7 SUPPLEMENTARY

7.1 PROOFS

Proof of Proposition 1. Note that $\mathbf{u}$ is the optimal solution to the lasso problem: $\mathbf{u} = \arg\min_{\mathbf{v}\in R^n} \frac{\tau s}{2}\|\mathbf{v} - \tilde{\mathbf{Z}}_{ki}^{(t)}\|_2^2 + \lambda\|\mathbf{v}\|_1$.

Suppose $k \in \Lambda_i$. Define $T_k(v) = \frac{\tau s}{2}(v - \tilde{\mathbf{Z}}_{ki}^{(t)})^2 + \lambda|v|$ for $v \in \mathbb{R}$, then $\mathbf{u}_k = \arg\min_{v\in\mathbb{R}} T_k(v)$. Since the two functions $H_k(v)$ and $T_k(v)$ only differ at $v = 0$, $\arg\min_{v\in\{\mathbf{u}_k,0\}} H_k(v)$ is the optimal solution to $\min_{v\in R} H_k(v)$ when $\mathbf{u}_k \neq 0$ or $\mathbf{u}_k = 0$ and $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} \geq 0$.

When $\mathbf{u}_k = 0$ and $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} < 0$, when $\varepsilon \to 0$ and $\varepsilon \neq 0$, $H_k(\varepsilon) \to \frac{\tau s}{2}(\tilde{\mathbf{Z}}_{ki}^{(t)})^2 + \gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}$ and $\inf_{v\in\mathbb{R}} H_k(v) = \frac{\tau s}{2}(\tilde{\mathbf{Z}}_{ki}^{(t)})^2 + \gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}$. Note that the infimum can never be achieved. Since $\inf_{v\in\mathbb{R}} H_k(v) < H_k(\mathbf{Z}_{ki}^{(t-1)})$, we can always find $\varepsilon \neq 0$ such that $H_k(\varepsilon) \leq H_k(\mathbf{Z}_{ki}^{(t-1)})$.

Suppose $k \notin \Lambda_i$, then $\mathbf{u}_k$ is the optimal solution to $\min_{v\in R} H_k(v)$ for $k \neq i$.

Define $H(\mathbf{v}) = \frac{\tau s}{2}\|\mathbf{v} - \tilde{\mathbf{Z}}^{i^{(t)}}\|_2^2 + \lambda\|\mathbf{v}\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{v})$. Based on the above argument, $H(\mathbf{Z}^{i^{(t)}}) \leq H(\mathbf{Z}^{i^{(t-1)}})$ which indicates that

$$\frac{\tau s}{2}\|\mathbf{Z}^{i^{(t)}} - \mathbf{Z}^{i^{(t-1)}}\|_2^2 + \langle \mathbf{Z}^{i^{(t)}} - \mathbf{Z}^{i^{(t-1)}}, \nabla Q(\mathbf{Z}^{i^{(t-1)}})\rangle \tag{29}$$

$$+ \lambda\|\mathbf{Z}^{i^{(t)}}\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^{i^{(t)}}) \leq \lambda\|\mathbf{Z}^{i^{(t-1)}}\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^{i^{(t-1)}}) \tag{30}$$

Also, since s is the Lipschitz constant for the gradient of function $Q(\cdot)$, we have

$$Q(\mathbf{Z}^{i^{(t)}}) \leq Q(\mathbf{Z}^{i^{(t-1)}}) + \langle \mathbf{Z}^{i^{(t)}} - \mathbf{Z}^{i^{(t-1)}}, \nabla Q(\mathbf{Z}^{i^{(t-1)}})\rangle \tag{31}$$

$$+ \frac{s}{2}\|\mathbf{Z}^{i^{(t)}} - \mathbf{Z}^{i^{(t-1)}}\|_2^2$$

Combining (29) and (31),

$$F(\mathbf{Z}^{i^{(t)}}) \leq F(\mathbf{Z}^{i^{(t-1)}}) - \frac{(\tau-1)s}{2}\|\mathbf{Z}^{i^{(t)}} - \mathbf{Z}^{i^{(t-1)}}\|_2^2$$

□

Proof of Lemma 1. We first prove that the sequences $\{\mathbf{Z}^{i^{(t)}}\}_t$ is bounded for any $1 \leq i \leq n$. By Proposition 1, the sequence $\{F(\mathbf{Z}^{i^{(t)}})\}_t$ decreases, so we have

$$F(\mathbf{Z}^{i^{(t)}}) = \|\mathbf{x}_i - \boldsymbol{X}\mathbf{Z}^{i^{(t)}}\|_2^2 + \lambda\|\mathbf{Z}^{i^{(t)}}\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^{i^{(t)}}) \leq F(\mathbf{Z}^{i^{(0)}})$$

for $t \geq 1$. Therefore,

$$\|\mathbf{Z}^{i^{(t)}}\|_1 \leq \frac{1}{\lambda}F(\mathbf{Z}^{i^{(0)}})$$

It follows that $\|\mathbf{Z}^{i^{(t)}}\|_1$ is bounded, and $\|\mathbf{Z}^{i^{(t)}}\|_2$ is also bounded. Since $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} \geq 0$ for all $k \in \Lambda_i$ and the indicator function $\mathbb{1}_{\cdot\neq 0}$ is semi-algebraic function, $\mathbf{R}_{\tilde{\mathbf{S}}}(\cdot)$ is also a semi-algebraic function and lower semicontinuous. Therefore, according to Theorem 1 by Bolte et al. (2014), $\{\mathbf{Z}^{i^{(t)}}\}_t$ converges to a critical point of $F(\mathbf{Z}^i)$, denoted by $\hat{\mathbf{Z}}^i$.

Let $\hat{\mathbf{v}} = 2\boldsymbol{X}^{(-i)^\top}(\boldsymbol{X}^{(-i)}\hat{\mathbf{Z}^i}_{-i} - \mathbf{x}_i) + \dot{\mathbf{P}}(\hat{\mathbf{Z}^i}_{-i}; b)$. For k such that $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} = 0$ or $k \notin \Lambda_i$, since $\hat{\mathbf{Z}}^i$ is a critical point of $F(\mathbf{Z}^i)$, $\hat{\mathbf{v}}_{k_{-i}} = 0$.

Now we consider the case that $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} \neq 0$ and $k \in \Lambda_i$. In the following text we denote by k_{-i} the index of the element of the vector $\mathbf{v}_{-i}$ corresponding to the element $\mathbf{v}_k$ of $\mathbf{v}$, i.e. $(\mathbf{v}_{-i})_{k_{-i}} = \mathbf{v}_k$ for any $\mathbf{v} \in \mathbb{R}^n$.

For $k \in \hat{\mathbf{S}}_i$, since $\hat{\mathbf{Z}}^i$ is a critical point of $F(\mathbf{Z}^i) = \|\mathbf{x}_i - \boldsymbol{X}\mathbf{Z}^i\|_2^2 + \lambda\|\mathbf{Z}^i\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i)$. then $\frac{\partial(Q+\lambda\|\mathbf{Z}^i\|_1)}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\hat{\mathbf{Z}}^i} = 0$ because $\frac{\partial\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i)}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\hat{\mathbf{Z}}^i} = 0$. Note that $\min_{k\in\hat{\mathbf{S}}_i}|\hat{\mathbf{Z}_k^i}| > b$, so $\frac{\partial\mathbf{T}}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\hat{\mathbf{Z}}^i} = 0$. It follows that $\hat{\mathbf{v}}_{k_{-i}} = 0$.

For $k \notin \hat{\mathbf{S}}_i$, since $\frac{dP_k}{d\mathbf{Z}_k^i}(\hat{\mathbf{Z}}_k^i+; b) = \frac{\gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b} + \lambda$ and $\frac{dP_k}{d\mathbf{Z}_k^i}(\hat{\mathbf{Z}}_k^i-; b) = -\frac{\gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b} - \lambda$, $\frac{\gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b} + \lambda \geq |\frac{\partial Q}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\hat{\mathbf{Z}}^i}|$, we can choose the k_{-i}-th element of $\dot{\mathbf{P}}(\hat{\mathbf{Z}}^i; b)$ such that $\hat{\mathbf{v}}_{k_{-i}} = 0$. Therefore, $\|\hat{\mathbf{v}}\|_2 = 0$, and $\hat{\mathbf{Z}}^i$ is a local solution to the problem (19).

Now we prove that $\mathbf{Z}^{i^*}$ is also a local solution to (19). Let $\mathbf{v}^* = 2\boldsymbol{X}^{(-i)^\top}(\boldsymbol{X}^{(-i)}\mathbf{Z}^{i^*}_{-i} - \mathbf{x}_i) + \dot{\mathbf{P}}(\mathbf{Z}^{i^*}_{-i}; b)$, and Q is defined as before. For k such that $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} = 0$ or or $k \notin \Lambda_i$, since $\mathbf{Z}^{i^*}$ is the globally optimal solution of $F(\mathbf{Z}^i)$, $\mathbf{v}^*_{k_{-i}} = 0$.

Again we consider the case that $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} \neq 0$ and $k \in \Lambda_i$.

For $k \in \mathbf{S}_i^*$, since $\mathbf{Z}^{i^*}$ is the globally optimal solution to problem (8), we also have $\frac{\partial(Q+\lambda\|\mathbf{Z}^i\|_1)}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\mathbf{Z}^{i^*}} = 0$. If it is not the case and $\frac{\partial(Q+\lambda\|\mathbf{Z}^i\|_1)}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\mathbf{Z}^{i^*}} \neq 0$, then we can change $\mathbf{Z}_k^i$ by a small amount in the direction of the gradient $\frac{\partial(Q+\lambda\|\mathbf{Z}^i\|_1)}{\partial\mathbf{Z}_k^i}$ at the point $\mathbf{Z}^i = \mathbf{Z}^{i^*}$ while $\mathbf{Z}_k^i$ is still nonzero, leading to a smaller value of the objective $F(\mathbf{Z}^i)$.

Note that $\min_{k\in\mathbf{S}_i^*}|\mathbf{Z}_k^{i^*}| > b$, so $\frac{\partial\mathbf{T}}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\hat{\mathbf{Z}}^i} = 0$, and it follows that $\mathbf{v}^*_{k_{-i}} = 0$.

For $k \notin \mathbf{S}_i^*$, since $\frac{\gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b} + \lambda \geq \max_{k\notin\hat{\mathbf{S}}_i}|\frac{\partial Q}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\mathbf{Z}^{i*}}|$, we can choose the k_{-i}-th element of $\dot{\mathbf{P}}(\mathbf{Z}^{i^*}; b)$ such that $\mathbf{v}^*_{k_{-i}} = 0$. It follows that $\|\mathbf{v}^*\|_2 = 0$, and $\mathbf{Z}^{i^*}$ is also a

local solution to the problem (19). □

Proof of Theorem 1. According to Lemma 1, both $\hat{\mathbf{Z}}^i$ and $\mathbf{Z}^{i*}$ are local solutions to problem (19). In the following text, let $\boldsymbol{\beta}_{\mathbf{I}}$ indicates a vector whose elements are those of $\boldsymbol{\beta}$ with indices in $\mathbf{I}$. Let $\Delta = \mathbf{Z}^{i*}_{-i} - \hat{\mathbf{Z}}^i_{-i}$, $\tilde{\Delta} = \dot{\mathbf{P}}(\mathbf{Z}^{i*}) - \dot{\mathbf{P}}(\hat{\mathbf{Z}}^i)$. By Lemma 1, we have

$$\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta + \tilde{\Delta}\|_2 = 0$$

It follows that

$$\begin{aligned}&2\Delta^\top \boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta + \Delta^\top\tilde{\Delta}\\ &\le \|\Delta\|_2\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta + \tilde{\Delta}\|_2 = 0\end{aligned}$$

Also, by the proof of Lemma 1, for $k \in \hat{\mathbf{S}}_i \cap \mathbf{S}^*_i$, $(2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta)_{k_{-i}} = 2\lambda\mathbb{I}_{\mathbf{Z}^{i*}_k\hat{\mathbf{Z}}^i_k<0} + 0\mathbb{I}_{\mathbf{Z}^{i*}_k\hat{\mathbf{Z}}^i_k>0}$. We now present another property on any nonconvex function P using the degree of nonconvexity in Definition 3: $\theta(t,\kappa) := \sup_s\{-\mathrm{sgn}(s-t)(\dot{P}(s;b)-\dot{P}(t;b)) - \kappa|s-t|\}$ on the regularizer $\mathbf{P}$. For any $s,t \in \mathbb{R}$, we have

$$-\mathrm{sgn}(s-t)\big(\dot{P}(s;b) - \dot{P}(t;b)\big) - \kappa|s-t| \le \theta(t,\kappa)$$

by the definition of θ. It follows that

$$\begin{aligned}&\theta(t,\kappa)|s-t| \ge -(s-t)\big(\dot{P}(s;b) - \dot{P}(t;b)\big) - \kappa(s-t)^2\\ &-(s-t)\big(\dot{P}(s;b) - \dot{P}(t;b)\big) \le \theta(t,\kappa)|s-t| + \kappa(s-t)^2\end{aligned} \tag{32}$$

Let $\hat{\mathbf{S}}^{-i}_i = \mathrm{supp}(\hat{\mathbf{Z}}^i_{-i})$, $\mathbf{S}^{-i*}_i = \mathrm{supp}(\mathbf{Z}^{i*}_{-i})$, $\mathbf{U}^{-i}_i = (\hat{\mathbf{S}}^{-i}_i \backslash \mathbf{S}^{-i*}_i) \cup (\mathbf{S}^{-i*}_i \backslash \hat{\mathbf{S}}^{-i}_i)$. Applying (32) with $P = P_k$ for $k = 1,\dots,n$, $k \ne i$, we have

$$\begin{aligned}&2\Delta^\top\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta \le -\Delta^\top\tilde{\Delta}\\ &= -\Delta^\top_{\mathbf{U}^{-i}_i}\tilde{\Delta}_{\mathbf{U}^{-i}_i} - \Delta^\top_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\\ &\le |(\mathbf{Z}^{i*}_{-i})_{\mathbf{U}^{-i}_i} - (\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i}|^\top\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\\ &+\kappa\|(\mathbf{Z}^{i*}_{-i})_{\mathbf{U}^{-i}_i} - (\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i}\|_2^2 + \|\Delta_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\\ &\le \|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2\|(\mathbf{Z}^{i*}_{-i})_{\mathbf{U}^{-i}_i} - (\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i}\|_2\\ &+\kappa\|\Delta\|_2^2 + \|\Delta\|_2\|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\\ &\le \|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2\|\Delta\|_2 + \kappa\|\Delta\|_2^2 + \|\Delta\|_2\|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\end{aligned} \tag{33}$$

On the other hand, $\Delta^\top\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta \ge \kappa_0^2\|\Delta\|_2^2$. It follows from (33) that

$$\begin{aligned}&2\kappa_0^2\|\Delta\|_2^2 \le \|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2\|\Delta\|_2 + \kappa\|\Delta\|_2^2\\ &+\|\Delta\|_2\|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\end{aligned}$$

When $\|\Delta\|_2 \ne 0$, we have

$$2\kappa_0^2\|\Delta\|_2 \le \|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2 + \kappa\|\Delta\|_2 + \|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i}_i}\|_2$$

$$\Rightarrow \|\Delta\|_2 \le \frac{\|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2 + \|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2}{2\kappa_0^2 - \kappa} \tag{34}$$

According to the definition of θ, it can be verified that $\theta((\hat{\mathbf{Z}}^i_{-i})_{k_{-i}},\kappa) = \max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa|\hat{\mathbf{Z}}_{ki} - b|\}$ for $k_{-i} \in \mathbf{U}^{-i}_i \cap \hat{\mathbf{S}}^{-i}_i$, and $\theta((\hat{\mathbf{Z}}^i_{-i})_{k_{-i}},\kappa) = \max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa b\}$ for $k_{-i} \in \mathbf{U}^{-i}_i \setminus \hat{\mathbf{S}}^{-i}_i$. Therefore,

$$\begin{aligned}&\|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2\\ &= \Big(\sum_{k\in\mathbf{U}_i\cap\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa|\hat{\mathbf{Z}}_{ki} - b|\})^2 +\\ &\sum_{k\in\mathbf{U}_i\setminus\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa b\})^2\Big)^{\frac{1}{2}}\end{aligned} \tag{35}$$

and it follows that

$$\begin{aligned}&\|\mathbf{Z}^{i*} - \hat{\mathbf{Z}}^i\|_2 = \|\Delta\|_2\\ &\le \frac{1}{2\kappa_0^2 - \kappa}\Bigg(\Big(\sum_{k\in\mathbf{U}_i\cap\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa|\hat{\mathbf{Z}}_{ki} - b|\})^2 +\\ &\sum_{k\in\mathbf{U}_i\setminus\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa b\})^2\Big)^{\frac{1}{2}} + \|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\Bigg)\end{aligned} \tag{36}$$

where $\tilde{\Delta}_{m_{-i}} = -(2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta)_{m_{-i}} = -2\lambda\mathbb{I}_{\mathbf{Z}^{i*}_m\hat{\mathbf{Z}}^i_m<0} - 0\mathbb{I}_{\mathbf{Z}^{i*}_m\hat{\mathbf{Z}}^i_m>0}$ for $m \in \hat{\mathbf{S}}_i \cap \mathbf{S}^*_i$. This proves the result of this theorem. □

Proof of Theorem 2. Let $\boldsymbol{Y} = \tilde{\boldsymbol{X}}$. By the proof of Lemma 1, we have

$$\|2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} + \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^i)\|_2 = 0$$

It follows that

$$\begin{aligned}&\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} + \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^i)\|_2\\ &= \|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} - 2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i}\\ &+2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} + \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^i)\|_2\\ &\le \|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} - 2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i}\|_2\\ &+\|2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} + \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^i)\|_2\\ &= \|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} - 2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i}\|_2\\ &\le \|2\boldsymbol{X}^{(-i)\top}(\boldsymbol{X}^{(-i)} - \boldsymbol{Y}^{(-i)})\tilde{\mathbf{Z}}^i_{-i}\|_2\\ &+\|2(\boldsymbol{X}^{(-i)} - \boldsymbol{Y}^{(-i)})^\top\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i}\|_2\end{aligned} \tag{37}$$

By $\tilde{F}(\mathbf{Z}^i) \le \tilde{F}(\mathbf{0})$, we have $\|\tilde{\mathbf{Z}}^i_{-i}\|_2 \le A$. Let $k_0 \ge 2$ and $p = k - k_0 \ge 4$. By Lemma 2, with probability at least $1 - 6e^{-p}$, $\|\boldsymbol{X} - \boldsymbol{Y}\|_2 \le C_{k,k_0}$. It follows from (37) that

$$\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} + \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^i)\|_2$$

$$\leq \sigma_{\max}(\boldsymbol{X})C_{k,k_0}A + C_{k,k_0}(\sigma_{\max}(\boldsymbol{X}) + C_{k,k_0})A$$
$$= C_{k,k_0}A(2\sigma_{\max}(\boldsymbol{X}) + C_{k,k_0})$$

Also, by Lemma 1,

$$\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\mathbf{Z}^{i*}_{-i} + \dot{\mathbf{P}}(\mathbf{Z}^{i*}_{-i})\|_2 = 0$$

Let $\Delta = \mathbf{Z}^{i*}_{-i} - \tilde{\mathbf{Z}}^{i}_{-i}$, $\tilde{\Delta} = \dot{\mathbf{P}}(\mathbf{Z}^{i*}) - \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^{i})$.

$$\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta + \tilde{\Delta}\|_2 \leq C_{k,k_0}A(2\sigma_{\max}(\boldsymbol{X}) + C_{k,k_0})$$

Now following the proof of Theorem 1, we have

$$\|\mathbf{Z}^{i*} - \tilde{\mathbf{Z}}^{i}\|_2 = \|\Delta\|_2$$
$$\leq \frac{1}{2\tau_0^2 - \tau}\Bigg(\Big(\sum_{k\in\mathbf{G}_i\cap\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa|\tilde{\mathbf{Z}}_{ki} - b|\})^2$$
$$+ \sum_{k\in\mathbf{G}_i\setminus\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa b\})^2\Big)^{\frac{1}{2}} + \|\mathbf{t}\|_2$$
$$+ C_{k,k_0}A(2\sigma_{\max}(\boldsymbol{X}) + C_{k,k_0})\Bigg) \quad (38)$$

□

7.2 MORE DETAILS IN THE PAPER AND MORE EXPERIMENTAL RESULTS

The SC baseline in this paper uses the self-tuning spectral clustering method (Zelnik-manor and Perona, 2005), and we choose this method due to its advantage of adaptively setting the kernel bandwidth for the Gaussian kernel similarity. More concretely, we construct a similarity matrix using Gaussian kernel, and the bandwidth of the Gaussian kernel similarity between two points is determined by the local statistics of the neighborhoods of these two points. We set the distance to the 7-th nearest neighbor as the local statistics, which is also the default choice suggested by the paper, then perform spectral clustering on such similarity matrix to obtain the clustering results for SC in Table 1. In addition, various sparse graph methods, including ℓ^1-Graph, ℓ^2-Rℓ^1-Graph and NRℓ^1-Graph, constructs a sparse graph upon which spectral clustering is applied to find the clusters.

It is worthwhile to mention the meaning of the condition that $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0$ for all $k \in \Lambda_i$ in (9). Let $k \in \Lambda_i$, if the number of point $\mathbf{x}_i$'s neighbors with zero k-th element of the sparse codes is larger than that with nonzero k-th element of the sparse codes, which indicates that the neighbors of $\mathbf{x}_i$ suggest that a zero k-th element of the sparse code of $\mathbf{x}_i$ is preferable, then $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0$ and $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}$ quantitatively measures the penalty if the sparse code element $\mathbf{Z}^i_k$ is nonzero while the neighbors of $\mathbf{x}_i$ suggest that $\mathbf{Z}^i_k = 0$ is preferable. The optimization helps point $\mathbf{x}_i$ make a sensible choice by considering the suggestion of its neighbors. We observe that $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0$ for all $k \in \Lambda_i$ happens in all the data sets used in the experiments.

The standard deviation values of the NMI by different clustering methods on the MNIST data is as follows. NRℓ^1-Graph-RP: 0.0118; NRℓ^1-Graph: 0.0137; ℓ^2-Rℓ^1-Graph: 0.0114; SMCE: 0.0166; ℓ^1-Graph: 0.0039; SC: 0.0071; KM: 0.0094. We have conducted paired t-test and conclude that both NRℓ^1-Graph and NRℓ^1-Graph-RP are statistically better than other baseline methods with p-value less than 0.05 in many cases. For example, the p-value of the paired t-test between the accuracy of NRℓ^1-Graph and SMCE is less than 0.05 on the COIL-20, COIL-100 and Yale-B data.

We also present clustering results on the first c clusters for COIL-100, CMU PIE and UMIST Face Data in Table 2, 3 and 4 respectively.

In order to investigate the parameter sensitivity of our model, namely how the performance of NRℓ^1-Graph varies with parameter γ and K, we vary γ and K and illustrate the result on the UMIST Face Database in Figure 2 and Figure 3 respectively in this supplementary. The performance of NRℓ^1-Graph is noticeably better than other competing algorithms over a relatively large range of both λ and K, which demonstrates the robustness of our algorithm with respect to the parameter settings. We also observe that a too small K (near to 1) results in under regularization, and a too big K (near to 15) or too big γ (close to 0.45) risks over regularization.

Table 2: Clustering Results on COIL-100 Database. c in the left column is the cluster number, i.e. the first c clusters of the entire data are used for clustering. c has the same meaning in the following tables.

COIL-100 / # Clusters	Measure	KM	SC	ℓ^1-Graph	SMCE	ℓ^2-Rℓ^1-Graph	NRℓ^1-Graph
c = 20	AC	0.5875	0.4493	0.5340	0.6208	0.6681	**0.9236**
	NMI	0.7448	0.6680	0.7681	0.7993	0.7933	**0.9610**
c = 40	AC	0.5774	0.4160	0.5819	0.6028	0.5944	**0.8771**
	NMI	0.7662	0.6682	0.7911	0.7919	0.7991	**0.9504**
c = 60	AC	0.5330	0.3225	0.5824	0.5877	0.6009	**0.7808**
	NMI	0.7603	0.6254	0.8310	0.7971	0.8310	**0.8924**
c = 80	AC	0.5062	0.3135	0.5380	0.5740	0.5632	**0.8177**
	NMI	0.7458	0.6071	0.8034	0.7931	0.8036	**0.9208**
c = 100	AC	0.4928	0.2833	0.5310	0.5625	0.5625	**0.7846**
	NMI	0.7522	0.5913	0.8015	0.8057	0.8059	**0.9238**

Table 3: Clustering Results on CMU PIE Data

CMU PIE / # Clusters	Measure	KM	SC	ℓ^1-Graph	SMCE	ℓ^2-Rℓ^1-Graph	NRℓ^1-Graph
c = 20	AC	0.1327	0.1288	0.2435	0.2321	0.3212	**0.3606**
	NMI	0.1220	0.1342	0.2895	0.2942	0.4007	**0.4876**
c = 40	AC	0.1054	0.0867	0.2443	0.1752	0.3412	**0.3555**
	NMI	0.1534	0.1422	0.3344	0.2976	0.4789	**0.4834**
c = 68	AC	0.0829	0.0718	0.2318	0.1603	0.3012	**0.3190**
	NMI	0.1865	0.1760	0.3378	0.3406	**0.5121**	0.4993

Table 4: Clustering Results on UMIST Face Data

UMIST Face / # Clusters	Measure	KM	SC	ℓ^1-Graph	SMCE	ℓ^2-Rℓ^1-Graph	NRℓ^1-Graph
c = 8	AC	0.4330	0.4789	0.4930	0.4695	0.5399	**0.6056**
	NMI	0.5373	0.5236	0.5516	0.5744	0.5721	**0.5749**
c = 12	AC	0.4478	0.4655	0.5195	0.4955	0.5706	**0.6246**
	NMI	0.6121	0.6049	0.6086	0.6445	0.6994	**0.7244**
c = 16	AC	0.4297	0.4539	0.4539	0.4747	0.4700	**0.6982**
	NMI	0.6343	0.6453	0.6582	0.6909	0.6714	**0.7816**
c = 20	AC	0.4216	0.4174	0.4417	0.4452	0.4991	**0.6765**
	NMI	0.6377	0.6095	0.6489	0.6641	0.6893	**0.7982**

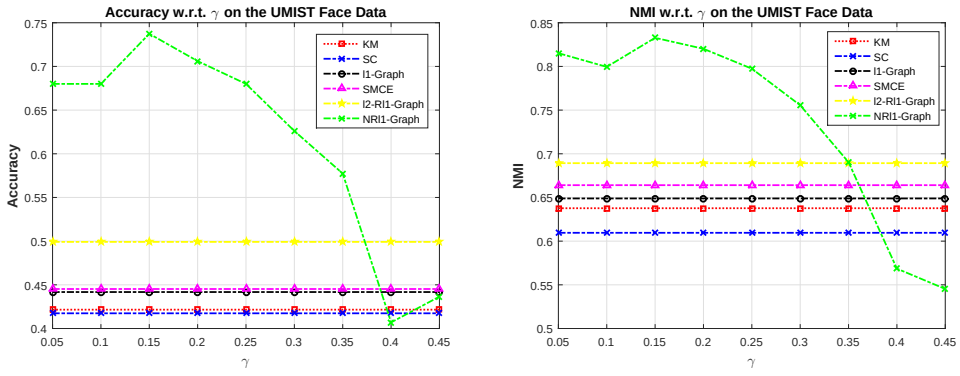

Figure 2: Clustering performance with different values of γ, i.e. the weight for the regularization term in NRℓ^1-Graph, on the UMIST Face Data. Left: Accuracy; Right: NMI

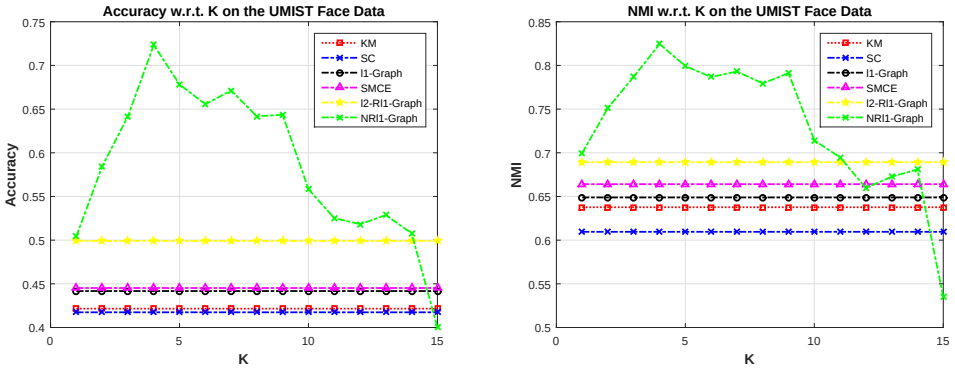

Figure 3: Clustering performance with different values of K, i.e. the number of nearest neighbors for the regularization term in NRℓ^1-Graph, on the UMIST Face Data. Left: Accuracy; Right: NMI