Data Min Knowl Disc
https://doi.org/10.1007/s10618-018-0564-z

# Robust finite mixture regression for heterogeneous targets

**Jian Liang**[1] · **Kun Chen**[2] · **Ming Lin**[3] · **Changshui Zhang**[1] · **Fei Wang**[4]

Received: 18 May 2017 / Accepted: 30 March 2018

**Abstract** Finite Mixture Regression (FMR) refers to the mixture modeling scheme which learns multiple regression models from the training data set. Each of them is in charge of a subset. FMR is an effective scheme for handling sample heterogeneity, where a single regression model is not enough for capturing the complexities of the conditional distribution of the observed samples given the features. In this paper, we propose an FMR model that (1) finds sample clusters and jointly models multiple incomplete mixed-type targets simultaneously, (2) achieves shared feature selection

Responsible editor: Pauli Miettinen.

✉ Jian Liang
liangjian12@mails.tsinghua.edu.cn

Kun Chen
kun.chen@uconn.edu

Ming Lin
linmin@umich.edu

Changshui Zhang
zcs@mails.tsinghua.edu.cn

Fei Wang
few2001@med.cornell.edu

[1] Department of Automation, State Key Lab of Intelligent Technologies and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing 100084, People's Republic of China

[2] Department of Statistics, University of Connecticut, Storrs, CT 06269, USA

[3] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

[4] Department of Healthcare Policy and Research, Cornell University, New York City, NY 10065, USA

Published online: 17 April 2018

among tasks and cluster components, and (3) detects anomaly tasks or clustered structure among tasks, and accommodates outlier samples. We provide non-asymptotic oracle performance bounds for our model under a high-dimensional learning framework. The proposed model is evaluated on both synthetic and real-world data sets. The results show that our model can achieve state-of-the-art performance.

**Keywords** Finite Mixture Regression · Mixed-type response · Incomplete targets · Anomaly detection · Task clustering

## 1 Introduction

Regression modeling, which refers to building models to learn conditional relationship between output targets and input features on some training samples, is a fundamental problem in statistics and machine learning. Some classical regression modeling approaches include least square regression, logistic regression and Poisson regression; see, e.g., Bishop (2006), Kubat (2015), Fahrmeir et al. (2013) and the references therein.

The aforementioned classic approaches usually train a single model of a single target over the entire data set. However, real-world problems can be much more complicated. In particular, the needs of utilizing high-dimensional features, population heterogeneity, and multiple interrelated targets are among the most prominent complications. To handle high-dimensional data, the celebrated regularized estimation approaches have undergone exciting developments in recent years; see, e.g., Fan and Lv (2010) and Huang et al. (2012). In the presence of population heterogeneity, the samples may form several distinct clusters corresponding to mixed relationships between the targets and the features. A popular modeling strategy in such a scenario is the *Finite Mixture Regression* (FMR) (McLachlan and Peel 2004), which is capable of adaptively learning multiple models, each of which is responsible for one subset/cluster of the data. FMR models have been widely used in market segmentation studies, patients' disease progression subtyping, motif gene-expression research, etc.; see, e.g., Städler et al. (2010), Khalili (2011), Khalili and Chen (2007), Doğru and Arslan (2017), and the references therein. The problem of joint learning for multiple targets is usually referred to as *Multi-Task Learning* (MTL) in machine learning or *multivariate learning* in statistics; see, e.g., Argyriou et al. (2007a), Argyriou et al. (2007b), Chen et al. (2011), and Gong et al. (2012b). We stress that the main definition of MTL considers tasks that do not necessarily share the same set of samples (and features), and that this paper focuses on a special case of MTL, where the multivariate outcomes are collected from the same set of samples and share the same set of features, whose reason will be explained later. There have also been multi-task FMR models, e.g., Wedel and DeSarbo (1995), Wang et al. (2004), Lim et al. (2016) and Bai et al. (2016), which mainly built on certain multivariate probability distribution such as Gaussian distribution or multivariate $t$ distribution.

Thus far, a comprehensive study on multi-task mixture-regression modeling with high-dimensional data is still lacking. To tackle this problem for handling real-world applications, there remain several challenges and practical concerns.

- *Task Heterogeneity* Current MTL approaches usually assume that the targets are of the same type. However, it is common that the multiple targets are of different types, such as continuous, binary, count, etc., which we refer to as task heterogeneity. For example, in anesthesia decision making (Tan et al. 2010), the anesthesia drugs will have impacts on multiple indicators of an anesthesia patient, such as anesthesia depth, blood pressures, heart rates, etc. The anesthesiologist needs to consider all those different aspects as well as their intrinsic dependence before making the decision.
- *Task Integration* As in the anesthesiology example, the multiple tasks are typically inter-related to each other, and the potential benefit from a MTL approach needs to be realized through properly exploring and taking advantage of these relationships. In existing high-dimensional MTL approaches, the tasks are usually integrated by assuming certain shared conditional mean structures between the targets and the features. The problem is more difficult in the presence of both task and population heterogeneities.
- *Task Robustness* Similar to the idea in the robust MTL approaches (Passos et al. 2012; Gong et al. 2012a; Chen et al. 2011), it is not always the case that jointly considering all tasks by assuming certain shared structures among them would be helpful. Certain tasks, referred to as anomaly tasks, may not follow the assumed shared structure and thus can ruin the overall model performance. More generally, tasks may even cluster into groups with different shared structures.

In this paper, we propose a novel method named HEterogeneous-target Robust MIxTure regression (HERMIT), to address the above challenges in a unified framework. Here we explain that we mainly consider the setting, where the multivariate outcomes are collected from the same set of samples and share the same set of features because our main objective is to learn potentially shared sample clusters and feature sets among tasks. Rigorous theoretical analysis and performance guarantees are provided. It is worthwhile to highlight the key aspects of our approach as follows.

- Our method handles mixed type of targets simultaneously. Each target follows an exponential dispersion family model (Jorgensen 1987), so that multiple different types of targets, e.g., continuous, binary, and counts, can be handled jointly. The tasks are naturally integrated through sharing the same clustering structure arising from population heterogeneity. Our theory allows HERMIT to cover sub-exponential distributions, including the commonly-encountered Bernoulli, Poisson and Gaussian as special cases.
- Our method imposes structural constraints in each mixture component of HERMIT, to deal with the curse of dimensionality and at the same time further take advantage of the interrelationship of the tasks. In particular, the group $\ell_1$ penalization is adopted to perform shared feature selection among tasks within each mixture component.
- Our method adopts three strategies for robustness. First, we adopt a mean-shift regularization technique (She and Chen 2017) to detect the outlier samples automatically and adjust for its outlying effects in model estimation. The second strategy measures discrepancy of different conditional distributions to detect anomaly tasks. The third strategy measures similarity between each pair of tasks

to discover a clustered structure among tasks. Moreover, our model can work with incomplete data and impute entry-wise missing values in the multiple targets.

The aforementioned key elements, e.g., multi-task learning, sample clustering, shared feature selection, and anomaly detection, are integrated in a unified mixture model setup, so that they can benefit from and reinforce each other. A generalized Expectation-Maximization (GEM) (Neal and Hinton 1998) algorithm is developed to conduct model estimation efficiently. For theoretical analysis, we generalize the results of Städler et al. (2010) to establish non-asymptotic oracle performance bounds for HERMIT under a high-dimensional learning framework. This is not trivial due to the non-convexity (due to the population heterogeneity) and the target heterogeneity of the problem.

The rest of this paper is organized as follows. Section 2 provides a brief review of the background and the related works to our method. Section 3 presents the details of the proposed HERMIT model and the computational algorithm. Section 4 discusses several extensions of our method. Section 5 shows the theoretical analysis. The empirical evaluations are presented in Sect. 6, followed by the discussions and conclusions in Sect. 7.

## 2 Background and related work

Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be the output target and $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ the input feature vector. GLMs (Nelder and Baker 1972) postulate that the conditional probability density function of $Y$ given $\mathbf{x}$ is

$$f(y \mid \mathbf{x}, \theta) = f(y \mid \varphi, \phi) = \exp\left\{\frac{y\varphi - b(\varphi)}{a(\phi)} + c(y, \phi)\right\},$$

where $\varphi = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ with $\boldsymbol{\beta}$ being the regression coefficient vector, $\phi$ is a dispersion parameter, and $a(\cdot), b(\cdot), c(\cdot)$ are known functions whose forms are determined by the specific distribution. Here we use $\theta$ to denote the collection of all the unknown parameters, i.e., $\theta = (\boldsymbol{\beta}, \phi)$. Least square regression, logistic regression and Poisson regression are all special cases of GLMs. In the presence of population heterogeneity, a standard finite mixture model of GLM postulates that the conditional probability density function of $Y$ given $\mathbf{x}$ is

$$f(y \mid \mathbf{x}, \theta) = \sum_{r=1}^{k} \pi_r f(y \mid \varphi_r, \phi_r),$$

where $\varphi_r = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_r$ with $\boldsymbol{\beta}_r$ being the regression coefficient vector for the $r$th mixture component, and $\pi_r > 0$ $(r = 1, \ldots, k)$ with $\sum_{r=1}^{k} \pi_r = 1$. So FMR model assumes that there are $k$ sub-populations, each of which admits a different conditional relationship between $Y$ and $\mathbf{x}$.

McLachlan and Peel (2004) introduced finite mixture of GLM models. Bartolucci and Scaccia (2005) considered a special case that $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ are different only in their first entries. Khalili and Chen (2007) discussed using sparse penalties such as Lasso and SCAD to perform feature selection for FMR models and showed asymptotic properties of the penalized estimators. Städler et al. (2010) reparameterized the finite mixture of Gaussian regression model and used $\ell_1$ penalization to achieve bounded log-likelihood and consistent feature selection. For multiple targets, Wedel and DeSarbo (1995) proposed finite mixture of GLM models with multivariate targets. These methods only consider the univariate-outcome case. Weruaga and Vía (2015) proposed multivariate Gaussian mixture regression and used $\ell_1$ penalty for sparseness of parameters. Besides mixture of GLMs, there have been many works on mixture of other continuous distributions such as $t$ and Laplace distributions, mainly motivated by the needs of robust estimation for handling heavy tailed or skewed continuous distribution; see, e.g., Wang et al. (2004), Doğru and Arslan (2017), Alfò et al. (2016), Doğru and Arslan (2016), Lim et al. (2016), Bai et al. (2016). However, these methods assume that the targets are of the same type, and only consider interrelationship among tasks with continuous outcomes. Additionally, they all assumed that their FMR model is shared by all the tasks.

In MTL, Kumar and Daumé (2012), Passos et al. (2012), Gong et al. (2012a), Chen et al. (2011), Jacob et al. (2009), Chen et al. (2010) and He and Lawrence (2011) proposed to tackle the problem that different groups of tasks share different information, providing methods to handle anomaly tasks, clustered structure or graph-based structure among tasks. Yang et al. (2009) proposed a multi-task framework to jointly learn tasks with output types of Gaussian and multinomial. Zhang et al. (2012) proposed a multi-modal multi-task model to predict clinical variables for regression and categorical variable for classification jointly. Li et al. (2014) proposed a heterogeneous multi-task learning framework to learn a pose-joint regressor and a sliding window body-part detector in a deep network architecture simultaneously. Nevertheless, these MTL methods cannot handle the heterogeneity of conditional relationship between features and targets.

By contrast, the proposed FMR framework HERMIT is effective for handling sample heterogeneity with mixed type of tasks whose interrelationship are harnessed by structural constraints. Non-asymptotic theoretical guarantees are provided. It also handles anomaly tasks or clustered structure among tasks, for the case that not all the tasks share the same FMR structure.

## 3 HEterogeneous-target Robust MIxTure regression

In this section, we first present the formulation of the main HERMIT model, followed by penalized likelihood estimators with sparse constraint and structural constraint, respectively. We then introduce the associated optimization procedures, and describe how to perform sample clustering and make imputation of the missing/unobserved outcomes on incomplete multi-target outcomes based on the main model. Hyperparameter tuning is discussed at last. Various extensions of the main methodology, including strategies to handle anomaly tasks or clustered tasks, will be introduced in Sect. 4.

### 3.1 Model formulation and estimation criterion

Let $\mathbf{Y} \in \mathbb{R}^{n\times m}$ be the output/target data matrix and $\mathbf{X} \in \mathbb{R}^{n\times d}$ the input/feature data matrix, consisting of $n$ independent samples $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$. As such, there are $m$ different targets with a common set of $d$ features. We allow $\mathbf{Y}$ to contain missing values at random; define $\Omega_i = \{j \in \{1, \ldots, m\}: y_{ij}$ is observed.$\}$ be the collection of indices of observed outcome in the $i$th sample $\mathbf{y}_i$ ($\Omega_i \neq \emptyset$), for $i = 1, \ldots, n$.

To model multiple types of targets, such as continuous, binary, count, etc., we allow $y_{ij}$ to potentially follow different distributions in the exponential-dispersion family, for each $j = 1, \ldots, m$. Specifically, we assume that given $\mathbf{x}_i$, the joint probability density function of $\tilde{\mathbf{y}}_i = \{y_{ij};\ j \in \Omega_i\}$ is

$$f(\tilde{\mathbf{y}}_i \mid \mathbf{x}_i, \theta) = f(y_{ij}, j \in \Omega_i \mid \mathbf{x}_i, \theta) = \sum_{r=1}^{k} \pi_r \prod_{j\in\Omega_i} f(y_{ij} \mid \varphi_{ijr}, \phi_{jr}), \tag{1}$$

where

$$f(y_{ij} \mid \varphi_{ijr}, \phi_{jr}) = \exp\left\{\frac{y_{ij}\varphi_{ijr} - b_j(\varphi_{ijr})}{a_j(\phi_{jr})} + c_j(y_{ij}, \phi_{jr})\right\},$$

$\varphi_{ijr}$ is the natural parameter for the $i$th sample of the $j$th target in the $r$th mixture component, $\phi_{jr}$ is the dispersion parameter of the $j$th target in the $r$th mixture component, and the functions $a_j, b_j, c_j$ $(j = 1, \ldots, m)$ are determined by the specific distribution of the $j$th target. Here, the key assumption is that the $m$ tasks all correspond to the same cluster structure (e.g., the $m$ tasks all have $k$ clusters) determined by the underlying population heterogeneity; given the shared cluster label (e.g., $r$), the tasks within each mixture component then become independent of each other (depicted by the product of their probability density functions). As such, by allowing cluster label sharing, the model provides an effective way to genuinely integrate the learning of the multiple tasks.

Following the setup of GLMs, we assume a linear structure in the natural parameters, i.e.,

$$\varphi_{ijr} = \mathbf{x}_i \boldsymbol{\beta}_{jr}, \tag{2}$$

where $\boldsymbol{\beta}_{jr}$ is the regression coefficient vector of the $j$th response in the $r$th mixture component. Since $\mathbf{x}_i$ is possibly of high dimensionality, the $\boldsymbol{\beta}_{jr}$s are potentially sparse vectors. For example, when the $\boldsymbol{\beta}_{jr}$s for $j = 1, \ldots, m$ share the same sparsity pattern, the tasks share the same set of relevant features within each mixture component. For $r = 1, \ldots, k$, write $\boldsymbol{\beta}_r \in \mathbb{R}^{d\times m} = [\boldsymbol{\beta}_{1r}, \boldsymbol{\beta}_{2r}, \ldots, \boldsymbol{\beta}_{mr}]$ and $\phi_r = [\phi_{1r}, \ldots, \phi_{mr}]^T$. Also write $\boldsymbol{\beta} \in \mathbb{R}^{(d\times m)\times k} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k]$. Let $\theta = \{\boldsymbol{\beta}, \phi_1, \ldots, \phi_k, \pi_1, \ldots, \pi_k\}$ collecting all the unknown parameters, with the parameter space given by $\Theta = \mathbb{R}^{(d\times m)\times k} \times \mathbb{R}_{>0}^{m\times k} \times \Pi$, where $\Pi = \{\pi; \pi_r > 0$ for $r = 1, \ldots, k$ and $\sum_{r=1}^{k} \pi_r = 1\}$.

The data log-likelihood of the proposed model is

$$\ell(\theta \mid \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{n} \log\left(\sum_{r=1}^{k} \pi_r \prod_{j\in\Omega_i} \exp\left\{\frac{y_{ij}\varphi_{ijr} - b_j(\varphi_{ijr})}{a_j(\phi_{jr})} + c_j(y_{ij}, \phi_{jr})\right\}\right). \tag{3}$$

The missing values in $\mathbf{Y}$ simply do not contribute to the likelihood, which follows the same spirit as in matrix completion (Candès and Recht 2009). The proposed model indeed possesses a genuine multivariate flavor, as the different outcomes share the same underlying latent cluster pattern of the heterogeneous population. We then propose to estimate $\theta$ by the following penalized likelihood criterion:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \; -\ell(\theta \mid \mathbf{Y}, \mathbf{X})/n + \mathcal{R}(\boldsymbol{\beta}; \lambda), \tag{4}$$

where $\mathcal{R}(\boldsymbol{\beta}; \lambda)$ is some certain penalty term on the regression coefficients with $\lambda$ being a tuning parameter.

We thus name our proposed method the HEterogeneous-target Robust MIxTure regression (HERMIT). The $\mathcal{R}(\boldsymbol{\beta}; \lambda)$ can be flexibly chosen based on specific needs of feature selection. The first sparse penalties adopted by our model is the $\ell_1$ norm (lasso-type) penalty,

$$\mathcal{R}(\boldsymbol{\beta}; \lambda, \pi) = \lambda \sum_{r=1}^{k} \pi_r^{\gamma} \|\boldsymbol{\beta}_r\|_1, \tag{5}$$

where $\lambda$ is the tuning parameter, $\|\cdot\|_1$ is the entry-wise $\ell_1$ norm, and $\pi_r^{\gamma}$s $(r = 1, \ldots, k)$ are the penalty weights with $\gamma \in \{0, 1/2, 1\}$ being a pre-specified constant. Here the penalty also depends on the unknown mixture proportions $\pi$; when the cluster sizes are expected to be imbalanced, using this weighted penalization with some $\gamma > 0$ is preferred (Städler et al. 2010). This entry-wise regularization approach allows the tasks to have independent set of relevant features. Alternatively, in order to enhance the integrative learning and potentially boost the performance of clustering, it could be beneficial to encourage the internal similarity within each sub-population. Then certain group-wise regularization of the features could be considered, which are widely adopted in multi-task learning. In particular, we consider a component-specific group sparsity pattern to achieve shared feature selection among different tasks, in which the group $\ell_1$ norm penalty is used (Gong et al. 2012a; Jalali et al. 2010),

$$\mathcal{R}(\boldsymbol{\beta}; \lambda, \pi) = \lambda \sum_{r=1}^{k} \pi_r^{\gamma} \|\boldsymbol{\beta}_r\|_{1,2}, \tag{6}$$

where $\|\cdot\|_{1,2}$ denotes the sum of the row $\ell_2$ norms of the enclosed matrix, and the weights are constructed as in (5). The shared feature set in each sub-population can be used to characterize the sub-population and render the whole model more interpretable.

### 3.2 Optimization

We propose a generalized EM (GEM) algorithm (Dempster et al. 1977) to solve the minimization problem in (4). For each $i = 1, \ldots, n$, define $(\delta_{i,1}, \ldots, \delta_{i,k})$ be a set of latent indicator variables, where $\delta_{i,r} = 1$ if the $i$th sample $(\mathbf{y}_i, \mathbf{x}_i)$ belongs to the $r$th component of the mixture model (1) and $\delta_{i,r} = 0$ otherwise. So $\sum_{r=1}^{k} \delta_{i,r} = 1, \forall i$. These indicators are not observed since the cluster labels of the samples are unknown.

Let $\delta$ denote the collection of all the indicator variables. By treating $\delta$ as missing, the EM algorithm proceeds by iteratively optimizing the conditional expectation of the complete log-likelihood criterion.

The complete log-likelihood is given by

$$\begin{aligned}\ell_c(\theta \mid \mathbf{Y}, \mathbf{X}, \delta) = & \sum_{r=1}^{k}\Bigg\{\sum_{i=1}^{n}\sum_{j\in\Omega_i}\delta_{i,r}\left(\frac{y_{ij}\varphi_{ijr} - b_j(\varphi_{ijr})}{a_j(\phi_{jr})} + c_j(y_{ij}, \phi_{jr})\right) \\ & + \sum_{i=1}^{n}\delta_{i,r}\log(\pi_r)\Bigg\},\end{aligned}$$

where $\varphi_{ijr} = \mathbf{x}_i\boldsymbol{\beta}_{jr}$, for $i = 1, \ldots, n,\ j = 1, \ldots, m$, and $r = 1, \ldots, k$. The conditional expectation of the penalized complete negative log-likelihood is then given by

$$Q_{pen}(\theta \mid \theta') = -\mathbb{E}[\ell_c(\theta \mid \mathbf{Y}, \mathbf{X}, \delta)|\mathbf{Y}, \mathbf{X}, \theta']/n + \mathcal{R}(\boldsymbol{\beta}; \lambda, \pi),$$

where $\mathcal{R}(\boldsymbol{\beta}; \lambda, \pi)$ can be any of the penalties in (5) or (6). It is easy to show that deriving $Q_{pen}(\theta \mid \theta')$ boils down to the computation of $\mathbb{E}[\delta_{i,r} \mid \mathbf{Y}, \mathbf{X}, \theta']$, which admits an explicit form.

The EM algorithm proceeds as follows. Let $\theta^{(0)}$ be some given initial values. We repeat the following steps for $t = 0, 1, 2, \ldots$, until convergence of the parameters or the pre-specified maximum number of iteration $T_{out}$ is reached.

*E-Step* Compute $\hat{\rho}_{i,r}^{(t+1)} = \mathbb{E}[\delta_{i,r} \mid \mathbf{Y}, \mathbf{X}, \theta^{(t)}]$. For $\varphi_{ijr}^{(t)} = \mathbf{x}_i\boldsymbol{\beta}_{jr}^{(t)}$,

$$\begin{aligned}&\hat{\rho}_{i,r}^{(t+1)} \\ &= \frac{\pi_r^{(t)}\prod_{j\in\Omega_i}\exp\left\{\left(y_{ij}\varphi_{ijr}^{(t)} - b_j\left(\varphi_{ijr}^{(t)}\right)\right)/a_j\left(\phi_{jr}^{(t)}\right) + c_j\left(y_{ij}, \phi_{jr}^{(t)}\right)\right\}}{\sum_{r'=1}^{k}\pi_{r'}^{(t)}\prod_{j\in\Omega_i}\exp\left\{\left(y_{ij}g_{ijr'}^{(t)} - b_j\left(g_{ijr'}^{(t)}\right)\right)/a_j\left(\phi_{jr'}^{(t)}\right) + c_j\left(y_{ij}, \phi_{jr'}^{(t)}\right)\right\}}.\end{aligned} \tag{7}$$

*M-Step* Minimize $Q_{pen}(\theta \mid \theta^{(t)})$.

(a) Update $\pi = (\pi_1, \ldots, \pi_k)$ by solving

$$\begin{aligned}&\pi^{(t+1)} = \arg\min_{\pi} -\frac{1}{n}\sum_{r=1}^{k}\sum_{i=1}^{n}\hat{\rho}_{i,r}^{(t+1)}\log(\pi_r) + \mathcal{R}(\boldsymbol{\beta}^{(t)}; \lambda, \pi) \\ &s.t. \sum_{r=1}^{k}\pi_r = 1, \pi_r > 0, \forall r.\end{aligned}$$

(b) Update $\boldsymbol{\beta}, \boldsymbol{\Phi}$.

$$(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\Phi}^{(t+1)}) = \arg\min_{\boldsymbol{\beta},\boldsymbol{\Phi}} -\frac{1}{n}\sum_{r=1}^{k}\sum_{i=1}^{n}\hat{\rho}_{i,r}^{(t+1)} \sum_{j\in\Omega_i}\left(\frac{y_{ij}\mathbf{x}_i\boldsymbol{\beta}_{jr} - b_j(\mathbf{x}_i\boldsymbol{\beta}_{jr})}{a_j(\phi_{jr})} + c_j(y_{ij},\phi_{jr})\right) + \mathcal{R}(\boldsymbol{\beta};\lambda,\pi^{(t+1)}).$$

For the problem in (a), Städler et al. (2010) proposed a procedure to lower the objective function by a feasible point, and we find that simply setting $\pi_r^{(t+1)} = \sum_{i=1}^{n}\hat{\rho}_{i,r}/n$ is good enough. For the problem in b), we use an accelerated proximal gradient (APG) method introduced in Nesterov et al. (2007) with the maximum number of iteration of $T_{in}$. The update steps by proximal operators correspond to the chosen penalty form. For the entry-wise $\ell_1$ norm penalty in (5),

$$\widehat{\boldsymbol{\beta}}_r^{(t+1)} = \operatorname{sign}\left(\widetilde{\boldsymbol{\beta}}_r^{(t)}\right) \circ \max\{0, |\widetilde{\boldsymbol{\beta}}_r^{(t)}| - \tau\lambda(\pi_r^{(t+1)})^{\gamma}\}, \tag{8}$$

where $\circ$ denotes entry-wise product, $\widetilde{\boldsymbol{\beta}}_r^{(t)} = \boldsymbol{\beta}_r^{(t)} + \tau\triangle\boldsymbol{\beta}_r^{(t)}$, $\tau$ denotes the step size, and $\triangle\boldsymbol{\beta}_r^{(t)}$ denotes the update direction of $\boldsymbol{\beta}_r^{(t)}$ determined by APG. For the group $\ell_1$ norm penalty in (6),

$$\widehat{\boldsymbol{\beta}}_{r,j}^{(t+1)} = \widetilde{\boldsymbol{\beta}}_{r,j}^{(t)} \circ \max\left\{0, 1 - \tau\lambda\left(\pi_r^{(t+1)}\right)^{\gamma} / \|\widetilde{\boldsymbol{\beta}}_{r,j}^{(t)}\|_2\right\},$$

where $\boldsymbol{\beta}_{r,j}$ denotes the $j$th column of $\boldsymbol{\beta}_r$. We adopt the active set algorithm in Städler et al. (2010) to speed up the computation.

The time complexity of our algorithm using the speed up technique is $O(T_{out}kmnsT_{in})$ with $s$ being the number of non-zero parameters. The algorithm performs well in practice, and we have not observed any convergence issues in our extensive numerical studies.

### 3.3 Clustering of samples and imputation of missing targets

From the model estimation, we can get estimates of both the conditional probabilities $p(\delta_{i,r} = 1 \mid y_{ij'}, j' \in \Omega_i, \mathbf{x}_i, \theta)$ and the conditional means $\mathbb{E}[Y_{ij} \mid \mathbf{x}_i, \theta, \delta_{i,r} = 1]$, where $\mathbb{E}[Y_{ij} \mid \mathbf{x}_i, \theta, \delta_{i,r} = 1] = \mu_j(\varphi_{ijr}) = b_j'(\mathbf{x}_i\boldsymbol{\beta}_{jr})$. Specifically, the conditional probabilities can be estimated by $p(\delta_{i,r} = 1 \mid y_{ij'}, j' \in \Omega_i, \mathbf{x}_i, \hat{\theta})$ which corresponds to (7), taking $t = T_{out}$. The conditional expectations can be estimated as $\mathbb{E}[Y_{ij} \mid \mathbf{x}_i, \hat{\theta}, \delta_{i,r} = 1] = b_j'(\mathbf{x}_i\hat{\boldsymbol{\beta}}_{jr})$.

For clustering the samples, we adopt the Bayes rule, i.e., for $i = 1, \ldots, n$,

$$\hat{r}_i = \underset{r}{\operatorname{argmax}}\ p(\delta_{i,r} = 1 \mid y_{ij'}, j' \in \Omega_i, \mathbf{x}_i, \hat{\theta}). \tag{9}$$

Following the idea of Jacobs et al. (1991), we propose to make imputation for the missing outcomes by

$$\hat{y}_{ij} = \sum_{r=1}^{k} p(\delta_{i,r} = 1 \mid y_{ij'}, j' \in \Omega_i, \mathbf{x}_i, \hat{\theta})\mathbb{E}[Y_j \mid \mathbf{x}_i, \hat{\theta}, \delta_{i,r} = 1], \text{ for } j \notin \Omega_i. \quad (10)$$

### 3.4 Tuning hyper-parameters

Unless otherwise specified, all the hyper-parameters, including regularization coefficients $\lambda$s and the number of clusters $k$, are tuned to maximize the data log-likelihood in (3) on the held-out validation data set. In other words, we fit models on training data with different specific hyper-parameter settings, and then the optimal model is chosen as the one that gives the highest log-likelihood in (3) of the held-out validation data set. This approach is fairly standard and has been widely used in existing works (Städler et al. 2010). Moreover, cross validation and various information criteria (Bhat and Kumar 2010; Aho et al. 2014) can also be applied to determine hyper-parameters.

## 4 Extensions

We provide several extensions of the proposed Hermit approach described in Sect. 3, including robust estimation against outlier samples, handling anomaly tasks or clustered structure among tasks, and modeling mixture probabilities for feature-based prediction.

### 4.1 Robust estimation

To perform robust estimation for parameters in the presence of outlier samples, we propose to adopt the mean shift penalization approach (She and Owen 2011). Specifically, we extend the natural parameter model to the following additive form,

$$\varphi_{ijr} = \mathbf{x}_i \boldsymbol{\beta}_{jr} + \zeta_{ijr}, \quad (11)$$

where $\zeta_{ijr}$ is a case-specific mean shift parameter to capture the potential deviation from the linear model. Apparently, when $\zeta_{ijr}$ is allowed to vary without any constraint, it can make the model fit as perfect as possible for every $y_{ijr}$. The merit of this approach is realized by assuming certain sparsity structure of the $\zeta_{ijr}$s, so that only a few of them have nonzero values corresponding to anomalies. Write $\boldsymbol{\zeta}_r \in \mathbb{R}^{n\times m} = [\boldsymbol{\zeta}_{1r}, \boldsymbol{\zeta}_{2r}, \ldots, \boldsymbol{\zeta}_{mr}]$ for $r = 1, \ldots, k$, and $\boldsymbol{\zeta} \in \mathbb{R}^{(n\times m)\times k} = [\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_r]$. We can then conduct joint model estimation and outlier detection by extending (4) to

$$(\hat{\theta}, \hat{\boldsymbol{\zeta}}) = \arg\min_{\theta\in\Theta,\zeta} \; -\ell(\theta \mid \mathbf{Y}, \mathbf{X})/n + \mathcal{R}(\boldsymbol{\beta}; \lambda_1) + \mathcal{R}(\boldsymbol{\zeta}; \lambda_2), \quad (12)$$

where, for example, the penalty on $\boldsymbol{\zeta}$ can be chosen as the group $\ell_1$ penalty,

$$\mathcal{R}(\boldsymbol{\zeta};\lambda_2)=\lambda_2\sum_{i=1}^{n}\sqrt{\sum_{jr}\zeta_{ijr}^2}, \tag{13}$$

so that entries of $\boldsymbol{\zeta}$ are nonzero for only a few data samples.

The proposed GEM algorithm can be readily extended to handle the inclusion of $\boldsymbol{\zeta}$, for which we omit the details.

### 4.2 Handling anomaly tasks

Besides outlier samples, certain tasks, referred to as anomaly tasks, may not follow the assumed shared structure and thus can ruin the overall model performance. To handle anomaly tasks, though it is also intuitive to adopt the approach above, our numerical study suggests that its performance is sensitive to the tuning parameters. Here, we adopt the idea of Koller (1996), by utilizing the estimated conditional probabilities to measure how well a task is concordant with the estimated mixture structure. Consider the $h$th task. The main idea is to measure the discrepancy between $p(\delta_{ir}=1\mid y_{ij}, j\in\Omega_i,\mathbf{x}_i,\theta)$, the conditional probability based on data from all observed targets, and $p(\delta_{ir}=1\mid y_{ih},\mathbf{x}_i,\theta)$, the conditional probability based on only the $h$th task. If $h$th task is an anomaly task, it is expected that the two conditional probabilities would differ more from each other (Koller 1996; Law et al. 2002).

For $r=1,\ldots,k$, $i=1,\ldots,n$, let

$$P_{\Omega,ir}=p(\delta_{ir}=1\mid y_{ij}, j\in\Omega_i,\mathbf{x}_i,\hat{\theta})=\frac{\hat{\pi}_r\prod_{j\in\Omega_i}f(y_{ij}\mid\hat{\varphi}_{ijr},\hat{\phi}_{jr})}{\sum_{r'=1}^{k}\hat{\pi}_{r'}\prod_{j\in\Omega_i}f(y_{ij}\mid\hat{\varphi}_{ijr'},\hat{\phi}_{jr'})},$$

$$P_{h,ir}=p(\delta_{ir}=1\mid y_{ih},\mathbf{x}_i,\hat{\theta})=\frac{\hat{\pi}_r f(y_{ih}\mid\hat{\varphi}_{ihr},\hat{\phi}_{hr})}{\sum_{r'=1}^{k}\hat{\pi}_{r'}f(y_{ih}\mid\hat{\varphi}_{ihr'},\hat{\phi}_{hr'})}. \tag{14}$$

Define $\mathbf{P}_\Omega=[P_{\Omega,ir}]_{n\times k}$ and $\mathbf{P}_h=[P_{h,ir}]_{n\times k}$. Then we define the concordant score of the $h$th task as

$$score(h)=-(D_{KL}(\mathbf{P}_\Omega\parallel\mathbf{P}_h)+D_{KL}(\mathbf{P}_h\parallel\mathbf{P}_\Omega))/(2n),\ \ h=1,\ldots,m, \tag{15}$$

where $D_{KL}$ is the widely used Kullback–Leibler divergence (Cover and Thomas 2012).

The tasks can then be ranked based on their concordant scores. As such, the detection of anomaly tasks boils down to a one-dimensional outlier detection problem. After anomaly tasks are detected, their FMR models can be built.

### 4.3 Handling clustered structure among tasks

In practice, tasks may be clustered into groups such that each task group has its own model structure. Here we assume that each cluster of tasks shares a FMR structure defined in (1), and propose to construct a similarity matrix to discover the potential cluster pattern among tasks.

We consider a two-stage strategy. First, each task learns a FMR model on the training data independently with the same pre-fixed $k$. Then we get $\mathbf{P}_h = [P_{h,ir}]_{n\times k}$ for all $h = 1, \ldots, m$, where

$$P_{h,ir} = p(\delta_{h,ir} = 1 \mid y_{ih}, \mathbf{x}_i, \hat{\theta}_h) = \frac{\hat{\pi}_{hr} f(y_{ih} \mid \hat{\varphi}_{ihr}, \hat{\phi}_{hr})}{\sum_{r'=1}^{k} \hat{\pi}_{hr'} f(y_{ih} \mid \hat{\varphi}_{ihr'}, \hat{\phi}_{hr'})},$$

and $\delta_{h,ir}(i = 1, \ldots, n, r = 1, \ldots, k)$ and $\hat{\pi}_{hr}(r = 1, \ldots, k)$ are the latent variables and the estimated prior probabilities of the $h$th task, respectively.

Second, we adopt Normalized Mutual Information (NMI) (Strehl and Ghosh 2002a) to measure the similarity between each pair of tasks. We choose NMI instead of Kullback–Leibler divergence because NMI can handle the case that the orders of clusters of two $k$-cluster structures are different. Specifically, given two methods to estimate latent variables, which are denoted by method $u$ and method $v$, let $\mathbf{P}_u = [P_{u,ir}]_{n\times k}$ and $\mathbf{P}_v = [P_{v,ir}]_{n\times k}$ denote the estimated probability of latent variables of method $u$ and method $v$, respectively, where $P_{u,ir} = p(\delta_{u,ir} = 1)$, $P_{v,ir} = p(\delta_{v,ir} = 1)$ for $i = 1, \ldots, n, r = 1, \ldots, k$. NMI is defined as

$$NMI(\mathbf{P}_u, \mathbf{P}_v) = \frac{I(\mathbf{P}_u, \mathbf{P}_v)}{\sqrt{I(\mathbf{P}_u, \mathbf{P}_u) I(\mathbf{P}_v, \mathbf{P}_v)}}, \quad u = 1, \ldots, m, v = 1, \ldots, m, \tag{16}$$

where $I(\mathbf{P}_u, \mathbf{P}_v)$ denotes the mutual information between $\mathbf{P}_u, \mathbf{P}_v$ such that

$$I(\mathbf{P}_u, \mathbf{P}_v) = \sum_{a=1}^{k}\sum_{b=1}^{k} p(\delta_{u,a} = 1, \delta_{v,b} = 1) \log\left(\frac{p(\delta_{u,a} = 1, \delta_{v,b} = 1)}{p(\delta_{u,a} = 1) p(\delta_{v,b} = 1)}\right).$$

Following Strehl and Ghosh (2002a), we approximate $p(\delta_{u,a} = 1, \delta_{v,b} = 1)$, $p(\delta_{u,a} = 1)$ and $p(\delta_{v,b} = 1)$ by

$$\begin{aligned}
p(\delta_{u,a} = 1) &\approx \frac{1}{n}\sum_{i=1}^{n} p(\delta_{u,ia} = 1) = \frac{1}{n}\sum_{i=1}^{n} P_{u,ia},\\
p(\delta_{v,b} = 1) &\approx \frac{1}{n}\sum_{i=1}^{n} p(\delta_{v,ib} = 1) = \frac{1}{n}\sum_{i=1}^{n} P_{v,ib},\\
p(\delta_{u,a} = 1, \delta_{v,b} = 1) &\approx \frac{1}{n}\sum_{i=1}^{n} p(\delta_{u,ia} = 1, \delta_{v,ib} = 1)\\
&\approx \frac{1}{n}\sum_{i=1}^{n} p(\delta_{u,ia} = 1) p(\delta_{v,ib} = 1) = \frac{1}{n}\sum_{i=1}^{n} P_{u,ia} P_{v,ib}.
\end{aligned}$$

As such, given the estimated models $\hat{\theta}_u, \hat{\theta}_v$ for the $u$th and the $v$th task, respectively, we treat $p(\delta_{u,ir} = 1 \mid y_{iu}, \mathbf{x}_i, \hat{\theta}_u)$ and $p(\delta_{v,ir} = 1 \mid y_{iv}, \mathbf{x}_i, \hat{\theta}_v)$ as $p(\delta_{u,ir} = 1)$ and

$p(\delta_{v,ir} = 1)$, respectively, for $i = 1, \ldots, n$, $r = 1, \ldots, k$. Then NMI between each pair of tasks are computed by (16). We note that for simplicity we set the pre-fixed $k$ to be the same, but in general $k$ can be different for different tasks by the definition of Mutual Information.

Given the similarity between each pair of tasks, any similarity-based clustering method can be applied to cluster $m$ tasks into groups. Empirically, the performance of task clustering is not sensitive to the pre-fixed $k$. As such, we set the pre-fixed $k$ to be 20 in this paper. We then apply the proposed HERMIT approach separately for each task group.

### 4.4 Modeling mixture probabilities

In real-applications, one may require to use $\mathbf{x}_i$ only to infer the latent variables $\delta_{i,r}$ and then to predict $\mathbf{y}_i$, for $i = 1, \ldots, n$, $r = 1, \ldots, k$. Here we further extend our method following the idea of Mixture-Of-Experts (MOE) (Yuksel et al. 2012) model; the only modification is that $\pi_r$ in (1) is assumed to be function of $\mathbf{x}_i$, for $i = 1, \ldots, n$.

To be specific, let $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k] \in \mathbb{R}^{d \times k}$ collect regression coefficient vectors for a multinomial linear model. We assume that given $\mathbf{x}_i$, the joint probability density function of $\tilde{\mathbf{y}}_i = \{y_{ij}; j \in \Omega_i\}$ in (1) is replaced by

$$\begin{aligned} f(\tilde{\mathbf{y}}_i \mid \mathbf{x}_i, \theta, \boldsymbol{\alpha}) &= f(y_{ij}, j \in \Omega_i \mid \mathbf{x}_i, \theta, \boldsymbol{\alpha}) \\ &= \sum_{r=1}^{k} p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \alpha_r) \prod_{j \in \Omega_i} f(y_{ij} \mid \varphi_{ijr}, \phi_{jr}), \end{aligned}$$

where

$$p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \alpha_r) = \frac{\exp(\mathbf{x}_i \boldsymbol{\alpha}_r)}{\sum_{r'=1}^{k} \exp(\mathbf{x}_i \boldsymbol{\alpha}_{r'})} \tag{17}$$

is referred to as the gating probability. All the other terms are defined the same as in (1).

Let $\theta_2 = \{\boldsymbol{\beta}, \phi_1, \ldots, \phi_k\}$, with the parameter space $\Theta_2 = \mathbb{R}^{(d \times m) \times k} \times \mathbb{R}_{>0}^{m \times k}$. The data log-likelihood of the MOE model is

$$\begin{aligned} \ell(\theta_2, \boldsymbol{\alpha} \mid \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{n} \log\Bigg( &\sum_{r=1}^{k} p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \alpha_r) \\ &\prod_{j \in \Omega_i} \exp\left\{ \frac{y_{ij}\varphi_{ijr} - b_j(\varphi_{ijr})}{a_j(\phi_{jr})} + c_j(y_{ij}, \phi_{jr}) \right\} \Bigg). \end{aligned} \tag{18}$$

The model estimation is conducted by extending (4) to

$$(\hat{\theta}_2, \hat{\boldsymbol{\alpha}}) = \arg\min_{\theta_2 \in \Theta_2, \boldsymbol{\alpha}} \; -\ell(\theta_2, \boldsymbol{\alpha} \mid \mathbf{Y}, \mathbf{X}) + \mathcal{R}(\boldsymbol{\beta}; \lambda_1)/n + \mathcal{R}(\boldsymbol{\alpha}; \lambda_2), \tag{19}$$

where, for example, the penalty on $\boldsymbol{\alpha}$ can be chosen as the lasso type penalty,

$$\mathcal{R}(\boldsymbol{\alpha}; \lambda_2) = \lambda_2 \|\boldsymbol{\alpha}\|_1. \tag{20}$$

The minimization problem in (19) is also solved by GEM, for which the optimization procedure is similar to that in Sect. 3.2. The differences occur at the E-step:

$$\hat{\rho}_{i,r}^{(t+1)} = \frac{p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \alpha_r^{(t)}) \prod_{j \in \Omega_i} f(y_{ij} \mid \varphi_{ijr}^{(t)}, \phi_{jr}^{(t)})}{\sum_{r'=1}^{k} p(\delta_{i,r'} = 1 \mid \mathbf{x}_i, \alpha_{r'}^{(t)}) \prod_{j \in \Omega_i} f(y_{ij} \mid \varphi_{ijr'}^{(t)}, \phi_{jr'}^{(t)})}, \tag{21}$$

where $f(y_{ij} \mid \varphi_{ijr}^{(t)}, \phi_{jr}^{(t)}) = \exp\{(y_{ij}\varphi_{ijr}^{(t)} - b_j(\varphi_{ijr}^{(t)}))/a_j(\phi_{jr}^{(t)}) + c_j(y_{ij}, \phi_{jr}^{(t)})\}$, at the optimization for $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^{(t+1)} = \arg\min_{\boldsymbol{\alpha}} -\frac{1}{n} \sum_{r=1}^{k} \sum_{i=1}^{n} \hat{\rho}_{i,r}^{(t+1)} \log p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \alpha_r) + \mathcal{R}(\boldsymbol{\alpha}^{(t)}; \lambda_2), \tag{22}$$

and at the computation of $\pi$: $\pi_r^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \alpha_r^{(t+1)})$ for $r = 1, \ldots, k$.

## 5 Theoretical analysis

We study the estimation and variable selection performance of HERMIT under the high-dimensional framework with $d \gg n$. Both $m$ and $k$, on the other hand, are considered as fixed. This is because usually, the number of interested targets and the number of desired clusters are not large in many real problems. Here we only present the setup and the main results on non-asymptotic oracle inequalities to bound the excess risk and false selection, leaving detailed derivations in the “Appendix” section. Our results generalize Städler et al. (2010) to cover mixture regression models with (1) multivariate, heterogeneous (mixed-type) and incomplete response and (2) shared feature grouping sparse structure. This is not trivial due to the non-convexity and the triple heterogeneity of the problem. It turns out that additional condition on the tail behaviors of the conditional density $f(\mathbf{y} \mid \mathbf{x}, \theta)$ is required. Fortunately, the required conditions are still satisfied by a broad range of distributions.

### 5.1 Notations and conditions on the conditional density

We firstly introduce some notations. Denote the regression parameters that are subject to regularization by $\beta = \mathrm{vec}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$, $\phi = \mathrm{vec}(\Phi_1, \ldots, \Phi_k)$, where $\mathrm{vec}(\cdot)$ is the vectorization operator. The other parameters in the mixture model are denoted by $\eta = \mathrm{vec}(\log(\phi), \log(\pi))$, where $\log(\cdot)$ is entry-wisely applied. Denote the true parameter by $\theta_0 = (\boldsymbol{\beta}_0, \Phi_{0,1}, \ldots, \Phi_{0,k}, \pi_{0,1}, \pi_{0,k-1})$ to be estimated under the FMR model defined in (1) and (2). In the sequel, we always use subscripts “0” to represent parameters or structures under the true model. To study sparsity recovery, denote the set of indices of non-zero entries of the true parameter by $S$. We use $\lesssim$ to indicate that the inequality holds up to some multiplicative numerical constants. To focus on the main idea, we consider the case of $\gamma = 0$ in the following analysis.

We define average excess risk for fixed design points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ based on Kullback–Leibler divergence as

$$\begin{aligned}\bar{\varepsilon}(\theta \mid \theta_0) &= \frac{1}{n}\sum_{i=1}^{n} \varepsilon(\theta \mid \mathbf{x}_i, \theta_0),\ \ \varepsilon(\theta \mid \mathbf{x}_i, \theta_0) \\ &= -\int \log\left(\frac{f(\tilde{\mathbf{y}}_i \mid \mathbf{x}_i, \theta)}{f(\tilde{\mathbf{y}}_i \mid \mathbf{x}_i, \theta_0)}\right) f(\tilde{\mathbf{y}}_i \mid \mathbf{x}_i, \theta_0) d\tilde{\mathbf{y}}_i,\end{aligned}$$

where $f(\tilde{\mathbf{y}}_i \mid \mathbf{x}_i, \theta)$ is defined in (1).

To impose the conditions on $f(\tilde{\mathbf{y}}_i \mid \mathbf{x}_i, \theta)$, denote $\psi_i = \text{vec}(\varphi_i, \eta)$, where $\varphi_i = \text{vec}(\{\varphi_{ijr};\ j \in \Omega_i, r = 1, \ldots, k\})$, and denote $\psi = \text{vec}(\psi_1, \ldots, \psi_n)$. As such, we may write $f(\tilde{\mathbf{y}}_i \mid \mathbf{x}_i, \theta) = f(\tilde{\mathbf{y}}_i \mid \psi_i)$, $\ell(\theta \mid \tilde{\mathbf{y}}_i, \mathbf{x}_i) = \log f(\tilde{\mathbf{y}}_i \mid \mathbf{x}_i, \theta) = \ell(\psi_i \mid \tilde{\mathbf{y}}_i)$, and $\bar{\varepsilon}(\psi \mid \psi_0) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon(\psi_i \mid \psi_{0,i}) = \bar{\varepsilon}(\theta \mid \theta_0)$.

Without loss of much generality, the model parameters are assumed to be in a bounded parameter space for a constant $K$:

$$\begin{aligned}\tilde{\Theta} \subset \Bigg\{\theta;\ &\max_{i=1,\ldots,n} \|\varphi_i(\mathbf{x}_i, \boldsymbol{\beta})\|_\infty \leq K,\ \max_{j=1,\ldots,m} |\log a_j(\phi)| \leq K, \\ &\max_{j=1,\ldots,m} \log|b_j'(\phi)| \leq K, \|\log\phi\|_\infty \leq K, \\ &-K \leq \log\pi_1 \leq 0, \ldots, -K \leq \log\pi_{k-1} \leq 0, \\ &\sum_{r=1}^{k-1}\pi_r < 1, \pi_k = 1 - \sum_{r=1}^{k-1}\pi_r\Bigg\}.\end{aligned} \tag{23}$$

We present the following conditions on $f(\tilde{\mathbf{y}}_i \mid \psi_i)$.

**Condition 1** *For some function $G_1(\cdot) \in \mathbb{R}$, for $i = 1, \ldots, n$,*

$$\sup_{\theta\in\tilde{\Theta}} \left\|\frac{\partial\ell(\psi_i \mid \tilde{\mathbf{y}}_i)}{\partial\psi_i}\right\|_\infty \leq G_1(\tilde{\mathbf{y}}_i).$$

**Condition 2** *For a constant $c_1 \geq 0$, and some constants $c_2, c_3, c_4, c_5 \geq 0$ depending $K$, and for $M > c_4$, we assume for $i = 1, \ldots, n$,*

$$\begin{aligned}\mathbb{E}[|G_1(\tilde{\mathbf{y}}_i)|1\{|G_1(\tilde{\mathbf{y}}_i)| > M\}] &\leq \left[c_3\left(\frac{M}{c_2}\right)^{c'} + c_5\right]\exp\left\{-\left(\frac{M}{c_2}\right)^{1/c_1}\right\}, \\ \mathbb{E}[|G_1(\tilde{\mathbf{y}}_i)|^2 1\{|G_1(\tilde{\mathbf{y}}_i)| > M\}] &\leq \left[c_3\left(\frac{M}{c_2}\right)^{c'} + c_5\right]^2\exp\left\{-2\left(\frac{M}{c_2}\right)^{1/c_1}\right\},\end{aligned}$$

*where $\tilde{\mathbf{y}}_i = \{y_{ij};\ j \in \Omega_i\}$, $c' = 2 + 3/c_1$ and $1\{\cdot\}$ denotes the indicator function.*

**Condition 3** *It holds that,*

$$\min_{i=1,\dots,n} \Lambda_{\min}(I(\psi_{0,i})) > 1/c_0 > 0,$$

*where $c_0$ is a constant, $\Lambda^2_{\min}(A)$ is the smallest eigenvalue of a symmetric, positive semi-definite matrix $A$ and for $i = 1, \dots, n$, $I(\psi_{0,i})$ is the Fisher information matrix such that*

$$I(\psi_{0,i}) = -\int \frac{\partial^2 \ell(\psi_{0,i} \mid \tilde{\mathbf{y}}_i)}{\partial \psi_{0,i} \partial \psi_{0,i}^T} f(\tilde{\mathbf{y}}_i \mid \psi_{0,i}) d\tilde{\mathbf{y}}_i.$$

The first condition follows from Städler et al. (2010), which aims to bound $\partial \ell(\psi_i \mid \tilde{\mathbf{y}}_i)/\partial \psi_i$ with known $\tilde{\mathbf{y}}_i$, for $i = 1, \dots, n$. The second condition is about the tail behaviors of $f(\tilde{\mathbf{y}}_i \mid \mathbf{x}_i, \theta)$. The third condition depicts the local convexity of $\ell$ at the point $\theta_0$. Condition 1 and 2 can cover a broad range of distributions for $f$, including but not limited to mixture of sub-exponential distributions, such as our proposed HERMIT model with known dispersion parameters, c.f., Lemma 1.

**Lemma 1** *Condition 1 and 2 hold for the heterogeneous mixture distribution $f(\tilde{\mathbf{y}}_i \mid \mathbf{x}_i, \theta)$ defined in (1) with known dispersion parameters.*

The following two quantities will be used.

$$\lambda_0 = \sqrt{mk} M_n \log n \sqrt{\frac{\log(d \vee n)}{n}}, \quad M_n = c_2 (\log n)^{c_1}, \tag{24}$$

where $c_1, c_2$ are the same constants as in Condition 2. More specifically, we choose $c_1 = 1/2, 0, 1$ for Gaussian, Bernoulli and Poisson task, respectively.

### 5.2 Results for Lasso-type estimator

Consider first the penalized estimator defined in (4) with the $\ell_1$ penalty in (5). Following Bickel et al. (2009) and Städler et al. (2010), we impose the following restricted eigenvalue condition on the design.

**Condition 4** (Restricted eigenvalue condition) *For all $\beta \in \mathbb{R}^{dmk}$ satisfying $\|\beta_{S^c}\|_1 \le 6\|\beta_S\|_1$, it holds that for some constant $\kappa \ge 1$,*

$$\|\beta_S\|_2^2 \le \kappa^2 \|\varphi\|_{Q_n}^2 = \frac{\kappa^2}{n} \sum_{i=1}^{n} \sum_{j \in \Omega_i} \sum_{r=1}^{k} (\mathbf{x}_i \boldsymbol{\beta}_{jr})^2.$$

**Theorem 1** *Consider the* HERMIT *model in (1) with $\theta_0 \in \tilde{\Theta}$ , and consider the penalized estimator (4) with the $\ell_1$ penalty in (5). Assume Conditions 1–4 hold. Suppose*

*$\sqrt{mk} \lesssim n/M_n$, and take $\lambda > 2T\lambda_0$ for some constant $T > 1$. For some constant $c > 0$ and large enough $n$, with probability*

$$1 - c\exp\left(-\frac{\log^2 n \log(d \vee n)}{c}\right) - \frac{1}{n}, \tag{25}$$

*we have*

$$\bar{\varepsilon}(\hat{\theta} \mid \theta_0) + 2(\lambda - T\lambda_0)\|\hat{\beta}_{S^c}\|_1 \leq 4(\lambda + T\lambda_0)^2\kappa^2 c_0^2 s \tag{26}$$

*where $s$ is the number of non-zero parameters of $w_0$.*

Theorem 1 suggests that the average excess risk has a convergence rate of the order $O(s\lambda_0^2) = O((\log n)^{2+2c_1}\log(d \vee n)smk/n)$, by taking $\lambda = 2T\lambda_0$ and using $\lambda_0$ and $M_n$ as defined in (24). Also, the degree of false selection measured by $\|\hat{\beta}_{S^c}\|_1$ converge to zero at rate $O(s\lambda_0) = O(s\sqrt{(\log n)^{2+2c_1}\log(d \vee n)mk/n})$.

Similar to Städler et al. (2010), under weaker conditions without the restricted eigenvalue assumption on the design, we still achieve the consistency for the average excess risk.

**Theorem 2** *Consider the* Hermit *model in* (1) *with $\theta_0 \in \tilde{\Theta}$, and consider the penalized estimator* (4) *with the $\ell_1$ penalty in* (5). *Assume Conditions* 1–3 *hold. Suppose*

$$\|\beta_0\|_1 = o\left(\sqrt{n/((\log n)^{2+2c_1}\log(d \vee n)mk)}\right),$$
$$\sqrt{mk} = o\left(\sqrt{n/((\log n)^{2+2c_1}\log(d \vee n))}\right)$$

*as $n \to \infty$, and take $\lambda = C\sqrt{(\log n)^{2+2c_1}\log(d \vee n)mk/n}$ for some constant $C > 0$ sufficiently large. For some constant $c > 0$ and large enough $n$, with the following probability $1 - c\exp\left(-\frac{(\log n)^2\log(d \vee n)}{c}\right) - \frac{1}{n}$, we have $\bar{\varepsilon}(\hat{\theta} \mid \theta_0) = o_P(1)$.*

### 5.3 Results for group-Lasso type estimator

Consider the following general form of the group $\ell_1$ penalty,

$$\mathcal{R}(\boldsymbol{\beta}) = \lambda \sum_{p=1}^{P} \|\boldsymbol{\beta}_{\mathcal{G}_p}\|_F, \tag{27}$$

where $\mathcal{G}_1, \ldots, \mathcal{G}_P$ are index collections such that $\mathcal{G}_p \bigcap \mathcal{G}_{p'} = \emptyset$ for $p \neq p'$ and $\bigcup_{p=1}^{P} \mathcal{G}_p = \bigcup_{l=1}^{d}\bigcup_{j=1}^{m}\bigcup_{r=1}^{k}(l, j, r)$ equals the universal set of indices of $\boldsymbol{\beta} \in \mathbb{R}^{(d\times m)\times k}$, i.e., $\boldsymbol{\beta}_{\mathcal{G}_p}$ is the $p$th group of $\boldsymbol{\beta}$. $\|\cdot\|_F$ denotes the Frobenius norm and here for $p = 1, \ldots, P$, $\|\boldsymbol{\beta}_{\mathcal{G}_p}\|_F = \sqrt{\sum_{(l,j,r)\in\mathcal{G}_p} w_{ljr}^2}$. This penalty form generalizes the row-wise group sparsity in (6).

Denote $\mathcal{I} = \{p : \boldsymbol{\beta}_{0,\mathcal{G}_p} = \mathbf{0}\}$ and $\mathcal{I}^c = \{p : \boldsymbol{\beta}_{0,\mathcal{G}_p} \neq \mathbf{0}\}$, where $\boldsymbol{\beta}_{0,\mathcal{G}_p}$ is the $p$th group of $\boldsymbol{\beta}_0$. Now denote by $s$ the size of $\mathcal{I}$, with some abuse of notation. We impose the following group-version restricted eigenvalue condition.

**Condition 5** *For all $\boldsymbol{\beta} \in \mathbb{R}^{(d\times m)\times k}$ satisfying*

$$\sum_{p\in\mathcal{I}^c} \|\boldsymbol{\beta}_{\mathcal{G}_p}\|_F \leq 6 \sum_{p\in\mathcal{I}} \|\boldsymbol{\beta}_{\mathcal{G}_p}\|_F,$$

*it holds that for some constant $\kappa \geq 1$,*

$$\sum_{p\in\mathcal{I}} \|\boldsymbol{\beta}_{\mathcal{G}_p}\|_F^2 \leq \kappa^2 \|\varphi\|_{Q_n}^2.$$

**Theorem 3** *Consider the* HERMIT *model in* (1) *with $\theta_0 \in \tilde{\Theta}$, and consider the penalized estimator* (4) *with the group $\ell_1$ penalty in* (27).

*(a) Assume Conditions* 1–3 *and* 5 *hold. Suppose $\sqrt{mk} \lesssim n/M_n$, and take $\lambda > 2T\lambda_0$ for some constant $T > 1$. For some constant $c > 0$ and large enough $n$, with the following probability $1 - c\exp\left(-\frac{(\log n)^2 \log(d\vee n)}{c}\right) - \frac{1}{n}$, we have*

$$\bar{\varepsilon}(\hat{\theta} \mid \theta_0) + 2(\lambda - T\lambda_0) \sum_{p\in\mathcal{I}^c} \|\widehat{\boldsymbol{\beta}}_{\mathcal{G}_p}\|_F \leq 4(\lambda + T\lambda_0)^2 \kappa^2 c_0^2 s.$$

*(b) Assume Conditions* 1–3 *hold (without Condition* 5*), and assume*

$$\sum_{p=1}^{P} \|\boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F = o\left(\sqrt{n/((\log n)^{2+2c_1} \log(d\vee n)mk)}\right),$$

$$\sqrt{mk} = o\left(\sqrt{n/((\log n)^{2+2c_1} \log(d\vee n))}\right)$$

*as $n \to \infty$. Let $\lambda = C\sqrt{(\log n)^{2+2c_1} \log(d\vee n)mk/n}$ for some $C > 0$ sufficiently large. Then for some constant $c > 0$ and large enough $n$, with the following probability $1 - c\exp\left(-\frac{(\log n)^2 \log(d\vee n)}{c}\right) - \frac{1}{n}$, we have $\bar{\varepsilon}(\hat{\theta} \mid \theta_0) = o_P(1)$.*

So the average excess risk has a convergence rate of $O(s\lambda_0^2)$, and the degree of false group selection, as measured by $\sum_{p\in\mathcal{I}^c} \|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}\|_F$, converges to zero at rate $O(s\lambda_0)$. The estimator in (4) using other group $\ell_1$ penalties such as (6) are special cases, so the results of Theorem 3 still apply.

*Remark* Our results can be extended to the mean-shifted natural parameter model as in (11), with a modified restricted eigenvalue condition. See the “Appendix” section for some details.

## 6 Experiments

In this section, we present empirical studies on both synthetic and real-world data sets.

### 6.1 Methods for comparison

We evaluate the following versions of the proposed HERMIT approach.

(1) Single task learning (**Single**): It is a special case of the HERMIT estimator (4) with (5), where each task is learned separately.
(2) Separately learning (**Sep**): It is a special case of the HERMIT estimator (4) with (5), where each type (Gaussian, Bernoulli or Poisson) of tasks is learned separately.
(3) Mixed learning with entry-wise sparsity (**Mix**): It is the proposed HERMIT estimator (4) with (5) where all the tasks are jointly learned. To compare with **Sep**, we allow different tuning parameters for different types of outcomes.
(4) Mixed learning with group sparsity (**Mix GS**): It is the proposed HERMIT estimator (4) with (6).
(5) Mixed learning Mixture-Of-Experts model with entry-wise sparsity (**Mix MOE**): It is the proposed HERMIT estimator (19) with (5) and (20).
(6) Mixed learning Mixture-Of-Experts model with group sparsity (**Mix MOE GS**): It is the proposed HERMIT estimator (19) with (6) and (20).

Besides the above FMR methods, we also evaluate several non-FMR multi-task methods below for comparison, some of which handle certain kinds of heterogeneities, such as anomaly tasks, clustered tasks and heterogeneous responses. Since they are non-FMR, they learn a single regression coefficient matrix $\boldsymbol{\beta} \in \mathbb{R}^{d\times m}$.

- **LASSO** $\ell_1$-norm multi-task regression with $\lambda\|\boldsymbol{\beta}\|_1$ as penalty. Each type of tasks are learned independently. It is a special case of **Sep** when pre-fixed $\hat{k} = 1$.
- **Sep L2** ridge multi-task regression with $\lambda\|\boldsymbol{\beta}\|_F^2$ as penalty. Each type of tasks are learned independently.
- **Group LASSO** $\ell_{1,2}$-norm multi-task regression with $\lambda\|\boldsymbol{\beta}\|_{1,2}$ as penalty (Yang et al. 2009), which handles heterogeneous responses, and is a special case of **Mix GS** when pre-fixed $\hat{k} = 1$.
- **TraceReg** trace-norm multi-task regression (Ji and Ye 2009).
- **Dirty** dirty model multi-task regression with $\lambda_1\|\mathbf{S}\|_1 + \lambda_2\|\mathbf{L}\|_{1,\infty}(\boldsymbol{\beta} = \mathbf{L}+\mathbf{S})$ as penalty (Jalali et al. 2010), handling entry-wise heterogeneity in $\boldsymbol{\beta}$ comparing with **Group LASSO**.
- **MSMTFL** multi-stage multi-task feature learning (Gong et al. 2012b) whose penalty is $\lambda_1 \sum_{l=1}^{d} \min(\|\boldsymbol{\beta}^l\|_1, \lambda_2)$, where $\boldsymbol{\beta}^l$ denotes the $l$th row of $\boldsymbol{\beta}$. It also handles entry-wise heterogeneity in $\boldsymbol{\beta}$ comparing with **Group LASSO**.
- **SparseTrace** multi-task regression, learning sparse and low-rank patterns with $\lambda_1\|\mathbf{S}\|_1 + \lambda_2\|\mathbf{L}\|_*(\boldsymbol{\beta} = \mathbf{L}+\mathbf{S})$ as penalty (Chen et al. 2012a), handling entry-wise heterogeneity in $\boldsymbol{\beta}$ comparing with **TraceReg**, where $\|\cdot\|_*$ denotes the nuclear norm of the enclosed matrix.
- **rMTFL** robust multi-task feature learning with $\lambda_1\|\mathbf{S}\|_{2,1} + \lambda_2\|\mathbf{L}\|_{1,2}(\boldsymbol{\beta} = \mathbf{L}+\mathbf{S})$ as penalty (Gong et al. 2012a), handling anomaly tasks comparing with **Group LASSO**.

- **RMTL** robust multi-task regression with $\lambda_1\|\mathbf{S}\|_{2,1} + \lambda_2\|\mathbf{L}\|_*(\boldsymbol{\beta} = \mathbf{L} + \mathbf{S})$ as penalty (Chen et al. 2011), handling anomaly tasks comparing with **TraceReg**.
- **CMTL** clustered multi-task learning (Zhou et al. 2011), handling clustered tasks.
- **GO-MTL** multi-task regression, handling overlapping clustered tasks (Kumar and Daumé 2012).

### 6.2 Experimental setting

In our experiments, for the E-step of GEM, we follow Städler et al. (2010) to initialize $\rho$. For the M-step, we initialize the entries of $\boldsymbol{\beta}$ from $\mathcal{N}(0, 10^{-10})$. We fix $\sigma = 1$ for Gaussian tasks, and set $\gamma = 1$. In the APG algorithm, step size is initialized by the Barzilai–Borwein rule (Barzilai and Borwein 1988) and updated by the TFOCS-style backtracking (Becker et al. 2011).

We terminate the APG algorithm with maximum iteration step $T_{in} = 200$ or when the relative $\ell_2$-norm distance of two consecutive parameters is less than $10^{-6}$. We terminate the GEM with maximum iteration step $T_{out} = 50$, or when the relative change of two consecutive $-\ell(\theta \mid \mathbf{Y}, \mathbf{X})/n$ is less than $10^{-6}$ or when the relative $\ell_\infty$-norm distance of two consecutive parameters is less than $10^{-3}$.

In the experiments on both simulated and real-world data sets, we partition the entire data set into three parts: a training set for model fitting, a validation set for tuning hyper-parameters and a testing set for testing the generalization performance of the selected models. The only exception is Sect. 6.4.1, where we do not generate testing data sets because the models are evaluated by comparing the estimation results to the ground truth.

In hyper-parameter tuning, the regularization parameters, i.e., $\lambda$s, are tuned from $[1e{-}6, 1e3]$, and the number of clusters are tuned from $\{1, \ldots, 10\}$. Hyper-parameters of the baseline methods are tuned according to the descriptions in their respective references.

All the experiments are replicated 100 times under each model setting.

### 6.3 Evaluation metrics

The prediction of latent variable is evaluated by Normalized Mutual Information (NMI) (Strehl and Ghosh 2002b; Fern and Brodley 2003; Strehl and Ghosh 2002a). In detail, we compute NMI scores by (16), treating estimated conditional probabilities $[\mathbf{P}_{\Omega,ir}]_{n\times k}$ defined in (14) and the ground truth latent variables $[\delta_{i,r}]_{n\times k}$ as $[P_{1,ir}]_{n\times k}$ and $[P_{2,ir}]_{n\times k}$, respectively.

For feature selection, firstly, the estimated components are reordered to make the best match with the true components. Then feature selection is evaluated by Area Under the ROC Curve (AUC) which is measured by the Wilcoxon–Mann–Whitney statistic provided by Hanley and McNeil (1982). Concretely, absolute values of vectorized estimated regression parameters, i.e., $\hat{\beta}_{\text{abs}} = |\text{vec}(\hat{\boldsymbol{\beta}})|$, and binarized vectorized ground truth regression parameters, i.e., $\beta_{0,\text{sign}} = \text{sign}(|\text{vec}(\boldsymbol{\beta}_0)|)$, are used as inputs to AUC, where $\boldsymbol{\beta}_0$ denotes the ground truth regression parameter and $\text{vec}(\cdot)$ is the vectorization operator.

In order to show the existence of mixed relationships between features and targets, imputation performance for incomplete targets is used to compare FMR methods with non-FMR MTL methods. Concretely, the goal is to predict one-half randomly chosen targets. The other half targets are allowed to be used. FMR methods use the other half targets to compute conditional probabilities $p(\delta_{i,r} = 1 \mid y_{ij'}, j' \in \Omega_i, \mathbf{x}_i, \hat{\theta})(i = 1, \ldots, n, r = 1, \ldots, k)$ and make prediction as stated in Sect. 3.3. Non-FMR MTL methods perform feature-based prediction.

Feature-based prediction performances are also compared between non-FMR MTL methods and our MOE methods, where only features are allowed to use to predict testing targets. For this case, the goal is to predict all the targets.

For target prediction, Gaussian outcomes are evaluated by nMSE (Chen et al. 2011; Gong et al. 2012a) which is defined as the mean of each task's mean squared error (MSE) divided by the variance of its target vector. Bernoulli outcomes are evaluated by average AUC (aAUC), which is defined as the mean AUC of each task. For Poisson tasks, we firstly compute the logarithms of outcomes, then use nMSE for evaluation.

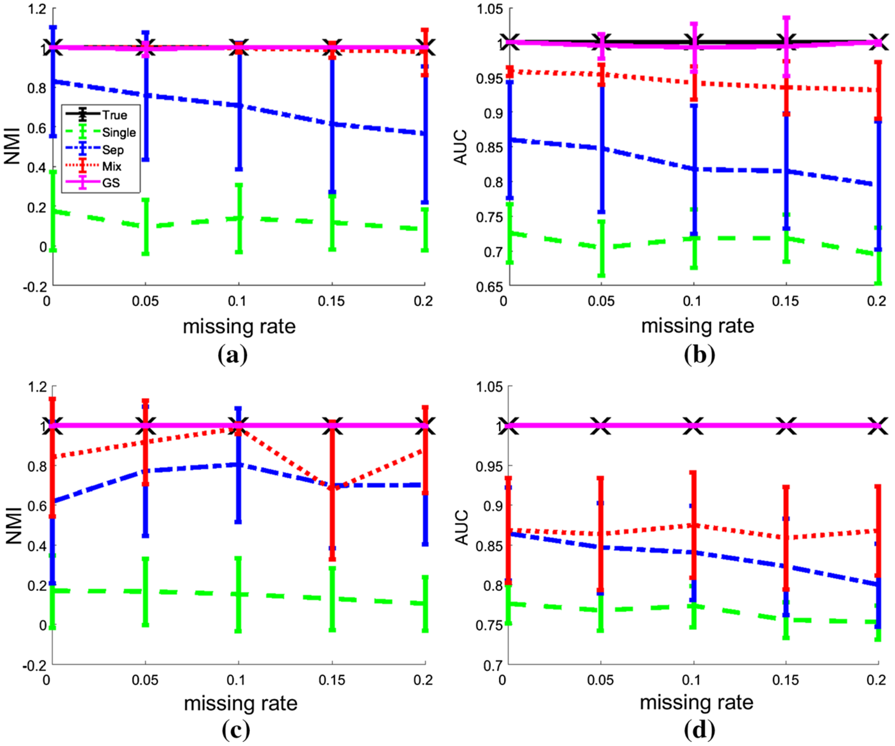
Since our objective functions in (4) and (19) are non-convex, estimated parameters may correspond to local minimums of the objective functions. Therefore, we try different initializations and report the results ranking the best 20% on the validation data set out of the 100 replications to avoid the results that may be stuck at local minimums, suggesting that one can always select any result within the best 20%.

## 6.4 Simulation

### 6.4.1 Latent variable prediction and feature selection

We consider both low dimensional case and high dimensional case for latent variable prediction and feature selection. For the low-dimensional case, we set the number of samples $n = 100$, feature dimension $d = 15$, number of non-zero features (sparsity) $s = 3$, and the number of tasks (responses) $m = 15$. The data set includes 3 Gaussian tasks, 10 Bernoulli tasks, and 2 Poisson tasks. The number of latent components $k = 2$. For $r = 1, \ldots, k$, in the $r$th component, the first row (biases) and the $(s(r-1)+2)$th to the $(sr+1)$th row (a block of $s$ rows) of the true $\boldsymbol{\beta}_r \in \mathbb{R}^{d \times m}$ are non-zero (to let different components have different sets of features). Non-zero parameters in $\boldsymbol{\beta}$ are in the range of $[-3, -1] \cup [1, 3]$ except that those of Poisson tasks are in the range of $[-0.3, -0.1] \cup [0.1, 0.3]$. The biases are all set to 1 except that those of Poisson tasks are set to 3. For Gaussian tasks, all $\sigma$s are set to 1. The entries of $\mathbf{X} \in \mathbb{R}^{d \times n}$ are drawn from $\mathcal{N}(0, 1)$ with the first dimension being 1. $\pi = (0.5, 0.5)$. Validation data is independently generated likewise and has $n$ samples. For the high-dimensional case, we set $n = 180$, $d = 320$ and $m = 20$. The data set includes 8 Gaussian tasks, 10 Bernoulli tasks, and 2 Poisson tasks. Other settings are the same as in the low-dimensional case. We set the pre-fixed $\hat{k}$ to be equal to the true $k = 2$. For targets of training data, we have tried different missing rates, which are in the range of $\{0, 0.05, 0.1, 0.15, 0.2\}$. We compare the performances of $\hat{\theta}$s estimated by **Single**, **Sep**, **Mix**, **Mix GS**, respectively, with that of $\theta_0$ (denoted by "**True**").

The results are shown in Fig. 1. The horizontal axis is the missing rates. Intuitively, larger missing rates may result in worse performances due to fewer data samples. **Single** provides poor results and is not sensitive to missing rate, because (1) data samples are deficient for single-task learning and (2) the influence of missing rate may be not significant when the number of samples is at this level. **Sep** outperforms **Single** and is affected significantly by missing rate, because (1) **Sep** uses the prior knowledge in data that multiple tasks share the same FMR structure and (2) **Sep** constructs separate FMR models such that tasks for each model are deficient, hence the advantage from joint learning multiple tasks can be easily affected when some targets are missing. Our HERMIT method **Mix** outperforms **Sep** and is robust against growing missing rate, because (1) **Mix** uses the prior knowledge in data that all the tasks share the same FMR structure and (2) **Mix** takes advantage of all the tasks, therefore, the number of tasks is then enough even some targets are missing. Our HERMIT method **Mix GS** outperforms **Mix**, even rivals the true model, and is also robust against growing missing rate, because (1) comparing with **Mix**, **Mix GS** further uses the prior knowledge in



**(a)** **(b)** **(c)** **(d)**

**Fig. 1** Latent variable prediction and feature selection performance. **a** and **b** are results on low-dimensional data; **c** and **d** are results on high-dimensional data. **a**, **c** Latent variable prediction accuracy. **b**, **d** Feature selection accuracy

data that all the tasks share the same feature space in each cluster, and (2) **Mix GS** takes advantage of all the tasks as well.

### *6.4.2 Performances when the pre-fixed $\hat{k}$ is different with the true $k$*

We consider testing the performance of target imputation when the pre-fixed $\hat{k}$ is different with the true $k$. Four data sets are generated with the true $k = 1, 2, 3, 4$, respectively. We set $n = 1000$, $d = 32$, and $m = 15$. There are 3 Gaussian tasks, 10 Bernoulli tasks, and 2 Poisson tasks. For each $k = 1, 2, 3, 4$, the sparsity $s$ is set to $\lfloor d/(2k) \rfloor$ such that the total numbers of relevant features for different data sets are the same. The values of the non-zero regression parameters for Gaussian and Bernoulli tasks in $\boldsymbol{\beta}$ are in the range of $[-6, -2] \cup [2, 6]$. We set $\pi_1 = \pi_2 = \cdots = \pi_k$. Validation and testing data are independently generated likewise and both have $n$ samples. We randomly set 20% of targets to be missing for all the training, validation and testing data. Other settings are the same as in Sect. 6.4.1. One intuitive thought is that when the pre-fixed $\hat{k}$ equals the true $k$, the imputation performance will be maximized. So we set the pre-fixed $\hat{k} \in \{1, 2, 3, 4, 5, 6, 7, 8\}$. We test **Mix** model in this experiment. Results by **Mix GS** model are similar.

In Fig. 2, the imputation performances are truly maximized when the pre-fixed $\hat{k}$ equals the true $k$. When pre-fixed $\hat{k}$ is larger than the true $k$, the imputation performances are similar. When the true $k > 1$ and when the pre-fixed $\hat{k}$ is less than the true $k$, the imputation performances grow with the pre-fixed $\hat{k}$. One may expect that when the pre-fixed $\hat{k}$ is larger than the true $k$, the performances will deteriorate, since imputation would be based on fewer data samples. We think it is because (1) the

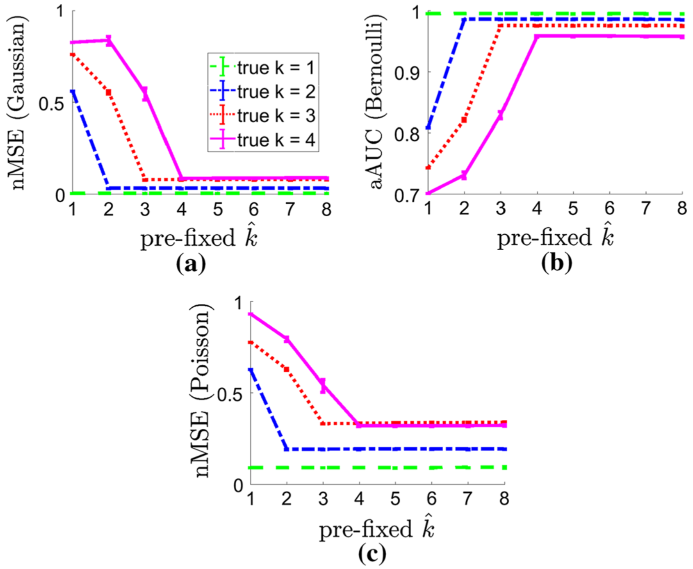


**Fig. 2** Imputation performance when the pre-fixed $\hat{k}$ is different with the true $k$. **a** nMSE of Gaussian targets. **b** aAUC of Bernoulli targets. **c** nMSE of log of Poisson targets

**Table 1** Comparison with non-FMR methods

| | nMSE | aAUC |
|---|---|---|
| LASSO | 0.6892 | 0.7384 |
| Mix | **0.1181** | 0.9525 |
| Group LASSO | 0.6850 | 0.7482 |
| Mix GS | 0.1212 | **0.9559** |
| Sep L2 | 0.6912 | 0.7355 |
| GO-MTL | 0.8055 | 0.7259 |
| CMTL | 0.6916 | 0.7344 |
| MSMTFL | 0.6890 | 0.7381 |
| TraceReg | 0.6913 | 0.7362 |
| SparseTrace | 0.6904 | 0.7374 |
| RMTL | 0.6913 | 0.7362 |
| Dirty | 0.6850 | 0.7482 |
| rMTFL | 0.6850 | 0.7482 |

Bold values indicate the best results in each setting

simulated data are simple, and (2) the information sharing among tasks renders the robustness of our Hermit method against decreasing sample size, which is consistent with the results in Sect. 6.4.1 when facing increasing missing rate (larger missing rate also indicates fewer data samples).

### *6.4.3 Comparison with non-FMR methods*

We compare the imputation performance of our Hermit methods **Mix** and **Mix GS** with all the non-FMR methods. We choose the data set used in Sect. 6.4.2 with the true $k = 3$. The Poisson targets are removed since many other methods are not able to handle them. The tuned $\hat{k} = 3$.

In Table 1, our Hermit methods **Mix** and **Mix GS** not only outperform their special cases, i.e., **LASSO** and **Group LASSO**, respectively, but also outperform other multi-task learning methods, including those handling certain kinds of heterogeneities.

### *6.4.4 Detection of anomaly tasks*

We set $n = 2000$. The number of tasks (responses) $m = 30$. The information about the true $k$s and numbers of different types of tasks is in Table 2. Other settings are the same as in Sect. 6.4.2. In Table 2, it can be seen that the true $k$ of the majority of tasks (the first 20 tasks) is 4. The first 20 tasks are referred to as concordant tasks, while the other 10 tasks are referred to as anomaly tasks.

We compute the concordant scores using (15) for the tasks. In Fig. 3, the concordant scores separate concordant tasks and anomaly tasks quite well. Scores of Poisson tasks are similar to scores of Bernoulli tasks, because they all provide less accurate information than Gaussian tasks do.

**Table 2** True *k*s and numbers of different types of tasks

| Group | True k | #Gaussian | #Bernoulli | #Poisson |
|---|---|---|---|---|
| 1 | 4 | 5 | 10 | 5 |
| 2 | 1 | 1 | 1 | 1 |
| 3 | 6 | 1 | 0 | 0 |
| 4 | 2 | 1 | 1 | 0 |
| 5 | 3 | 0 | 1 | 1 |
| 6 | 5 | 1 | 1 | 0 |

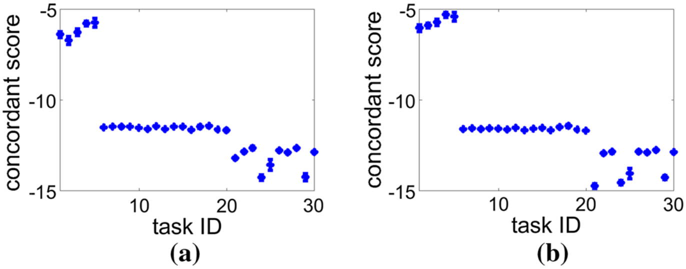


**Fig. 3** Concordant scores of tasks, which are associated with Table 2. **a** Estimated by **Mix**; **b** estimated by **Mix GS**. The first 20 tasks are concordant tasks, the last 10 tasks are anomaly tasks. The first 5 tasks are Gaussian tasks, the subsequent 10 tasks are Bernoulli tasks and then the subsequent 5 tasks are Poisson tasks

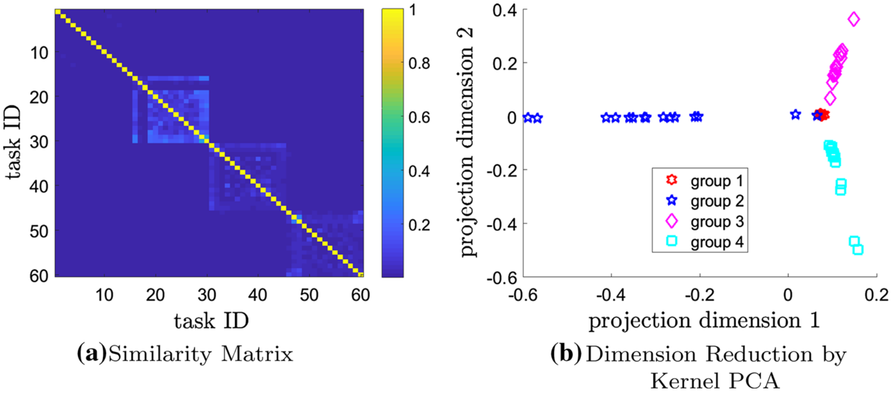
### *6.4.5 Handling clustered relationship among tasks*

We construct 4 groups of tasks. The total number of tasks (responses) $m = 60$. The information about the true $k$s and numbers of different types of tasks is in Table 3. Other settings are the same as in Sect. 6.4.4. We first apply **Single** for each task, setting $\hat{k} = 20$. Then we apply the strategy in Sect. 4.3 to construct a similarity matrix by NMI defined in (16). Kernel PCA (Schölkopf et al. 1998; Van Der Maaten et al. 2009) is then applied using the similarity matrix as the kernel matrix. The similarity matrix and the result of Kernel PCA are shown in Fig. 4.

In Fig. 4a, Group 2,3 and 4 can be recognized as three groups. In Group 1, each task shows no similarity with other tasks, because with the true $k = 1$, the data samples can be randomly partitioned into $\hat{k} = 20$ sub-populations, which results in low NMI scores. In Fig. 4b, basically, 4 groups of tasks are clustered into 4 different regions.

**Table 3** True *k*s and numbers of different types of tasks. Tasks are clustered into 4 groups

| Group | True k | #Gaussian | #Bernoulli | #Poisson |
|---|---|---|---|---|
| 1 | 1 | 3 | 10 | 2 |
| 2 | 2 | 3 | 10 | 2 |
| 3 | 3 | 3 | 10 | 2 |
| 4 | 4 | 3 | 10 | 2 |



**(a)** Similarity Matrix

**(b)** Dimension Reduction by Kernel PCA

**Fig. 4** **a** Similarity matrix among tasks described in Table 3; **b** relationship among tasks shown by Kernel PCA

### 6.4.6 *Handling outlier samples*

We choose the data set used in Sect. 6.4.2 with the true $k = 2$, then randomly shuffle the data pairs $(\mathbf{y}_i, \mathbf{x}_i)$, for $i = 1, \ldots, n$, and contaminate the true targets by the following procedure. For outlier ratio $p_{\text{outlier}} = 0, 1, 2, 5, 8, 10\%$, (1) for Gaussian targets, set all the targets of $p_{\text{outlier}}$ of data samples to be 100; (2) for Bernoulli targets, set all the targets of $p_{\text{outlier}}$ of data samples to be 1. Such contamination is only performed on training and validation data, leaving testing data clean.

Then we evaluate two groups of methods. For the group of non-robust methods, we choose our HERMIT methods **Mix** and **Mix GS**. For the group of robust methods, we firstly run the robust version of the non-robust methods by adding $\boldsymbol{\zeta}$ in the natural parameter models as (11) and adding (13) as the additional penalty, then we clean the data by removing $p_{\text{outlier}}$ of data samples associated with the largest value of $\sqrt{\sum_{jr} \zeta_{ijr}^2}$ $(i \in \{1, \ldots, n\})$. Finally, we run their non-robust version of methods on the "cleaned" data, respectively. We follow Gong et al. (2012a) to adopt such two-stage strategy.

The imputation performances are reported in Table 4, from where it can be seen that, (1) when $p_{\text{outlier}} = 0\%$, robust methods are over-parameterized and may underperform non-robust methods; (2) when $p_{\text{outlier}} > 0\%$, robust methods significantly outperform non-robust methods.

### 6.4.7 *Feature-based prediction by MOE*

We set the true $k = 3$. The true $\boldsymbol{\alpha} \in \mathbb{R}^{d \times k}$, whose first four rows are non-zero. The non-zero entries of $\boldsymbol{\alpha}$ are drawn from $\mathcal{N}(0, 1)$. Number of data samples $n = 1000$. For all $i = 1, \ldots, n, r = 1, \ldots, k$, the $i$th data sample coming from the $r$th sub-population obeys a multinomial distribution with the probability defined in (17). Other settings are the same as in Sect. 6.4.3.

**Table 4** Comparison between methods handling outlier samples on synthetic data

| | 0% | 1% | 2% | 5% | 8% | 10% |
|---|---|---|---|---|---|---|
| *nMSE for Gaussian* | | | | | | |
| **Mix** | | | | | | |
| Non-robust | 0.0625 | 0.6754 | 0.6894 | 1.0122 | 1.3250 | 1.4953 |
| Robust | **0.0620** | **0.0627** | **0.0626** | **0.0737** | **0.0632** | **0.0635** |
| **Mix GS** | | | | | | |
| Non-robust | 0.0658 | 0.6434 | 0.6741 | 0.7505 | 1.0736 | 1.2939 |
| Robust | **0.0599** | **0.0611** | **0.0673** | **0.0694** | **0.0602** | **0.0607** |
| *aAUC* | | | | | | |
| **Mix** | | | | | | |
| Non-robust | **0.9571** | 0.7954 | 0.7961 | 0.7982 | 0.7986 | 0.7981 |
| Robust | 0.9570 | **0.9571** | **0.9574** | **0.9519** | **0.9568** | **0.9567** |
| **Mix GS** | | | | | | |
| Non-robust | 0.9509 | 0.7979 | 0.7984 | 0.7982 | 0.7979 | 0.7952 |
| Robust | **0.9581** | **0.9577** | **0.9519** | **0.9482** | **0.9578** | **0.9574** |
| *nMSE for Poisson* | | | | | | |
| **Mix** | | | | | | |
| Non-robust | 0.2089 | 0.7368 | 0.6905 | 0.6528 | 0.6642 | 0.6736 |
| Robust | **0.2086** | **0.2105** | **0.2099** | **0.2345** | **0.2222** | **0.2230** |
| **Mix GS** | | | | | | |
| Non-robust | 0.2136 | 0.7416 | 0.7795 | 0.6587 | 0.6688 | 0.8665 |
| Robust | **0.2087** | **0.2109** | **0.2169** | **0.2202** | **0.2212** | **0.2236** |

Bold values indicate the best results in each setting

We compare our HERMIT methods **Mix MOE** and **Mix MOE GS**. The prediction performances are shown in Table 5, which are consistent with the results in Sect. 6.4.3.

We further show in Table 6 the concordance between $p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \hat{\alpha}_r)$ and $p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \alpha_{0,r})$, where $\boldsymbol{\alpha}_0$ denotes the true $\boldsymbol{\alpha}$, for all $i = 1, \ldots, n$, $r = 1, \ldots, k$, for both training and testing data. In (22), $\boldsymbol{\alpha}$ is optimized by partially minimizing the discrepancy between $p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \hat{\alpha}_r)$ and $\hat{\rho}_{i,r}^{(t+1)}$ for $t = 0, \ldots, T-1$. As such we also show the concordance between $p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \hat{\alpha}_r)$ and $\hat{\rho}_{i,r}^{(T)} = p(\delta_{i,r} = 1 \mid y_{ij'}, j' \in \Omega_i, \mathbf{x}_i, \hat{\theta}_2)$. The concordances are measured by NMI defined in (16). We use NMI instead of KL-divergence, because NMI is normalized to the range of [0, 1].

Both $p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \alpha_{0,r})$ and $p(\delta_{i,r} = 1 \mid y_{ij'}, j' \in \Omega_i, \mathbf{x}_i, \hat{\theta}_2)$ are approximated accurately on the training data. The approximation accuracies are lower on the testing data because the deficiency of data samples comparing with the dimension.

### 6.4.8 Scalability

We discuss the scalability of our method for increasing number of features and tasks. The running time is evaluated. We choose the data set used in Sect. 6.4.2 with the true $k = 4$. The sparsity $s$ is fixed to be 4. We set the number of features $d \in \{32, 64, 128, 256, 512\}$ and the number of tasks $m \in \{15, 30, 60, 120, 240\}$. The ratios

**Table 5** Prediction performances based on only features

| | nMSE | aAUC |
|---|---|---|
| LASSO | 0.6390 | 0.7834 |
| Mix MOE | 0.0656 | 0.9466 |
| Group LASSO | 0.6348 | 0.7878 |
| Mix MOE GS | **0.0579** | **0.9502** |
| Sep L2 | 0.6481 | 0.7794 |
| GO-MTL | 0.6946 | 0.7778 |
| CMTL | 0.6496 | 0.7796 |
| MSMTFL | 0.6397 | 0.7831 |
| TraceReg | 0.6509 | 0.7790 |
| SparseTrace | 0.6473 | 0.7805 |
| RMTL | 0.6511 | 0.7797 |
| Dirty | 0.6348 | 0.7878 |
| rMTFL | 0.6483 | 0.7787 |

Bold values indicate the best results in each setting

**Table 6** Approximation performances based on only features

| | Training | | Testing | |
|---|---|---|---|---|
| | $C(\hat{\boldsymbol{\alpha}} \parallel \boldsymbol{\alpha}_0)$ | $C(\hat{\boldsymbol{\alpha}} \parallel \hat{\theta}_2)$ | $C(\hat{\boldsymbol{\alpha}} \parallel \boldsymbol{\alpha}_0)$ | $C(\hat{\boldsymbol{\alpha}} \parallel \hat{\theta}_2)$ |
| Mix MOE | 0.9863 | 0.9918 | 0.8440 | 0.8457 |
| Mix MOE GS | **0.9962** | **0.9933** | **0.8455** | **0.8516** |

$C(\hat{\boldsymbol{\alpha}} \parallel \boldsymbol{\alpha}_0)$ denotes the concordance between $p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \hat{\alpha}_r)$ and $p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \alpha_{0,r})$, where $\boldsymbol{\alpha}_0$ denotes the true $\boldsymbol{\alpha}$. $C(\hat{\boldsymbol{\alpha}} \parallel \hat{\theta}_2)$ denotes the concordance between $p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \hat{\alpha}_r)$ and $p(\delta_{i,r} = 1 \mid y_{ij'}, j' \in \Omega_i, \mathbf{x}_i, \hat{\theta}_2)$. The concordances are measured by NMI defined in (16)
Bold values indicate the best results in each setting

between the numbers of Gaussian, Bernoulli and Poisson tasks are the same as in Sect. 6.4.2. We randomly generate 100 data sets for each pair of $(d, m)$. We report the results of the method **Mix GS** only, as the results of **Mix** are similar. In each case, $\hat{k}$ is tuned and is equal to the true $k = 4$. The estimated parameters in different cases may have different numbers of relevant features. As such, in order to provide a fair comparison, we report the running time per feature, i.e., running time divided by the number of non-zero features of the estimated parameters in each case.

In Fig. 5, both dimension $d$ and the number of tasks $m$ have no significant influence on running time per feature, especially when $d$ and $m$ is large, which is consistent with our time-complexity analysis in Sect. 3.2.

### 6.5 Application

On the real-world data sets, we first demonstrate the existence of the heterogeneity of conditional relationship, then report the superiority of our HERMIT method over other methods considered in Sect. 6.1. We further interpret the advantage of our method

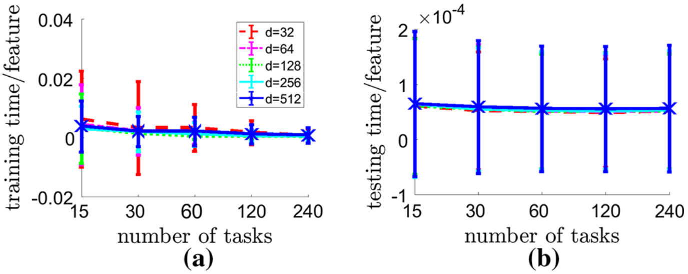


**Fig. 5** Running time per feature when dimension and the number of tasks grow. **a** Reports the running time per feature for training the model; **b** reports the running time per feature for testing. **a** Training time. **b** Testing time

by presenting the selected features. Effectiveness of anomaly-task detection and task clustering strategy is also validated.

### *6.5.1 Data description*

Both real-world data sets introduced in the following are longitudinal surveys for elder patients, which includes a set of questions. Some of the question answers are treated as input features and some of the questions related to indices of geriatric assessments are treated as targets.

*LSOA II Data* This data is from the Second Longitudinal Study of Aging (LSOA II).[1] LSOA II is a collaborative study by the National Center for Health Statistics (NCHS) and the National Institute of Aging conducted from 1994–2000. A national representative sample of 9447 subjects 70 years of age and over were selected and interviewed. Three separated interviews were conducted during the periods of 1994–1996, 1997–1998, and 1999–2000, respectively. The interviews are referred to as WAVE 1, WAVE 2, and WAVE 3 interviews, respectively. We use data WAVE 2 and WAVE 3, which includes a total of 4299 sample subjects and 44 targets, and 188 features are extracted from WAVE 2 interview.

Among the targets, specifically, three self-rated health measures, including overall health status, memory status and depression status, can be regarded as continuous outcomes; there are 41 binary outcomes, which fall into several categories: fundamental daily activity, extended daily activity, social involvement, medical condition, on cognitive ability, and sensation condition. The features include records of demographics, family structure, daily personal care, medical history, social activity, health opinion, behavior, nutrition, health insurance and income and assets, the majority of which are binary measurements. Both targets and features have missing values due to non-response and questionnaire filtering. The average missing value rates in targets and features are 13.7 and 20.2%, respectively. For the missing values in features, we adopt the following procedure for pre-processing. For continuous features, the missing values are imputed with sample mean. For binary features, a better approach is to treat missing as a third category as it may also carry important information; as such, two

[1] https://www.cdc.gov/nchs/lsoa/lsoa2.htm.

dummy variables are created from each binary feature with missing values (the third one is not necessary.) This results in totally $d = 293$ features. We randomly select 30% of the samples for training, 30% for validation and the rest for testing.

*easySHARE Data* This data is a simplified data set from the Survey of Heath, Aging, and Retirement in Europe (SHARE).[2] SHARE includes multidisciplinary and cross-national panel databases on health, socio-economic status, and social and family networks of more than 85,000 individuals from 20 European countries aged 50 or over. Four waves of interviews were conducted during 2004–2011, and are referred to as WAVE 1 to WAVE 4 interviews. We use WAVE 1 and WAVE 2, which includes 20,449 sample persons and 15 targets (among which 11 are binary, and 4 are continuous), and totally 75 features are constructed from WAVE 1 interview.

The targets are from four interview modules: social support, mental health, functional limitation indices and cognitive function indices. The features cover a wide range of assessments, including demographics, household composition, social support and network, physical health, mental health, behavior risk, healthcare, occupation and income. Detailed description features are not listed in this paper. Both targets and features have missing values due to non-response and questionnaire filtering. The average missing value rates in targets and features are 6.9 and 5.1%, respectively. The same pre-processing procedure as that for LSOA II Data has been adopted and results in totally $d = 118$ features. We randomly select 10% of the samples for training, 10% for validation and the rest for testing.

### 6.5.2 Comparison with FMR method

In this experiment, we compare our proposed HERMIT methods **Mix** and **Mix GS** which handle mixed type of outcomes with **Sep** which learns different types of tasks separately. **Single** is abandoned because it learns each task independently and is not able to use targets of other tasks to help increasing imputation performance.

Results are reported in Fig. 6, where (1) for both the real data sets, basically, the best pre-fixed $\hat{k} > 1$, except for Bernoulli tasks of easySHARE data, suggesting that the heterogeneity of conditional relationship exists in LSOA II data and the Gaussian tasks of easySHARE data; (2) FMR models benefit Gaussian targets more than Bernoulli targets; (3) **Mix** and **Mix GS** outperform **Sep** in Gaussian tasks. However, their performances are comparable with **Sep** in Bernoulli tasks, which may be because that the number of Bernoulli tasks are much more than that of Gaussian tasks such that the benefit from Gaussian tasks is limited.

### 6.5.3 Comparison with non-FMR methods

In this experiment we test imputation performance, comparing our HERMIT methods **Mix** and **Mix GS** with all the non-FMR methods.

Results are reported in Table 7, where (1) our HERMIT methods **Mix** and **Mix GS** not only outperform their special cases, **LASSO** and **Group LASSO**, respectively, but also outperform other methods, including those handling certain kinds of heterogeneities,

[2] http://www.share-project.org/data-access-documentation.html.

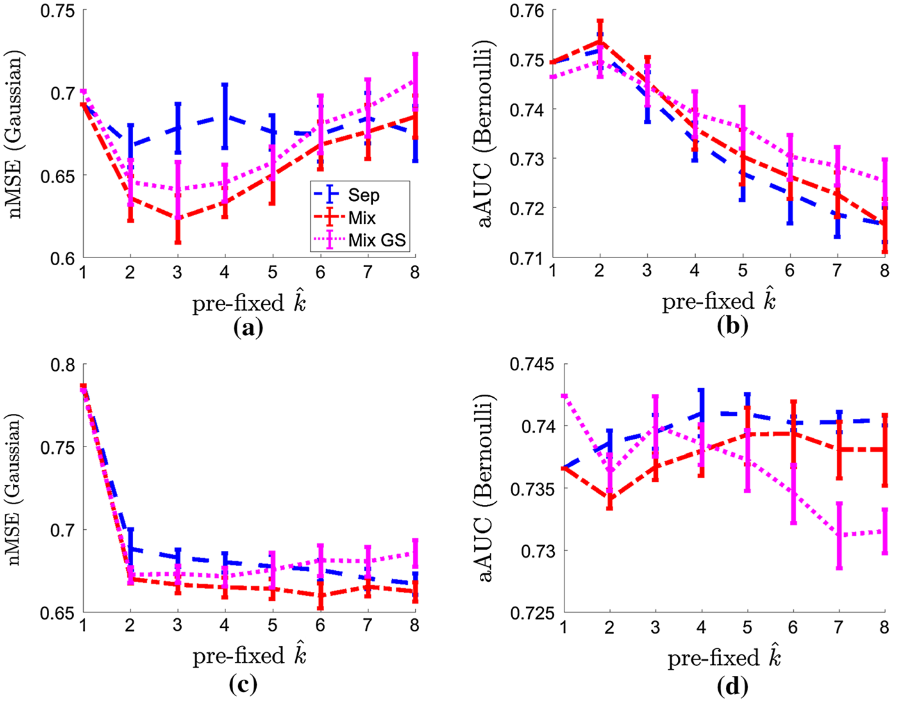


**Fig. 6** Comparison with **Sep** on real data sets. **a** and **b** are results on LSOA II data, **c** and **d** are results on easySHARE data. **a**, **c** nMSE of Gaussian targets. **b**, **d** aAUC of Bernoulli targets

except for aAUC on easySHARE, the reason of which has been discussed in Sect. 6.5.2; (2) **Mix GS** increases nMSE by 9.76 and 14.37% on LSOA II data and easySHARE data, respectively, comparing with its non-FMR version **Group LASSO**. The similar improvements by **Mix** are witnessed as well.

### 6.5.4 *Feature selection*

We consider demonstrating the advantage of our HERMIT method on feature selection. We compare our HERMIT method **Mix GS** with its non-FMR version **Group LASSO**. Both methods select shared features across tasks. We collect the unique features that only selected for each sub-population.

For LSOA II data set, the tuned $\hat{k} = 2$. **Mix GS** selects $47/294$ features (summing up selected features of both sub-populations), while **Group LASSO** selects $48/294$ features. Descriptions of unique features of both sub-populations are listed in Table 8. Sub-population 1 seems considering worse condition of patients.

For easySHARE data set, the tuned $\hat{k} = 5$. **Mix GS** selects $58/118$ features, while **Group LASSO** selects $57/118$ features. Descriptions of unique features of two sub-populations are listed in Table 9. Sub-population 1 seems considering more about personality and experience, while sub-population 2 seems considering more about politics and education.

**Table 7** Comparison with non-FMR methods on real data sets

| | LSOA II | | easySHARE | |
|---|---|---|---|---|
| | nMSE | aAUC | nMSE | aAUC |
| LASSO | 0.7051 | 0.7474 | 0.7869 | 0.7386 |
| Mix | 0.6408 | **0.7525** | 0.6601 | 0.7419 |
| Group LASSO | 0.6975 | 0.7413 | 0.7897 | 0.7413 |
| Mix GS | **0.6294** | 0.7481 | **0.6548** | 0.7402 |
| Sep L2 | 0.7176 | 0.7392 | 0.7796 | 0.7464 |
| GO-MTL | 0.8516 | 0.6972 | 0.8231 | 0.7288 |
| CMTL | 0.8186 | 0.7089 | 0.7958 | 0.7364 |
| MSMTFL | 0.7028 | 0.7473 | 0.7803 | 0.7411 |
| TraceReg | 0.7150 | 0.7408 | 0.7809 | **0.7496** |
| SparseTrace | 0.6972 | 0.7475 | 0.7791 | 0.7475 |
| RMTL | 0.7145 | 0.7418 | 0.7808 | 0.7496 |
| Dirty | 0.7032 | 0.7480 | 0.7781 | 0.7486 |
| rMTFL | 0.6953 | 0.7418 | 0.7781 | 0.7486 |

Bold values indicate the best results in each setting

**Table 8** Descriptions of unique features of each sub-population of LSOA II data

| Sub-population 1 ($\pi_1 = 70.09\%$) | Sub-population 2 ($\pi_2 = 29.91\%$) |
|---|---|
| Able or prevented to leave house | Times seen doctor in past 3 months |
| Have problems with balance | Easier or harder to walk 1/4 mile |
| Total number of living children | **Widowed** |
| Easier/harder than before: in/out of bed | **Follow regular physical routine** |
| **#(ADL activities) SP is unable to perform** | **Present social activities** |
| Easier or harder to walk 10 steps | **Ever had a stress test** |
| Do you take aspirin | **Do you take vitamins** |
| **Often troubled with pain** | **Necessary to use steps or stairs** |
| Visit homebound friend for others | **Had flu shot** |
| Ever had a hysterectomy | **Ever had cataract surgery** |
| | **Physical activity more/less/same** |

Features are sorted in descending order by the $\|\boldsymbol{\beta}_r^l\|_2$ where $\boldsymbol{\beta}_r^l$ is the $l$th row of $\boldsymbol{\beta}_r$ $(l = 1, \ldots, d)$. The bold denote the features that are not selected by **Group LASSO**

#(·) number of the enclosed events, *ADL* Activity of Daily Livings. *SP* Standardized Patients

For both real data sets, our HERMIT method **Mix GS** recalls more useful features than **Group LASSO** does.

### 6.5.5 Detection of anomaly tasks

We firstly use (15) to compute concordant scores of tasks, which are reported in Fig. 7. Clear separations are witnessed on both data sets. We select one-third of tasks with highest scores as concordant tasks and another third with lowest scores as anomaly

**Table 9** Descriptions of unique features of two sub-populations of easySHARE data

| Sub-population 1 ($\pi_1 = 21.31\%$) | Sub-population 2 ($\pi_2 = 26.36\%$) |
|---|---|
| **Fatigue** | **Taken part in a political organization** |
| Guilt | **Attended an educational or training course** |
| Enjoyment | **Taken part in religious organization** |
| Suicidality | **None of social activities** |
| Tearfullness | **Cared for a sick or disabled adult** |
| **Interest** | **Done voluntary or charity work** |
| **Current job situation:sick** | **Education: lower secondary** |
| | **Education: first tertiary** |
| | **Education: post secondary** |
| | **Education: upper secondary** |
| | **Education: primary** |
| | **Education: second tertiary** |

Features are sorted in descending order by the $\|\boldsymbol{\beta}_r^l\|_2$ where $\boldsymbol{\beta}_r^l$ is the $l$th row of $\boldsymbol{\beta}_r$ ($l = 1, \ldots, d$). The bold denote the features that are not selected by **Group LASSO**

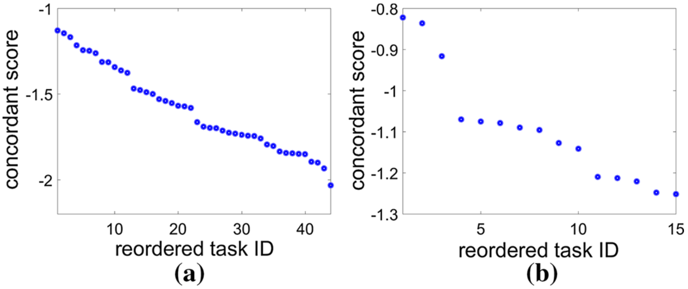


**Fig. 7** Concordant scores of tasks, which were estimated by **Mix GS**. The tasks are reordered according to the scores. **a** LSOA II data set. **b** easySHARE data set

tasks. The descriptions of the concordant and anomaly tasks are listed in Tables 10 and 11, respectively.

The concordant tasks detected by our methods seem truly correlated with each other intuitively. And the information of detected anomaly tasks is diverse and seems different with that of concordant tasks.

For each data set, we apply our HERMIT method **Mix** (and **Mix GS**) to build two models for non-anomaly tasks (the first two-third tasks) and anomaly tasks, respectively. For LSOA II data set, the tuned $\hat{k} = 4$ and 1 for non-anomaly tasks and anomaly tasks, respectively. For easySHARE data set, the tuned $\hat{k} = 6$ and 2 for non-anomaly tasks and anomaly tasks, respectively.

Averaged imputation performances are shown in Table 12. By providing separate models to handle anomaly tasks, the performances improve significantly, where **Mix GS** outperforms **Mix**, maybe because the non-anomaly tasks share some relevant features.

**Table 10** Descriptions of tasks of LSOA II data

| Concordant tasks (top 7) | Anomaly tasks (top 8) |
|---|---|
| Have difficulty dressing | Go to movies, sports, events, etc. |
| Have difficulty doing light housewrk | Now have asthma |
| Have difficulty using toilet | Now have arthritis |
| Have difficulty managing medication | Now have hypertension |
| Have difficulty bathing or showering | Injured from fall(s) |
| Have difficulty managing money | Memory of year |
| Have difficulty preparing meals | Have deafness |
| | Get together with relatives |

**Table 11** Descriptions of tasks of easySHARE data

| Concordant tasks (top 5) | Anomaly tasks (top 5) |
|---|---|
| Activities of daily living index | Numeracy score |
| Instrumental activities of daily living indices | Gone to sport social or other kind of club |
| Mobility index | Recall of words first trial |
| Appetite | Give help to others outside the household |
| Orientation to date | Provided help to family friends or neighbors |

**Table 12** Comparison for imputation performances

| | LSOA II | | easySHARE | |
|---|---|---|---|---|
| | nMSE | aAUC | nMSE | aAUC |
| Mix—All tasks | 0.6408 | 0.7525 | 0.6601 | 0.7419 |
| Mix—Handle anomalies | 0.5979 | 0.7602 | 0.6569 | 0.7370 |
| Mix GS—All tasks | 0.6294 | 0.7481 | 0.6548 | 0.7402 |
| Mix GS—Handle anomalies | **0.5923** | **0.7649** | **0.6462** | **0.7447** |

"All tasks" denotes building one FMR model for all the tasks. "Handle anomalies" denotes building two models for non-anomaly tasks and anomaly tasks, respectively

Bold values indicate the best results in each setting

### *6.5.6 Handling clustered relationship among tasks*

We adopt the same strategy as that in Sect. 6.4.5 to construct a similarity matrix and perform dimension reduction for each of the real-world data sets.

For LSOA II data set, the similarity matrix and results of 2D reduction are shown in Fig. 8. In Fig. 8b, tasks are partitioned into groups. We apply k-means algorithm to separate the tasks into 4 groups. Tasks in Group 1 are mainly about current status. The descriptions of tasks of Group 2 are "how often felt sad or depressed in the past 12 months" and "self rated memory". Tasks in Group 3 and 4 are about having difficulty

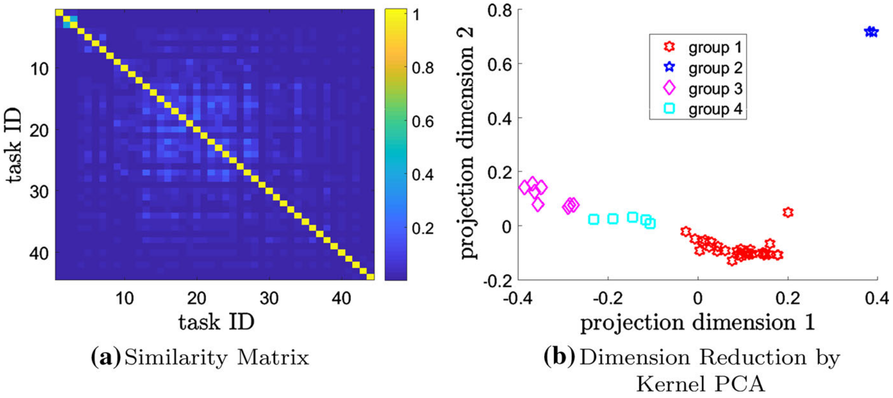


**Fig. 8** Clustered Relationship among tasks on LSOA II. **a** Similarity matrix among tasks. First three tasks are Gaussian tasks. Other tasks are Bernoulli tasks. **b** Relationship among tasks shown by Kernel PCA

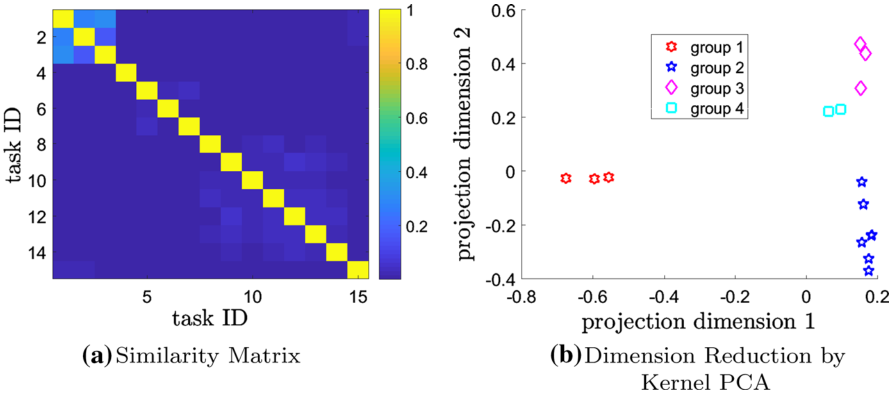


**Fig. 9** Clustered relationship among tasks on easySHARE. **a** Similarity matrix among tasks. First four tasks are Gaussian tasks. Other tasks are Bernoulli tasks. **b** Relationship among tasks shown by Kernel PCA

performing some certain actions. Group 3 is similar to Group 4, which can be reflected by Fig. 8b.

For easySHARE data set, the similarity matrix and results of 2D reduction are shown in Fig. 9. In Fig. 9b, tasks are partitioned into groups as well. We also apply k-means algorithm to separate the tasks into 4 groups. The descriptions of tasks for each group are shown in Table 13, where descriptions of 4 types of interview modules are basically separated into 4 groups, respectively. The only “misclassified” task with the description of “Orientation to date” seems to be more related to other tasks in Group 2 than the tasks in Group 4.

For each data set, we further apply our Hermit methods **Mix** and **Mix GS** for each group of tasks. For LSOA II data set, tuned $\hat{k} = 3, 3, 5$ and 2 for Group 1,2,3 and 4, respectively. For easySHARE data set, tuned $\hat{k} = 5, 2, 1$ and 1 for Group 1, 2, 3 and 4, respectively. Imputation performances are shown in Table 14. Performances increase by building separate models for each group, suggesting that separate models for clustered tasks are more accurate.

**Table 13** Clustered tasks of easySHARE

| Group | Targets | Interview module |
|---|---|---|
| 1 | Activities of daily living index | Functional limitation indices |
| | Instrumental activities of daily living index | Functional limitation indices |
| | Mobility index | Functional limitation indices |
| 2 | Depression | Mental health |
| | Pessimism | Mental health |
| | Sleep | Mental health |
| | Irritability | Mental health |
| | Appetite | Mental health |
| | Concentration | Mental health |
| | Orientation to date | Cognitive function indices |
| 3 | Provided help to family friends or neighbors | Social support and network |
| | Gone to sport social or other kind of club | Social support and network |
| | Give help to others outside the household | Social support and network |
| 4 | Recall of words score | Cognitive function indices |
| | Numeracy score | Cognitive function indices |

15 tasks are clustered into 4 groups

**Table 14** Comparison for imputation performances

| | LSOA II | | easySHARE | |
|---|---|---|---|---|
| | nMSE | aAUC | nMSE | aAUC |
| Mix—All tasks | 0.6408 | 0.7525 | 0.6601 | 0.7419 |
| Mix—Clustered tasks | 0.6370 | **0.7592** | 0.6552 | 0.7439 |
| Mix GS—All tasks | 0.6294 | 0.7481 | 0.6548 | 0.7402 |
| Mix GS—Clustered tasks | **0.6202** | 0.7559 | **0.6533** | **0.7474** |

"All tasks" denotes building one FMR model for all the tasks. "Clustered tasks" denotes building different FMR models for different groups of tasks
Bold values indicate the best results in each setting

### *6.5.7 Feature-based prediction by MOE*

We compare the methods using only features to predict targets on both real-world data sets. Our proposed MOE type of HERMIT methods, **Mix MOE** and **Mix MOE GS**, are compared with the non-FMR methods. We also integrate our strategies to handle anomaly tasks and clustered structure among tasks in both our proposed MOE type of HERMIT methods **Mix MOE** and **Mix MOE GS**. Concretely, we use the anomaly-task detection results in Sect. 6.5.5 and the task clustering results in Sect. 6.5.6.

The prediction results are reported in Table 15. Our proposed HERMIT method **Mix MOE** and **Mix MOE GS** outperform baseline methods on LSOA II and on Gaussian tasks of easySHARE, which is consistent with the results in Table 7. In addition, by integrating our task clustering strategy, our proposed HERMIT methods **Mix MOE TC**

**Table 15** Comparison for prediction performance with non-FMR methods on real data sets

| | LSOA II | | easySHARE | |
|---|---|---|---|---|
| | nMSE | aAUC | nMSE | aAUC |
| LASSO | 0.7051 | 0.7474 | 0.7869 | 0.7386 |
| Group LASSO | 0.6975 | 0.7413 | 0.7897 | 0.7413 |
| MSMTFL | 0.7028 | 0.7473 | 0.7803 | 0.7411 |
| Sep L2 | 0.7176 | 0.7392 | 0.7796 | 0.7464 |
| GO-MTL | 0.8516 | 0.6972 | 0.8231 | 0.7288 |
| CMTL | 0.8186 | 0.7089 | 0.7958 | 0.7364 |
| TraceReg | 0.7150 | 0.7408 | 0.7809 | **0.7496** |
| SparseTrace | 0.6972 | 0.7475 | 0.7791 | 0.7475 |
| RMTL | 0.7145 | 0.7418 | 0.7808 | 0.7496 |
| Dirty | 0.7032 | 0.7480 | 0.7781 | 0.7486 |
| rMTFL | 0.6953 | 0.7418 | 0.7781 | 0.7486 |
| Mix MOE | 0.6935 | **0.7504** | 0.7991 | 0.7395 |
| Mix MOE GS | 0.7054 | 0.7438 | 0.7774 | 0.7387 |
| Mix MOE Robust | 0.6906 | 0.7436 | 0.7642 | 0.7351 |
| Mix MOE GS Robust | 0.6981 | 0.7430 | 0.7668 | 0.7344 |
| Mix MOE TC | **0.6859** | 0.7333 | **0.7584** | 0.7389 |
| Mix MOE GS TC | 0.6925 | 0.7379 | 0.7657 | 0.7367 |

“Robust” denotes adopting the strategy to handle anomaly tasks. “TC” denotes task clustering strategy

Bold values indicate the best results in each setting

**Table 16** The concordance between $p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \hat{\alpha}_r)$ and $p(\delta_{i,r} = 1 \mid y_{ij'}, j' \in \Omega_i, \mathbf{x}_i, \hat{\theta}_2)$

| | LSOA II | | easySHARE | |
|---|---|---|---|---|
| | Training | Testing | Training | Testing |
| Mix MOE | 0.2745 | 0.1301 | 0.1314 | 0.1068 |
| Mix MOE GS | 0.1060 | 0.0673 | 0.2527 | 0.2054 |

The concordances are measured by NMI defined in (16)

and **Mix MOE GS TC** outperform other methods on Gaussian targets, while providing comparable results on Bernoulli tasks. **Mix MOE TC** and **Mix MOE GS TC** even outperform our proposed HERMIT methods **Mix MOE Robust** and **Mix MOE GS Robust** on Gaussian targets, suggesting that it is more accurate to build a specific model for each cluster of tasks.

Comparing Table 15 with Table 7, our MOE methods do not rival our FMR methods. We investigate the reason by showing the concordance between $p(\delta_{i,r} = 1 \mid \mathbf{x}_i, \hat{\alpha}_r)$ and $p(\delta_{i,r} = 1 \mid y_{ij'}, j' \in \Omega_i, \mathbf{x}_i, \hat{\theta}_2) = \hat{\rho}_{i,r}^{(T)}$ ($\hat{\rho}_{i,r}^{(T)}$ is defined in equation 21) in Table 16. In Table 16, the concordances of conditional probabilities measured by NMI are generally low, especially comparing with the results in Table 6, suggesting that on both real-world data sets, it is difficult to learn the mixture probabilities.

## 7 Discussions and conclusions

In this paper, we propose a novel model HERMIT to explore heterogeneities of conditional relationship, output type and shared information among tasks. Based on multivariate-target FMR and MOE models, our model jointly learns tasks with mixed type of output, allows incomplete data in the output, imposes inner component-wise group $\ell_1$ constraint and handles anomaly tasks and clustered structure among tasks. These key elements are integrated in a unified generalized mixture model setup so that they can benefit from and reinforce each other to discover the triple heterogeneities in data. Rigorous theoretical analyses under the high dimensional framework are provided.

We mainly consider the special setting of MTL, where the multivariate outcomes share the same set of instances and the same set of features because our main objective is to learn potentially shared sample clusters and feature sets among tasks. However, as stressed in the introduction, the main definition of MTL considers tasks that do not necessarily share the same set of samples/instances and the same set of features, such as distributed learning systems (different tasks have entirely different data instances, see Jin et al. 2006 and Boyd et al. 2011) and multi-source learning systems (different tasks have entirely different feature spaces, see Zhang and Yeung 2011 and Jin et al. 2015). For such cases, one can define the specific expected shared information among tasks and then extend our methodology. For example, although tasks do not share the same instances, they could share the same mixture model structure. Then for the distributed learning systems, our model in Sect. 3 can still be applied. Additionally, the tasks could still share the pattern/sparsity in feature selection even though the feature sets are different, e.g., Liu et al. (2009) and Gong et al. (2012b). Then one can build FMR models for the tasks in which the regression coefficient vectors of the tasks share the same sparsity pattern achieved by group $\ell_1$ penalization. The case of multi-source learning systems can also be handled similarly by embedding features into a shared feature space, e.g., Zhang and Yeung (2011) and Jin et al. (2015).

There are many interesting future directions. It is worthwhile to explore the theoretical and empirical performance of non-convex penalties. Meanwhile, different components should share some features, and overlapping cluster pattern of conditional relationship should also be considered in real applications, both of which require further investigation. It is also interesting to explore other low-dimensional structures in the natural parameters, e.g., low-rank structure and its sparse composition (Chen et al. 2012b). Our strategies on handling anomaly tasks and clustered structure among tasks require two stages. It is worthwhile to explore one-stage models to handle such task heterogeneities during a whole learning process. More complicated structure among tasks, such as graph-based structure, should also be explored. Our theoretical results cover our method introduced in Sect. 3 and robust estimation introduced in Sect. 4.1. Nonetheless, theoretical guarantees for other extensions in Sect. 4 are still challenging due to joint learning complicated relationship among tasks and population heterogeneity, which will be focused on in our future research.

**Acknowledgements** The authors would like to thank the editors and reviewers for their valuable suggestions on improving this paper. This work of Jian Liang and Changshui Zhang is (jointly or partly) funded by National Natural Science Foundation of China under Grant No. 61473167 and Beijing Natural Science Foundation under Grant No. L172037. Kun Chen's work is partially supported by U.S. National Science Foundation under Grants DMS-1613295 and IIS-1718798. The work of Fei Wang is supported by National Science Foundation under Grants IIS-1650723 and IIS-1716432.

## Appendix A: Definitions

**Definition 1** $Z = (Z_1, \dots, Z_{m'})^{\mathrm{T}} \in \mathbb{R}^{m'}$ has a sub-exponential distribution with parameters $(\sigma, v, t)$ if for $M > t$, it holds

$$\mathbb{P}(\|Z\|_\infty > M) \le \begin{cases} \exp\left(-\frac{M^2}{\sigma^2}\right), & t \le M \le \frac{\sigma^2}{v} \\ \exp\left(-\frac{M}{v}\right), & M > \frac{\sigma^2}{v}. \end{cases}$$

## Appendix B: The empirical process

In order to prove the first part of Theorem 1 that the bound in (26) has the probability in (25), we firstly follow Städler et al. (2010) to define the empirical process for fixed data points $\mathbf{x}_1, \dots, \mathbf{x}_n$. For $\tilde{\mathbf{y}}_i = (y_{ij}, j \in \Omega_i)^{\mathrm{T}} \in \mathbb{R}^{|\Omega_i|}$ and $X = (X_1, \dots, X_d)$, let

$$V_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} (\ell_\theta(\mathbf{x}_i, \tilde{\mathbf{y}}_i) - \mathbb{E}[\ell_\theta(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \mid X = \mathbf{x}_i]) .$$

By fixing some $T \ge 1$ and $\lambda_0 \ge 0$, we define an event $\mathcal{T}$ below, upon which the bound in (26) can be proved. So the probability of the event $\mathcal{T}$ is the probability in (25).

$$\mathcal{T} = \left\{ \sup_{\theta \in \tilde{\Theta}} \frac{|V_n(\theta) - V_n(\theta_0)|}{(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 + \|\eta - \eta_0\|_2) \vee \lambda_0} \le T\lambda_0 \right\}. \tag{21}$$

It can be seen that, (21) defines a set of the parameter $\theta$, and the bound in (26) will be proved with $\hat{\theta}$ in the set.

For group-lasso type estimator, define an event similar to that in (21) in the following.

$$\mathcal{T}_{group} = \left\{ \sup_{\theta \in \tilde{\Theta}} \frac{|V_n(\theta) - V_n(\theta_0)|}{\left(\sum_p \|\boldsymbol{\beta}_{\mathcal{G}_p} - \boldsymbol{\beta}_{0,\mathcal{G}_p}\|_2 + \|\eta - \eta_0\|_2\right) \vee \lambda_0} \le T\lambda_0 \right\}. \tag{22}$$

## Appendix C: Lemmas

In order to show that the probability of event $\mathcal{T}$ is large, we firstly invoke the following lemma.

**Lemma 2** *Under Condition 2, for model (1) with $\theta_0 \in \tilde{\Theta}$, $M_n$ and $\lambda_0$ defined in (24), some constants $c_6, c_7$ depending on $K$, and for $n \geq c_7$, we have*

$$\mathbb{P}_{\mathbf{X}}\left(\frac{1}{n}\sum_{i=1}^{n} F(\tilde{\mathbf{y}}_i) > c_6\lambda_0^2/(mk)\right) \leq \frac{1}{n},$$

*where $\mathbb{P}_{\mathbf{X}}$ denote the conditional probability given $(X_1^{\mathrm{T}}, \ldots, X_n^{\mathrm{T}})^{\mathrm{T}} = (\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}} = \mathbf{X}$, and $F(\tilde{\mathbf{y}}_i) = G_1(\tilde{\mathbf{y}}_i)1\{G_1(\tilde{\mathbf{y}}_i) > M_n\} + \mathbb{E}[G_1(\tilde{\mathbf{y}}_i)1\{G_1(\tilde{\mathbf{y}}_i) > M_n\} \mid X = \mathbf{x}_i], \forall i$.*

A proof is given in "Appendix F" section.

Then we can follow the Corollary 1 in Städler et al. (2010) to show that the probability of event $\mathcal{T}$ is large below.

**Lemma 3** *Use Lemma 2. For model (1) with $\theta_0 \in \tilde{\Theta}$, some constants $c_7, c_8, c_9, c_{10}$ depending on $K$, for $\mathcal{T}$ is defined in (21), and for all $T \geq c_{10}$ we have*

$$\mathbb{P}_{\mathbf{X}}(\mathcal{T}) \geq 1 - c_9 \exp\left(-\frac{T^2(\log n)^2 \log(d \vee n)}{c_8}\right) - \frac{1}{n}, \forall n \geq c_7.$$

A proof is given in "Appendix G" section.

## Appendix D: Corollaries for models considering outlier samples

When considering outlier samples and modifying the natural parameter model as in (11), we can show in this section the similar results.

First, as $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ are treated in the similar way, we denote them together by $\boldsymbol{\xi} \in \mathbb{R}^{((d+n)\times m)\times k}$, and $\xi = vec(\boldsymbol{\xi}) \in \mathbb{R}^{(d+n)mk}$ such that for all $r = 1, \ldots, k$,

$$\begin{aligned} \boldsymbol{\varphi}_r &= \mathbf{X}\boldsymbol{\beta}_r + \boldsymbol{\zeta}_r \;\Rightarrow \boldsymbol{\varphi}_r = \mathbf{A}\boldsymbol{\xi}_r, \\ \mathbf{A} &= [\mathbf{X}, \mathbf{I}_n] \in \mathbb{R}^{n\times(d+n)},\; \boldsymbol{\xi}_r = [\boldsymbol{\beta}_r^{\mathrm{T}}, \boldsymbol{\zeta}_r^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{(d+n)\times m}, \end{aligned}$$

where $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is a identity matrix.

Thus it can be seen that the modification only results in new design matrix and regression coefficient matrix, therefore, we can apply Theorems 1–3 to have similar results for the modified models.

For lasso-type penalties, denote the set of indices of non-zero entries of $\beta_0$ by $S_\beta$, and the set of indices of non-zero entries of $\zeta_0$ by $S_\zeta$, where $\zeta = \mathrm{vec}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_k)$. Denote by $s = |S_\beta| + |S_\zeta|$. Then for entry-wise $\ell_1$ penalties in (5) (for $\boldsymbol{\beta}$) with $\gamma = 0$ and $\mathcal{R}(\boldsymbol{\zeta}) = \lambda\|\zeta\|_1$ (for $\boldsymbol{\zeta}$), we need the following modified restricted eigenvalue condition.

**Condition 6** *For all $\beta \in \mathbb{R}^{dmk}$ and all $\zeta \in \mathbb{R}^{nmk}$ satisfying $\|\beta_{S_\beta^c}\|_1 + \|\zeta_{S_\zeta^c}\|_1 \leq 6(\|\beta_{S_\beta}\|_1 + \|\zeta_{S_\zeta}\|_1)$, it holds for some constant $\kappa \geq 1$ that,*

$$\|\beta_{S_\beta}\|_2^2 + \|\zeta_{S_\zeta}\|_2^2 \leq \kappa^2\|\varphi\|_{Q_n}^2 = \frac{\kappa^2}{n}\sum_{i=1}^{n}\sum_{j\in\Omega_i}\sum_{r=1}^{k}(\mathbf{x}_i\boldsymbol{\beta}_{jr} + \zeta_{ijr})^2.$$

**Corollary 1** *Consider the* HERMIT *model in* (1) *with* $\theta_0 \in \tilde{\Theta}$, *and consider the penalized estimator* (12) *with the* $\ell_1$ *penalties in* (5) *and* $\mathcal{R}(\boldsymbol{\zeta}) = \lambda\|\zeta\|_1$.

*(a) Assume Conditions* 1–3 *and* 6 *hold. Suppose* $\sqrt{mk} \lesssim n/M_n$, *and take* $\lambda > 2T\lambda_0$ *for some constant* $T > 1$. *For some constant* $c > 0$ *and large enough* $n$, *with probability* $1 - c\exp\left(-\frac{(\log n)^2 \log(d \vee n)}{c}\right) - \frac{1}{n}$, *we have*

$$\bar{\varepsilon}(\hat{\theta} \mid \theta_0) + 2(\lambda - T\lambda_0)\left(\|\hat{\beta}_{S_\beta^c}\|_1 + \|\hat{\zeta}_{S_\zeta^c}\|_1\right) \le 4(\lambda + T\lambda_0)^2\kappa^2 c_0^2 s,$$

*(b) Assume Conditions* 1–3 *hold (without Condition* 6*), assume*

$$\|\beta_0\|_1 + \|\zeta_0\|_1 = o\left(\sqrt{n/((\log n)^{2+2c_1}\log(d \vee n)mk)}\right),$$
$$\sqrt{mk} = o\left(\sqrt{n/((\log n)^{2+2c_1}\log(d \vee n))}\right)$$

*as* $n \to \infty$. *If* $\lambda = C\sqrt{(\log n)^{2+2c_1}\log(d \vee n)mk/n}$ *for some* $C > 0$ *sufficiently large, and for some constant* $c > 0$ *and large enough* $n$, *with the following probability* $1 - c\exp\left(-\frac{(\log n)^2 \log(d \vee n)}{c}\right) - \frac{1}{n}$, *we have* $\bar{\varepsilon}(\hat{\theta} \mid \theta_0) = o_P(1)$.

For group-lasso type penalties, denote

$$\mathcal{I}_\beta = \{p : \boldsymbol{\beta}_{0,\mathcal{G}_{\beta,p}} = \mathbf{0}\},\ \mathcal{I}_\beta^c = \{p : \boldsymbol{\beta}_{0,\mathcal{G}_{\beta,p}} \neq \mathbf{0}\},$$
$$\mathcal{I}_\zeta = \{q : \boldsymbol{\zeta}_{0,\mathcal{G}_{\zeta,q}} = \mathbf{0}\},\ \mathcal{I}_\zeta^c = \{q : \boldsymbol{\zeta}_{0,\mathcal{G}_{\zeta,q}} \neq \mathbf{0}\},$$

where $\boldsymbol{\beta}_{0,\mathcal{G}_{\beta,p}}$ and $\boldsymbol{\zeta}_{0,\mathcal{G}_{\zeta,q}}$ denote the $p$th group of $\boldsymbol{\beta}_0$ and the $q$th group of $\boldsymbol{\zeta}_0$, respectively. Now denote $s = |\mathcal{I}_\beta| + |\mathcal{I}_\zeta|$ with some abuse of notation.

Then for group $\ell_1$ penalties in (27) (for $\boldsymbol{\beta}$) and $\mathcal{R}(\boldsymbol{\zeta}) = \sum_q^Q \|\boldsymbol{\zeta}_{\mathcal{G}_{\zeta,q}}\|_F$ (for $\boldsymbol{\zeta}$), we need the following modified restricted eigenvalue condition.

**Condition 7** *For all* $\boldsymbol{\beta} \in \mathbb{R}^{d\times mk}$ *and all* $\boldsymbol{\zeta} \in \mathbb{R}^{n\times mk}$ *satisfying*

$$\sum_{p\in\mathcal{I}_\beta^c} \|\boldsymbol{\beta}_{\mathcal{G}_{\beta,p}}\|_F + \sum_{q\in\mathcal{I}_\zeta^c} \|\boldsymbol{\zeta}_{\mathcal{G}_{\zeta,q}}\|_F \le 6\left(\sum_{p\in\mathcal{I}_\beta} \|\boldsymbol{\beta}_{\mathcal{G}_{\beta,p}}\|_F + \sum_{q\in\mathcal{I}_\zeta} \|\boldsymbol{\zeta}_{\mathcal{G}_{\zeta,q}}\|_F\right),$$

*it holds that for some constant* $\kappa \ge 1$,

$$\sum_{p\in\mathcal{I}_\beta} \|\boldsymbol{\beta}_{\mathcal{G}_{\beta,p}}\|_F^2 + \sum_{q\in\mathcal{I}_\zeta} \|\boldsymbol{\zeta}_{\mathcal{G}_{\zeta,q}}\|_F^2 \le \kappa^2\|\varphi\|_{Q_n}^2 = \frac{\kappa^2}{n}\sum_{i=1}^n \sum_{j\in\Omega_i}\sum_{r=1}^k (\mathbf{x}_i\boldsymbol{\beta}_{jr} + \zeta_{ijr})^2.$$

**Corollary 2** *Consider the* HERMIT *model in* (1) *with* $\theta_0 \in \tilde{\Theta}$, *and consider estimator* (12) *with the group* $\ell_1$ *penalties in* (27) *and* $\mathcal{R}(\boldsymbol{\zeta}) = \sum_q^Q \|\boldsymbol{\zeta}_{\mathcal{G}_{\zeta,q}}\|_F$.

(a) *Assume Conditions 1–3 and 7 hold. Suppose* $\sqrt{mk} \lesssim n/M_n$, *and take* $\lambda > 2T\lambda_0$ *for some constant* $T > 1$. *For some constant* $c > 0$ *and large enough n, with probability* $1 - c\exp\left(-\frac{(\log n)^2 \log(d \vee n)}{c}\right) - \frac{1}{n}$, *we have*

$$\bar{\varepsilon}(\hat{\theta} \mid \theta_0) + 2(\lambda - T\lambda_0)\left(\sum_{p \in \mathcal{I}_\beta^c} \|\hat{\boldsymbol{\beta}}_{\mathcal{G}_{\beta,p}}\|_F + \sum_{q \in \mathcal{I}_\zeta^c} \|\hat{\boldsymbol{\zeta}}_{\mathcal{G}_{\zeta,q}}\|_F\right) \le 4(\lambda + T\lambda_0)^2 \kappa^2 c_0^2 s,$$

(b) *Assume Conditions 1–3 hold (without Condition 7), assume*

$$\sum_{p=1}^{P} \|\boldsymbol{\beta}_{0,\mathcal{G}_{\beta,p}}\|_F + \sum_{q=1}^{Q} \|\boldsymbol{\zeta}_{0,\mathcal{G}_{\zeta,q}}\|_F = o\left(\sqrt{n/((\log n)^{2+2c_1} \log(d \vee n) mk)}\right),$$
$$\sqrt{mk} = o\left(\sqrt{n/((\log n)^{2+2c_1} \log(d \vee n))}\right)$$

*as* $n \to \infty$. *If* $\lambda = C\sqrt{(\log n)^{2+2c_1} \log(d \vee n) mk/n}$ *for some* $C > 0$ *sufficiently large, and for some constant* $c > 0$ *and large enough n, with the following probability* $1 - c\exp\left(-\frac{(\log n)^2 \log(d \vee n)}{c}\right) - \frac{1}{n}$, *we have* $\bar{\varepsilon}(\hat{\theta} \mid \theta_0) = o_P(1)$.

## Appendix E: Proof of Lemma 1

*Proof* For non-negative continuous variable $X$, we have

$$\begin{aligned}
\mathbb{E}[X1\{X > M\}] &= \int_M^\infty t f_X(t)dt = \int_M^\infty \int_0^t f_X(t)dxdt \\
&= \int_0^M \int_M^\infty f_X(t)dtdx + \int_M^\infty \int_x^\infty f_X(t)dtdx \\
&= M\mathbb{P}(X > M) + \int_M^\infty \mathbb{P}(X > x)dx.
\end{aligned}$$

Similarly, we have $\mathbb{E}[X^2 1\{X > M\}] = M^2\mathbb{P}(X > M) + \int_M^\infty 2x\mathbb{P}(X > x)dx$. For $X$ sub-exponential with parameters $(\sigma, v, t)$ such that for $M > t$

$$\mathbb{P}(X > M) \le \begin{cases} \exp\left(-\frac{M^2}{\sigma^2}\right), \ t \le M \le \frac{\sigma^2}{v} \\ \exp\left(-\frac{M}{v}\right), \quad M \ge \frac{\sigma^2}{v}, \end{cases}$$

we have the following.

If $M \le \frac{\sigma^2}{v}$, we have

$$\mathbb{E}[X1\{X > M\}] = M\mathbb{P}(X > M) + \int_M^\infty \mathbb{P}(X > x)dx$$

$$
\begin{aligned}
&\leq M\exp\biggl(-\frac{M^2}{\sigma^2}\biggr)+\int_M^{\frac{\sigma^2}{v}}\exp\biggl(-\frac{x^2}{\sigma^2}\biggr)dx+\int_{\frac{\sigma^2}{v}}^{\infty}\exp\biggl(-\frac{x}{v}\biggr)dx\\
&\leq M\exp\biggl(-\frac{M^2}{\sigma^2}\biggr)+\biggl(\frac{\sigma^2}{v}-M\biggr)\exp\biggl(-\frac{M^2}{\sigma^2}\biggr)+v\exp\biggl(-\frac{M}{v}\biggr)\\
&= M\exp\biggl(-\frac{M^2}{\sigma^2}\biggr)+v\exp\biggl(-\frac{M}{v}\biggr)\leq (M+v)\exp\biggl(-\frac{M^2}{\sigma^2}\biggr),
\end{aligned}
$$

and similarly, $\mathbb{E}[X^2 1\{X > M\}] \leq \Bigl(M^2+2v^2+2\sigma^2\Bigr)\exp\Bigl(-\frac{M^2}{\sigma^2}\Bigr)$.

If $M > \frac{\sigma^2}{v}$, we have $\mathbb{E}[X1\{X > M\}] \leq (M+v)\exp\Bigl(-\frac{M}{v}\Bigr)$ and $\mathbb{E}[X^2 1\{X > M\}] \leq (M^2+2v^2+2vM)\exp\Bigl(-\frac{M}{v}\Bigr)$.

Then for some constants $c_1, c_2, c_3, c_4, c_5 > 0$, for non-negative continuous variable $X$ which is sub-exponential with parameters $(\sigma, v, t)$, for $M > c_4 > t$ and $c' = 2+\frac{3}{c_1}$, we have

$$
\begin{aligned}
\mathbb{E}[X1\{X > M\}] &\leq \biggl[c_3\biggl(\frac{M}{c_2}\biggr)^{c'}+c_5\biggr]\exp\biggl\{-\biggl(\frac{M}{c_2}\biggr)^{1/c_1}\biggr\},\\
\mathbb{E}[X^2 1\{X > M\}] &\leq \biggl[c_3\biggl(\frac{M}{c_2}\biggr)^{c'}+c_5\biggr]^2\exp\biggl\{-2\biggl(\frac{M}{c_2}\biggr)^{1/c_1}\biggr\}.
\end{aligned}
$$

If $t \leq M \leq \frac{\sigma^2}{v}$, $c_1 = 1/2$, $c_2 = \sqrt{2}\sigma$, $c_3 = 16\sigma^8$. And if $M \geq \frac{\sigma^2}{v}$, $c_1 = 1$, $c_2 = 2v$, $c_3 = 32v^5$. And $c_5 = \sqrt{2}(v+\sigma)$.

For non-negative discrete variables, the result is the same.

The result of Lemma 1 follows from the result above, $\tilde{\mathbf{y}}_i$ has a finite mixture distribution for $i = 1, \ldots, n$ and the following.

When dispersion parameter $\phi$ is known, for a constant $c_K$ depending on $K$, we have

$$
G_1(\tilde{\mathbf{y}}_i) = e^K \max_{j\in\Omega_i}|y_{ij}| + c_K,\ \ i = 1, \ldots, n.
$$

□

## Appendix F: Proof of Lemma 2

*Proof* Under Condition 2, $M_n = c_2(\log n)^{c_1}$, and $\lambda_0$ defined in (24), for a constant $c_6$ depending on $K$, for $i = 1, \ldots, n$, we have

$$
\begin{aligned}
&\mathbb{E}[|G_1(\tilde{\mathbf{y}}_i)|1\{|G_1(\tilde{\mathbf{y}}_i)| > M_n\}] \leq c_6\lambda_0^2/(mk),\\
&\mathbb{E}[|G_1(\tilde{\mathbf{y}}_i)|^2 1\{|G_1(\tilde{\mathbf{y}}_i)| > M_n\}] \leq c_6^2\lambda_0^4/(mk)^2.
\end{aligned}
$$

The we can get

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}}\left(\frac{1}{n}\sum_{i=1}^{n}G_1(\tilde{\mathbf{y}}_i)1\{G_1(\tilde{\mathbf{y}}_i)>M_n\}+\mathbb{E}[G_1(\tilde{\mathbf{y}}_i)1\{G_1(\tilde{\mathbf{y}}_i)>M_n\}]>3c_6\lambda_0^2/(mk)\right)\\
&\quad\le\mathbb{P}_{\mathbf{X}}\left(\frac{1}{n}\sum_{i=1}^{n}G_1(\tilde{\mathbf{y}}_i)1\{G_1(\tilde{\mathbf{y}}_i)>M_n\}-\mathbb{E}[G_1(\tilde{\mathbf{y}}_i)1\{G_1(\tilde{\mathbf{y}}_i)>M_n\}]>c_6\lambda_0^2/(mk)\right)\\
&\quad\le\frac{\mathbb{E}[|G_1(\tilde{\mathbf{y}}_i)|^2 1\{|G_1(\tilde{\mathbf{y}}_i)|>M_n\}]}{n}\frac{m^2k^2}{c_6^2\lambda_0^4}\le\frac{1}{n}.
\end{aligned}$$

□

## Appendix G: Proof of Lemma 3

*Proof* We follow Städler et al. (2010) to give a Entropy Lemma and then prove Lemma 3.

We use the following norm $\|\cdot\|_{P_n}$ introduced in the Proof of Lemma 2 in Städler et al. (2010) and use $H(\cdot,\mathcal{H},\|\cdot\|_{P_n})$ as the entropy of covering number [see Van de Geer (2000)] which is equipped the metric induced by the norm for a collection $\mathcal{H}$ of functions on $\mathcal{X}\times\mathcal{Y}$,

$$\|h(\cdot,\cdot)\|_{P_n}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}h^2(\mathbf{x}_i,\tilde{\mathbf{y}}_i)}.$$

Define $\tilde{\Theta}(\epsilon)=\{\theta\in\tilde{\Theta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1+\|\eta-\eta_0\|_2\le\epsilon\}$.

**Lemma 4** *(Entropy Lemma) For a constant $c_{12}>0$, for all $u>0$ and $M_n>0$, we have*

$$\begin{aligned}
&H\left(u,\left\{(\ell_\theta-\ell_{\theta^\star})1\{G_1\le M_n\}:\theta\in\tilde{\Theta}(\epsilon)\right\},\|\cdot\|_{P_n}\right)\\
&\quad\le c_{12}\frac{mk\epsilon^2M_n^2}{u^2}\log\left(\frac{\sqrt{mk}\epsilon M_n}{u}\right).
\end{aligned}$$

*Proof* (For Entropy Lemma) The difference between this proof and that of Entropy Lemma in the proof of Lemma 2 of Städler et al. (2010) is in the notations and the effect of multivariate responses.

For multivariate responses we have for $i=1,\ldots,n$,

$$\begin{aligned}
|\ell_\theta(\mathbf{x}_i,\tilde{\mathbf{y}}_i)-\ell_{\theta'}(\mathbf{x}_i,\tilde{\mathbf{y}}_i)|^2&\le G_1^2(\tilde{\mathbf{y}}_i)\|\psi_i-\psi_i'\|_1^2\le d_\psi G_1^2(\tilde{\mathbf{y}}_i)\|\psi_i-\psi_i'\|_2^2\\
&=d_\psi G_1^2(\tilde{\mathbf{y}}_i)\left[\sum_{r=1}^{k}\sum_{j\in\Omega_i}|\mathbf{x}_i(\boldsymbol{\beta}_{rj}-\boldsymbol{\beta}'_{rj})|^2+\|\eta-\eta'\|_2^2\right],
\end{aligned}$$

where $d_\psi = (2m+1)k$ is the maximum of dimension of $\psi_i$ for $i = 1, \ldots, n$.

Under the definition of the norm $\|\cdot\|_{P_n}$ we have

$$\begin{aligned}&\|(\ell_\theta - \ell_{\theta'})1\{G_1 \le M_n\}\|_{P_n}^2\\ &\quad \le d_\psi M_n^2 \left[\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{k}\sum_{j\in\Omega_i} |\mathbf{x}_i(\boldsymbol{\beta}_{rj} - \boldsymbol{\beta}'_{rj})|^2 + \|\eta - \eta'\|_2^2\right].\end{aligned}$$

Then by the result of Städler et al. (2010) we have

$$H(u, \{\eta \in \mathbb{R}^{d_\eta} : \|\eta - \eta_0\|_2 \le \epsilon\}, \|\cdot\|_2) \le d_\eta \log\left(\frac{5\epsilon}{u}\right),$$

where $d_\eta = (m+1)k$ is the dimension of $\eta$.

And we follow Städler et al. (2010) to apply Lemma 2.6.11 of Van Der Vaart and Wellner (1996) to give a bound as

$$\begin{aligned}&H\left(2u, \left\{\sum_{r=1}^{k}\sum_{j\in\Omega_i} \mathbf{x}_i(\boldsymbol{\beta}_{rj} - \boldsymbol{\beta}_{0,rj}) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \le \epsilon\right\}, \|\cdot\|_{P_n}\right)\\ &\quad \le \left(\frac{\epsilon^2}{u^2} + 1\right)\log(1 + kmd).\end{aligned}$$

Thus we can get

$$\begin{aligned}&H\left(3\sqrt{d_\psi} M_n u, \left\{(\ell_\theta - \ell_{\theta_0})1\{G_1 \le M_n\} : \theta \in \tilde{\Theta}(\epsilon)\right\}, \|\cdot\|_{P_n}\right)\\ &\quad \le \left(\frac{\epsilon^2}{u^2} + 1 + d_\eta\right)\left(\log(1 + kmd) + \log\left(\frac{5\epsilon}{u}\right)\right).\end{aligned}$$

□

Now we turn to prove Lemma 3.

We follow Städler et al. (2010) to use the truncated version of the empirical process below.

$$\begin{aligned}&V_n^{trunc}(\theta)\\ &\quad = \frac{1}{n}\sum_{i=1}^{n}\Big(\ell_\theta(\mathbf{x}_i, \tilde{\mathbf{y}}_i)1\{G_1(\tilde{\mathbf{y}}_i) \le M_n\} - \mathbb{E}[\ell_\theta(\mathbf{x}_i, \tilde{\mathbf{y}}_i)1\{G_1(\tilde{\mathbf{y}}_i) \le M_n\} \mid X = \mathbf{x}_i].\Big)\end{aligned}$$

We follow Städler et al. (2010) to apply the Lemma 3.2 in Van de Geer (2000) and a conditional version of Lemma 3.3 in Van de Geer (2000) to the class

$$\left\{ (\ell_\theta - \ell_{\theta_0}) 1\{G_1 \le M_n\} : \theta \in \tilde{\Theta}(\epsilon) \right\}, \forall \epsilon > 0.$$

For some constants $\{c_t\}_{t>12}$ depending on $K$ and $\Lambda_{\max}$ in Condition 2 of Städler et al. (2010), using the notation of Lemma 3.2 in Van de Geer (2000), we follow Städler et al. (2010) to choose $\delta = c_{13} T \epsilon \lambda_0$ and $R = c_{14}(\sqrt{mk}\epsilon \wedge 1) M_n$.

Thus we by choosing $M_n = c_2 (\log n)^{c_1}$ we can satisfy the condition of Lemma 3.2 of Van de Geer (2000) to have

$$\begin{aligned}
&\int_{\epsilon/c_{15}}^{R} H^{1/2}\left(u, \left\{ (\ell_\theta - \ell_{\theta^\star}) 1\{G_1 \le M_n\} : \theta \in \tilde{\Theta}(\epsilon) \right\}, \|\cdot\|_{P_n}\right) du \vee R \\
&= \int_{\epsilon/c_{15}}^{c_{14}\sqrt{mk}(\epsilon \wedge 1) M_n} c_{12} \left( \frac{\sqrt{mk}\epsilon M_n}{u} \right) \log^{1/2} \left( \frac{\sqrt{mk}\epsilon M_n}{u} \right) du \vee (c_{14}(\epsilon \wedge 1) M_n) \\
&\le \frac{2}{3} c_{12} \sqrt{mk}\epsilon M_n \left[ \log^{3/2}(c_{15}\sqrt{mk} M_n) - \log^{3/2} \left( \frac{\sqrt{mk}\epsilon M_n}{c_{14}\sqrt{mk}(\epsilon \wedge 1) M_n} \right) \right] \\
&\quad \vee (c_{14}\sqrt{mk}(\epsilon \wedge 1) M_n) \\
&\le \frac{2}{3} c_{12} \sqrt{mk}\epsilon M_n \log^{3/2}(c_{15}\sqrt{mk} M_n) \\
&\le c_{16}\sqrt{mk}\epsilon M_n \log^{3/2}(n) \quad \left( \text{by choosing } M_n = c_2 (\log n)^{c_1}, \text{ and } \sqrt{mk} \le c_{17} \frac{n}{M_n} \right) \\
&\le c_{18} \sqrt{n} T \epsilon \lambda_0 \le \sqrt{n}(\delta - \epsilon).
\end{aligned}$$

Now for the rest we can apply Lemma 3.2 of Van de Geer (2000) to give the same result with Lemma 2 of Städler et al. (2010).

So we have

$$\sup_{\theta \in \tilde{\Theta}} \frac{|V_n^{trunc}(\theta) - V_n^{trunc}(\theta_0)|}{(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 + \|\eta - \eta_0\|_2) \vee \lambda_0} \le 2 c_{23} T \lambda_0$$

with probability at least $1 - c_9 \exp\left[ -\frac{T^2 (\log n)^2 \log(d \vee n)}{c_8^2} \right]$.

At last, for the case when $G_1(\tilde{\mathbf{y}}_i) > M_n$, for $i = 1, \ldots, n$, we have

$$|(\ell_\theta(\mathbf{x}_i, \tilde{\mathbf{y}}_i) - \ell_{\theta_0}(\mathbf{x}_i, \tilde{\mathbf{y}}_i)) 1\{G_1(\tilde{\mathbf{y}}_i) > M_n\}| \le d_\psi K G_1(\tilde{\mathbf{y}}_i) 1\{G_1(\tilde{\mathbf{y}}_i) > M_n\},$$

and

$$\begin{aligned}
&\frac{|(V_n^{trunc}(\theta) - V_n^{trunc}(\theta_0)) - (V_n(\theta) - V_n(\theta_0))|}{(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 + \|\eta - \eta_0\|_2) \vee \lambda_0} \\
&\quad \le \frac{d_\psi K}{n \lambda_0} \sum_{i=1}^{n} \left( G_1(\tilde{\mathbf{y}}_i) 1\{G_1(\tilde{\mathbf{y}}_i) > M_n\} + \mathbb{E}[G_1(\tilde{\mathbf{y}}_i) 1\{G_1(\tilde{\mathbf{y}}_i) > M_n\} \mid X = \mathbf{x}_i] \right).
\end{aligned}$$

Then the probability of the following inequality under our model is given in Lemma 2.

$$\frac{d_\psi K}{n\lambda_0}\sum_{i=1}^{n}\Big(G_1(\tilde{\mathbf{y}}_i)1\{G_1(\tilde{\mathbf{y}}_i) > M_n\} + \mathbb{E}[G_1(\tilde{\mathbf{y}}_i)1\{G_1(\tilde{\mathbf{y}}_i) > M_n\} \mid X = \mathbf{x}_i]\Big)$$
$$\leq c_{23}T\lambda_0,$$

where $d_\psi = 2(m+1)k$. □

## Appendix H: Proof of Theorem 1

*Proof* This proof mostly follows that of Theorem 3 of Städler et al. (2010). The only difference is in the notations. As such, we omit the details. □

## Appendix I: Proof of Theorem 2

*Proof* This proof also mostly follows that of Theorem 5 of Städler et al. (2010). The difference is in the notations and the choice of $M_n$.

If the event $\mathcal{T}$ happens, with $M_n = c_2(\log n)^{c_1}$ for some constants $0 \leq c_1, c_2 < \infty$, where $c_2$ depends on $K$,

$$\lambda_0 = \sqrt{mk}M_n \log n\sqrt{\log(d \vee n)/n} = c_2\sqrt{mk\log^{2+2c_1}\log(d \vee n)/n},$$

we have

$$\begin{aligned}\bar{\varepsilon}(\hat{\psi} \mid \psi_0) + \lambda\|\hat{\beta}\|_1 \leq T\lambda_0[(\|\hat{\beta} - \beta_0\|_1 + \|\eta - \eta_0\|_2) \vee \lambda_0]\\ + \lambda\|\beta_0\|_1 + \bar{\varepsilon}(\psi_0 \mid \psi_0).\end{aligned}$$

Under the definition of $\theta \in \tilde{\Theta}$ in (23) we have $\|\eta - \eta_0\|_2 \leq 2K$. And as $\bar{\varepsilon}(\psi_0 \mid \psi_0) = 0$ we have for $n$ sufficiently large.

$$\begin{aligned}&\bar{\varepsilon}(\hat{\psi} \mid \psi_0) + \lambda\|\hat{\beta}\|_1 \leq T\lambda_0(\|\hat{\beta}\|_1 + \|\beta_0\|_1 + 2K) + \lambda\|\beta_0\|_1\\ &\to \bar{\varepsilon}(\hat{\psi} \mid \psi_0) + (\lambda - T\lambda_0)\|\hat{\beta}\|_1 \leq T\lambda_0 2K + (\lambda + T\lambda_0)\|\beta_0\|_1\end{aligned}$$

As $C > 0$ sufficiently large we have $\lambda \geq 2T\lambda_0$.

And using the condition on $\|\beta_0\|_1$ and $\sqrt{mk}$, we have both $T\lambda_0 2K = o(1)$ and $(\lambda + T\lambda_0)\|\beta_0\|_1 = o(1)$, so we have $\bar{\varepsilon}(\hat{\psi} \mid \psi_0) \to 0$ $(n \to \infty)$.

At last, as the event $\mathcal{T}$ has large probability, we have $\bar{\varepsilon}(\hat{\theta}_\lambda \mid \theta_0) = o_P(1)$ $(n \to \infty)$. □

## Appendix J: Proof of Theorem 3

*Proof* First we discuss the bound for the probability of $\mathcal{T}_{group}$ in (22).

The difference between $\mathcal{T}_{group}$ and $\mathcal{T}$ in (21) is only related to the following entropy of the Entropy Lemma in the proof of Lemma 3.

$$H\left(2u, \left\{\sum_{r=1}^{k} \sum_{j \in \Omega_i} \mathbf{x}_i(\boldsymbol{\beta}_{rj} - \boldsymbol{\beta}_{0,rj}) : \sum_{p} \|\boldsymbol{\beta}_{\mathcal{G}_p} - \boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F \leq \epsilon\right\}, \|\cdot\|_{P_n}\right), \\ \text{for } i = 1 \ldots, n,$$

where $\sum_p \|\boldsymbol{\beta}_{\mathcal{G}_p} - \boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F \leq \epsilon$ still maintains a convex hull for $\boldsymbol{\beta}$ in the metric space equipped with the metric induced by the norm $\|\cdot\|_{P_n}$ defined in the proof of Lemma 3. Thus it still satisfies the Condition of Lemma 2.6.11 of Van Der Vaart and Wellner (1996) which can still be applied to give

$$H\left(2u, \left\{\sum_{r=1}^{k} \sum_{j \in \Omega_i} \mathbf{x}_i(\boldsymbol{\beta}_{rj} - \boldsymbol{\beta}_{0,rj}) : \sum_{p} \|\boldsymbol{\beta}_{\mathcal{G}_p} - \boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F \leq \epsilon\right\}, \|\cdot\|_{P_n}\right) \\ \leq \left(\frac{\epsilon^2}{u^2} + 1\right) \log(1 + kmd), \text{ for } i = 1 \ldots, n.$$

So the probability of event $\mathcal{T}_{group}$ remains the same form with that in Lemma 3.

Then we discuss the bound for the average excess risk and feature selection.

If the event $\mathcal{T}_{group}$ happens, we have

$$\bar{\varepsilon}(\hat{\psi} \mid \psi_0) + \lambda \sum_p \|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}\|_F \leq T\lambda_0\left[\left(\sum_{\mathcal{G}_p} \|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p} - \boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F + \|\eta - \eta_0\|_2\right) \vee \lambda_0\right] \\ + \lambda \sum_p \|\boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F + \bar{\varepsilon}(\psi_0 \mid \psi_0).$$

Using Condition 3 we have $\bar{\varepsilon}(\psi_0 \mid \psi_0) = 0$ and $\bar{\varepsilon}(\hat{\psi} \mid \psi_0) \geq \|\hat{\psi} - \psi_0\|_{Q_n}^2 / c_0^2$.
**Case 1** When the following is true:

$$\sum_p \|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p} - \boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F + \|\hat{\eta} - \eta_0\|_2 \leq \lambda_0,$$

we have

$$\bar{\varepsilon}(\hat{\psi} \mid \psi_0) \leq T\lambda_0^2 + \lambda \sum_p \|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p} - \boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F + \bar{\varepsilon}(\psi_0 \mid \psi_0) \leq (\lambda + T\lambda_0)\lambda_0.$$

**Case 2** When the following is true:

$$\sum_p \|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p} - \boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F + \|\hat{\eta} - \eta_0\|_2 \geq \lambda_0, \\ T\lambda_0\|\hat{\eta} - \eta_0\|_2 \geq (\lambda + T\lambda_0) \sum_{p \in \mathcal{I}} \|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p} - \boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F.$$

As $\sum_{p\in\mathcal{I}^c}\|\boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F=0$, we have

$$\begin{aligned}\bar{\varepsilon}(\hat{\psi}\mid\psi_0)+(\lambda-T\lambda_0)\sum_{p\in\mathcal{I}^c}\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}\|_F&\le 2T\lambda_0\|\hat{\eta}-\eta_0\|_2\\ &\le 2T^2\lambda_0^2c_0^2+\|\hat{\eta}-\eta_0\|_2^2/(2c_0^2)\le 2T^2\lambda_0^2c_0^2+\bar{\varepsilon}(\hat{\psi}\mid\psi_0)/2.\end{aligned}$$

Then we get

$$\bar{\varepsilon}(\hat{\psi}\mid\psi_0)+2(\lambda-T\lambda_0)\sum_{p\in\mathcal{I}^c}\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}\|_F\le 4T^2\lambda_0^2c_0^2.$$

**Case 3** When the following is true:

$$\begin{aligned}&\sum_p\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}-\boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F+\|\hat{\eta}-\eta_0\|_2\ge\lambda_0,\\ &T\lambda_0\|\hat{\eta}-\eta_0\|_2\le(\lambda+T\lambda_0)\sum_{p\in\mathcal{I}}\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}-\boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F,\end{aligned}$$

we have

$$\bar{\varepsilon}(\hat{\psi}\mid\psi_0)+(\lambda-T\lambda_0)\sum_{p\in\mathcal{I}^c}\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}\|_F\le 2(\lambda+T\lambda_0)\sum_{p\in\mathcal{I}}\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}-\boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F.$$

Thus we have

$$\sum_{p\in\mathcal{I}^c}\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}\|_F\le 6\sum_{p\in\mathcal{I}}\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}-\boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F,$$

so we can use the Condition 5 for $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0$ to have

$$\begin{aligned}\bar{\varepsilon}(\hat{\psi}\mid\psi_0)+(\lambda-T\lambda_0)\sum_{p\in\mathcal{I}^c}\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}\|_F&\le 2(\lambda+T\lambda_0)\sqrt{s}\sum_{p\in\mathcal{I}}\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}-\hat{\boldsymbol{\beta}}_{0,\mathcal{G}_p}\|_F\\ \le 2(\lambda+T\lambda_0)\sqrt{s}\kappa\|\hat{\varphi}-(\varphi_0)\|_{Q_n}&\le 2(\lambda+T\lambda_0)^2s\kappa^2c_0^2+\|\hat{\varphi}-(\varphi_0)\|_{Q_n}^2/(2c_0^2)\\ \le 2(\lambda+T\lambda_0)^2s\kappa^2c_0^2+\bar{\varepsilon}(\hat{\psi}\mid\psi_0)/2.\end{aligned}$$

So we have

$$\bar{\varepsilon}(\hat{\psi}\mid\psi_0)+2(\lambda-T\lambda_0)\sum_{p\in\mathcal{I}^c}\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_p}\|_F\le 4(\lambda+T\lambda_0)^2s\kappa^2c_0^2.$$

And without restricted eigenvalue Condition 5, we can prove similarly as in "Appendix I" section, assuming event $\mathcal{T}_{group}$ happens and using the condition on $\sum_p\|\boldsymbol{\beta}_{0,\mathcal{G}_p}\|_F$ and $\sqrt{mk}$. □

## References

Aho K, Derryberry D, Peterson T (2014) Model selection for ecologists: the worldviews of AIC and BIC. Ecology 95(3):631–636

Alfò M, Salvati N, Ranallli MG (2016) Finite mixtures of quantile and M-quantile regression models. Stat Comput 27:1–24

Argyriou A, Evgeniou T, Pontil M (2007a) Multi-task feature learning. In: Advances in neural information processing systems, pp 41–48

Argyriou A, Pontil M, Ying Y, Micchelli CA (2007b) A spectral regularization framework for multi-task structure learning. In: Advances in neural information processing systems, pp 25–32

Bai X, Chen K, Yao W (2016) Mixture of linear mixed models using multivariate t distribution. J Stat Comput Simul 86(4):771–787

Bartolucci F, Scaccia L (2005) The use of mixtures for dealing with non-normal regression errors. Comput Stat Data Anal 48(4):821–834

Barzilai J, Borwein JM (1988) Two-point step size gradient methods. IMA J Numer Anal 8(1):141–148

Becker SR, Candès EJ, Grant MC (2011) Templates for convex cone problems with applications to sparse signal recovery. Math Program Comput 3(3):165–218

Bhat HS, Kumar N (2010) On the derivation of the Bayesian information criterion. School of Natural Sciences, University of California, Oakland

Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of Lasso and Dantzig selector. Ann Stat 37:705–1732

Bishop CM (2006) Pattern recognition. Mach Learn 128:1–58

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends® Mach Learn 3(1):1–122

Candès EJ, Recht B (2009) Exact matrix completion via convex optimization. Found Comput Math 9(6):717–772

Chen X, Kim S, Lin Q, Carbonell JG, Xing EP (2010) Graph-structured multi-task regression and an efficient optimization method for general fused lasso. ArXiv preprint arXiv:1005.3579

Chen J, Zhou J, Ye J (2011) Integrating low-rank and group-sparse structures for robust multi-task learning. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 42–50

Chen J, Liu J, Ye J (2012a) Learning incoherent sparse and low-rank patterns from multiple tasks. ACM Trans Knowl Discov Data (TKDD) 5(4):22

Chen K, Chan KS, Stenseth NC (2012b) Reduced rank stochastic regression with a sparse singular value decomposition. J R Stat Soc Ser B (Stat Methodol) 74(2):203–221

Cover TM, Thomas JA (2012) Elements of information theory. Wiley, Hoboken

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodol) 39:1–38

Doğru FZ, Arslan O (2016) Robust mixture regression using mixture of different distributions. In: Agostinelli C, Basu A, Filzmoser P, Mukherjee D (eds) Recent advances in robust statistics: theory and applications. Springer, New Delhi, pp 57–79

Doğru FZ, Arslan O (2017) Parameter estimation for mixtures of skew Laplace normal distributions and application in mixture regression modeling. Commun Stat Theory Methods 46(21):10,879–10,896

Fahrmeir L, Kneib T, Lang S, Marx B (2013) Regression: models, methods and applications. Springer, Berlin

Fan J, Lv J (2010) A selective overview of variable selection in high dimensional feature space. Stat Sin 20(1):101–148

Fern XZ, Brodley CE (2003) Random projection for high dimensional data clustering: a cluster ensemble approach. In: Proceedings of the 20th international conference on machine learning (ICML-03), pp 186–193

Gong P, Ye J, Zhang C (2012a) Robust multi-task feature learning. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 895–903

Gong P, Ye J, Zhang C (2012b) Multi-stage multi-task feature learning. In: Advances in neural information processing systems, pp 1988–1996

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36

He J, Lawrence R (2011) A graph-based framework for multi-task multi-view learning. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp 25–32

Huang J, Breheny P, Ma S (2012) A selective review of group selection in high-dimensional models. Stat Sci 27(4):481–499

Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Comput 3(1):79–87

Jacob L, Vert J, Bach FR (2009) Clustered multi-task learning: a convex formulation. In: Advances in neural information processing systems, pp 745–752

Jalali A, Sanghavi S, Ruan C, Ravikumar PK (2010) A dirty model for multi-task learning. In: Advances in neural information processing systems, pp 964–972

Ji S, Ye J (2009) An accelerated gradient method for trace norm minimization. In: Proceedings of the 26th annual international conference on machine learning. ACM, pp 457–464

Jin R, Goswami A, Agrawal G (2006) Fast and exact out-of-core and distributed k-means clustering. Knowl Inf Syst 10(1):17–40

Jin X, Zhuang F, Pan SJ, Du C, Luo P, He Q (2015) Heterogeneous multi-task semantic feature learning for classification. In: Proceedings of the 24th ACM international on conference on information and knowledge management. ACM, pp 1847–1850

Jorgensen B (1987) Exponential dispersion models. J R Stat Soc Ser B (Methodol) 49:127–162

Khalili A (2011) An overview of the new feature selection methods in finite mixture of regression models. J Iran Stat Soc 10(2):201–235

Khalili A, Chen J (2007) Variable selection in finite mixture of regression models. J Am Stat Assoc 102(479):1025–1038

Koller D (1996) Toward optimal feature selection. In: Proceedings of the 13th international conference on machine learning, pp 284–292

Kubat M (2015) An introduction to machine learning. Springer, Berlin

Kumar A, Daumé III H (2012) Learning task grouping and overlap in multi-task learning. In: Proceedings of the 29th international conference on machine learning. Omnipress, pp 1723–1730

Lim H, Narisetty NN, Cheon S (2016) Robust multivariate mixture regression models with incomplete data. J Stat Comput Simul 87:1–20

Law MH, Jain AK, Figueiredo M (2002) Feature selection in mixture-based clustering. In: Advances in neural information processing systems, pp 625–632

Li S, Liu ZQ, Chan AB (2014) Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 482–489

Liu J, Ji S, Ye J (2009) Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In: Proceedings of the 25th conference on uncertainty in artificial intelligence. AUAI Press, pp 339–348

McLachlan G, Peel D (2004) Finite mixture models. Wiley, Hoboken

Neal RM, Hinton GE (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Jordan MI (ed) Learning in graphical models. Springer, Dordrecht, pp 355–368

Nelder JA, Baker RJ (1972) Generalized linear models. Encyclopedia of statistical sciences. Wiley, Hoboken

Nesterov Y et al (2007) Gradient methods for minimizing composite objective function. Technical report, UCL

Passos A, Rai P, Wainer J, Daumé III H (2012) Flexible modeling of latent task structures in multitask learning. In: Proceedings of the 29th international conference on machine learning. Omnipress, pp 1283–1290

Schölkopf B, Smola A, Müller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10(5):1299–1319

She Y, Chen K (2017) Robust reduced-rank regression. Biometrika 104(3):633–647

She Y, Owen AB (2011) Outlier detection using nonconvex penalized regression. J Am Stat Assoc 106(494):626–639

Städler N, Bühlmann P, Van De Geer S (2010) $\ell_1$-penalization for mixture regression models. Test 19(2):209–256

Strehl A, Ghosh J (2002a) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3(Dec):583–617

Strehl A, Ghosh J (2002b) Cluster ensembles: a knowledge reuse framework for combining partitionings. In: 18th national conference on artificial intelligence. American Association for Artificial Intelligence, pp 93–98

Tan Z, Kaddoum R, Le Yi Wang HW (2010) Decision-oriented multi-outcome modeling for anesthesia patients. Open Biomed Eng J 4:113

Van de Geer SA (2000) Applications of empirical process theory, vol 91. Cambridge University Press, Cambridge

Van Der Maaten L, Postma E, Van den Herik J (2009) Dimensionality reduction: a comparative. J Mach Learn Res 10:66–71

Van Der Vaart AW, Wellner JA (1996) Weak convergence. Springer, Berlin

Wedel M, DeSarbo WS (1995) A mixture likelihood approach for generalized linear models. J Classif 12(1):21–55

Weruaga L, Vía J (2015) Sparse multivariate gaussian mixture regression. IEEE Trans Neural Netw Learn Syst 26(5):1098–1108

Wang HX, bing Zhang Q, Luo B, Wei S (2004) Robust mixture modelling using multivariate t-distribution with missing information. Pattern Recognit Lett 25(6):701–710

Yang X, Kim S, Xing EP (2009) Heterogeneous multitask learning with joint sparsity constraints. In: Advances in neural information processing systems, pp 2151–2159

Yuksel SE, Wilson JN, Gader PD (2012) Twenty years of mixture of experts. IEEE Trans Neural Netw Learn Syst 23(8):1177–1193

Zhang D, Shen D, Initiative ADN et al (2012) Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer’s disease. NeuroImage 59(2):895–907

Zhang Y, Yeung DY (2011) Multi-task learning in heterogeneous feature spaces. In: 25th AAAI conference on artificial intelligence and the 23rd innovative applications of artificial intelligence conference, AAAI-11/IAAI-11, San Francisco, CA, 7–11 August 2011, Code 87049, Proceedings of the National Conference on Artificial Intelligence, p 574

Zhou J, Chen J, Ye J (2011) Clustered multi-task learning via alternating structure optimization. In: Advances in neural information processing systems, pp 702–710