Supporting Information for "Subset Selection in Linear Regression using Sequentially Normalized Least Squares: Asymptotic Theory": Appendix S1: Proofs of Lemmas

Jussi Määttä* Daniel F. Schmidt† Teemu Roos*

This supplement provides proofs for the lemmas presented in our article "Subset Selection in Linear Regression using Sequentially Normalized Least Squares: Asymptotic Theory".

The numbering of equations and lemmas used in this appendix matches that of the main article.

The proofs will use the following assumptions from the article:

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

$$\sup_n z_n z_n' = \alpha < \infty, \qquad \text{and} \tag{8}$$

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

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

$$\log\hat{\tau}_n = \log\hat{\sigma}_n^2(\gamma) - \left(\frac{\log n}{n}\right)\left[\frac{|\gamma|\sigma_*^2 + \mathrm{tr}((R\Lambda R)^- \tilde{G})}{\hat{\sigma}_n^2(\gamma)}\right] + o\left(\frac{\log n}{n}\right) \quad a.s. \tag{A.1}$$

Proof. We start by showing a correspondence between our setting and that of Theorem 4.1.1 in Wei (1992). In his notation, Wei considers the regression model $y_i = f((X_i)) + \alpha_i$, where $X_i = (x_{i1}, x_{i2}, \dots) \in l^2$ and $f(X_i) = \sum_{j=1}^{\infty}\theta_j x_{ij}$ for some $(\theta_1, \theta_2, \dots) \in l^2$. He considers an incorrect model where one selects only p variables, denoting $x_i = (x_{i1}, \dots, x_{ip})'$.

Our setting corresponds to the case where only the first q elements of each X_i are allowed to be non-zero. Moreover, where Wei considers the vector x_i truncated

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

†Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Level 3, 207 Bouverie Street, VIC 3010, Australia. E-mail: dschmidt@unimelb.edu.au

from an l^2 sequence, we use the diagonal matrix R to set all elements not included in the model to zero. Wei's p equals our $|\gamma|$. Note also that we use row vectors where Wei uses column vectors. Wei's matrix Γ corresponds to our $R\Lambda R$; his requirement that Γ is nonsingular corresponds to our assumption (5) and our use of the Moore–Penrose pseudoinverse for $R\Lambda R$ is not a problem. Wei's vector $\gamma \in \mathbb{R}^{p\times 1}$ to our $R\Lambda\beta_*$. His and ours $\tilde{G}$ denote the same matrix, with our δ_i^2 matching Wei's $[f(X_i) - \beta' x_i]^2$.

Noting the above considerations, Theorem 4.1.1 in Wei (1992) gives the asymptotic (almost sure) form of the sum of squared errors for the predictive least squares method as

$$\sum_{i=m+1}^{n} e_i^2 = n\hat{\sigma}_n^2 + (\log n)[|\gamma|\sigma_*^2 + \mathrm{tr}((R\Lambda R)^{-}\tilde{G})](1 + o(1)). \tag{A.2}$$

We now require a similar result for the SNLS errors (instead of the predictive least squares errors), namely for the sum

$$\sum_{i=m+1}^{n} \hat{e}_i^2 = \sum_{i=m+1}^{n} (e_i^2 - 2d_i e_i^2 + d_i^2 e_i^2). \tag{A.3}$$

The first term inside the parentheses, involving the predictive least squares errors, e_i^2 is given by (A.2) above. The middle term appears in Theorem 2.1 in Wei (1992):

$$\sum_{i=m+1}^{n} e_i^2 = n\hat{\sigma}_n^2 - m\hat{\sigma}_m^2 + \sum_{i=m+1}^{n} d_i e_i^2, \tag{A.4}$$

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

By (A.2) and (A.4), we have

$$\sum_{t=m+1}^{n} d_t e_t^2 = (|\gamma|\sigma_*^2 + \mathrm{tr}((R\Lambda R)^{-}\tilde{G}))\log n + o(\log n) \quad \text{a.s.} \tag{A.5}$$

From (A.3) and (A.4), it follows that almost surely

$$\begin{aligned}\sum_{t=m+1}^{n} \hat{e}_t^2 &= n\hat{\sigma}_n^2 - m\hat{\sigma}_m^2 - \sum_{t=m+1}^{n} d_t e_t^2 + \sum_{t=m+1}^{n} d_t^2 e_t^2 \\ &= n\hat{\sigma}_n^2 - (|\gamma|\sigma_*^2 + \mathrm{tr}((R\Lambda R)^{-}\tilde{G}))\log n + \sum_{t=m+1}^{n} d_t^2 e_t^2 + o(\log n),\end{aligned} \tag{A.6}$$

where the second equality follows from (A.5).

We will now show that the term $\sum_{t=m+1}^{n} d_t^2 e_t^2$ is negligible. The proof uses a similar technique to the proof of Theorem 4 in Rissanen *et al.* (2010). Start by noting that by (5), the limit $\lim_{n\to\infty} \left\{nJ_n^{-1}\right\} = \Sigma^{-1}$ exists, with Σ a positive definite matrix. Denote the smallest eigenvalue of Σ by $\lambda_{\min} > 0$; this implies that the greatest eigenvalue of Σ^{-1} is $\lambda_{\min}^{-1}$. By (8) we now have

$$t d_t = x_t \left(t J_t^{-1} \right) x_t^{\mathrm{T}} \leq \alpha(\lambda_{\min}^{-1} + o(1)) = O(1),$$

and $d_t = O(1/t)$.

By (A.5), for any $s > m$, we must have $\sum_{t=s}^{n} d_t e_t^2 = O(\log(n/s))$ a.s., and it follows that

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

where the inequality holds by $0 \leq d_t \leq 1$ and $d_t = O(1/t)$. Letting $s = \log n$ yields the bound

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

Having shown that the left-hand side in the previous equation is smaller than the leading terms in (A.6), we obtain that

$$\frac{1}{n\hat{\sigma}_n^2} \sum_{t=m+1}^{n} \hat{e}_t^2 = 1 - \left(\frac{\log n}{n}\right) \left[\frac{|\gamma|\sigma_*^2 + \mathrm{tr}((R\Lambda R)^{-}\tilde{G})}{\hat{\sigma}_n^2}\right] + o\left(\frac{\log n}{n}\right) \text{ a.s.}$$

Taking the logarithm and applying Taylor expansion $\log(1+x) = x + O(x^2)$ to the right-hand side yields

$$\log \sum_{i=m+1}^{n} \hat{e}_i^2 = \log n\hat{\sigma}_n^2 - \left(\frac{\log n}{n}\right) \left[\frac{|\gamma|\sigma_*^2 + \mathrm{tr}((R\Lambda R)^{-}\tilde{G})}{\hat{\sigma}_n^2}\right] + o\left(\frac{\log n}{n}\right) \quad \text{a.s.}$$

Subtracting $\log(n-m) = \log n + O(1/n)$ 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 claim. □

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.

Proof. First of all, we define

$$
\begin{aligned}
\tilde{\beta}_n &= ((Z_nR)^{\mathrm{T}}(Z_nR))^{-}(Z_nR)^{\mathrm{T}}y_{1:n} \\
&= ((Z_nR)^{\mathrm{T}}(Z_nR))^{-}(Z_nR)^{\mathrm{T}}\left(Z_n\beta_* + \sigma_*\varepsilon^{(n)}\right) \\
&= \left(\frac{1}{n}(Z_nR)^{\mathrm{T}}(Z_nR)\right)^{-}\left(\frac{1}{n}Z_n^{\mathrm{T}}Z_n\right)\beta_* \\
&\quad + \left(\frac{1}{n}(Z_nR)^{\mathrm{T}}(Z_nR)\right)^{-}\left(\frac{1}{n}Z_n^{\mathrm{T}}\sigma_*\varepsilon^{(n)}\right).
\end{aligned}
\tag{A.7}
$$

Note that $\tilde{\beta}_n$ always contains q elements; if we omit the $q-|\gamma|$ elements that are not related to the variables in γ, we arrive at $\hat{\beta}_n$. Also note that $(RAR)^{-}R = (RAR)^{-}$ for all matrices A; we often use this to simplify our expressions.

The first term in (A.7) tends to $(R\Lambda R)^{-}\Lambda\beta_*$ as $n \to \infty$. For the second term, we consider separately the random variables

$$
S_n^{(k)} = \left(\frac{1}{n}Z_n^{\mathrm{T}}\sigma_*\varepsilon^{(n)}\right)_k = \frac{\sigma_*}{n}\sum_{i=1}^{n} z_{i,k}\varepsilon_i
$$

for each $k = 1, 2, \ldots, q$. Since $E[z_{i,k}\varepsilon_i] = 0$ and $\mathrm{var}[z_{i,k}\varepsilon_i] = (z_{i,k}\sigma_*)^2$ and the $z_{i,k}$ are bounded by assumption (8), it follows from the Kolmogorov criterion for the strong law of large numbers (Feller, 1968, p. 259) that

$$
S_n^{(k)} \to 0 \text{ a.s. for all } k = 1, 2, \ldots, q
$$

as $n \to \infty$. Hence, $\tilde{\beta}_n \to (R\Lambda R)^{-}\Lambda\beta_*$ a.s. as $n \to \infty$. Notice also that if the model γ is correct, the result simplifies to $\tilde{\beta}_n \to \beta_*$ a.s., and on the other hand, if γ is not correct, then $(R\Lambda R)^{-} \neq \Lambda^{-1}$ and therefore $\tilde{\beta}_n \not\to \beta_*$ a.s.

Moving on to $\hat{\sigma}_n^2$, we first decompose

$$
\begin{aligned}
\hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - z_i\tilde{\beta}_n\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - z_i\beta_* + z_i\beta_* - z_i\tilde{\beta}_n\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left((y_i - z_i\beta_*)^2 + (z_i(\beta_* - \tilde{\beta}_n))^2 + 2\,(y_i - z_i\beta_*)\,z_i(\beta_* - \tilde{\beta}_n)\right) \\
&= \frac{\sigma_*^2}{n}\sum_{i=1}^{n}\varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^{n}(z_i(\beta_* - \tilde{\beta}_n))^2 \\
&\quad + \frac{2\sigma_*}{n}\sum_{i=1}^{n}\varepsilon_i z_i\beta_* - \frac{2\sigma_*}{n}\sum_{i=1}^{n}\varepsilon_i z_i\tilde{\beta}_n.
\end{aligned}
\tag{A.8}
$$

By the strong law of large numbers, the first term tends to σ_*^2 almost surely.

The second term of (A.8) is nonnegative, and if γ is correct, it tends to zero almost surely (here we use the boundedness assumption (8)). On the other hand, if γ is not correct, then $\beta_* - \tilde{\beta}_n$ tends to a non-zero limit; this implies that the second term cannot tend to zero, because then assumption (5) would be violated. The constant ξ in the statement of the lemma is the a.s. limit of the second term.

Using similar techniques as above, the third term is easily seen to converge to zero almost surely.

To prove our claim, it now remains to show that the final term of (A.8) also vanishes almost surely. Expanding with (A.7), we have

$$\begin{aligned}\frac{2\sigma_*}{n}\sum_{i=1}^{n}\varepsilon_i z_i\tilde{\beta}_n &= \frac{2\sigma_*}{n}\sum_{i=1}^{n}\varepsilon_i z_i\left(\frac{1}{n}(Z_nR)^{\mathrm{T}}(Z_nR)\right)^{-}\left(\frac{1}{n}Z_n^{\mathrm{T}}Z_n\right)\beta_* \\ &\quad + \frac{2\sigma_*}{n}\sum_{i=1}^{n}\varepsilon_i z_i\left(\frac{1}{n}(Z_nR)^{\mathrm{T}}(Z_nR)\right)^{-}\left(\frac{1}{n}Z_n^{\mathrm{T}}\sigma_*\varepsilon^{(n)}\right).\end{aligned} \tag{A.9}$$

The first term in (A.9) tends to zero almost surely by similar arguments as above. To see the behavior of the second term more clearly, we write

$$\begin{aligned}&\frac{2\sigma_*}{n}\sum_{i=1}^{n}\varepsilon_i z_i\left(\frac{1}{n}(Z_nR)^{\mathrm{T}}(Z_nR)\right)^{-}\left(\frac{1}{n}Z_n^{\mathrm{T}}\sigma_*\varepsilon^{(n)}\right)\\ &= \frac{2\sigma_*^2}{n^2}\sum_{i=1}^{n}\varepsilon_i\left[z_i\left(\frac{1}{n}(Z_nR)^{\mathrm{T}}(Z_nR)\right)^{-}Z_n^{\mathrm{T}}\right]\varepsilon^{(n)}\\ &= \frac{2\sigma_*^2}{n^2}\sum_{i=1}^{n}\varepsilon_i\sum_{j=1}^{n}\underbrace{\left[z_i\left(\frac{1}{n}(Z_nR)^{\mathrm{T}}(Z_nR)\right)^{-}Z_n^{\mathrm{T}}\right]_j}_{=:\eta_{i,j,n}}\varepsilon_j\\ &= \frac{2\sigma_*^2}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\eta_{i,j,n}\varepsilon_i\varepsilon_j + \frac{2\sigma_*^2}{n^2}\sum_{i=1}^{n}\eta_{i,i,n}\varepsilon_i^2.\end{aligned} \tag{A.10}$$

The terms $\eta_{i,j,n}$ can be written as

$$\eta_{i,j,n} = \sum_{k=1}^{q}\sum_{\ell=1}^{q}[Z_n]_{ik}[Z_n]_{j\ell}\left[\left(\frac{1}{n}(Z_nR)^{\mathrm{T}}(Z_nR)\right)^{-}\right]_{k\ell}$$

from which it is easy to see that by assumptions (5) and (8) they are bounded. Since ε_i and ε_j are independent when $i \neq j$, it follows that the both terms of (A.10) tend to zero almost surely as $n \to \infty$. □

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

Proof. If $\gamma = \{1, 2, \ldots, q\}$, then $(R\Lambda R)^{-}\Lambda\beta_* = I_q\beta_* = \beta_*$, $\delta = 0$, and $\tilde{G} = 0$, implying that the trace is also zero.

The same argument can be extended to the case $\gamma \subsetneq \{1, 2, \dots, q\}$ as follows. Let $\ell = q - |\gamma|$ and assume without loss of generality that $\gamma = \{\ell+1, \ell+2, \dots, q\}$. Write

$$\Lambda = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \quad A \in \mathbb{R}^{\ell \times \ell}.$$

Then

$$\begin{aligned}(R\Lambda R)^{-}\Lambda\beta_* &= \begin{bmatrix} 0 & 0 \\ 0 & D \end{bmatrix}^{-} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \beta_* = \begin{bmatrix} 0 & 0 \\ 0 & D^{-} \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \beta_* \\ &= \begin{bmatrix} 0 & 0 \\ D^{-}C & I_{q-\ell} \end{bmatrix} \beta_* = \beta_*,\end{aligned}$$

since $\beta_*(i) = 0$ for $1 \le i \le \ell$. □

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*

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

for all $n \ge N$.

Proof. In the following, we will denote by $\lambda_*(\cdot)$ and $\lambda^*(\cdot)$ the smallest and largest eigenvalues of a given matrix.

Let $p = |\gamma_2|$. By Lemma 3.3 and equation (3.9) of Wei (1992),

$$\log |J_n(\gamma_2)| \ge \log p^{-1}\lambda_*(J_n(\gamma_2)) + \log |J_n(\gamma_1)|.$$

Hence,

$$\begin{aligned}\log |J_n(\gamma_2)| - \log |J_n(\gamma_1)| &\ge \log p^{-1}\lambda_*(J_n(\gamma_2)) \\ &= \log n + \log \frac{\lambda_*(n^{-1}J_n(\gamma_2))}{p},\end{aligned}$$

and the claim follows by assumption (5). □

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.

Proof. By Lemmas 1 and 3, we have almost surely

$$\begin{aligned}\log \hat{\tau}_n(\gamma_2) - \log \hat{\tau}_n(\gamma_1) &= \log \hat{\sigma}_n^2(\gamma_2) - \log \hat{\sigma}_n^2(\gamma_1) \\ &\quad - \left(\frac{\log n}{n}\right)\left(\frac{|\gamma_2|\sigma_*^2}{\hat{\sigma}_n^2(\gamma_2)} - \frac{|\gamma_1|\sigma_*^2}{\hat{\sigma}_n^2(\gamma_1)}\right) \\ &\quad + o\left(\frac{\log n}{n}\right).\end{aligned}$$

Noting that $|\gamma_1| = |\gamma_2| - 1$, this can be written as

$$
\begin{aligned}
&\log \hat{\tau}_n(\gamma_2) - \log \hat{\tau}_n(\gamma_1) \\
=\ &\log \hat{\sigma}_n^2(\gamma_2) - \log \hat{\sigma}_n^2(\gamma_1) \\
&\quad - \left(\frac{\log n}{n}\right)\left(\frac{|\gamma_2|\sigma_*^2}{\hat{\sigma}_n^2(\gamma_1)\hat{\sigma}_n^2(\gamma_2)}\right)\left(\hat{\sigma}_n^2(\gamma_1) - \hat{\sigma}_n^2(\gamma_2)\right) \\
&\quad - \left(\frac{\log n}{n}\right)\left(\frac{\sigma_*^2}{\hat{\sigma}_n^2(\gamma_1)}\right) + o\left(\frac{\log n}{n}\right) \quad \text{a.s.} \\
=\ &\log \hat{\sigma}_n^2(\gamma_2) - \log \hat{\sigma}_n^2(\gamma_1) \\
&\quad - \left(\frac{\log n}{n}\right)\left(\frac{|\gamma_2|}{\sigma_*^2} + o(1)\right)\left(\hat{\sigma}_n^2(\gamma_1) - \hat{\sigma}_n^2(\gamma_2)\right) \\
&\quad - \left(\frac{\log n}{n}\right)(1 + o(1)) + o\left(\frac{\log n}{n}\right) \quad \text{a.s.}
\end{aligned}
$$

Since $\left(\hat{\sigma}_n^2(\gamma_2) - \hat{\sigma}_n^2(\gamma_1)\right) \to 0$ a.s., we have

$$
\log \hat{\sigma}_n^2(\gamma_2) - \log \hat{\sigma}_n^2(\gamma_1) = \left(\hat{\sigma}_n^2(\gamma_2) - \hat{\sigma}_n^2(\gamma_1)\right)(1 + o(1)) \quad \text{a.s.}
$$

so we may further simplify to

$$
\begin{aligned}
&\log \hat{\tau}_n(\gamma_2) - \log \hat{\tau}_n(\gamma_1) \\
&\qquad = \left(\hat{\sigma}_n^2(\gamma_2) - \hat{\sigma}_n^2(\gamma_1)\right)(1 + o(1)) - \frac{\log n}{n} + o\left(\frac{\log n}{n}\right) \quad \text{a.s.}
\end{aligned}
$$

□

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

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

Proof. Denote $\Lambda_n = (1/n) Z_n^{\mathrm{T}} Z_n$. Using the notation $\tilde{\beta}_n(\gamma_i)$ introduced in the proof of Lemma 2, we have by definition

$$
\begin{aligned}
\hat{\sigma}_n^2(\gamma_2) - \hat{\sigma}_n^2(\gamma_1) &= \frac{1}{n}\sum_{i=1}^{n}\Big[(y_i - z_i\tilde{\beta}_n(\gamma_2))^2 - (y_i - z_i\tilde{\beta}_n(\gamma_1))^2\Big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big[(z_i\tilde{\beta}_n(\gamma_2))^2 - (z_i\tilde{\beta}_n(\gamma_1))^2 \\
&\qquad + 2y_i z_i(\tilde{\beta}_n(\gamma_1) - \tilde{\beta}_n(\gamma_2))\Big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\tilde{\beta}_n(\gamma_2)^{\mathrm{T}} z_i^{\mathrm{T}} z_i \tilde{\beta}_n(\gamma_2) - \frac{1}{n}\sum_{i=1}^{n}\tilde{\beta}_n(\gamma_1)^{\mathrm{T}} z_i^{\mathrm{T}} z_i \tilde{\beta}_n(\gamma_1) \\
&\qquad + \frac{2}{n}\sum_{i=1}^{n}(z_i\beta_* + \sigma_*\varepsilon_i) z_i(\tilde{\beta}_n(\gamma_1) - \tilde{\beta}_n(\gamma_2)) \\
&= \tilde{\beta}_n(\gamma_2)^{\mathrm{T}}\Lambda_n\tilde{\beta}_n(\gamma_2) - \tilde{\beta}_n(\gamma_1)^{\mathrm{T}}\Lambda_n\tilde{\beta}_n(\gamma_1) \\
&\qquad + 2\tilde{\beta}_n(\gamma_1)^{\mathrm{T}}\Lambda_n\beta_* - 2\tilde{\beta}_n(\gamma_2)^{\mathrm{T}}\Lambda_n\beta_* \\
&\qquad + \frac{2\sigma_*}{n}(\varepsilon^{(n)})^{\mathrm{T}} Z_n(\tilde{\beta}_n(\gamma_1) - \tilde{\beta}_n(\gamma_2))
\end{aligned}
$$

which we simplify as

$$
\begin{aligned}
\hat{\sigma}_n^2(\gamma_2) - \hat{\sigma}_n^2(\gamma_1) &= f_n(\gamma_2) - f_n(\gamma_1), \\
f_n(\gamma_i) &= \tilde{\beta}_n(\gamma_i)^{\mathrm{T}}\Lambda_n\tilde{\beta}_n(\gamma_i) - 2\tilde{\beta}_n(\gamma_i)^{\mathrm{T}}\Lambda_n\beta_*\frac{2\sigma_*}{n}(\varepsilon^{(n)})^{\mathrm{T}} Z_n\tilde{\beta}_n(\gamma_i).
\end{aligned}
$$

By the expansion (A.7), we have

$$
\tilde{\beta}_n(\gamma_i) = Q_n(\gamma_i)\Lambda_n\beta_* + \frac{1}{n}\sigma_* Q_n(\gamma_i) Z_n^{\mathrm{T}}\varepsilon^{(n)},
$$

and since the matrices $Q_n(\gamma_i)$ and Λ_n are symmetric and it holds that

$$
Q_n(\gamma_i)\Lambda_n Q_n(\gamma_i) = Q_n(\gamma_i) R(\gamma_i)\Lambda_n R(\gamma_i) Q_n(\gamma_i) = Q_n(\gamma_i),
$$

we get

$$
\begin{aligned}
f_n(\gamma_i) &= \beta_*^{\mathrm{T}}\Lambda_n Q_n(\gamma_i)\Lambda_n Q_n(\gamma_i)\Lambda_n\beta_* + \frac{\sigma_*}{n}\beta_*^{\mathrm{T}}\Lambda_n Q_n(\gamma_i)\Lambda_n Q_n(\gamma_i) Z_n^{\mathrm{T}}\varepsilon^{(n)} \\
&\qquad + \frac{\sigma_*}{n}(\varepsilon^{(n)})^{\mathrm{T}} Z_n Q_n(\gamma_i)\Lambda_n Q_n(\gamma_i)\Lambda_n\beta_* \\
&\qquad + \frac{\sigma_*^2}{n^2}(\varepsilon^{(n)})^{\mathrm{T}} Z_n Q_n(\gamma_i)\Lambda_n Q_n(\gamma_i) Z_n^{\mathrm{T}}\varepsilon^{(n)} \\
&\qquad - 2\beta_*^{\mathrm{T}}\Lambda_n Q_n(\gamma_i)\Lambda_n\beta_* - \frac{2\sigma_*}{n}(\varepsilon^{(n)})^{\mathrm{T}} Z_n Q_n(\gamma_i)\Lambda_n\beta_* \\
&\qquad - \frac{2\sigma_*}{n}(\varepsilon^{(n)})^{\mathrm{T}} Z_n Q_n(\gamma_i)\Lambda_n\beta_* - \frac{2\sigma_*^2}{n^2}(\varepsilon^{(n)})^{\mathrm{T}} Z_n Q_n(\gamma_i) Z_n^{\mathrm{T}}\varepsilon^{(n)} \\
&= -\beta_*^{\mathrm{T}}\Lambda_n Q_n(\gamma_i)\Lambda_n\beta_* - \frac{2\sigma_*}{n}\beta_*^{\mathrm{T}}\Lambda_n Q_n(\gamma_i) Z_n^{\mathrm{T}}\varepsilon^{(n)} \\
&\qquad - \frac{\sigma_*^2}{n^2}(\varepsilon^{(n)})^{\mathrm{T}} Z_n Q_n(\gamma_i) Z_n^{\mathrm{T}}\varepsilon^{(n)}.
\end{aligned}
$$

Note then that since both γ_1 and γ_2 are correct, we have $\beta_*^{\mathrm{T}} R(\gamma_i) = \beta_*^{\mathrm{T}}$. Hence,

$$\beta_*^{\mathrm{T}} \Lambda_n Q_n(\gamma_i) = \left[\beta_*^{\mathrm{T}} R(\gamma_i)\right] \left[\frac{1}{n} Z_n^{\mathrm{T}} Z_n\right] \left[R(\gamma_i) \left(\frac{1}{n} R(\gamma_i) Z_n^{\mathrm{T}} Z_n R(\gamma_i)\right)^{-}\right] = \beta_*^{\mathrm{T}}$$

and we may conclude that

$$f_n(\gamma_2) - f_n(\gamma_1) = \frac{\sigma_*^2}{n^2} (\varepsilon^{(n)})^{\mathrm{T}} Z_n \left[Q_n(\gamma_1) - Q_n(\gamma_2)\right] Z_n^{\mathrm{T}} \varepsilon^{(n)}$$

which was our claim. □

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*

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

Proof. We use the notation from Lemma 6. Note first that the matrix M_n is symmetric. We have assumed that $E[\varepsilon^{(n)}] = \overline{0}$ and $\mathrm{var}[\varepsilon^{(n)}] = I_n$. By Lemma 6,

$$E\left[\frac{n}{2}\left(\hat{\sigma}_n^2(\gamma_2) - \hat{\sigma}_n^2(\gamma_1)\right)\right] = \frac{\sigma_*^2}{2} E\left[(\varepsilon^{(n)})^{\mathrm{T}} M_n \varepsilon^{(n)}\right]$$

and Theorem 1.5 of Seber & Lee (2003) gives

$$E\left[(\varepsilon^{(n)})^{\mathrm{T}} M_n \varepsilon^{(n)}\right] = \mathrm{tr}(M_n).$$

This can be expanded as

$$\begin{aligned}
\mathrm{tr}(M_n) &= \mathrm{tr}\left(\frac{1}{n} Z_n \left(Q_n(\gamma_1) - Q_n(\gamma_2)\right) Z_n^{\mathrm{T}}\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[Z_n \left(Q_n(\gamma_1) - Q_n(\gamma_2)\right) Z_n^{\mathrm{T}}\right]_{ii} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{n} [Z_n]_{ij} \left[\left(Q_n(\gamma_1) - Q_n(\gamma_2)\right) Z_n^{\mathrm{T}}\right]_{ji} \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{q} \sum_{k=1}^{q} [Z_n]_{ij} \left[Q_n(\gamma_1) - Q_n(\gamma_2)\right]_{jk} \left[Z_n^{\mathrm{T}}\right]_{ki} \right\}
\end{aligned}$$

which is the mean of n terms. These terms are uniformly bounded, because we have assumed (8), q is a constant, and the $q \times q$ matrices $Q_n(\gamma_i)$ converge by assumption (5). Hence, $\mathrm{tr}(M_n) = O(1)$.

Introduce then the assumption (12). Denote $\mu_4 = E[\varepsilon_i^4] < \infty$, and let $\overline{a}_n$ be a column vector of the diagonal elements of M_n. Then we have (Atiqullah, 1962; Seber & Lee, 2003, Theorem 1.6)

$$\mathrm{var}[(\varepsilon^{(n)})^{\mathrm{T}} M_n \varepsilon^{(n)}] = (\mu_4 - 3)\overline{a}_n^{\mathrm{T}}\overline{a}_n + 2\,\mathrm{tr}(M_n^2).$$

Denote by $\|\cdot\|_F$ the Frobenius norm. We have

$$\begin{aligned}
\mathrm{tr}(M_n^2) = \mathrm{tr}(M_n^{\mathrm{T}} M_n) = \|M_n\|_F^2 &= \frac{1}{n^2}\|Z_n(Q_n(\gamma_1) - Q_n(\gamma_2))Z_n^{\mathrm{T}}\|_F^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\left[Z_n(Q_n(\gamma_1) - Q_n(\gamma_2))Z_n^{\mathrm{T}}\right]_{ij}\right\}^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\sum_{k=1}^{q}\sum_{\ell=1}^{q}[Z_n]_{ik}\,[(Q_n(\gamma_1) - Q_n(\gamma_2))]_{k\ell}\,[Z_n^{\mathrm{T}}]_{\ell j}\right\}^2
\end{aligned}$$

which is the mean of n^2 uniformly bounded terms; therefore $\mathrm{tr}(M_n^2) = O(1)$.

Finally, we have

$$\overline{a}_n^{\mathrm{T}}\overline{a}_n = \frac{1}{n^2}\sum_{i=1}^{n}\left\{\sum_{k=1}^{q}\sum_{\ell=1}^{q}[Z_n]_{ik}\,[Q_n(\gamma_1) - Q_n(\gamma_2)]_{k\ell}\,[Z_n^{\mathrm{T}}]_{\ell i}\right\}^2 = o(1).$$

and the claim follows. □

References

Atiqullah, M. (1962). The estimation of residual variance in quadratically balanced least-squares problems and the robustness of the F-test. *Biometrika* **49**(1/2), 83–91.

Feller, W. (1968). *An introduction to probability theory and its applications*, vol. 1. John Wiley & Sons, Inc., New York, NY, 3rd edn.

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

Seber, G. A. F. & Lee, A. J. (2003). *Linear regression analysis*. Wiley, 2nd edn.

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