect of the predictive distribution by replacing the Gaussian distribution in PLS by a Student's $t$-distribution with a constant scale parameter. The resulting criterion is

$$\begin{aligned}&\mathrm{Hybrid}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma) := \\ &-\log\prod_{t=m+2}^{n} g\left(y_t \mid \nu = t-m-1, \mu = \boldsymbol{z}_t\widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}, \lambda^2\right)\end{aligned} \tag{14}$$

where $\lambda^2 > 0$ is a fixed constant. The only difference between (14) and SNLS is that the scale parameter estimator has been replaced by a fixed value. Since the variance of the non-standardized Student's $t$-distribution is $\nu\lambda^2/(\nu-2)$ for $\nu > 2$, the value $\lambda^2$ may be interpreted as the noise variance.

The hybrid can also be written in the form

$$\begin{aligned}&\mathrm{Hybrid}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma) = h(n,m,\lambda^2) \\ &+ \sum_{t=m+2}^{n}\left(\frac{t-m}{2}\right)\log\left(1 + \frac{(y_t - \boldsymbol{z}_t\widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1})^2}{\lambda^2(t-m-1)}\right)\end{aligned} \tag{15}$$

where the function $h(n,m,\lambda^2)$ hides terms independent of the data. Equation (15) is highly reminiscent of PLS in its traditional form (4). Note that unlike with PLS, the scale parameter $\lambda^2$ does in general affect the relative scores assigned by the hybrid to various models.

The following theorem shows that the hybrid is asymptotically equivalent to PLS.

**Theorem 2.** *Denote*

$$\widehat{\gamma}_{\mathrm{PLS}}(n) = \arg\min_{\gamma\in\Gamma}\mathrm{PLS}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma) \quad \textit{and}$$

$$\widehat{\gamma}_{\mathrm{Hybrid}}(n) = \arg\min_{\gamma\in\Gamma}\mathrm{Hybrid}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma).$$

*Then we have*

$$\lim_{n\to\infty}\Pr\left[\widehat{\gamma}_{\mathrm{PLS}}(n) = \widehat{\gamma}_{\mathrm{Hybrid}}(n)\right] = 1.$$

## Numerical Experiments

In order to evaluate the performance of various criteria, we performed an experiment with synthetic data sets. For each pair $(n,k)$, with $n = 100, 120, \ldots, 200$ and $k = 1, 2, \ldots, 10$, we generated one thousand data sets as follows. First, the elements of a vector $\boldsymbol{\mu} \in \mathbb{R}^9$ are drawn independently from $\mathrm{Cauchy}(0,1)$, and a covariance matrix $\boldsymbol{\Sigma}$ is drawn from $\mathrm{Wishart}(9^{-1}\boldsymbol{I}_9, 9)$ in order to produce correlations between variables. Then the row vectors $\boldsymbol{z}_{i,2:10}$ are drawn from the multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The intercept is included by setting each $z_{i,1}$ to one. The first $k$ elements of $\boldsymbol{\beta}$ are drawn from the distribution obtained by restricting $\mathcal{N}(0,1)$ to the domain outside the interval $(-1,1)$, and the remaining $10-k$ elements of $\boldsymbol{\beta}$ are set to zeros. The response vector is obtained by $\boldsymbol{y}_{1:n} = \boldsymbol{Z}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{1:n}$, where the elements of $\boldsymbol{\varepsilon}_{1:n}$ are drawn independently from the zero-mean Laplace distribution with a standard deviation of two.

The purpose of disallowing coefficients from the interval $(-1,1)$ is to make the model selection problem more tractable by ensuring that the effects of the variables in the true model are non-negligible even for small sample sizes. For the noise terms, we have used the Laplace distribution because its heavy tails should provide more challenge to the robustness of the various criteria. The Laplace distribution also has a finite fourth moment, as required by the consistency proof of SNLS.

For every data set, we computed the scores of all subsets $\gamma \supseteq \{1\}$ given by BIC (2), PLS (4), SNLS (10), SNLSa (13), and the hybrid (14). For the hybrid criterion, we used three different values for the scale parameter: $\lambda^2 = 0.01$, $\lambda^2 = 1$, and $\lambda^2 = 100$. We then noted the rank assigned to the true subset $\gamma = \{1, 2, \ldots, k\}$ by each criterion. Figure 1 shows the median rank of the true model. The error bars show the upper and lower quartiles. We display the pooled results for all $k = 1, 2, \ldots, 10$, and also the individual plots for $k = 1, 5, 10$.

It should be noted that our inclusion of BIC among the criteria studied is primarily to provide a baseline. Since BIC is not a predictive criterion, its results are not directly comparable to those of PLS, SNLS and the hybrid.

In the case $k = 1$, where the responses are just a constant plus noise, all criteria except the hybrid with $\lambda^2 = 0.01$ perform extremely well. BIC and SNLSa slightly outperform the others. PLS tends to favor simple models and accordingly performs better than SNLS.

For $k = 5$, the results for the various criteria are quite similar, again with the exception of the hybrid with $\lambda^2 = 0.01$. BIC and SNLSa are still the winners, with the hybrid with $\lambda^2 = 1$ being slightly better than the rest.

In the complex model case $k = 10$, one starts to see more differences: BIC is matched with the hybrid with $\lambda^2 \in \{0.01, 1\}$. PLS is clearly the worst of the criteria considered, with the rest clustered in between.

When the results for all $k = 1, 2, \ldots, 10$ are combined, it becomes clear that BIC performs the best, with SNLSa and the hybrid with $\lambda^2 = 1$ competing for the second place. PLS and the hybrid with $\lambda^2 = 0.01$ perform slightly worse than others. The rest of the criteria are quite close to each other.

## Discussion and Open Problems

We have proposed a new subset selection criterion for linear regression. The criterion is essentially a robust variant of the PLS criterion, and it can also be interpreted as a simplification of SNLS; hence, it may be viewed as a PLS/SNLS hybrid. We have shown that the new criterion is asymptotically equivalent to PLS and thus consistent.

Our numerical results indicate that the hybrid usually outperforms PLS, especially when the underlying model is complex. Of the criteria than can be written in the log-product form, the hybrid with $\lambda^2 = 1$ outperforms not only PLS but also SNLS. This is perhaps not surprising as the parameter $\lambda^2$ can be interpreted as an estimate of the noise variance and the variance of the noise used in the simulated data was always $\mathrm{Var}[\varepsilon_i] = 4$. However, it should be noted that the hybrid performed quite well with $\lambda^2 = 100$, so the choice of the scale parameter appears to be quite robust at least to overestimation.

Our results indicate that it is possible for a sequential criterion to match and even exceed the performance of a batch criterion such as BIC (see the $k = 10$ case in Figure 1). We are optimistic that future developments in sequential criteria will bring further improvements.

As for directions for future research, we have made some preliminary experiments with a modification of the hybrid that optimizes the scale parameter $\lambda^2$ to minimize eq. (14). The optimization problem might be convex, but we do not think it has a closed-form solution. Proving or disproving consistency of this criterion, as well as evaluating its empirical performance, would be an interesting topic of study. Alternatively, one might try replacing the scale parameter estimator of SNLS by the simpler $\widehat{\sigma}^2_{n,\gamma}$ and see if it affects consistency or subset selection performance. Also of interest is the effect of the Fisher information ratio in the SNLS scale estimator.

## Acknowledgments

The authors thank the anonymous reviewers for useful comments.

This work was funded by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN).

## Appendix: Proofs of Theorems

The proofs of our theorems require the following result concerning the PLS error terms $e_t$.

**Lemma 3.** *The summands of PLS satisfy* $\mathrm{E}[e^2_{t,\gamma}] = \sigma^2 + \mathrm{E}[(\boldsymbol{z}_t(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}))^2]$, *and if we assume* (6) *and* (7)*, then* $\sup_t \mathrm{E}[e^2_{t,\gamma}] < \infty$.

*Proof.* We decompose

$$\begin{aligned} e^2_{t,\gamma} &= \left(y_t - \boldsymbol{z}_t\widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}\right)^2 \\ &= \left(y_t - \boldsymbol{z}_t\boldsymbol{\beta} + \boldsymbol{z}_t\boldsymbol{\beta} - \boldsymbol{z}_t\widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}\right)^2 \\ &= \left[\varepsilon_t + \boldsymbol{z}_t\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}\right)\right]^2 \\ &= \underbrace{\varepsilon^2_t}_{\text{(a)}} + \underbrace{2\varepsilon_t\boldsymbol{z}_t\boldsymbol{\beta}}_{\text{(b)}} - \underbrace{2\varepsilon_t\boldsymbol{z}_t\widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}}_{\text{(c)}} + \underbrace{\left[\boldsymbol{z}_t\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}\right)\right]^2}_{\text{(d)}}. \end{aligned} \tag{16}$$

Clearly $\mathrm{E}[\varepsilon^2_t] = \sigma^2$ and $\mathrm{E}[2\varepsilon_t\boldsymbol{z}_t\boldsymbol{\beta}] = 0$. The expected value of (c) is also zero due to the fact that the random variables $\varepsilon_t$ and $\boldsymbol{z}_t\widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}$ are independent.

Under assumptions (6) and (7), it can be shown that the vectors $\widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}$ converge to a limit almost surely (Määttä, Schmidt, and Roos 2015). This and (6) imply that the terms $[\boldsymbol{z}_t(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1})]^2$ are uniformly bounded with respect to $t$. □

*Proof of Theorem 1.* Consider first the case where we include assumption (8). Then the claim follows from Theorem 4.1.1 of Wei (1992), which states that

$$\begin{aligned} &\mathrm{PLS}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma) = \\ &\qquad n\widehat{\sigma}^2_\gamma + (\log n)\left[p\sigma^2 + \mathrm{tr}(\Gamma^{-1}\widetilde{G})\right](1 + o(1)) \quad \text{a.s.} \end{aligned}$$

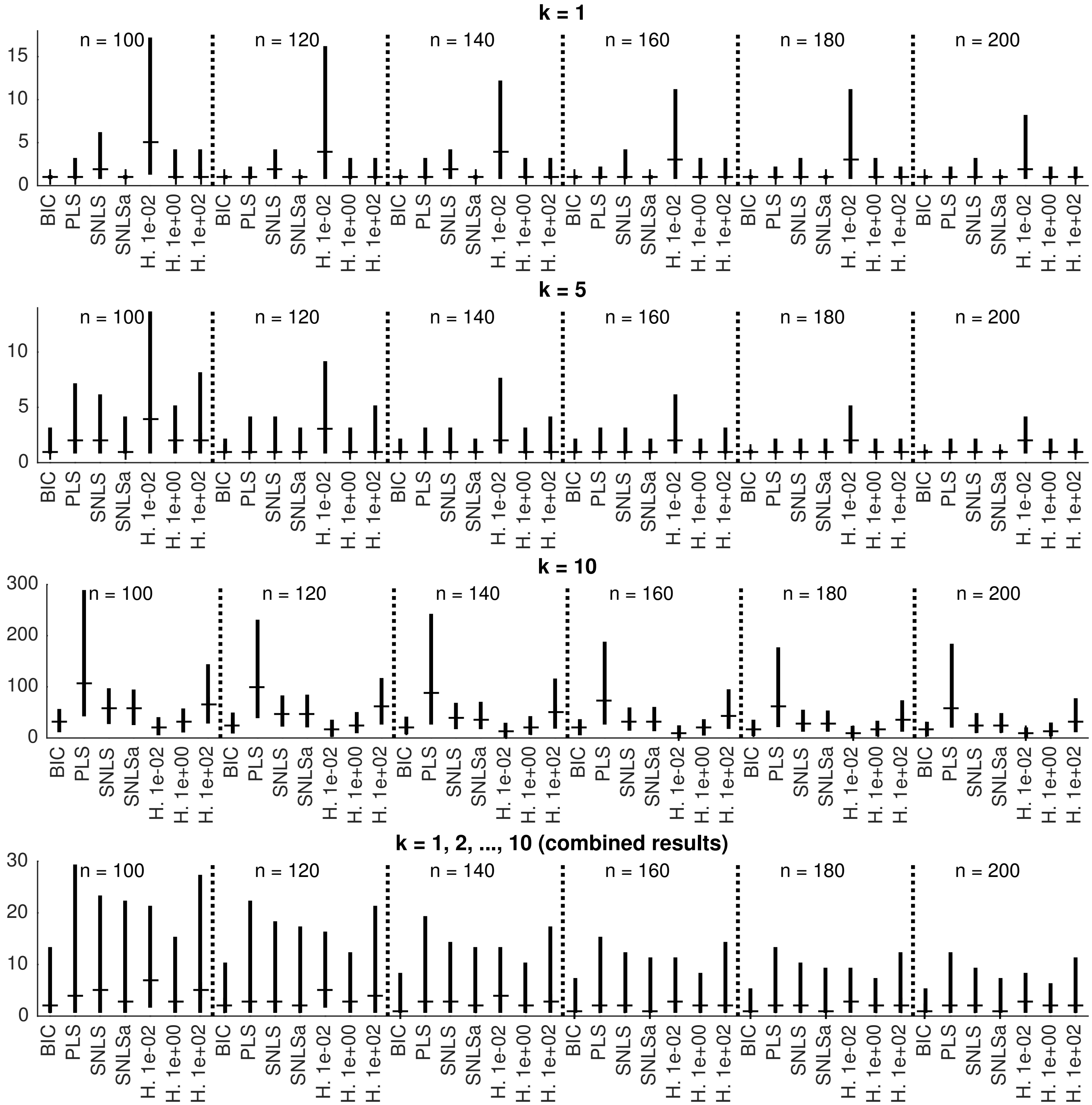

Figure 1: The rank of the true model, as given by various criteria, plotted against increasing sample sizes. Displayed are the median ranks over 1000 experiments; the error bars show the upper and lower quartiles.

where $\widehat{\sigma}^2_\gamma$ is defined as in (3), $p = |\gamma|$, and $\Gamma^{-1}$ and $\widetilde{G}$ are limit matrices (independent of $n$) whose existence is implied by assumptions (7) and (8), respectively (Määttä, Schmidt, and Roos 2015). It is known that $\widehat{\sigma}^2_\gamma \to \sigma^2 + \xi$ almost surely as $n \to \infty$, where the value of $\xi \geq 0$ depends on $\gamma$. The claim follows with $\xi = \xi_1 = \xi_2$.

For the case where assumption (8) does not hold, we separately consider each of the components (a)–(d) of (16).

*(a)* By the strong law of large numbers, it holds almost surely that $(n-m)^{-1}\sum_{t=m+1}^{n} \varepsilon_t^2 \to \sigma^2$ as $n \to \infty$.

*(b)* The terms $2\varepsilon_t \boldsymbol{z}_t \boldsymbol{\beta}$ have zero mean, and their variances are bounded because of assumption (6), so the Kolmogorov Criterion for the strong law of large numbers (Feller 1968, p. 259) applies and $(n-m)^{-1}\sum_{t=m+1}^{n} 2\varepsilon_t \boldsymbol{z}_t \boldsymbol{\beta} \to 0$ almost surely.

*(c)* The random variables $2\varepsilon_t \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}$ are more tricky, because they are not independent for different values of $t$. We will use a variation of the weak law of large numbers that applies to a sequence of dependent random variables with a common expected value, bounded variances, and pairwise covariances that are upper-bounded by zero (Feller 1968, p. 262). Since $\varepsilon_t$ and $\boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}$ are independent, the expectation $\mathrm{E}[2\varepsilon_t \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}]$ equals zero, and

$$\mathrm{Var}\left[2\varepsilon_t \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}\right] = 4\sigma^2 \,\mathrm{Var}\left[\boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}\right].$$

Now, define the diagonal matrix $\boldsymbol{R}^{(\gamma)} \in \mathbb{R}^{q\times q}$ by setting $R^{(\gamma)}_{ii} = 1$ if $i \in \gamma$ and $R^{(\gamma)}_{ii} = 0$ otherwise. It is known that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1} = &\left(\frac{1}{t-1}\left(\boldsymbol{Z}_{t-1}\boldsymbol{R}^{(\gamma)}\right)^{\mathrm{T}}\left(\boldsymbol{Z}_{t-1}\boldsymbol{R}^{(\gamma)}\right)\right)^{-} \\ &\times \left(\frac{1}{t-1}\boldsymbol{Z}^{\mathrm{T}}_{t-1}\boldsymbol{Z}_{t-1}\right)\boldsymbol{\beta} \\ &+ \left(\frac{1}{t-1}\left(\boldsymbol{Z}_{t-1}\boldsymbol{R}^{(\gamma)}\right)^{\mathrm{T}}\left(\boldsymbol{Z}_{t-1}\boldsymbol{R}^{(\gamma)}\right)\right)^{-} \\ &\times \left(\frac{1}{t-1}\boldsymbol{Z}^{\mathrm{T}}_{t-1}\boldsymbol{\varepsilon}_{1:t-1}\right)\end{aligned}$$

where $(\,\cdot\,)^{-}$ denotes the Moore–Penrose pseudoinverse (Määttä, Schmidt, and Roos 2015, proof of Lemma 2). This and assumptions (6) and (7) imply that we may write

$$\boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1} = c^{(\gamma)}_t + \frac{1}{t-1}\sum_{i=1}^{t-1} C^{(\gamma)}_{t,i}\varepsilon_i, \tag{17}$$

where the numbers $c^{(\gamma)}_t$ and $C^{(\gamma)}_{t,i}$ are bounded constants. From (17) it is easy to see that $\mathrm{Var}[\boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}]$ and hence $\mathrm{Var}[2\varepsilon_t \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}]$ are bounded. Moreover, when $s < t$,

$$\begin{aligned}&\mathrm{Cov}[-2\varepsilon_s \boldsymbol{z}_s \widehat{\boldsymbol{\beta}}^{(\gamma)}_{s-1}, -2\varepsilon_t \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}] \\ &= \mathrm{E}[4\varepsilon_s \boldsymbol{z}_s \widehat{\boldsymbol{\beta}}^{(\gamma)}_{s-1}\varepsilon_t \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}] \\ &= \mathrm{E}[\varepsilon_t]\ \mathrm{E}[4\varepsilon_s \boldsymbol{z}_s \widehat{\boldsymbol{\beta}}^{(\gamma)}_{s-1} \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}] = 0.\end{aligned}$$

Hence the weak law of large number holds:

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n-m}\sum_{t=m+1}^{n} 2\varepsilon_t \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1} = 0.$$

*(d)* By Lemma 3, the terms $[\boldsymbol{z}_t(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1})]^2$ are uniformly bounded with respect to $t$.

Combining the contributions of components (a)–(c), we have

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n-m}\sum_{t=m+1}^{n}\left(\varepsilon_t^2 + 2\varepsilon_t \boldsymbol{z}_t \boldsymbol{\beta} - 2\varepsilon_t \boldsymbol{z}_t \widehat{\boldsymbol{\beta}}^{(\gamma)}_{t-1}\right) = \sigma^2.$$

Since the components (d) are non-negative and bounded, the claim follows. □

*Proof of Theorem 2.* Define $a_n = \mathrm{Hybrid}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma) - h(n, m, \lambda^2)$ and $b_n = \mathrm{PLS}(\boldsymbol{y}_{1:n}, \boldsymbol{Z}_n, \gamma)$. In other words,

$$a_n = \sum_{t=m+2}^{n}\left(\frac{t-m}{2}\right)\log\left(1 + \frac{e^2_{t,\gamma}}{\lambda^2(t-m-1)}\right) \quad \text{and}$$

$$b_n = \sum_{t=m+1}^{n} e^2_{t,\gamma}.$$

Note that the sequence $(b_n)$ is monotone and, by Theorem 1, almost surely divergent. Therefore, the classical Stolz–Cesàro theorem (Stolz 1885; Cesàro 1888) holds with probability one:

$$\begin{aligned}\liminf_{n\to\infty}\frac{a_{n+1}-a_n}{b_{n+1}-b_n} \leq \liminf_{n\to\infty}\frac{a_n}{b_n} \leq \limsup_{n\to\infty}\frac{a_n}{b_n} \\ \leq \limsup_{n\to\infty}\frac{a_{n+1}-a_n}{b_{n+1}-b_n}.\end{aligned}$$

Recall that $x/(1+x) \leq \log(1+x) \leq x$ for $x > -1$ (Abramowitz and Stegun 1972). We immediately have

$$\begin{aligned}&\limsup_{n\to\infty}\frac{a_{n+1}-a_n}{b_{n+1}-b_n} \\ &= \limsup_{n\to\infty}\left(\frac{n-m+1}{2e^2_{n+1,\gamma}}\right)\log\left(1 + \frac{e^2_{n+1,\gamma}}{\lambda^2(n-m)}\right) \\ &\leq \limsup_{n\to\infty}\left(\frac{n-m+1}{2e^2_{n+1,\gamma}}\right)\left(\frac{e^2_{n+1,\gamma}}{\lambda^2(n-m)}\right) \\ &= \limsup_{n\to\infty}\frac{n-m+1}{2\lambda^2(n-m)} = \frac{1}{2\lambda^2}\end{aligned}$$

almost surely (that is, unless $e^2_t = 0$ for infinitely many $n$, an event of probability zero[1]).

Going to the other direction,

$$\begin{aligned}&\operatorname*{p-lim\,inf}_{n\to\infty}\frac{a_{n+1}-a_n}{b_{n+1}-b_n} \\ &= \operatorname*{p-lim\,inf}_{n\to\infty}\left(\frac{n-m+1}{2e^2_{n+1,\gamma}}\right)\log\left(1 + \frac{e^2_{n+1,\gamma}}{\lambda^2(n-m)}\right) \\ &\geq \operatorname*{p-lim\,inf}_{n\to\infty}\left(\frac{n-m+1}{2e^2_{n+1,\gamma}}\right)\left(\frac{e^2_{n+1,\gamma}}{e^2_{n+1,\gamma} + \lambda^2(n-m)}\right) \\ &= \operatorname*{p-lim\,inf}_{n\to\infty}\frac{n-m+1}{2e^2_{n+1,\gamma} + 2\lambda^2(n-m)},\end{aligned}$$

[1] Note that the random variables $\varepsilon_t$ may be discrete, so it is possible that $\Pr[\varepsilon_t = 0] > 0$. However, if $\varepsilon_t \neq 0$ for infinitely many $t$ (this has probability 1), then this forces $e^2_t$ to be non-zero for some $t$, and henceforth it can never become exactly zero again.

and for every $0 < \delta < 1/(2\lambda^2)$,

$$\begin{aligned}
&\Pr\left[\left|\frac{n-m+1}{2e_{n+1,\gamma}^2 + 2\lambda^2(n-m)} - \frac{1}{2\lambda^2}\right| \geq \delta\right] \\
&= \Pr\left[\frac{\left|\lambda^2 - e_{n+1,\gamma}^2\right|}{2\lambda^2 e_{n+1,\gamma}^2 + 2(n-m)\lambda^4} \geq \delta\right] \\
&\leq \Pr\left[\frac{\lambda^2 + e_{n+1,\gamma}^2}{2\lambda^2 e_{n+1,\gamma}^2 + 2(n-m)\lambda^4} \geq \delta\right] \\
&= \Pr\left[e_{n+1,\gamma}^2 \geq \frac{2(n-m)\lambda^4\delta - \lambda^2}{1 - 2\lambda^2\delta}\right].
\end{aligned}$$

By Lemma 3, $\mathrm{E}[e_{n+1,\gamma}^2]$ is uniformly bounded with regard to $n$. Hence, Markov's inequality implies that the above tends to zero as $n \to \infty$, that is,

$$\operatorname*{p-lim}_{n\to\infty} \frac{n-m+1}{2e_{n+1,\gamma}^2 + 2\lambda^2(n-m)} = \frac{1}{2\lambda^2}.$$

Combining the results above, the Stolz–Cesàro theorem implies that

$$\operatorname*{p-lim}_{n\to\infty} \frac{a_n}{b_n} = \frac{1}{2\lambda^2}$$

and our claim follows. □

# References

Abramowitz, M., and Stegun, I. A. 1972. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, 10th edition.

Akaike, H. 1978. A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* 30(Part A):9–14.

Bartlett, P. L.; Koolen, W. M.; Malek, A.; Takimoto, E.; and Warmuth, M. K. 2015. Minimax fixed-design linear regression. In *Proceedings of the 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, 226–239.

Cesàro, E. 1888. Sur la convergence des sèries. *Nouvelles annales de mathématiques* 3(7):49–59.

Dawid, A. P. 1984. Statistical theory: The prequential approach. *J. Roy. Statist. Soc. Ser. A* 147(2):278–292.

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications*, volume 1. New York, NY: John Wiley & Sons, 3rd edition.

Määttä, J.; Schmidt, D. F.; and Roos, T. 2015. Subset selection in linear regression using sequentially normalized least squares: Asymptotic theory. *Scand. J. Stat.*

McQuarrie, A. D. R., and Tsai, C.-L. 1998. *Regression and Time Series Model Selection*. Singapore: World Scientific.

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

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

Schwarz, G. 1978. Estimating the dimension of a model. *Ann. Statist.* 6(2):461–464.

Stolz, O. 1885. *Vorlesungen über allgemeine Arithmetik: Nach den neueren Ansichten*. Leipzig: B. G. Teubner.

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