Biometrics xx, 000–000 DOI: 000
Xx 0000

Web-Based Supplementary Materials

"Risk Prediction for Heterogeneous Populations with Application to Hospital Admission Prediction"

Jared Huling[1], Menggang Yu[2,*], Muxuan Liang[1], and Maureen Smith[3]

[1]Department of Statistics, University of Wisconsin-Madison, WI 53706, USA

[2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, WI, 53792, USA

[3] Department of Population Health Sciences,
University of Wisconsin School of Medicine and Public Health, Madison, WI, 53792, USA

**email:* meyu@biostat.wisc.edu

Web Appendix A. Statement of Theorems for Generalized Linear Models

For convenience, we first restate the theorems for generalized linear models and corresponding background information here before we present the proofs.

The density of a generalized linear model with canonical link given single observation $(y_k, \boldsymbol{x}_k)$ for subpopulation k can be written as:

$$f_k(y_k|\boldsymbol{x}_k, \theta_k) = h(y_k)\exp(y_k\theta_k - \phi(\theta_k)), \tag{1}$$

and our sparsity-inducing estimator is defined by the following problem

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\Big[\sum_{k=1}^{K}\frac{1}{N}\big\{-\boldsymbol{Y}_k^\top(\boldsymbol{X}_k\boldsymbol{\beta}_{k,\cdot}) + \boldsymbol{e}_k^\top\phi(\boldsymbol{X}_k\boldsymbol{\beta}_{k,\cdot})\big\}\Big] + \lambda P(\boldsymbol{\beta}). \tag{2}$$

We assume the following regularity conditions:

(C.1) $\boldsymbol{I}^k = \mathrm{E}_k[\phi''(\boldsymbol{x}_k\boldsymbol{\beta}^0_{k,\cdot})\boldsymbol{x}_k\boldsymbol{x}_k^\top]$ is finite and postive definite, where $\mathrm{E}_k[\cdot]$ is the expectation w.r.t $\boldsymbol{x}_k$ under the measure of subpopulation k.

(C.2) For subpopulation k, there is a sufficiently large enough open set $\mathcal{O}_k$ that contains $\boldsymbol{\beta}^0_{k,\cdot}$ such that $\forall \boldsymbol{\beta}_{k,\cdot} \in \mathcal{O}_k$,

$$|\phi'''(\boldsymbol{x}_k\boldsymbol{\beta}_{k,\cdot})| \leqslant M_k(\boldsymbol{x}_k) < \infty,$$

and

$$\mathrm{E}_k[M_k(\boldsymbol{x}_k)|x_{k,j}x_{k,l}x_{k,m}|] < \infty,$$

for all $1 \leqslant j, l, m \leqslant p$.

(C.3) $0 < \inf_{k=1,\ldots,K}\liminf_{N\to+\infty}\frac{n_k}{N} \leqslant \sup_{k=1,\ldots,K}\limsup_{N\to+\infty}\frac{n_k}{N} < 1$.

Theorem 1: *Assume the data are generated under the model represented by (1) and that our estimator is given by (2). Furthermore, assume that the non-zero patterns $\mathcal{Z}$ induced by the specified group structure $\mathcal{G}$ contain the true zero pattern. Assume conditions (C.1) - (C.3) and let $\lambda_{G,j} = ||\hat{\boldsymbol{\beta}}^{MLE}_{G,j}||_2^{-\gamma}$ for some $\gamma > 0$ such that $N^{(\gamma+1)/2}\lambda \to \infty$. If $\sqrt{N}\lambda \to 0$, then we* *have the following:*

$$P(\hat{J}_{\cdot,j} = J_{\cdot,j}) \to 1 \text{ as } N \to \infty, \tag{3}$$

and

$$\sqrt{n_k}(\hat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}^0_{k,\cdot}) \xrightarrow{d} \boldsymbol{Z}_k, \tag{4}$$

where $\boldsymbol{Z}_{k,J_{k,\cdot}} \sim N_{|J_{k,\cdot}|}(0, (\boldsymbol{I}^k_{J_{k,\cdot}J_{k,\cdot}})^{-1})$ *and* $\boldsymbol{Z}_{k,J^c_{k,\cdot}} = \mathbf{0}$.

Theorem 2: *Assume the data are generated under the model represented by (1) and that our estimator is given by (2). Here we do not necessarily assume that the group structure is correctly specifiied. Assume conditions (C.1) - (C.3) and let* $\lambda_{G,j} = ||\hat{\boldsymbol{\beta}}^{MLE}_{G,j}||_2^{-\gamma}$ *for some* $\gamma > 0$ *such that* $N^{(\gamma+1)/2}\lambda \to \infty$. *If* $\sqrt{N}\lambda \to 0$, *then we have the following:*

$$P(\hat{J}_{\cdot,j} = H_{\cdot,j}) \to 1 \text{ as } N \to \infty, \tag{5}$$

and

$$\sqrt{n_k}(\hat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}^0_{k,\cdot}) \xrightarrow{d} \boldsymbol{Z}_k, \tag{6}$$

where $\boldsymbol{Z}_{k,H_{k,\cdot}} \sim N_{|H_{k,\cdot}|}(0, (\boldsymbol{I}^k_{H_{k,\cdot}H_{k,\cdot}})^{-1})$ *and* $\boldsymbol{Z}_{k,H^c_{k,\cdot}} = \mathbf{0}$.

Web Appendix B. Additional Theorems for Semiparametric Linear Models

We now present new theorems not presented in the main text of our paper pertaining to semiparametric linear models. To loosen the restriction of normality required for our theorems for generalized linear models, we make separate set of assumptions. We assume that the observed response $\boldsymbol{Y}_k$ for subpopulation k follows the semiparametric linear model

$$\boldsymbol{Y}_k = \boldsymbol{X}_k\boldsymbol{\beta}^0_{k,\cdot} + \boldsymbol{\epsilon}_k \tag{7}$$

where $\boldsymbol{\epsilon}_k$ is a vector of *iid* errors with zero mean and finite variance σ^2_k. The resulting group lasso estimator is the solution of the following problem

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{k=1}^{K} \frac{1}{2N} \left\| \boldsymbol{Y}_k - \boldsymbol{X}_k \boldsymbol{\beta}_{k,\cdot} \right\|_2^2 + \lambda P(\boldsymbol{\beta}) \tag{8}$$

where $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{G \in \mathcal{G}} \lambda_{G,j} ||\boldsymbol{\beta}_{G,j}||_2$. To perform adaptive estimation, we take $\lambda_{G,j} = ||\hat{\boldsymbol{\beta}}_{G,j}^{MLE}||_2^{-\gamma}$, where $\gamma > 0$.

We assume the following classical conditions

(D.1) $\lim_{n \to \infty} \frac{1}{n_k} \boldsymbol{X}_k^\top \boldsymbol{X}_k \to \boldsymbol{Q}^k$ where $\boldsymbol{Q}^k$ is positive definite

(D.2) The error ϵ_k is a n_k-dimensional vector of random errors of subpopulation k. And the random errors of subpopulation k are *iid* with mean zero and finite variance σ_k^2. All subpopulations are independent.

(D.3) $0 < \inf_{k=1,\dots,K} \liminf_{n \to +\infty} \frac{n_k}{N} \leqslant \sup_{k=1,\dots,K} \limsup_{n \to +\infty} \frac{n_k}{N} < 1$.

Theorem 3:

Assume the data are generated under the model described in (7) and that our estimator is given by (8). Furthermore, assume the non-zero patterns $\mathcal{Z}$ induced by the specified group structure $\mathcal{G}$ contain the true zero pattern. Assume conditions (D.1) - (D.3) and let $\lambda_{G,j} = ||\hat{\boldsymbol{\beta}}_{G,j}^{OLS}||_2^{-\gamma}$ for some $\gamma > 0$ such that $N^{(\gamma+1)/2}\lambda \to \infty$. If $\sqrt{N}\lambda \to 0$, then we have the following:

$$P(\hat{J}_{\cdot,j} = J_{\cdot,j}) \to 1 \text{ as } N \to \infty \tag{9}$$

and

$$\sqrt{n_k}(\hat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}_{k,\cdot}^0) \xrightarrow{d} \boldsymbol{Z}_k \tag{10}$$

where $\boldsymbol{Z}_{k,J_{k,\cdot}} \sim N_{|J_{k,\cdot}|}(0, \sigma_k^2 (\boldsymbol{Q}_{J_{k,\cdot}J_{k,\cdot}}^k)^{-1})$ and $\boldsymbol{Z}_{k,J_{k,\cdot}^c} = \boldsymbol{0}$.

The following result pertains to cases where the group structure has been misspecified. This theorem is analogous to Theorem 2.

Theorem 4:

Assume the data are generated under the model described in (7) and that our estimator is given by (8). Furthermore, assume conditions (D.1) - (D.3) and let $\lambda_{G,j} = ||\hat{\boldsymbol{\beta}}_{G,j}^{MLE}||_2^{-\gamma}$ *for some* $\gamma > 0$ *such that* $N^{(\gamma+1)/2}\lambda \to \infty$. *If* $\sqrt{N}\lambda \to 0$, *then we have the following:*

$$P(\hat{J}_{\cdot,j} = J_{\cdot,j}) \to 1 \text{ as } N \to \infty, \tag{11}$$

and

$$\sqrt{n_k}(\hat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}_{k,\cdot}^0) \xrightarrow{d} \boldsymbol{Z}_k, \tag{12}$$

where $\boldsymbol{Z}_{k,H_{k,\cdot}} \sim N_{|H_{k,\cdot}|}(0, \sigma_k^2(\boldsymbol{Q}^k_{H_{k,\cdot}H_{k,\cdot}})^{-1})$ *and* $\boldsymbol{Z}_{k,H_{k,\cdot}^c} = \boldsymbol{0}$.

Web Appendix C. Proofs

In this section, we prove Theorems 1-4 of our paper. We will first prove the results for semiparametric linear models before proving the results for generalized linear models for simplicity of presentation.

Proof. [Proof of Theorem 4] We first prove asymptotic results for the linear model and then extend them to generalized linear models later. We begin by showing (10). Let $\boldsymbol{\beta}_{k,\cdot} = \boldsymbol{\beta}_{k,\cdot}^0 + \frac{1}{\sqrt{n_k}}\boldsymbol{u}_{k,\cdot}$, where $\boldsymbol{u}_{k,\cdot} \in \mathbb{R}^p$. We can write the objective function (8) multiplied by N as a function of $\boldsymbol{u} = (\boldsymbol{u}_{1,\cdot}, \ldots, \boldsymbol{u}_{K,\cdot})$ as follows:

$$F_N(\boldsymbol{u}) = \sum_{k=1}^{K} \frac{1}{2}\left|\left|\frac{1}{\sqrt{n_k}}\boldsymbol{X}_k\boldsymbol{u}_{k,\cdot} + \boldsymbol{\epsilon}_k\right|\right|_2^2 + \lambda N \sum_{j=1}^{p}\sum_{G\in\mathcal{G}} \lambda_{G,j}||\boldsymbol{\beta}_{G,j}^0 + \frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}||_2,$$

where $\boldsymbol{u}_{k,j} = \sqrt{n_k}(\boldsymbol{\beta}_{k,j} - \boldsymbol{\beta}_{k,j}^0)$, $\frac{1}{\sqrt{n_G}}$ is a $|G|$-dimensional vector with $\frac{1}{\sqrt{n_G}} = (\frac{1}{\sqrt{n_{G_1}}}, \ldots, \frac{1}{\sqrt{n_{G_m}}})$ if $G = \{G_1, \ldots, G_m\} \subset \{1, \cdots, K\}$, and $\circ$ represents coordinate-wise multiplication of two vectors. Let $\hat{\boldsymbol{u}}^{(N)} = \operatorname{argmin}_{\boldsymbol{u}} F_N(\boldsymbol{u})$ and note that $\hat{\boldsymbol{u}}_{k,\cdot}^{(N)} = \sqrt{n_k}(\hat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}_{k,\cdot}^0)$, where $\hat{\boldsymbol{\beta}}$ is the minimizer of the objective function (8). Thus, to investigate the asymptotic distribution of

$\hat{\boldsymbol{\beta}}$ is equivalent to investigating the asymptotic distribution of $\hat{\boldsymbol{u}}^{(N)}$. Now, we let

$$\begin{align}
D_N(\boldsymbol{u}) &= F_N(\boldsymbol{u}) - F_N(\mathbf{0}) \tag{13}\\
&= \sum_{k=1}^{K} \left(\frac{1}{2}\boldsymbol{u}_{k,\cdot}^{\top} \left(\frac{1}{n_k}\boldsymbol{X}_k{}^{\top}\boldsymbol{X}_k \right) \boldsymbol{u}_{k,\cdot} - \frac{1}{\sqrt{n_k}}\boldsymbol{u}_{k,\cdot}^{\top}\boldsymbol{X}_k\boldsymbol{\epsilon}_k \right) \nonumber\\
&\quad + \sqrt{N}\lambda\sqrt{N} \left(\sum_{j=1}^{p}\sum_{G\in\mathcal{G}} \lambda_{G,j}||\boldsymbol{\beta}_{G,j}^{0} + \frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}||_2 - \sum_{j=1}^{p}\sum_{G\in\mathcal{G}} \lambda_{G,j}||\boldsymbol{\beta}_{G,j}^{0}||_2 \right) \nonumber\\
&= \sum_{k=1}^{K} \left(\frac{1}{2}\boldsymbol{u}_{k,\cdot}^{\top} \left(\frac{1}{n_k}\boldsymbol{X}_k{}^{\top}\boldsymbol{X}_k \right) \boldsymbol{u}_{k,\cdot} - \frac{1}{\sqrt{n_k}}\boldsymbol{u}_{k,\cdot}^{\top}\boldsymbol{X}_k\boldsymbol{\epsilon}_k \right) \nonumber\\
&\quad + \sqrt{N}\lambda \sum_{j=1}^{p}\sum_{G\in\mathcal{G}_{H_{\cdot,j}}} \lambda_{G,j}\sqrt{N} \left(\left\|\boldsymbol{\beta}_{G,j}^{0} + \frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}\right\|_2 - \left\|\boldsymbol{\beta}_{G,j}^{0}\right\|_2 \right) \tag{14}\\
&\quad + \sqrt{N}\lambda \sum_{j=1}^{p}\sum_{G\in\mathcal{G}_{H^c_{\cdot,j}}} \lambda_{G,j}\sqrt{N} \left(\left\|\boldsymbol{\beta}_{G,j}^{0} + \frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}\right\|_2 - \left\|\boldsymbol{\beta}_{G,j}^{0}\right\|_2 \right), \tag{15}
\end{align}$$

where $\mathcal{G}_{H_{\cdot,j}}$ is the set of all groups G with any $k \in G$ such that $\boldsymbol{\beta}_{k,j}^{0} \neq 0$ and $\mathcal{G}_{H^c_{\cdot,j}}$ is the set of all groups G such that $\boldsymbol{\beta}_{k,j}^{0} = 0$ for all $k \in G$. To obtain the asymptotic distribution of $\hat{\boldsymbol{u}}^{(N)}$, we first investigate the asymptotic property of $D_N(\boldsymbol{u})$ for every fix $\boldsymbol{u} \in \mathbb{R}^{Kp}$.

For all $G \in \mathcal{G}_{H_{\cdot,j}}$, we have $\lambda_{G,j} \xrightarrow{p} ||\boldsymbol{\beta}_{G,j}^{0}||_2^{-\gamma}$ and by taking the directional derivative in the direction of $\boldsymbol{u}_{G,j}$, we have

$$\sqrt{N} \left(\left\|\boldsymbol{\beta}_{G,j}^{0} + \frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}\right\|_2 - \left\|\boldsymbol{\beta}_{G,j}^{0}\right\|_2 \right) = \sqrt{N}\frac{1}{\sqrt{n_G}} \circ \frac{\boldsymbol{u}_{G,j}{}^{\top}\boldsymbol{\beta}_{G,j}^{0}}{||\boldsymbol{\beta}_{G,j}^{0}||_2} + o_p(1).$$

Then because $\sqrt{N}\lambda = o(1)$, we have by Slutsky's theorem that

$$\sqrt{N}\lambda\lambda_{G,j}\sqrt{N} \left(\left\|\boldsymbol{\beta}_{G,j}^{0} + \frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}\right\|_2 - \left\|\boldsymbol{\beta}_{G,j}^{0}\right\|_2 \right) = o_p(1).$$

For all $G \in \mathcal{G}_{H^c_{\cdot,j}}$, because $N^{\gamma/2}||\hat{\boldsymbol{\beta}}_{G,j}^{MLE}||_2^{\gamma} = O_p(1)$ and $\sqrt{N} \left(\left\|\boldsymbol{\beta}_{G,j}^{0} + \frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}\right\|_2 - \left\|\boldsymbol{\beta}_{G,j}^{0}\right\|_2 \right) = ||\sqrt{N}\frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}||_2$, we have

$$\lambda\lambda_{G,j}\sqrt{N}||\sqrt{N}\frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}||_2 = ||\sqrt{N}\frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}||_2\lambda\frac{N^{(\gamma+1)/2}}{(\sqrt{N}||\hat{\boldsymbol{\beta}}_{G,j}^{MLE}||_2)^{\gamma}} \xrightarrow{p} \infty, \tag{16}$$

if $\boldsymbol{u}_{G,j} \neq \mathbf{0}$, and,

$$\lambda\lambda_{G,j}\sqrt{N}||\sqrt{N}\frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}||_2 = ||\sqrt{N}\frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}||_2 \lambda \frac{N^{(\gamma+1)/2}}{(\sqrt{N}||\hat{\boldsymbol{\beta}}_{G,j}^{MLE}||_2)^{\gamma}} = 0, \tag{17}$$

if $\boldsymbol{u}_{G,j} = \mathbf{0}$.

To facilitate the notations, for any vector $\mathbf{a}_{k,\cdot} \in R^p$ associated with subpopulation k and a index set $I_{k,\cdot} \subset \{1, \ldots, p\}$ associated with subpopulation k, $\mathbf{a}_{I_{k,\cdot}}$ represents a $|I_{k,\cdot}|$ dimensional vector with elements in $\mathbf{a}_{k,\cdot}$ indexed by $I_{k,\cdot}$.

Since the term (13) converges in distribution to $\sum_{k=1}^{K} \left(\frac{1}{2}\boldsymbol{u}_{k,\cdot}^{\top}\boldsymbol{Q}^k\boldsymbol{u}_{k,\cdot} + \boldsymbol{u}_{k,\cdot}^{\top}\boldsymbol{W}_{k,\cdot}\right)$, where $\boldsymbol{W}_{k,\cdot} \sim N_p(0, \sigma_k^2\boldsymbol{Q}^k)$, using Slutsky's theorem we have that $D_N(\boldsymbol{u}) \xrightarrow{d} D(\boldsymbol{u})$ for each $\boldsymbol{u}$, where

$$D(\boldsymbol{u}) = \begin{cases} \sum_{k=1}^{K} \left(\frac{1}{2}\boldsymbol{u}_{H_{k,\cdot}}^{\top}\boldsymbol{Q}_{H_{k,\cdot}H_{k,\cdot}}^{k}\boldsymbol{u}_{H_{k,\cdot}} + \boldsymbol{u}_{H_{k,\cdot}}^{\top}\boldsymbol{W}_{H_{k,\cdot}}\right) & \text{if } \boldsymbol{u}_{H_{k,\cdot}^c} = \mathbf{0}, \forall k = 1, \ldots, K, \\ \infty & \text{otherwise.} \end{cases}$$

It is clear that $D_N(\boldsymbol{u})$ is convex and the unique minimum of $D(\boldsymbol{u})$ is $((\boldsymbol{Q}_{H_{k,\cdot}H_{k,\cdot}}^{k})^{-1}\boldsymbol{W}_{H_{k,\cdot}}, \mathbf{0})$. By the epiconvergence results of Geyer (1994) and Knight and Fu (2000), we have the following:

$$\hat{\boldsymbol{u}}_{H_{k,\cdot}}^{(N)} \xrightarrow{d} (\boldsymbol{Q}_{H_{k,\cdot}H_{k,\cdot}}^{k})^{-1}\boldsymbol{W}_{H_{k,\cdot}} \text{ and } \hat{\boldsymbol{u}}_{H_{k,\cdot}^c}^{(N)} \xrightarrow{d} \mathbf{0} \tag{18}$$

where $(\boldsymbol{Q}_{H_{k,\cdot}H_{k,\cdot}}^{k})^{-1}\boldsymbol{W}_{H_{k,\cdot}} \sim N_{|H_{k,\cdot}|}(0, \sigma_k^2(\boldsymbol{Q}_{H_{k,\cdot}H_{k,\cdot}}^{k})^{-1})$. Hence (10) is verified.

We now show selection consistency. For any $j \in H_{k,\cdot}$, we have by the asymptotic normality (10) and thus it follows that $P(k \in \hat{J}_{\cdot,j}) \to 1$. To verify (9) it is equivalent to show that for any k' such that $j \in H_{k',\cdot}^c$ which implies that $k' \notin \text{Hull}(J_{\cdot,j})$, $P(k' \in \hat{J}_{\cdot,j}) \to 0$. Suppose $\hat{\boldsymbol{\beta}}_{k',j} \neq \mathbf{0}$ and $j \in H_{k',\cdot}^c$, there exists $G_0 \subset \{1, \ldots, K\}$ such that $k' \in G_0 \subset H_{\cdot,j}^c$. Assume without lose of generality that $G_0 = \{1, \ldots, k_0\}$. Then by the KKT optimality conditions as derived in Lee and Xing (2014) and Jenatton et al. (2011) we know that

$$\mathbf{0} = \begin{pmatrix} -\boldsymbol{X}_{1,j}(\boldsymbol{Y}_1 - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{1,\cdot}) \\ \vdots \\ -\boldsymbol{X}_{k_0,j}(\boldsymbol{Y}_{k_0} - \boldsymbol{X}_{k_0}\hat{\boldsymbol{\beta}}_{k_0,\cdot}) \end{pmatrix} + \lambda \begin{pmatrix} \mathbf{v}_{1,j} \\ \vdots \\ \mathbf{v}_{k_0,j} \end{pmatrix} = \Phi + \lambda\mathbf{v},$$

where $\boldsymbol{X}_{k,j}$ is the jth covariate for subpopulation k, $\mathbf{v}_{k',j} = N(\sum_{G\in\mathcal{G}s.t.k'\in G} \lambda_{G,j} \frac{\hat{\boldsymbol{\beta}}_{k',j}}{||\hat{\boldsymbol{\beta}}_{G,j}||_2})$ is the subgradient of $n\sum_{G\in\mathcal{G}} \lambda_{G,j}||\hat{\boldsymbol{\beta}}_{G,j}||_2$ with respect to $\hat{\boldsymbol{\beta}}_{k',j}$. Due to the asymptotic normality of $\sqrt{n_k}(\boldsymbol{\beta}^0_{k,\cdot} - \hat{\boldsymbol{\beta}}_{k,\cdot})$ and condition (D.1)(D.3), we have

$$\left\| \frac{1}{\sqrt{N}}\Phi \right\|_2 = O_p(1).$$

In addition, by the same arguments that show (16) and definition of G_0, we have

$$||\sqrt{N}\lambda\lambda_{G,j}\frac{\hat{\boldsymbol{\beta}}_{G_0,j}}{||\hat{\boldsymbol{\beta}}_{G_0,j}||_2}||_2 \xrightarrow{p} \infty.$$

Thus we can see that the probability of the KKT conditions holding vanishes,

$$P(k' \in \hat{J}_{\cdot,j}) \leqslant P(||\Phi||_2 = ||\lambda\mathbf{v}||_2) \to 0.$$

Hence, we have established selection consistency.

Proof. [Proof of Theorem 2]

We will start from asymptotic normality (4). Let $\boldsymbol{\beta}_{k,\cdot} = \boldsymbol{\beta}^0_{k,\cdot} + \frac{1}{\sqrt{n_k}}\boldsymbol{u}_{k,\cdot}$, where $\boldsymbol{u}_{k,\cdot} \in \mathbb{R}^p$. We define the following function:

$$F_N(\boldsymbol{u}) = \sum_{k=1}^{K}\left(-\boldsymbol{Y}_k^\top(\boldsymbol{X}_k(\boldsymbol{\beta}^0_{k,\cdot} + \frac{\boldsymbol{u}_{k,\cdot}}{\sqrt{n_k}})) + \boldsymbol{e}_k^\top\phi(\boldsymbol{X}_k(\boldsymbol{\beta}^0_{k,\cdot} + \frac{\boldsymbol{u}_{k,\cdot}}{\sqrt{n_k}}))\right)$$
$$+ N\lambda\sum_{j=1}^{p}\sum_{G\in\mathcal{G}}\lambda_{G,j}||\boldsymbol{\beta}^0_{G,j} + \frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}||_2.$$

Same as linear case, let $\hat{\boldsymbol{u}}^{(N)} = \text{argmin}_{\boldsymbol{u}} F_N(\boldsymbol{u})$ and $\hat{\boldsymbol{u}}^{(N)}_{k,\cdot} = \sqrt{n_k}(\hat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}^0_{k,\cdot})$. Now, we define that $D_N(\boldsymbol{u}) = F_N(u) - F_N(0)$. We notice that $D_N(\boldsymbol{u})$ is a convex function. By Taylor expansion, we obtain

$$D_N(\boldsymbol{u}) = \sum_{k=1}^{K} S_{1,k}^{(N)} + \sum_{k=1}^{K} S_{2,k}^{(N)} + S_3^{(n)} + S_4^{(N)} + \sum_{k=1}^{K} S_{5,k}^{(N)},$$

with

$$S_{1,k}^{(N)} = -\frac{1}{\sqrt{n_k}}[\boldsymbol{Y}_k - \phi'(\boldsymbol{X}_k\boldsymbol{\beta}_{k,\cdot}^0)]^\top(\boldsymbol{X}_k\boldsymbol{u}_{k,\cdot}),$$
$$S_{2,k}^{(N)} = \frac{1}{n_k}\frac{1}{2}\boldsymbol{u}_{k,\cdot}^\top(\boldsymbol{X}_k{}^\top\mathrm{diag}\{\phi''(\boldsymbol{X}_k\boldsymbol{\beta}_{k,\cdot}^0)\}\boldsymbol{X}_k)\boldsymbol{u}_{k,\cdot},$$
$$S_3^{(N)} = \sqrt{N}\lambda\sum_{j=1}^{p}\sum_{G\in\mathcal{G}_{H_{\cdot,j}}}\lambda_{G,j}\sqrt{N}\left(\left\|\boldsymbol{\beta}_{G,j}^0 + \frac{1}{\sqrt{n_G}}\circ\boldsymbol{u}_{G,j}\right\|_2 - \left\|\boldsymbol{\beta}_{G,j}^0\right\|_2\right),$$
$$S_4^{(N)} = \sqrt{N}\lambda\sum_{j=1}^{p}\sum_{G\in\mathcal{G}_{H^c_{\cdot,j}}}\lambda_{G,j}\sqrt{N}\left(\left\|\boldsymbol{\beta}_{G,j}^0 + \frac{1}{\sqrt{n_G}}\circ\boldsymbol{u}_{G,j}\right\|_2 - \left\|\boldsymbol{\beta}_{G,j}^0\right\|_2\right),$$
$$S_{5,k}^{(N)} = n_k^{-3/2}\frac{1}{6}\phi'''(\boldsymbol{X}_k\tilde{\boldsymbol{\beta}}_{k,\cdot})(\boldsymbol{X}_k\boldsymbol{u}_{k,\cdot}\circ\boldsymbol{X}_k\boldsymbol{u}_{k,\cdot}\circ\boldsymbol{X}_k\boldsymbol{u}_{k,\cdot}),$$

where $\tilde{\beta}_{k,\cdot}$ is between $\boldsymbol{\beta}_{k,\cdot}^0$ and $\boldsymbol{\beta}_{k,\cdot}^0 + \frac{1}{\sqrt{n_k}}\boldsymbol{u}_{k,\cdot}$. $\phi'(\boldsymbol{X}_k\boldsymbol{\beta}_{k,\cdot}^0)$ is a n_k-dimensional vector with transformation ϕ' on each entry of $\boldsymbol{X}_k\boldsymbol{\beta}_{k,\cdot}^0$, and similar notations are adopted with $\phi''(\boldsymbol{X}_k\boldsymbol{\beta}_{k,\cdot}^0)$ and $\phi'''(\boldsymbol{X}_k\tilde{\boldsymbol{\beta}}_{k,\cdot})$. To investigate the asymptotic property of $\hat{\boldsymbol{u}}^{(N)}$, we can find the limit of $D_N(\boldsymbol{u})$ for each fixed $\boldsymbol{u}$. By the properties of exponential families, we have

$$\mathrm{E}_k S_{1,k}^{(N)} = 0,$$

and

$$\mathrm{Var}_k S_{1,k}^{(N)} = \boldsymbol{u}_{k,\cdot}^\top\boldsymbol{I}^k(\boldsymbol{\beta}_{k,\cdot}^0)\boldsymbol{u}_{k,\cdot}.$$

Thus, by central limit theorem, the first term $S_{1,k}^{(N)} \xrightarrow{d} \boldsymbol{u}_{k,\cdot}^\top N_p(\boldsymbol{0}, \boldsymbol{I}^k(\boldsymbol{\beta}_{k,\cdot}^0))$. By LLN, it is obvious that $S_{2,k}^{(N)} \xrightarrow{p} \frac{1}{2}\boldsymbol{u}_{k,\cdot}^\top\boldsymbol{I}^k(\boldsymbol{\beta}_{k,\cdot}^0))\boldsymbol{u}_{k,\cdot}$. Let us discuss about the behaviour of $S_3^{(N)}$ and $S_4^{(N)}$.

Our discussion here is the same as that in the proof of Theorem 4. We directly use the following two results from Theorem 4:

For all $G \in \mathcal{G}_{H_{\cdot,j}}$, $\sqrt{N}\lambda\lambda_{G,j}\sqrt{N}\left(\left\|\boldsymbol{\beta}_{G,j}^0 + \frac{1}{\sqrt{n_G}}\circ\boldsymbol{u}_{G,j}\right\|_2 - \left\|\boldsymbol{\beta}_{G,j}^0\right\|_2\right) \xrightarrow{p} 0$. Thus, $S_3^{(N)} = o_p(1)$.

For all $G \in \mathcal{G}^c_{H_{.,j}}$, $\sqrt{N}\lambda\lambda_{G,j}\sqrt{N}\left(\left\|\boldsymbol{\beta}^0_{G,j} + \frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}\right\|_2 - \left\|\boldsymbol{\beta}^0_{G,j}\right\|_2\right) \xrightarrow{p} \infty$, if $\boldsymbol{u}_{G,j} \neq \mathbf{0}$, and

$$\sqrt{N}\lambda\lambda_{G,j}\sqrt{N}\left(\left\|\boldsymbol{\beta}^0_{G,j} + \frac{1}{\sqrt{n_G}} \circ \boldsymbol{u}_{G,j}\right\|_2 - \left\|\boldsymbol{\beta}^0_{G,j}\right\|_2\right) \xrightarrow{p} 0,$$

if $\boldsymbol{u}_{G,j} = \mathbf{0}$. Thus, $S_4^{(N)} = o_p(1)$ if $\boldsymbol{u}_{G,j} = \mathbf{0}$ for all j, and $S_4^{(N)} \xrightarrow{p} \infty$ otherwise.

The fifth term can be bounded by the condition (C.2) as

$$6\sqrt{n_k}S^{(N)}_{5,k} \leqslant n_k^{-1}M_k(\boldsymbol{X}_k)(|\boldsymbol{X}_k\boldsymbol{u}_{k,\cdot}| \circ |\boldsymbol{X}_k\boldsymbol{u}_{k,\cdot}| \circ |\boldsymbol{X}_k\boldsymbol{u}_{k,\cdot}|) \xrightarrow{p} \mathrm{E}_k[M_k(\boldsymbol{x}_k)|\boldsymbol{x}_k\boldsymbol{u}_{k,\cdot}|^3] < \infty,$$

where $|\cdot|$ takes absolute value of each entry of a vector, and $M_k(\boldsymbol{X}_k)$ is a n_k-dimensional vector with transformation $M_k(\cdot)$ on each row of $\boldsymbol{X}_k$.

By using Slutsky's theorem we have that $D_N(\boldsymbol{u}) \xrightarrow{d} D(\boldsymbol{u})$ for every $\boldsymbol{u}$, where

$$D(\boldsymbol{u}) = \begin{cases} \sum_{k=1}^K \left(\frac{1}{2}\boldsymbol{u}^\top_{H_{k,\cdot}}\boldsymbol{I}^k_{H_{k,\cdot}H_{k,\cdot}}\boldsymbol{u}_{H_{k,\cdot}} + \boldsymbol{u}^\top_{H_{k,\cdot}}\boldsymbol{W}_{H_{k,\cdot}}\right) & \text{if } \boldsymbol{u}_{H^c_{k,\cdot}} = \mathbf{0}, \forall k = 1,\ldots,K, \\ \infty & \text{otherwise.} \end{cases}$$

where $\boldsymbol{W}_{H_{k,\cdot}} = N(\mathbf{0}, \boldsymbol{I}^k_{H_{k,\cdot}H_{k,\cdot}})$. For $D_N(\boldsymbol{u})$ is convex and $D(\boldsymbol{u})$ is strictly convex with unique minimum $((\boldsymbol{I}^k_{H_{k,\cdot}H_{k,\cdot}})^{-1}\boldsymbol{W}_{H_{k,\cdot}}, \mathbf{0})$. By the same argument in the proof of Theorem 3, we have

$$\hat{\boldsymbol{u}}^{(N)}_{H_{k,\cdot}} \xrightarrow{d} (\boldsymbol{I}^k_{H_{k,\cdot}H_{k,\cdot}})^{-1}\boldsymbol{W}_{H_{k,\cdot}} \text{ and } \hat{\boldsymbol{u}}^{(N)}_{H^c_{k,\cdot}} \xrightarrow{d} \mathbf{0}. \tag{19}$$

where $(\boldsymbol{I}^k_{H_{k,\cdot}H_{k,\cdot}})^{-1}\boldsymbol{W}_{H_{k,\cdot}} \sim N_{|H_{k,\cdot}|}(0, (\boldsymbol{I}^k_{H_{k,\cdot}H_{k,\cdot}})^{-1})$, and hence (4) is verified.

Now we show the model selection consistency. For any $j \in H_{k,\cdot}$, the asymptotic normality implies that $P(k \in \hat{J}_{j,\cdot}) \to 1$. It is sufficient to show that for any k' such that $j \in H^c_{k',\cdot}$ which implies that $k' \notin \mathrm{Hull}(J_{\cdot,j})$, $P(k' \in \hat{J}_{\cdot,j}) \to 0$. Suppose $\hat{\boldsymbol{\beta}}_{k',j} \neq \mathbf{0}$ and $j \in H^c_{k',\cdot}$, there exists $G_0 \subset \{1,\ldots,K\}$ such that $k' \in G_0 \subset H^c_{\cdot,j}$. Assume without lose of generality that $G_0 = \{1,\ldots,k_0\}$. Then by optimality condition of Lemma 14 in Jenatton et al. (2011), we must have

$$\mathbf{0} = \begin{pmatrix} \boldsymbol{X}_{1,j}(\boldsymbol{Y}_1 - \phi'(\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{1,\cdot})) \\ \vdots \\ \boldsymbol{X}_{k_0,j}(\boldsymbol{Y}_1 - \phi'(\boldsymbol{X}_{k_0}\hat{\boldsymbol{\beta}}_{k_0,\cdot})) \end{pmatrix} + \lambda \begin{pmatrix} \mathbf{v}_{1,j} \\ \vdots \\ \mathbf{v}_{k_0,j} \end{pmatrix} = \Phi + \lambda\mathbf{v},$$

where $\boldsymbol{X}_{k,j}$ is the jth covariate for subpopulation k,

$$\mathbf{v}_{k',j} = N(\sum_{G \in \mathcal{G} s.t. k' \in G} \lambda_{G,j} \frac{\hat{\boldsymbol{\beta}}_{k',j}}{||\hat{\boldsymbol{\beta}}_{G,j}||_2})$$

is the subgradient of $N\sum_{G\in\mathcal{G}} \lambda_{G,j}||\hat{\boldsymbol{\beta}}_{G,j}||_2$ with respect to $\hat{\boldsymbol{\beta}}_{k',j}$.

Thus, $P(k' \in \hat{J}^{\cdot,j}) \leqslant P(||\frac{1}{\sqrt{N}}\Phi||_2 = ||\frac{1}{\sqrt{N}}\lambda\mathbf{v}||_2)$.

Notice that

$$\frac{1}{\sqrt{n_k}}\boldsymbol{X}_{k,j}(\boldsymbol{Y}_k - \phi'(\boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k, \cdot)) = \boldsymbol{T}_1^{(N)} + \boldsymbol{T}_2^{(N)} + \boldsymbol{T}_3^{(N)}, \tag{20}$$

with

$$\begin{aligned} \boldsymbol{T}_1^{(N)} &= \boldsymbol{X}_{k,j}(\boldsymbol{Y}_k - \phi'(\boldsymbol{X}_k\boldsymbol{\beta}^0_{k,\cdot}))/\sqrt{n_k}, \\ \boldsymbol{T}_2^{(N)} &= (\frac{1}{n_k}\boldsymbol{X}_k{}^\top \mathrm{diag}\{\phi''(\boldsymbol{X}_k\boldsymbol{\beta}^0_{k,\cdot})\}\boldsymbol{X}_{k,j})\sqrt{n_k}(\boldsymbol{\beta}^0_k - \hat{\boldsymbol{\beta}}_k), \\ \boldsymbol{T}_3^{(N)} &= \frac{1}{n_k}\phi'''(\boldsymbol{X}_k\tilde{\beta}_{k,\cdot})\left(\boldsymbol{X}_{k,j} \circ (\boldsymbol{X}_k\sqrt{n_k}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})) \circ (\boldsymbol{X}_k\sqrt{n_k}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}))\right)/\sqrt{n^k}, \end{aligned}$$

where $\tilde{\beta}_{k,\cdot}$ is between $\hat{\boldsymbol{\beta}}_{k,\cdot}$ and $\boldsymbol{\beta}^0_{k,\cdot}$.

By the previous analysis, the first term $\boldsymbol{T}_1^{(N)} = O_p(1)$. By LLN and normality (4), the second term is also $O_p(1)$. By the regularity condition (C.2)(C.3) and normality (4), the third term is $O_p(\frac{1}{\sqrt{N}})$. Thus, $||\frac{1}{\sqrt{N}}\Phi||_2 = O_p(1)$. However, by the same arguments that show (16) and definition of G_0, we have $||\sqrt{N}\lambda\lambda_{G,j}\frac{\hat{\boldsymbol{\beta}}_{G_0,j}}{||\hat{\boldsymbol{\beta}}_{G_0,j}||_2}||_2 \xrightarrow{p} \infty$.

It implies that $||\frac{1}{\sqrt{N}}\lambda\mathbf{v}||_2 \xrightarrow{p} \infty$. Thus, $P(k' \in \hat{J}_{\cdot,j}) \to 0$.

Web Appendix D. Computational Details for ADMM Algorithm

We utilize an alternating direction method of multipliers (ADMM) (Glowinski and Marroco, 1975; Gabay and Mercier, 1976; Boyd et al., 2011) algorithm for maximization. The ADMM algorithm works by decomposing an objective function and solving the decomposed subproblems iteratively. Each subproblem has a computationally tractable form. ADMM solves problems of the form

$$\begin{aligned}&\text{minimize } f(\boldsymbol{\beta}) + P(\boldsymbol{\gamma})\\ &\text{subject to } A\boldsymbol{\beta} + B\boldsymbol{\gamma} = c\end{aligned}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{Kp}$, $\boldsymbol{\gamma} \in \mathbb{R}^{m}$, $A \in \mathbb{R}^{r\times Kp}$, $B \in \mathbb{R}^{r\times m}$, and $c \in \mathbb{R}^{r}$. Both f and P must be convex functions. Typically f represents some loss function and P represents a penalty. In the simplest case the constraint is of the form $\boldsymbol{\beta} = \boldsymbol{\gamma}$. To solve the above problem, the augmented Lagrangian is formed as:

$$\begin{aligned}L_\rho(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = f(\boldsymbol{\beta}) + P(\boldsymbol{\gamma}) + \boldsymbol{\nu}^\top(A\boldsymbol{\beta} + B\boldsymbol{\gamma} - c)\\ + (\rho/2)||A\boldsymbol{\beta} + B\boldsymbol{\gamma} - c||_2^2\end{aligned}$$

and ADMM iterates by alternatingly minimizing with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and updating the Lagrangian parameter $\boldsymbol{\nu}$

$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, L_\rho(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\nu}^{(t)}) \tag{21}$$

$$\boldsymbol{\gamma}^{(t+1)} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}}\, L_\rho(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}, \boldsymbol{\nu}^{(t)}) \tag{22}$$

$$\boldsymbol{\nu}^{(t+1)} = \boldsymbol{\nu}^{(t)} + \rho(A\boldsymbol{\beta}^{(t+1)} + B\boldsymbol{\gamma}^{(t+1)} - c)$$

where t indexes the iteration number. ADMM has been shown to converge for any $\rho > 0$.

The following describes ADMM applied to the overlapping group lasso problem. Let $m = \sum_{G\in\mathcal{G}} |G|$, $g = |\mathcal{G}|$, and suppose $\mathcal{G} = \{G_1, \cdots, G_g\}$, let $F = (F_1, \ldots, F_g)$ be an $m\times Kp$ matrix where F_l is a $|G_l|\times Kp$ matrix with the (i, j)th entry equals to 1 if j is the i^{th} element of group

G_l, and 0 otherwise, $\forall j = 1, \ldots, g$. Then $F\boldsymbol{\beta}$ is a vector of length m comprised components of $\boldsymbol{\beta}$ where each element of $\boldsymbol{\beta}$ appears in $F\boldsymbol{\beta}$ the total number of times it appears in any group. For example, if p=1, K=3, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$ and $\mathcal{G} = \{\{1,2\},\{2,3\}\}$, then

$$F = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \text{ and } F\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

In this example, the penalty $P(\boldsymbol{\gamma})$ is

$$P(\boldsymbol{\gamma}) = \lambda(||(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)||_2 + ||(\boldsymbol{\gamma}_3, \boldsymbol{\gamma}_4)||_2).$$

In general, the penalty $P(\boldsymbol{\gamma})$ is

$$P(\boldsymbol{\gamma}) = \lambda \sum_{l=1}^{g} \lambda_l ||\boldsymbol{\gamma}_l||_2,$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_g)$ and $\boldsymbol{\gamma}_l$ is a $|G_l|$-dimensional vector. Note that for the nonoverlapping group lasso, $A = I_{Kp}$. The ADMM algorithm for the overlapping group lasso is constructed by taking $A = F$ as defined above, $B = -I_m$, and $c = \mathbf{0}$. When $f(\boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$, step (21) for the overlapping group lasso is simply the solution of $(\boldsymbol{X}^\top\boldsymbol{X} + \rho F^\top F)\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{Y} + F^\top\boldsymbol{\nu}^{(t)} + \rho F^\top\boldsymbol{\gamma}^{(t)}$. For our case, $\boldsymbol{X}$ is a diagonal block matrix of dimension $n \times Kp$ with $\boldsymbol{X}_k$ the kth block, and $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K)$. When $f(\boldsymbol{\beta})$ is the negative log-likelihood, step (21) can be carried out by Newton-Raphson or other standard optimization techniques. As step (22) is group-separable, it can be minimized by minimizing with respect to each group $\boldsymbol{\gamma}_l$ independently. This is achieved by the block soft-thresholding operator $S_{\lambda\lambda_{G_l}/\rho}((F\boldsymbol{\beta}^{(t+1)})_l - \boldsymbol{\nu}_l^{(t)}/\rho)$, where $S_\lambda(\boldsymbol{u}) = \boldsymbol{u}\,(1 - \lambda/||\boldsymbol{u}||_2)_+$ and $(F\boldsymbol{\beta}^{(t+1)})_l$ and $\boldsymbol{\nu}_l^{(t)}$ are defined in the same way as $\boldsymbol{\gamma}_l$. Our convergence criterion is the same as suggested in Section 3.3.1 of Boyd et al. (2011) with $\epsilon^{\text{abs}} = \epsilon^{\text{rel}} = 10^{-5}$. Further details on an improvement of this algorithm for GLMs is described in the Online Supplementary Materials.

Web Appendix E. Additional Computational Details for GLMs

While the ADMM algorithm we described in the Appendix can be used for any convex negative log-likelihood, it is inefficient for high dimensional problems for generalized linear models due to the requirement of the iterative minimization with respect to $\boldsymbol{\beta}$ when there is no analytical solution. Instead, we consider using the ADMM algorithm inside a proximal Newton algorithm. Proximal Newton methods are a generalization of Newton-type methods for nonsmooth objectives. Proximal Newton methods are essentially an iteratively reweighted penalized least squares procedure where a quadratic approximation to the likelihood plus a penalty is solved in each iteration. This quadratic subproblem can be solved using the ADMM algorithm described above for weighted least squares plus the overlapping group lasso penalty. For more details and theory on proximal Newton methods see Lee et al. (2012); Schmidt (2010); Patriksson (2013). A well-known example of a proximal algorithm is `glmnet` (Friedman et al., 2010).

Web Appendix F. Additional Simulation Results

First we present simulation results pertaining to the variable selection properties of our estimator for logistic regression models. The simulation setup is the same as described in Section 4 of the main text. While prediction is the primary interest of this paper, we also investigate the variable selection properties of vennLasso and vennLasso Adaptive. We investigate the average number of false positives, number of false negatives, sensitivity, and specificity in Figure 1. The vennLasso adaptive method performs better than the vennLasso in terms of false negatives and performs similarly to the vennLasso in terms of false positives. The vennLasso adaptive method tends to perform best in terms of variable selection among all methods and has balanced performance across the selection measures. Note that we did not report these results for the Interaction Lasso and Interaction HierLasso methods because a covariate with non-zero coefficient in the interaction models can potentially influence the outcomes of all subpopulations, regardless whether the subpopulation indicators interact with the covariate or not. For example, the main effects in these models are relevant across all subpopulations. Hence, while these models can flexibly capture heterogeneity, the specific meanings of the estimated effects have an entirely different interpretation. Therefore it is not very satisfactory or sensible to report these results for the interaction models.

The coverage results of the vennLasso Adaptive method are presented in Table 1. The average empirical coverages are typically higher than the nominal level, but improve with larger sample sizes. The coverage of a particular covariate is considered to be zero if a truly nonzero coefficient is estimated to be zero and hence the coverage simulation is also a reflection of the selection properties of our penalty.

[Figure 1 about here.]

[Table 1 about here.]

Here we present additional simulation results with the same simulation setup as described in Section 4 of the main text, but where the total sparsity of the true coefficients is smaller and larger than the results from in the main text. Specifically, here we choose the overall average coefficient sparsity to be 0.75 and 0.95. The results here indicate that the superiority of vennLasso is largely unaffected by the overall coefficient sparsity.

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

Furthermore, we present computing time results for the analysis of the hospital admissions data. All tuning parameters for all methods are chosen via 10-fold cross validation and the model for each fold is computed in parallel using a total of 10 computing cores. Computing times are listed in Table 2

[Table 2 about here.]

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* **3,** 1–122.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33,** 1–22.

Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications* **2,** 17–40.

Geyer, C. (1994). On the asymptotics of constrained m-estimation. *The Annals of Statistics* **22,** 1993–2010.

Glowinski, R. and Marroco, A. (1975). Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique* **9,** 41–76.

Jenatton, R., Audibert, J.-Y., and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research* **12,** 2777–2824.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28,** 1356–1378.

Lee, J., Sun, Y., and Saunders, M. (2012). Proximal newton-type methods for convex optimization. In *Advances in Neural Information Processing Systems*, pages 836–844.

Lee, S. and Xing, E. P. (2014). Screening rules for overlapping group lasso. Technical report.

Patriksson, M. (2013). *Nonlinear programming and variational inequality problems: a unified approach*, volume 23. Springer Science & Business Media.

Schmidt, M. (2010). *Graphical model structure learning with l1-regularization*. PhD thesis, Citeseer.

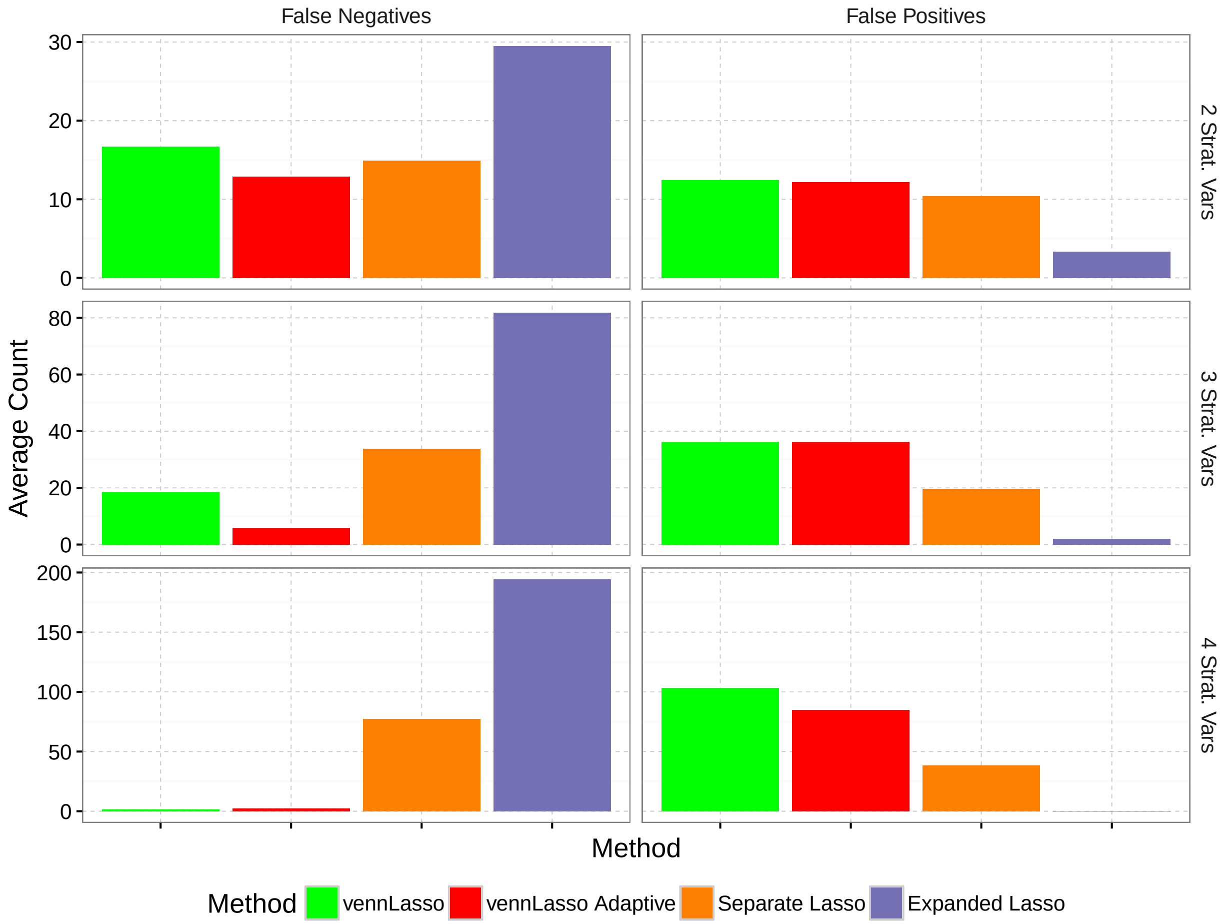

Figure 1. The number of covariates is set to 100 and the average sparsity of the coefficients is 0.875 for this simulation. The results depicted above are for the setting with 250 observations per subpopulation, and a maximum effect size $c = 1$.

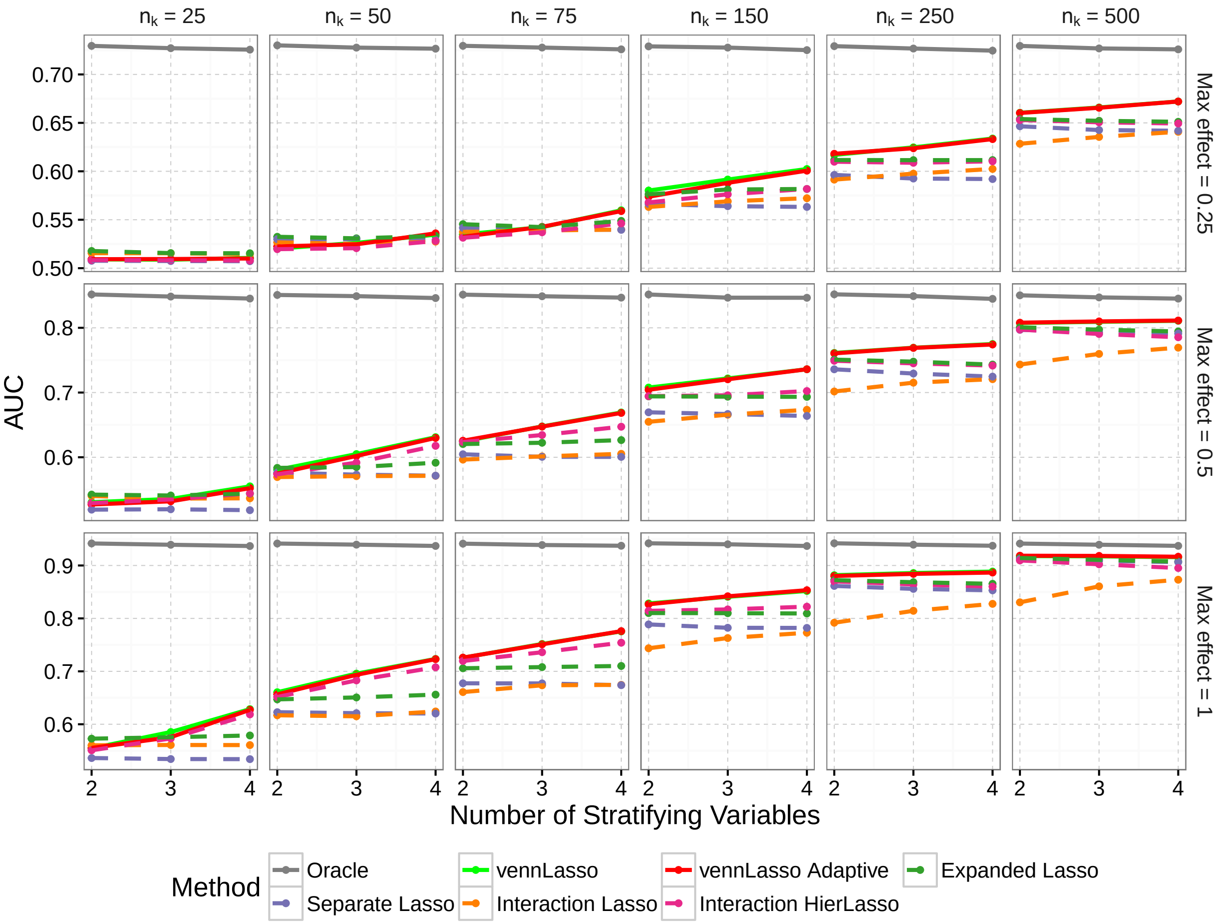

Figure 2. The number of covariates is set to 100 and the average sparsity of the coefficients is 0.75 for this simulation. The number of observations listed is the number of observations per subpopulation. Hence, the number of coefficients to be estimated and the number of total observations increase together, but their ratio is consistent.

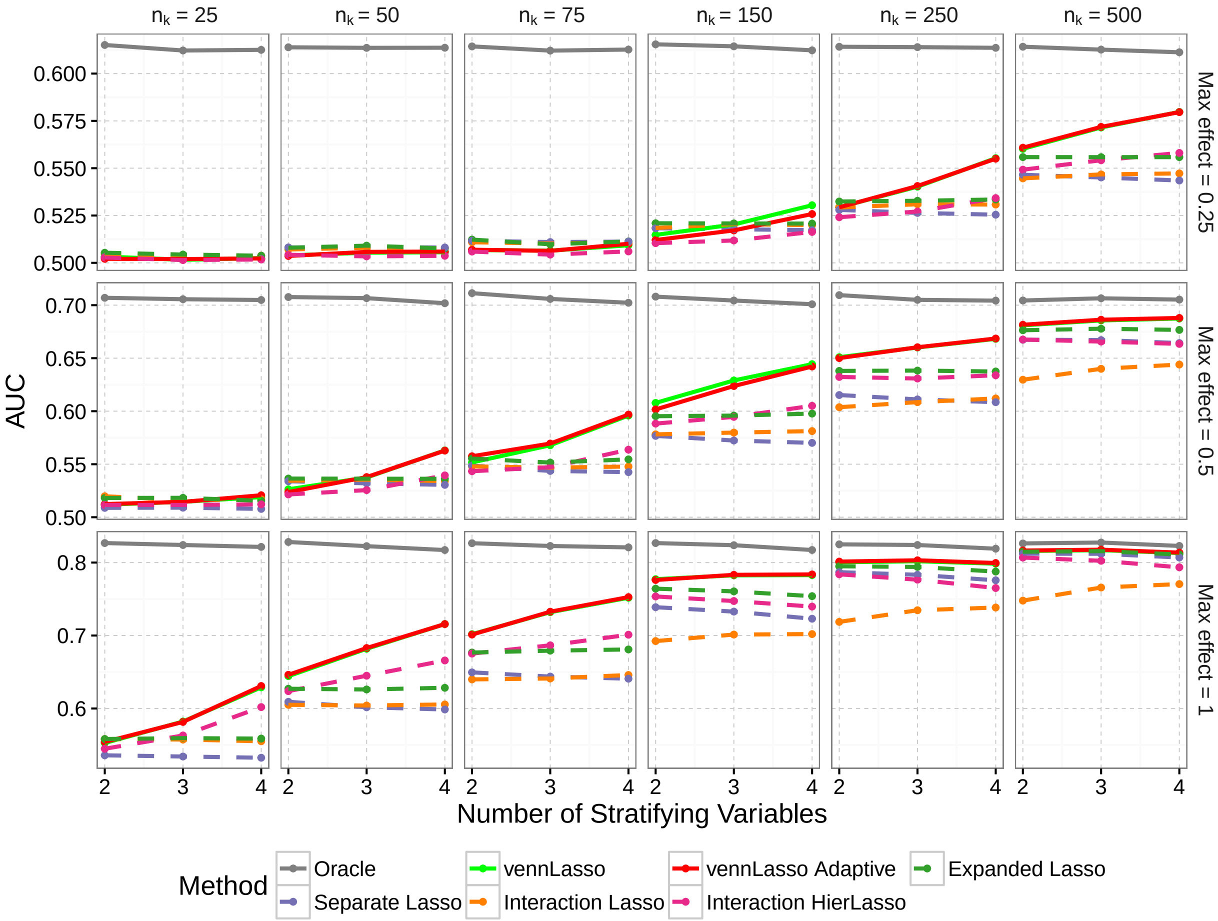

Figure 3. The number of covariates is set to 100 and the average sparsity of the coefficients is 0.95 for this simulation. The number of observations listed is the number of observations per subpopulation. Hence, the number of coefficients to be estimated and the number of total observations increase together, but their ratio is consistent.

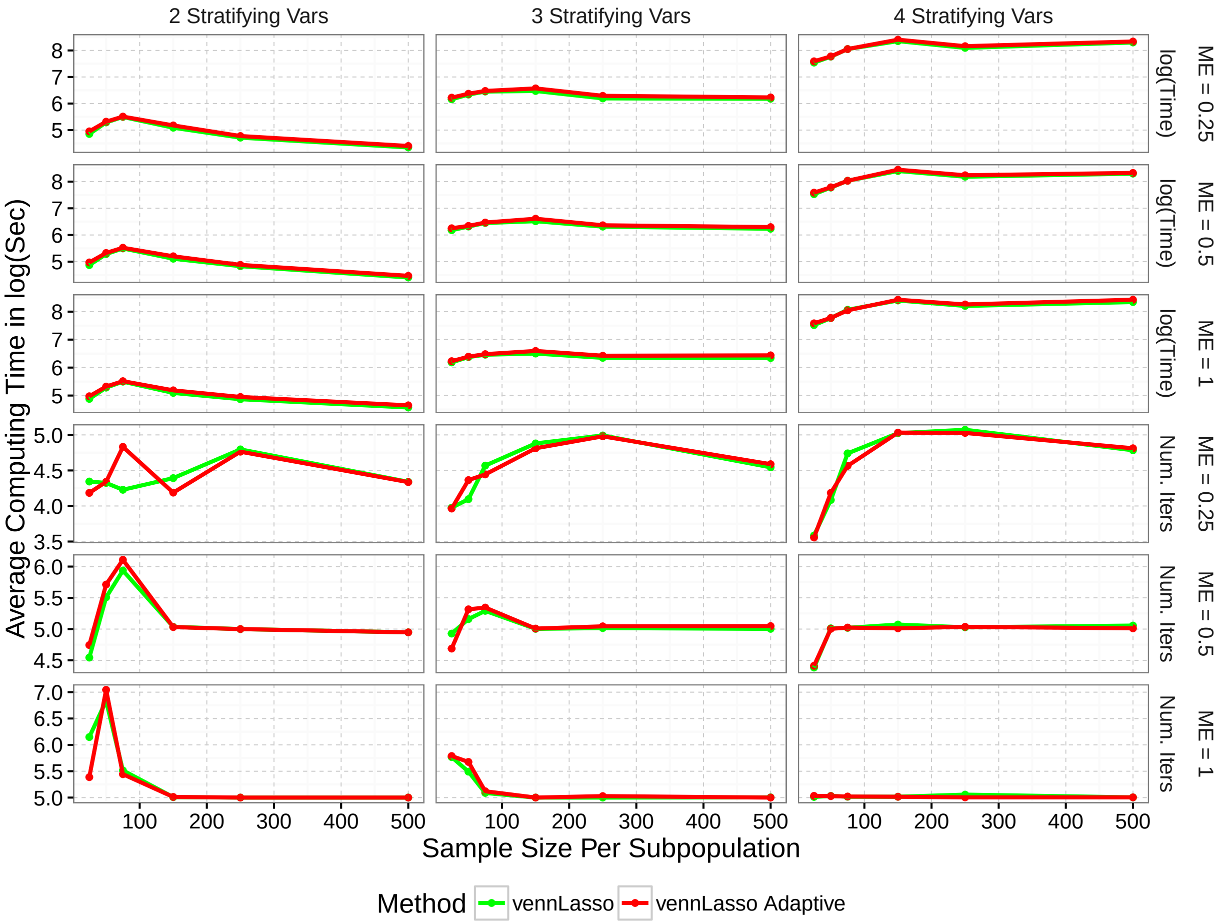

Figure 4. The number of covariates is set to 100 and the average sparsity of the coefficients is 0.875 for this simulation. Displayed above are the average computing times for the vennLasso and vennLasso Adaptive for the simulations. The number of iterations displayed are the number of outside Newton iterations of the proximal Newton algorithm described in the Appendix.

Table 1
Empirical coverage results of the vennLasso Adaptive penalty under the same simulation settings as the simulation presented in Section 4 of the main text of this paper. The numbers above are the average coverage over 500 simulations.

		Coverage		
		Max Effect Size		
n_k	Conditions	0.25	0.5	1
150	2	0.918	0.958	0.988
	3	0.931	0.977	0.994
	4	0.955	0.989	0.996
250	2	0.946	0.984	0.993
	3	0.976	0.991	0.989
	4	0.988	0.992	0.964
500	2	0.970	0.982	0.973
	3	0.983	0.982	0.976
	4	0.989	0.981	0.976

Table 2
All computation times shown below for the admissions data are performed on a Microsoft Windows Server 2008 Enterprise system with 64 Intel® Xeon® E5-4650 CPU cores and 768 Gigabytes of RAM.

Method	Computing Time (sec)
vennLasso	3874
vennLasso Adaptive	4119
Separate Lasso	44
Expanded Lasso	517
Interaction Lasso	678
Interaction HierLasso	4395