Web-based Supplementary Materials for

"Model Selection and Inference for Censored Lifetime Medical Expenditures"

Brent A. Johnson[1,*], **Qi Long**[2], **Yijian Huang**[2], **Kari Chansky**[3], **Mary Redman**[3]

[1]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, U.S.A.

[2]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, U.S.A.

[3]The Fred Hutchinson Cancer Research Center, Seattle, Washington U.S.A.

[*]*email*: brent_johnson@urmc.rochester.edu

Web Appendix A: Large Sample Theory for $\widehat{\boldsymbol{\beta}}_\lambda$

To facilitate a statement about the asymptotic properties of the mark-scale regularized coefficient estimator, we need some additional notation, including a simple expression for asymptotic covariance of $n^{1/2}(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0)$. Without loss of generality, let $\widehat{\boldsymbol{\vartheta}}_0$ be a root to the weighted log-rank estimating function, $\mathbb{S}_\vartheta(\boldsymbol{\theta}) = o_p(n^{1/2})$, and $\widehat{\boldsymbol{\beta}}_\lambda$ be a root of the penalized estimating function $\mathbb{S}_{\beta,\lambda}(\mathbf{b}, \widehat{\boldsymbol{\vartheta}}_0)$, $\mathbb{S}_{\beta,\lambda}(\mathbf{b}, \mathbf{v}) = (S_{\beta,\lambda,1}(\mathbf{b}, \mathbf{v}), \ldots, S_{\beta,\lambda,d}(\mathbf{b}, \mathbf{v}))^{\mathrm{T}}$. Next, partition the asymptotic slope matrix

$$\nabla \left[\lim_n n^{-1} \begin{pmatrix} \mathbb{S}_\beta(\boldsymbol{\theta}) \\ \mathbb{S}_\vartheta(\boldsymbol{\theta}) \end{pmatrix} \right] = \begin{pmatrix} \boldsymbol{\Gamma}_{\beta\beta} & \boldsymbol{\Gamma}_{\beta\vartheta} \\ \boldsymbol{\Gamma}_{\vartheta\beta} & \boldsymbol{\Gamma}_{\vartheta\vartheta} \end{pmatrix} = \boldsymbol{\Gamma}$$

and define

$$\boldsymbol{\Xi} = E\left[\left\{ \mathbf{s}(\mathbf{O}_i, \boldsymbol{\theta}_0) + \boldsymbol{\Gamma}_{\beta\vartheta}\boldsymbol{\Gamma}_{\vartheta\vartheta}^{-1} \int_{-\infty}^{\infty} \mathbf{h}_i(u, \boldsymbol{\vartheta}_0)\, dM_{i,\vartheta}(u, \boldsymbol{\vartheta}_0) \right\}^{\otimes 2} \right],$$

where $\mathbb{S}(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{s}(\mathbf{O}_i, \boldsymbol{\theta})$, $\mathbf{O}_i = (Y_i\delta_i, X_i, \delta_i, \mathbf{z}_i)$, $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathrm{T}}$ for a column vector $\mathbf{a}$, $\mathbf{h}_i(u, \boldsymbol{\vartheta})$ is a mean-zero function of the data and $M_{i,\vartheta}(u, \boldsymbol{\vartheta})$ is the time-scale martingale process for the ith subject (e.g. Tsiatis, 1990; Wei et al., 1990; Ying, 1993).

Lemma 1 *Under conditions (A)–(E) in Huang (2002, Appendix) and conditions A.1–A.2 in Section 2,*

(a) there exist $n^{1/2}$-consistent roots to $\mathbb{S}_\lambda(\boldsymbol{\theta})$ such that $\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$;

(b) $\lim_n P(\widehat{\beta}_{\lambda,j} \neq 0,\ \textit{for } j \in \mathcal{A}^c) = 0$*;*

(c) If we let $\boldsymbol{\beta}_{\mathcal{A}} = \{\beta_{0j}|\beta_{0j} \neq 0\}$, $\widehat{\boldsymbol{\beta}}_{\mathcal{A}} = \{\widehat{\beta}_{\lambda,j}|\beta_{0j} \neq 0\}$, $\mathbb{S}_{\mathcal{A}}(\boldsymbol{\theta}) = \{S_{\beta,\lambda,j}(\boldsymbol{\theta})|\beta_{0j} \neq 0\}$ *and* $d_0 = |\mathcal{A}|$, *then*

$$n^{1/2}(\boldsymbol{\Gamma}^{\dagger}_{\mathcal{A}\mathcal{A}} + \boldsymbol{\Sigma}_{\mathcal{A}}) \left\{\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}} + (\boldsymbol{\Gamma}^{\dagger}_{\mathcal{A}\mathcal{A}} + \boldsymbol{\Sigma}_{\mathcal{A}})^{-1}\mathbf{g}\right\} \longrightarrow_d N\left(0, \boldsymbol{\Xi}^{\dagger}_{\mathcal{A}}\right)$$

where $\boldsymbol{\Sigma}_{\mathcal{A}}$ *is the* $d_0 \times d_0$ *sub-matrix of* $diag\{-(\partial/\partial\beta_j)q_\lambda(|\beta_j|)sign(\beta_j)\}$, $\mathbf{g} = -\{q_\lambda(|\beta_j|)sign(\beta_j),\ j \in \mathcal{A}\}$, $\boldsymbol{\Gamma}^{\dagger}_{\mathcal{A}\mathcal{A}}$ *is the upper-left* $d_0 \times d_0$ *sub-matrix of* $\boldsymbol{\Gamma}^{\dagger}$, $\boldsymbol{\Gamma}^{\dagger}_{\mathcal{A}\vartheta}$ *is the upper-right* $d_0 \times d$ *sub-matrix of* $\boldsymbol{\Gamma}^{\dagger}$,

$$\boldsymbol{\Gamma}^{\dagger} = \nabla[\lim_n n^{-1}\{\mathbb{S}^{\mathrm{T}}_{\mathcal{A}}(\boldsymbol{\theta}), \mathbb{S}^{\mathrm{T}}_{\vartheta}(\boldsymbol{\theta})\}^{\mathrm{T}}],$$

$$\boldsymbol{\Xi}^{\dagger}_{\mathcal{A}} = E\left[\left\{\mathbf{s}_{\mathcal{A}}(\mathbf{O}_i, \boldsymbol{\theta}) + \boldsymbol{\Gamma}^{\dagger}_{\mathcal{A}\vartheta}\boldsymbol{\Gamma}^{-1}_{\vartheta\vartheta}\int_{-\infty}^{\infty} \mathbf{h}_i(u, \boldsymbol{\vartheta})\, dM_{i,\vartheta}(u, \boldsymbol{\vartheta})\right\}^{\otimes 2}\right],$$

and $\mathbb{S}_{\mathcal{A}}(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{s}_{\mathcal{A}}(\mathbf{O}_i, \boldsymbol{\theta})$.

Proof. Under conditions (A)–(E) in Huang (2002), the $2d$-estimating function $\mathbb{S}(\boldsymbol{\theta})$ is, almost surely, asymptotically linear in a neighborhood of the true value $\boldsymbol{\theta}_0$. Because conditions A.1–A.2 agree with condition C.2 of Johnson et al. (2008), the conclusions of Lemma 1 follow by an extension of Theorem 1 from Johnson et al. (2008). □

Web Appendix B: Simulation Results With 2 Predictors

To investigate the operating characteristics of the resampling procedure in finite samples without regularization, we repeated one of the simulation exercises in Huang (2002). We simulated data according to joint lifetime model

$$Y = z_1\beta_1 + z_2\beta_2 + \varepsilon_Y, \qquad\qquad T = z_1\vartheta_1 + z_2\vartheta_2 + \varepsilon_T,$$

with $d = 2$ independent covariates z_1 and z_2, where z_1 is Bernoulli with equal probability on -1 and 1 and z_2 is $\mathrm{Un}(-1, 1)$. The coefficients are $(\vartheta_1, \vartheta_2) = (1, 1)$ and $(\beta_1, \beta_2) = (1, 1)$, and the error vectors $\boldsymbol{\varepsilon} = (\varepsilon_Y, \varepsilon_T)^{\mathrm{T}}$ are independent and identically distributed bivariate normal with $E(\varepsilon_T) = E(\varepsilon_Y) = 0$, $\mathrm{var}(\varepsilon_T) = \mathrm{var}(\varepsilon_Y) = 1$ and $\mathrm{cov}(\varepsilon_Y, \varepsilon_T) = 0.5$. The censoring distribution was $\mathrm{Un}(0, 3.25)$ and $\mathrm{Un}(0, 0.9)$ which lead to 20% and 40% average censoring levels, respectively.

The simulation results for 500 Monte Carlo data sets and using $B = 500$ resamples are displayed in Table 1. We find that the performance of the point and interval estimator perform somewhat better when the censoring is 20% compared with 40%. For example, when the sample size was $n = 80$, the coverage probabilities for the log-rank estimator were roughly 2 points closer to the nominal level when censoring was 20% compared to 40%. There are some finite sample differences between results for the Bernoulli and uniform covariate but no strong trends to report. The log-rank estimator was more precise than the Gehan estimator. For example, the log-rank estimator was 10-15% more efficient than the Gehan estimator at $n = 80$ and 20% censoring. The Wald and percentile method both performed similarly in this simulation study and we saw no strong advantage to one versus the other. Finally, the results in Table 1 suggest the coverage probabilities are close to the nominal level but the asymptotic properties are not fully satisfied at $n = 80$. We conducted additional numerical studies at $n = 100$ and larger sample sizes and the empirical coverage probabilities are very close to 95%. More importantly, the results presented in Table 1 suggest that the resampling technique works well even for moderately-sizes samples.

Web Appendix C: Additional Simulation Results for Interval Estimators

In our simulation studies, we were interested in the performance of interval estimators across 4 models, Model 1–4, of varying complexity, with Model 4 being most complex. In the interest of space, only simulation results for lasso and adaptive lasso were presented in the manuscript. Here, we provide a graphical presentation of simulation results for all estimators. These estimators include bridge (Frank and Friedman, 1993), lasso (Tibshirani, 1996), hard thresholding (Antoniadis, 1997), scad (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive lasso (Zou, 2006), and logarithmic penalty (Zou and Li, 2008; Johnson et al., 2008). Each figure contains four panels: empirical coverage probabilities (ECP) for the inactive sets in the upper left-hand panel, ECP for active sets in upper right-hand panel, average expected length for the inactive set in lower left panel, and average expected interval length for the active set in lower right-hand panel. In summary, we find that a normal approximation to the variance of $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}})$ works well across a variety of simulation scenarios although the interval width can be unnecessarily wide at times. The percentile method gave confidence intervals that covered nominally in Model 1 but became too narrow in Models 2–4. High density regions (HDR) similarly performed well in Model 1 and not as well in Models 2–4. The simplified HDR method performed worst among all methods and generally sacrificed coverage probability for narrower intervals.

References

Antoniadis, A. (1997). Wavelets in statistics: A review (with discussion). *Journal of the Italian Statistical Association* **6,** 97–144.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35,** 109–148.

Huang, Y. (2002). Calibration regression of censored lifetime medical cost. *Journal of the American Statistical Association* **97,** 318–327.

Johnson, B. A., Lin, D.-Y., and Zeng, D. (2008). Penalized estimating functions and variable seleciton in semiparametric regression models. *Journal of the American Statistical Association* **103,** 672–680.

Minnier, J., Tian, L., and Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association* **106,** 1371–1382.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58(1),** 267–288.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics* **18(1),** 354–372.

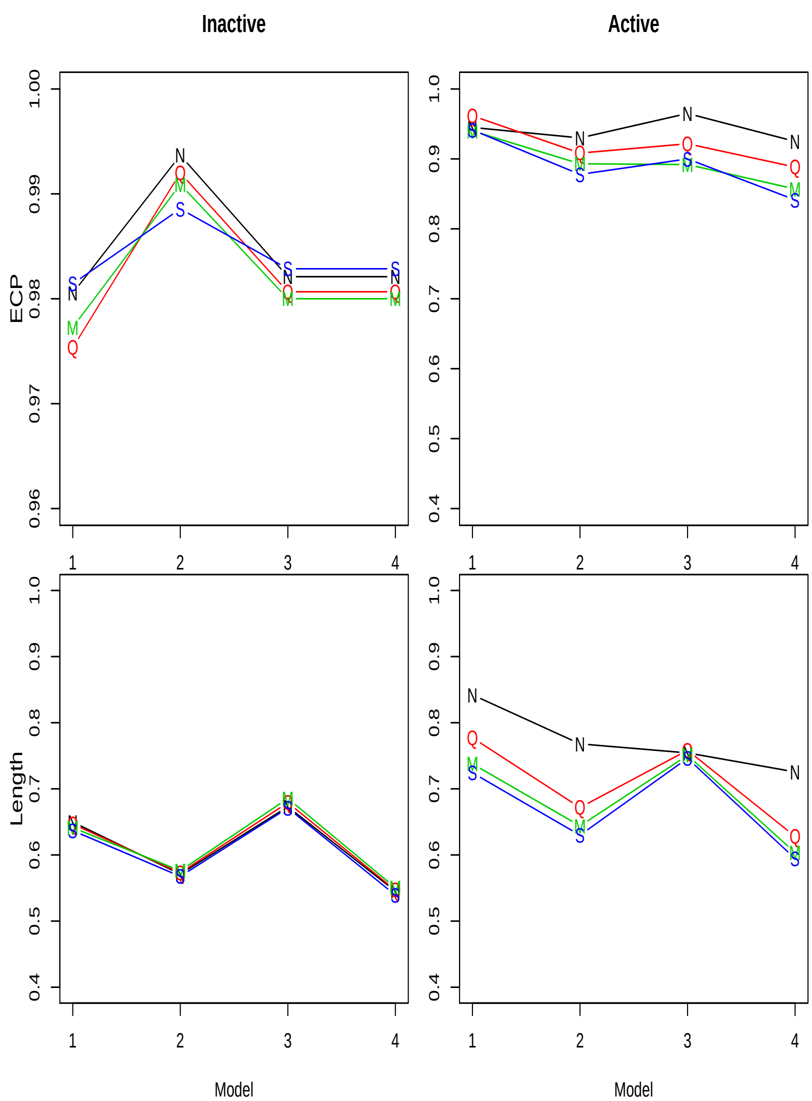

Figure 1: Simulation results of inference procedures for adaptive lasso, including confidence intervals based on normal approximation (N), percentile method (Q), highest density region by Minnier et al. (2011, M) and the simplified method (S).

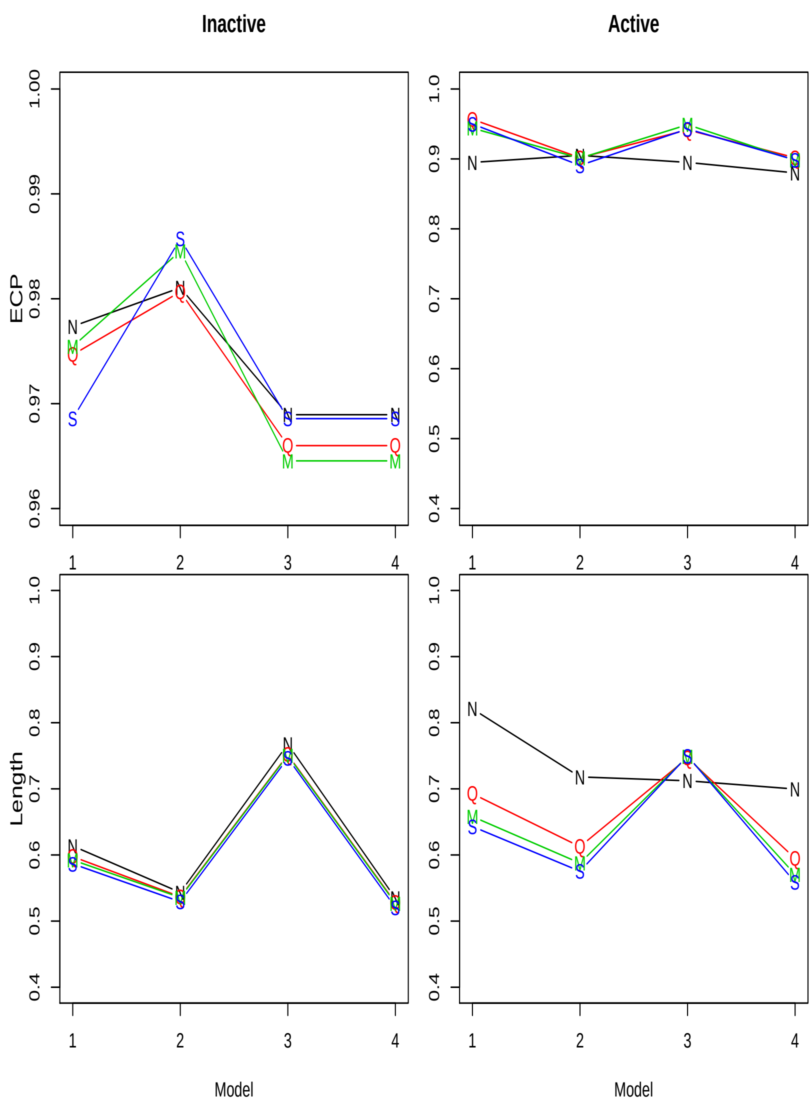

Figure 2: Simulation results of inference procedures for lasso, including confidence intervals based on normal approximation (N), percentile method (Q), highest density region by Minnier et al. (2011, M) and the simplified method (S).

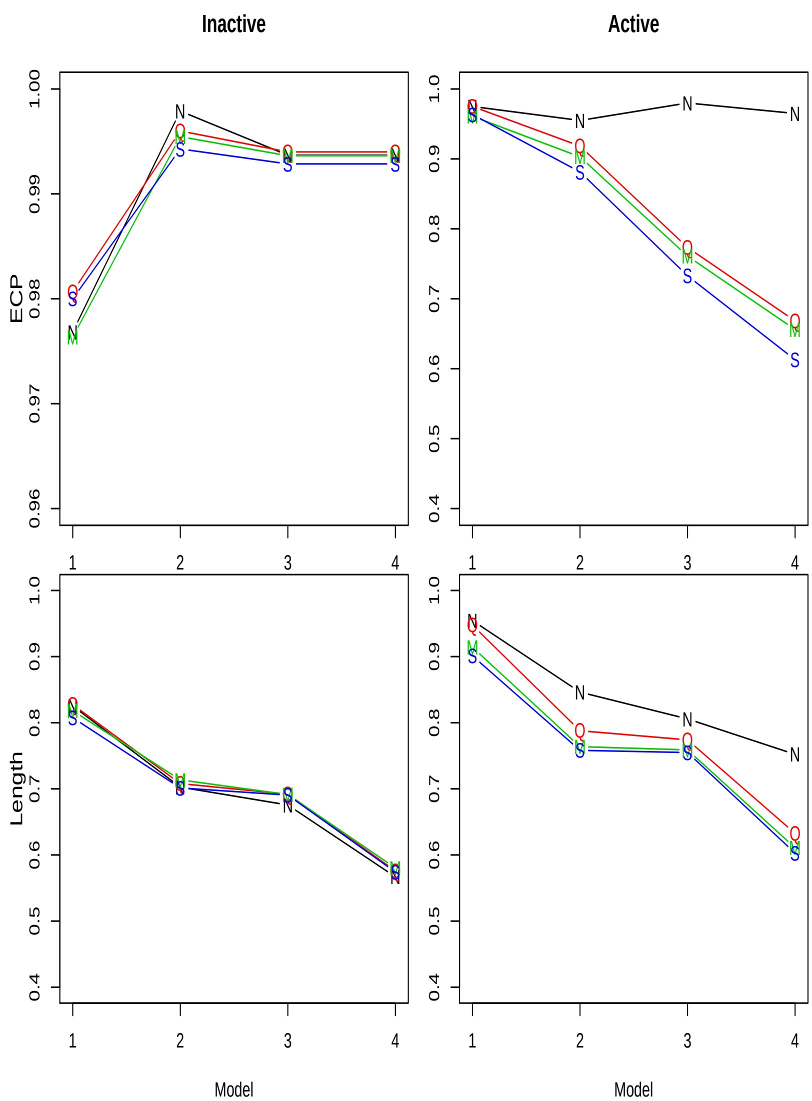

Figure 3: Simulation results of inference procedures for scad, including confidence intervals based on normal approximation (N), percentile method (Q), highest density region by Minnier et al. (2011, M) and the simplified method (S).

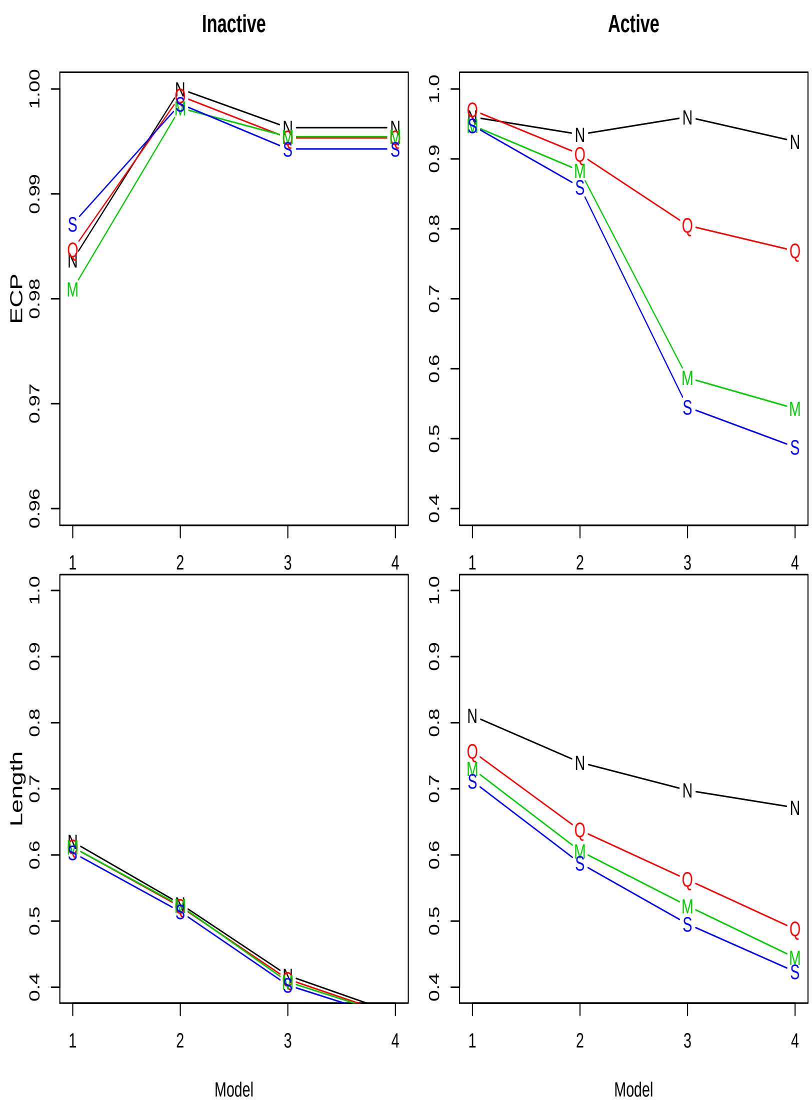

Figure 4: Simulation results of inference procedures for bridge, including confidence intervals based on normal approximation (N), percentile method (Q), highest density region by Minnier et al. (2011, M) and the simplified method (S).

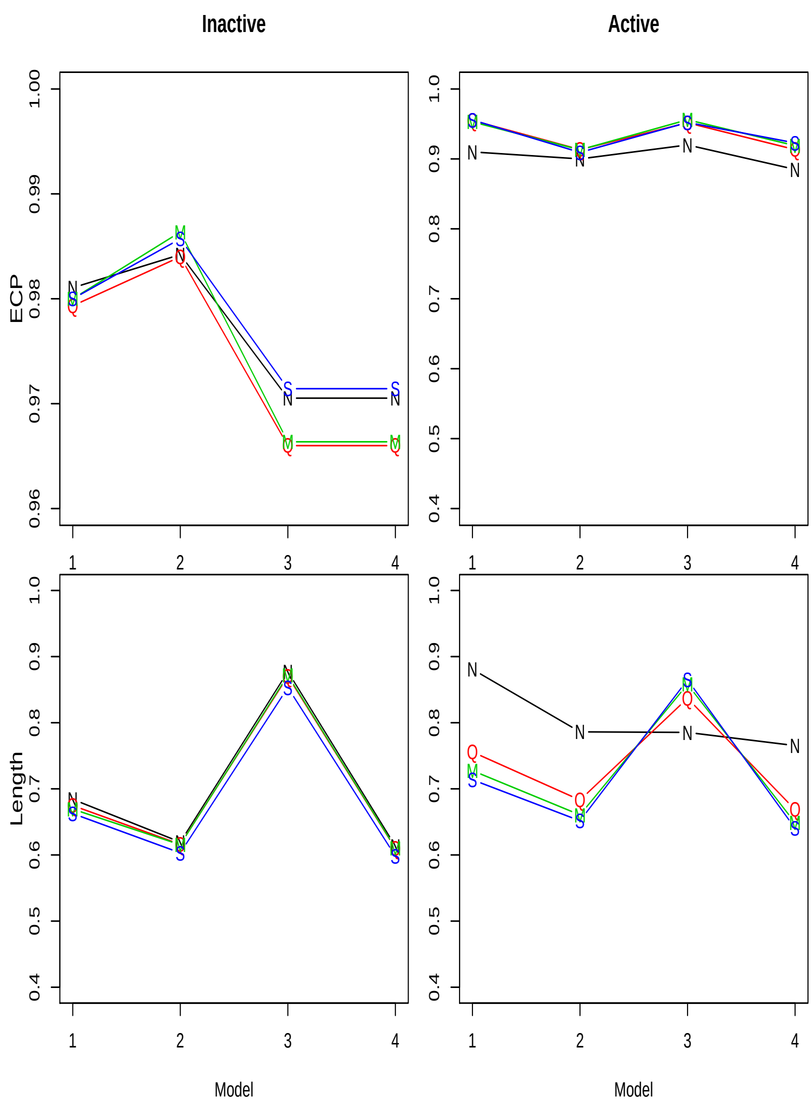

Figure 5: Simulation results of inference procedures for elastic net, including confidence intervals based on normal approximation (N), percentile method (Q), highest density region by Minnier et al. (2011, M) and the simplified method (S).

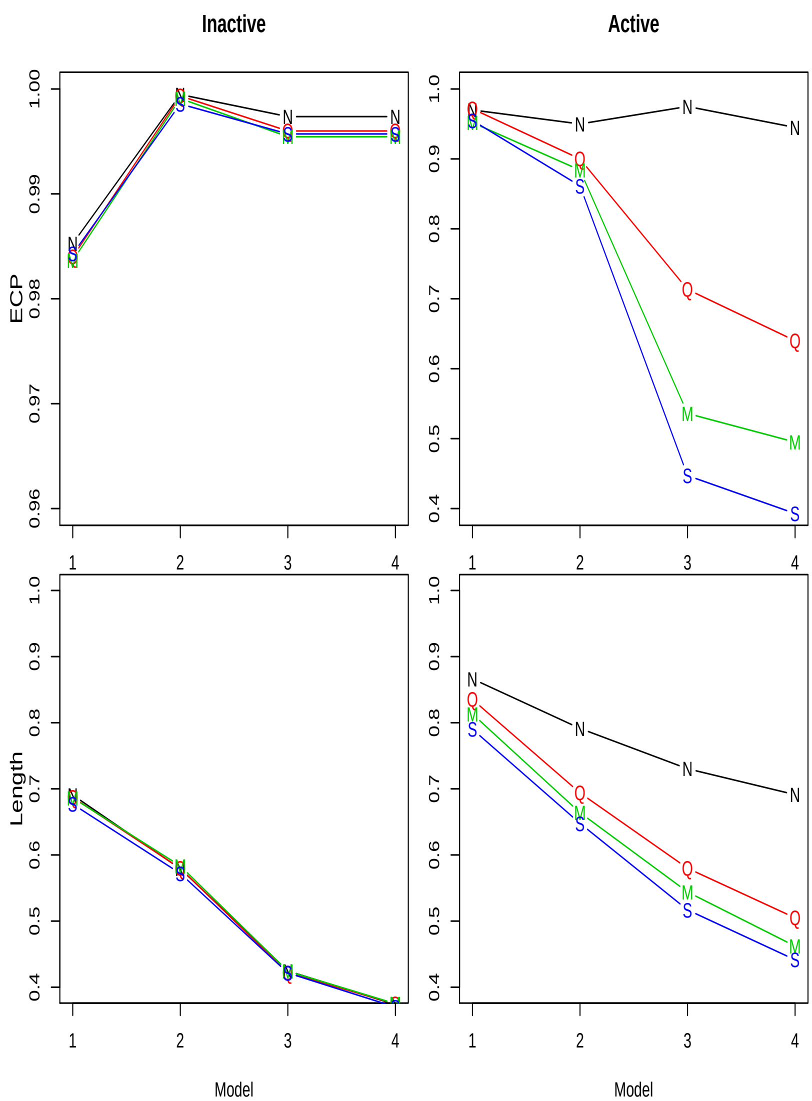

Figure 6: Simulation results of inference procedures for log, including confidence intervals based on normal approximation (N), percentile method (Q), highest density region by Minnier et al. (2011, M) and the simplified method (S).

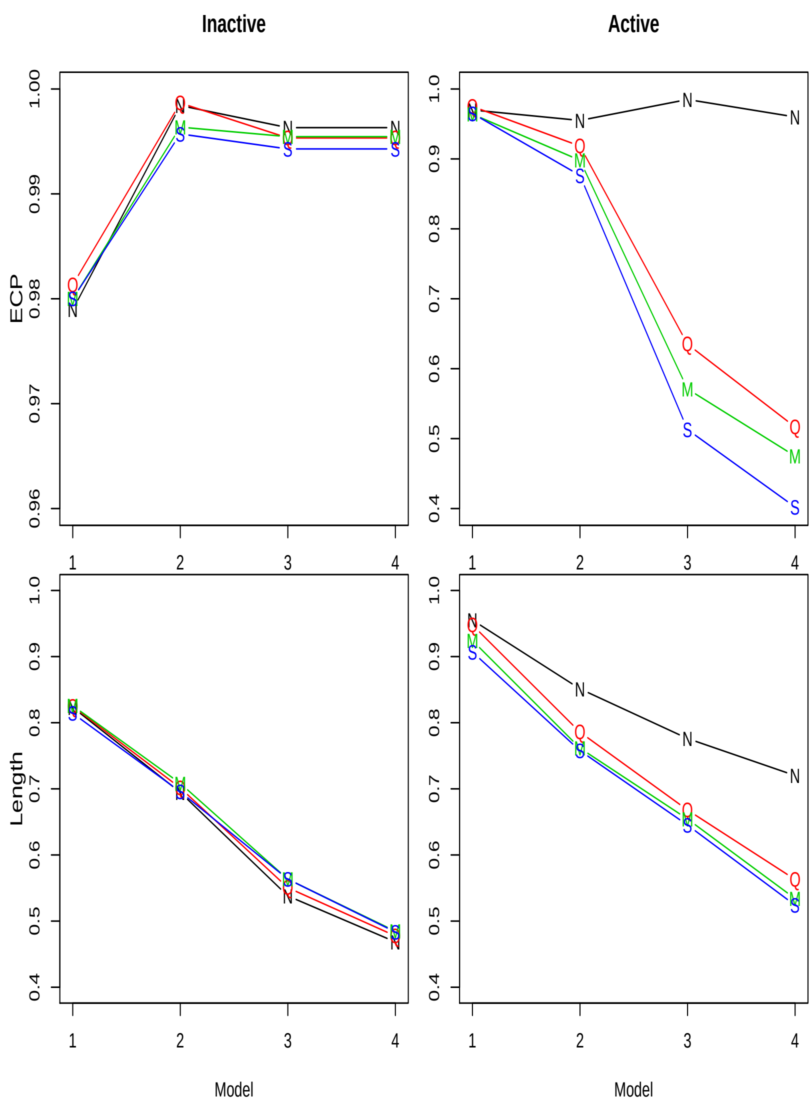

Figure 7: Simulation results of inference procedures for hard thresholding, including confidence intervals based on normal approximation (N), percentile method (Q), highest density region by Minnier et al. (2011, M) and the simplified method (S).

Wei, L. J., Ying, Z., and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77,** 845–851.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *Annals of Statistics* **21,** 76–99.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36,** 1509–1533.

Table 1: Simulations results with $d = 2$ predictors. Table entries include Monte Carlo bias and standard error (SSE), the average of the standard error estimates (SEE), the empirical coverage probabilities (ECP) for Wald-type intervals and using the percentile method. All table entries are multiplied by 1000 except the empirical coverage probabilities which are reported as percentages.

			Gehan					Log-rank				
						ECP					ECP	
n	Cens.	β	Bias	SSE	SEE	Wald	Pcntl.	Bias	SSE	SEE	Wald	Pcntl.
40	20%	β_1	-17	209	193	92.0	92.8	-16	188	178	91.8	93.2
		β_2	14	365	331	92.0	91.4	8	322	307	93.6	93.2
	40%	β_1	9	291	247	87.6	85.8	4	278	236	88.8	88.2
		β_2	28	481	404	88.0	88.0	32	436	382	90.8	89.2
60	20%	β_1	-9	178	160	91.4	91.2	-14	158	146	91.8	92.6
		β_2	-8	323	273	88.0	88.2	-2	291	251	89.2	88.4
	40%	β_1	23	230	200	90.8	89.4	18	219	193	92.2	91.0
		β_2	9	383	323	88.6	89.2	23	344	300	90.6	89.0
80	20%	β_1	-6	163	142	91.0	91.4	-4	143	129	93.0	92.2
		β_2	-11	248	241	94.0	94.0	-1	222	219	94.6	94.6
	40%	β_1	-15	199	177	90.4	90.4	-15	183	168	90.6	90.8
		β_2	-6	332	282	89.8	89.2	-10	287	259	92.8	92.0