Supplement to "Post-selection inference in generalized linear models via parametric programming"

University of South Carolina

Supplementary Material

This supplement is organized as follows: Section S1 provides additional simulation results, Section S1.1 showing comparisons of selected models $\hat{M}_\lambda^{\text{mle}}$ and $\hat{M}_\lambda$ on simulated data sets and Section S1.2 giving a comparison of the performance of our proposed method versus data splitting. Section S2 gives proofs of our main results.

S1 Additional simulation results

S1.1 Agreement of two selected models

Here we present a simulation study of the agreement between selected models $\hat{M}_\lambda^{\text{mle}}$ and $\hat{M}_\lambda$. We generate logistic, Poisson, and beta regression data sets (as described more fully in Sections 4.1 and 4.2) with $p = 20$ and $|M_0| = 3$ under the sample sizes $n \in \{100, 200, 400, 600, 800, 1000\}$. At each sample size n we obtain both $\hat{M}_\lambda^{\text{mle}}$ and $\hat{M}_\lambda$ at λ equal to $n(0.02)$, $n(0.04)$, and $n(0.10)$ for the logistic, Poisson, and beta regressions, respectively.

Table 1 gives average model sizes $|\hat{M}_\lambda^{\text{mle}}|$ and $|\hat{M}_\lambda|$ across 1000 simulated data sets as well as the proportion of these data sets for which $\hat{M}_\lambda^{\text{mle}}$ and $\hat{M}_\lambda$ agreed in terms of the inclusion of important covariates, that is for which the event $\hat{M}_\lambda^{\text{mle}} \cap M_0 = \hat{M}_\lambda \cap M_0$ occurred, where $M_0 = \{j : \beta_j \neq 0\}$.

We find agreement between $\hat{M}_\lambda^{\text{mle}}$ and $\hat{M}_\lambda$ to be quite high under these settings, suggesting that it is meaningful to select a model using L_1-penalization of a least-squares criterion in the pseudo-data $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$, that is via (2.4).

S1.2 Parametric programming versus data splitting

Here we compare our PPL method against the data splitting approach. While data splitting is a valid post-selection inference strategy, it is known to suffer

Table 1: Proportions of model selection agreement between $\hat{M}_\lambda^{\text{mle}}$ and $\hat{M}_\lambda$, and average model sizes across 1000 simulated data sets.

Model	n	100	200	400	600	800	1000
Logistic	agreement	0.96	0.99	1.00	1.00	1.00	1.00
	$\lvert\hat{M}_\lambda^{\text{mle}}\rvert$	15.2	14.0	12.1	10.8	9.6	8.8
	$\lvert\hat{M}_\lambda\rvert$	15.2	14.0	12.1	10.8	9.7	8.8
Poisson	agreement	0.99	1.00	1.00	1.00	1.00	1.00
	$\lvert\hat{M}_\lambda^{\text{mle}}\rvert$	13.1	13.1	12.0	11.0	10.1	9.5
	$\lvert\hat{M}_\lambda\rvert$	14.4	13.6	12.2	11.2	10.2	9.6
Beta	agreement	1.00	1.00	1.00	1.00	1.00	1.00
	$\lvert\hat{M}_\lambda^{\text{mle}}\rvert$	14.5	13.0	10.9	9.4	8.2	7.4
	$\lvert\hat{M}_\lambda\rvert$	14.4	12.7	10.4	8.9	7.7	6.9

from a loss of statistical power due to the reduction in sample size available for inference. We conduct this comparison in the context of Poisson regression. We specifically avoid logistic regression for this comparison because data splitting with binary outcomes often leads to unstable inference; splitting the sample size in half increases the likelihood of complete separation in the inference set, particularly when the responses are unbalanced (mostly zeros or mostly ones, thus making maximum likelihood estimation infeasible.

We generate Poisson regression data under the same settings as before, and under $\lambda = 53$ (one of the values in the grid), we compare inferences from the PPL method with those obtained by data splitting. For the data splitting method, we randomly partitioned each data set into two equal halves, each with sample size 250, performing variable selection using the first half and conducting stan-

dard Wald-type inferences in the selected model using the second half of the data. We note that to ensure a strictly comparable evaluation of conditional inference between these methods, we restricted our analysis to realizations where exactly five variables were selected.

Table 2 summarizes the results. Both methods maintain the nominal 95% coverage probability over all coefficients. However, an advantage in statistical power is observed for the PPL method in that for nonzero coefficients the PPL method produces narrower confidence intervals on average than those produced by data splitting. This reduction in width demonstrates that the PPL method effectively utilizes the full sample size, leading to more efficient inference without sacrificing validity. For the selected noise variables, the comparable interval widths and empirical coverage probabilities between the two methods suggest that the PPL method preserves the nominal Type I error rate as effectively as the data splitting procedure, while simultaneously achieving superior statistical efficiency for the nonzero coefficients.

S2 Proofs of main results

We will make use of the following result, which follows from the Lindeberg Central Limit Theorem (Athreya and Lahiri, 2006). First, for a generic $n \times 1$ vector $\mathbf{v} = (v_1, \ldots, v_n)^\top$, let $\|\mathbf{v}\|_\infty \equiv \max_{1 \leq i \leq n} |v_i|$ and $\|\mathbf{v}\|^2 \equiv \sum_{i=1}^n v_i^2$.

Table 2: Average lower and upper bounds, average width, and empirical coverage probability of 95% confidence intervals from PPL and data spitting.

Coefficient	PPL			Data Splitting		
	95% CI	Width	Coverage	95% CI	Width	Coverage
Nonzero coefficients						
β_1	[0.85, 1.16]	0.30	95.1	[0.81, 1.18]	0.38	94.1
β_2	[0.85, 1.16]	0.31	95.0	[0.81, 1.19]	0.38	95.1
β_3	[−1.16, −0.85]	0.31	96.2	[−1.19, −0.82]	0.38	95.5
Coefficients equal to zero (Average)						
$\{\beta_j, j \notin M_0\}$	[−0.19, 0.19]	0.36	96.7	[−0.18, 0.18]	0.36	94.4

Lemma 1. *For each $n \geq 1$, let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ be an $n \times 1$ vector of independent random variables with zero mean and unit variance and let $\mathbf{a} = (a_1, \ldots, a_n)^\top$ be an $n \times 1$ vector of real numbers. Then*

$$\frac{\mathbf{a}^\top \boldsymbol{\varepsilon}}{\sqrt{\mathbf{a}^\top \mathbf{a}}} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$ provided $\|\mathbf{a}\|_\infty / \|\mathbf{a}\| \to 0$ as $n \to \infty$.

We can now prove Theorem 1.

Proof of Theorem 1. To prove claim (i), we recall that

$$g_{M,j,n}(\boldsymbol{\beta}) = \frac{\mathbf{c}_{M,j}^\top (\mathbf{z}_0 - \mathbf{U}_0 \boldsymbol{\beta})}{\sqrt{\mathbf{c}_{M,j}^\top \mathbf{c}_{M,j}}}.$$

Now we use the expression $\mathbf{z}_0 = \mathbf{U}_0 \boldsymbol{\beta} + \boldsymbol{\xi}_0$ to write $\mathbf{c}_{M,j}^\top (\mathbf{z}_0 - \mathbf{U}_0 \boldsymbol{\beta}) = \mathbf{c}_{M,j}^\top \boldsymbol{\xi}_0$. Now $\mathbf{c}_{M,j}^\top \boldsymbol{\xi}_0 = \mathbf{c}_{M,j}^\top \boldsymbol{\xi}$ since $\boldsymbol{\xi}_0 = (\mathbf{I} - \mathbf{P}_0)\boldsymbol{\xi}$ and by the fact that $(\mathbf{I} - \mathbf{P}_0)$ is

idempotent. Therefore we may write

$$\frac{\mathbf{c}_{M,j}^\top(\mathbf{z}_0 - \mathbf{U}_0\boldsymbol{\beta})}{\sqrt{\mathbf{c}_{M,j}^\top \mathbf{c}_{M,j}}} = \frac{\mathbf{c}_{M,j}^\top \boldsymbol{\xi}}{\sqrt{\mathbf{c}_{M,j}^\top \mathbf{c}_{M,j}}},$$

which converges in distribution to a $\mathcal{N}(0, a(\phi))$ random variable as $n \to \infty$ by Lemma 1, due Assumption 1(i).

We prove claim (ii) as follows: Fix $\epsilon, \eta > 0$ with $\epsilon\eta \leq C\delta$, where C and δ are the constants appearing in Assumption 1(ii). Moreover, suppose $n > n_0$. To make notation more concise, set $\tilde{\mathbf{t}} \equiv (t_0, \mathbf{t}^\top)^\top$ and $\tilde{\mathbf{t}}_0 \equiv (\beta_0, \boldsymbol{\beta}^\top)^\top$. Now, for any $\tilde{\mathbf{t}}$ such that $\|\tilde{\mathbf{t}} - \tilde{\mathbf{t}}_0\| \leq \delta$, we may write

$$\begin{aligned} P(|g_n(\tilde{\mathbf{t}}) - g_n(\tilde{\mathbf{t}}_0)| > \eta) &\leq \frac{\mathbb{E}|g_n(\tilde{\mathbf{t}}) - g_n(\tilde{\mathbf{t}}_0)|}{\eta} \\ &\leq \frac{C\|\tilde{\mathbf{t}} - \tilde{\mathbf{t}}_0\|}{\eta} \end{aligned}$$

where the first inequality comes from Markov's inequality and the second from Assumption 1(ii). From here we see that for any $\delta^* < (\epsilon\eta)/C \leq \delta$ we have

$$\sup_{\|\tilde{\mathbf{t}} - \tilde{\mathbf{t}}_0\| \leq \delta^*} P(|g_n(\tilde{\mathbf{t}}) - g_n(\tilde{\mathbf{t}}_0)| > \eta) \leq \frac{C\delta^*}{\eta} < \epsilon.$$

Now, setting $\hat{\mathbf{t}}_n \equiv (\hat{\beta}_0, \hat{\boldsymbol{\beta}}^\top)^\top$ we have

$$\begin{aligned} P(|g_n(\hat{\mathbf{t}}_n) - g_n(\tilde{\mathbf{t}}_0)| > \eta) &= P(|g_n(\hat{\mathbf{t}}_n) - g_n(\tilde{\mathbf{t}}_0)| > \eta \cap \|\hat{\mathbf{t}}_n - \mathbf{t}_0\| \leq \delta^*) \\ &\quad + P(|g_n(\hat{\mathbf{t}}_n) - g_n(\tilde{\mathbf{t}}_0)| > \eta \cap \|\hat{\mathbf{t}}_n - \mathbf{t}_0\| > \delta^*) \\ &< \epsilon + P(\|\hat{\mathbf{t}}_n - \mathbf{t}_0\| > \delta^*). \end{aligned}$$

Since the above holds for all $n > n_0$ and since $P(\|\hat{\mathbf{t}}_n - \mathbf{t}_0\| > \delta^*) \to 0$ we have

$$\limsup_{n\to\infty} P(|g_n(\hat{\mathbf{t}}_n) - g_n(\tilde{\mathbf{t}}_0)| > \eta) < \epsilon.$$

This proves the claim. □

Proof of Corollary 1. Recall that $\hat{\mathbf{c}}_{M,j}^\top \hat{\mathbf{z}}_0 = (\hat{\boldsymbol{\beta}}_M)_j$ and $\mathbf{c}_{M,j}^\top \mathbf{z}_0 = (\tilde{\boldsymbol{\beta}}_M)_j$. Now, if $M \supset \{j : \beta_j \neq 0\}$ then $\hat{\mathbf{c}}_{M,j}^\top \hat{\mathbf{U}}_0 \boldsymbol{\beta} = (\boldsymbol{\beta}_M)_j$ by the fact that $\hat{\mathbf{U}}_0 \boldsymbol{\beta} = \hat{\mathbf{U}}_{0,M} \boldsymbol{\beta}_M$ and by the definition of $\hat{\mathbf{c}}_{M,j}^\top$, and, likewise, $\mathbf{c}_{M,j}^\top \mathbf{U}_0 \boldsymbol{\beta} = (\boldsymbol{\beta}_M)_j$ by the fact that $\mathbf{U}_0 \boldsymbol{\beta} = \mathbf{U}_{0,M} \boldsymbol{\beta}_M$ and by the definition of $\mathbf{c}_{M,j}^\top$. Theorem 1(i) therefore gives

$$\frac{\mathbf{c}_{M,j}^\top (\mathbf{z}_0 - \mathbf{U}_0 \boldsymbol{\beta})}{\sqrt{\mathbf{c}_{M,j}^\top \mathbf{c}_{M,j}}} = \frac{(\tilde{\boldsymbol{\beta}}_M)_j - (\boldsymbol{\beta}_M)_j}{\sqrt{\mathbf{e}_j^\top (\mathbf{U}_{0,M}^\top \mathbf{U}_{0,M})^{-1} \mathbf{e}_j}} \xrightarrow{d} \mathcal{N}(0, a(\phi))$$

as $n \to \infty$. We also have

$$\frac{\hat{\mathbf{c}}_{M,j}^\top (\hat{\mathbf{z}}_0 - \hat{\mathbf{U}}_0 \boldsymbol{\beta})}{\sqrt{\hat{\mathbf{c}}_{M,j}^\top \hat{\mathbf{c}}_{M,j}}} = \frac{(\hat{\boldsymbol{\beta}}_M)_j - (\boldsymbol{\beta}_M)_j}{\sqrt{\mathbf{e}_j^\top (\hat{\mathbf{U}}_{0,M}^\top \hat{\mathbf{U}}_{0,M})^{-1} \mathbf{e}_j}}$$

so that Theorem 1(ii) gives

$$\left| \frac{(\hat{\boldsymbol{\beta}}_M)_j - (\boldsymbol{\beta}_M)_j}{\sqrt{\mathbf{e}_j^\top (\hat{\mathbf{U}}_{0,M}^\top \hat{\mathbf{U}}_{0,M})^{-1} \mathbf{e}_j}} - \frac{(\tilde{\boldsymbol{\beta}}_M)_j - (\boldsymbol{\beta}_M)_j}{\sqrt{\mathbf{e}_j^\top (\mathbf{U}_{0,M}^\top \mathbf{U}_{0,M})^{-1} \mathbf{e}_j}} \right| \xrightarrow{p} 0$$

as $n \to \infty$. The claim follows from Slutzky's theorem. □

Bibliography

Athreya, K. B. and S. N. Lahiri (2006). *Measure theory and probability theory*. Springer.