Supporting Information for

"Post-Stratification Fusion Learning in Longitudinal Data Analysis"

by Lu Tang and Peter X.-K. Song

Web Appendix A. Model Selection

Model selection to determine the number of clusters is done by the tunning of λ, whose value can be selected based on standard model selection criteria, such as AIC, BIC or GCV. We propose to use the extended Bayesian information criterion (EBIC) (Chen and Chen, 2012) of the form:

$$\text{EBIC}(\lambda) = \widetilde{\boldsymbol{S}}(\widehat{\boldsymbol{\theta}}_\lambda)^T \widetilde{\boldsymbol{G}}(\widehat{\boldsymbol{\theta}}_\lambda)^{-1} \widetilde{\boldsymbol{S}}(\widehat{\boldsymbol{\theta}}_\lambda) + \log(N) \sum_{j=1}^{p} \text{df}(\widehat{\boldsymbol{\theta}}_{\lambda j\cdot}) + 2\log \sum_{j=1}^{p} \binom{K}{\text{df}(\widehat{\boldsymbol{\theta}}_{\lambda j\cdot})},$$

where

$$\widetilde{\boldsymbol{G}}(\widehat{\boldsymbol{\theta}}_\lambda) = \text{block-diag}\{\boldsymbol{S}_1(\boldsymbol{C}_1\boldsymbol{D}^{-1}\widehat{\boldsymbol{\theta}}_\lambda)\boldsymbol{S}_1(\boldsymbol{C}_1\boldsymbol{D}^{-1}\widehat{\boldsymbol{\theta}}_\lambda)^T, \ldots, \boldsymbol{S}_K(\boldsymbol{C}_K\boldsymbol{D}^{-1}\widehat{\boldsymbol{\theta}}_\lambda)\boldsymbol{S}_K(\boldsymbol{C}_K\boldsymbol{D}^{-1}\widehat{\boldsymbol{\theta}}_\lambda)^T\},$$

and $\text{df}(\widehat{\boldsymbol{\theta}}_j)$ denotes the number of unique nonzero values in $\widehat{\boldsymbol{\theta}}_j$. Starting from $\lambda = 0$, we fit a path of solutions $\widehat{\boldsymbol{\theta}}_\lambda$ for a sequence of $\lambda \geqslant 0$. To accelerate computation in calculating the solution paths for a sequence of λ values, for the next value λ, we employ the warm-start technique and use $\widehat{\boldsymbol{\theta}}$ from the current value of λ as the initial value of the iterative algorithm. The initial value of the search algorithm is $\widehat{\boldsymbol{\theta}}_{\lambda=0}$

Web Appendix B. Regularity Conditions

For the results in Section 4.3, we require the following regularity conditions, which are similar to the conditions required in Theorem 1 of Wang et al. (2012).

(C1) $\boldsymbol{X}_{kit}$ are uniformly bounded for $k = 1, \ldots, K$, $i = 1, \ldots, n_k$ and $t = 1, \ldots, T$. The eigenvalues of $n_k^{-1} \sum_{i=1}^{n_k} \boldsymbol{X}_{ki}\boldsymbol{X}_{ki}^T$ are uniformly bounded away from zero and $+\infty$ for $k = 1, \ldots, K$.

(C2) There exists a positive-definite matrix $\overline{\boldsymbol{R}}_k$ such that the estimated working correlation matrix $\widehat{\boldsymbol{R}}_k$ satisfies $||\widehat{\boldsymbol{R}}_k^{-1} - \overline{\boldsymbol{R}}_k^{-1}||_2 = O_p(\sqrt{s/n_{\sup}K})$, $k = 1, \ldots, K$.

(C3) Let $\boldsymbol{\epsilon}_{ki}(\boldsymbol{\theta}) = \boldsymbol{A}_{ki}^{-1/2}(\boldsymbol{C}_k\boldsymbol{D}^{-1}\boldsymbol{\theta})(\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_{ki}(\boldsymbol{C}_k\boldsymbol{D}^{-1}\boldsymbol{\theta}))$. Assume $E(||\boldsymbol{\epsilon}_{ki}(\boldsymbol{\theta})||^{2+\delta})$ is bounded by a finite positive constant, for all k, i and some $\delta > 0$; and there exist positive constants M such that $E[\exp(M|\epsilon_{kit}(\boldsymbol{\theta}_*)|)|\boldsymbol{X}_{ki}]$ is bounded by a finite positive constant, uniformly in $k = 1, \ldots, K$, $i = 1, \ldots, n_k$ and $t = 1, \ldots, T$.

(C4) For some $\Delta > 0$, $B = \{\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_*|| \leqslant \Delta\sqrt{p/n_{\sup}}\}$. On B, for $k = 1, \ldots, K$, $i = 1, \ldots, n_k$ and $t = 1, \ldots, T$, $\nabla_{\boldsymbol{\theta}}\mu(\boldsymbol{X}_{kit}^T\boldsymbol{C}_k\boldsymbol{D}^{-1}\boldsymbol{\theta})$ are uniformly bounded away from 0 and ∞; $\nabla^2_{\boldsymbol{\theta}}\mu(\boldsymbol{X}_{kit}^T\boldsymbol{C}_k\boldsymbol{D}^{-1}\boldsymbol{\theta})$ and $\nabla^3_{\boldsymbol{\theta}}\mu(\boldsymbol{X}_{kit}^T\boldsymbol{C}_k\boldsymbol{D}^{-1}\boldsymbol{\theta})$ are uniformly bounded by a finite positive constant M.

(C5) Assuming $\min_{j\in\mathcal{A}}|\theta_{*j}|/\lambda \to \infty$ as $N \to \infty$ and $s^3N^{-1} = o(1)$, $\lambda \to 0$, $s^2(\log N)^4 = o(N\lambda^2)$, $\log(pK) = o(N\lambda^2/(\log N)^2)$, $pKs^4(\log N)^6 = o(N^2\lambda^2)$ and $pKs^3(\log N)^8 = o(N^2\lambda^4)$.

Condition (C1) is standard for regularized regressions, and is require here uniformly across strata. Condition (C2) specifies the limits of working correlation matrices exist in individual strata. In particular, when $\boldsymbol{\tau}$ and correlation structure are the same across strata, this assumption can be relaxed to $||\widehat{\boldsymbol{R}}^{-1} - \overline{\boldsymbol{R}}^{-1}||_2 = O_p(\sqrt{s/N})$. This is satisfied when the method of moments is used to estimate $\widehat{\boldsymbol{\tau}}$, and $\overline{\boldsymbol{R}} = \boldsymbol{R}_0$. Conditions (C3)-(C5) are standard for regularized GEE, and generally satisfied for Gaussian, binary and Poisson distributions.

Web Appendix C. Proofs for Section 4.3

Web Appendix C.1 *Proof of Theorem 1*

By treating (5) as a penalized GEE with respect to $\boldsymbol{\theta}$, where the true value $\boldsymbol{\theta}_*$ is sparse with s nonzero elements, we may derive similar asymptotic properties by reparametrization. We follow the proof of Theorem 1 of Wang et al. (2012). By conditions (C1)-(C5), the first part, se-

lection consistency can be easily established thus omitted. Because $\boldsymbol{D}^*$ is invertible, the oracle asymptotic normality for $\widehat{\boldsymbol{\theta}}$ states that $\sqrt{N}(\widetilde{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{*\mathcal{A}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \widetilde{\boldsymbol{H}}^*_{\mathcal{A}}(\boldsymbol{\theta}_*)^{-1}\boldsymbol{M}^*_{\mathcal{A}}(\boldsymbol{\theta}_*)\{\widetilde{\boldsymbol{H}}^*_{\mathcal{A}}(\boldsymbol{\theta}_*)^T\}^{-1})$. Since $\widetilde{\boldsymbol{S}}(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{D}_*^{-1}\boldsymbol{\theta}) = (\boldsymbol{S}_1^T(\boldsymbol{C}_1\boldsymbol{D}_*^{-1}\boldsymbol{\theta}), \ldots, \boldsymbol{S}_K^T(\boldsymbol{C}_K\boldsymbol{D}_*^{-1}\boldsymbol{\theta}))^T$, by reparameterization we have,

$$
\begin{aligned}
\widetilde{\boldsymbol{H}}^*(\boldsymbol{\theta}_*) &= -\frac{\partial \widetilde{\boldsymbol{S}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&= -\left(\frac{\partial \boldsymbol{S}_1^T(\boldsymbol{C}_1\boldsymbol{D}_*^{-1}\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \ldots, \frac{\partial \boldsymbol{S}_K^T(\boldsymbol{C}_K\boldsymbol{D}_*^{-1}\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \\
&= -\left(\left(\frac{\partial \boldsymbol{S}_1(\boldsymbol{C}_1\boldsymbol{D}_*^{-1}\boldsymbol{\theta})}{\partial \boldsymbol{C}_1\boldsymbol{D}_*^{-1}\boldsymbol{\theta}}\frac{\partial \boldsymbol{C}_1\boldsymbol{D}_*^{-1}\boldsymbol{\theta}}{\partial \boldsymbol{\theta}}\right)^T, \ldots, \left(\frac{\partial \boldsymbol{S}_K(\boldsymbol{C}_K\boldsymbol{D}_*^{-1}\boldsymbol{\theta})}{\partial \boldsymbol{C}_K\boldsymbol{D}_*^{-1}\boldsymbol{\theta}}\frac{\partial \boldsymbol{C}_K\boldsymbol{D}_*^{-1}\boldsymbol{\theta}}{\partial \boldsymbol{\theta}}\right)^T\right)^T \\
&= \left((\boldsymbol{H}_1(\boldsymbol{C}_1\boldsymbol{D}_*^{-1}\boldsymbol{\theta})\boldsymbol{C}_1\boldsymbol{D}_*^{-1})^T, \ldots, (\boldsymbol{H}_K(\boldsymbol{C}_K\boldsymbol{D}_*^{-1}\boldsymbol{\theta})\boldsymbol{C}_K\boldsymbol{D}_*^{-1})^T\right)^T \\
&= \text{block-diag}\{\boldsymbol{H}_1(\boldsymbol{C}_1\boldsymbol{D}_*^{-1}\boldsymbol{\theta}_*), \ldots, \boldsymbol{H}_K(\boldsymbol{C}_K\boldsymbol{D}_*^{-1}\boldsymbol{\theta}_*)\}\boldsymbol{C}\boldsymbol{D}_*^{-1} \\
&= \text{block-diag}\{\boldsymbol{H}_1(\boldsymbol{\beta}_{*\cdot 1}), \ldots, \boldsymbol{H}_K(\boldsymbol{\beta}_{*\cdot K})\}\boldsymbol{C}\boldsymbol{D}_*^{-1},
\end{aligned}
$$

where $\boldsymbol{H}_k(\boldsymbol{\beta}_{*\cdot k}) = \sum_{i=1}^{n_k} \boldsymbol{X}_{ki}^T\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_{*\cdot k})\overline{\boldsymbol{R}}^{-1}\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_{*\cdot k})\boldsymbol{X}_{ki}$ is the sensitivity matrix of $\boldsymbol{S}_k(\boldsymbol{\beta}_{*\cdot k})$, $k = 1, \ldots, K$, $\boldsymbol{C} = (\boldsymbol{C}_1^T, \ldots, \boldsymbol{C}_K^T)^T$, and

$$
\begin{aligned}
\boldsymbol{M}^*(\boldsymbol{\theta}_*) &= \text{var}(\widetilde{\boldsymbol{S}}(\boldsymbol{\theta}_*)) \\
&= \text{block-diag}\{\text{var}(\boldsymbol{S}_1(\boldsymbol{C}_1\boldsymbol{D}_*^{-1}\boldsymbol{\theta}_*)), \ldots, \text{var}(\boldsymbol{S}_K(\boldsymbol{C}_K\boldsymbol{D}_*^{-1}\boldsymbol{\theta}_*))\} \\
&= \text{block-diag}\{\boldsymbol{M}_1(\boldsymbol{\beta}_{*\cdot 1}), \ldots, \boldsymbol{M}_K(\boldsymbol{\beta}_{*\cdot K})\},
\end{aligned}
$$

where $\boldsymbol{M}_k(\boldsymbol{\beta}_{*\cdot k}) = \sum_{i=1}^{n_k} \boldsymbol{X}_{ki}^T\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_{*\cdot k})\overline{\boldsymbol{R}}^{-1}\boldsymbol{R}_*\overline{\boldsymbol{R}}^{-1}\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_{*\cdot k})\boldsymbol{X}_{ki}$ is the variability matrix of $\boldsymbol{S}_k(\boldsymbol{\beta}_{*\cdot k})$, $k = 1, \ldots, K$. Therefore, the second part of asymptotic results follows.

Web Appendix C.2 *Proof of Theorem 2*

First, we present the following lemma regarding the estimation of parameter ordering to assist the proof of Theorem 2.

LEMMA 1: *If $\widehat{\boldsymbol{D}}$ is estimated based on the unpenalized estimator $\widehat{\boldsymbol{\beta}}$ from* (1) *and $K = O(N^{1/4-\xi})$ where $\xi \in (0, 1/4]$, then for $\forall \epsilon > 0$, $\lim_N P(||\widehat{\boldsymbol{D}} - \boldsymbol{D}_*|| \geqslant \epsilon) = 0$ for some $\boldsymbol{D}_*$ such that $\boldsymbol{\theta}_* = \boldsymbol{D}_*\boldsymbol{\beta}_*$.*

Proof. First, we define $\boldsymbol{D}_k$ the submatrix of $\boldsymbol{D}$ that only contains the rows and columns relevant to parameters in $\boldsymbol{\beta}_{\cdot k}, k = 1, \ldots, K$. The dimension of $\boldsymbol{D}_k$ is of same order as p since it only involves adjacent contrasts. Similar to Tang and Song (2016), we examine the probability of incorrectly estimating the order. Note that existence and consistency of solution to the unpenalized GEE (1) as p diverges has been established in Wang (2011, Theorem 3.6). Combining results from both, the probability of getting the wrong ordering for coefficient estimates from different clusters tend to zero at rate $O_p(\sqrt{p/N})$, under the condition that $N^{-1}p^2 = o(1)$. Since $N^{-1}K^4 = o(1)$, we have that $K = o(N^{1/2}/p^{1/2})$. Thus, jointly considering all K diagonal blocks of $\widehat{\boldsymbol{D}}$, we have

$$\begin{aligned} P(||\widehat{\boldsymbol{D}} - \boldsymbol{D}_*|| \geqslant \epsilon) &= P(\sum_{k=1}^{K} ||\widehat{\boldsymbol{D}}_k - \boldsymbol{D}_{*k}||^2 \geqslant \epsilon^2) \\ &\leqslant \sum_{k=1}^{K} P(||\widehat{\boldsymbol{D}}_k - \boldsymbol{D}_{*k}||^2 \geqslant \epsilon^2/K) \\ &\leqslant \sum_{k=1}^{K} P(||\widehat{\boldsymbol{D}}_k - \boldsymbol{D}_{*k}||^2 \geqslant \epsilon^2 p^{1/2}/N^{1/2}) \\ &= O_p(p^{1/2}N^{-1/2}K) = o(1). \end{aligned}$$

Note that $\boldsymbol{D}_*$ is not unique since some parameters are common across strata within $\boldsymbol{\beta}_*$ thus ties may occur. Here, we resolve the ties by ordering the parameters based on first occurrence. This will not affect our results since $\boldsymbol{\beta}_*$ are uniquely defined given $\boldsymbol{\theta}_*$ and $\boldsymbol{D}_*$ pairs. Below we present the proof to Theorem 2 following a result in Tang and Song (2016).

By Corollary 1, $P(\widehat{\boldsymbol{\theta}}_{\mathcal{A}^c} = \mathbf{0}|\widehat{\boldsymbol{D}} = \boldsymbol{D}_*) \to 1$. Combine with Lemma 1, we have

$$\begin{aligned} P(\widehat{\boldsymbol{\theta}}_{\mathcal{A}^c} = \mathbf{0}) = P(\widehat{\boldsymbol{\theta}}_{\mathcal{A}^c} = \mathbf{0}|\widehat{\boldsymbol{D}} = \boldsymbol{D}_*)P(\widehat{\boldsymbol{D}} = \boldsymbol{D}_*) \\ + P(\widehat{\boldsymbol{\theta}}_{\mathcal{A}^c} = \mathbf{0}|\widehat{\boldsymbol{D}} \neq \boldsymbol{D}_*)P(\widehat{\boldsymbol{D}} \neq \boldsymbol{D}_*) \to 1 \end{aligned}$$

as $N \to \infty$. Thus, we have selection consistency.

Similarly, the estimator $\widehat{\boldsymbol{\theta}}_{\mathcal{A}}$ can be written as

$$\widehat{\boldsymbol{\theta}}_{\mathcal{A}} = \widetilde{\boldsymbol{\theta}}_{\mathcal{A}} 1\{\widehat{\boldsymbol{D}} = \boldsymbol{D}_*\} + \widehat{\boldsymbol{\theta}}_{\mathcal{A}} 1\{\widehat{\boldsymbol{D}} \neq \boldsymbol{D}_*\}.$$

Therefore,

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{*\mathcal{A}}) = \sqrt{N}(\widetilde{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{*\mathcal{A}})1\{\widehat{\boldsymbol{D}} = \boldsymbol{D}_*\} + \sqrt{N}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{*\mathcal{A}})1\{\widehat{\boldsymbol{D}} \neq \boldsymbol{D}_*\}.$$

Based on results given in Proposition 1, we can show that $\sqrt{N}(\widetilde{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{*\mathcal{A}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{A}})$. From Lemma 1 as $K = O(N^{1/2-\xi})$ and $n_{\text{inf}} \to \infty$, we have $P(\widehat{\boldsymbol{D}} = \boldsymbol{D}_*) \to 1$ and $P(\widehat{\boldsymbol{D}} \neq \boldsymbol{D}_*) \to 0$, as $N \to \infty$. Hence, the asymptotic normality result follows.

Web Appendix C.3 *Proof of Corollary 2*

For all $k, k' \in \{k, k' | g_j(k) = g_j(k')\}$, we show $P(\widehat{\beta}_{jk} = \widehat{\beta}_{jk'}) \to 1$.

$$\begin{aligned} P(|\widehat{\beta}_{jk} - \widehat{\beta}_{jk'}| > 0) &\leqslant P(|\widehat{\beta}_{jk} - \widehat{\beta}_{jk'} - (\beta_{*jk} - \beta_{*jk'})| > 0) \\ &\leqslant P(|\widehat{\beta}_{jk} - \beta_{*jk}| > 0) + P(|\widehat{\beta}_{jk'} - \beta_{*jk'}| > 0) \to 0 \end{aligned}$$

Since $\beta_{*jk} = \beta_{*jk'} = 0$ or $\beta_{*jk} = \beta_{*jk'} \neq 0$, thus by Slutsky's Theorem we prove the result.

Web Appendix D. Additional Simulation Results

Web Appendix D.1 *The Linear Model*

For the linear model with exchangeable working correlation matrix, boxplots of estimate number of coefficient clusters are shown in Web Figure 1.

[Figure 1 about here.]

Web Appendix D.2 *The Logistic Model*

We simulate correlated data with binary outcomes from the following marginal mean model:

$$\text{logit}\{E(Y_{kij})\} = \beta_{k1}X_{kij1} + \beta_{k2}X_{kij2} + \beta_{k3}X_{kij3} + \beta_{k4}X_{ki4} + \beta_{k5}X_{ki5}$$

for $k = 1, \ldots, K$, $i = 1, \ldots, n$, and $j \in \mathcal{L}_k$. The covariates are simulated in the same way as in the linear model, and outcomes with the exchangeable correlation with $\tau = 0.6$ are simulated by R package `SimCorMultRes`. We use the true coefficient values $\boldsymbol{\beta}_1 = (1, \ldots, 1)^T, \boldsymbol{\beta}_2 =$

$(0.6, \ldots, 0.6)^T, \boldsymbol{\beta}_3 = (0.6, \ldots, 0.6)^T, \boldsymbol{\beta}_4 = (0, \ldots, 0)^T, \boldsymbol{\beta}_5 = (0, \ldots, 0)^T$, and create heterogeneous parameter clusters in $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_4$ using the same procedure as in the linear model, with $\delta = 0, 0.2$ and 0.6. The comparison between choices of working correlation matrices in terms of efficiency evaluation is similar to the linear model, thus only results for the logistic model with exchangeable working correlation are presented. For the logistic model with exchangeable working correlation matrix, mean squared errors of coefficient estimates are shown in Web Table 1; histograms of estimated coefficient values are shown in Web Figure 2; boxplots of estimate number of coefficient clusters are shown in Web Figure 3. All results are based on 100 replications.

[Table 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

Web Appendix E. Summary Statistics of IHS Covariates

Summary of variables in their original scales, stratified by missing-data patterns, is shown in Web Table 2.

[Table 2 about here.]

Web Appendix F. Comparison of LASSO, SCAD and MCP

We provide an example comparing LASSO, SCAD and MCP penalties in fusion learning. Compared to the convex LASSO penalty, SCAD and MCP penalties are nonconvex functions that are designed to protect large coefficients from being unduly shrunk through a very slowly growing function for large value. In the context of fusion learning, different shrinkage mechanisms can be visualized in Web Figure 4 through their respective solution paths. Clearly, nonconvex penalties on the coefficient contrasts shows a clearer separation between

parameter subgroups than LASSO. Both SCAD and MCP taper off quickly for large β values, leading to dendrogram-like solution paths for clear separation of clusters.

[Figure 4 about here.]

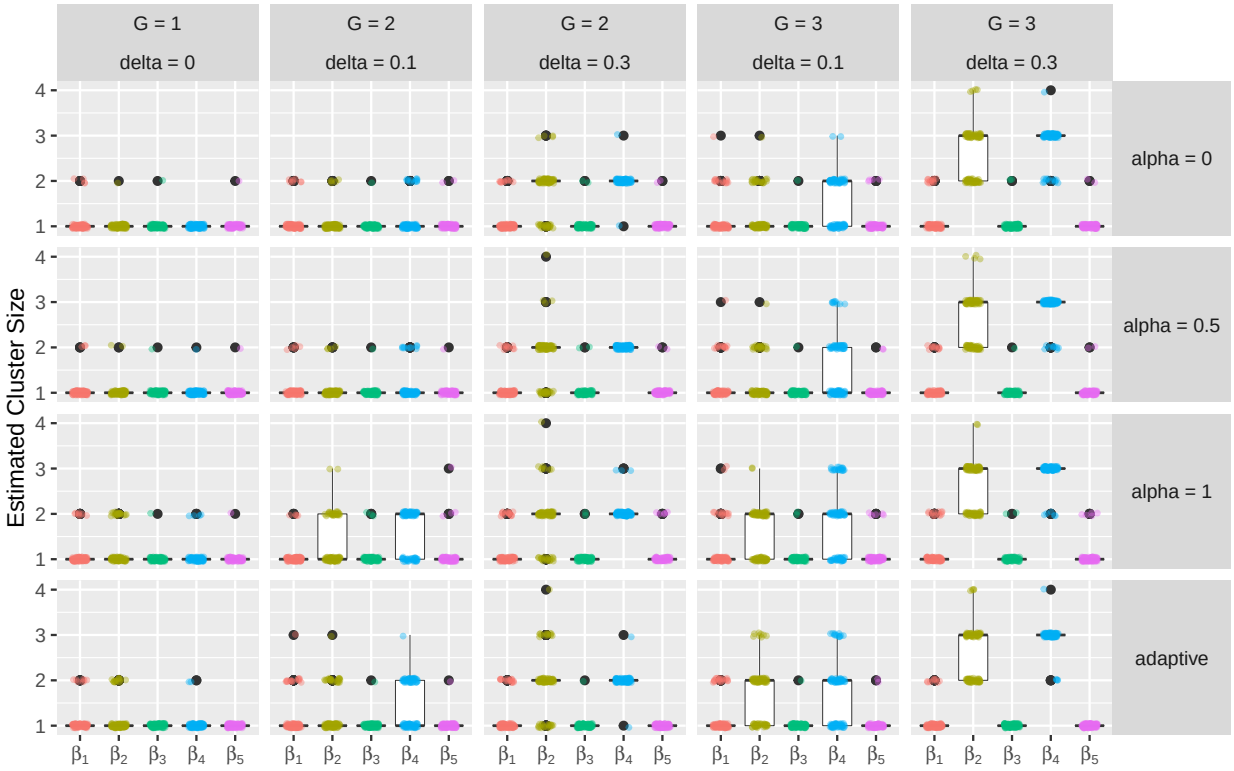
References

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-p sparse GLM. *Statistica Sinica* **22,** 555–574.

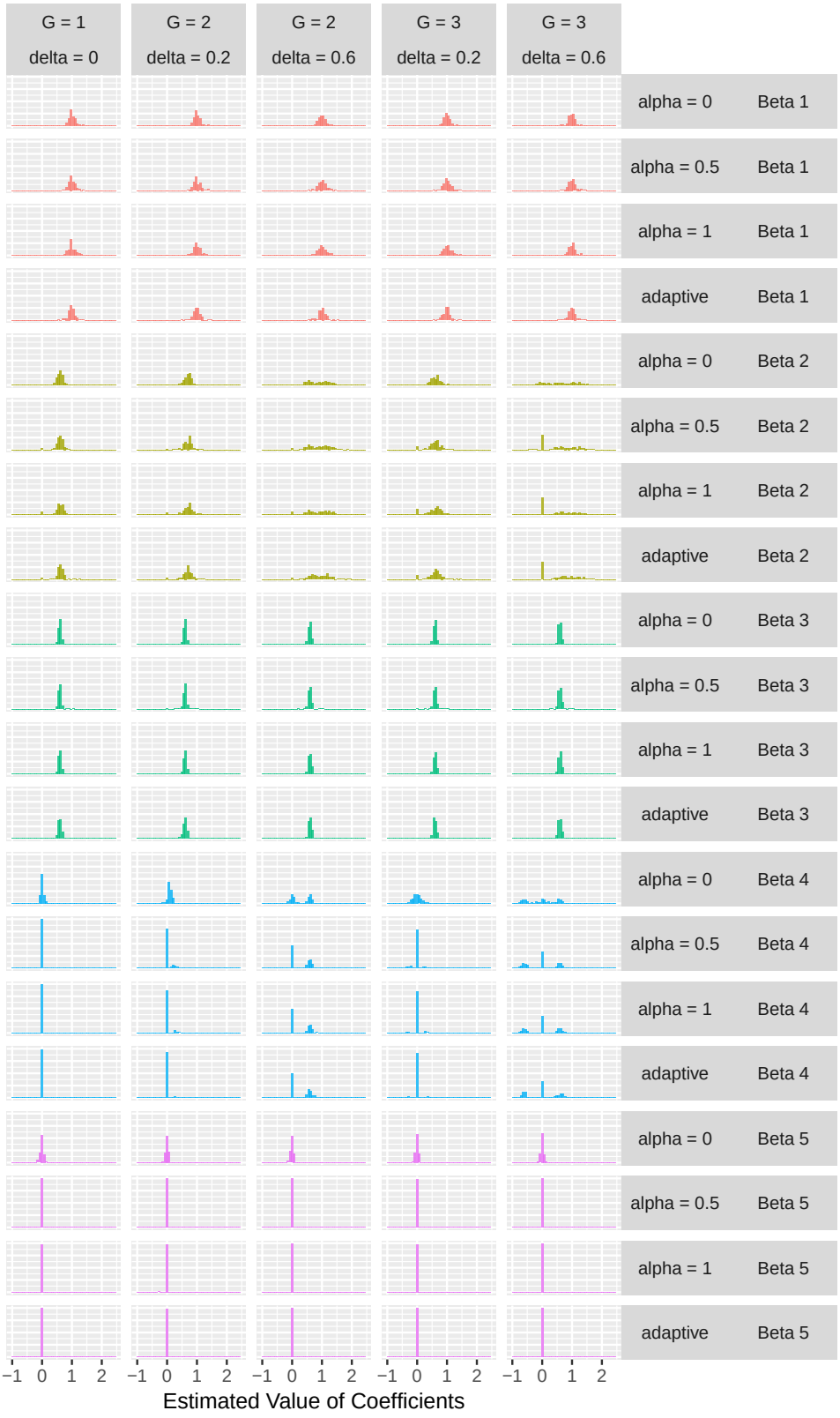
Tang, L. and Song, P. X. (2016). Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration. *The Journal of Machine Learning Research* **17,** 3915–3937.

Wang, L. (2011). Gee analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics* **39,** 389–417.

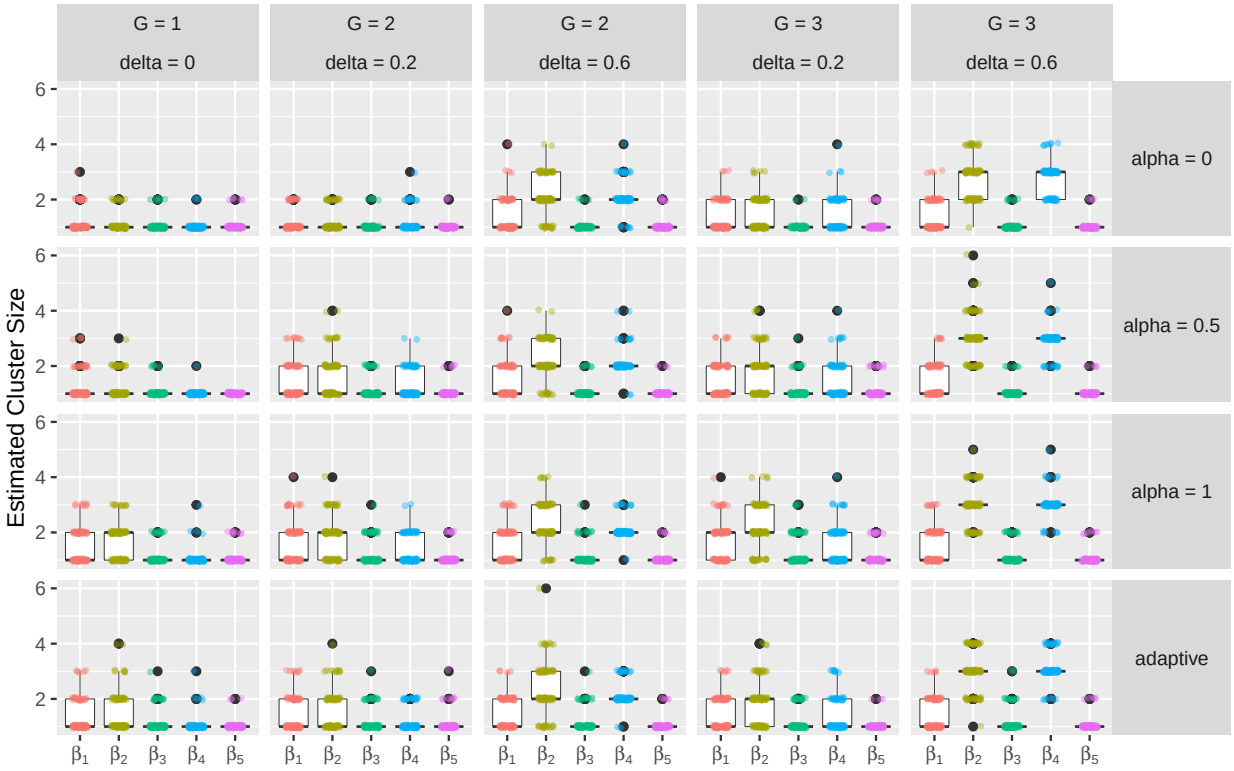
Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68,** 353–360.



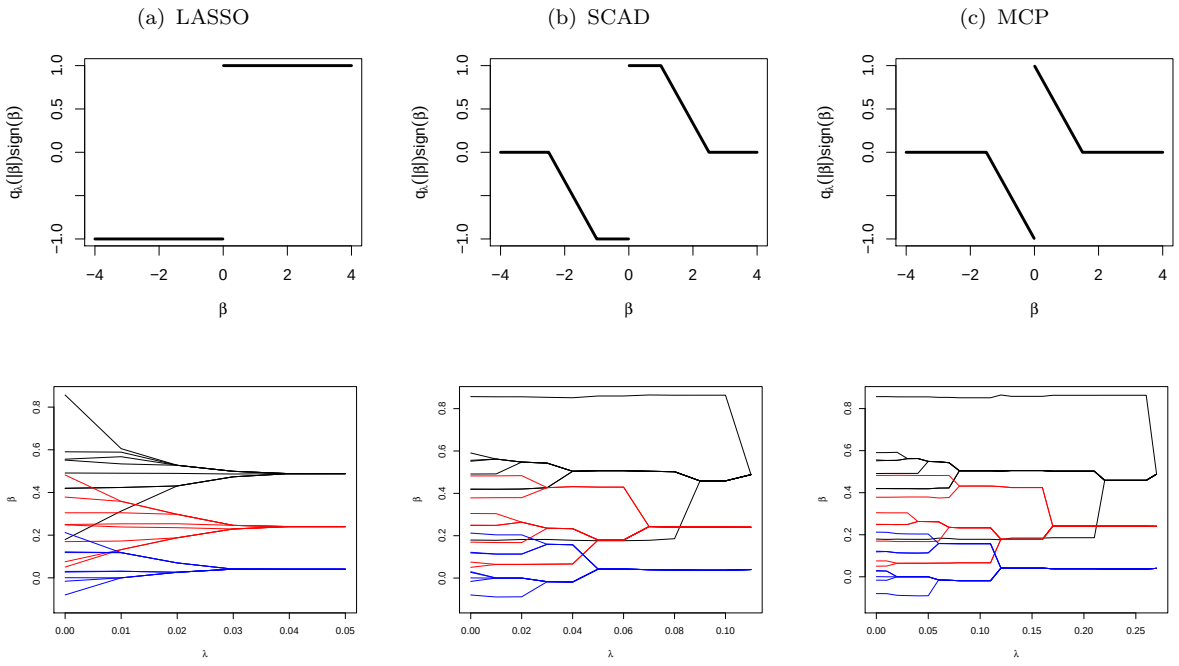
Web Figure 1. Boxplots of estimated cluster sizes in the linear model with exchangeable working correlation matrix. Bolded horizontal bars represent the medians.



Web Figure 2. Histograms of coefficient estimates in the logistic model with exchangeable working correlation matrix.



Web Figure 3. Boxplots of estimated cluster sizes in the logistic model with exchangeable working correlation matrix. Bolded horizontal bars represent the medians.



Web Figure 4. Plots of penalty functions (top) and solution paths (bottom) in an example of GEE fusion learning with $p = 3$ and $K = 8$, for LASSO, SCAD and MCP, respectively. Each color in the solution path plots represents stratum-specific estimates from one covariate.

Web Table 1

Mean squared error ($\times 100$) *of GEE coefficient estimates in the logistic model with exchangeable working correlation matrix (i.e., working correlation matrix correctly specified).*

G	δ	Method	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_3$	$\boldsymbol{\beta}_4$	$\boldsymbol{\beta}_5$
			$\boldsymbol{\beta}_1 = \mathbf{1}$	$\boldsymbol{\beta}_2 = \mathbf{0.6}$	$\boldsymbol{\beta}_3 = \mathbf{0.6}$	$\boldsymbol{\beta}_4 = \mathbf{0}$	$\boldsymbol{\beta}_5 = \mathbf{0}$
1	0.0	Homogeneous	0.837	0.639	0.179	0.198	0.212
1	0.0	Stratified	6.457	5.366	2.257	1.876	1.817
1	0.0	Two-stage	1.299	1.158	0.446	0.375	0.393
1	0.0	Proposed ($\alpha = 0$)	1.797	1.135	0.329	0.234	0.301
1	0.0	Proposed ($\alpha = 0.5$)	2.191	1.901	0.302	0.029	0.000
1	0.0	Proposed ($\alpha = 1$)	3.431	4.049	0.625	0.074	0.139
1	0.0	Proposed (adaptive)	1.953	2.536	0.719	0.124	0.062
			$\boldsymbol{\beta}_1 = \mathbf{1}$	$\boldsymbol{\beta}_2 \in \{0.6, 0.6+\delta\}^8$	$\boldsymbol{\beta}_3 = \mathbf{0.6}$	$\boldsymbol{\beta}_4 \in \{0, \delta\}^8$	$\boldsymbol{\beta}_5 = \mathbf{0}$
2	0.2	Homogeneous	0.789	1.558	0.197	1.14	0.207
2	0.2	Stratified	6.355	5.564	2.359	1.896	1.853
2	0.2	Two-stage	2.513	3.067	1.086	1.47	0.7
2	0.2	Proposed ($\alpha = 0$)	1.523	2.227	0.326	1.316	0.245
2	0.2	Proposed ($\alpha = 0.5$)	2.548	3.304	0.524	1.985	0.064
2	0.2	Proposed ($\alpha = 1$)	3.213	4.794	0.786	1.993	0.131
2	0.2	Proposed (adaptive)	2.400	3.862	0.853	2.002	0.166
2	0.6	Homogeneous	0.859	8.87	0.242	8.404	0.214
2	0.6	Stratified	6.605	6.827	2.509	2.316	1.863
2	0.6	Two-stage	6.599	6.916	2.503	2.403	1.86
2	0.6	Proposed ($\alpha = 0$)	2.637	6.426	0.318	1.558	0.309
2	0.6	Proposed ($\alpha = 0.5$)	3.174	6.962	0.367	1.189	0.052
2	0.6	Proposed ($\alpha = 1$)	3.973	7.947	0.677	1.133	0.095
2	0.6	Proposed (adaptive)	3.043	7.554	0.703	1.293	0.084
			$\boldsymbol{\beta}_1 = \mathbf{1}$	$\boldsymbol{\beta}_2 \in \{0.6, 0.6 \pm \delta\}^8$	$\boldsymbol{\beta}_3 = \mathbf{0.6}$	$\boldsymbol{\beta}_4 \in \{0, \pm\delta\}^8$	$\boldsymbol{\beta}_5 = \mathbf{0}$
3	0.2	Homogeneous	0.771	3.054	0.188	2.665	0.182
3	0.2	Stratified	6.527	5.455	2.257	1.852	1.81
3	0.2	Two-stage	4.663	4.79	1.645	2.179	1.344
3	0.2	Proposed ($\alpha = 0$)	2.314	3.755	0.242	2.377	0.285
3	0.2	Proposed ($\alpha = 0.5$)	2.838	4.623	0.444	2.533	0.172
3	0.2	Proposed ($\alpha = 1$)	3.939	5.587	0.898	2.628	0.189
3	0.2	Proposed (adaptive)	2.373	5.140	0.644	2.719	0.081
3	0.6	Homogeneous	1.066	22.673	0.311	22.283	0.198
3	0.6	Stratified	6.392	6.546	2.419	2.263	1.806
3	0.6	Two-stage	6.392	6.546	2.419	2.263	1.806
3	0.6	Proposed ($\alpha = 0$)	2.836	7.685	0.37	2.986	0.285
3	0.6	Proposed ($\alpha = 0.5$)	3.119	7.322	0.334	2.006	0.066
3	0.6	Proposed ($\alpha = 1$)	3.445	7.119	0.671	1.755	0.117
3	0.6	Proposed (adaptive)	2.466	6.443	0.765	1.732	0.118

Web Table 2

Summary statistics for suicidal ideation data. Means (and standard deviations) are reported for continuous covariates and percentages are reported for binary covariates.

Pattern	0011	0110	0111	1011	1100	1101	1110	1111
Sample size	41	68	120	128	150	141	249	1570
Baseline								
Age	27.9(3.5)	27.2(2.1)	27.6(3)	27.8(3.5)	27.6(3)	27.7(3)	27.3(2.3)	27.4(2.5)
Female (%)	56.1	47.1	38.3	54.7	54.0	49.6	47.0	51.3
Baseline SI (%)	4.9	4.4	4.2	4.7	3.3	3.5	3.6	2.7
Baseline PHQ score	2.6(3.1)	1.8(2.3)	2.8(3.5)	2.7(3.1)	2.5(2.8)	3.1(3.4)	2.6(3)	2.4(2.7)
Time Dependent								
PHQ score	5.6(4.8)	6.1(4.3)	5.1(4.1)	6.4(4.9)	6.4(4.6)	6.4(4.7)	5.6(4.5)	5.3(4.2)
GAD score	4.1(4.4)	4.8(4)	3.9(4.2)	5(4.5)	5.2(4.5)	5.7(4.9)	4.5(4.4)	4.4(4.2)
MEDERR (%)	19.5	21.3	20.0	15.9	26.3	22.5	19.0	18.2
HOUR	64.7(17)	64.7(18.1)	66.3(17.5)	63.6(17.5)	65.5(18.6)	65.3(20)	63.8(19.2)	63.8(18.5)
SI (%)	8.5	11.0	8.1	9.6	10.3	10.9	9.2	6.8