Supplemental Material to "Identifying Alzheimer's Disease-Related Brain Regions from Multi-Modality Neuroimaging Data using Sparse Composite Linear Discrimination Analysis"

I – How does the formulation (5) serve the purpose of the composite parameterization

It is interesting to see how the penalty function, $\sqrt{\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)}\right|}$, will serve the purpose of the composite parameterization, i.e., common information and specific information can be estimated separately and simultaneously. To illustrate this point, let's look at LASSO first. It has been pointed out in [40] that, the variable selection ability of LASSO depends on the singularity of the parameters when the parameters are equal to zero. For example, consider the LASSO problem [27], whose penalized likelihood function is $\mathrm{LS}(\boldsymbol{\beta})+\lambda\sum_i|\beta_i|$, and the ridge regression, whose penalized likelihood function is $\mathrm{LS}(\boldsymbol{\beta})+\lambda\sum_i\beta_i^2$. Here, $\mathrm{LS}(\boldsymbol{\beta})$ represents the least square error function. It is well known that LASSO is capable of performing variable selection, i.e., forcing many parameters in the $\boldsymbol{\beta}$ estimate to be exactly zero, while ridge regression is not. The reason is that the penalized likelihood function of LASSO is singular when $\beta_i=0$ [40]. Similarly, the penalty function used in SCLDA is singular when all the variables under the square root are zero (corresponding to $\delta_k=0$), which facilitates the variable selection on the common information level. Furthermore, this penalty function is also singular when any variable under the square root is zero (corresponding to any $\gamma_{k,l}^{(m)}=0$), which further facilitates the variable selection on the specific information level. It is worth mentioning that L2/L1 regularization employed in multitask feature selection methods, which will be $\sqrt{\sum_{l=1}^{q}\sum_{m=1}^{M}\left(\theta_{k,l}^{(m)}\right)^2}$ in our case, is only singular when $\sum_{l=1}^{q}\sum_{m=1}^{M}\left(\theta_{k,l}^{(m)}\right)^2=0$. This tends to produce an "all-in-all-out" solution, i.e., when one variable is selected, all the other variables in the square root will be selected, since there is no more singularity. Thus, the L2/L1 regularization is not capable of performing the variable selection on the specific information level, which limits its practical use in some applications.

II – The details on the DC programming for solving (5)

The optimization problem (5) is a non-convex optimization problem that is difficult to solve. We address this problem by using an iterative two-stage procedure known as Difference of Convex functions (DC) programming [39]. The basic idea behind the DC programming is that, at the first stage of every iteration, a surrogate convex objective function is proposed to bound the non-convex objective function at the current solution; then, at the second stage, a new solution is obtained by maximizing this surrogate convex objective function. This process iterates until a certain convergence rule is met. It is worth mentioning that the DC programming shares the same spirit as the Expectation-Maximization (EM) algorithm or Minorization-Maximization (MM) algorithms that have been widely used in statistics and machine learning.

We adopt the DC algorithm to solve (5) following [39], in which the DC programming is to solve a least-square problem with non-convex penalty terms. In our problem, the essential task is to find a decomposition of $\sum_k\sqrt{\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)}\right|}$ as a sum of two convex functions. As suggested in [39], the decomposition, $\sqrt{\theta_{k,l}^{(m)}}=\left|\theta_{k,l}^{(m)}\right|-\left(\left|\theta_{k,l}^{(m)}\right|-\sqrt{\theta_{k,l}^{(m)}}\right)$, can be used for the non-convex penalty term $\sqrt{\theta_{k,l}^{(m)}}$. Use

$$\sqrt{\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)}\right|} = \sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)}\right| - \left(\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)}\right| - \sqrt{\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)}\right|}\right),$$

as a decomposition for $\sqrt{\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)}\right|}$. Based on the theory developed in [39], this decomposition results in the following objective function

$$\widetilde{\boldsymbol{\Theta}}^{(t+1)} = \operatorname{argmin}_{\boldsymbol{\Theta}} l_3\left(\boldsymbol{\Theta}|\left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(M)}\right\}\right) =$$

$$\operatorname{argmin}_{\boldsymbol{\Theta}}\left\{-l_0\left(\boldsymbol{\Theta}|\left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(M)}\right\}\right) + \sum_k \lambda_k^{(t+1)} \sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)}\right|\right\}, \quad (6)$$

as the surrogate convex objective function, where

$$\lambda_k^{(t+1)} = \lambda - h'(z)|_{z=\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)(t)}\right|}, \; h(z) = \lambda\left(z - \sqrt{z}\right),$$

$\widetilde{\boldsymbol{\Theta}}^{(t+1)}$ is the solution at iteration $t+1$ and $\theta_{k,l}^{(m)(t)}$ is a corresponding element. It is shown that this decomposition produces a surrogate convex objective function with L1- penalty, which can be solved by many existing efficient algorithms developed for LASSO-type problems. A complete procedure for the DC programming is depicted in Figure 1.

Initialize: Let $t = 0$;
Repeat
Calculate $\lambda_k^{(t)} = \lambda - h'(z)|_{z=\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)(t)}\right|}$;
for $1 \le k, l \le p,\ 1 \le m \le M$;
Solve (6) and get $\widetilde{\boldsymbol{\Theta}}^{(t)}$;
Let $t = t + 1$;
Until converge

Figure 1: The DC programming for solving (5)

The optimization problem (6) is a standard L1-regularization type problem, whose objective function is a sum of a smooth likelihood/lease square error function and an L1- penalty on the parameters. This problem can be solved by many efficient numeric algorithms in the literature [25, 26]. In our case, the two-metric method is employed [25].

It has been pointed out that the DC programming for solving many non-convex regularization problems can be closely linked to the adaptive Lasso formulation [39]. To illustrate this point in our case, we note that

$$h'(z)|_{z=\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)}\right|} = \lambda - \frac{\lambda}{2\left(\sqrt{\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)}\right|} + \epsilon\right)}$$

is in our DC programming, where ϵ is a user-specified number for numerical stability consideration. This produces a new regularization parameter for iteration $t+1$,

$$\lambda_k^{(t+1)} = \frac{\lambda}{2\left(\sqrt{\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)(t)}\right|} + \epsilon\right)},$$

which is inversely proportional to the magnitude of $\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\theta_{k,l}^{(m)(t)}\right|$. This implies that, at each iteration, the DC programming essentially reweights the regularization parameters of each

L1-regularization type problem (6). Specifically, at iteration $t+1$, the new regularization parameters associated with the zero $\theta_{k,l}^{(m)}$'s identified at the previous iteration will increase drastically, while the new regularization parameters associated with the non-zero $\theta_{k,l}^{(m)}$'s identified at the previous iteration will decrease proportionally to the sum of the absolute magnitudes of these $\theta_{k,l}^{(m)}$'s. which belong to the same variable. In this manner, the shrinkage effect imposed on the non-zero $\theta_{k,l}^{(m)}$'s by the L1-regularization is effectively alleviated. This explains why the proposed SCLDA has the capability of preserving weak-effect features.

III – Simulation procedure

For a given combination of values of these parameters, simulation data can be generated in the following way: First, we generate a $1\times p$ vector, $\boldsymbol{\beta}$, with l elements being randomly selected to be 1 and the other elements being zero. Second, to generate a feature vector for dataset i, $\boldsymbol{\beta}_i$, we randomly change $(100-s)\%$ of the non-zero elements of $\boldsymbol{\beta}$ to be zero, and, at the same time, we randomly change the same number of zero elements to be 1. The nonzero elements of $\boldsymbol{\beta}_i$ correspond to the features of dataset i. In this manner, the larger the $s\%$, the more overlapping of the features across the data sources. We further randomly pick up half of the nonzero elements of each $\boldsymbol{\beta}_i$ and modify them in such a way: if the element is 1, then its new value is sampled from the uniform distribution U[0,0.5]; if the element is -1, then its new value is sampled from the uniform distribution U[-0.5,0]. This is to mimic the true situation in the real application in section 5, where a large portion of weak-effect features are present. After that, we use the resulting $\boldsymbol{\beta}_i$ and $-\boldsymbol{\beta}_i$ as the mean vectors of two classes of dataset i (for simplicity, we only investigate 2-class problems for the simulation). And, as we assume heterogeneous covariance matrices for different classes, two covariance matrices are independently generated from a Wishart distribution with degrees of freedom to be n and scale matrix to be an Identity matrix. Each covariance matrix is further diagonalized to make the diagonal elements being 0.2 (the corresponding signal-to-noise ratio is about 5). With these mean vectors and covariance matrices, we generate the dataset i by sampling one sub-dataset for each class from its corresponding multivariate Gaussian distribution, and then combining these two sub-datasets as dataset i.

IV – Compare SCLDA, SLDA and MSLDA under different sample sizes

Furthermore, we conduct an experiment to compare SCLDA, SLDA and MSLDA under different sample sizes. As shown in Figure 3, SCLDA performs significantly better when sample sizes are small, which confirms that SCLDA is more statistically efficient in recovering the features.

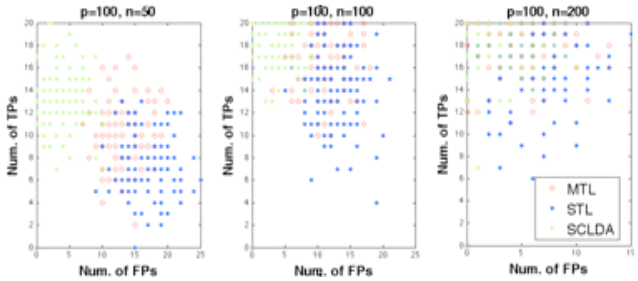

Figure 2: Average numbers of TPs vs FPs for proposed SCLDA (green symbols "+"), SLDA (blue symbols "*") and MSLDA (red symbols "o") with task relatedness $s\% = 90\%, m = 2$

V– Selection of λ and q

We first discuss the optimal selection of λ and q for the case when there is only one data source. This optimal selection can be obtained by maximizing Akaike's Information Criteria (AIC) [1]:

$$AIC = s(\lambda, q) - \frac{d}{n_{avg}},$$

where $s(\lambda, q)$ is the cross validation classification accuracy associated with λ and q, d is the number of nonzero parameters in $\widetilde{\boldsymbol{\Theta}}$, and n_{avg} is the average number of observations per class. For the general case where there are M data sources, the optimal selection of λ and q can be obtained by maximizing the average AIC, which is:

$$AIC_{avg} = \sum_{i=1}^{M} AIC_i / M.$$

where AIC_i is the AIC value for the i^{th} data source. Both the simulation studies in section 4 and real application in section 5 reveal that this criterion can produce accurate and meaningful model selection results.

[1] Chiang, L., Russell, E.R. 2001. *Fault Detection and Diagnosis in Industrial Systems.* Springer, London.

VI – Proof of Theorem 1

Proof: We first demonstrate that any local optima of $l_1\left(\boldsymbol{\Gamma}, \boldsymbol{\Psi} | \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(M)}\right\}\right)$, $\hat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Psi}}$, corresponds to a local optima of

$l_1'\left(\boldsymbol{\Gamma}, \boldsymbol{\Psi} | \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(M)}\right\}\right) =$

$\sum_{m=1}^{M}\left\{\frac{N^{(m)}}{2}\log\left|\boldsymbol{\theta}_{p-q}^{(m)\,T}\mathbf{T}^{(m)}\boldsymbol{\theta}_{p-q}^{(m)}\right| + \sum_{j=1}^{J}\frac{N_j}{2}\log\boldsymbol{\theta}_q^{(m)}\mathbf{W}_j^{(m)}\boldsymbol{\theta}_q^{(m)} - N^{(m)}\log\left|\boldsymbol{\theta}^{(m)}\right|\right\} + \sum_k \delta_k +$

$\eta\sum_k\sum_{l=1}^{q}\sum_{m=1}^{M}\gamma_{k,l}^{(m)}$, which is denoted $\tilde{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\Psi}}$. To show this, we insert $\tilde{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\Psi}}$ into l_1', which is

$l_1'\left(\tilde{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\Psi}} | \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(M)}\right\}\right)$,

$= \sum_{m=1}^{M}\left\{\frac{N^{(m)}}{2}\log\left|\tilde{\boldsymbol{\theta}}_{p-q}^{(m)\,T}\mathbf{T}^{(m)}\tilde{\boldsymbol{\theta}}_{p-q}^{(m)}\right| + \sum_{j=1}^{J}\frac{N_j}{2}\log\tilde{\boldsymbol{\theta}}_q^{(m)}\mathbf{W}_j^{(m)}\tilde{\boldsymbol{\theta}}_q^{(m)} - N^{(m)}\log\left|\tilde{\boldsymbol{\theta}}^{(m)}\right|\right\} + \sum_k \tilde{\delta}_k +$

$\eta\sum_k\sum_{l=1}^{q}\sum_{m=1}^{M}\tilde{\gamma}_{k,l}^{(m)}$, which can be further written as

$l_1'\left(\tilde{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\Psi}} | \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(M)}\right\}\right) =$

$\sum_{m=1}^{M}\left\{\frac{N^{(m)}}{2}\log\left|\tilde{\boldsymbol{\theta}}_{p-q}^{(m)\,T}\mathbf{T}^{(m)}\tilde{\boldsymbol{\theta}}_{p-q}^{(m)}\right| + \sum_{j=1}^{J}\frac{N_j}{2}\log\tilde{\boldsymbol{\theta}}_q^{(m)}\mathbf{W}_j^{(m)}\tilde{\boldsymbol{\theta}}_q^{(m)} - N^{(m)}\log\left|\tilde{\boldsymbol{\theta}}^{(m)}\right|\right\} + \lambda_1\sum_k\frac{\tilde{\delta}_k}{\lambda_1} +$

$\lambda_2\sum_k\sum_{l=1}^{q}\sum_{m=1}^{M}\lambda_1\tilde{\gamma}_{k,l}^{(m)}$,

$= l_1\left(\lambda_1\tilde{\boldsymbol{\Gamma}}, \frac{\tilde{\boldsymbol{\Psi}}}{\lambda_1} | \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(M)}\right\}\right) \geq l_1\left(\hat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\Psi}} | \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(M)}\right\}\right)$. The last inequality holds since $\hat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Psi}}$ is a local optima of l_1. Similarly, we insert $\hat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\Psi}}$ into l_1 and we can prove

$$l_1\left(\hat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\Psi}} | \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(M)}\right\}\right) = l_1'\left(\frac{\hat{\boldsymbol{\Gamma}}}{\lambda_1}, \lambda_1\widehat{\boldsymbol{\Psi}} | \left\{\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(M)}\right\}\right) \geq l_1'\left(\tilde{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\Psi}} | \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(M)}\right\}\right),$$

Therefore, we have

$$l_1\left(\hat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\Psi}} | \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(M)}\right\}\right) = l_1'\left(\frac{\hat{\boldsymbol{\Gamma}}}{\lambda_1}, \lambda_1\widehat{\boldsymbol{\Psi}} | \left\{\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(M)}\right\}\right) = l_1'\left(\tilde{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\Psi}} | \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(M)}\right\}\right).$$

This has an implication that $\left\{\frac{\hat{\boldsymbol{\Gamma}}}{\lambda_1}, \lambda_1\widehat{\boldsymbol{\Psi}}\right\}$ is also a local optimizer of $l_1'\left(\boldsymbol{\Gamma}, \boldsymbol{\Psi} | \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(M)}\right\}\right)$ and $\tilde{\delta}_k\tilde{\gamma}_{k,l}^{(m)} = \hat{\delta}_k\hat{\gamma}_{k,l}^{(m)}$, $\eta = \lambda_1\lambda_2$.

Now we demonstrate that any local optima of $l_1'\left(\mathbf{\Gamma},\mathbf{\Psi}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)$ corresponds to a local optima of $l_2\left(\mathbf{\Theta}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)$, $\widehat{\mathbf{\Theta}}$. Using the same idea above, we can obtain that

$l_1'\left(\tilde{\mathbf{\Gamma}},\widetilde{\mathbf{\Psi}}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)$,

$\geq \sum_{m=1}^{M}\left\{\frac{N^{(m)}}{2}\log\left|\widetilde{\boldsymbol{\theta}}_{p-q}^{(m)\,T}\mathbf{T}^{(m)}\widetilde{\boldsymbol{\theta}}_{p-q}^{(m)}\right|+\sum_{j=1}^{J}\frac{N_j}{2}\log\widetilde{\boldsymbol{\theta}}_{q}^{(m)}\mathbf{W}_j^{(m)}\widetilde{\boldsymbol{\theta}}_{q}^{(m)}-N^{(m)}\log\left|\widetilde{\boldsymbol{\theta}}^{(m)}\right|\right\}+$

$2\ \sum_k\sqrt{\eta\tilde{\delta}_k\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\tilde{\gamma}_{k,l}^{(m)}\right|}$,

$=l_2\left(\widetilde{\mathbf{\Theta}}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)\geq l_2\left(\widehat{\mathbf{\Theta}}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)$.

On the other hand, let $\widehat{\mathbf{\Psi}}=\left\{\hat{\delta}_k=\sqrt{\eta\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\hat{\theta}_{k,l}^{(m)}\right|},1\leq k\leq p\right\}$ and $\widehat{\mathbf{\Gamma}}=\left\{\hat{\gamma}_{k,l}^{(m)}=\frac{\hat{\theta}_{k,l}^{(m)}}{\hat{\delta}_k},1\leq k\leq p,\ 1\leq l\leq p,1\leq m\leq M\right\}$, then we can obtain

$l_2\left(\widehat{\mathbf{\Theta}}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)$,

$=\sum_{m=1}^{M}\left\{\frac{N^{(m)}}{2}\log\left|\widehat{\boldsymbol{\theta}}_{p-q}^{(m)\,T}\mathbf{T}^{(m)}\widehat{\boldsymbol{\theta}}_{p-q}^{(m)}\right|+\sum_{j=1}^{J}\frac{N_j}{2}\log\widehat{\boldsymbol{\theta}}_{q}^{(m)}\mathbf{W}_j^{(m)}\widehat{\boldsymbol{\theta}}_{q}^{(m)}-N^{(m)}\log\left|\widehat{\boldsymbol{\theta}}^{(m)}\right|\right\}+$

$\sum_k\sqrt{\eta\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\hat{\theta}_{k,l}^{(m)}\right|}+\sqrt{\eta}\sum_k\sum_{l=1}^{q}\sum_{m=1}^{M}\frac{\left|\hat{\theta}_{k,l}^{(m)}\right|}{\sqrt{\sum_{l=1}^{q}\sum_{m=1}^{M}\left|\hat{\theta}_{k,l}^{(m)}\right|}}$,

$=\sum_{m=1}^{M}\left\{\frac{N^{(m)}}{2}\log\left|\widehat{\boldsymbol{\theta}}_{p-q}^{(m)\,T}\mathbf{T}^{(m)}\widehat{\boldsymbol{\theta}}_{p-q}^{(m)}\right|+\sum_{j=1}^{J}\frac{N_j}{2}\log\widehat{\boldsymbol{\theta}}_{q}^{(m)}\mathbf{W}_j^{(m)}\widehat{\boldsymbol{\theta}}_{q}^{(m)}-N^{(m)}\log\left|\widehat{\boldsymbol{\theta}}^{(m)}\right|\right\}+\ \sum_k\hat{\delta}_k+$

$\lambda\sum_k\sum_{l=1}^{q}\sum_{m=1}^{M}\hat{\gamma}_{k,l}^{(m)}$,

$=l_1'\left(\widehat{\mathbf{\Theta}}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)\geq l_1'\left(\widetilde{\mathbf{\Theta}}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)$.

Thus, we have
$l_2\left(\hat{\mathbf{\Gamma}},\widehat{\mathbf{\Psi}}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)=l_1'\left(\hat{\mathbf{\Gamma}},\widehat{\mathbf{\Psi}}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)=l_1'\left(\tilde{\mathbf{\Gamma}},\widetilde{\mathbf{\Psi}}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)$, which implies that $\left\{\hat{\mathbf{\Gamma}},\widehat{\mathbf{\Psi}}\right\}$ is also a local optimizer of $l_2\left(\mathbf{\Gamma},\mathbf{\Psi}|\left\{\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\dots,\mathbf{Z}^{(M)}\right\}\right)$, with $\lambda=2\sqrt{\eta}=2\sqrt{\lambda_1\lambda_2}$ and $\tilde{\delta}_k\tilde{\gamma}_{k,l}^{(m)}=\hat{\theta}_{k,l}^{(m)}$.