Web-based Supplementary Materials for “Score test variable screening”

by Sihai Dave Zhao and Yi Li

Web Appendix A: Theoretical properties of score test screening

Theoretical justifications can be easier to derive for score test-based screening compared to Wald test-based screening. The main task is to find a finite-sample bound for $\mathrm{P}\{|U_j^M(0)| \geq \gamma_n\}$, which can often be done by applying Bernstein-type inequalities. In contrast, Wald test-based screening requires deriving non-asymptotic tail bounds for the marginal estimators, which can be considerably more involved. We will give a sufficient condition that, under certain assumptions, will guarantee sure screening and false positive control.

We assume throughout that the covariates have mean 0 and variance 1. Let $u_j^M(\beta_j)$ be the limiting marginal estimating equation, such that $U_j^M(\beta_j) \to u_j^M(\beta_j)$.

Assumption 1 *There exists some constant* $c_1 > 0$ *such that* $\min_{j \in \mathcal{M}} |u_j^M(0)| \geq c_1 n^{-\kappa}$ *with* $0 < \kappa < 1/2$.

Assumption 2 $\|\mathbf{u}(\mathbf{0})\|_2^2 = \sqrt{\sum_j u_j^M(0)^2}$.

Assumption 3 *The negative Jacobian* $\mathbf{i}(\boldsymbol{\beta}) = -\partial \mathbf{u} / \partial \boldsymbol{\beta}$ *exists.*

Assumption 4 *There exists some constant* $c_3 > 0$ *such that* $\|\boldsymbol{\beta}_0\|_2 \leq c_3$.

Assumption 1 ensures that the limiting numerator of the marginal score test for $H_0 : \beta_{0j} = 0$, is still large enough to detect. For example, for generalized linear models when $K_i = 1$, $u_j^M(\beta_j) = n^{-1} \sum_i \mathrm{E}\{X_{ij}(Y_i - g^{-1}(X_{ij}\beta_j)\}$ with g equal to the canonical link function, so Assumption 1 is equivalent to assuming that $|\mathrm{cov}\{g^{-1}(\mathbf{X}_i^T \boldsymbol{\beta}_0), X_{ij}\}| \geq c_1 n^{-\kappa}$. Fan and Song (2010) make exactly this assumption to prove the sure screening property in their Theorem 4(ii). Assumption 2 relates the marginal expected estimating equations to the full expected estimating equation. This is a mild assumption, because frequently $u_j\{(\mathbf{0})\} = u_j^M(0)$, where u_j is the j^{th} component of $\mathbf{u}$. This holds, for example, for generalized estimating equations when $\mathrm{E}(\mathbf{Y}_i \mid \mathbf{X}_i) = \mu(\mathbf{X}_i \boldsymbol{\beta}_0)$ for some mean function μ (Zeger et al., 1988). Assumption 3 can hold even if the sample estimating equation $\mathbf{U}$ is nondifferentiable. Assumption 4 merely requires that there exist an upper bound on the size of the true $\boldsymbol{\beta}_0$ that does not grow with n, which is a reasonable condition

Next we give a sufficient condition which will ensure good screening properties.

Condition 1 *For* κ *from Assumption 1 and any constant* $c_2 > 0$, $p_n \mathrm{P}(|U_j^M(0) - u_j^M(0)| \geq c_2 n^{-\kappa}) \to 0$.

Condition 1 requires that that the probability that $U_j^M(0)$ is not close to $u_j^M(0)$ approaches 0 faster than p_n approaches ∞, so that we can use $U_j^M(0)$ for screening in high dimensions. This condition must be separately verified for different regression models. Condition 1 will often hold even when p_n grows exponentially in n.

For many estimating equations, to verify Condition 1 we need additional assumptions on the tails of $\mathbf{X}_i$ and $\mathbf{Y}_i$ and on the rate of p_n, such as the following:

Assumption 5 *There exist constants $l_0, l_1, \eta > 0$ such that for all j, $\mathrm{P}(|X_{ij}| > s) \leq l_0 \exp(-l_1 s^\eta)$ for sufficiently large s.*

Tail conditions of this type were also assumed in Fan and Song (2010) and Gorst-Rasmussen and Scheike (2013). When $U_j^M(0)$ is a sum of independent random variables, we can appeal to the usual Bernstein's inequality. A similar approach applies when $U_j^M(0)$ is a U-statistic (Hoeffding, 1963). Establishing this condition for more complicated $U_j^M(0)$, such as the marginal score equations for the Cox model, is more involved (Gorst-Rasmussen and Scheike, 2013).

Under these assumptions, and given the sufficient condition, score test screening has the following theoretical guarantees:

Theorem 1 (Sure screening) *Let $\gamma_n = c_1 n^{-\kappa}/2$. Under Assumption 1, if Condition 1 is satisfied, then $\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \to 1$.*

Theorem 2 (False positive control) *Let $\gamma_n = c_1 n^{-\kappa}/2$. Under Assumptions 1–4, if Condition 1 is satisfied, then $\mathrm{P}(|\hat{\mathcal{M}}| \leq 16 c_3^2 \sigma_{\max}^{*2} / c_1^2 n^{-2\kappa}) \to 1$, where $\sigma_{\max}(\mathbf{A})$ is the largest singular value of the matrix $\mathbf{A}$ and $\sigma_{\max}^* = \sup_{0<t<1} \sigma_{\max}\{\mathbf{i}(t\boldsymbol{\beta}_0)\}$.*

Theorem 1 shows that marginal score testing maintains the sure screening property, and is thus an attractive alternative to marginal Wald testing. Theorem 2 shows that the number of selected covariates is not too large, with high probability. For example, if $\sigma_{\max}^*$ increased only polynomially in n, $|\hat{\mathcal{M}}|$ would increase polynomially. At the same time, p_n can frequently be allowed to increase exponentially in n. Thus the false positive rate would decrease quickly to zero.

The presence of $\sigma_{\max}^*$ in Theorem 2 reflects the dependence of $|\hat{\mathcal{M}}|$ on the degree of collinearity of our data. The collinearity of estimating equations not only depends on the design matrix, but also varies across the parameter space. For example, Mackinnon and Puterman (1989) and Lesaffre and Marx (1993) showed that generalized linear models can be collinear even if their design matrices are not, and vice versa. In our situation, we are concerned with collinearity along the line segment between $\boldsymbol{\beta}_0$ and $\mathbf{0}$.

Proof of Theorem 1

The event $\{\mathcal{M} \subseteq \hat{\mathcal{M}}\}$ equals $\{\min_{j \in \mathcal{M}} |U_j^M(0)| \geq \gamma_n\}$, so it is easy to see that

$$\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \geq 1 - \sum_{j \in \mathcal{M}} \mathrm{P}(|U_j^M(0)| < \gamma_n).$$

By the triangle inequality, we know that for all j, $|u_j^M(0)| \leq |U_j^M(0) - u_j^M(0)| + |U_j^M(0)|$, and by Assumption 1 we see that $c_1 n^{-\kappa} - |U_j^M(0)| \leq |U_j^M(0) - u_j^M(0)|$ for all $j \in \mathcal{M}$. Therefore, $|U_j^M(0)| < \gamma_n$ for $j \in \mathcal{M}$ implies $|U_j^M(0) - u_j^M(0)| \geq c_1 n^{-\kappa}/2$, so that

$$\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \geq 1 - s_n \mathrm{P}(|U_j^M(0) - u_j^M(0)| \geq c_1 n^{-\kappa}/2).$$

The right-hand side goes to 1 if Condition 1 is satisfied. □

Proof of Theorem 2

If Condition 1 is satisfied, then

$$\mathrm{P}\{\max_j |U_j^M(0) - u_j^M(0)| \geq c_1 n^{-\kappa}/4\} \geq 1 - p_n \mathrm{P}\{|U_j^M(0) - u_j^M(0)| \geq c_1 n^{-\kappa}/4\} \to 1.$$

On the event $\max_j |U_j(0) - u_j(0)| \leq c_1[n/m]^{-\kappa}/4$, $|U_j(0)| \geq \gamma_n$ implies that $|u_j(0)| \geq c_1[n/m]^{-\kappa}/4$. This means that

$$\begin{aligned} |\hat{\mathcal{M}}| &= |\{j : |U_j(0)| \geq \gamma_n\}| \leq |\{j : |u_j(0)| \geq c_1[n/m]^{-\kappa}/4\}| \\ &\leq \sum_j u_j^M(0)^2/(c_1 n^{-\kappa}/4)^2 = \|\mathbf{u}(\mathbf{0})\|_2^2 16/c_1^2 n^{-2\kappa}, \end{aligned}$$

where the last equality follows from Assumption 2. Using the generalization of the mean value theorem to vector-valued functions (Hall and Newell, 1979) and Assumptions 3 and 4,

$$\|\mathbf{u}(\mathbf{0})\|_2 = \|\mathbf{u}(\boldsymbol{\beta}_0) - \mathbf{u}(\mathbf{0})\|_2 \leq \sup_{0<t<1} \|\mathbf{i}(t\boldsymbol{\beta}_0)\|_2 \|\boldsymbol{\beta}_0\|_2 \leq c_3 \sup_{0<t<1} \sigma_{\max}\{\mathbf{i}(t\boldsymbol{\beta}_0)\} = c_3 \sigma^*_{\max},$$

which implies that $|\hat{\mathcal{M}}| \leq 16 c_3^2 \sigma^{*2}_{\max}/c_1^2 n^{-2\kappa}$. $\square$

Verifying Condition 1 for censored quantile regression

Define

$$Z_i^{(1)} = X_{ij}\left[\tau I\{h(Y_i) > \beta_{int}\} - (1-\tau)\frac{I\{h(Y_i) \leq \beta_{int}\}\delta_i}{S_{h(C)}\{h(Y_i)\}} S_{h(C)}(\beta_{int})\right] - u_j^M(0)$$

and

$$Z_i^{(2)} = X_{ij}(1-\tau) I\{h(Y_i) \leq \beta_{int}\}\delta_i \left[\frac{S_{h(C)}(\beta_{int})}{S_{h(C)}\{h(Y_i)\}} - \frac{\hat{S}_{h(C)}(\beta_{int})}{\hat{S}_{h(C)}\{h(Y_i)\}}\right].$$

We assume that β_{int} is either known or has been estimated from an independent dataset, so that in the remainder of the proof we can treat it as a constant. Then

$$\mathrm{P}(|U_j^M(0) - u_j^M(0)| \geq 2t) \leq \mathrm{P}(n^{-1}|\sum_i Z_i^{(1)}| \geq t) + \mathrm{P}(n^{-1}|\sum_i Z_i^{(2)}| \geq t).$$

To bound the term containing $Z_i^{(1)}$ we first note that $\mathrm{E}(Z_i^{(1)}) = 0$ by the definition of $u_j^M(0)$. Also, by assumption $S_{h(C)}(\beta_{int}) > 0$, so the term $I\{h(Y_i) \leq \beta_{int}\}\delta_i / S_{h(C)}\{h(Y_i)\}$, which is nonzero only when $h(Y_i) \leq \beta_{int}$, can be at most $S_{h(C)}(\beta_{int})^{-1}$. Therefore when $|X_{ij}| \leq M$ for all i, j, where $M > 0$, $|Z_i^{(1)}| \leq 2M$. Using Bernstein's inequality van der Vaart and Wellner (1996) and Assumption 5,

$$\mathrm{P}(n^{-1}|Z_1^{(1)} + \ldots + Z_n^{(1)}| \geq t) \leq 2\exp\left(-\frac{1}{2}\frac{t^2 n}{4M^2 + 2Mt/3}\right) + n l_0 \exp(-l_1 M^\eta).$$

To bound the term containing $Z_i^{(2)}$, we first note that

$$\mathrm{P}(n^{-1}|Z_1^{(2)} + \ldots + Z_n^{(2)}| \geq t) \leq \mathrm{P}(\max_i |Z_i^{(2)}| \geq t) \leq n\mathrm{P}(|Z_i^{(2)}| \geq t).$$

Since $Z_i^{(2)} = 0$ when $h(Y_i) > \beta_{int}$, $\mathrm{P}(|Z_i^{(2)}| \geq t) = \mathrm{P}\{|Z_i^{(2)}| \geq t \cap h(Y_i) \leq \beta_{int}\}$. For notational convenience let $S_{int} = S_{h(C)}(\beta_{int})$, $\hat{S}_{int} = \hat{S}_{h(C)}(\beta_{int})$, $S_Y = S_{h(C)}\{h(Y_i)\}$, and $\hat{S}_Y = \hat{S}_{h(C)}\{h(Y_i)\}$. Then

$$\begin{aligned}
\mathrm{P}(|Z_i^{(2)}| \geq t) &\leq \mathrm{P}\left\{\left|\frac{S_{int}}{S_Y} - \frac{\hat{S}_{int}}{\hat{S}_Y}\right| \geq \frac{t}{M(1-\tau)} \cap h(Y_i) \leq \beta_{int}\right\} \\
&\leq \mathrm{P}\left\{|S_{int}\hat{S}_Y - \hat{S}_{int}S_Y| \geq \frac{tS_Y\hat{S}_Y}{M(1-\tau)} \cap h(Y_i) \leq \beta_{int}\right\}.
\end{aligned}$$

Now let $\varepsilon_{int} = |\hat{S}_{int} - S_{int}|$ and $\varepsilon_Y = |\hat{S}_Y - S_Y|$. Then

$$\begin{aligned}
\mathrm{P}(|Z_i^{(2)}| \geq t) \leq &\mathrm{P}\left\{S_{int}\varepsilon_Y + S_Y\varepsilon_{int} \geq \frac{tS_Y(S_Y - \varepsilon_Y)}{M(1-\tau)} \cap h(Y_i) \leq \beta_{int}\right\} \\
\leq &\mathrm{P}\left[\left\{S_{int} + \frac{tS_Y}{M(1-\tau)}\right\}\varepsilon_Y + S_Y\varepsilon_{int} \geq \frac{tS_Y^2}{M(1-\tau)} \cap h(Y_i) \leq \beta_{int}\right] \\
\leq &\mathrm{P}\left[\varepsilon_Y \geq \frac{tS_Y^2}{2M(1-\tau)}\left\{S_{int} + \frac{tS_Y}{M(1-\tau)}\right\}^{-1} \cap h(Y_i) \leq \beta_{int}\right] + \\
&\mathrm{P}\left[\varepsilon_{int} \geq \frac{tS_Y}{2M(1-\tau)} \cap h(Y_i) \leq \beta_{int}\right] \\
\leq &\mathrm{P}\left[\varepsilon_Y \geq \frac{tS_{int}^2}{2M(1-\tau)}\left\{S_{int} + \frac{t}{M(1-\tau)}\right\}^{-1}\right] + \\
&\mathrm{P}\left\{\varepsilon_{int} \geq \frac{tS_{int}}{2M(1-\tau)}\right\},
\end{aligned}$$

where the last inequality follows because $h(Y_i) \leq \beta_{int}$ implies that $S_Y \geq S_{int}$. Now by the theorem of Bitouzé et al. (1999), there exists some constant C such that

$$\mathrm{P}(n^{1/2}\|S_{h(T)}(\hat{S}_{h(C)} - S_{h(C)})\|_\infty \geq \lambda) \leq 2.5\exp(-2\lambda^2 + C\lambda),$$

where $S_{h(T)}$ is the survival function of the $h(T_i)$. When $h(Y_i) \leq \beta_{int}$, $S_{h(T)}\{h(Y_i)\} \geq S_{h(T)}(\beta_{int})$, so we can apply this theorem to

$$\mathrm{P}\left[n^{1/2}S_{h(T)}\{h(Y_i)\}\varepsilon_Y \geq \frac{tn^{1/2}S_{int}^2}{2M(1-\tau)}\left\{S_{int} + \frac{tS_{h(T)}\{\beta_{int}\}}{M(1-\tau)}\right\}^{-1}\right]$$

and

$$\mathrm{P}\left\{n^{1/2}S_{h(T)}(\beta_{int})\varepsilon_{int} \geq \frac{tn^{1/2}S_{int}S_{h(T)}(\beta_{int})}{M(1-\tau)}\right\}$$

to bound $\mathrm{P}(n^{-1}|Z_1^{(2)} + \ldots + Z_n^{(2)}| \geq t)$.

By setting $t = c_2 n^{-\kappa}/2$ and $M = n^{(1-2\kappa)/(\eta+2)}$ and combining the previous tail bounds, we can conclude that $\mathrm{P}(|U_j^M(0) - u_j^M(0)| \geq 2t) \leq O\{\exp(-n^{(1-2\kappa)\eta/(\eta+2)})\}$. □

References

D. Bitouzé, B. Laurent, and P. Massart. A Dvoretzky–Kiefer–Wolfowitz type inequality for the Kaplan–Meier estimator. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 35, pages 735–763. Elsevier, 1999.

J. Fan and R. Song. Sure independence screening in generalized linear models and NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

A. Gorst-Rasmussen and T. H. Scheike. Independent screening for single-index hazard rate models with ultra-high dimensional features. *Journal of the Royal Statistical Society, Ser. B*, 75:217–245, 2013.

W. S. Hall and M. L. Newell. The mean value theorem for vector valued functions: a simple proof. *Mathematics Magazine*, 52(3):157–158, 1979.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

E. Lesaffre and B. D. Marx. Collinearity in generalized linear regression. *Communications in Statistics – Theory and Methods*, 22:1933–1952, 1993.

M. J. Mackinnon and M. L. Puterman. Collinearity in generalized linear models. *Communications in Statistics – Theory and Methods*, 18:3463–3472, 1989.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer, New York, 1996.

S. L. Zeger, K.-Y. Liang, and P. S. Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.