Mutual Transfer Learning for Massive Data

Ching-Wei Cheng [1] **Xingye Qiao** [2] **Guang Cheng** [1]

Abstract

In the transfer learning problem, the target and the source data domains are typically known. In this article, we study a new paradigm called mutual transfer learning where among many heterogeneous data domains, every data domain could potentially be the target of interest, and it could also be a useful source to help the learning in other data domains. However, it is important to note that given a target not every data domain can be a successful source; only data sets that are similar enough to be thought as from the same population can be useful sources for each other. Under this mutual learnability assumption, a confidence distribution fusion approach is proposed to recover the mutual learnability relation in the transfer learning regime. Our proposed method achieves the same oracle statistical inferential accuracy as if the true learnability structure were known. It can be implemented in an efficient parallel fashion to deal with large-scale data. Simulated and real examples are analyzed to illustrate the usefulness of the proposed method.

1. Introduction

Transfer Learning (TL) aims to leverage knowledge from a related domain (called source domain) to improve the predictive and inferential performance in a target domain, on which the availability of the training data is limited. Source data are drawn from a source distribution P on $\mathcal{S}^P$, and a relatively small quantity of labeled or unlabeled data are from the target distribution Q on $\mathcal{S}^Q$. We focus on the homogeneous TL setting with $\mathcal{S}^P = \mathcal{S}^Q$, on which P and Q are different but related distributions. See surveys of TL in Pan & Yang (2009); Zhuang et al. (2019).

In the TL problem, it is typically known *a priori* which data domain is the target and which data will be used as the source. However, this is not necessarily the case in many contemporary applications. For example, one of the 4V's for Big Data is variety. Big Data are often collections of multiple data sets (called "data units" hereafter) from different data domains. These data units are of the same nature but are distributed differently, due to, for example, that they are collected in different time periods, at different locations, or possibly using different data collection devices. It hence may be desirable to leverage TL techniques to transfer knowledge between these related data domains so as to improve the predictive performance or statistical inference in a target domain. However, in the applications we consider in this article and many other cases, literally, every data domain could potentially be a target of interest, and every data domain could potentially be useful as a source to help the learning performance in other domains. In other words, the target and source domains are not known *a priori*.

Moreover, due to the variety among the data units, it is quite possible that given a target domain, although all data units could be useful for transferring knowledge to some extent, some data units are less useful as source data than the others. Hence, a second feature of the problem considered here is that one must identify source data units that are more useful for each possible target. While the existing TL literature is fairly diverse and extensive, fewer efforts have been made to make successful transfers in this setting.

In this article, we consider a mutual transfer learning (MTL) framework in which every data unit could potentially be the target. The goal of MTL is to simultaneously identify useful source data units for each target domain and use such information to improve learning performance. We propose a mutual learnability assumption, which suggests that only data sets that are similar enough to be thought as from the same population can be useful sources for each other. Under this assumption, we propose a confidence distribution fusion approach to recover the mutual learnability structure in the transfer learning regime, using regression problem as a working example. Incorporating the recovered learnability structure, we fit a statistical model for both predictive and inferential purposes. Our proposed method achieves the same oracle statistical inferential accuracy as if the true learnability structure were known.

[1]Department of Statistics, Purdue University [2]Department of Mathematical Sciences, Binghamton University. Correspondence to: Xingye Qiao <qiao@math.binghamton.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

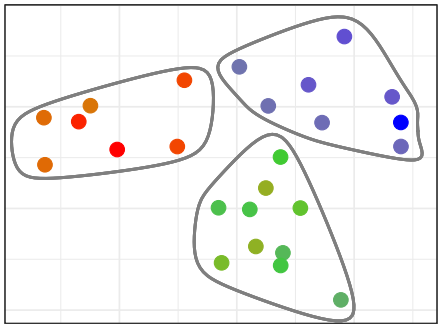

Figure 1. Illustration of the two-layer learnability. The data units from the same subgroup have more learnability between them.

Suppose there are M data units from M related domains. For each potential target domain, there exist $M-1$ source data units to transfer from. We consider a regression problem with two sets of features: the global features and the heterogeneous features. While all the data units could be useful to learn the role of the global features in explaining the response variable, it is unreasonable to assume that all the $M-1$ source data units are equally useful to learn the heterogeneous features. Specifically, given the target, there may be a subgroup of data units that are more useful than the others. This could be explained by a two-layer learnability model, shown in Figure 1, in which different behaviors for data units are coded by colors. The data units bounded together are assumed from the same subgroup, within which it is easier to transfer useful information to learn the heterogeneous features, hence more learnability. Data from different groups are significantly different, and the learnability between groups are limited to only the global features.

To motivate our statistical model, we analyzed NOAA's nClimDiv database. It consists of monthly temperature, precipitation, and several indices for drought severity, and contains a total of $N = 503{,}616$ observations from $M = 344$ climate divisions. Each climate division is viewed as a data domain. The monthly average temperature is the response of interest, and there are 8 features within which we have designated 5 as global features and 3 heterogeneous features. Our proposed method has identified 5 subgroups shown in different colors in Figure 2. Knowledge about the global features can be learned from all climate divisions, while that about heterogeneous features can only be learned from identified source data in the same subgroup. Whether a given feature is global or heterogeneous is typically suggested by domain knowledge or preliminary inspection of the data. In this real data example, we checked if the kernel density for the coefficient estimates of a feature have a multimodal distribution to determine whether it is global.

We use the following statistical model to characterize this two-layer structure. Suppose that M data units come from S exclusive subgroups, and that the i-th data unit consists of

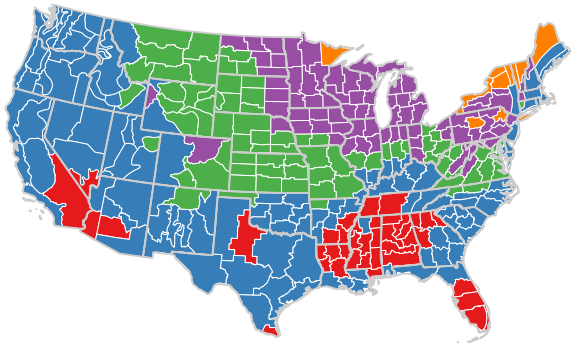

Figure 2. Learnability recovered for the 344 climate divisions.

n_i observations. Response Y_i from the ith data domain is

$$\begin{aligned}\mathsf{Y}_i &= \mathbf{X}^\top\boldsymbol{\beta}_0 + \mathbf{Z}^\top\boldsymbol{\vartheta}_i + \varepsilon \\ &= \mathbf{X}^\top\boldsymbol{\beta}_0 + \mathbf{Z}^\top\boldsymbol{\theta}_{i0} + \mathbf{Z}^\top\boldsymbol{u}_i + \varepsilon, \quad i = 1,\ldots,M, \quad (1)\end{aligned}$$

where the global features $\mathbf{X} \in \mathbb{R}^p$ and the heterogeneous features $\mathbf{Z} \in \mathbb{R}^q$ correspond to coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{\vartheta}_i$, which have different learnability. $\boldsymbol{\beta}_0$ is the same among all data domains and all data units can be used to learn it. $\boldsymbol{\vartheta}_i = \boldsymbol{\theta}_{i0} + \boldsymbol{u}_i \in \mathbb{R}^q$ represents a unit-specific effect. Its first component $\boldsymbol{\theta}_{i0}$ is equal among data units in the same subgroup as the ith data unit. That is, $\boldsymbol{\theta}_{i0} = \boldsymbol{\theta}_{j0}$, if and only if the i-th and j-th data units are from the same subgroup. Knowledge about $\boldsymbol{\theta}_{i0}$ can only be transferred between data units in the same subgroup. The second term $\boldsymbol{u}_i$ explains the subtle difference between data units in the same subgroup, as seen from Figure 1. This random effect $\boldsymbol{u}_i$ cannot be transferred. We assume that $\{\boldsymbol{u}_i\}_{i=1}^M$ are independent and identically distributed with $\mathsf{E}(\boldsymbol{u}_i) = \mathbf{0}$ and $\mathsf{Var}(\boldsymbol{u}_i) = \boldsymbol{\Psi}$ with the minimum eigenvalue $\tau = \lambda_{\min}(\boldsymbol{\Psi}) > 0$. Distinct values of $\{\boldsymbol{\theta}_{i0}\}_{i=1}^M$ are denoted as $\{\boldsymbol{\alpha}_{s0}\}_{s=1}^S$. With a membership label $L_i \in \{1,\ldots,S\}$, indicating which subgroup unit i belongs to, i.e., $\boldsymbol{\theta}_{i0} = \boldsymbol{\alpha}_{L_i 0}$, (1) is re-formulated as

$$\mathsf{Y}_i = \mathbf{X}^\top\boldsymbol{\beta}_0 + \mathbf{Z}^\top\boldsymbol{\alpha}_{L_i 0} + \mathbf{Z}^\top\boldsymbol{u}_i + \varepsilon, \; i = 1,\ldots,M. \quad (2)$$

Our goal in this article is two-fold. First, we aim to recover the learnability structure, that is, the subgroup membership for each data unit. Second, we aim to transfer knowledge to improve learning performance for all the data domains by incorporating this revealed structure. We develop a *confidence distribution (CD) fusion approach*. Specifically, we first obtain individual unit estimates, and then combine their CD densities, as defined in Liu et al. (2015), using a pairwise concave fusion penalty for $\boldsymbol{\theta}_i$'s. We prove that the resulting estimate is exactly the same as that based on the full-data method, which is often considered as gold standard in the meta-analysis (e.g. Debray et al., 2015). However, the full-data method has much larger computational costs as discussed in Section 3.2.

The proposed estimator is proven to asymptotically achieve the highest statistical inferential accuracy as if the true learnability structure were known. Such an oracle result holds when the minimum subsample size diverges fast enough and a minimal signal condition is met. The above learnability recovery issue was not considered in Liu et al. (2015).

An alternating direction method of multipliers (ADMM, Boyd et al., 2010) algorithm was established for the proposed method. The derived algorithm is particularly suitable for large-scale data based on parallel computing. The empirical performance of the proposed approach is examined through simulation studies, demonstrating that the most reliable results are produced under the minimax concave penalty (MCP, Zhang, 2010).

The rest of the article is organized as follows. In Section 2, we formalize the proposed statistical model (2) and develop a CD fusion-based MTL approach to recovery the learnability structure and mutually transfer knowledge to estimate the coefficients. We derive an ADMM algorithm to implement the proposed MTL approach and then analyze some computational considerations in Section 3. Theoretical properties of the proposed estimator are established in Section 4. The finite-sample properties of the proposed approach are evaluated in Section 5 via simulation experiments. Section 6 investigates the nClimDiv database to illustrate the practical usefulness of the proposed method. Detailed proofs to all theoretical results are provided in the supplementary material.

Notations. Let $\|\boldsymbol{v}\| \triangleq \sqrt{\boldsymbol{v}^\top \boldsymbol{v}}$ and $\|\boldsymbol{v}\|_\infty \triangleq \max_{1\le l\le r}|v_l|$ be the L_2 and L_∞ norms of a vector $\boldsymbol{v} = (v_1,\ldots,v_r)^\top$. For a symmetric matrix $\boldsymbol{M}$, let $\lambda_{\min}(\boldsymbol{M})$ and $\lambda_{\max}(\boldsymbol{M})$ denote its minimum and maximum eigenvalues. For a matrix $\boldsymbol{M} \in \mathbb{R}^{r\times d}$, let $[\boldsymbol{M}]_{jk}$ be its (j,k)-th element, $[\boldsymbol{M}]_{j\cdot}$ be its j-th row vector, and $[\boldsymbol{M}]_{\cdot k}$ be its k-th column vectors. Let $\|\boldsymbol{M}\| \triangleq \max_{\boldsymbol{v}\in\mathbb{R}^d,\|\boldsymbol{v}\|=1}\|\boldsymbol{M}\boldsymbol{v}\|$, $\|\boldsymbol{M}\|_\infty \triangleq \max_{1\le j\le r}\sum_{k=1}^d |[\boldsymbol{M}]_{jk}|$, and $\|\boldsymbol{M}\|_{\max} \triangleq \max_{1\le j\le r,1\le k\le d}|[\boldsymbol{M}]_{jk}|$. If $\boldsymbol{M}$ has full column rank, define $\boldsymbol{P}_{\boldsymbol{M}} \triangleq \boldsymbol{M}(\boldsymbol{M}^\top\boldsymbol{M})^{-1}\boldsymbol{M}^\top$ and $\boldsymbol{P}_{\boldsymbol{M}}^\perp \triangleq \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{M}}$. For positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \gg b_n$ if $a_n^{-1}b_n = o(1)$. Let $a\wedge b \triangleq \min(a,b)$ for any $a,b\in\mathbb{R}$.

2. Methodology

We formalize the two-layer learnability framework, and introduce a confidence distribution fusion approach for learnability recovery and mutual transfer learning.

2.1. Two-Layer Learnability Structure

The n_i data points in the ith unit can be expressed as

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta}_0 + \boldsymbol{z}_i\boldsymbol{\theta}_{i0} + \boldsymbol{z}_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i, \quad i = 1,\ldots,M, \tag{3}$$

where $\boldsymbol{y}_i = (y_{i1},\ldots,y_{in_i})^\top$, $\boldsymbol{x}_i$ ($\boldsymbol{z}_i$) is an $n_i\times p$ ($n_i\times q$) data matrix for the global (heterogeneous) features, and $\boldsymbol{\varepsilon}_i$ is the vector of noises with zero mean and covariance $\sigma_\varepsilon^2\boldsymbol{I}_{n_i}$. Let $N = \sum_{i=1}^M n_i$ be the total sample size of the full data, which can be expressed as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{Z}\boldsymbol{\Theta}_0 + \boldsymbol{Z}\boldsymbol{U} + \boldsymbol{\mathcal{E}}, \tag{4}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_i^\top)_{i=1,\ldots,M}^\top \in \mathbb{R}^N$ is the stacked response, $\boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_i^\top)_{i=1,\ldots,M}^\top \in \mathbb{R}^N$ is the stacked error, $\boldsymbol{X} = (\boldsymbol{x}_i^\top)_{i=1,\ldots,M}^\top$ and $\boldsymbol{Z} = \mathrm{diag}\left[(\boldsymbol{z}_i)_{i=1,\ldots,M}\right]$ are stacked $N\times p$ and $N\times Mq$ data matrices, and $\boldsymbol{\Theta}_0 = (\boldsymbol{\theta}_{i0}^\top)_{i=1,\ldots,M}^\top$ and $\boldsymbol{U} = (\boldsymbol{u}_i^\top)_{i=1,\ldots,M}^\top$ are Mq-dimensional vectors of coefficients for the heterogeneous features and random effects. Recall that $\boldsymbol{\beta}_0$ is a global coefficient vector shared by all data units. Ignoring the mutual learnability assumption, the above model is a linear mixed-effects model, which can be estimated by the generalized least square method (GLS Rothenberg, 1984; Anh & Chelliah, 1999), which amounts to minimize

$$\begin{aligned} &L_N^{\mathrm{GLS}}(\boldsymbol{\beta},\boldsymbol{\Theta}) \\ &\triangleq \frac{1}{2N}\sum_{i=1}^M (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{\theta}_i)^\top \boldsymbol{W}_i (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{\theta}_i) \end{aligned}$$

where $\boldsymbol{W}_i \triangleq \mathrm{Cov}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{z}_i)^{-1} = (\sigma_\varepsilon^2\boldsymbol{I}_{n_i} + \boldsymbol{z}_i\boldsymbol{\Psi}\boldsymbol{z}_i^\top)^{-1}$. For simplicity, we assume that $\boldsymbol{\Psi}$ and σ_ε^2 are known. Otherwise, these variance components can be consistently estimated through the restricted maximum likelihood (REML) method (Richardson & Welsh, 1994; Jiang, 1996). However, the standard GLS approach does not take into account the different learnability and does not conduct learnability recovery. To fuse together $\boldsymbol{\theta}_i$'s, we may add a pairwise concave fusion penalty to the objective function:

$$Q_N(\boldsymbol{\beta},\boldsymbol{\Theta}) \triangleq L_N^{\mathrm{GLS}}(\boldsymbol{\beta},\boldsymbol{\Theta}) + \sum_{1\le i<j\le M} p_\gamma(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|;\lambda), \tag{5}$$

where $p_\gamma(t;\lambda)$ is a concave penalty function with a tuning parameter $\lambda > 0$ and a concavity parameter $\gamma > 0$. A natural estimate is thus defined as

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}(\lambda)\\ \widehat{\boldsymbol{\Theta}}(\lambda)\end{pmatrix} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\Theta}\in\mathbb{R}^{Mq}}{\arg\min}\ Q_N(\boldsymbol{\beta},\boldsymbol{\Theta}). \tag{6}$$

The above full-data estimation is referred as the individual participant data (IPD) method, the "gold standard" in the meta-analysis context (e.g. Debray et al., 2015). However, for large-scale data, the IPD method requires very expensive computation; see Section 3.2 for the comparisons. This motivates the proposed CD fusion approach below.

In this article, we consider four popular concave fusion penalty functions: L_1 or the Lasso penalty (Tibshirani,

1996), minimax concave penalty (MCP, Zhang (2010)), smoothly clipped absolute deviation penalty (SCAD, Fan & Li (2001)) and truncated Lasso penalty (TLP, Shen et al. (2012)). The L_1 penalty is known to produce biased estimates (Zhao & Yu, 2006). In our simulations, it tends to produce either many subgroups or no subgroup. Instead, MCP, SCAD and TLP are more appropriate for recovering the learnability structure since they enjoy selection consistency. See Section S.1 for definitions of these penalties.

2.2. MTL Estimator using Confidence Distribution Fusion

We propose a CD fusion approach which produces the MTL estimator $\left(\widehat{\boldsymbol{\beta}}(\lambda)^\top, \widehat{\boldsymbol{\Theta}}(\lambda)^\top\right)^\top$ with much less computational costs than the IPD method, while obtaining the same result (see Theorem 2.1.) We start with obtaining individual unit estimates, and then merge them through their CD densities as defined in Liu et al. (2015). Given data in each unit, the first step is to obtain individual unit estimates for $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}_i^\top)^\top$ via the GLS method, for each i, as

$$\begin{pmatrix} \breve{\boldsymbol{\beta}}_i \\ \breve{\boldsymbol{\theta}}_i \end{pmatrix} = \left[(\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i (\boldsymbol{x}_i, \boldsymbol{z}_i)\right]^{-1} (\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i \boldsymbol{y}_i. \quad (7)$$

Recall that $\boldsymbol{\Psi}$ and σ_ε^2 are assumed known. Let $\boldsymbol{V}_i$ denote the squared root matrix of $\boldsymbol{W}_i$ such that $\boldsymbol{W}_i = \boldsymbol{V}_i^2$, and then (7) is equivalent to the ordinary least square solution of the unit data $\boldsymbol{V}_i\boldsymbol{y}_i = \boldsymbol{V}_i\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{V}_i\boldsymbol{z}_i\boldsymbol{\theta}_i + \boldsymbol{V}_i\boldsymbol{\varepsilon}_i$ such that $\mathsf{Var}(\boldsymbol{V}_i\boldsymbol{y}_i) = \boldsymbol{I}_{n_i}$. According to He & Shao (2000), as n_i grows, we have, conditional on $(\boldsymbol{x}_i, \boldsymbol{z}_i)$,

$$\left[(\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i (\boldsymbol{x}_i, \boldsymbol{z}_i)\right]^{\frac{1}{2}} \left[\begin{pmatrix} \breve{\boldsymbol{\beta}}_i \\ \breve{\boldsymbol{\theta}}_i \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\theta}_{i0} \end{pmatrix}\right] \overset{D}{\Longrightarrow} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

Following Liu et al. (2015), we define the combined CD density as $h(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \prod_{i=1}^M h_i(\boldsymbol{\beta}, \boldsymbol{\theta}_i)$, where $h_i(\boldsymbol{\beta}, \boldsymbol{\theta}_i)$ is the CD density for the ith unit. We then define an objective function which consists of $-\log h(\boldsymbol{\beta}, \boldsymbol{\Theta})$ (with additive constant terms omitted) and a pairwise concave fusion penalty, i.e., $Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta})$

$$\triangleq \frac{1}{2N} \sum_{i=1}^M \begin{pmatrix} \breve{\boldsymbol{\beta}}_i - \boldsymbol{\beta} \\ \breve{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \end{pmatrix}^\top (\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i (\boldsymbol{x}_i, \boldsymbol{z}_i) \begin{pmatrix} \breve{\boldsymbol{\beta}}_i - \boldsymbol{\beta} \\ \breve{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \end{pmatrix} + \sum_{1 \le i < j \le M} p_\gamma\left(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|; \lambda\right). \quad (8)$$

The following theorem validates that the CD-based objective function $Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ produces exactly the same minimizer, $\left(\widehat{\boldsymbol{\beta}}(\lambda)^\top, \widehat{\boldsymbol{\Theta}}(\lambda)^\top\right)^\top$, as $Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta})$ does.

Theorem 2.1. *$Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ differs from $Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta})$ only by a constant.*

Remarkably, our CD fusion approach does not require the statistical model in each data unit to be the same. This allows our method to be generalized to other learning tasks. For example, consider a generalized linear model for the i-th unit

$$\phi_i\left(\mathsf{E}\left[\mathsf{Y}|\mathbf{X}, \mathbf{Z}\right]\right) = \mathbf{X}^\top \boldsymbol{\beta}_0 + \mathbf{Z}^\top \boldsymbol{\theta}_{i0} + \mathbf{Z}^\top \boldsymbol{u}_i,$$

where the link function $\phi_i(\cdot)$ may differ between units (e.g., identity link for linear model with continuous response, logit link for logistic model with binary response, log link for Poisson regression with count data, etc.) Given the individual unit estimates $(\breve{\boldsymbol{\beta}}_i, \breve{\boldsymbol{\theta}}_i)$'s and their limit distributions, a modified CD fusion MTL approach can still be applied with Theorem 2.1 holds in an asymptotic way.

3. Computation

We discuss the derived ADMM algorithm, analyze its convergence property, and compare our MTL method with the IPD method in terms of various computational considerations.

3.1. An ADMM Algorithm

We note that the fusion penalty function cannot be written as the sum of individual functions for $\boldsymbol{\theta}_i$. Hence, we introduce a new set of parameter $\boldsymbol{\delta}_{ij} = \boldsymbol{\theta}_i - \boldsymbol{\theta}_j$ and rewrite $Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ as the following constrained optimization problem:

$$\begin{aligned} & L_0(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\delta}) \\ \triangleq & \frac{1}{2N} \sum_{i=1}^M \begin{pmatrix} \breve{\boldsymbol{\beta}}_i - \boldsymbol{\beta} \\ \breve{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \end{pmatrix}^\top (\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i (\boldsymbol{x}_i, \boldsymbol{z}_i) \begin{pmatrix} \breve{\boldsymbol{\beta}}_i - \boldsymbol{\beta} \\ \breve{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \end{pmatrix} \\ & + \sum_{1 \le i < j \le M} p_\gamma\left(\|\boldsymbol{\delta}_{ij}\|; \lambda\right), \\ & \text{subject to } \boldsymbol{\theta}_i - \boldsymbol{\theta}_j - \boldsymbol{\delta}_{ij} = \boldsymbol{0}, \quad 1 \le i < j \le M, \end{aligned} \quad (9)$$

where $\boldsymbol{\delta} = (\boldsymbol{\delta}_{ij}^\top)_{i<j}^\top$. This reformulation allows the adoption of the ADMM algorithm. See the details of the derivation in the supplementary material.

For regularization parameters (γ, λ), the best pair could be searched over a two-dimensional grid using some criteria, such as Bayesian Information Criterion (BIC, Schwarz (1978)) and generalized cross-validation (GCV Craven & Wahba, 1978). To avoid this computationally expensive search, we simply fix $\gamma = 3.7$ for MCP and SCAD as suggested in Fan & Li (2001), and choose $\gamma = 1.85$ for TLP to mimic the degree of penalization in MCP under $\gamma = 3.7$. These choices of the penalty parameters have been shown to be practically useful in many statistical applications. For λ, we employ the modified BIC (Wang et al., 2009) that is defined in high-dimensional data setting:

$$\mathrm{BIC} \triangleq \text{log-likelihood} + \frac{C_N \log N}{N}\left(p + \widehat{S}(\lambda) q\right), \quad (10)$$

where C_N is a positive number which can depend on N, and $\widehat{S}(\lambda)$ is the estimated number of subgroups using the CD fusion approach. Following Wang et al. (2009), we use $C_N = \log\log\left(p + \widehat{S}(\lambda)q\right)$ for implementation.

In the following proposition, we study the convergence property of the ADMM algorithm by examining whether the primal residual vanishes. We choose to stop the updates when $\|\boldsymbol{r}^{[k+1]}\|^2 < 10^{-6}$ in our study.

Proposition 3.1 (Convergence). *The primal residual $\boldsymbol{r}^{[k]} = \boldsymbol{B}^\top\boldsymbol{\Theta}^{[k]} - \boldsymbol{\delta}^{[k]}$ of the proposed ADMM, where $\boldsymbol{B}$ is a matrix such that $\boldsymbol{B}^\top\boldsymbol{\Theta} = (\boldsymbol{\theta}_i^\top - \boldsymbol{\theta}_j^\top)_{i<j}^\top$, satisfies that $\lim_{k\to\infty}\left\|\boldsymbol{r}^{[k]}\right\|^2 = 0$ for MCP, SCAD and TLP.*

3.2. Computational Considerations

Communication Cost

We compare the communication costs of the proposed MTL approach and the IPD method. For the former, each unit only needs to pass a $(p+q)$-dimensional vector $(\breve{\boldsymbol{\beta}}_i^\top, \breve{\boldsymbol{\theta}}_i^\top)^\top$ and a $(p+q)\times(p+q)$ precision matrix $(\boldsymbol{x}_i, \boldsymbol{z}_i)^\top\boldsymbol{W}_i(\boldsymbol{x}_i, \boldsymbol{z}_i)$ to a central computer node, leading to a communication cost $O\left[(p+q)^2\right]$. On the other hand, the latter requires each unit to pass a $n_i \times (p+q+1)$ data matrix $(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$, resulting in a communication cost $O\left[n_i(p+q)\right]$. Since the unit GLS estimates require that $(\boldsymbol{x}_i, \boldsymbol{z}_i)^\top\boldsymbol{W}_i(\boldsymbol{x}_i, \boldsymbol{z}_i)$ are invertible, we must have $p+q < n_i$ for all $i = 1, \ldots, M$. Accordingly, the communication cost has been much reduced in the proposed MTL approach.

Computation Time and Peak Memory

For computation time and peak memory used during the implementations of both approaches, we provide empirical evaluation using the simulation cases considered in Section 5. We only present the results obtained on the first generated dataset of each simulation cases; see Section 5.1 for details about the simulation settings. L_1 penalty is excluded due to its poor performance to be shown in Section 5.

Figure 3 displays the computation time (upper panel) and the maximum memory used (lower panel). It can be seen that the CD-based MTL approach outperforms the IPD method in terms of computation speed and memory usage in all cases, especially when M is large (e.g., Cases 5–9). Specifically, compared to IPD, MTL generally takes about $3/4$ to $4/5$ of the computation time, and requires roughly just a half of the memory usage. In addition, the three types of penalties result in similar performance. We note that the computational performance under different S values is not comparable because different ranges of λ were used. In summary, the proposed approach provides significant savings on computational resources. Next, we will show it also offers the best possible statistical inferential performance.

4. Theoretical Guarantees

We show that the proposed estimator has the same limiting distribution as the *oracle* estimator, which transfers the knowledge as if the true learnability structure were known.

4.1. Oracle Estimator

Consider an $Mq \times Sq$ label matrix $\boldsymbol{A}$ such that $\boldsymbol{\Theta}_0 = \boldsymbol{A}\boldsymbol{\alpha}_0$, where $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{s0}^\top)_{s=1,\ldots,S}^\top$. The oracle model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{Z}\boldsymbol{A}\boldsymbol{\alpha}_0 + \boldsymbol{Z}\boldsymbol{U} + \boldsymbol{\mathcal{E}}$ is defined assuming $\boldsymbol{A}$ were known. The resulting oracle estimator for $(\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$ is

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}\\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}\end{pmatrix} = \left[(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})^\top\boldsymbol{W}(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})\right]^{-1}(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})^\top\boldsymbol{W}\boldsymbol{Y},$$

where $\boldsymbol{W} = \mathrm{diag}\left[(\boldsymbol{W}_i)_{i=1,\ldots,M}\right]$ is a $N\times N$ block diagonal matrix. Below we present a concrete example for $\boldsymbol{A}$.

Example 4.1. Suppose $M = 5, S = 2, \boldsymbol{\theta}_{1,0} = \boldsymbol{\theta}_{2,0} = \boldsymbol{\alpha}_{1,0}$ and $\boldsymbol{\theta}_{3,0} = \boldsymbol{\theta}_{4,0} = \boldsymbol{\theta}_{5,0} = \boldsymbol{\alpha}_{2,0}$. Then we have

$$\begin{pmatrix}\boldsymbol{\theta}_{1,0}\\ \boldsymbol{\theta}_{2,0}\\ \boldsymbol{\theta}_{3,0}\\ \boldsymbol{\theta}_{4,0}\\ \boldsymbol{\theta}_{5,0}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\alpha}_{1,0}\\ \boldsymbol{\alpha}_{1,0}\\ \boldsymbol{\alpha}_{2,0}\\ \boldsymbol{\alpha}_{2,0}\\ \boldsymbol{\alpha}_{2,0}\end{pmatrix} = \underbrace{\begin{bmatrix}\boldsymbol{I}_q & \\ \boldsymbol{I}_q & \\ & \boldsymbol{I}_q\\ & \boldsymbol{I}_q\\ & \boldsymbol{I}_q\end{bmatrix}_{5q\times 2q}}_{\boldsymbol{A}} \underbrace{\begin{pmatrix}\boldsymbol{\alpha}_{1,0}\\ \boldsymbol{\alpha}_{2,0}\end{pmatrix}_{2q\times 1}}_{\boldsymbol{\alpha}_0}.$$

The following regularity conditions on the design matrices, $\boldsymbol{U}$ and noise $\boldsymbol{\mathcal{E}}$ are needed.

Assumption 4.1 (Design Matrix). (i) The rows of the design matrices $\boldsymbol{f}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)$'s have sub-Gaussian tails in the sense that, for any $\boldsymbol{a} \in \mathbb{R}^{p+q}$, $P(|\boldsymbol{a}^\top[\boldsymbol{f}_i]_{k\cdot}| > \|\boldsymbol{a}\|t) \le 2\exp(-c_f t^2)$ for all $i = 1, \ldots, M, k = 1, \ldots, n_i$, and some absolute constant $c_f > 0$.

(ii) Let $\mathsf{F} = (\mathsf{X}^\top, \mathsf{Z}^\top)^\top$. There exists some absolute constant $0 < C_f < 1$ such that

$$C_f \le \lambda_{\min}\left(\mathsf{E}\left[\mathsf{F}\mathsf{F}^\top\right]\right) \le \lambda_{\max}\left(\mathsf{E}\left[\mathsf{F}\mathsf{F}^\top\right]\right) \le C_f^{-1}.$$

Figure 3. Computation time (upper panel) and peak memory usage (lower panel) for the CD-based MTL and the IPD methods.

Assumption 4.2 ($\boldsymbol{U}$ and $\boldsymbol{\mathcal{E}}$)**.** The coefficient $\boldsymbol{U}$ and the noise vectors $\boldsymbol{\mathcal{E}}$ have sub-Gaussian tails in the sense that $P(|\boldsymbol{a}^\top \boldsymbol{U}| > \|\boldsymbol{a}\|t) \le 2\exp(-c_e t^2)$ and $P(|\boldsymbol{b}^\top \boldsymbol{\mathcal{E}}| > \|\boldsymbol{b}\|t) \le 2\exp(-c_e t^2)$ for any vectors $\boldsymbol{a} \in \mathbb{R}^{Mq}, \boldsymbol{b} \in \mathbb{R}^N$ and $t, c_e > 0$.

Let $n_{\min} \triangleq \min_{1\le i\le M} n_i$, $g_{\min} \triangleq \min_{1\le s\le S} \sum_{i:L_i=s} n_i$, and $g_{\max} \triangleq \max_{1\le s\le S} \sum_{i:L_i=s} n_i$ be the minimum unit size, and the minimum and maximum total sample size in a subgroup. Theorem 4.1 gives a non-asymptotic upper bound and the limiting distribution for the oracle estimator.

Theorem 4.1 (Properties of the Oracle Estimator)**.** *Given Assumptions 4.1 and 4.2, suppose* $g_{\min} \gg (p+Sq)^{1/2}N^{3/4}$, $M = o(\exp(n_{\min}))$ *and* $p + Sq = o\big(\exp(g_{\max})\big)$.

(i). With probability at least $1-(p+Sq)(4N^{-1}+e^{-g_{\max}}) - 2Me^{-n_{\min}}$*, we have,*

$$\left\| \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{OR}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}} - \boldsymbol{\alpha}_0 \end{pmatrix} \right\| \le \phi_N, \tag{11}$$

where $\phi_N = \sqrt{5}c_e^{-1/2}c_f^{1/2}C_f^{-1}\left(\tau \wedge \sigma_\varepsilon^2\right)^{-1/2}\sigma_\varepsilon g_{\min}^{-1} \times \sqrt{(p+Sq)N\log N}$*, and* $\tau = \lambda_{\min}(\boldsymbol{\Psi})$.

(ii). For any sequence of $(p+Sq)$*-vectors* $\{\boldsymbol{a}_N\}$ *with* $\|\boldsymbol{a}_N\| = 1$*, we have as* $N \to \infty$*,*

$$\sigma_N^{-1}(\boldsymbol{a}_N)\boldsymbol{a}_N^\top \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{OR}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}} - \boldsymbol{\alpha}_0 \end{pmatrix} \xRightarrow{D} \mathcal{N}(0,1), \tag{12}$$

with $\sigma_N(\boldsymbol{a}_N) \triangleq \sqrt{\boldsymbol{a}_N^\top \left[(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})^\top \boldsymbol{W}(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})\right]^{-1}\boldsymbol{a}_N}$.

It is important to note that $g_{\min} \to \infty$. Since $g_{\min} \le N/S$, the condition $g_{\min} \gg N^{3/4}(p+Sq)^{1/2}$ implies that $S\sqrt{p+Sq} = o(N^{1/4})$. This imposes an upper bound on how fast (S, p, q) can grow with N. In particular, if p and q are assumed to be fixed, we have $S = o(N^{1/6})$.

To appreciate the upper bound (11), we make the following assumptions: $n_i \equiv n$ (thus $N = Mn$), $g_{\min} \asymp g_{\max} \asymp N/S$ and fixed (p, q). In this case, the upper bound in (11) admits the rate $S^{3/2}\sqrt{(\log M + \log n)/(Mn)}$. It becomes $o_P(1)$ if $S^3 = o\big[Mn/(\log M + \log n)\big]$, which suggests S cannot grow faster than $M^{1/3}$.

4.2. Theoretical Properties of the MTL Estimator

We prove that the proposed MTL estimator is asymptotically equivalent to the oracle estimator, and hence achieves the highest level of statistical inferential accuracy. More importantly, we also show that through MTL, the proposed estimator is asymptotically more efficient than the IPD estimated in (7) that is based on the full data. These results require two additional conditions: one is on the minimal signal defined as $\Delta_N \triangleq \min_{s\ne s'} \|\boldsymbol{\alpha}_s - \boldsymbol{\alpha}_{s'}\|$; the other is on the penalty function $p_\gamma(t;\lambda)$.

Assumption 4.3 (Penalty Function)**.** $p_\gamma(t;\lambda)$ is symmetric with $p_\gamma(0;\lambda) = 0$, and is non-decreasing and concave for $t \in [0,\infty)$. There exists a constant $a > 0$ such that $p_\gamma(t;\lambda)$ is a constant for all $t \ge a\lambda$. $\frac{\partial}{\partial t}p_\gamma(0;\lambda)$ exists and is continuous except for a finite number of t and $\frac{\partial}{\partial t}p_\gamma(0+;\lambda) = \lambda$.

This assumption is satisfied by such non-convex penalties as MCP, SCAD and TLP with $a = \gamma$ and is considered in Ma & Huang (2017; 2016) as well. Along with proper minimal signal condition, it ensures that the penalty term does not push $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ from different subgroups toward each other.

Theorem 4.2 (Oracle Property)**.** *Suppose the conditions in Theorem 4.1 and Assumption 4.3 hold and* $S \ge 2$*. If* $\phi_N \ll \lambda \ll a^{-1}\Delta_N$*, where* a *is defined in Assumption 4.3 and* ϕ_N *is given in Theorem 4.1, then there exists a local minimizer* $\big(\widehat{\boldsymbol{\beta}}(\lambda)^\top, \widehat{\boldsymbol{\Theta}}^\top(\lambda)\big)^\top$ *of (8) satisfying*

$$P\left(\begin{pmatrix} \widehat{\boldsymbol{\beta}}(\lambda) \\ \widehat{\boldsymbol{\Theta}}(\lambda) \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{OR}} \\ \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}} \end{pmatrix}\right) \to 1.$$

Let $\widehat{\boldsymbol{\alpha}}(\lambda)$ and $\widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}$ be the distinct values of $\widehat{\boldsymbol{\Theta}}(\lambda)$ and $\widehat{\boldsymbol{\Theta}}_{\mathrm{OR}}$, respectively. Theorem 4.2 implies that the proposed estimator has the same limiting distribution as the oracle estimator given in (12). Moreover, an interesting efficiency boosting phenomenon is discovered in (ii) of Corollary 4.1 below

Corollary 4.1. *Suppose the conditions in Theorem 4.2 hold.*

(i) (Asymptotic Normality) For any sequence of $(p+Sq)$*-vectors* $\{\boldsymbol{a}_N\}$ *with* $\|\boldsymbol{a}_N\| = 1$*, we have as* $N \to \infty$*,*

$$\sigma_N^{-1}(\boldsymbol{a}_N)\boldsymbol{a}_N^\top \begin{pmatrix} \widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}_0 \end{pmatrix} \xRightarrow{D} \mathcal{N}(0,1),$$

where $\sigma_N(\boldsymbol{a}_N)$ *is given in Theorem 4.1.*

(ii) (Efficiency Boosting) For any p*-vector* $\boldsymbol{v}_p$ *and* q*-vector* $\boldsymbol{v}_q$*, we have for all* $i = 1, \ldots, M$*,*

$$\boldsymbol{v}_p^\top \mathsf{Cov}\big(\widehat{\boldsymbol{\beta}}(\lambda)\big)\boldsymbol{v}_p \le \boldsymbol{v}_p^\top \mathsf{Cov}\big(\breve{\boldsymbol{\beta}}_i\big)\boldsymbol{v}_p \quad \textit{and}$$

$$\boldsymbol{v}_q^\top \mathsf{Cov}\big(\widehat{\boldsymbol{\theta}}_i(\lambda)\big)\boldsymbol{v}_q \le \boldsymbol{v}_q^\top \mathsf{Cov}\big(\breve{\boldsymbol{\theta}}_i\big)\boldsymbol{v}_q.$$

To estimate the limit covariance matrix $\sigma_N(\boldsymbol{a}_N)$, we need to replace the unknown label matrix $\boldsymbol{A}$ (under which $\boldsymbol{\Theta}_0 = \boldsymbol{A}\boldsymbol{\alpha}_0$) by its estimate $\widehat{\boldsymbol{A}}$. This can be trivially done after $\widehat{\boldsymbol{\alpha}}(\lambda)$ and $\widehat{\boldsymbol{\Theta}}(\lambda)$ are computed. Another challenge is that the direct computation of $(\boldsymbol{X}, \boldsymbol{Z}\widehat{\boldsymbol{A}})^\top \boldsymbol{W}(\boldsymbol{X}, \boldsymbol{Z}\widehat{\boldsymbol{A}})$, which requires $O\big[N^2(p + \widehat{S}(\lambda)q)\big]$ operations, is infeasible when N is large. A more efficient way is to take the multivariate GLS approach of Becker & Wu (2007) as follows. Let $\widehat{\boldsymbol{G}}_s, s = 1, \ldots, \widehat{S}(\lambda)$ be $(p+q) \times (p + \widehat{S}(\lambda)q)$ matrices based on the estimated label matrix $\widehat{\boldsymbol{A}}$ such that

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}(\lambda) \\ \widehat{\boldsymbol{\theta}}_i(\lambda) \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}(\lambda) \\ \widehat{\boldsymbol{\alpha}}_{\widehat{L}_i}(\lambda) \end{pmatrix} = \widehat{\boldsymbol{G}}_{L_i} \begin{pmatrix} \widehat{\boldsymbol{\beta}}(\lambda) \\ \widehat{\boldsymbol{\alpha}}(\lambda) \end{pmatrix}, \ i = 1, \ldots, M.$$

After straightforward algebra, we have

$$(\boldsymbol{X}, \boldsymbol{Z}\widehat{\boldsymbol{A}})^\top \boldsymbol{W}(\boldsymbol{X}, \boldsymbol{Z}\widehat{\boldsymbol{A}}) = \sum_{i=1}^{M} \widehat{\boldsymbol{G}}_{\widehat{L}_i}^\top \boldsymbol{W}_i(\boldsymbol{x}_i, \boldsymbol{z}_i)\widehat{\boldsymbol{G}}_{\widehat{L}_i}.$$

Note that $(\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i(\boldsymbol{x}_i, \boldsymbol{z}_i)$'s are $(p+q)\times(p+q)$ precision matrices given by data units under the MTL approach. Hence, the right-hand-side in the equation above only requires $O\big[M(p+\widehat{S}(\lambda)q)^2(p+q)\big]$ operations.

5. Simulation Experiments

We now evaluate the MTL method using simulations.

5.1. Simulation Setup

Table 1. Simulation Settings.

Simulation case	n	M	S	SNR
Case 1	1024	50	2	4.399
Case 2	1024	50	3	8.838
Case 3	2048	50	2	4.399
Case 4	2048	50	3	8.838
Case 5	2048	100	2	4.399
Case 6	2048	100	3	8.838
Case 7	2048	100	5	8.734
Case 8	2048	150	5	8.734
Case 9	2048	150	7	9.260

Table 1 summarizes nine simulation settings (with $p = 5$ global features and $q = 3$ heterogeneous features), and each has 100 replications, where the signal-to-noise ratio (SNR) is defined in Section S.10. The largest total sample size is 307,200. For simplicity, we consider equal unit sizes $n_i \equiv n$ for $i = 1, \ldots, M$. We let the number of units in each subgroup to be $(M_1, \ldots, M_S) = \mathbf{1}_S + \textit{Multinomial}(M - S, \mathbf{1}_S/S)$. The coordinates of $\boldsymbol{\beta}_0$ were generated from $\textit{Uniform}(-2, 2)$ independently. To mimic the different coefficient values for the heterogeneous features between subgroups, we generated $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{1,0}^\top, \ldots, \boldsymbol{\alpha}_{S,0}^\top)^\top$, where $\boldsymbol{\alpha}_{s,0} = (\alpha_{s,0,1}, \alpha_{s,0,2}, \alpha_{s,0,3})^\top$, in a way to guarantee the minimal signal condition. The values of $(\alpha_{1,0,1}, \ldots, \alpha_{S,0,1})$ were assigned to evenly spaced grid points over $[-S^{1.4}/2, S^{1.4}/2]$, denoted as $(\alpha_1^*, \ldots, \alpha_S^*)$, and then $\boldsymbol{\alpha}_0$ was generated by

$$\mathrm{vec}\begin{bmatrix} \alpha_1^* & \alpha_2^* & \alpha_3^* & \ldots & \alpha_{S-1}^* & \alpha_S^* \\ \alpha_S^* & \alpha_1^* & \alpha_2^* & \ldots & \alpha_{S-2}^* & \alpha_{S-1}^* \\ \alpha_{S-1}^* & \alpha_S^* & \alpha_1^* & \ldots & \alpha_{S-3}^* & \alpha_{S-2}^* \end{bmatrix},$$

where $\mathrm{vec}(\cdot)$ denotes the vectorization operator. If S is odd, we further add 1 to all coordinates of $\boldsymbol{\alpha}_0$ to avoid subgroup effect coordinates being 0. The data matrices $\boldsymbol{f}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)$ consist of independent rows with each generated from 8-dimensional normal $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{F}})$, where the diagonal elements of $\boldsymbol{\Sigma}_{\mathbf{F}}$ are 1 and off-diagonal elements are 0.3.

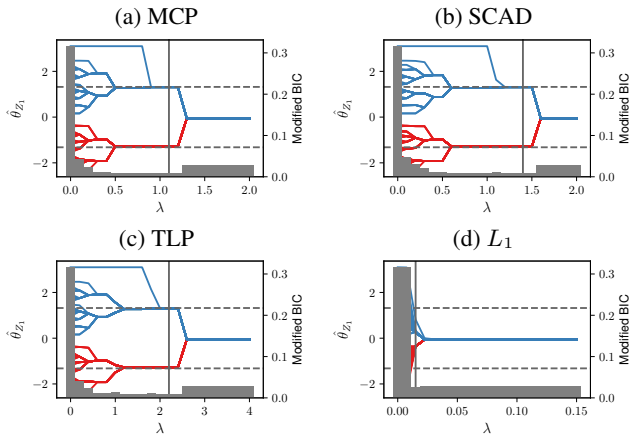

Figure 4. Solution paths of $(\theta_{1,1}, \ldots, \theta_{M,1})$ for the first dataset in Case 5 ($n = 256, M = 100, S = 2$). Colors stand for true subgroups, horizontal dashed lines represent true values of the first coordinate of $\boldsymbol{\alpha}_s$'s, and vertical line displays best λ selected by modified BIC, whose values are displayed by gray bars.

Moreover, $\boldsymbol{u}_i$ and $\boldsymbol{\varepsilon}_i$ follows $\mathcal{N}(\mathbf{0}, 0.3\boldsymbol{I})$ and $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, respectively. Finally, $\boldsymbol{Y}$ was generated from the oracle model.

Figure 4 shows the solution paths of $(\theta_{1,1}, \ldots, \theta_{M,1})$, the first coordinate of $\boldsymbol{\theta}_i$'s. It can be seen that, under MCP, SCAD and TLP, the learnability structure can be well recovered using the modified BIC given in (10) to choose the tuning parameter λ, denoted as $\widehat{\lambda}$. On the other hand, although the L_1 penalty gets the correct number of subgroups, the coefficient estimates are far away from the truth.

5.2. Learnability Structure Recovery

We use the Normalized Mutual Information (NMI; Fred & Jain, 2003) as the performance measure for learnability structure recovery. NMI$\in [0, 1]$ and larger values imply more similar groupings. Let $\mathbb{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots\}$ and $\mathbb{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots\}$ denote two partitions of $\{1, \ldots, M\}$. NMI is defined as $\mathrm{NMI}(\mathbb{C}, \mathbb{D}) \triangleq \dfrac{2I(\mathbb{C}; \mathbb{D})}{H(\mathbb{C}) + H(\mathbb{D})}$, where $I(\mathbb{C}; \mathbb{D}) \triangleq \sum_{k,l} \big(|\mathcal{C}_k \cap \mathcal{D}_l|/M\big) \log(M|\mathcal{C}_k \cap \mathcal{D}_l|/|\mathcal{C}_k||\mathcal{D}_l|)$ is the mutual information between $\mathbb{C}$ and $\mathbb{D}$, and $H(\mathbb{C}) \triangleq -\sum_k (|\mathcal{C}_k|/M) \log(|\mathcal{C}_k|/M)$ is the entropy of $\mathbb{C}$. The percentage of perfect recoveries among the 100 replications is also reported.

We also consider an *ad hoc* approach based on performing K-means on the unit estimates $\widehat{\boldsymbol{\theta}}_i$ with modified BIC to select the optimal subgroup size, to compare with our method. Specifically, for any given number of subgroup size $S = K$, the K-means algorithm is performed to obtain an estimated subgroup labels, and then the oracle estimation can be computed using the estimated label and variance components. Subsequently, the modified BIC values are calculated by plugging in these K-means-based estimates. In our analysis, we run through $K = 1, \ldots, 10$ and choose

Table 2. Learnability structure recovery. Short version of Table S.1.

	Method	Mean, Median (Min,Max) of $\widehat{S}$	NMI	Perfect Recover
Case 1	MCP	2, 2 (2, 2)	0.998	0.99
$S = 2$	SCAD	2, 2 (2, 2)	0.998	0.99
	TLP	2, 2 (2, 2)	0.998	0.99
	L_1	2, 2 (1, 4)	0.947	0.62
	K-Mns	2, 2 (2, 3)	0.989	0.95
Case 7	MCP	5, 5 (5, 6)	0.999	0.98
$S = 5$	SCAD	5, 5 (5, 5)	0.999	0.99
	TLP	5, 5 (5, 6)	0.999	0.99
	L_1	63.2, 100 (1, 100)	0.587	0.00
	K-Mns	5, 5 (5, 5)	0.976	0.37
Case 9	MCP	7, 7 (7, 7)	0.999	0.99
$S = 7$	SCAD	7.3, 7 (7, 9)	0.998	0.77
	TLP	7.2, 7 (7, 9)	0.998	0.81
	L_1	150, 150 (150, 150)	0.620	0.00
	K-Mns	7, 7 (7, 7)	0.978	0.20

the K value such that the modified BIC is minimized.

Table 2 includes results for select cases (Cases 2, 5, 7; see the full table in the supplementary material). It can be seen that MCP, SCAD and TLP result in desirable recovery performance, while L_1 penalty does not. MCP works very well in all cases. Although the high NMI values suggest that SCAD and TLP still capture the true subgroup structure well, their performance gets worse when S is large (e.g., Case 7) in terms of the standard deviation of $\widehat{S}$ and perfect recovery rate. This echoes with the discussion following Theorem 4.1 that the theoretical properties hold only when S does not grow too fast. Due to the poor performance, L_1 penalty is no longer considered in the subsequent analysis. K-means performs well for small $M = 50$, but poorly for larger $M = 100$ and 150, even though the number of subgroups are correctly obtained (e.g., Case 7).

5.3. Parameter Estimation

To evaluate the parameter estimation, we define the root mean squared error (RMSE) between two vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^d$ as $\mathrm{RMSE}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{1}{\sqrt{d}}\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|$. In Figure 5, we empirically examine how close the proposed estimate is to both true value and oracle estimate. Memory outage occurred in some cases when computing oracle estimates. To resolve this issue, the multivariate GLS approach of Becker & Wu (2007) is applied to compute the oracle estimates.

Firstly, from the top panel in Figure 5, we can easily see $\mathrm{RMSE}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)$ reasonably increases (decreases) with S (M). This confirms the intuition that more S suggests more challenges in mutual transfer. On the other hand, increasing M allows to estimate the heterogeneous features using more data units, which improves their estimation accuracy. The third panel in Figure 5 indicates that $\mathrm{RMSE}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$ depends on $N = Mn$ but is independent of S. Note that $\boldsymbol{\beta}$ is

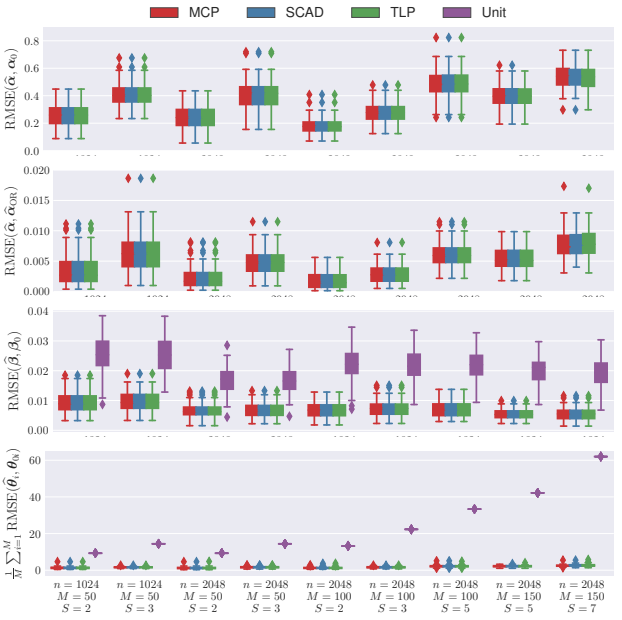

Figure 5. Evaluation of parameter estimation.

the global parameter shared over all data units. This result numerically verifies that the proposed method effectively merges the common information provided from data units, regardless of how many different subgroups they come from.

The results in Theorem 4.2 and (ii) in Corollary 4.1 can be verified as well. Firstly, the smaller values of $\mathrm{RMSE}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}})$ (the second panel) and $\mathrm{RMSE}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}_{\mathrm{OR}})$ (which are all close to 0 and omitted here) imply that $(\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\alpha}}^\top)^\top$ is fairly close to their oracle counterpart $(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}^\top, \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}^\top)^\top$. The third panel in Figure 5 indicates that the estimation accuracy of $\widehat{\boldsymbol{\beta}}$ is better than the best unit estimate for common effect. Similar observation applies to the estimation of $\boldsymbol{\theta}_i$'s as demonstrated in the bottom panel in Figure 5. Both panels support the efficiency boosting result shown in Corollary 4.1. We also studied the asymptotic covariance approximation and the inferential accuracy. See Section S.11.

6. Real Data Example

NOAA's nClimDiv database[1] were analyzed to demonstrate MTL method's practical usefulness. The monthly average temperature is the response of interest. We categorize the 8 features to global ones and heterogeneous ones by inspecting the kernel densities. Intuitively, the distribution of the unit-level coefficient estimate for a heterogeneous feature is more likely to be multimodal or widespread (as opposed to unimodal or concentrated) across units. In this way we have chosen intercept, PCPN and ZNDX as the heterogeneous features (i.e., $q = 3$ and $p = 5$) and the rest as global

[1] Available at ftp://ftp.ncdc.noaa.gov/pub/data/cirs/climdiv/.

features; see details in Section S.11.2. Figure 2 displays the recovered learnability structure, where the five resulting subgroups generally follow a geographical pattern, although we do not use any spatial information of the climate divisions. It may appear strange that northeastern parts of New England are in the same subgroup as Texas and Southern California. However, this pattern actually coincides with NOAA's outlook for temperature, precipitation, and drought: blue and red groups together follow the temperature outlook[2], and red subgroup stands out due to severe drought conditions[3].

Table 3. The parameter estimates using the proposed MTL method with MCP. The average asymptotic standard deviation (ASD), after multiplied by 100, is shown in the parenthesis.

Color [#(units)]	**Heterogeneous Features (ASD$\times$100)**		
	$\widehat{\boldsymbol{\alpha}}_{\text{Intercept}}$	$\widehat{\boldsymbol{\alpha}}_{\text{PCPN}}$	$\widehat{\boldsymbol{\alpha}}_{\text{ZNDX}}$
Red [41]	64.97 (13.2)	-0.37 (9.5)	-0.07 (9.5)
Blue [132]	49.53 (7.1)	0.85 (5.3)	-1.51 (5.3)
Green [79]	35.32 (8.9)	5.44 (6.9)	-4.05 (6.8)
Purple [81]	24.74 (9.2)	7.28 (6.8)	-5.16 (6.7)
Orange [11]	9.90 (32.3)	9.14 (19.3)	-6.54 (18.6)

Coef. for Global Features (ASD$\times$100)				
$\widehat{\beta}_{\text{Summer}}$	$\widehat{\beta}_{\text{Fall}}$	$\widehat{\beta}_{\text{Winter}}$	$\widehat{\beta}_{\text{PDSI}}$	$\widehat{\beta}_{\text{PHDI}}$
18.26 (2)	4.06 (2)	-15.12 (2)	0.18 (1)	0.20 (1)

From the parameter estimates tabulated in Table 3, we can see the general pattern that the average temperature tends to drop from southwestern areas to northeastern areas. ZNDX has the similar tendency as the intercept does, but PCPN behaves in an opposite direction. For the global feature estimates, there is no surprise to see the seasonal effects, and PDSI and PHDI have positive but relatively small (compared to the heterogeneous features) influence on the temperature.

Additional prediction-driven analysis was conducted by partitioning the nClimDiv dataset into two parts: data in 1895-2000 was used for training and 2001-2016 for testing. Our model did not perform as well as the baseline unit-level regression model in terms of the RMSE for the test data. This may be caused by a sub-optimal choice of the tuning parameter λ which leads to fewer subgroups than what would be needed for better prediction performance. Recall that we did not choose λ to minimize the cross-validation prediction error, but used BIC in order to get a parsimonious model. On the other hand, with the simulation results and optimality theory presented in the paper, our model dominates the baseline model in terms of statistical inference accuracy. This may suggest a fundamental difference between inferential analysis and predictive analysis.

[2]E.g., see the first figure in https://www.climate.gov/news-features/videos/noaas-2019-20-winter-outlook-temperature-precipitation-and-drought.

[3]See the United States Drought Monitor Animation at https://droughtmonitor.unl.edu/Maps/Animations.aspx

Acknowledgements

This work was completed while Guang Cheng was a member of Institute for Advanced Study (IAS) at Princeton University in Fall 2019. Guang Cheng would like to acknowledge the hospitality of IAS, and the financial support from the National Science Foundation (DMS-1712907, DMS-1811812, and DMS-1821183), the Office of Naval Research (ONR N00014-18-2759), and the Adobe Data Science Fund.

References

Anh, V. V. and Chelliah, T. Estimated generalized least squares for random coefficient regression models. *Scandinavian Journal of Statistics*, 26(1):31–46, 1999.

Becker, B. J. and Wu, M.-J. The synthesis of regression slopes in meta-analysis. *Statistical Science*, 22(3):414–429, 2007.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.

Craven, P. and Wahba, G. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1978.

Debray, T. P. A., Moons, K. G. M., van Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H. H., Reitsma, J. B., and on behalf of the GetReal methods review group. Get real in individual participant data (IPD) meta-analysis: A review of the methodology. *Research Synthesis Methods*, 6(4):293–309, 2015.

Eckstein, J. A practical test for univariate and multivariate normality. Technical report, RUTCOR Research Report RRR 32-2012, 12 2012.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Fred, A. L. and Jain, A. K. Robust data clustering. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 128–136, 2003.

Ghadimi, E., Teixeira, A., Shames, I., and Johansson, M. Optimal parameter selection for the alternating direction method of multipliers (admm): Quadratic problems. *IEEE Transactions on Automatic Control*, 60(3):644–658, 2015.

Harville, D. A. *Matrix Algebra From a Statistician's Perspective*. Springer, corrected edition edition, 2000. ISBN 978-0387949789.

He, X. and Shao, Q.-M. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135, 2000.

Hsu, D., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

Jiang, J. REML estimation: Asymptotic behavior and related topics. *The Annals of Statistics*, 24(1):255–286, 1996.

Liu, D., Liu, R. Y., and Xie, M. Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. *Journal of the American Statistical Association*, 110(509):326–340, 2015.

Ma, S. and Huang, J. Estimating subgroup-specific treatment effects via concave fusion. arXiv preprint, 2016.

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

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2009.

Richardson, A. and Welsh, A. Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models. *Australian Journal of Statistics*, 36(1):31–43, 1994.

Rothenberg, T. J. Approximate normality of generalized least squares estimates. *Econometrica*, 52(4):811–825, 1984.

Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Shen, X., Pan, W., and Zhu, Y. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288, 1996.

Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. C. and Kutyniok, G. (eds.), *Compressed Sensing*, chapter 5, pp. 210–268. Cambridge University Press, Cambridge, 2012.

Wang, H., Li, B., and Leng, C. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2): 894–942, 2010.

Zhao, P. and Yu, B. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *arXiv preprint arXiv:1911.02685*, 2019.