When can Multi-Site Datasets be Pooled for Regression? Hypothesis Tests, ℓ_2-consistency and Neuroscience Applications (Supplementary Material)

Hao Henry Zhou [1] Yilin Zhang [1] Vamsi K. Ithapu [1] Sterling C. Johnson [1 2] Grace Wahba [1] Vikas Singh [1]

In this supplementary document we first present the proofs of all the technical results in Section 2 and 3 of the main paper. We then expand upon the Section 4 and present extra experiments to strengthen the evaluations from the main paper.

Remarks on transformations in pre-processing step: For all $i \in \{1, ..., k\}$, after applying the transformation (shift correction), we pool (X_i, y_i) together to estimate β^*. Note that in general the transformation (shift correction) should *not* depend on the responses y_i, otherwise we get a dependence on the noise. To see this, notice that $y_i = X_i\beta_i + \epsilon_i$ where X_i is the transformed set of features. But when the transformation depends on y_i, then X_i will also depend on ϵ_i, which causes a poor estimation of β^* (and β_i). In situations where the transformations must involve y_i, a sensible strategy is to separate each site's dataset into two parts, where one part from each site is used to learn the transformation, and the other part (after applying the learned transformation) is used for pooling towards β^* estimation and conducting our hypothesis test.

1. Proof of Section 2

We now provide the proofs of the results presented in the main paper.

Theorem 2.1. *$\tau_i = \frac{\sigma_1}{\sigma_i}$ achieve the smallest variance in $\hat{\beta}$.*

Proof. The choice of τ_i leads to weighted least squares, which is known to be the best linear unbiased estimator (BLUE) under uncorrelated heteroscedastic errors. The variance of $\hat{\beta}$ is equivalent to the case when $\Delta\beta_i = 0$. In the latter case, BLUE condition holds and setting τ_i to the above value achieves lowest variance. The equivalence between variances under two cases completes the proof. □

Lemma 2.2. *For multi-site model, we have*

$$\frac{\|Bias_\beta\|_2^2}{\|G^{-1/2}\Delta\beta\|_2^2} \le \|(\hat{\Sigma}_1^k)^{-2}(\hat{\Sigma}_2^k(n_1\hat{\Sigma}_1)^{-1}\hat{\Sigma}_2^k + \hat{\Sigma}_2^k)\|_*, \tag{1}$$

$$Var_\beta = \sigma_1^2 \left\|(n_1\hat{\Sigma}_1)^{-1} - (n_1\hat{\Sigma}_1 + \hat{\Sigma}_2^k)^{-1}\right\|_*. \tag{2}$$

Proof. The estimation from single site model is unbiased, and it has the following variance.

$$Var_1 = tr((X_1^T X_1)^{-1})\sigma_1^2 = tr((n_1\hat{\Sigma}_1)^{-1})\sigma_1^2 \tag{3}$$

[1]University of Wisconsin-Madison [2]William S. Middleton Memorial Veteran's Affairs Hospital. Correspondence to: Hao Zhou <hzhou@stat.wisc.edu>, Vikas Singh <vsingh@biostat.wisc.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

The estimation error from multi-sites model has the following closed form expression

$$\begin{aligned}\hat{\beta}-\beta^* = &\left(\begin{pmatrix} X_1 \\ \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}^T \begin{pmatrix} X_1 \\ \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}\right)^{-1} \begin{pmatrix} \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}^T \begin{pmatrix} \tau_2 X_2 (\Delta\beta_2) \\ .. \\ \tau_k X_k (\Delta\beta_k) \end{pmatrix} \\ &+ \left(\begin{pmatrix} X_1 \\ \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}^T \begin{pmatrix} X_1 \\ \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}\right)^{-1} \begin{pmatrix} X_1 \\ \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}^T \begin{pmatrix} \epsilon_1 \\ \tau_2 \epsilon_2 \\ .. \\ \tau_k \epsilon_k \end{pmatrix}\end{aligned} \tag{4}$$

First term in the summation from (4) is bias, while second term is variance. We can see that our choice of $\tau_i = \frac{\sigma_1}{\sigma_i}$ resolves heteroscedastic errors issue among sites. We further simplify bias and variance terms, and obtain

$$Var_2 = tr((n_1 \hat{\Sigma}_1 + \sum_{i=2}^{k} n_i \tau_i^2 \hat{\Sigma}_i)^{-1})\sigma_1^2 \tag{5}$$

The reduced variance statement is proved. For the bias term, it is equivalent as shown below.

$$\left(\begin{pmatrix} X_1 \\ \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}^T \begin{pmatrix} X_1 \\ \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}\right)^{-1} \begin{pmatrix} \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}^T \begin{pmatrix} \tau_2 X_2 & 0 & ... & 0 \\ 0 & \tau_3 X_3 & ... & 0 \\ 0 & 0 & ... & \tau_k X_k \end{pmatrix} G^{1/2} \left\{ G^{-1/2} \begin{pmatrix} \Delta\beta_2 \\ .. \\ \Delta\beta_k \end{pmatrix} \right\} \tag{6}$$

A one step Cauchy Schwartz inequality is then applied. Then our final proof is to show $\|..\|_F^2$ on

$$\left(\begin{pmatrix} X_1 \\ \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}^T \begin{pmatrix} X_1 \\ \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}\right)^{-1} \begin{pmatrix} \tau_2 X_2 \\ .. \\ \tau_k X_k \end{pmatrix}^T \begin{pmatrix} \tau_2 X_2 & 0 & ... & 0 \\ 0 & \tau_3 X_3 & ... & 0 \\ 0 & 0 & ... & \tau_k X_k \end{pmatrix} G^{1/2} \tag{7}$$

is equal to right side of the bias relaxation in (1).
It is easy to see that $\|A\|_F^2 = \|A^T A\|_*$. Based on this, we can see the first term of matrix inverse contributes the $(\hat{\Sigma}_1^k)^{-2}$ in (1). Let the other part in (7) be L. We have

$$LL^T = \begin{pmatrix} \tau_2^2 X_2^T X_2 \\ .. \\ \tau_k^2 X_k^T X_k \end{pmatrix}^T G \begin{pmatrix} \tau_2^2 X_2^T X_2 \\ .. \\ \tau_k^2 X_k^T X_k \end{pmatrix} = \begin{pmatrix} n_2 \tau_2^2 \hat{\Sigma}_2 \\ .. \\ n_k \tau_k^2 \hat{\Sigma}_k \end{pmatrix}^T G \begin{pmatrix} n_2 \tau_2^2 \hat{\Sigma}_2 \\ .. \\ n_k \tau_k^2 \hat{\Sigma}_k \end{pmatrix} \tag{8}$$

After some manipulations, this becomes $(\hat{\Sigma}_2^k (n_1 \hat{\Sigma}_1)^{-1} \hat{\Sigma}_2^k + \hat{\Sigma}_2^k)$. The bias part is proved. □

Theorem 2.3. **a):** *The multi-sites model has smaller MSE of $\hat{\beta}$ than single-site model whenever*

$$H_0 : \left\| G^{-1/2} \Delta\beta \right\|_2^2 \leq \sigma_1^2. \tag{9}$$

b): *Further, we have the following test statistic,*

$$\left\| \frac{G^{-1/2} \Delta\hat{\beta}}{\sigma_1} \right\|_2^2 \sim \chi^2_{(k-1)*p} \left(\left\| \frac{G^{-1/2} \Delta\beta}{\sigma_1} \right\|_2^2 \right), \tag{10}$$

where $\|G^{-1/2}\Delta\beta/\sigma_1\|_2$ is called a "condition value".

Proof. **(a):** Based on Lemma 2.2, the theorem is proved when right side in (9) is replaced by

$$\frac{\sigma_1^2 \left\|(n_1\hat{\Sigma}_1)^{-1} - (n_1\hat{\Sigma}_1 + \hat{\Sigma}_2^k)^{-1}\right\|_*}{\|(\hat{\Sigma}_1^k)^{-2}(\hat{\Sigma}_2^k(n_1\hat{\Sigma}_1)^{-1}\hat{\Sigma}_2^k + \hat{\Sigma}_2^k)\|_*} \tag{11}$$

We first calculate the numerator

$$\begin{aligned}\sigma_1^2 \left\|(n_1\hat{\Sigma}_1)^{-1} - (n_1\hat{\Sigma}_1 + \hat{\Sigma}_2^k)^{-1}\right\|_* &= \sigma_1^2 \left\|\left[(n_1\hat{\Sigma}_1)^{-1}(n_1\hat{\Sigma}_1 + \hat{\Sigma}_2^k) - I\right](n_1\hat{\Sigma}_1 + \hat{\Sigma}_2^k)^{-1}\right\|_* \\ &= \sigma_1^2 \left\|(n_1\hat{\Sigma}_1)^{-1}\hat{\Sigma}_2^k(n_1\hat{\Sigma}_1 + \hat{\Sigma}_2^k)^{-1}\right\|_*\end{aligned} \tag{12}$$

The denominator is then given by

$$\|(\hat{\Sigma}_1^k)^{-2}(\hat{\Sigma}_2^k(n_1\hat{\Sigma}_1)^{-1}\hat{\Sigma}_2^k + \hat{\Sigma}_2^k)\|_* = \|(\hat{\Sigma}_1^k)^{-2}((\hat{\Sigma}_2^k + n_1\hat{\Sigma}_1)(n_1\hat{\Sigma}_1)^{-1}\hat{\Sigma}_2^k)\|_* \tag{13}$$

$$\text{Remember } \hat{\Sigma}_1^k = \hat{\Sigma}_2^k + n_1\hat{\Sigma}_1, \text{, we continue} \tag{14}$$

$$= \|((\hat{\Sigma}_2^k + n_1\hat{\Sigma}_1)^{-1}(n_1\hat{\Sigma}_1)^{-1}\hat{\Sigma}_2^k)\|_* = \|((n_1\hat{\Sigma}_1)^{-1}\hat{\Sigma}_2^k(n_1\hat{\Sigma}_1 + \hat{\Sigma}_2^k)^{-1})\|_* \tag{15}$$

The last step uses the property of $\|..\|_*$ norm. The proof is completed by noticing the simplified form of numerator and denominator. It is clear now that the right side in (9) is exactly σ_1^2.

(b): First, we show $\sigma_1^2 G$ is the covariance matrix of $\Delta\hat{\beta}$. We have

$$cov(\Delta\hat{\beta}_i, \Delta\hat{\beta}_j) = cov(\hat{\beta}_i, \hat{\beta}_j) - cov(\hat{\beta}_i, \hat{\beta}_1) - cov(\hat{\beta}_1, \hat{\beta}_j) + cov(\hat{\beta}_1, \hat{\beta}_1) \tag{16}$$

$$\text{Since each site is independent from other site, we have} \tag{17}$$

$$cov(\Delta\hat{\beta}_i, \Delta\hat{\beta}_j) = cov(\hat{\beta}_1, \hat{\beta}_1) = \sigma_1^2(n_1\hat{\Sigma}_1)^{-1}\text{for } i \neq j \tag{18}$$

$$cov(\Delta\hat{\beta}_i, \Delta\hat{\beta}_i) = cov(\hat{\beta}_i, \hat{\beta}_i) + cov(\hat{\beta}_1, \hat{\beta}_1) = \sigma_1^2((n_1\hat{\Sigma}_1)^{-1} + (n_i(\sigma_1^2/\sigma_i^2)\hat{\Sigma}_i)^{-1}) = \sigma_1^2((n_1\hat{\Sigma}_1)^{-1} + (n_i\tau_i^2\hat{\Sigma}_i)^{-1}) \tag{19}$$

$\Delta\hat{\beta}$ follows Gaussian distribution since it is a linear transformation of Gaussian distribution. It's expectation is $\Delta\beta$ since each $\hat{\beta}_i$ is an unbiased estimator. Hence, we have

$$\Delta\hat{\beta} \sim N(\Delta\beta, \sigma_1^2 G) \tag{20}$$

This distribution result, and noticing the connection between Gaussian and non-central χ^2 distributions completes the proof. □

Corollary 2.4. *For the case where we have two participating sites, the condition* (9) *from Theorem* 2.3 *reduces to*

$$H_0 : \Delta\beta^T((n_1\hat{\Sigma}_1)^{-1} + (n_2\tau_2^2\hat{\Sigma}_2)^{-1})^{-1}\Delta\beta \leq \sigma_1^2. \tag{21}$$

Proof. The proof is follows by noticing the form of G when $k = 2$. □

Theorem 2.5. *Analysis in Section 2.1 holds for β in model with Z confounding features, when we replace $\hat{\Sigma}_i$ with*

$$\tilde{\Sigma}_i = \hat{\Sigma}_{xx_i} - \hat{\Sigma}_{xz_i}(\hat{\Sigma}_{zz_i})^{-1}\hat{\Sigma}_{zx_i}. \tag{22}$$

Proof. Define $\gamma^T = (\gamma_1^T, ..., \gamma_k^T)$, $X_{all}^T = \left(X_1^T, \tau_2 X_2^T, ..., \tau_k X_k^T\right)$, $Z_{all} = Diag\left(Z_1, \tau_2 Z_2, ..., \tau_k Z_k\right)$. We have

$$\begin{aligned}\begin{pmatrix}\hat{\beta} \\ \hat{\gamma}\end{pmatrix} - \begin{pmatrix}\beta^* \\ \gamma^*\end{pmatrix} = \begin{pmatrix}X_{all}^T X_{all} & X_{all}^T Z_{all} \\ Z_{all}^T X_{all} & Z_{all}^T Z_{all}\end{pmatrix}^{-1}\begin{pmatrix}X_{all}^T \\ Z_{all}^T\end{pmatrix}\begin{pmatrix}0 \\ \tau_2 X_2(\Delta\beta_2) \\ .. \\ \tau_k X_k(\Delta\beta_k)\end{pmatrix} + \\ \begin{pmatrix}X_{all}^T X_{all} & X_{all}^T Z_{all} \\ Z_{all}^T X_{all} & Z_{all}^T Z_{all}\end{pmatrix}^{-1}\begin{pmatrix}X_{all}^T \\ Z_{all}^T\end{pmatrix}\begin{pmatrix}\epsilon_1 \\ \tau_2\epsilon_2 \\ .. \\ \tau_k\epsilon_k\end{pmatrix}\end{aligned} \tag{23}$$

Using sub-matrix inverse property, we obtain

$$\begin{pmatrix} X_{all}^T X_{all} & X_{all}^T Z_{all} \\ Z_{all}^T X_{all} & Z_{all}^T Z_{all} \end{pmatrix}^{-1} \begin{pmatrix} X_{all}^T \\ Z_{all}^T \end{pmatrix} = \begin{pmatrix} (\tilde{X}_{all}^T \tilde{X}_{all})^{-1} \tilde{X}_{all}^T \\ (\tilde{Z}_{all}^T \tilde{Z}_{all})^{-1} \tilde{Z}_{all}^T \end{pmatrix} \tag{24}$$

We then have

$$\tilde{Z}_{all} = (I - X_{all}(X_{all}^T X_{all})^{-1} X_{all}^T) Z_{all} \tag{25}$$

$$\tilde{X}_{all} = (I - Z_{all}(Z_{all}^T Z_{all})^{-1} Z_{all}^T) X_{all} = \begin{pmatrix} (I - Z_1(Z_1^T Z_1)^{-1} Z_1^T) X_1 \\ (I - Z_2(Z_2^T Z_2)^{-1} Z_2^T) X_2 \\ .. \\ (I - Z_k(Z_k^T Z_k)^{-1} Z_k^T) X_k \end{pmatrix} \tag{26}$$

Define

$$H_{Z_i} = (I - Z_i(Z_i^T Z_i)^{-1} Z_i^T) \tag{27}$$

Hence, we have

$$\begin{aligned} \hat{\beta} - \beta^* &= (\tilde{X}_{all}^T \tilde{X}_{all})^{-1} \tilde{X}_{all}^T \begin{pmatrix} 0 \\ \tau_2 X_2(\Delta\beta_2) \\ .. \\ \tau_k X_k(\Delta\beta_k) \end{pmatrix} + (\tilde{X}_{all}^T \tilde{X}_{all})^{-1} \tilde{X}_{all}^T \begin{pmatrix} \epsilon_1 \\ \tau_2 \epsilon_2 \\ .. \\ \tau_k \epsilon_k \end{pmatrix} \\ &= (\tilde{X}_{all}^T \tilde{X}_{all})^{-1} \sum_{i=2}^{k} \tau_i \tilde{X}_i^T X_i (\Delta\beta_i) + (\tilde{X}_{all}^T \tilde{X}_{all})^{-1} \tilde{X}_{all}^T \begin{pmatrix} \epsilon_1 \\ \tau_2 \epsilon_2 \\ .. \\ \tau_k \epsilon_k \end{pmatrix} \end{aligned} \tag{28}$$

We also observe that

$$\tilde{X}_i^T X_i = X_i^T H_{Z_i} X_i = X_i^T H_{Z_i}^2 X_i = \tilde{X}_i^T \tilde{X}_i \tag{29}$$

Therefore, we can apply our previous results to a subset of parameters if we replace X_i by $\tilde{X}_i$. Since our results only depend on $\hat{\Sigma}_i$, we only need to replace it by

$$\frac{1}{n_i} \tilde{X}_i^T \tilde{X}_i = \frac{1}{n_i} X_i^T H_{Z_i} X_i = \hat{\Sigma}_{xx_i} - \hat{\Sigma}_{xz_i} (\hat{\Sigma}_{zz_i})^{-1} \hat{\Sigma}_{zx_i} \tag{30}$$

This proves the theorem. □

2. Proof of Section 3

Definition 3.1. *The m-sparse minimal and maximal eigenvalues of C, denoted by $\phi_{\min}(m)$ and $\phi_{\max}(m)$, are*

$$\min_{\nu:\|\nu\|_0 \leq \lceil m \rceil} \frac{\nu^T C \nu}{\nu^T \nu} \quad \textit{and} \quad \max_{\nu:\|\nu\|_0 \leq \lceil m \rceil} \frac{\nu^T C \nu}{\nu^T \nu} \tag{31}$$

We first list down the two key theorem statements that we prove in this section.

Theorem 3.2. *Let $0 \leq \alpha \leq 0.4$. Assume there exist constants $0 \leq \rho_{\min} \leq \rho_{\max} \leq \infty$ such that*

$$\begin{aligned} &\liminf_{n\to\infty} \phi_{\min}\left(s_p \left(1 + \frac{2\alpha}{1 - 2\alpha} \right)^2 \right) \geq \rho_{\min}, \textit{ and} \\ &\limsup_{n\to\infty} \phi_{\max}(s_p + \min\{\sum_{i=1}^{k} n_i, kp\}) \leq \rho_{\max}. \end{aligned} \tag{32}$$

Then, for $\lambda \propto \sigma\sqrt{\bar{n}\log(kp)}$, there exists a constant $\omega > 0$ such that, with probability converging to 1 for $n \to \infty$,

$$\frac{1}{k} \|\hat{B}^\lambda - B^*\|_F^2 \leq \omega \sigma^2 \frac{\bar{s}\log(kp)}{\bar{n}}, \tag{33}$$

where $\bar{s} = \{(1-\alpha)\sqrt{s_p} + \alpha\sqrt{s_h/k}\}^2$, σ is the noise level.

Theorem 3.3. *Let $0.4 \leq \tilde{\alpha} \leq 1$. Assume there exist constants $0 \leq \rho_{\min} \leq \rho_{\max} \leq \infty$ such that*

$$\begin{aligned} &\liminf_{n\to\infty} \phi_{\min}\left(s_h\left(1+\frac{(1-\tilde{\alpha})}{\tilde{\alpha}}\right)^2\right) \geq k_{\min}, \text{ and} \\ &\limsup_{n\to\infty} \phi_{\max}(s_h + \min\{\sum_{i=1}^{k} n_i, kp\}) \leq k_{\max}. \end{aligned} \tag{34}$$

Then, for $\tilde{\lambda} \propto \sigma\sqrt{\bar{n}\log(kp)}$, there exists a $\omega > 0$ such that, with probability converging to 1 for $n \to \infty$, we have

$$\frac{1}{k}\|\hat{B}^{\lambda} - B^{*}\|_F^2 \leq \omega\sigma^2 \frac{\tilde{s}\log(kp)}{\bar{n}}, \tag{35}$$

with $\tilde{s} = \{(1-\tilde{\alpha})\sqrt{s_p/k} + \tilde{\alpha}\sqrt{s_h/k}\}^2$ instead of $\bar{s}$.

Comment about Theorem 3.3: We do not penalize by $\sqrt{k}$ when the sparsity patterns across sites share few of the features. To see this, first observe that when sparsity patterns are similar, most of the groups we have are non-sparse, and the effects of $\sqrt{k}\|\beta^j\|_2$ and $\|\beta^j\|_1$ have the same scale. This is simply because, $\sqrt{k}\sqrt{a_1^2+...+a_k^2}$ is close to $|a_1|+...+|a_k|$ whenever $|a_1|, ..., |a_k|$ are close. However when sparsity patterns across sites share few features only, most of the groups are going to be sparse. For these groups, we should use $\|\beta^j\|_2$, because in this setting $\sqrt{a_1^2+0+...+0}$ is close to $|a_1|+0+...+0$.

3.1. Proof of Theorem 3.2:

We follow the proof procedure from Lasso (Meinshausen & Yu, 2009) and group Lasso (Liu & Zhang, 2009) results. Let B^{λ} be the estimator under the absence of noise, i.e., $B^{\lambda} = \hat{B}^{\lambda,0}$, where $\hat{B}^{\lambda,\xi}$ is defined as in (37). The ℓ_2-distance can then be bounded by $\|\hat{B}^{\lambda} - B^{*}\|_F^2 \leq 2\|\hat{B}^{\lambda} - B^{\lambda}\|_F^2 + 2\|B^{\lambda} - B^{*}\|_F^2$. The first term on the right-hand side represents the variance of the estimation, while the second term represents the bias. The bias contribution follows directly from Lemma 3.4 below, and the variance bound term follows from Lemma 3.9.

De-noised response. For $0 < \xi < 1$, we define a de-noised version of the response variable as follows,

$$Y_i(\xi) = X_i\beta_i + \xi\epsilon_i \tag{36}$$

We can regulate the amount of noise with the parameter ξ.

For $\xi = 0$, only the signal is retained. The original observations with the full amount of noise are recovered for $\xi = 1$. Now consider for $0 \leq \xi \leq 1$ the estimator $\hat{B}^{\lambda,\xi}$,

$$\begin{aligned} \hat{B}^{\lambda,\xi} &= \arg\min_{B} \sum_{i=1}^{k} \|Y_i(\xi) - X_i\beta_i\|_2^2 + \lambda\Lambda(B) \\ \Lambda(B) &= (1-\alpha)\sqrt{k}\sum_{j=1}^{p}\|\beta^j\|_2 + \alpha\sum_{j=1}^{p}\|\beta^j\|_1 \end{aligned} \tag{37}$$

The ordinary sparse multi-site Lasso estimate is recovered under the full amount of noise so that $\hat{B}^{\lambda,1} = \hat{B}^{\lambda}$. Using the notation from the previous results, we have $\hat{B}^{\lambda,0} = B^{\lambda}$, for the estimate in the absence of noise. The definition of the de-noised version of the sparse multi-site Lasso estimator will be helpful for the proof as it allows to characterize the variance of the estimator.

3.1.1. Part I of proof – Dealing with bias

Let P_* be the set of nonzero groups of B^*, i.e., $P_* = \{j : \beta^j \neq 0\}$. The cardinality of P_* is denoted by s_p. For each j in P_*, let H_j be the set of nonzero elements of β_j, i.e., $H_j = \{i : \beta_i^j \neq 0\}$. The number of all nonzero elements of B is denoted by s_h. For the following, let B^{λ} be the estimator $\hat{B}^{\lambda,0}$ with no noise (as defined in (37)). For each λ, the solution B^{λ} can be written as $B^{\lambda} = B^{*} + \Gamma^{\lambda}$. We define γ^j and γ_i to be j-th column and i-th row of Γ. γ is the transpose of the unfolded vector of Γ by row. Denote $\lambda_2 = \lambda(1-\alpha)$ and $\eta = \frac{\alpha}{1-\alpha}$. Then

$$\Gamma^{\lambda} = \arg\min_{\Gamma} f(\Gamma) \tag{38}$$

The function $f(\Gamma)$ is given by

$$f(\Gamma) = \bar{n}\gamma^T C\gamma + \lambda_2 \left\{ \sum_{j \in P_*^C} (\sqrt{K}\|\gamma^j\|_2 + \eta\|\gamma^j\|_1) + \sum_{j \in P_*} \sqrt{K}(\|\beta^j + \gamma^j\|_2 - \|\beta^j\|_2) \right\} + \\ \lambda_2 \left\{ \sum_{j \in P_*} \eta(\|\beta^j_{H_j} + \gamma^j_{H_j}\|_1 - \|\beta^j H_j)\| + \sum_{j \in P_*} \eta\|\gamma^j H_j^C\|_1 \right\} \tag{39}$$

The matrix Γ^λ is the bias of the sparse multi-site Lasso estimator. We derive first a bound on the Frobenius norm of Γ^λ.

Lemma 3.4. *Assume conditions in Theorem3.2. The Frobenius norm of Γ^λ is then bounded for sufficiently large values of $\bar{n}$, given a constant $\omega_1 > 0$, by*

$$\|\Gamma^\lambda\|_F^2 \leq \omega_1 \sigma^2 \frac{k\bar{s}\log(kp)}{\bar{n}} \tag{40}$$

Proof. $f(\Gamma) = 0$ whenever $\Gamma = 0$ following the definition from (39). For the true solution Γ^λ, it follows hence that $f(\Gamma^\lambda) \leq 0$. For notational simplicity, we drop the super-script λ from here on. Using $\gamma^T C\gamma \geq 0$, we have

$$\left\{ \sum_{j \in P_*^C} (\sqrt{k}\|\gamma^j\|_2) + \sum_{j \in P_*^C} (\eta\|\gamma^j\|_1) + \sum_{j \in P_*} \eta\|\gamma^j_{H_j^C}\|_1 \right\} \leq \left\{ \sum_{j \in P_*} \sqrt{k}\|\gamma^j\|_2 + \sum_{j \in P_*} \eta\|\gamma^j_{H_j}\|_1 \right\} \tag{41}$$

Since $|P_*| = s_p$, $\sum_{j \in P_*} |H_j| = s_h$. It follows that $\sum_{j \in P_*} \|\gamma^j\|_2 \leq \sqrt{s_p}\|\gamma\|_2$, $\sum_{j \in P_*} \|\gamma^j_{H_j}\|_1 \leq \sqrt{s_h}\|\gamma\|_2$, and hence, using (41),

$$\Lambda(\Gamma) \leq 2\{(1-\alpha)\sqrt{ks_p} + \alpha\sqrt{s_h}\}\|\gamma\|_2 = 2\sqrt{k\bar{s}}\|\gamma\|_2 \tag{42}$$

Using $f(\Gamma) \leq 0$ again and (42), it follows that

$$\bar{n}\gamma^T C\gamma \leq 2\lambda\sqrt{k\bar{s}}\|\gamma\|_2 \tag{43}$$

Now consider $\gamma^T C\gamma$. Bounding this term from below and plugging the result into (42) will yield the desired upper bound on the Frobenius norm of Γ. Let $\|\gamma^{(1)}\| \geq \|\gamma^{(2)}\| \geq ... \geq \|\gamma^{(p)}\|$ be the ordered columns of Γ. Let u_n for $n \in N$ be a sequence of positive integers, to be chosen later, and define $U = \{j : \|\gamma^j\|_2 \geq \|\gamma^{(u_n)}\|_2\}$. Define $\gamma(U)$ and $\gamma(U^C)$ by setting $\gamma^j(U) = \gamma^j 1\{i \notin U\}$ and $\gamma^j(U^C) = \gamma^j 1\{i \in U\}$, followed by unfolding Γ. Then quantity $\gamma^T C\gamma$ can be written as $\gamma^T C\gamma = \|a + b\|_2^2$, where $a := \bar{n}^{-1/2}X\gamma(U)$, $b := \bar{n}^{-1/2}X\gamma(U^C)$, $X = DIAG(X_1, ..., X_k)$. Then

$$\gamma^T C\gamma = \|a + b\|_2^2 \geq (\|a\|_2 - \|b\|_2)^2 \tag{44}$$

Before proceeding, we need to bound the norm $\|\gamma(U^C)\|_2$ as a function of u_n. Assume $l = \sum_{j=1}^p \|\gamma^j\|_2$. It holds for every $j = 1, ..., p$ that $\|\gamma^{(j)}\|_2 \leq l/j$. Hence,

$$\|\gamma(U^C)\|_2^2 \leq (\sum_{j=1}^p \|\gamma^j\|_2)^2 \sum_{j=u_n+1}^p \frac{1}{j^2} \tag{45}$$

Therefore, we have

$$\|\gamma(U^C)\|_2 \leq \sum_{j=1}^p \|\gamma^j\|_2 \sqrt{\frac{1}{u_n}} \leq \|\gamma\|_1 \sqrt{\frac{1}{u_n}} \tag{46}$$

Based on (42), $\Lambda(\Gamma) = (1-\alpha)\sqrt{k}\sum_{j=1}^p \|\gamma^j\|_2 + \alpha\|\gamma\|_1$, and (46), it follows that

$$\|\gamma(U^C)\|_2^2 \leq 4\|\gamma\|_2^2 \left\{ \frac{1}{u_n} \left(\frac{\sqrt{k\bar{s}}}{(1-\alpha)\sqrt{k} + \alpha} \right)^2 \right\} \tag{47}$$

By definition, since $\gamma(U)$ has only u_n nonzero groups,

$$\|a\|_2^2 = \|\gamma(U)^T C\gamma(U)\|_2^2 \geq \phi_{\min}(u_n)\|\gamma(U)\|_2^2 \geq \\ \phi_{\min}(u_n)\|\gamma\|_2^2 \left(1 - 4\left\{ \frac{1}{u_n} \left(\frac{k\bar{s}}{(1-\alpha)\sqrt{k} + \alpha} \right)^2 \right\} \right) \tag{48}$$

Here we explain why we obtain $\phi_{\min}(u_n)$ instead of $\phi_{\min}(ku_n)$. We denote $\phi^i_{\min}(m)$ to be m-sparse of $\bar{n}^{-1}X_i^TX_i$. Then $\phi_{\min}(m) = \min_{i=1}^k \phi^i_{\min}(m)$ because of block structure. Since we have u_n nonzero groups, instead of arbitrary ku_n nonzero elements, we obtain a higher value $\phi_{\min}(u_n) = \min_{i=1}^k \phi^i_{\min}(u_n)$ instead of $\phi_{\min}(ku_n)$. This is the one place where we consider the block structure of multi-site design.

As $\gamma(U^C)$ has at most $\min\{\sum_{i=1}^k n_i, kp\}$ nonzero groups, using again (47), (42) and the block structure of multi-site design,

$$\|b\|_2^2 \leq 4\phi_{\max}(\min\{\sum_{i=1}^k n_i, kp\})\|\gamma\|_2^2 \left\{\frac{1}{u_n}\left(\frac{\sqrt{k\bar{s}}}{(1-\alpha)\sqrt{k}+\alpha}\right)^2\right\} \tag{49}$$

Using (49), (48) and (44), along with $\phi_{\max}(\min\{\sum_{i=1}^k n_i, kp\}) \geq \phi_{\min}(u_n)$,

$$\gamma^TC\gamma \geq \phi_{\min}(u_n)\|\gamma\|_2^2 \times \left(1 - 4\sqrt{\frac{\phi_{\max}(\min\{\sum_{i=1}^k n_i, kp\})}{\phi_{\min}(u_n)}\left\{\frac{1}{u_n}(\frac{\sqrt{k\bar{s}}}{(1-\alpha)\sqrt{k}+\alpha})^2\right\}}\right) \tag{50}$$

Using conditions in Theorem 3.2 and setting $u_n = \left(\frac{\sqrt{k\bar{s}}}{(1-\alpha)\sqrt{k}+\alpha}\right)^2$, it follows that

$$\gamma^TC\gamma \geq \rho_{\min}\left(1 - 4\sqrt{\frac{\rho_{\max}}{\rho_{\min}}}\right)\|\gamma\|_2^2 \tag{51}$$

Using this result together with (43), which says that $\gamma^TC\gamma \leq 2\bar{n}^{-1}\lambda\sqrt{k\bar{s}}\|\gamma\|_2$, we have the following for large $\bar{n}$,

$$\|\Gamma\|_F^2 = \|\gamma\|_2^2 \leq \frac{1}{(\rho_{\min} - 4\sqrt{\rho_{\min}\rho_{\max}})^2}\frac{\lambda^2 k\bar{s}}{\bar{n}^2} \tag{52}$$

The proof of Lemma 3.4 is completed by noticing λ in Theorem 3.2. □

3.1.2. Part II of proof – Dealing with variance

The proof for the variance part is two-fold. We first derive a bound on the variance, which is a function of the number of nonzero groups. We then bound the number of nonzero groups, taking into account the bound on the bias derived above.

Variance of restricted OLS: Before considering the sparse multi-site Lasso estimator, a trivial bound is shown for the variance of a restricted OLS estimation. For every subset $\psi \subset \{1, , p\}$, we use it to select a subset of columns from design matrix X_i for task i. These columns form a matrix $X_{i\psi}$. Define $X_\psi = DIAG(X_{1\psi}, X_{2\psi}, ..., X_{k\psi})$, and the restricted OLS-estimator with the noise vector $\epsilon^T = (\epsilon_1, ..., \epsilon_k)^T$ is

$$\hat{\theta}^\psi = (X_\psi^TX_\psi)^{-1}X_\psi^T\epsilon \tag{53}$$

The ℓ_2-norm of this estimator can be bounded.

Lemma 3.5. *Let m_p be a sequence with $m_p = o(\bar{n})$ and $m_p \to \infty$ for $\bar{n} \to \infty$. It holds with probability converging to 1 for $n \to \infty$*

$$\max_{\psi:|\psi|\leq m_p}\|\hat{\theta}^\psi\|_2^2 \leq \frac{2\log kp}{\bar{n}}\frac{km_p}{\phi^2_{\min}(m_p)}\sigma^2 \tag{54}$$

Proof. We refer the readers to Lemma 3 in (Meinshausen & Yu, 2009) and Lemma 3 in (Liu & Zhang, 2009) for the proof. Here, we again use block design structure of multi-site problem, the same as in (48), to obtain $\phi_{\min}(m_p)$ instead of $\phi_{\min}(km_p)$. □

The variance of the sparse multi-site Lasso estimator can be bounded by the variance of restricted OLS estimators, using bounds on the number of active groups.

Lemma 3.6. *If, for a fixed value of λ, the number of nonzero groups of de-noised estimators $\hat{B}^{\lambda,\xi}$ is for every $0 \le \xi \le 1$ bounded by m, then*

$$\sup_{0\le\xi\le1} \|\hat{B}^{\lambda,0} - \hat{B}^{\lambda,\xi}\|_F^2 \le \mathcal{C} \max_{\psi:|\psi|\le m} \|\hat{\theta}^{\psi}\|_2^2 \tag{55}$$

with $\mathcal{C}$ as a generic constant.

Proof. We refer the readers to Lemma 4 and Lemma 5 in (Liu & Zhang, 2009) for the proof. □

Let $A^P_{\lambda,\xi}$ be the set of variables in nonzero groups of the de-noised estimator $\hat{B}^{\lambda,\xi}$. Define m_p to be the largest number of nonzero groups over all values of $0 \le \xi \le 1$. Then we have $km_p = \sup_{0\le\xi\le1} |A^P_{\lambda,\xi}|$.

Lemma 3.7. *Given $0 \le \alpha \le 0.5$, we have*

$$|A^P_{\lambda,\xi}|\lambda^2(1-2\alpha)^2 \le \|2X^T_{A^p_{\lambda,\xi}}(Y - X\hat{\beta}^{\lambda,\xi})\|_2^2 \tag{56}$$

where we defined before that $X = DIAG(X_1, ..., X_k)$, $Y^T = (Y_1^T, ..., Y_k^T)$. $\hat{\beta}^{\lambda,\xi}$ is the transpose of unfolded vector of $\hat{B}^{\lambda,\xi}$ by rows. $X_{A^p_{\lambda,\xi}}$ is X_ψ when $\psi = A^p_{\lambda,\xi}$

Proof. The conditions for the solution of sparse multi-site Lasso are presented in (Simon et al., 2013). We use $\hat{\beta}$ rather than $\hat{\beta}^{\lambda,\xi}$ for notational simplicity in this proof. We continue to use our notation $\hat{\beta}^j$ to refer the j-th column (here it is a group) of $\hat{B}$, and $\hat{\beta}^j_i$ to refer the i-th element (task) in $\hat{\beta}^j$. We define $X^j = DIAG(X_1^j, ..., X_k^j)$ and X_i^j to be the j-th column of X_i for task i. In other words, we allow for $(k-1)p$ number of 0 in X_i^j.

$$\begin{aligned}
&-2{X_i^j}^T(Y - X\hat{\beta}) + \lambda\left\{\alpha\frac{\hat{\beta}^j_i}{\|\hat{\beta}^j_i\|_2} + (1-\alpha)\frac{\hat{\beta}^j_i}{\|\hat{\beta}^j\|_2/\sqrt{k}}\right\} = 0, \text{ when } \hat{\beta}^j_i \ne 0,\ \hat{\beta}^j \ne 0,\\
&-2{X_i^j}^T(Y - X\hat{\beta}) + \lambda(1-\alpha)\frac{\hat{\beta}^j_i}{\|\hat{\beta}^j\|_2/\sqrt{k}} = \lambda\alpha v_i^j, \text{ with } \|v_i^j\|_2 \le 1, \text{ when } \hat{\beta}^j_i = 0,\ \hat{\beta}^j \ne 0,\\
&\left\|-2{X^j}^T(Y - X\hat{\beta})\right\|_2 \le \lambda\sqrt{k}, \text{ when } \hat{\beta}^j = 0.
\end{aligned} \tag{57}$$

Let $D^P_{\lambda,\xi} = \{j \in 1, 2, ..., p|\text{group j is active for } \hat{B}^{\lambda,\xi}\}$. For each j in $D^P_{\lambda,\xi}$, we define $\hat{\beta}^j_*$ to be the vector of all $\hat{\beta}^j_i \ne 0$. Their corresponding columns X_i^js from X^j, would form a matrix X^j_*. For each j in $D^P_{\lambda,\xi}$, we define $\hat{\beta}^j_{*C}$ to be the vector of all $\hat{\beta}^j_i = 0$. Their corresponding columns X_i^js from X^j, would form a matrix X^j_{*C}. Then, from (57),

$$\sum_{j=1}^{D^P_{\lambda,\xi}} \|2{X_*^j}^T(Y - X\hat{\beta})\|_2^2 \ge \lambda^2(1-\alpha)^2 k \sum_{j=1}^{D^P_{\lambda,\xi}} \frac{\|\hat{\beta}^j_*\|_2^2}{\|\hat{\beta}^j\|_2^2} \tag{58}$$

Based on the fact that $\|a+b\|_2^2 \ge (\|a\|_2 - \|b\|_2)^2$

$$\begin{aligned}
\sum_{j=1}^{D^P_{\lambda,\xi}} \|2{X^j_{*C}}^T(Y - X\hat{\beta})\|_2^2 &\ge \sum_{j=1}^{D^P_{\lambda,\xi}} \left(\lambda(1-\alpha)\sqrt{k}\frac{\|\hat{\beta}^j_{*C}\|_2}{\|\hat{\beta}^j\|_2} - \lambda\alpha\|v^j_{*C}\|_2\right)^2\\
&= \sum_{j=1}^{D^P_{\lambda,\xi}} \left\{\lambda^2(1-\alpha)^2 k\frac{\|\hat{\beta}^j_{*C}\|_2^2}{\|\hat{\beta}^j\|_2^2} + \lambda^2\alpha^2\|v^j_{*C}\|_2^2 - 2\lambda^2\alpha(1-\alpha)\sqrt{k}\frac{\|\hat{\beta}^j_{*C}\|_2}{\|\hat{\beta}^j\|_2}\|v^j_{*C}\|_2\right\}\\
&\ge \sum_{j=1}^{D^P_{\lambda,\xi}} \left\{\lambda^2(1-\alpha)^2 k\frac{\|\hat{\beta}^j_{*C}\|_2^2}{\|\hat{\beta}^j\|_2^2} + \lambda^2\alpha^2\|v^j_{*C}\|_2^2 - \lambda^2\alpha(1-\alpha)\left[k\frac{\|\hat{\beta}^j_{*C}\|_2^2}{\|\hat{\beta}^j\|_2^2} + \|v^j_{*C}\|_2^2\right]\right\}\\
&= \lambda^2(1-\alpha)(1-2\alpha)k\sum_{j=1}^{D^P_{\lambda,\xi}} \frac{\|\hat{\beta}^j_{*C}\|_2^2}{\|\hat{\beta}^j\|_2^2} - \lambda^2\alpha(1-2\alpha)\sum_{j=1}^{D^P_{\lambda,\xi}} \|v^j_{*C}\|_2^2
\end{aligned} \tag{59}$$

Based on (58) and (59), we have

$$\|2X_{A^p_{\lambda,\xi}}^T(Y - X\hat{\beta})\|_2^2 = \sum_{j=1}^{D^P_{\lambda,\xi}} \|2{X^j}^T(Y - X\hat{\beta})\|_2^2 = \sum_{j=1}^{D^P_{\lambda,\xi}} \|2{X^j_*}^T(Y - X\hat{\beta})\|_2^2 + \sum_{j=1}^{D^P_{\lambda,\xi}} \|2{X^j_{*^C}}^T(Y - X\hat{\beta})\|_2^2 \tag{60}$$

$$\geq \lambda^2(1-\alpha)^2 k \sum_{j=1}^{D^P_{\lambda,\xi}} \frac{\|\hat{\beta}^j_*\|_2^2}{\|\hat{\beta}^j\|_2^2} + \lambda^2(1-\alpha)(1-2\alpha)k \sum_{j=1}^{D^P_{\lambda,\xi}} \frac{\|\hat{\beta}^j_{*^C}\|_2^2}{\|\hat{\beta}^j\|_2^2} - \lambda^2\alpha(1-2\alpha) \sum_{j=1}^{D^P_{\lambda,\xi}} \|v^j_{*^C}\|_2^2 \tag{61}$$

$$\geq \lambda^2(1-\alpha)(1-2\alpha)k \sum_{j=1}^{D^P_{\lambda,\xi}} \frac{\|\hat{\beta}^j_*\|_2^2 + \|\hat{\beta}^j_{*^C}\|_2^2}{\|\hat{\beta}^j\|_2^2} - \lambda^2\alpha(1-2\alpha) \sum_{j=1}^{D^P_{\lambda,\xi}} \|v^j_{*^C}\|_2^2 \tag{62}$$

$$\geq \lambda^2(1-\alpha)(1-2\alpha)k|D^P_{\lambda,\xi}| - \lambda^2\alpha(1-2\alpha)k|D^P_{\lambda,\xi}| \tag{63}$$

$$= \lambda^2(1-2\alpha)^2 k|D^P_{\lambda,\xi}| = \lambda^2(1-2\alpha)^2|A^P_{\lambda,\xi}| \tag{64}$$

□

The next lemma provides an asymptotic upper bound on the number of selected variables, the proof of which is similar to Lemma 5 in (Meinshausen & Yu, 2009).

Lemma 3.8. *Assume conditions in Theorem 3.2, with probability converging to 1 for $n \to \infty$,*

$$\sup_{0\leq\xi\leq 1} |A^P_{\lambda,\xi}| \leq \left\{\left(1 + \frac{\alpha}{1-2\alpha}\right)\sqrt{ks_p} + \frac{\alpha}{1-2\alpha}\sqrt{s_h}\right\}^2 \tag{65}$$

Proof. Based on Lemma 3.7,

$$(1-2\alpha)^2 km_p = (1-2\alpha)^2 \sup_{0\leq\xi\leq 1} |A^P_{\lambda,\xi}| \leq \frac{1}{\lambda^2} \sup_{0\leq\xi\leq 1} \|2X^T_{A^p_{\lambda,\xi}}(Y - X\hat{\beta}^{\lambda,\xi})\|_2^2 \tag{66}$$

We decompose the right side into two parts and then have

$$(1-2\alpha)^2 km_p \leq \left(\frac{1}{\lambda} \sup_{0\leq\xi\leq 1} \|2X^T_{A^p_{\lambda,\xi}} X(\beta^* - \hat{\beta}^{\lambda,\xi})\|_2 + \frac{1}{\lambda} \sup_{0\leq\xi\leq 1} \|2X^T_{A^p_{\lambda,\xi}}\epsilon\|_2\right)^2 \tag{67}$$

Similarly, we know from proof in Lemma 3.5 that

$$\sup_{0\leq\xi\leq 1} \|2X^T_{A^p_{\lambda,\xi}}\epsilon\|_2^2 \leq 2km_p \log(kp)\sigma^2\bar{n} \tag{68}$$

Based on the definition of λ, there exists a constant $\varpi_1 > 0$, such that

$$\frac{\sup_{0\leq\xi\leq 1}\|2X^T_{A^p_{\lambda,\xi}}\epsilon\|_2^2}{\lambda^2} \leq \varpi_1^2 km_p \tag{69}$$

Therefore, we have

$$(1-2\alpha)^2 km_p \leq \left(\frac{1}{\lambda} \sup_{0\leq\xi\leq 1} \|2X^T_{A^p_{\lambda,\xi}} X(\beta^* - \hat{\beta}^{\lambda,\xi})\|_2 + \varpi_1\sqrt{km_p}\right)^2 \tag{70}$$

Define $F^P_{\lambda,\xi} = \{i : \beta^*_i \neq 0\} \cup A^P_{\lambda,\xi}$. Based on the block trick we used in proof of Lemma 3.4,

$$\|X^T_{A^p_{\lambda,\xi}} X(\beta^* - \hat{\beta}^{\lambda,\xi})\|_2^2 \leq \|X^T_{F^p_{\lambda,\xi}} X_{F^p_{\lambda,\xi}}(\beta^* - \hat{\beta}^{\lambda,\xi})\|_2^2 \leq \bar{n}^2\phi^2_{\max}(s_p + \min\{\sum_{i=1}^k n_i, kp\})\|\beta^* - \hat{\beta}^{\lambda,\xi}\|_2^2 \tag{71}$$

From the assumption on $\phi_{\max}(s_p + \min\{\sum_{i=1}^k n_i, kp\})$, we know

$$\|X^T_{A^p_{\lambda,\xi}} X(\beta^* - \hat{\beta}^{\lambda,\xi})\|_2^2 \leq \bar{n}^2\rho^2_{\max}\|\beta^* - \hat{\beta}^{\lambda,\xi}\|_2^2 \tag{72}$$

Therefore, we have

$$(1-2\alpha)^2 km_p \le \left(\frac{2}{\lambda}\bar{n}\rho_{\max} \sup_{0\le\xi\le1} \|\beta^* - \hat{\beta}^{\lambda,\xi}\|_2 + \varpi_1\sqrt{km_p}\right)^2 \tag{73}$$

$$\le \left(\frac{2}{\lambda}\bar{n}\rho_{\max}\|\beta^* - \hat{\beta}^{\lambda,0}\|_2 + \frac{2}{\lambda}\bar{n}\rho_{\max} \sup_{0\le\xi\le1} \|\hat{\beta}^{\lambda,0} - \hat{\beta}^{\lambda,\xi}\|_2 + \varpi_1\sqrt{km_p}\right)^2 \tag{74}$$

Because β is the unfolded vector of B, actually $\sup_{0\le\xi\le1}\|\hat{\beta}^{\lambda,0} - \hat{\beta}^{\lambda,\xi}\|_2 = \sup_{0\le\xi\le1}\|\hat{B}^{\lambda,0} - \hat{B}^{\lambda,\xi}\|_F$. From Lemmas 3.5 and 3.6, definition of λ and the assumption on $\phi_{\min}$, we obtain the bound

$$\frac{4\bar{n}^2\rho_{\max}^2}{\lambda^2}\sup_{0\le\xi\le1}\|\hat{\beta}^{\lambda,0} - \hat{\beta}^{\lambda,\xi}\|_2^2 \le \mathcal{C}\frac{4\bar{n}^2\rho_{\max}^2}{\lambda^2}\frac{2\log(kp)}{\bar{n}}\frac{km_p}{\phi_{\min}^2(m_p)}\sigma^2 \le \varpi_2^2 km_p \tag{75}$$

Here, $\varpi_2 > 0$ is a constant. We define $\varpi = \varpi_1 + \varpi_2$. Now, we obtain

$$(1-2\alpha)^2 km_p \le \left(\frac{2}{\lambda}\bar{n}\rho_{\max}\|\beta^* - \hat{\beta}^{\lambda,0}\|_2 + \varpi\sqrt{km_p}\right)^2 \tag{76}$$

By setting the constant term in λ large enough, we can have $\varpi/(1-2\alpha) \le 5\varpi \le 0.026$, and hence

$$km_p \le (18/17.5)^2(2\rho_{\max})^2\frac{\bar{n}^2\|\beta^* - \hat{\beta}^{\lambda,0}\|_2^2}{(1-2\alpha)^2\lambda^2} \le \left\{\left(1+\frac{\alpha}{1-2\alpha}\right)\sqrt{ks_p} + \frac{\alpha}{1-2\alpha}\sqrt{s_h}\right\}^2 \tag{77}$$

The last inequality is obtained by plugging in Lemma 3.4. The constant can be 1 by setting the constant term in λ large enough. □

Follow from Lemmas 3.5,3.6, and 3.8, the next lemma bounds the variance part of the sparse multi-sites Lasso estimator:

Lemma 3.9. *Assume conditions in Theorem3.2, there exists a constant $\omega_2 > 0$, with probability converging to 1 for $n \to \infty$,*

$$\|B^\lambda - \hat{B}^{\lambda,1}\|_F^2 = \|\hat{B}^{\lambda,0} - \hat{B}^{\lambda,1}\|_F^2 \le \omega_2\sigma^2\frac{k\bar{s}\log(kp)}{\bar{n}} \tag{78}$$

Proof. We have defined B^λ as the estimator $\hat{B}^{\lambda,0}$ with no noise before Lemma 3.4.
Based on Lemmas 3.5 and 3.6

$$\|\hat{B}^{\lambda,0} - \hat{B}^{\lambda,1}\|_F^2 \le \frac{2\log kp}{\bar{n}}\frac{km_p}{\phi_{\min}^2(m_p)}\sigma^2 \tag{79}$$

Based on Lemma 3.8, assumption on $\phi_{\min}$ and $0 \le \alpha \le 0.4$,

$$\|\hat{B}^{\lambda,0} - \hat{B}^{\lambda,1}\|_F^2 \le \frac{2\log kp}{\bar{n}}\frac{km_p}{\phi_{\min}^2(m_p)}\sigma^2 \le \omega_2\sigma^2\frac{k\bar{s}\log(kp)}{\bar{n}} \tag{80}$$

□

The lemma 3.4 and 3.9 together complete the proof of Theorem 3.2

3.2. Proof of Theorem 3.3:

The proof is similar to that of Theorem 3.2. Recall that in this case, however, we do not penalize $\sqrt{k}$ on group penalty. Hence, we have the following result about bias contribution of Theorem 3.3.

Lemma 3.10. *Assume conditions in Theorem 3.3. The Frobenius norm of Γ^λ is then bounded for sufficiently large values of $\bar{n}$, given a constant $\omega_1 > 0$, by*

$$\|\Gamma^\lambda\|_F^2 \le \omega_1\sigma^2\frac{k\tilde{s}\log(kp)}{\bar{n}} \tag{81}$$

Proof. The proof procedure is same as Lemma 3.4. But instead of (42), we now have

$$\Lambda(\Gamma^\lambda) \leq 2\{(1-\tilde{\alpha})\sqrt{s_p} + \tilde{\alpha}\sqrt{s_h}\}\|\gamma^\lambda\|_2 = 2\sqrt{k\tilde{s}}\|\gamma^\lambda\|_2 \tag{82}$$

because we do not have $\sqrt{k}$ penalization on group penalty. Hence, in Lemma 3.10, we have $\tilde{s} = \{(1-\tilde{\alpha})\sqrt{s_p/k} + \tilde{\alpha}\sqrt{s_h/k}\}^2$, instead of $\bar{s} = \{(1-\tilde{\alpha})\sqrt{s_p} + \tilde{\alpha}\sqrt{s_h/k}\}^2$. □

For restricted OLS estimation, we redefine few things here. For every subset $\psi \subset \{1, ..., kp\}$ with $|\psi| \leq \sum_{i=1}^{k} n_i$, we define X_ψ to be the combination of columns from design matrix X, where $X = DIAG(X_1, X_2, ..., X_k)$. The restricted OLS-estimator of the noise vector $\epsilon^T = (\epsilon_1, ..., \epsilon_k)^T$ is then given by,

$$\hat{\theta}^\psi = (X_\psi^T X_\psi)^{-1} X_\psi^T \epsilon \tag{83}$$

For the variance contribution, the proof is similar to that of Theorem 3.2. We present the required Lemmas for Theorem 3.3 here.

Lemma 3.11. *Let m_n be a sequence with $m_n = o(k\bar{n})$ and $m_n \to \infty$ for $\bar{n} \to \infty$. It holds with probability converging to 1 for $n \to \infty$*

$$\max_{\psi:|\psi|\leq m_n} \|\hat{\theta}^\psi\|_2^2 \leq \frac{2\log kp}{\bar{n}} \frac{m_n}{\phi_{\min}^2(m_n)} \sigma^2 \tag{84}$$

Lemma 3.12. *If, for a fixed value of λ, the number of active variables of de-noised estimators $\hat{B}^{\lambda,\xi}$ is for every $0 \leq \xi \leq 1$ bounded by m, then*

$$\sup_{0\leq\xi\leq1} \|\hat{B}^{\lambda,0} - \hat{B}^{\lambda,\xi}\|_F^2 \leq \mathcal{C} \max_{\psi:|\psi|\leq m} \|\hat{\theta}^\psi\|_2^2 \tag{85}$$

with $\mathcal{C}$ as a generic constant.

Let $A_{\lambda,\xi}^1$ be the set of active variables of the de-noised estimator $\hat{B}^{\lambda,\xi}$. Let m_n to be the largest number of active variables over all values of $0 \leq \xi \leq 1$. Then we have $m_n = \sup_{0\leq\xi\leq1} |A_{\lambda,\xi}^1|$.

Lemma 3.13. *For any $0 \leq \tilde{\alpha} \leq 1$, we have*

$$|A_{\lambda,\xi}^1|\lambda^2\tilde{\alpha}^2 \leq \|2X_{A_{\lambda,\xi}^1}^T (Y - X\hat{\beta}^{\lambda,\xi})\|_2^2 \tag{86}$$

where we defined before that $X = DIAG(X_1, ..., X_k)$, $Y^T = (Y_1^T, ..., Y_k^T)$. $\hat{\beta}^{\lambda,\xi}$ is the transpose of unfolded vector of $\hat{B}^{\lambda,\xi}$ by rows. $X_{A_{\lambda,\xi}^1}$ is X_ψ when $\psi = A_{\lambda,\xi}^1$

Lemma 3.14. *Assume conditions in Theorem 3.3, with probability converging to 1 for $n \to \infty$,*

$$\sup_{0\leq\xi\leq1} |A_{\lambda,\xi}^1| \leq \left\{\sqrt{s_h} + \frac{1-\tilde{\alpha}}{\tilde{\alpha}}\sqrt{s_p}\right\}^2 \tag{87}$$

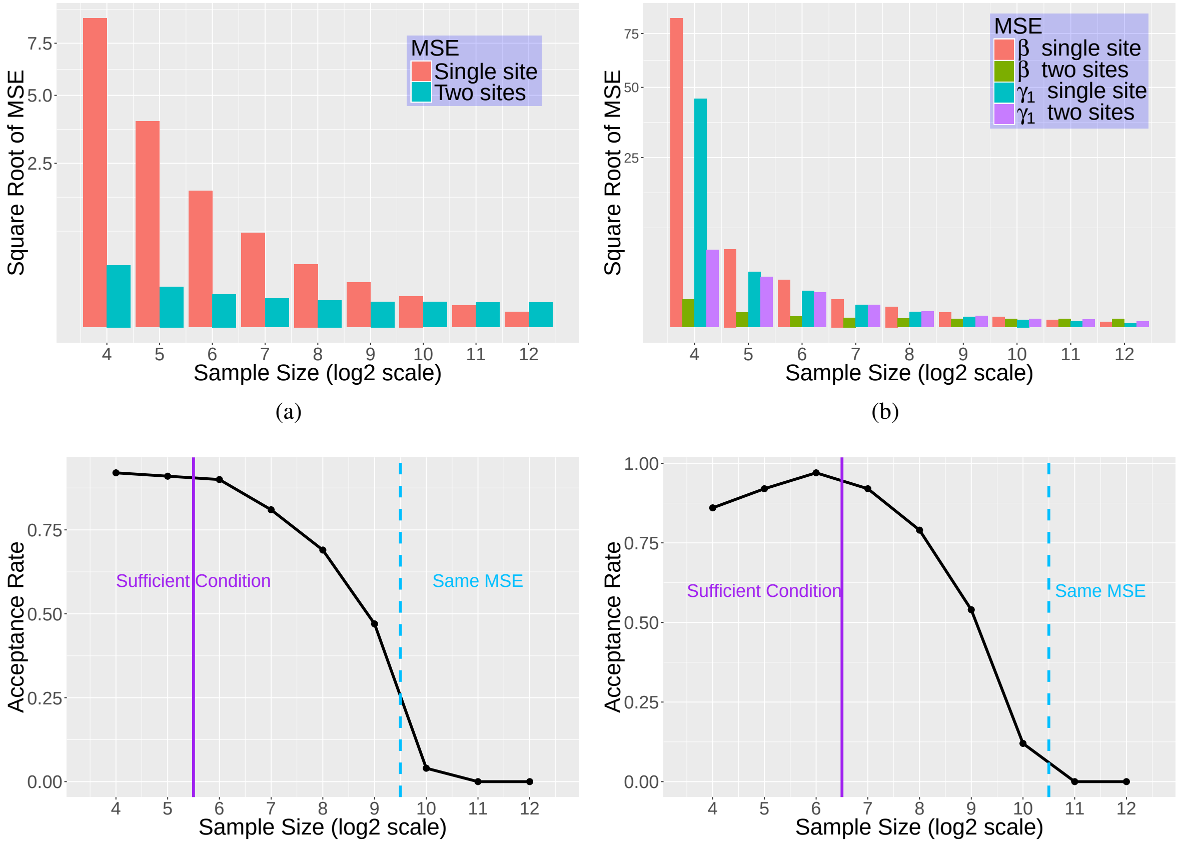
Lemma 3.15. *Assume conditions in Theorem3.3, there exists a constant $\omega_2 > 0$, with probability converging to 1 for $n \to \infty$,*

$$\|B^\lambda - \hat{B}^{\lambda,1}\|_F^2 = \|\hat{B}^{\lambda,0} - \hat{B}^{\lambda,1}\|_F^2 \leq \omega_2\sigma^2 \frac{k\tilde{s}\log(kp)}{\bar{n}} \tag{88}$$

Lemma 3.10 and Lemma 3.15 complete the proof of Theorem 3.3

3. Extra set of simulations (corresponding to Section 4.1 in the main paper)

3.1. Hypothesis Test Simulation when $p = 6$



(a)

(b)

(c)

(d)

Figure 1. The figure is similar to the simulations done in Figure 3 (which is also the one Figure 3 in the main paper). However, here the dimension p of β is 6 instead of 3. (a,c) are MSE of $\hat{\beta}$ and the corresponding acceptance rate of our hypothesis test (from Section2.1). (b,d) are MSE of $\hat{\beta}$ and $\hat{\gamma}_1$ and the corresponding acceptance rate (from Section2.2). These are based on 100 bootstrap repetitions. The solid line in (c,d) represents the point where the condition from Theorem 2.3 is equal to 1. The dotted line is when MSE of $\hat{\beta}$ is the same for single-site and multi-site models.

3.2. Sparse Multi-Sites Lasso Simulation

Table 1. Add multi-sites Lasso on Lasso.

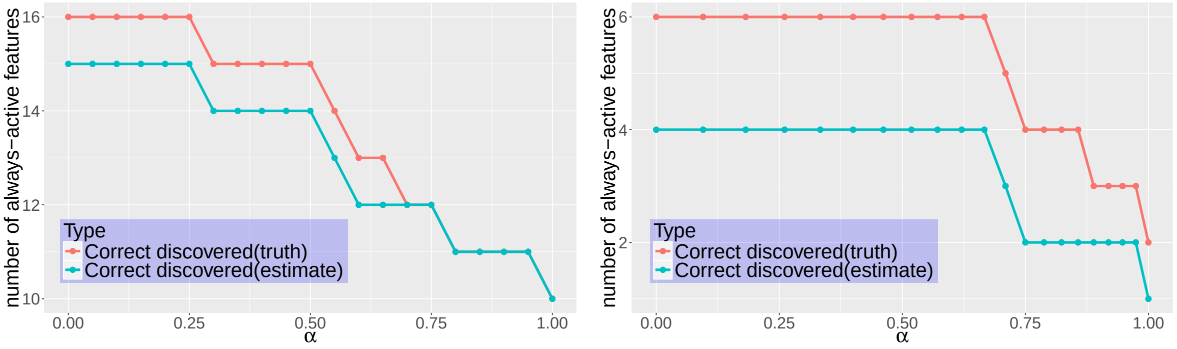
α	0	0.05	0.95	0.97 (OUR)	1
CDR	0.1423	0.1463	0.2747	0.2863	0.2955
CDV	78	78	75	75	73
CDG	5	5	3	3	1

We report correctly discovered number of active variables (CDV), ratio of CDV and total number of discovered variables (CDR), and correctly discovered number of always-active features (CDG).
From Table1 and Table 2 we see that our chosen α helps sparse multi-sites Lasso to discover more or preserve always-active features. The number and rate of correctly discovered number of active variables given by our chosen α are also among the best.

Table 2. Add Lasso on multi-sites Lasso.

α	0	0.05	0.25 (OUR)	0.95	1
CDR	0.2292	0.2381	0.2453	0.2841	0.2885
CDV	80	80	79	75	73
CDG	16	16	15	11	11

3.3. Figure Examples for Choosing α



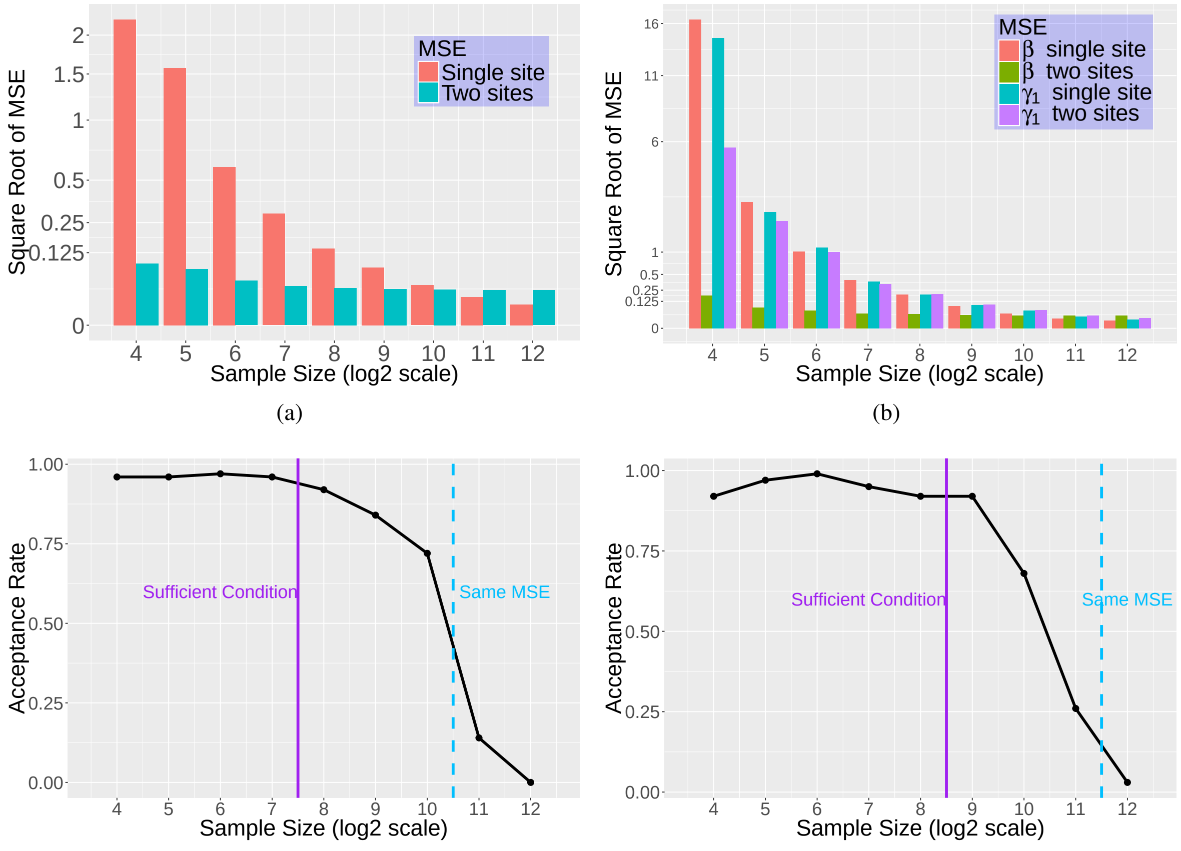
(a)

(b)

Figure 2. These plots show that *site-active* set from simultaneous inference provides information of always-active features (which is then used to choose the hyper-parameters α an λ). In (a), we add Lasso on multi-sties Lasso, and $\alpha = 0.25$ is chosen. Similarly, in Figure(b), we add multi-sites Lasso on Lasso, and $\alpha = 0.97$ is chosen.

We here point out a caveat about our choice of α when sparsity patterns share few features and always-active features exist. In this setting, we do want to discover more always-active features. Hence, we decrease α from 1 and stop at the point where we just select one more always-active feature. In other words, we choose the α left to the one described in main body.

4. Longer version of Section 4 from the main paper



(a)

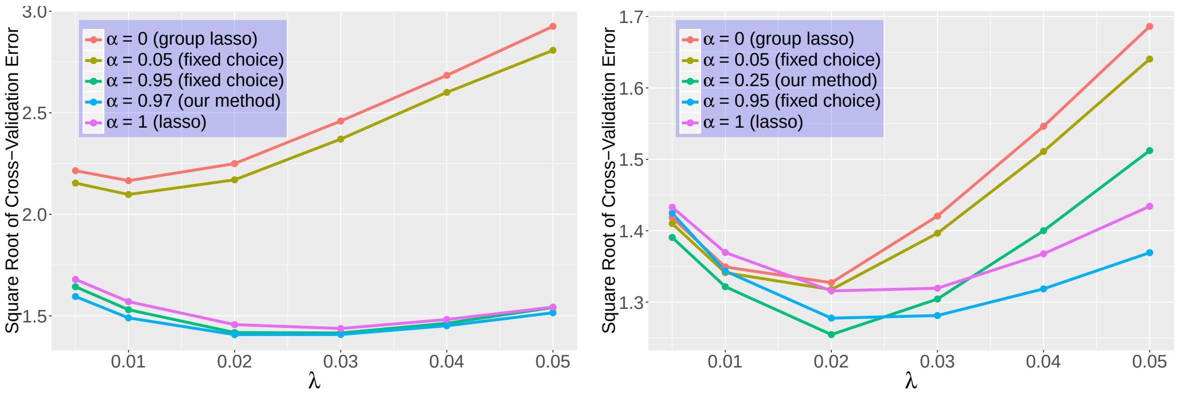
(b)

(c)

(d)

Figure 3. (a,c) are MSE of $\hat{\beta}$ and the corresponding acceptance rate of our hypothesis test (from Section2.1). (b,d) are MSE of $\hat{\beta}$ and $\hat{\gamma}_1$ and the corresponding acceptance rate from Section 2.2). These are based on 100 bootstrap repetitions. The solid line in (c,d) represents the point where the condition from Theorem 2.3 is equal to 1. The dotted line is when MSE of $\hat{\beta}$ is the same for single-site and multi-site models.

We now provide few more details about the different curves observed in Figure 3, beyond what is reported in the main paper due to space constraint. First, we check whether the gap between the sufficient condition (from Theorem 2.3) and the point where single-site and multi-site models have same MSE is small. The solid lines in Figure 3(c,d) correspond to the point where the condition value defined in Theorem 2.3 is equal to 1. The dotted lines (where condition value is approximately 3.3) are the points where the MSE of multi-site model starts to increase above the MSE of single-site one. In other words, to the left of the dotted lines that MSE of $\hat{\beta}$ from multi-sites model is smaller than single-site model. To the right of these lines it is larger. We see that the gap is reasonably small. We then check the type I error of our hypothesis test. On the left side of solid lines, the sufficient condition holds and our hypothesis test accepts the combination with high rate around 95%, i.e., the type I error is well-controlled. Further, the power of our hypothesis test is evident when MSE of $\hat{\beta}$ from multi-sites model is worse than single-site model. Though our sufficient condition is conservative for the combination, by noticing that χ^2 test is progressive, our test has a high power on the right side of dotted line. In the regime between the two lines, the multi-sites model has slightly better MSE of $\hat{\beta}$ compared to single-site model, and our hypothesis test accepts the combination with high rate.

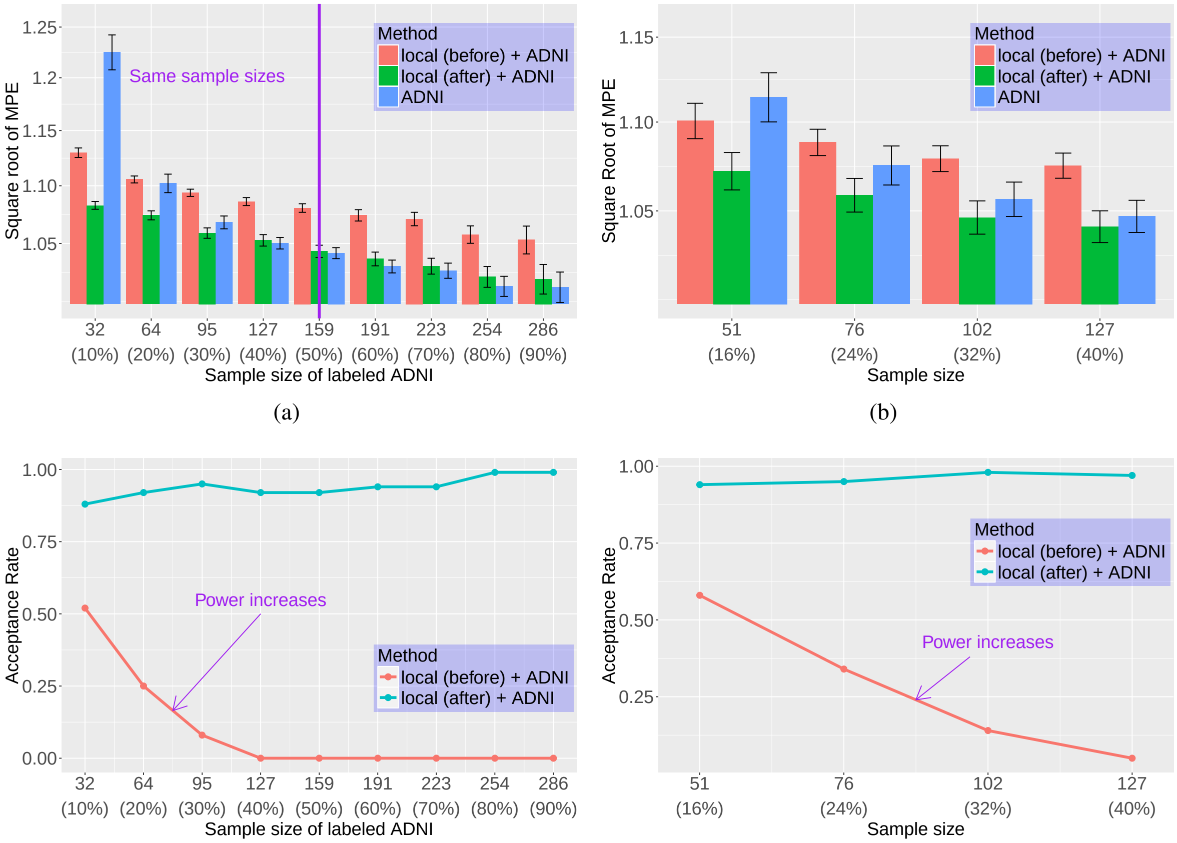


(a) (b)

Figure 4. (a) shows the solution path of λ when sparsity patterns share few features across sites, and group Lasso penalty is added to balance Lasso penalty. (b) shows the alternate regime where sparsity patters are similar. The ℓ_2 loss is plotted, based on 10-fold cross validation.

4.1. AD dataset details

The two datasets we use are – an open-source Alzheimer's Disease Neuroimage Initiative (ADNI) dataset, and a local dataset (ADlocal). ADNI is an open consortium with the goal of understanding AD related cognitive decline, and in the process, develop clinical interventions aimed at delaying the disease onset. ADlocal corresponds to a recent (smaller) initiative local study for the AD related decline. We used 318 samples from ADNI and 156 samples from ADlocal. The input variables are 8 Cerebrospinal fluid (CSF) protein levels, and the response is hippocampus volume. The CSF proteins are "1-38-Tr", "1-40-Tr", "1-42-Tr", "NFL", "AB42", "$htau$", "$ptau_{181}$", and "Neurogranin". The two datasets have different age and diagnosis distributions, and hence, we subsample 81 samples from either of sites to control age and diagnosis variation. Using these 81 samples from each dataset, we perform domain adaptation (using a maximum mean discrepancy objective as a measure of distance between the two marginals) and transform CSF proteins from ADlocal to match ADNI. The transformed data are then used to evaluate our proposed framework. The results in Figure 5 are already explained in the main body.



(a) (b)

(c) (d)

Figure 5. Evaluating combined models. (a,c) In this first setting x-axis represents number/fraction of ADNI labeled samples used in training along with ADlocal labeled data. The dotted line in (a) is where the sample sizes of ADNI and ADlocal in training datasets match. y-axis shows square root of mean prediction error (computed on the remaining unused ADNI data) scaled by estimated noise level in ADNI responses. Error bars give 95% confidence interval. (c) shows the acceptance rate of our hypothesis test. (b,d) show the same evaluations for the alternate setting where *equal* number of ADNI and ADlocal samples are used for training.

References

Liu, Han and Zhang, Jian. Estimation consistency of the group lasso and its applications. In *AISTATS*, pp. 376–383, 2009.

Meinshausen, Nicolai and Yu, Bin. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pp. 246–270, 2009.

Simon, Noah, Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.