ROBUST MODEL SELECTION IN GENERALIZED LINEAR MODELS

Samuel Müller and A. H. Welsh

University of Sydney and Australian National University

Supplementary Material

This note contains the proof of Theorem 3.1 and some additional theoretical results on the monotonicity of trace$(\Sigma_\alpha \Gamma_\alpha)$.

S1. Proof of Theorem 3.1.

The proof of this result is similar to that given in Müller and Welsh (2005). The main term we need to deal with is the bootstrap term

$$M_n^{(2)}(\alpha) \;=\; \frac{1}{n}\,\mathrm{E}_* \sum_{i=1}^{n} w_{\alpha i}\rho\Big\{ (y_i - h[x_{\alpha i}^T\{\widehat{\beta}^*_{\alpha,m} - \mathrm{E}_*(\widehat{\beta}^*_{\alpha,m} - \widehat{\beta}_\alpha)\}])/\widehat{\sigma}_i \Big\},$$

where $\widehat{\beta}_\alpha$ and $\widehat{\sigma}_i = \widehat{\sigma} v(x_{\alpha_f i}^T \widehat{\beta}_{\alpha_f})$ are constant with respect to the bootstrap. We make a Taylor expansion of ρ as a function of $\widehat{\beta}^*_{\alpha,m} - \mathrm{E}_*(\widehat{\beta}^*_{\alpha,m} - \widehat{\beta}_\alpha)$ about $\widehat{\beta}_\alpha$, to obtain

$$\begin{aligned} M_n^{(2)}(\alpha) \;&=\; \frac{1}{n}\sum_{i=1}^{n} w_{\alpha i}\rho[\{y_i - h(x_{\alpha i}^T\widehat{\beta}_\alpha)\}/\widehat{\sigma}_i] \\ &\quad + \mathrm{E}_*\,\frac{1}{2n}\sum_{i=1}^{n} \widehat{\sigma}_i^{-2} w_{\alpha i} x_{\alpha i}^T (\widehat{\beta}^*_{\alpha,m} - \mathrm{E}_*\,\widehat{\beta}^*_{\alpha,m})(\widehat{\beta}^*_{\alpha,m} - \mathrm{E}_*\,\widehat{\beta}^*_{\alpha,m})^T x_{\alpha i} \\ &\quad \times \big(h'(x_{\alpha i}^T\bar{\beta}_\alpha)^2 \psi'[\{y_i - h(x_{\alpha i}^T\bar{\beta}_\alpha)\}/\widehat{\sigma}_i] - h''(x_{\alpha i}^T\bar{\beta}_\alpha)\psi[\{y_i - h(x_{\alpha i}^T\bar{\beta}_\alpha)\}/\widehat{\sigma}_i]\big) \\ &=\; T_1 + T_2, \end{aligned}$$

where $|\bar{\beta}_\alpha - \widehat{\beta}_\alpha| \le |\widehat{\beta}^*_{\alpha,m} - \widehat{\beta}_\alpha|$. This equation is analogous to (9) in Müller and Welsh (2005) except that we have eliminated the linear term by using the bias-adjusted bootstrap estimator. We consider T_1 and T_2 in turn.

Order of T_2: Let

$$\begin{aligned} \bar{H}_{\alpha i} \;&=\; h'(x_{\alpha i}^T\bar{\beta}_\alpha)^2\psi'[\{y_i - h(x_{\alpha i}^T\bar{\beta}_\alpha)\}/\widehat{\sigma}_i] - h''(x_{\alpha i}^T\bar{\beta}_\alpha)\psi[\{y_i - h(x_{\alpha i}^T\bar{\beta}_\alpha)\}/\widehat{\sigma}_i] \\ &=\; h'(x_{\alpha i}^T\bar{\beta}_\alpha)^2\psi'[\{\epsilon_i + h(x_{\alpha i}^T\beta_\alpha) - h(x_{\alpha i}^T\bar{\beta}_\alpha)\}/\widehat{\sigma}_i] - h''(x_{\alpha i}^T\bar{\beta}_\alpha) \\ &\quad \times \psi[\{\epsilon_i + h(x_{\alpha i}^T\beta_\alpha) - h(x_{\alpha i}^T\bar{\beta}_\alpha)\}/\widehat{\sigma}_i], \end{aligned}$$

and write

$$\begin{aligned}
T_2 &= \mathrm{E}_* \frac{1}{2n}\sum_{i=1}^{n} \widehat{\sigma}_i^{-2} w_{\alpha i} x_{\alpha i}^T (\widehat{\beta}^*_{\alpha,m} - \mathrm{E}_* \widehat{\beta}^*_{\alpha,m})(\widehat{\beta}^*_{\alpha,m} - \mathrm{E}_* \widehat{\beta}^*_{\alpha,m})^T x_{\alpha i} \bar{H}_{\alpha i} \\
&= \frac{1}{2n}\sum_{i=1}^{n} \sigma_i^{-2} w_{\alpha i} x_{\alpha i}^T \operatorname{Var}_*(\widehat{\beta}^*_{\alpha,m}) x_{\alpha i} (h'^2_{\alpha i} \mathrm{E}\,\psi'_i - h''_{\alpha i} \mathrm{E}\,\psi_i) \\
&\quad + \frac{1}{2n}\sum_{i=1}^{n} \sigma_i^{-2} w_{\alpha i} x_{\alpha i}^T \operatorname{Var}_*(\widehat{\beta}^*_{\alpha,m}) x_{\alpha i} (h'^2_{\alpha i}\psi'_i - h''_{\alpha i}\psi_i - h'^2_{\alpha i} \mathrm{E}\,\psi'_i + h''_{\alpha i} \mathrm{E}\,\psi_i) \\
&\quad + \mathrm{E}_* \frac{1}{2n}\sum_{i=1}^{n} w_{\alpha i} x_{\alpha i}^T (\widehat{\beta}^*_{\alpha,m} - \mathrm{E}_* \widehat{\beta}^*_{\alpha,m})(\widehat{\beta}^*_{\alpha,m} - \mathrm{E}_* \widehat{\beta}^*_{\alpha,m})^T \\
&\qquad \times x_{\alpha i} (\widehat{\sigma}_i^{-2} \bar{H}_{\alpha i} - \sigma_i^{-2} h'^2_{\alpha i}\psi'_i + \sigma_i^{-2} h''_{\alpha i}\psi_i).
\end{aligned}$$

Then

$$\begin{aligned}
&\frac{1}{2n}\sum_{i=1}^{n} \sigma_i^{-2} w_{\alpha i} x_{\alpha i}^T \operatorname{Var}_*(\widehat{\beta}^*_{\alpha,m}) x_{\alpha i} (h'^2_{\alpha i} \mathrm{E}\,\psi'_i - h''_{\alpha i} \mathrm{E}\,\psi_i) \\
&\quad = \frac{1}{2n} \operatorname{trace}\left\{ \operatorname{Var}_*(\widehat{\beta}^*_{\alpha,m}) \sum_{i=1}^{n} \sigma_i^{-2} w_{\alpha i} x_{\alpha i} x_{\alpha i}^T (h'^2_{\alpha i} \mathrm{E}\,\psi'_i - h''_{\alpha i} \mathrm{E}\,\psi_i) \right\} \\
&\quad = \frac{\kappa^c}{2m} \operatorname{trace}(\Sigma_\alpha \Gamma_\alpha) + o_p(m^{-1})
\end{aligned}$$

by condition (iii) and the first part of condition (i). Similarly

$$\frac{1}{2n}\sum_{i=1}^{n} \sigma_i^{-2} w_{\alpha i} x_{\alpha i}^T \operatorname{Var}_*(\widehat{\beta}^*_{\alpha,m}) x_{\alpha i} (h'^2_{\alpha i}\psi'_i - h''_{\alpha i}\psi_i - h'^2_{\alpha i} \mathrm{E}\,\psi'_i + h''_{\alpha i} \mathrm{E}\,\psi_i) = o_p(m^{-1})$$

by condition (iii) and the second part of condition (i). Finally,

$$|\,\mathrm{E}_* \frac{1}{2n}\sum_{i=1}^{n} w_{\alpha i} x_{\alpha i}^T (\widehat{\beta}^*_{\alpha,m} - \mathrm{E}_* \widehat{\beta}^*_{\alpha,m})(\widehat{\beta}^*_{\alpha,m} - \mathrm{E}_* \widehat{\beta}^*_{\alpha,m})^T x_{\alpha i} (\widehat{\sigma}_i^{-2} \bar{H}_{\alpha i} - \sigma_i^{-2} h'^2_{\alpha i}\psi'_i + \sigma_i^{-2} h''_{\alpha i}\psi_i)|$$

is $o_p(m^{-1})$ provided

$$\max_{1\le i\le n} \sup_{\epsilon} \sup_{|t-\beta_\alpha|\le n^{-1/2}C} |\widehat{\sigma}_i^{-2} h'(x_{\alpha i}^T t)^2 \psi'[\{\epsilon + h(x_{\alpha i}^T \beta_\alpha) - h(x_{\alpha i}^T t)\}/\widehat{\sigma}_i] - \sigma_i^{-2} h'^2(x_{\alpha i}^T \beta_\alpha)\psi'(\epsilon/\sigma_i)|,$$

$$\max_{1\le i\le n} \sup_{\epsilon} \sup_{|t-\beta_\alpha|\le n^{-1/2}C} |\widehat{\sigma}_i^{-2} h''(x_{\alpha i}^T t)\psi[\{\epsilon + h(x_{\alpha i}^T \beta_\alpha) - h(x_{\alpha i}^T t)\}/\widehat{\sigma}_i] - \sigma_i^{-2} h''(x_{\alpha i}^T \beta_\alpha)\psi(\epsilon/\sigma_i)|$$

are $o_p(1)$. Conditions (v)-(viii) ensure that these requirements hold.

Order of T_1: Let $|\widetilde{\beta}_\alpha - \beta_\alpha| \leq |\widehat{\beta}_\alpha - \beta_\alpha|$, $|\widetilde{\beta}_{\alpha_f} - \beta_{\alpha_f}| \leq |\widehat{\beta}_{\alpha_f} - \beta_{\alpha_f}|$ and $|\widetilde{\sigma} - \sigma| \leq |\widehat{\sigma} - \sigma|$. Recall that $\sigma_i = \sigma v(h_{\alpha_f i})$ and write

$$D(y_i, h_{\alpha i}, \sigma_i) = \begin{pmatrix} -x_{\alpha i} h'_{\alpha i} \sigma_i^{-1} \psi\{(y_i - h_{\alpha i})/\sigma_i\} \\ -\sigma_i^{-2} v(h^{-1}(h_{\alpha_f i}))(y_i - h_{\alpha i})\psi\{(y_i - h_{\alpha i})/\sigma_i\} \\ -x_{\alpha_f i}\sigma_i^{-2}\sigma v'(h^{-1}(h_{\alpha_f i}))(y_i - h_{\alpha i})\psi\{(y_i - h_{\alpha i})/\sigma_i\} \end{pmatrix}$$

Then

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n w_{\alpha i}\rho\{(y_i - \widehat{h}_{\alpha i})/\widehat{\sigma}_i\} \\
&= \frac{1}{n}\sum_{i=1}^n w_{\alpha i}\rho(\epsilon_i/\sigma_i) + \frac{1}{n}\sum_{i=1}^n w_{\alpha i}(\widehat{\beta}_\alpha - \beta_\alpha, \widehat{\sigma} - \sigma, \widehat{\beta}_{\alpha_f} - \beta_{\alpha_f})^T D(y_i, h_{\alpha i}, \sigma_i) \\
&\quad + \frac{1}{n}\sum_{i=1}^n w_{\alpha i}(\widehat{\beta}_\alpha - \beta_\alpha, \widehat{\sigma} - \sigma, \widehat{\beta}_{\alpha_f} - \beta_{\alpha_f})^T\{D(y_i, \widetilde{h}_{\alpha i}, \widetilde{\sigma}_i) - D(y_i, h_{\alpha i}, \sigma_i)\} \\
&= \frac{1}{n}\sum_{i=1}^n w_{\alpha i}\rho(\epsilon_i/\sigma_i) + O_p(n^{-1/2}),
\end{aligned}$$

provided

$$\max_{1\leq i\leq n}\sup_\epsilon \sup_{|t-\beta_\alpha|\leq n^{-1/2}C} |\widehat{\sigma}_i^{-1}h'(x_{\alpha i}^T t)\psi[\{\epsilon + h(x_{\alpha i}^T\beta_\alpha) - h(x_{\alpha i}^T t)\}/\widehat{\sigma}_i] - \sigma_i^{-1}h'_{\alpha i}\psi(\epsilon/\sigma_i)| = o_p(1)$$

$$\begin{aligned}
&\max_{1\leq i\leq n}\sup_\epsilon \sup_{|t-\beta_\alpha|\leq n^{-1/2}C} |\widehat{\sigma}^{-1}v(x_{\alpha_f i}^T\widehat{\beta}_{\alpha_f})^{-2}v'(x_{\alpha_f i}^T\widehat{\beta}_{\alpha_f})\{\epsilon + h(x_{\alpha i}^T\beta_\alpha) - h(x_{\alpha i}^T t)\} \\
&\quad \times\psi[\{\epsilon + h(x_{\alpha i}^T\beta_\alpha) - h(x_{\alpha i}^T t)\}/\widehat{\sigma}_i] - \sigma^{-1}v(x_{\alpha_f i}^T\beta_{\alpha_f i})^{-2}v'(x_{\alpha_f i}^T\beta_{\alpha_f})\epsilon\psi(\epsilon/\sigma_i)| = o_p(1)
\end{aligned}$$

$$\begin{aligned}
\max_{1\leq i\leq n}\sup_\epsilon \sup_{|t-\beta_\alpha|\leq n^{-1/2}C} &|\widehat{\sigma}^{-2}v(x_{\alpha_f i}^T\widehat{\beta}_{\alpha_f})^{-1}\{\epsilon + h(x_{\alpha i}^T\beta_\alpha) - h(x_{\alpha i}^T t\}\psi[\{\epsilon + h(x_{\alpha i}^T\beta_\alpha) \\
&\qquad -h(x_{\alpha i}^T t)\}/\widehat{\sigma}_i] - \sigma^{-2}v(x_{\alpha_f i}^T\beta_{\alpha_f})^{-1}\epsilon\psi(\epsilon/\sigma_i)| = o_p(1).
\end{aligned}$$

As for T_2, these results follow from conditions (v)-(vii).

Putting both terms together, it follows that

$$M_n^{(2)}(\alpha) \;=\; \frac{1}{n}\sum_{i=1}^n w_{\alpha i}\rho\{(y_i - h(x_{\alpha i}^T\widehat{\beta}_\alpha))/\widehat{\sigma}_i\} + \frac{\kappa^c}{2m}\,\mathrm{trace}(\Sigma_\alpha\Gamma_\alpha) + o_p(m^{-1}),$$

and the proof is completed as in Müller and Welsh (2005). □

S2. The Cantoni-Ronchetti estimator

Consider the Mallows quasi–likelihood estimator defined in Cantoni and Ronchetti ((2001), Section 2.2) as the solution of the estimating equations

$$X_\alpha^T A^{1/2} \Psi = 0,$$

where Ψ is the n-vector with elements $\Psi_i = \psi_c(r_{\alpha i}) - \mathrm{E}\,\psi_c(r_{\alpha i})$, with

$$r_i = \{y_i - h(x_{\alpha i}^T \beta_\alpha)\}/v(x_{\alpha i}^T \beta_\alpha),$$

the Pearson residuals, and $\psi_c = \min\{c, \max(-c, r)\}$, the Huber function, and A is the $n \times n$ diagonal matrix with diagonal elements

$$a_{ii} = w(x_{\alpha i})^2 \frac{1}{\sigma^2 v(x_{\alpha i}^T \beta_\alpha)^2} h'(x_{\alpha i}^T \beta_\alpha)^2.$$

When $w(x_i) = 1$, the estimator is called the Huber quasi–likelihood estimator. In general we do not require that $\psi_c = \rho' = \psi$ or that $w_{\alpha i} = w(x_{\alpha i})$. Cantoni and Ronchetti ((2001), Appendix B) show that the estimator has an asymptotic normal distribution with asymptotic variance $\Sigma_\alpha = M_\alpha^{-1} Q_\alpha M_\alpha^{-1}$, where

$$Q_\alpha = \frac{1}{n} X_\alpha^T A^{1/2} \Sigma A^{1/2} X_\alpha \quad \text{and} \quad M_\alpha = \frac{1}{n} X_\alpha^T B X_\alpha,$$

with $\Sigma = \mathrm{Var}(\Psi)$ and B a diagonal matrix with diagonal elements

$$b_{ii} \;=\; w(x_{\alpha i}) \frac{1}{\sigma v(x_{\alpha i}^T \beta_\alpha)} h'(x_{\alpha i}^T \beta_\alpha)^2 \,\mathrm{E}\, r_{\alpha i} \psi_c(r_{\alpha i}).$$

Using the generalized inverse for which

$$X_\alpha X_\alpha^- = \mathrm{blockdiag}(I, 0) = E_{n,p_\alpha},$$

we have to show that $\mathrm{trace}(M_\alpha^{-1} Q_\alpha M_\alpha^{-1} \Gamma_\alpha)$ is monotone in p_α. Indeed,

$$\begin{aligned}
\mathrm{trace}(M_\alpha^{-1} Q_\alpha M_\alpha^{-1} \Gamma_\alpha) &= \mathrm{trace}(X_\alpha^- B^{-1} X_\alpha^{-T} X_\alpha^T A^{1/2} \Sigma A^{1/2} X_\alpha X_\alpha^- B^{-1} X_\alpha^{-T} X_\alpha^T W_{\Gamma_\alpha} X_\alpha) \\
&= \mathrm{trace}(X_\alpha X_\alpha^- B^{-1} X_\alpha^{-T} X_\alpha^T A^{1/2} \Sigma A^{1/2} X_\alpha X_\alpha^- B^{-1} X_\alpha^{-T} X_\alpha^T W_{\Gamma_\alpha}) \\
&= \mathrm{trace}(E_{n,p_\alpha} B^{-1} E_{n,p_\alpha} A^{1/2} \Sigma A^{1/2} E_{n,p_\alpha} B^{-1} E_{n,p_\alpha} W_{\Gamma_\alpha}) \\
&= \frac{1}{2} \sum_{i=1}^{p_\alpha} \mathrm{Var}(\Psi_i) \frac{a_{ii}(h_{\alpha i}^{\prime 2} \,\mathrm{E}\, \psi_i' - h_{\alpha i}'' \,\mathrm{E}\, \psi_i)}{b_{ii}^2},
\end{aligned}$$

so this function is monotone in p_α under the same conditions as the maximum likelihood estimator.

S3. The monotonicty condition

For the log link which is often used in Poisson and gamma models,

$$h(\eta_{\alpha i}) = h'(\eta_{\alpha i}) = h''(\eta_{\alpha i}) = \exp(\eta_{\alpha i}) > 0,$$

and for the reciprocal link which is often used in gamma models,

$$h(\eta_{\alpha i}) = \frac{1}{\eta_{\alpha i}}, \quad h'(\eta_{\alpha i}) = -\frac{1}{\eta_{\alpha i}^2}, \quad h''(\eta_{\alpha i}) = \frac{2}{\eta_{\alpha i}^3} > 0.$$

However, for many right skewed distributions like the Poisson and gamma, anti-symmetric ψ functions with sufficiently large b truncate more of the upper tail than the lower tail so $\mathrm{E}\,\psi_i \le 0$. To see this, note that for $\rho(z) = \min(z^2, b^2)$, we can write

$$\begin{aligned}
\mathrm{E}\,\psi_i &= \int_{-\mu/\sigma}^{\infty} \psi(z) dF(\sigma z + \mu) \\
&= \int_{-\min(b,\mu/\sigma)}^{b} 2z dF(\sigma z + \mu) \\
&= -\int_{-\mu/\sigma}^{-\min(b,\mu/\sigma)} 2z dF(\sigma z + \mu) - \int_{b}^{\infty} 2z dF(\sigma z + \mu) \\
&\le 0
\end{aligned}$$

provided b is large enough to ensure that $\int_{-\mu/\sigma}^{-\min(b,\mu/\sigma)} z dF(\sigma z+\mu) + \int_b^\infty z dF(\sigma z+\mu) \ge 0$. It follows that $h''_{\alpha i}\,\mathrm{E}\,\psi_i \le 0$ and (11) holds in these cases.

For the logistic link

$$h(\eta_{\alpha i}) = \frac{\exp(\eta_{\alpha i})}{1+\exp(\eta_{\alpha i})}, \quad h'(\eta_{\alpha i}) = \frac{\exp(\eta_{\alpha i})}{(1+\exp(\eta_{\alpha i}))^2}, \quad h''(\eta_{\alpha i}) = \frac{\exp(\eta_{\alpha i}) - \exp(2\eta_{\alpha i})}{(1+\exp(\eta_{\alpha i}))^3},$$

so that $h_{\alpha i} < 1/2$, $h''_{\alpha i} > 0$ if $\eta_{\alpha_i} < 0$ and $h_{\alpha i} > 1/2$, $h''_{\alpha i} < 0$ if $\eta_{\alpha_i} > 0$, and we need a more careful analysis. The Bernoulli model can be left or right skewed depending on the value of $h_{\alpha i}$, so $\mathrm{E}\,\psi_i$ can be positive or negative. Fortunately, for anti-symmetric ψ,

$$\begin{aligned}
\mathrm{E}\,\psi_i &= \mathrm{E}\,\psi\{(y_i - h_{\alpha_i})/h'_{\alpha_i}\} \\
&= \psi(-h_{\alpha_i}/h'_{\alpha_i})(1-h_{\alpha_i}) + \psi\{(1-h_{\alpha_i})/h'_{\alpha_i}\}h_{\alpha_i} \\
&= -\psi(h_{\alpha_i}/h'_{\alpha_i})(1-h_{\alpha_i}) + \psi\{(1-h_{\alpha_i})/h'_{\alpha_i}\}h_{\alpha_i}
\end{aligned}$$

from which $\mathrm{E}\,\psi_i \le 0$ if $\eta_{\alpha_i} < 0$ and $\mathrm{E}\,\psi_i \ge 0$ if $\eta_{\alpha_i} > 0$, so that $h''_{\alpha i}\,\mathrm{E}\,\psi_i \le 0$ and the monotonicity condition holds.