Supplementary Material for "Learning Coefficient Heterogeneity over Networks: A Distributed Tree-Based Fused-Lasso Approach"

Xin Zhang
Department of Statistics
Iowa State University
Ames, IA, 50011
xinzhang@iastate.edu

Jia Liu
Department of Computer Science
Iowa State University
Ames, IA, 50011
jialiu@iastate.edu

Zhengyuan Zhu
Department of Statistics
Iowa State University
Ames, IA, 50011
zhuz@iastate.edu

1 Detailed derivations for the proposed generalized ADMM updating rules

First, given the primal and dual pair $\mathbf{w}^t, \mathbf{z}^t$, to determine the weight difference $\mathbf{\Delta}^{t+1}$, we have:

$$\begin{aligned}
\mathbf{\Delta}^{t+1} &= \arg\min_{\mathbf{\Delta}} L_\tau(\mathbf{w}^t, \mathbf{\Delta}, \mathbf{z}^t) \\
&\overset{(a)}{=} \arg\min_{\mathbf{\Delta}} \lambda_N \sum_{l=1}^{K-1}\sum_{p=1}^{d} [\hat{\boldsymbol{\pi}}_l]_p \big|[\mathbf{\Delta}_l]_p\big| - \langle \mathbf{z}^t, -\mathbf{\Delta}\rangle + \frac{\tau}{2}\|\underline{\mathbf{H}}\mathbf{w}^t - \mathbf{\Delta}\|^2 \\
&\overset{(b)}{=} \arg\min_{\mathbf{\Delta}} \lambda_N \sum_{l=1}^{K-1}\sum_{p=1}^{d} [\hat{\boldsymbol{\pi}}_l]_p \big|[\mathbf{\Delta}_l]_p\big| + \mathbf{\Delta}^\top \mathbf{z}^t + \frac{\tau}{2}\mathbf{\Delta}^\top\mathbf{\Delta} - \tau\mathbf{\Delta}^\top\underline{\mathbf{H}}\mathbf{w}^t \\
&= \arg\min_{\mathbf{\Delta}} \lambda_N \sum_{l=1}^{K-1}\sum_{p=1}^{d} [\hat{\boldsymbol{\pi}}_l]_p \big|[\mathbf{\Delta}_l]_p\big| + \frac{\tau}{2}\mathbf{\Delta}^\top\mathbf{\Delta} - \tau\mathbf{\Delta}^\top\Big[\underline{\mathbf{H}}\mathbf{w}^t - \frac{1}{\tau}\mathbf{z}^t\Big] \\
&\overset{(c)}{=} \arg\min_{\mathbf{\Delta}} \lambda_N \sum_{l=1}^{K-1}\sum_{p=1}^{d} [\hat{\boldsymbol{\pi}}_l]_p \big|[\mathbf{\Delta}_l]_p\big| + \frac{\tau}{2}\Big\|\mathbf{\Delta} - (\underline{\mathbf{H}}\mathbf{w}^t - \frac{1}{\tau}\mathbf{z}^t)\Big\|^2,
\end{aligned}$$

where (a) follows from (15) by ignoring constant terms; (b) follows from expanding $\|\underline{\mathbf{H}}\mathbf{w}^t - \mathbf{\Delta}\|^2$ and ignoring constant terms; and (c) follows from adding $\frac{\tau}{2}\|\underline{\mathbf{H}}\mathbf{w}^t - \frac{1}{\tau}\mathbf{z}^t\|^2$ and forming the square term. To compute the element $[\mathbf{\Delta}_l^{t+1}]_p$, the subgradient can be evaluated as:

$$g([\mathbf{\Delta}_l]_p) = \lambda_N[\hat{\boldsymbol{\pi}}_l]_p \nabla|[\mathbf{\Delta}_l]_p| + \tau\Big[[\mathbf{\Delta}_l]_p - ([\mathbf{w}^t_{s(l)}]_p - [\mathbf{w}^t_{e(l)}]_p - \frac{1}{\tau}[\mathbf{z}^t_l]_p)\Big], \tag{A1}$$

Setting the subgradient to zero, we have that

$$[\mathbf{\Delta}_l^{t+1}]_p = S_{\lambda_N[\hat{\boldsymbol{\pi}}_l]_p/\tau}\Big([\mathbf{w}^t_{s(l)}]_p - [\mathbf{w}^t_{e(l)}]_p - \frac{1}{\tau}[\mathbf{z}^t_l]_p\Big), \tag{A2}$$

where $S_{\lambda_N[\hat{\boldsymbol{\pi}}_l]_p/\tau}$ is the soft-thresholding operator (i.e., $S_a(x) = \text{sign}(x)(|x|-a)_+$). To simplify the notation, we define $S_{\lambda_N\hat{\boldsymbol{\pi}}_l/\tau}$ as the coordinate-wise soft-thresholding operator with $[\lambda_N\hat{\boldsymbol{\pi}}_l/\tau]_p = \lambda_N[\hat{\boldsymbol{\pi}}_l]_p/\tau$. Hence, for the l-th edge with end nodes $s(l)$ and $e(l)$, it follows that:

$$\mathbf{\Delta}_l^{t+1} = S_{\lambda_N\hat{\boldsymbol{\pi}}_l/\tau}\Big(\mathbf{w}^t_{s(l)} - \mathbf{w}^t_{e(l)} - \frac{1}{\tau}\mathbf{z}^t_l\Big). \tag{A3}$$

Next, we derive the updating rule for $\mathbf{w}^{t+1}$. In the classical ADMM, it can be shown that:

$$\begin{aligned}
\mathbf{w}^{t+1} &= \arg\min_{\mathbf{w}} L_\tau(\mathbf{w}, \mathbf{\Delta}^{t+1}, \mathbf{z}^t) \\
&= [\mathbf{X}^\top\mathbf{X} + \tau\mathbf{L}\otimes\mathbf{I}_d]^{-1}[\mathbf{X}^\top\mathbf{y} + \underline{\mathbf{H}}^\top(\tau\mathbf{\Delta}^{t+1} + \mathbf{z}^t)].
\end{aligned} \tag{A4}$$

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Unfortunately, the matrix inverse in (A4) *cannot* be computed in distributed fashion due to the coupled structure of the Laplacian matrix $\mathbf{L}$. Here, we adopt the generalized ADMM studied in [1], with which the updating can be implemented in parallel. Instead of directly solving the subproblem $\mathbf{w}^{t+1} = \arg\min_{\mathbf{w}} L_\tau(\mathbf{w}, \boldsymbol{\Delta}^{t+1}, \mathbf{z}^t)$, we add a quadratic term $\frac{1}{2}(\mathbf{w}-\mathbf{w}^t)^\top \mathbf{P}(\mathbf{w}-\mathbf{w}^t)$ in the subproblem:

$$\begin{aligned}
\mathbf{w}^{t+1} &= \arg\min_{\mathbf{w}} L_\tau(\mathbf{w}, \boldsymbol{\Delta}^{t+1}, \mathbf{z}^t) + \frac{1}{2}(\mathbf{w}-\mathbf{w}^t)^\top \mathbf{P}(\mathbf{w}-\mathbf{w}^t) \\
&= \arg\min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{K} \|\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i\|^2 - \langle \mathbf{z}^t, \underline{\mathbf{H}}\mathbf{w}\rangle + \frac{\tau}{2}\|\underline{\mathbf{H}}\mathbf{w} - \boldsymbol{\Delta}^{t+1}\|^2 + \frac{1}{2}(\mathbf{w}-\mathbf{w}^t)^\top \mathbf{P}(\mathbf{w}-\mathbf{w}^t) \\
&= \arg\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^\top[\mathbf{X}^\top\mathbf{X} + \tau\underline{\mathbf{H}}^\top\underline{\mathbf{H}} + \mathbf{P}]\mathbf{w} - \mathbf{w}^\top[\mathbf{X}^\top\mathbf{y} + \underline{\mathbf{H}}^\top(\tau\boldsymbol{\Delta}^{t+1} + \mathbf{z}^t) + \mathbf{P}\mathbf{w}^t] \\
&= [\mathbf{X}^\top\mathbf{X} + \tau\underline{\mathbf{H}}^\top\underline{\mathbf{H}} + \mathbf{P}]^{-1}[\mathbf{X}^\top\mathbf{y} + \underline{\mathbf{H}}^\top(\tau\boldsymbol{\Delta}^{t+1} + \mathbf{z}^t) + \mathbf{P}\mathbf{w}^t].
\end{aligned}$$

Now, we choose the matrix $\mathbf{P} = -\tau\underline{\mathbf{H}}^\top\underline{\mathbf{H}} + \mathbf{D} = -\tau\mathbf{L}\otimes\mathbf{I}_d + \mathbf{D}$, where the diagnoal matrix $\mathbf{D} = \text{diag}(D_1, \cdots, D_K)\otimes\mathbf{I}_d$ with positive scalars D_i for node i and $\mathbf{L} = \mathbf{H}^\top\mathbf{H}$ is the Laplacian matrix for the MST$_s$. It then follows that

$$\mathbf{w}^{t+1} = [\mathbf{X}^\top\mathbf{X} + \mathbf{D}]^{-1}[\mathbf{X}^\top\mathbf{y} + \underline{\mathbf{H}}^\top(\tau\boldsymbol{\Delta}^{t+1} + \mathbf{z}^t) + \mathbf{P}\mathbf{w}^t], \tag{A5}$$

and for each node, plugging in $\mathbf{P} = -\tau\mathbf{L}\otimes\mathbf{I}_d + \mathbf{D}$, the local coefficient can be updated as

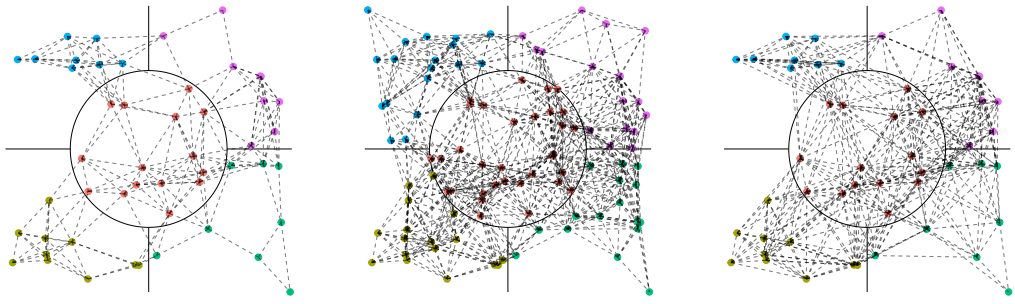
$$\mathbf{w}_i^{t+1} = [\mathbf{X}_i^\top\mathbf{X}_i + D_i\mathbf{I}_d]^{-1}\Big[\mathbf{X}_i^\top\mathbf{y}_i + \sum_{v_i\in e_l}[\mathbf{H}]_{li}(\tau\boldsymbol{\Delta}_l^{t+1} + \mathbf{z}_l^t) + (D_i - \tau\text{deg}(i)\mathbf{w}_i^t + \tau\sum_{j\in\mathcal{N}_i}\mathbf{w}_j^t\Big], \tag{A6}$$

where $v_i \in e_l$ means node v_i is an end node of edge e_l, deg(i) is the degree of the node v_i (i.e., deg$(i) = |\mathcal{N}_i|$). Thus, the updating of $\mathbf{w}_i^{t+1}$ only requires the local and connected neighbor's information, which facilitates *distributed* implementation. Lastly, the dual variables $\mathbf{z}^{t+1}$ can be updated as $\mathbf{z}^{t+1} = \mathbf{z}^t - \tau(\underline{\mathbf{H}}\mathbf{w}^{t+1} - \Delta^{t+1})$, and hence for the l-th edge, the corresponding dual update is:

$$\mathbf{z}_l^{t+1} = \mathbf{z}_l^t - \tau\Big(\mathbf{w}_{s(l)}^{t+1} - \mathbf{w}_{e(l)}^{t+1} - \boldsymbol{\Delta}_l^{t+1}\Big). \tag{A7}$$

2 Further numerical results

In this section, we empirically examine the statistical performance of our proposed estimator $\hat{\mathbf{w}}_{\text{MST}s}$. In Section 2.1, we compare our MST$_s$ method with serveral exising methods on the random generated networks. A comparison is provided in Section 2.2 to study the impact of the choice of regularization on clustering accuracy and costs. Each simulation result is based on 100 independent repetitions and the random seeds are from 1 to 100.



(a) (b) (c)

Figure 1: Simulation network settings: (a) random design with 50 nodes and the radius 0.5; (c) random design with 100 nodes and the radius 0.5;(c) random design with 50 nodes and the radius 0.75. The solid lines show the underlying partition and the dashed lines represent the edges in network graphs.

Table 1: The results of Simulation 2 with BIC criterion. The results are based on 100 repetitions.

Case	Method	MSE($\hat{\mathbf{w}}$)	$\hat{S}$	Sensitivity	Specificity
n=50 K=50	Laplacian	0.0329	NA	NA	NA
	Graph	0.0123	7.28	0.9325	1
	MST_d	0.0134	11.87	0.6777	1
	MST_s	0.0097	5.55	0.9681	1
n=100 K=50	Laplacian	0.0154	NA	NA	NA
	Graph	0.0061	6.83	0.9449	1
	MST_d	0.0067	11.58	0.7051	1
	MST_s	0.0039	5.35	0.9759	1
n=50 K=100	Laplacian	0.0331	NA	NA	NA
	Graph	0.0107	6.94	0.9717	1
	MST_d	0.0132	18.62	0.3855	1
	MST_s	0.0055	5.89	0.9140	1

2.1 Simulation 1: random design

In this part of simulation, we consider the following network setting (see Figure 1 (a-b)): The nodes are uniformly located in the space $[-1, 1]^2$ and the numbers of the nodes are 50 and 100, respectively. There are five underlying clusters, as shown in Figure 1 (a-b). The covariate $\mathbf{x}$ are generated from multivariate normal distribution with zero mean and covariance $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$. The random error ε follows the standard normal distribution. The true coefficients are randomly generated as $\mathbf{w}_{G_1,*} = [4.59, 2.60, -5.12]^\top$, $\mathbf{w}_{G_2,*} = [-2.88, 1.51, 0.59]^\top$, $\mathbf{w}_{G_3,*} = [3.04, 0.53, -4.74]^\top$, $\mathbf{w}_{G_4,*} = [-8.09, -3.20, -2.45]^\top$ and $\mathbf{w}_{G_5,*} = [-0.28, -4.25, -1.28]^\top$. In the connected graph, if the distance of two nodes is smaller than 0.5, then there is an edge between them. Note that with 0.5 as the radius, Assumption 1 is satisfied (see in Figure 1 (a-b)).

We focus on four different types of regularizer: 1) the Laplacian regularizer [6], which can be regarded as a variant of ℓ_2 penalty; 2) the Graph ℓ_1 regularizer as the penalty in (2), which considers all the edges in the graph; 3) the MST_d ℓ_1 regularizer proposed in [3], in which the MST is generated according to spatial distances; 4) our MST_s ℓ_1 regularizer, which generates an MST based on model similarity. We use the Bayesian information criterion (BIC) to select the tuning parameter λ_N. Note that the BIC is widely used in the related works, including homogeneity pursuit methods [2, 3] and subgroup analysis [4, 5]. We simulate three cases: 1) the total number of nodes is $K = 50$ and each node contains $n = 50$ samples; 2) the total number of nodes is $K = 50$ and each node contains $n = 100$ samples; 3) the total number of nodes is $K = 100$ and each node contains $n = 50$ samples. Note that Cases 1) and 2) have different local sample sizes, Cases 1) and 3) have different numbers of nodes and network structures, and Cases 2) and 3) have the same total sample size.

we compare in terms of the following performance metrics: 1) the accuracy of model estimation, $\text{MSE}(\hat{\mathbf{w}}) = \frac{1}{K}\sum_{i=1}^{K} \|\hat{\mathbf{w}}_i - \mathbf{w}_{i,*}\|_2^2$; 2) the estimated group number $\hat{S}$; 3) sensitivity, which measures the proportion of node pairs from the same cluster that are correctly identified; 4) specificity, which measures the proportion of node pairs from the different clusters that are correctly identified. Note that the values of sensitivity and specificity are in the range $[0, 1]$. The closer to 1, the better the prediction is. The simulation results are reported in Table 1 and Figure 2.

From Table 1 and Figure 2, we can see that our MST_s ℓ_1 method outperforms the other methods under all the three circumstances: First of all, we can see that the the MSE from the Laplacian regularizer is higher than those of the other ℓ_1 based regularizer and also the Laplacian regularizer cannot find the nodes' membership. This is because the Laplacian penalty, which is a variant of ℓ_2 penalty, cannot shrink the coefficient difference to zero when two nodes are from the same cluster. Compared to the Graph ℓ_1 regularization, our method improves the efficiency by reducing about $21\%, 36\%, 49\%$ in MSE for the three cases, respectively, while the estimated cluster numbers are closer to five, which is the true group number. Keeping the sample nodes and doubling the local samples in Case 2, the MSE of our proposed regularization reduces to the half of Case 1, which validates our Theorem 1. Comparing Cases 1 and 2, the estimation efficiencies for all three regularizations are improved. This is because by adding more nodes, the total sample size is larger. However, for Cases 2 and 3, although

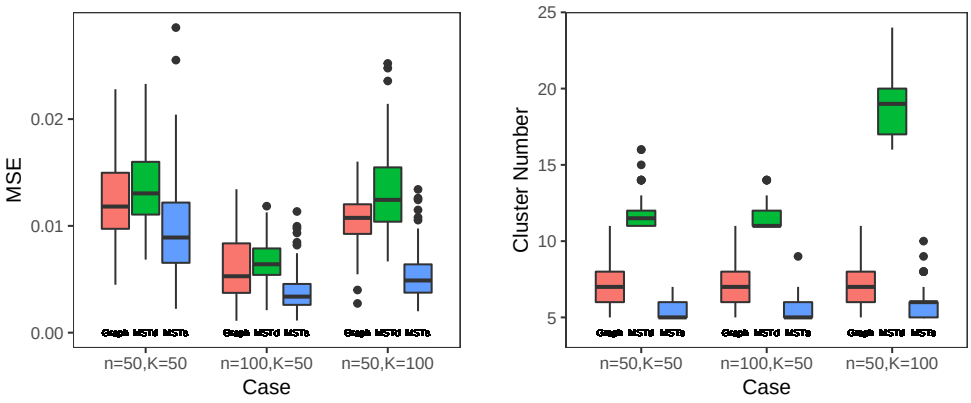


Figure 2: The boxplots of MSEs of $\hat{\mathbf{w}}$ and the estimated group numbers $\hat{S}$ using the three ℓ_1 penalty methods under three cases.

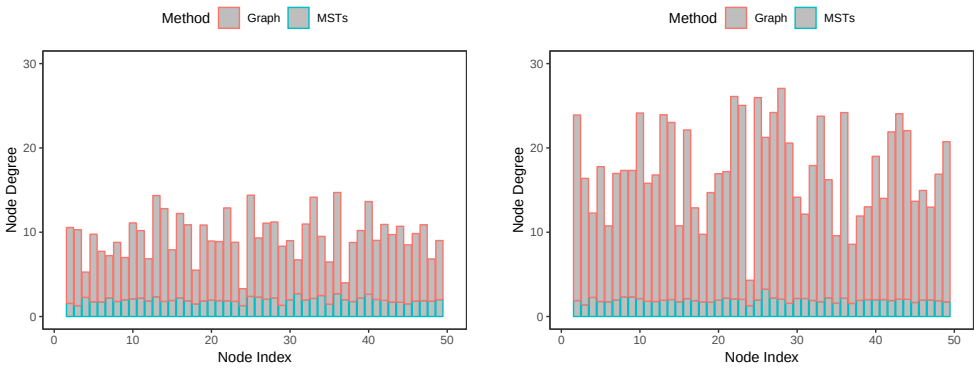
they have the same total sample size, the estimation gets better with fewer node and simpler network topology. Additionally, note that the estimated cluster numbers of the MST$_d$ ℓ_1 regularization is much worse than the Graph ℓ_1 and our regularizations. This is because the MST constructed by the spatial distant cannot guarantee that the nodes from the same group are connected in the tree. This encourages us to use model similarity as the weights when constructing the MST in our MST$_s$ method.

2.2 Simulation 2: network complexity

In this section, we use simulations to illustrate the impact of the choice of regularization on the accuracy, computation time and communication cost. The computations are performed on a Windows computer with a 2.93 GHz Intel(R) Core(TM) i7 CPU processor and 16.0 GB memory. We focus on the two regularizations: the Graph ℓ_1 regularization method and our MST$_s$ ℓ_1 regularization method. In the distributed algorithm, the nodes need to update and store the local $\mathbf{w}_i$, $\mathbf{\Delta}_{i,l}$ and $\mathbf{z}_{i,l}$ in each iteration. Note that the numbers of $\{\mathbf{\Delta}_{i,l}\}_l$ and $\{\mathbf{z}_{i,l}\}_l$ are the same as those the penalty terms associated with node i. Meanwhile, the nodes are required to send the local $\mathbf{w}_i$ to their neighbor nodes in the graph or MST. Clearly, the amount of data being transmitted grows as the graph becomes denser. Here, we consider 50 nodes with the same setting as in Simulation 2. Each node contains 50 samples. We adjust the network denseness by changing the connection radius threshold value r. Two setting are compared, $r = 0.50$ and $r = 0.75$ (See Figure 1 (a) and Figure 1 (c)).

As discussed above, the costs for computation and communication depend on the node degrees of the nodes in the graph or MST. Based on the simulation setting, the connected degrees are shown in Figure 3. The node degrees are deterministic for the graph ℓ_1 regularization. For MST$_s$ ℓ_1 regularization, the node degrees are stochastic because the trees are varying with the local samples. Thus, we repeat 100 trials and compute the average degrees for the nodes. In the case with $r = 0.50$, the maximum degrees for the graph ℓ_1 and MST$_s$ ℓ_1 regularizations are 12 and 2.72, respectively; while in the case with $r = 0.75$, the corresponding maximum degrees are 25 and 3.25, respectively.

Next, with the same local samples from the above 100 simulations, we compare the accuracy and costs for the two regularizations. The MSEs and the estimated group number $\hat{S}$ are used to measure the accuracy. Here we only consider the synchronous algorithm for the computation time approximation. Note that the node with more edges take longer time to calculate more varibles. Thus, the computation time for each iteration is the time for the nodes with the maximum node degree, and the total computation time is the summation of the running times of all interations. The communication cost is defined as the total amount of transmitted messages, which is proportional to the product of the iterations and the edges. We set the baseline as the average computation time and the average communication cost for the MST$_s$ ℓ_1 method under $r = 0.50$. The boxplots for the accuracy, the computation time ratios, and the communication cost ratios are shown in Figure 4. We can see that our MST$_s$ ℓ_1 method outperforms in all aspects. By pruning redundant edges, our MST$_s$ ℓ_1 method enjoys both lower computation and communication costs, as well as the higher



(a) $r = 0.50$ (b) $r = 0.75$

Figure 3: The barcharts of the node degrees for the two setting2 in Simulation 3.

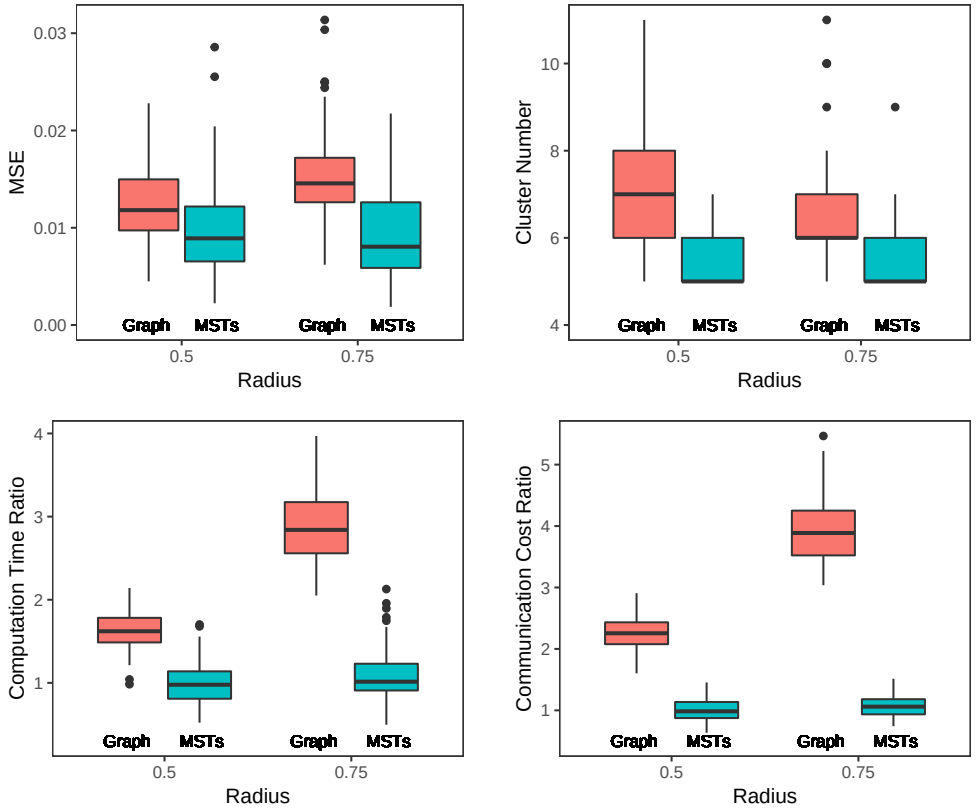


Figure 4: The comparison boxplots of MSEs of $\hat{\mathbf{w}}$, the estimated group numbers $\hat{S}$, the computation time ratio and the communication cost ratio of the graph ℓ_1 regularization and MST$_s$ ℓ_1 regularization.

estimation accuracy. In contrast, for the Graph ℓ_1 regularization, more edges in the graph result in longer computation time and higher communication cost, as well as less accurate estimation.

3 Proofs of the theoretical results

3.1 Proof of Lemma 1

First, we show that under Assumption 1, as the local sample size $n \to \infty$, for any node v_i, the corresponding neighbor node with the minimum weight is from the same cluster with probability 1, i.e., $\lim_{n\to\infty} \mathbb{P}(v_i \sim v_j) = 1$, $j = \arg\min_j \tilde{s}_{i,j}, \forall v_i$, where the notation "$\sim$" means being in the same cluster. Note $\hat{\mathbf{w}}_{i,OLS}$ a root-n consistent estimator of $\mathbf{w}_i$. Thus, the weights for v_i are

$$\tilde{s}_{i,j} = \begin{cases} O_p\Big(\frac{1}{\sqrt{n}}\Big), & \text{if}(v_i, v_j) \in \text{E and } v_i \sim v_j, \\ \|\hat{\mathbf{w}}_{G_k} - \hat{\mathbf{w}}_{G_l}\| + O_p\Big(\frac{1}{\sqrt{n}}\Big), & \text{if}(v_i, v_j) \in \text{E and } v_i \in G_k, v_j \in G_l, \\ \infty, & \text{otherwise.} \end{cases} \tag{A8}$$

where G_k and G_l represent different underlying clusters, $\hat{\mathbf{w}}_{G_k}$ and $\hat{\mathbf{w}}_{G_l}$ are their corresponding coefficients, repectively. Thus, for any node v_i, the event that its corresponding neighboring node with the smallest edge weight is from the same cluster happens with probability 1 as $n \to \infty$.

With the above result, we will show that there is no isolated node in the MST_s with probability 1. We prove it by contradiction. Suppose there is one isolated node v_i. From Assumption 1, we know that there exists a node v_j from the neighbors of v_i in G, such that $\tilde{s}_{i,j} = \min_k \tilde{s}_{i,k}$ and $v_i \sim v_j$. Since v_i is an isolated node, the edge (v_i, v_j) is not in the MST_s. Now, we add the edge (v_i, v_j) to the MST_s. By the property of the spanning tree, adding one more edge to the MST_s would create a cycle. Thus, after adding the edge (v_i, v_j), we have one cycle C in the new graph: $\text{MST}_s + (v_i, v_j)$. Since v_i is an isolated node, in the MST_s, v_i is connected with a node from different cluster, denoted as v_l, and the edge $(v_i, v_l) \in C$. From the weight (A8), it holds that $\tilde{s}_{i,l} > \tilde{s}_{i,j}$ with probability 1. This suggests that $\tilde{s}_{i,j}$ is not the largest weight in C and there exist another edge in C with the largest weight among all the edges in C. By the cycle property of the minimum spanning tree, the edge with largest weight in C cannot be included in the MST_s, contradicting to v_i being an isolated node in the MST_s. Therefore, there is no isolated node in the MST_s with probability 1 as $n \to \infty$.

3.2 Proof of Theorem 1

The proof of Theorem 1 is inspired by [7]. Recall the objective function on $\boldsymbol{\Delta}$:

$$L_{\text{MST}_s}(\boldsymbol{\Delta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\tilde{\underline{\mathbf{H}}}^{-1}\boldsymbol{\Delta}\|^2 + \lambda_N \sum_{p=1}^{dK} [\hat{\boldsymbol{\pi}}]_p |[\Delta]_p|, \tag{A9}$$

and weights

$$[\hat{\boldsymbol{\pi}}]_p = \begin{cases} 1/[\tilde{\underline{\mathbf{H}}}\mathbf{w}_{OLS}]_p^{\gamma}, & \text{if } p \bmod K \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{A10}$$

We first prove the asymptotic normality result. Let $\boldsymbol{\Delta} = \boldsymbol{\Delta}_* + \mathbf{u}/\sqrt{N}$ and

$$\Psi_{\text{MST}_s}(\mathbf{u}) = \frac{1}{2}\Big\|\mathbf{y} - \mathbf{X}\tilde{\underline{\mathbf{H}}}^{-1}\Big(\boldsymbol{\Delta}_* + \frac{\mathbf{u}}{\sqrt{N}}\Big)\Big\|^2 + \lambda_N \sum_{p=1}^{dK} [\hat{\boldsymbol{\pi}}]_p \Big|[\boldsymbol{\Delta}_* + \frac{\mathbf{u}}{\sqrt{N}}]_p\Big|. \tag{A11}$$

Denote $\hat{\mathbf{u}}_{\text{MST}_s} = \Psi_{\text{MST}_s}(\mathbf{u})$. Then, we have $\widehat{\boldsymbol{\Delta}}_{\text{MST}_s} = \boldsymbol{\Delta}_* + \hat{\mathbf{u}}_{\text{MST}_s}/\sqrt{N}$. Define $V_N(\mathbf{u}) = \Psi_{\text{MST}_s}(\mathbf{u}) - \Psi_{\text{MST}_s}(\mathbf{0})$, where

$$V_N(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top\Big[\frac{1}{N}(\mathbf{X}\tilde{\underline{\mathbf{H}}}^{-1})^\top\mathbf{X}\tilde{\underline{\mathbf{H}}}^{-1}\Big]\mathbf{u} - \frac{\boldsymbol{\varepsilon}^\top\mathbf{X}\tilde{\underline{\mathbf{H}}}^{-1}}{\sqrt{N}}\mathbf{u} + \frac{\lambda_N}{\sqrt{N}}\sum_{p=1}^{dK}[\hat{\boldsymbol{\pi}}]_p\sqrt{N}\Big(\Big|\Big[\Delta_* + \frac{\mathbf{u}}{\sqrt{N}}\Big]_p\Big| - |[\Delta_*]_p|\Big). \tag{A12}$$

With Assumption 2, we have that

$$\frac{1}{N}(\mathbf{X}\tilde{\underline{\mathbf{H}}}^{-1})^\top\mathbf{X}\tilde{\underline{\mathbf{H}}}^{-1} \xrightarrow{p} \mathbf{C} \text{ and } \frac{\boldsymbol{\varepsilon}^\top\mathbf{X}\tilde{\underline{\mathbf{H}}}^{-1}}{\sqrt{N}} \xrightarrow{d} \mathbf{W} = \mathcal{N}(0, \sigma^2\mathbf{C}).$$

In what follows, we derive the limiting behavior of the third term in (A12). If $[\Delta_*]_p \neq 0$ and $p \bmod K \neq 0$, then $[\hat{\boldsymbol{\pi}}]_p \xrightarrow{p} |[\Delta_*]_p|^{-\gamma}$ and $\sqrt{N}(|[\Delta_* + \frac{\mathbf{u}}{\sqrt{N}}]_p| - |[\Delta_*]_p|) \to [\mathbf{u}]_p \text{sign}([\Delta_*]_p)$. By the Slutsky's theorem, we have $\frac{\lambda_N}{\sqrt{N}}[\hat{\boldsymbol{\pi}}]_p\sqrt{N}(|[\Delta_* + \frac{\mathbf{u}}{\sqrt{N}}]_p| - |[\Delta_*]_p|) \xrightarrow{p} 0$ with $\lambda_N/\sqrt{N} \to 0$. If $[\Delta_*]_p = 0$ and $p \bmod K \neq 0$, then it holds that

$$\sqrt{N}(|[\Delta_* + \frac{\mathbf{u}}{\sqrt{N}}]_p| - |[\Delta_*]_p|) \to |[\mathbf{u}]_p| \tag{A13}$$

$$\frac{\lambda_N}{\sqrt{N}}\hat{\boldsymbol{\pi}}_p = \frac{\lambda_N}{\sqrt{N}}N^{\gamma/2}|\sqrt{N}[\widehat{\Delta}_{\text{OLS}}]_p|^{-\gamma} = O_p(1)\lambda_N N^{(\gamma-1)/2} \to \infty. \tag{A14}$$

Thus, with $[\Delta_*]_p = 0$ and $p \bmod K \neq 0$, we have

$$\frac{\lambda_N}{\sqrt{N}}[\hat{\boldsymbol{\pi}}]_p\sqrt{N}(|[\Delta_* + \frac{\mathbf{u}}{\sqrt{N}}]_p| - |[\Delta_*]_p|) \tag{A15}$$

$$= O_p(1)|[\mathbf{u}]_p|\lambda_N N^{(\gamma-1)/2}] = \begin{cases} 0, & \text{if } [\mathbf{u}]_p = 0, \\ \infty, & \text{otherwise.} \end{cases} \tag{A16}$$

If $p \bmod K = 0$, with the above weight $\hat{\boldsymbol{\pi}}_p = 0$, we have $\frac{\lambda_N}{\sqrt{N}}[\hat{\boldsymbol{\pi}}]_p\sqrt{N}(|[\Delta_* + \frac{\mathbf{u}}{\sqrt{N}}]_p| - |[\Delta_*]_p|) = 0$. Thus, by the Slutsky's theorem, it holds that $V_N(\mathbf{u}) \xrightarrow{d} V(\mathbf{u})$, where

$$V(\mathbf{u}) = \begin{cases} \frac{1}{2}\mathbf{u}_{\mathcal{A}_*}^\top \mathbf{C}_{\mathcal{A}_*}\mathbf{u}_{\mathcal{A}_*} - \mathbf{u}_{\mathcal{A}_*}^\top \mathbf{W}_{\mathcal{A}_*}, & \text{if } [\mathbf{u}]_p = 0,\ \forall p \notin \mathcal{A}_*, \\ \infty, & \text{otherwise.} \end{cases} \tag{A17}$$

Note V_N is convex and the unique minimum of V is $\tilde{\mathbf{u}}$, where $[\tilde{\mathbf{u}}]_{\mathcal{A}_*} = \mathbf{C}_{\mathcal{A}_*}^{-1}\mathbf{W}_{\mathcal{A}_*}$ and $[\tilde{\mathbf{u}}]_{\mathcal{A}_*^c} = \mathbf{0}$. Following the same line as in [7], we have

$$[\hat{\mathbf{u}}_{\text{MST}_s}]_{\mathcal{A}_*} \xrightarrow{d} \mathbf{C}_{\mathcal{A}_*}^{-1}\mathbf{W}_{\mathcal{A}_*} \text{ and } [\hat{\mathbf{u}}_{\text{MST}_s}]_{\mathcal{A}_*^c} \xrightarrow{d} \mathbf{0}. \tag{A18}$$

With $\mathbf{W}_{\mathcal{A}_*} = \mathcal{N}(0, \sigma^2\mathbf{C}_{\mathcal{A}_*})$, the asymptotic normality part is proved.

Next, we show the consistency part. For $p \in \mathcal{A}_*$, the asymptotic normality result shows that $[\widehat{\Delta}_{\text{MST}_s}]_p \xrightarrow{p} [\Delta_*]_p$, therefore, $\mathbb{P}(p \in \hat{\mathcal{A}}_N) \to 1$. Then, we need to show $\forall p' \notin \mathcal{A}_*, \mathbb{P}(p \in \hat{\mathcal{A}}_N) \to 0$. Consider the event $p' \in \hat{\mathcal{A}}_N$. With the KKT optimality conditions, it holds that

$$[\mathbf{X}\underline{\tilde{\mathbf{H}}}^{-1}]_{p'}^\top[\mathbf{y} - \mathbf{X}\underline{\tilde{\mathbf{H}}}^{-1}\widehat{\Delta}_{\text{MST}_s}] = \pm\lambda_N[\hat{\boldsymbol{\pi}}]_{p'}, \tag{A19}$$

where $[\cdot]_{\cdot p}$ represents the pth column of the matrix. Note that on the RHS, we have

$$\lambda_N[\hat{\boldsymbol{\pi}}]_{p'}/\sqrt{N} = \lambda_N N^{(\gamma-1)/2}|\sqrt{N}\widehat{\Delta}_{\text{OLS}}|^{-\gamma} \xrightarrow{p} \infty, \tag{A20}$$

while on the LHS, we have

$$2\frac{[\mathbf{X}\underline{\tilde{\mathbf{H}}}^{-1}]_{\cdot p'}^\top[\mathbf{y} - \mathbf{X}\underline{\tilde{\mathbf{H}}}^{-1}\widehat{\Delta}_{\text{MST}_s}]}{\sqrt{N}} = 2\frac{[\mathbf{X}\underline{\tilde{\mathbf{H}}}^{-1}]_{\cdot p'}^\top\mathbf{X}\underline{\tilde{\mathbf{H}}}^{-1}[\Delta_* - \widehat{\Delta}_{\text{MST}_s}]}{\sqrt{N}} + 2\frac{[\mathbf{X}\underline{\tilde{\mathbf{H}}}^{-1}]_{\cdot p'}^\top\boldsymbol{\varepsilon}}{\sqrt{N}}. \tag{A21}$$

With the asymptotic normality result and the Slutsky's theorem, $\lambda_N[\hat{\boldsymbol{\pi}}]_{p'}/\sqrt{N}$ asymptotically follows a normal distribution. Thus, we finally have

$$\mathbb{P}(p' \in \hat{\mathcal{A}}_N) \leq \mathbb{P}([\mathbf{X}\underline{\tilde{\mathbf{H}}}^{-1}]_{p'}^\top[\mathbf{y} - \mathbf{X}\underline{\tilde{\mathbf{H}}}^{-1}\widehat{\Delta}_{\text{MST}_s}] = \pm\lambda_N[\hat{\boldsymbol{\pi}}]_{p'}) \to 0. \tag{A22}$$

3.3 Proof of Theorem 2

We prove the convergence following the framework of [1]. Recalling the constrained objective function (16), for notational simplicity, we denote $f(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^K \|\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i\|^2$ and $g(\boldsymbol{\Delta}) = \lambda_N \sum_{l=1}^{K-1}\sum_{p=1}^d [\hat{\boldsymbol{\pi}}_l]_p|[\boldsymbol{\Delta}_l]_p|$. Next, we will provide a convergence analysis for our proposed generalized ADMM method for the problem in the following form:

$$\begin{aligned} &\text{Minimize} \quad f(\mathbf{w}) + g(\boldsymbol{\Delta}) \\ &\text{subject to} \quad \underline{\mathbf{H}}\mathbf{w} - \boldsymbol{\Delta} = \mathbf{0}. \end{aligned} \tag{A23}$$

Note that $f(\mathbf{w})$ is the ordinary least square problem, and with Assumption 2, it is strongly convex with a convex modulus $\nu_f > 0$ and has Lipschitz continuous gradients; for $g(\Delta)$, it is a convex function and the corresponding convex modulus $\nu_g = 0$; the constraint $\underline{\mathbf{H}}\mathbf{w} - \boldsymbol{\Delta} = \mathbf{0}$ is linearly independent and $\underline{\mathbf{H}}$ is full row rank.

Now, we state Lemma 2.1 and Theorem 2.2 from [1] as follows (with our notation).

Lemma. *The sequence $\{\mathbf{u}^t\}$ obeys the followings optimality conditions at each iteration:*

$$\underline{\mathbf{H}}^\top \mathbf{z}^{t+1} + \mathbf{P}(\mathbf{w}^t - \mathbf{w}^{t+1}) = \nabla f(\mathbf{w}^{t+1}) \tag{A24}$$

$$-\mathbf{z}^{t+1} + \tau\underline{\mathbf{H}}(\mathbf{w}^t - \mathbf{w}^{t+1}) \in \nabla g(\Delta^{t+1}) \tag{A25}$$

Theorem. *If the matrix $\mathbf{P}$ satisfies that $\mathbf{P} \succ \mathbf{0}$, then there exists $\eta > 0$ such that*

$$\|\mathbf{u}^t - \mathbf{u}_*\|_{\mathbf{G}}^2 - \|\mathbf{u}^{t+1} - \mathbf{u}_*\|_{\mathbf{G}}^2 \geq \eta\|\mathbf{u}^t - \mathbf{u}^{t+1}\|_{\mathbf{G}}^2 + 2\nu_f\|\mathbf{w}^{t+1} - \mathbf{w}_*\|^2. \tag{A26}$$

Note that the RHS of (A26) is positive. Hence, $\|\mathbf{u}^{t+1} - \mathbf{u}_*\|_{\mathbf{G}}^2$ and $\mathbf{u}^{t+1}$ are bounded. With the boundedness of sequence $\{\mathbf{u}^t\}$, it follows that there exists a converging subsequence $\{\mathbf{u}^{t_j}\}$ of $\{\mathbf{u}^t\}$. Let $\bar{\mathbf{u}} = \lim_{j\to\infty} \mathbf{u}^{t_j}$. In what follows, we will show that $\bar{\mathbf{u}} = (\bar{\mathbf{w}}^\top, \bar{\Delta}^\top, \bar{\mathbf{z}}^\top)^\top$ is a KKT point. Let $\mathbf{u}_*$ denote an arbirary KKT point for the problem (A23).

From (A26), it can be seen that $\|\mathbf{u}^t - \mathbf{u}_*\|_{\mathbf{G}}^2$ is monotonically nonincreasing and converging. Also, due to $\eta > 0$, $\|\mathbf{u}^t - \mathbf{u}^{t+1}\|_{\mathbf{G}}^2 \to 0$. With the structure of $\mathbf{G}$, it holds that $\mathbf{z}^t - \mathbf{z}^{t+1} \to \mathbf{0}$, and equivalently,

$$\underline{\mathbf{H}}\mathbf{w}^{t+1} - \boldsymbol{\Delta}^{t+1} = \mathbf{0}, \tag{A27}$$

from the updating rule of $\mathbf{z}$. Taking limit on (A27) over the subsequence, we have

$$\underline{\mathbf{H}}\bar{\mathbf{w}} - \bar{\Delta} \to \mathbf{0}, \tag{A28}$$

Since $\|\mathbf{u}^t - \mathbf{u}^{t+1}\|_{\mathbf{G}}^2 \to \mathbf{0}$, it follows that $\mathbf{w}^t - \mathbf{w}^{t+1} \to \mathbf{0}$. From (A24) and (A25), we have that: 1) $\underline{\mathbf{H}}^\top \bar{\mathbf{z}} = \nabla f(\bar{\mathbf{w}})$; and 2) $-\bar{\mathbf{z}} \in \nabla g(\bar{\boldsymbol{\Delta}})$. With (A28), we have that $\bar{\mathbf{u}}$ is a KKT point and thus we have $\mathbf{u}_* = \bar{\mathbf{u}}$. Since $\mathbf{u}^{t_j} \to \mathbf{u}_*$ and the convergence of $\|\mathbf{u}^t - \mathbf{u}_*\|_{\mathbf{G}}^2$, we have $\|\mathbf{u}^t - \mathbf{u}_*\|_{\mathbf{G}}^2 \to \mathbf{0}$.

Next, we prove the linear convergence rate of our algorithm. From (A26), we have that

$$\|\mathbf{u}^t - \mathbf{u}_*\|_{\mathbf{G}}^2 - \|\mathbf{u}^{t+1} - \mathbf{u}_*\|_{\mathbf{G}}^2 \geq \eta\|\mathbf{u}^t - \mathbf{u}^{t+1}\|_{\mathbf{G}}^2 + 2\nu_f\|\mathbf{w}^{t+1} - \mathbf{w}_*\|^2. \tag{A29}$$

Hence, we need to show that there exists some $\delta > 0$, such that

$$\eta\|\mathbf{u}^t - \mathbf{u}^{t+1}\|_{\mathbf{G}}^2 + 2\nu_f\|\mathbf{w}^{t+1} - \mathbf{w}_*\|^2 \geq \delta\|\mathbf{u}^{t+1} - \mathbf{u}_*\|_{\mathbf{G}}^2, \tag{A30}$$

which is equivalent to

$$\eta\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{D}}^2 + \frac{\eta}{\tau}\|\mathbf{z}^t - \mathbf{z}^{t+1}\|^2 + 2\nu_f\|\mathbf{w}^{t+1} - \mathbf{w}_*\|^2 \geq \delta\|\mathbf{w}^{t+1} - \mathbf{w}_*\|_{\mathbf{D}}^2 + \frac{\delta}{\tau}\|\mathbf{z}^{t+1} - \mathbf{z}_*\|^2 \tag{A31}$$

To this end, we state Lemma 3.2 in [1] as follows:

Lemma. *Suppose that ∇f is Lipschitz continuous with constant L_f. For all $\mu > 1$, we have*

$$\|\mathbf{z}^{t+1} - \mathbf{z}_*\|^2 \leq c_1\|\mathbf{w}^{t+1} - \mathbf{w}_*\|^2 + c_2\|\mathbf{w}^t - \mathbf{w}^{t+1}\|^2, \tag{A32}$$

where $c_1 = L_f^2(1 - \frac{1}{\mu})^{-1}\lambda_{\min}^{-1}(\underline{\mathbf{H}}\underline{\mathbf{H}}^\top) > 0$ and $c_2 = \mu\|\mathbf{P}\|^2\lambda_{\min}^{-1}(\underline{\mathbf{H}}\underline{\mathbf{H}}^\top) > 0$.

Applying the above lemma to the RHS of (A31), we have

$$\begin{aligned}\eta\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{D}}^2 + \frac{\eta}{\tau}\|\mathbf{z}^t - \mathbf{z}^{t+1}\|^2 + 2\nu_f\|\mathbf{w}^{t+1} - \mathbf{w}_*\|^2 \\ \geq c_3\|\mathbf{w}^t - \mathbf{w}^{t+1}\|^2 + \frac{\eta}{\tau}\|\mathbf{z}^t - \mathbf{z}^{t+1}\|^2 + 2\nu_f\|\mathbf{w}^{t+1} - \mathbf{w}_*\|^2,\end{aligned} \tag{A33}$$

where $c_3 = \eta \min D_i$. With c_4 and c_5 satisfying $c_3 - c_1c_5 \geq c_4$ and $2\nu_f \geq c_2c_5$, it follows that

$$\begin{aligned}&\eta\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{D}}^2 + \frac{\eta}{\tau}\|\mathbf{z}^t - \mathbf{z}^{t+1}\|^2 + 2\nu_f\|\mathbf{w}^{t+1} - \mathbf{w}_*\|^2 \\ &\qquad \geq c_4\|\mathbf{w}^t - \mathbf{w}^{t+1}\|^2 + c_5\|\mathbf{z}^{t+1} - \mathbf{z}_*\|^2 + \frac{\eta}{\tau}\|\mathbf{z}^t - \mathbf{z}^{t+1}\|^2 \\ &\qquad \geq \frac{c_4}{\max D_i}\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{D}}^2 + \frac{c_5}{\tau}\tau\|\mathbf{z}^{t+1} - \mathbf{z}_*\|^2 \\ &\qquad \geq \delta\|\mathbf{u}^{t+1} - \mathbf{u}_*\|_{\mathbf{G}}^2,\end{aligned} \tag{A34}$$

where $\delta = \min\{c_4/\max D_i, c_5/\tau\}$. This completes the proof.

References

[1] Wei Deng and Wotao Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.

[2] Zheng Tracy Ke, Jianqing Fan, and Yichao Wu. Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509):175–194, 2015.

[3] Furong Li and Huiyan Sang. Spatial homogeneity pursuit of regression coefficients for large datasets. *Journal of the American Statistical Association*, (just-accepted):1–37, 2018.

[4] Shujie Ma and Jian Huang. A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423, 2017.

[5] Shujie Ma, Jian Huang, and Zhiwei Zhang. Exploration of heterogeneous treatment effects via concave fusion. *arXiv preprint arXiv:1607.03717*, 2018.

[6] Weiran Wang, Jialei Wang, Mladen Kolar, and Nathan Srebro. Distributed stochastic multi-task learning with graph regularization. *arXiv preprint arXiv:1802.03830*, 2018.

[7] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.