Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

# High-Dimensional Multi-Task Learning using Multivariate Regression and Generalized Fiducial Inference

Zhenyu Wei & Thomas C. M. Lee

**To cite this article:** Zhenyu Wei & Thomas C. M. Lee (2022): High-Dimensional Multi-Task Learning using Multivariate Regression and Generalized Fiducial Inference, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2022.2090946

**To link to this article:** https://doi.org/10.1080/10618600.2022.2090946

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 00, NO. 0, 1–15
https://doi.org/10.1080/10618600.2022.2090946

Taylor & Francis
Taylor & Francis Group



# High-Dimensional Multi-Task Learning using Multivariate Regression and Generalized Fiducial Inference

Zhenyu Wei and Thomas C. M. Lee

Department of Statistics, University of California, Davis, Davis, CA

**ABSTRACT**

Over the past decades, the Multi-Task Learning (MTL) problem has attracted much attention in the artificial intelligence and machine learning communities. However, most published work in this area focuses on point estimation; that is, estimating model parameters and/or making predictions. This article studies another important aspect of the MTL problem: uncertainty quantification for model choices and predictions. To be more specific, this article approaches the MTL problem with multivariate regression and develops a novel method for deriving a probability density function on the space of all potential regression models. With this density function, point estimates, as well as confidence and prediction ellipsoids, can be obtained for quantities of interest, such as future observations. The proposed method, termed GMTask, is based on the generalized fiducial inference (GFI) framework and is shown to enjoy desirable theoretical properties. Its promising empirical properties are illustrated via a sequence of numerical experiments and applications to two real datasets. Supplementary materials for this article are available online.

**ARTICLE HISTORY**
Received September 2021
Accepted June 2022

**KEYWORDS**
Confidence ellipsoids; GMTask; Large $p$ small $n$; Prediction ellipsoids; Uncertainty quantification

## 1. Introduction

Multi-task learning (MTL) (e.g., Caruana 1997), a subfield of machine learning, aims to jointly learn multiple related tasks with the hope to improve the learning for each individual task. In the past two decades, many approaches have been developed to exploit the common structure shared amongst the tasks. One popular approach is to model the problem with multivariate regression and estimate the parameters through regularization. For example, a novel method called calibrated multivariate regression was proposed by Liu, Wang, and Zhao (2014). The LASSO method of Tibshirani (1996) was extended to the multi-task $L_1/L_2$ LASSO by Obozinski, Taskar, and Jordan (2006). In addition, low rank modeling has also been used; for example, see Yuan et al. (2007). Finally, methods that use different norms (e.g., $L_0$, $L_1$ and/or $L_2$) for regularization have also been investigated; for example, see Evgeniou and Pontil (2007) and Seneviratne and Solo (2012). Notice that these methods allow the high-dimensional setting, and hence regularization is required. For the low-dimensional setting, regularization is not necessary.

Another approach is to use deep neural networks, where the goal is to learn feature representations by using linear or nonlinear transformations of the original features. For example, the method of Zhang et al. (2014) learns common feature representations among different tasks by sharing the first multiple layers, and followed by several task specific layers. The authors of Misra et al. (2016) start out with two separate identical network architectures for two tasks, then use what they refer to as a cross-stitch operation to learn related feature representation for different tasks. In contrast to the multivariate regression approach, where the same data are shared by all the tasks, in deep learning the features of different tasks could come from different datasets.

This article follows the multivariate regression framework. Let $m$ be the number of tasks, $p$ be the number of features (or predictors), and $n$ be the number of observations. The $i$th observation is written as $(\boldsymbol{x}_i, \boldsymbol{Y}_i)$, where $\boldsymbol{x}_i$ is a feature vector of length $p$ and $\boldsymbol{Y}_i$ is a response vector of length $m$. Multivariate regression adopts the following model:

$$\boldsymbol{Y}_{n\times m} = \boldsymbol{X}_{n\times p}\boldsymbol{B}_{p\times m} + \boldsymbol{E}_{n\times m}, \tag{1}$$

where $\boldsymbol{Y}_{n\times m} = (\boldsymbol{Y}_{(1)}, \ldots, \boldsymbol{Y}_{(m)}) = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)^T$ is the response matrix, $\boldsymbol{X}_{n\times p} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is the design matrix, $\boldsymbol{B}_{p\times m} = (\boldsymbol{\beta}_{(1)}, \ldots, \boldsymbol{\beta}_{(m)}) = (\beta_{ij})$ is the regression coefficient parameter matrix, and $\boldsymbol{E}_{n\times m} = (\boldsymbol{\varepsilon}_{(1)}, \ldots, \boldsymbol{\varepsilon}_{(m)}) = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)^T$ is the noise matrix. It is assumed that $\boldsymbol{\varepsilon}_i$ is iid from $N_m(\boldsymbol{0}, \boldsymbol{\Sigma})$ with unknown covariance matrix $\boldsymbol{\Sigma}$, and that $\boldsymbol{E}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent. This article will study the case when $p \gg n$, the so-called high-dimensional scenario. Under this situation, it often assumes that all the regression tasks share a common sparsity pattern; that is, many rows of $\boldsymbol{B}$ are zero vectors.

Up-to-date, most existing work for the MTL problem using multivariate regression focuses on the issues of selecting a model, estimating parameters of the selected model, and making predictions from the selected model. In other words, the major focus has been on point estimation. This article looks at the MTL problem from a different angle: it examines the issue of uncertainty quantification. More specifically, this article

**CONTACT** Thomas C. M. Lee tcmlee@ucdavis.edu Department of Statistics, University of California, Davis, Davis, CA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

applies the generalized fiducial inference (GFI) methodology of Hannig et al. (2016) to perform statistical inference for the MTL problem. In addition to providing point estimates for quantities of interest, the new method also offers various uncertainty measures, such as prediction ellipsoids for future observations. To the best of our knowledge, this article is one of the earliest that systematically addresses uncertainty quantification for the MTL problem when $p \gg n$.

Equation (1) can be represented as an ensemble of univariate linear regression model

$$\boldsymbol{Y}_{(t)} = \boldsymbol{X}\boldsymbol{\beta}_{(t)} + \boldsymbol{\varepsilon}_{(t)}, \quad t = 1, \ldots, m.$$

It is also called the single-task learning problem, which trains a regression model for each task separately. However, it does not use the information that the responses are related amongst different tasks. Therefore, compared with multi-task learning, multiple univariate tasks perform poorly for uncertainty quantification when the correlation between tasks or the number of tasks $m$ are relatively large.

A closely related work is Koner and Williams (2021), where GFI is also applied to the multivariate regression problem. More precisely, the authors apply the epsilon admissible subsets (EAS) approach to perform group variable selection for high-dimensional multivariate regression. In general, the EAS approach can be seen as a way to approximate the so-called generalized fiducial distribution (more below). The focus of Koner and Williams (2021) is to perform variable selection, which is somehow different from the goal of the current article—uncertainty quantification. Nevertheless, the EAS method of Koner and Williams (2021) can be extended to quantify uncertainties, and is shown to be consistent under some mild regularity conditions. Note that the EAS approach was first proposed by Williams and Hannig (2019) to solve the high-dimensional univariate regression problem, and later was applied to the vector autoregressive processes by Williams, Xie, and Hannig (2019).

Before proceeding, it is useful to highlight a major difference between MTL and transfer learning: transfer learning aims to improve the performance of one task (the target task) by borrowing knowledge from other learned tasks (the source tasks), while in multi-task learning the target and the source tasks are learned simultaneously by borrowing knowledge from each other.

The rest of this article is organized as follows. Section 2 provides background on the GFI methodology. Then in Section 3 this methodology is applied to the MTL problem to develop the proposed method for uncertainty quantification. The resulting method is termed GMTask, short for Generalized fiducial inference for Multi-Task, and Section 4 examines the theoretical properties of GMTask. Empirical performance of GMTask is illustrated via numerical experiments and applications to two real data examples in Sections 5 and 6, respectively. Finally, concluding remarks are offered in Section 7 and technical details are delayed to the supplementary materials.

## 2. Background of Generalized Fiducial Inference

The original fiducial inference idea was first introduced by Fisher in 1930s (Fisher 1930) with the goal to assign an appropriate statistical distribution on the parameter space when there is no prior information and hence the classical Bayes' theorem is not applicable. For readers interested in the history of fiducial inference, please refer to Hannig et al. (2016) and Hannig and Lee (2009).

In the recent two decades, there have been lots of efforts devoted to reformulating the fiducial concepts. Some modern modifications include Dempster-Shafer's theory (Dempster 2008) and its related work inferential models (Martin and Liu 2015), confidence distributions (Xie and Singh 2013; Schweder and Hjort 2016), and generalized inference (Weerahandi 1995). In particular, the GFI framework is one of the successful modern formulations of Fisher's fiducial inference idea. This framework has been successfully applied to various statistical learning problems, including wavelet regression (Hannig and Lee 2009) and ultrahigh-dimensional regression (Lai, Hannig, and Lee 2015).

In the GFI framework, the relationship between the data $\boldsymbol{Y}$ and the parameters $\boldsymbol{\theta}$ can be expressed as

$$\boldsymbol{Y} = \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{U}),$$

where $\boldsymbol{G}(\cdot,\cdot)$ is a deterministic function, $\boldsymbol{U}$ is a random component whose distribution is completely known (e.g., iid $N(0,1)$) and independent with $\boldsymbol{\theta}$.

The essential idea behind the philosophy of GFI is the so-called switch principle, which is also the idea behind maximum likelihood estimation: assume $\boldsymbol{y}$ is the observed data of $\boldsymbol{Y}$, then the likelihood is a function of $\boldsymbol{\theta}$, so $\boldsymbol{y}$ is treated as fixed while $\boldsymbol{\theta}$ is random. With this in mind, for any given $\boldsymbol{y}$ if we assume the inverse mapping of $\boldsymbol{G}$ always exists, we can define the following set:

$$\boldsymbol{Q}_{\boldsymbol{y}}(\boldsymbol{u}) = \{\boldsymbol{\theta} : \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{u})\}, \tag{2}$$

where $\boldsymbol{u}$ is a realization of $\boldsymbol{U}$. There are two scenarios that inverse mapping may not exist: no $\boldsymbol{\theta}$ or more than one $\boldsymbol{\theta}$ in the set of $\{\boldsymbol{\theta} : \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{u})\}$. For the first case, we can remove those values of $\boldsymbol{u}$ for which there is no solution and re-normalized the density function of $\boldsymbol{u}$. For the second case, it was suggested by Hannig (2009) that any one of the solutions will give satisfactory results, so we can randomly select one. These two strategies guarantee the existence of the inverse mapping.

Thus, since the distribution of $\boldsymbol{U}$ is known, one can always generate a random sample $\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2, \ldots$, and then obtain a sample of $\boldsymbol{\theta}$ by (2):

$$\tilde{\boldsymbol{\theta}}_1 = \boldsymbol{Q}_{\boldsymbol{y}}(\tilde{\boldsymbol{u}}_1), \quad \tilde{\boldsymbol{\theta}}_2 = \boldsymbol{Q}_{\boldsymbol{y}}(\tilde{\boldsymbol{u}}_2), \ldots$$

We call $\{\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots\}$ a fiducial sample of $\boldsymbol{\theta}$. Its function is similar to Bayesian posterior samples, and we can use it to perform statistical inference like constructing the confidence interval for $\boldsymbol{\theta}$.

The procedure above implicitly defines a density function for $\boldsymbol{\theta}$, which is termed generalized fiducial density (GFD). Indeed, the GFD can be formally defined as the following limit:

$$\lim_{\epsilon \to 0} \left[ \operatorname*{argmin}_{\boldsymbol{\theta}} ||\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{u})|| \,\middle|\, \min_{\boldsymbol{\theta}} ||\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{u})|| \le \epsilon \right].$$

It was shown by Hannig et al. (2016) that, under some reasonable smoothness assumptions, the GFD denoted as $r(\boldsymbol{\theta}|\boldsymbol{y})$ is absolutely continuous and is given by

$$r(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}, \boldsymbol{\theta}) J(\boldsymbol{y}, \boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(\boldsymbol{y}, \boldsymbol{\theta}') J(\boldsymbol{y}, \boldsymbol{\theta}') d\boldsymbol{\theta}'}, \tag{3}$$

where $f(\boldsymbol{y},\boldsymbol{\theta})$ is the likelihood function, and

$$J(\boldsymbol{y},\boldsymbol{\theta}) = D\left(\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{\theta},\boldsymbol{u})\Big|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})}\right), \tag{4}$$

where $D(\boldsymbol{A}) = det(\boldsymbol{A}^T\boldsymbol{A})^{\frac{1}{2}}$ and $\boldsymbol{u} = \boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})$ is the value of $\boldsymbol{u}$ such that $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\theta},\boldsymbol{u})$.

Equations (3) and (4) show the interesting connection between GFI and Bayesian methodology: $r(\boldsymbol{\theta}|\boldsymbol{y})$ in (3) behaves like a Bayesian posterior with $J(\boldsymbol{y},\boldsymbol{\theta})$ as the "prior." However, as $J(\boldsymbol{y},\boldsymbol{\theta})$ depends on $\boldsymbol{y}$, technically it is not a prior density. Note that $J(\boldsymbol{y},\boldsymbol{\theta})$ is very similar to the Jeffreys' prior.

Sometimes it is not possible to calculate the $r(\boldsymbol{\theta}|\boldsymbol{y})$ in (3) analytically and $r(\boldsymbol{\theta}|\boldsymbol{y})$ is only known up to a normalizing constant. In this case, one may resort to use the MCMC methods to generate a fiducial sample from $r(\boldsymbol{\theta}|\boldsymbol{y})$, which is often computationally demanding.

When the model dimension is unknown, Equation (3) is not applicable, which is the situation for the current problem. To solve this problem, a method was proposed by Hannig and Lee (2009) to introduce a penalty in the GFI framework. In particular, it can be shown that the fiducial probability of any model $M$ is proportional to

$$r(M) \propto e^{-q(M)} \int_{\boldsymbol{\Theta}} f_M(\boldsymbol{y},\boldsymbol{\theta})\boldsymbol{J}_M(\boldsymbol{y},\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{5}$$

where $f_M(\boldsymbol{y},\boldsymbol{\theta})$ is the likelihood, $\boldsymbol{J}_M(\boldsymbol{y},\boldsymbol{\theta})$ is the Jacobian (4), and $q(M)$ is the penalty term associated with the model $M$. We will use the minimum description length (MDL) principle (Rissanen 1989, 2007) to derive the penalty term $q(M)$, which shows excellent theoretical and empirical properties.

## 3. GMTask: GFI for Multi-Task Learning

This section applies the above GFI framework to the multi-task learning problem using model (1). We will first calculate $\boldsymbol{J}_M(\boldsymbol{y},\boldsymbol{\theta})$ in (4), then $f_M(\boldsymbol{y},\boldsymbol{\theta})$, and finally $r(M)$ in (5). We will then develop a practical algorithm for simulating fiducial samples from the resulting $r(M)$.

### 3.1. Derivation of r(M)

For this multi-task learning problem, $\boldsymbol{\theta}$ contains three components: $\{M, \boldsymbol{B}_M, \boldsymbol{\Sigma}_M\}$, where $M$ denotes a candidate model that collects a group of predictors, $\boldsymbol{B}_M$ is the parameter matrix of the significant predictors coefficients, and $\boldsymbol{\Sigma}$ is the noise covariance matrix. The data generating function is now

$$\boldsymbol{Y} = \boldsymbol{G}(M,\boldsymbol{B}_M,\boldsymbol{\Sigma},\boldsymbol{U}) = \boldsymbol{X}_M\boldsymbol{B}_M + \boldsymbol{U}\boldsymbol{\Sigma}^{\frac{1}{2}}, \quad M \in \mathcal{M}, \tag{6}$$

where $\boldsymbol{X}_M$ is the design matrix for $M$, and $\mathcal{M}$ is a collection of candidate models. We denotes

$$\hat{\boldsymbol{B}}_M = (\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^T\boldsymbol{y}$$

as the least square estimates of $\boldsymbol{B}_M$, and

$$\boldsymbol{S}_M = (\boldsymbol{y} - \boldsymbol{X}_M\hat{\boldsymbol{B}}_M)^T(\boldsymbol{y} - \boldsymbol{X}_M\hat{\boldsymbol{B}}_M).$$

Using the vectorization of $\boldsymbol{B}_M$ and $\boldsymbol{Y}$ to reformat (6), we obtain

$$\boldsymbol{J}_M(\boldsymbol{y},\boldsymbol{\theta}) = |\det(\boldsymbol{X}_M^T\boldsymbol{X}_M)|^{\frac{m}{2}}\operatorname{tr}(\boldsymbol{S}_M\boldsymbol{\Sigma}^{-1})^{\frac{1}{2}}, \tag{7}$$

where $m$ is the number of tasks, and $\operatorname{tr}(\cdot)$ is the trace function of the matrix.

Next, direct calculations show that the likelihood function is

$$f_M(\boldsymbol{y},\boldsymbol{\theta}) \propto |\boldsymbol{\Sigma}|^{-\frac{n}{2}}\exp\left\{\operatorname{tr}\left[-\frac{1}{2}\left(\boldsymbol{S}_M + (\boldsymbol{B}_M - \hat{\boldsymbol{B}}_M)^T\boldsymbol{X}_M^T\boldsymbol{X}_M(\boldsymbol{B}_M - \hat{\boldsymbol{B}}_M)\right)\boldsymbol{\Sigma}^{-1}\right]\right\}. \tag{8}$$

For the penalty term $q(M)$ in (5), we use the MDL principle to derive its expression, which gives

$$\begin{aligned} q(M) &= \frac{|M|m}{2}\log n + |M|\log p \\ &= |M|\left(\frac{m}{2}\log n + \log p\right), \end{aligned} \tag{9}$$

where $|M|$ is the number of significant predictors in model $M$.

To have a stabler performance, we approximate $\operatorname{tr}(\boldsymbol{S}_M\boldsymbol{\Sigma}^{-1})$ by its expectation. Then substituting (7), (8), and (9) into (5), we obtain the following generalized fiducial probability for any candidate model $M$

$$\begin{aligned} r(M) &\propto R(M) \\ &= \Gamma_m\left(\frac{n - |M| - m}{2}\right) \times |\pi\boldsymbol{S}_M|^{-\frac{n-|M|-m}{2}} \\ &\quad \times [(n - |M|)m]^{\frac{1}{2}} \times (n^{-\frac{m}{2}}p^{-1})^{|M|}, \end{aligned} \tag{10}$$

where $\Gamma_m(\cdot)$ is multivariate gamma function.

### 3.2. Practical Generation of Fiducial Sample

In this section, we propose a procedure to practically generate fiducial sample of $\{M, \boldsymbol{B}, \boldsymbol{\Sigma}\}$ for the multi-task learning problem. The main idea is to first generate a $M$ from (10), then a $\boldsymbol{\Sigma}_M$ from (12), and lastly a $\boldsymbol{B}_M$ from (14).

*Generation of M*: Notice that the total number of possible candidate models is huge ($2^p$ and $p \gg n$) and that, under a sparsity assumption, many of these models have negligible values of $r(M)$. Therefore, a logical way to reduce the computational burden is to only consider a small subset of the candidate models that have nonnegligible values of $r(M)$. In the sequel, we denote this collection of candidate models as $\mathcal{M}^*$. As to be shown below, by doing so the loss in statistical efficiency is, if any, minimal.

Now we propose a two-stage method to obtain $\mathcal{M}^*$. In the first stage a fast screening procedure is applied to remove a large number of insignificant predictors, so that the number of predictors is reduced from $p$ to $p'$, where $p' < n$. Our screening procedure is inspired by the sure independence screening (SIS) procedure in Fan and Lv (2008), which ranks the predictors according to their absolute values of marginal correlations with the response. For the current problem where we have multivariate responses, we suggest calculating the sum of marginal correlation across different tasks for each predictor; this is reasonable when we assume different tasks share the same sparsity pattern. We call this procedure Multi-task SIS, and in practice we set $p' = n - 1$. Notice that Multi-task SIS can be skipped when $p = O(n)$, and but it saves a huge amount of computational time when $p \gg n$.

To further reduce the candidate model space, where the total number of models is $2^{p'}$, in the second stage we apply the multi-task $L_1/L_2$ Lasso of Obozinski, Taskar, and Jordan (2006) to the remaining $p'$ predictors that survived Multi-task SIS. The $L_1/L_2$ block-regularization is commonly referred to as the group Lasso, and the objective function to minimize is

$$\frac{1}{2n}||\boldsymbol{Y}-\boldsymbol{XB}||_F^2+\lambda\sum_{i=1}^{p}\sqrt{\sum_{j=1}^{m}\beta_{ij}^2},$$

where $||\cdot||_F$ is the Frobenius norm. Therefore, one can perform row selection by solving the optimization problem above. We take the sequence of nested models that lies on the regularization path or so-called solution path to form $\mathcal{M}^*$. Note that the multi-task $L_1/L_2$ Lasso solution path can be derived quickly by the coordinate descent method (Friedman, Hastie, and Tibshirani 2010). For the purpose not to missing any candidate models with nonnegligible values of $r(M)$, we repeat the $L_1/L_2$ Lasso procedure on the randomly selected subset of the data multiple times, and take all the models that lie on the different resulting solution paths to form $\mathcal{M}^*$. By using this two-stage method, it is reasonable to expect that $\sum_{M\in\mathcal{M}^*} r(M)$ is very close to 1 and the size of $\mathcal{M}^*$ is substantially smaller than $2^p$.

Once $\mathcal{M}^*$ is obtained, for each $M\in\mathcal{M}^*$, we compute

$$R(M)=\Gamma_m(\tfrac{n-|M|-m}{2})\times|\pi\boldsymbol{S}_M|^{-\frac{n-|M|-m}{2}}\times[(n-|M|)m]^{\frac{1}{2}}\times(n^{-\frac{m}{2}}p^{-1})^{|M|}.$$

Then, for each $M\in\mathcal{M}^*$, the generalized fiducial probability $r(M)$ can be well approximated by

$$\hat{r}(M)=\frac{R(M)}{\sum_{M'\in\mathcal{M}^*}R(M')},\tag{11}$$

and we can generate a $M$ from it.

*Generation of* $\boldsymbol{\Sigma}_M$*:* For any given $M$, it can be shown that the generalized fiducial distribution of $\boldsymbol{\Sigma}_M$ conditional on $M$ is, for example, Triantafyllopoulos (2011)

$$\boldsymbol{\Sigma}_M\sim W_m^{-1}(n-|M|,\boldsymbol{S}_M),\tag{12}$$

where $W_m^{-1}(\cdot,\cdot)$ is the inverse Wishart distribution with the first parameter as the degrees of freedom and second one as the scale matrix. Therefore, $\boldsymbol{\Sigma}_M$ can be generated from (12) once $M$ is given.

*Generation of* $\boldsymbol{B}_M$*:* Finally, it is straightforward to show that the generalized fiducial distribution of $\boldsymbol{B}_M$ conditional on $M$ and $\boldsymbol{\Sigma}_M$ is

$$\text{vec}(\boldsymbol{B}_M)\sim N\big(\text{vec}(\hat{\boldsymbol{B}}_M),\boldsymbol{\Sigma}_M\otimes(\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\big),\tag{13}$$

where vec$(\cdot)$ is vectorization of the matrix, $\hat{\boldsymbol{B}}_M$ is the maximum likelihood estimates of $\boldsymbol{B}_M$, and $\otimes$ is the Kronecker product. Equation (13) has a equivalent form in matrix normal distribution which leads to a more efficient sampling algorithm:

$$\boldsymbol{B}_M\sim MN\big(\hat{\boldsymbol{B}}_M,(\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1},\boldsymbol{\Sigma}_M\big),\tag{14}$$

where $MN(\cdot,\cdot,\cdot)$ is the matrix normal distribution with the first parameter as the mean, the second one as the among-row covariance matrix, and the last one as the among-column covariance matrix.

To sum up, we can generate a fiducial sample $\{\tilde{M},\boldsymbol{B}_{\tilde{M}},\boldsymbol{\Sigma}_{\tilde{M}}\}$ by the following steps:

1. Draw a model $\tilde{M}\in\mathcal{M}^*$ from (11).
2. Generate $\boldsymbol{\Sigma}_{\tilde{M}}$ from (12) given $\tilde{M}$.
3. Sample $\boldsymbol{B}_{\tilde{M}}$ from (14) given $\tilde{M}$ and $\boldsymbol{\Sigma}_{\tilde{M}}$.

Notice that in carrying out the above three steps, no MCMC methods are required, and hence the generation of a fiducial sample is fast.

### 3.3. Point Estimates, Confidence Ellipsoids, and Prediction Ellipsoids

Repeating the above procedure multiple times, one can obtain multiple copies of the fiducial sample $\{\tilde{M},\tilde{\boldsymbol{B}},\tilde{\boldsymbol{\Sigma}}\}$, which is similar to Bayesian posterior sample and can be used for inference. We remark that the multiple copies of $\{\tilde{M},\tilde{\boldsymbol{B}},\tilde{\boldsymbol{\Sigma}}\}$ do not necessary share the same candidate model, as they are drawn from (11).

For any $\boldsymbol{x}_i$, making inference on the conditional mean $\boldsymbol{\mu}_{\boldsymbol{x}_i}=E(\boldsymbol{Y}_i|\boldsymbol{x}_i)=\boldsymbol{B}^T\boldsymbol{x}_i$ can be achieved by using the fiducial sample $\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}_i}=\tilde{\boldsymbol{B}}^T\boldsymbol{x}_i$. Denote the sample mean of these $\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}_i}$'s as $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}_i}$ and their sample covariance matrix as $\hat{\boldsymbol{S}}_{\boldsymbol{x}_i}$. Then one can naturally use $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}_i}$ as a point estimate for $\boldsymbol{\mu}_{\boldsymbol{x}_i}$, while for a $100(1-\alpha)\%$ confidence ellipsoid for $\boldsymbol{\mu}_{\boldsymbol{x}_i}$, one can use the following set (Johnson and Wichern 2002)

$$\{\boldsymbol{\mu}:(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}_i})^T\hat{\boldsymbol{S}}_{\boldsymbol{x}_i}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}_i})\le\chi_m^2(\alpha)\},\tag{15}$$

where $\chi_m^2(\alpha)$ is the upper $(100\alpha)$th percentile of a $\chi^2$-distribution with $m$ degrees of freedom.

One can use a similar approach to obtain a point estimate and a prediction ellipsoid for any *new (or future)* observation $\boldsymbol{Y}_i$ at $\boldsymbol{x}_i$. When comparing to its conditional expectation $\boldsymbol{\mu}_{\boldsymbol{x}_i}=E(\boldsymbol{Y}_i|\boldsymbol{x}_i)=\boldsymbol{B}^T\boldsymbol{x}_i$, the new observation $\boldsymbol{Y}_i$ has a higher variability, which can be accounted for by suitably adding a noise term to the fiducial sample. More specifically, instead of using $\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}_i}=\tilde{\boldsymbol{B}}^T\boldsymbol{x}_i$, we use $\tilde{\boldsymbol{\mu}}^*_{\boldsymbol{x}_i}=\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}_i}+\tilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{Z}$, where $\boldsymbol{Z}\sim N(\boldsymbol{0},\boldsymbol{I}_m)$. Let the sample mean of these $\tilde{\boldsymbol{\mu}}^*_{\boldsymbol{x}_i}$'s be $\hat{\boldsymbol{\mu}}^*_{\boldsymbol{x}_i}$ and their sample covariance matrix be $\hat{\boldsymbol{S}}^*_{\boldsymbol{x}_i}$. Then, as before, one can use $\hat{\boldsymbol{\mu}}^*_{\boldsymbol{x}_i}$ as a point estimate for $\boldsymbol{Y}_i$ and the following set as a $100(1-\alpha)\%$ prediction ellipsoid for $\boldsymbol{Y}_i$ (Johnson and Wichern 2002):

$$\{\boldsymbol{\mu}:(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}^*_{\boldsymbol{x}_i})^T\hat{\boldsymbol{S}}^{*-1}_{\boldsymbol{x}_i}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}^*_{\boldsymbol{x}_i})\le\chi_m^2(\alpha)\}.\tag{16}$$

## 4. Theoretical Properties

This section presents some asymptotic properties of the above generalized fiducial based method. We assume that $p$ is diverging and the size of the true model is either fixed or diverging. Before presenting our theorem, we first provide the following necessary notations and assumptions, which are standard in related work, for example, Hannig et al. (2016) and Yanagihara et al. (2015).

### 4.1. Preliminaries and Notations

Let $M$ be any model and $M_0$ be the true model, while $|M|$ and $|M_0|$ be the number of predictors in $M$ and $M_0$, respectively. To reduce the candidate model space, we only consider $M\in\mathcal{M}$

where $\mathcal{M} = \{M : |M| \le c|M_0|\}$ for a fixed finite constant $c > 1$; that is, the model whose size is comparable to the true model.

Denote $\mathcal{M}_+ = \{M : M_0 \subset M\}$, $\mathcal{M}_- = \{M : M_0 \not\subseteq M\}$ and $\mathcal{M} = \mathcal{M}_- \cup \{M_0\} \cup \mathcal{M}_+$. Define

$$\boldsymbol{S}_M = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H}_M)\boldsymbol{Y}, \hat{\boldsymbol{\Sigma}}_M = \frac{1}{n}\boldsymbol{S}_M,$$

where $\boldsymbol{H}_M = \boldsymbol{X}_M(\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^T$ is the projection matrix to the subspace spanned by the columns of $\boldsymbol{X}_M$. For simplicity, we write $\boldsymbol{X}_{M_0}$ as $\boldsymbol{X}_0$, same for $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$. We assume that the data are generated from the following true model:

$$\boldsymbol{Y} \sim N_{n\times m}(\boldsymbol{X}_0\boldsymbol{B}_0, \boldsymbol{\Sigma}_0 \otimes \boldsymbol{I}_n).$$

In order to prove the consistency of our method, we need to describe the noncentrality matrix. For a model $M \in \mathcal{M}$, let a $m \times m$ noncentrality matrix denoted by

$$\Delta(M) = \boldsymbol{\Sigma}_0^{-\frac{1}{2}}\boldsymbol{B}_0^T\boldsymbol{X}_0^T(\boldsymbol{I}_n - \boldsymbol{H}_M)\boldsymbol{X}_0\boldsymbol{B}_0\boldsymbol{\Sigma}_0^{-\frac{1}{2}}.$$

### 4.2. Assumptions

We need the following assumptions.

(A1) The true model $M_0 \in \mathcal{M}$, where $\mathcal{M} = \{M : |M| \le c|M_0|\}$ for a fixed finite constant $c > 1$.

(A2) $\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}_M^T\boldsymbol{X}_M$ exists and is positive definite for all $M \in \mathcal{M}$, and $\lim_{n\to\infty} \frac{1}{n}\Delta(M) = \boldsymbol{\Psi}_M$ exists and is not the zero matrix for all $M \in \mathcal{M}_-$.

(A3) When $p$ is too large, a variable screening procedure can be used to reduce the size of $\mathcal{M}$ in practice. That screening procedure should result in a class of candidate models $\mathcal{M}^*$ that satisfies:

$$P(M_0 \in \mathcal{M}^*) \to 1 \quad \text{and} \quad \log(|\mathcal{M}_j^*|) = o(j\log n), \tag{17}$$

where $\mathcal{M}_j^*$ denotes the set of all submodels in $\mathcal{M}^*$ of size $j$. The first condition in (17) ensures that the true model is contained in $\mathcal{M}^*$, while the second condition in (17) implies that the size of model space $\mathcal{M}^*$ is not too large.

### 4.3. Main Result

We establish the following theorem and its proof can be found in the Appendix.

*Theorem 1.* Assume A1–A2 hold. As $n \to \infty, p \to \infty, \log p = o(n)$ and $|M_0| + m = o(\log n)$, we have

$$\max_{M\neq M_0, M\in\mathcal{M}} \frac{r(M)}{r(M_0)} \xrightarrow{P} 0. \tag{18}$$

Moreover, if A3 also holds, we have

$$\frac{r(M_0)}{\sum_{M\in\mathcal{M}^*} r(M)} \xrightarrow{P} 1. \tag{19}$$

Theorem 1 shows that the true model $M_0$ has the highest generalized fiducial probability among all the candidate models in $\mathcal{M}$. However, Equation (18) does not imply (19) in general since the model candidate space can be very large as $p$ goes to infinity. If we can constrain the candidate models in a class that satisfies (17), then the true model will be selected with probability tending to 1.

In practice, we use the multi-task $L_1/L_2$ Lasso to generate the candidate models as discussed in Section 3.2. The resulting model space satisfies (17), as the multi-task Lasso is selection consistent for some $\lambda$ as shown in Obozinski, Wainwright, and Jordan (2008). With Theorems 2 and 3 of Hannig et al. (2016), Theorem 1 also implies that the confidence ellipsoids and prediction ellipsoids constructed using the generalized fiducial density (10) will have correct asymptotic coverage rates, and the generalized fiducial distribution and the derived point estimates are consistent.

## 5. Simulation Experiments

A sequence of simulation experiments has been performed to evaluate the practical performance of GMTask, and its relative merits when compared to other approaches.

### 5.1. Uncertainty Quantification

Inspired by the setting in Fan, Guo, and Hao (2012) and Bedrick and Tsai (1994), we consider the following data-generating model

$$\boldsymbol{Y}_i = b(X_1 + \cdots + X_k)\mathbb{1}_m + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{\varepsilon}_i$ is iid as $N_m(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\mathbb{1}_m$ is a $m \times 1$ vector of ones, $k$ is the number of significant predictors with nonzero coefficients, and $b$ is the coefficient value that controls the signal-to-noise ratio. The covariates $X_i$'s are generated from standard normal distribution with $\text{cor}(X_i, X_j) = \rho_X^{|i-j|}$. We set the covariance matrix $\boldsymbol{\Sigma} = (1-\rho_\Sigma)\boldsymbol{I}_m + \rho_\Sigma\boldsymbol{J}_m$, where $\boldsymbol{J}_m$ is a $m \times m$ matrix of ones. Two combinations of $(n, p, m, k)$ were used: (200, 2000, 2, 3) and (200, 2000, 3, 3). For each combination setting, we used three different $b$, two different $\rho_X$ and three different $\rho_\Sigma$: $b = (1/\sqrt{k}, 2/\sqrt{k}, 3/\sqrt{k})$, $\rho_X = (0, 0.5)$ and $\rho_\Sigma = (0.3, 0.6, 0.8)$. Therefore, a total of $2 \times 3 \times 2 \times 3 = 36$ experimental settings were considered. The number of repetitions for each setting was 1000.

For each simulated dataset, we applied five methods to obtain various estimates $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{\Sigma}}$ for the coefficient matrix $\boldsymbol{B}$ and the covariance matrix $\boldsymbol{\Sigma}$, respectively. Then, we constructed the prediction ellipsoids for new $\boldsymbol{Y}_i$'s and the confidence ellipsoids for the conditional mean $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ for 50 randomly selected new designed points $\boldsymbol{x}_i$'s. The five methods were

- GMTask: the proposed GFI based method with 10,000 fiducial samples of $\{M, \boldsymbol{B}, \boldsymbol{\Sigma}\}$;
- V-L0LS-CD: the method of Seneviratne and Solo (2012);
- AICc: the method of Bedrick and Tsai (1994);
- BIC: similar to AICc but uses BIC to select the final model; and
- Oracle: the method that uses the true model.

Of course, the last method, Oracle, is not applicable in practice, but it is used as a benchmark comparison here. For the last four

methods, we first applied them to perform model selection in $\mathcal{M}^*$, then used the classical multivariate linear model theory (Johnson and Wichern 2002) to the final selected model to build the ellipsoids. For GMTask, we used (15) and (16) to construct the ellipsoids.

For all five methods, we constructed 90%, 95%, and 99% prediction ellipsoids and confidence ellipsoids for all 1000 simulated datasets from each simulation setting. To evaluate the relative performances, we calculated the average empirical coverage rates for these prediction ellipsoids and confidence ellipsoids. Those results that correspond to $\rho_\Sigma = 0.6$ are summarized in Tables 1–2. The remaining results are similar and can be found in a separate supplementary materials. It can be seen that the proposed GMTask was nearly as good as the oracle method, and it outperformed other non-oracle methods significantly especially for the confidence ellipsoids of the conditional mean function $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$.

### 5.2. Point Estimates

Furthermore, we also calculated Mean Squared Prediction Error (MSPE) and Mean Squared Mahalanobis Distance (MSMD) to compare the performances of predicting a new observation $\boldsymbol{Y}_i$ at $\boldsymbol{x}_i$, defined, respectively, as

$$\text{MSPE}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i)^T(\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i)$$

and

$$\text{MSMD}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i)^T\boldsymbol{D}_i^{-1}(\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i).$$

Here, $\hat{\boldsymbol{Y}}_i$, a $m \times 1$ vector, is a generic notation that denotes the prediction of $\boldsymbol{Y}_i$ by any one of the five methods. And $\boldsymbol{D}_i$ is $\text{cov}(\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_{i,\text{oracle}}) = (1 + \boldsymbol{x}_{i0}^T(\boldsymbol{X}_0^T\boldsymbol{X}_0)^{-1}\boldsymbol{x}_{i0})\boldsymbol{\Sigma}_0$, where $\boldsymbol{x}_{i0}$ is the true model $M_0$ subset of $\boldsymbol{x}_i$ and $\hat{\boldsymbol{Y}}_{i,\text{oracle}} = \hat{\boldsymbol{B}}_{\text{oracle}}^T\boldsymbol{x}_{i0}$ is the oracle prediction for $\boldsymbol{Y}_i$. Note that when $\boldsymbol{D}_i$ is an identity matrix, MSMD reduces to MSPE.

Similarly, to measure the performance of predicting the mean function $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ at $\boldsymbol{x}_i$, we used Mean Squared Error (MSE) and Mean Squared Mahalanobis Distance (MSMD), defined, respectively, as

$$\text{MSE}\big(E(\boldsymbol{Y}), \hat{\boldsymbol{Y}}\big) = \frac{1}{n}\sum_{i=1}^{n}\big(E(\boldsymbol{Y}_i) - \hat{\boldsymbol{Y}}_i\big)^T\big(E(\boldsymbol{Y}_i) - \hat{\boldsymbol{Y}}_i\big)$$

and

$$\text{MSMD}\big(E(\boldsymbol{Y}), \hat{\boldsymbol{Y}}\big) = \frac{1}{n}\sum_{i=1}^{n}\big(E(\boldsymbol{Y}_i) - \hat{\boldsymbol{Y}}_i\big)^T\tilde{\boldsymbol{D}}_i^{-1}\big(E(\boldsymbol{Y}_i) - \hat{\boldsymbol{Y}}_i\big),$$

where $E(\boldsymbol{Y}_i) = \boldsymbol{B}_0^T\boldsymbol{x}_{i0}$ and $\tilde{\boldsymbol{D}}_i = \text{cov}(\hat{\boldsymbol{Y}}_{i,\text{oracle}}) = \boldsymbol{x}_{i0}^T(\boldsymbol{X}_0^T\boldsymbol{X}_0)^{-1}\boldsymbol{x}_{i0}\boldsymbol{\Sigma}_0$. Again, MSMD reduces to MSE when $\tilde{\boldsymbol{D}}_i$ is an identity matrix.

The results for $\rho_\Sigma = 0.6$ are summarized in Tables 3 and 4, while the remaining results are given in the supplement. From these tables, one can see that the MSE, MSPE and MSMD of the GMTask estimates are usually not much larger than those from Oracle. For the conditional mean $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$, the GMTask estimates outperformed the other non-oracle estimates significantly.

### 5.3. Misspecified Models

In all the previous experiments we only consider the situation that the true predictors are amongst the $p$ predictors that are available to the methods. Now we consider the case that some of the true predictors are not available. In other words, model (1) is incorrect, as $E(\boldsymbol{Y})$ cannot be represented as a linear combination of the $p$ predictors $X_i$'s.

The following four models were used to generate the noisy data:

1: $\boldsymbol{Y}_i = b(X_1 + X_2 + X_{2001})\mathbb{1}_m + \boldsymbol{\varepsilon}_i$, where $X_{2001}$ is not one of the $p = 2000$ available predictors $X_i$'s;
2: $\boldsymbol{Y}_i = b(X_1 + X_2 + X_{1999}X_{2000})\mathbb{1}_m + \boldsymbol{\varepsilon}_i$, where $X_{1999}X_{2000}$ is an interaction term;
3: $\boldsymbol{Y}_i = b(X_1 + X_2 + X_{2000}^2)\mathbb{1}_m + \boldsymbol{\varepsilon}_i$, where $X_{2000}^2$ is the squared term of $X_{2000}$;
4: $\boldsymbol{Y}_i = b(X_1 + X_2 + X_1^2)\mathbb{1}_m + \boldsymbol{\varepsilon}_i$, where $X_1^2$ is the squared term of $X_1$.

Except for the model formulations, the remaining experimental parameters were identical to Section 5.1. The average empirical coverage rates of the prediction ellipsoids for predicting $\boldsymbol{Y}_i$ with $\rho_\Sigma = 0.6$ are reported in Tables 5–8. Results for the other values of $\rho_\Sigma$ are given in the supplementary materials. When compared with other methods, GMTask provided more reliable results.

## 6. Real Data Application: Polymerase Chain Reaction Data

An experiment was conducted by Lan et al. (2006) to examine the genetics of two inbred mouse populations: B6 and BTBR. The expression levels of 22,575 genes of 60 mice were measured. Some physiological phenotypes, including the numbers of stearoyl-CoA desaturase 1 (SCD1) and phosphoenopyruvate carboxykinase (PEPCK), were also measured by quantitative real-time polymerase chain reaction. Since $n = 60$, $p = 22575$ and $m = 2$, it is a high dimensional multi-task learning problem as $p \gg n$. A so-called credible approach was used by Bondell and Reich (2012) to predict each of these two phenotypes independently based on the gene expression data, and therefore it was a single task method.

The following procedure was conducted to evaluate the empirical coverage rates produced by the four methods compared before: GMTask AICc, BIC and V-L0LS-CD. First, we left out the first observation as the test point. Then we applied GMTask and the other methods to the remaining 59 observations to construct the 90%, 95%, and 99% prediction ellipsoids for this first observation. We repeated this leave-one-out process with the remaining 59 observations and the resulting empirical coverages of these prediction ellipsoids are summarized in Table 9. It can be seen that GMTask is the preferred method as its coverage rates of the 90%, 95%, and 99% prediction ellipsoids are all close to the nominal levels. Also notice that the volume of the 90% GMTask prediction ellipsoid is smaller than that of the 99% AICc prediction ellipsoid, while their empirical coverage rates are both 0.900. Overall, in terms of making predictions, GMTask is more accurate for quantifying the uncertainties for this dataset.

**Table 1.** Empirical coverage rates of the prediction ellipsoids for $\boldsymbol{Y}_i$ when $\rho_\Sigma = 0.6$.

| | | | 90% | 95% | 99% |
|---|---|---|---|---|---|
| $m = 2$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.898**(11.822) | **0.949**(15.381) | **0.990**(23.644) |
| | | AICc | 0.871(11.306) | 0.930(14.762) | 0.984(22.884) |
| | | BIC | 0.880(11.454) | 0.936(14.956) | 0.985(23.183) |
| | | V-L0LS-CD | 0.801(10.314) | 0.873(13.469) | 0.950(20.883) |
| | | Oracle | 0.899(11.843) | 0.950(15.462) | 0.991(23.967) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.897**(11.787) | **0.947**(15.335) | **0.989**(23.573) |
| | | AICc | 0.872(11.308) | 0.929(14.765) | 0.983(22.888) |
| | | BIC | 0.882(11.482) | 0.937(14.992) | 0.986(23.238) |
| | | V-L0LS-CD | 0.864(11.258) | 0.924(14.700) | 0.980(22.787) |
| | | Oracle | 0.897(11.807) | 0.949(15.417) | 0.990(23.896) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.898**(11.843) | **0.949**(15.408) | **0.989**(23.686) |
| | | AICc | 0.873(11.415) | 0.931(14.905) | 0.984(23.105) |
| | | BIC | 0.885(11.615) | 0.941(15.166) | 0.986(23.507) |
| | | V-L0LS-CD | 0.880(11.545) | 0.935(15.074) | 0.984(23.367) |
| | | Oracle | 0.898(11.862) | 0.949(15.488) | 0.990(24.007) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.898**(11.854) | **0.948**(15.422) | **0.990**(23.707) |
| | | AICc | 0.870(11.285) | 0.930(14.735) | 0.983(22.842) |
| | | BIC | 0.879(11.428) | 0.935(14.921) | 0.985(23.129) |
| | | V-L0LS-CD | 0.844(10.935) | 0.907(14.279) | 0.971(22.136) |
| | | Oracle | 0.898(11.847) | 0.949(15.468) | 0.990(23.976) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.902**(11.800) | **0.950**(15.353) | **0.989**(23.601) |
| | | AICc | 0.877(11.350) | 0.933(14.821) | 0.984(22.975) |
| | | BIC | 0.888(11.539) | 0.941(15.066) | 0.987(23.353) |
| | | V-L0LS-CD | 0.883(11.487) | 0.937(14.999) | 0.985(23.250) |
| | | Oracle | 0.903(11.819) | 0.950(15.432) | 0.989(23.919) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.897**(11.804) | **0.947**(15.358) | **0.989**(23.609) |
| | | AICc | 0.873(11.407) | 0.932(14.895) | 0.984(23.090) |
| | | BIC | 0.885(11.618) | 0.940(15.169) | 0.987(23.512) |
| | | V-L0LS-CD | 0.885(11.616) | 0.940(15.167) | 0.986(23.510) |
| | | Oracle | 0.897(11.823) | 0.948(15.437) | 0.989(23.927) |
| $m = 3$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.898**(40.284) | **0.947**(56.304) | **0.988**(98.485) |
| | | AICc | 0.876(38.891) | 0.934(54.691) | 0.984(96.996) |
| | | BIC | 0.893(39.987) | 0.945(56.228) | 0.988(99.706) |
| | | V-L0LS-CD | 0.891(39.902) | 0.943(56.108) | 0.987(99.495) |
| | | Oracle | 0.899(40.477) | 0.949(56.917) | 0.989(100.925) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.900**(40.328) | **0.949**(56.365) | **0.989**(98.592) |
| | | AICc | 0.881(39.080) | 0.936(54.957) | 0.985(97.468) |
| | | BIC | 0.897(40.210) | 0.947(56.541) | 0.989(100.260) |
| | | V-L0LS-CD | 0.897(40.250) | 0.948(56.597) | 0.989(100.361) |
| | | Oracle | 0.901(40.530) | 0.951(56.991) | 0.990(101.057) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.898**(40.207) | **0.948**(56.196) | **0.988**(98.295) |
| | | AICc | 0.878(39.070) | 0.936(54.942) | 0.985(97.441) |
| | | BIC | 0.895(40.091) | 0.947(56.374) | 0.989(99.965) |
| | | V-L0LS-CD | 0.895(40.145) | 0.947(56.449) | 0.989(100.098) |
| | | Oracle | 0.899(40.404) | 0.949(56.814) | 0.989(100.744) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.899**(40.393) | **0.949**(56.456) | **0.989**(98.750) |
| | | AICc | 0.875(38.790) | 0.934(54.549) | 0.985(96.745) |
| | | BIC | 0.895(39.961) | 0.946(56.192) | 0.989(99.643) |
| | | V-L0LS-CD | 0.894(40.019) | 0.946(56.273) | 0.988(99.786) |
| | | Oracle | 0.900(40.451) | 0.951(56.879) | 0.990(100.859) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.899**(40.195) | **0.948**(56.180) | **0.989**(98.268) |
| | | AICc | 0.880(39.038) | 0.935(54.898) | 0.986(97.363) |
| | | BIC | 0.895(40.074) | 0.947(56.350) | 0.989(99.923) |
| | | V-L0LS-CD | 0.896(40.162) | 0.947(56.474) | 0.989(100.141) |
| | | Oracle | 0.899(40.392) | 0.949(56.797) | 0.990(100.713) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.898**(40.256) | **0.947**(56.265) | **0.989**(98.415) |
| | | AICc | 0.879(39.246) | 0.937(55.190) | 0.986(97.881) |
| | | BIC | 0.896(40.228) | 0.946(56.567) | 0.989(100.307) |
| | | V-L0LS-CD | 0.896(40.291) | 0.947(56.654) | 0.989(100.461) |
| | | Oracle | 0.898(40.449) | 0.948(56.877) | 0.989(100.855) |

NOTE: Numbers in parentheses are averaged volumes of the ellipsoids. This table is for the case when $(n, p, k) = (200, 2000, 3)$. Best results are bolded (other than Oracle's).

## 7. Conclusion

In this article, we studied the problem of uncertainty quantification in multi-task learning under the "large $p$ small $n$" scenario. We adopted the GFI framework to perform statistical inference for this problem. In particular, we developed GMTask, a method for generating fiducial samples that can be used

**Table 2.** Empirical coverage rates of the confidence ellipsoids for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ when $\rho_\Sigma = 0.6$.

| | | | 90% | 95% | 99% |
|---|---|---|---|---|---|
| $m = 2$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.916**(0.190) | **0.961**(0.247) | **0.993**(0.380) |
| | | AICc | 0.529(0.317) | 0.619(0.414) | 0.763(0.641) |
| | | BIC | 0.603(0.251) | 0.683(0.327) | 0.812(0.507) |
| | | V-L0LS-CD | 0.454(0.515) | 0.529(0.672) | 0.657(1.043) |
| | | Oracle | 0.906(0.179) | 0.954(0.234) | 0.992(0.362) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.901**(0.188) | **0.948**(0.244) | **0.990**(0.375) |
| | | AICc | 0.527(0.319) | 0.618(0.417) | 0.770(0.646) |
| | | BIC | 0.628(0.241) | 0.709(0.315) | 0.834(0.488) |
| | | V-L0LS-CD | 0.601(0.304) | 0.676(0.396) | 0.796(0.614) |
| | | Oracle | 0.894(0.179) | 0.945(0.234) | 0.990(0.363) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.905**(0.187) | **0.952**(0.244) | **0.991**(0.374) |
| | | AICc | 0.545(0.327) | 0.639(0.427) | 0.789(0.662) |
| | | BIC | 0.675(0.231) | 0.752(0.302) | 0.867(0.467) |
| | | V-L0LS-CD | 0.690(0.253) | 0.763(0.331) | 0.863(0.512) |
| | | Oracle | 0.899(0.180) | 0.948(0.235) | 0.990(0.365) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.911**(0.200) | **0.952**(0.260) | **0.986**(0.399) |
| | | AICc | 0.522(0.320) | 0.610(0.418) | 0.758(0.647) |
| | | BIC | 0.577(0.256) | 0.661(0.334) | 0.798(0.517) |
| | | V-L0LS-CD | 0.531(0.384) | 0.608(0.501) | 0.734(0.777) |
| | | Oracle | 0.900(0.180) | 0.947(0.234) | 0.988(0.363) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.907**(0.187) | **0.955**(0.243) | **0.990**(0.373) |
| | | AICc | 0.546(0.324) | 0.638(0.423) | 0.788(0.656) |
| | | BIC | 0.656(0.235) | 0.735(0.306) | 0.854(0.475) |
| | | V-L0LS-CD | 0.685(0.256) | 0.758(0.334) | 0.856(0.518) |
| | | Oracle | 0.901(0.179) | 0.951(0.234) | 0.990(0.363) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.902**(0.183) | **0.950**(0.238) | **0.988**(0.366) |
| | | AICc | 0.553(0.325) | 0.648(0.424) | 0.799(0.657) |
| | | BIC | 0.700(0.221) | 0.776(0.289) | 0.884(0.448) |
| | | V-L0LS-CD | 0.742(0.228) | 0.810(0.298) | 0.897(0.461) |
| | | Oracle | 0.896(0.178) | 0.946(0.233) | 0.988(0.361) |
| $m = 3$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.901**(0.094) | **0.950**(0.132) | **0.989**(0.230) |
| | | AICc | 0.554(0.202) | 0.648(0.284) | 0.798(0.505) |
| | | BIC | 0.784(0.109) | 0.846(0.153) | 0.921(0.271) |
| | | V-L0LS-CD | 0.808(0.116) | 0.866(0.163) | 0.928(0.289) |
| | | Oracle | 0.899(0.093) | 0.950(0.130) | 0.990(0.231) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.902**(0.095) | **0.951**(0.132) | **0.990**(0.232) |
| | | AICc | 0.560(0.206) | 0.657(0.290) | 0.808(0.514) |
| | | BIC | 0.818(0.104) | 0.878(0.146) | 0.943(0.259) |
| | | V-L0LS-CD | 0.842(0.104) | 0.898(0.147) | 0.954(0.260) |
| | | Oracle | 0.898(0.093) | 0.949(0.131) | 0.990(0.233) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.898**(0.095) | **0.951**(0.133) | **0.992**(0.233) |
| | | AICc | 0.570(0.205) | 0.667(0.288) | 0.817(0.510) |
| | | BIC | 0.814(0.104) | 0.876(0.146) | 0.944(0.259) |
| | | V-L0LS-CD | 0.836(0.103) | 0.895(0.145) | 0.954(0.258) |
| | | Oracle | 0.895(0.094) | 0.950(0.132) | 0.991(0.234) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.896**(0.099) | **0.942**(0.138) | **0.981**(0.242) |
| | | AICc | 0.547(0.205) | 0.640(0.288) | 0.788(0.510) |
| | | BIC | 0.778(0.109) | 0.842(0.153) | 0.919(0.271) |
| | | V-L0LS-CD | 0.817(0.111) | 0.874(0.156) | 0.935(0.276) |
| | | Oracle | 0.898(0.093) | 0.949(0.131) | 0.990(0.232) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.893**(0.094) | **0.945**(0.131) | **0.988**(0.230) |
| | | AICc | 0.563(0.207) | 0.658(0.292) | 0.814(0.517) |
| | | BIC | 0.809(0.104) | 0.869(0.146) | 0.941(0.259) |
| | | V-L0LS-CD | 0.839(0.103) | 0.896(0.145) | 0.956(0.257) |
| | | Oracle | 0.890(0.093) | 0.944(0.131) | 0.988(0.232) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.887**(0.094) | **0.942**(0.132) | **0.989**(0.230) |
| | | AICc | 0.572(0.207) | 0.671(0.291) | 0.826(0.516) |
| | | BIC | 0.828(0.101) | 0.890(0.142) | 0.955(0.253) |
| | | V-L0LS-CD | 0.846(0.101) | 0.905(0.141) | 0.966(0.251) |
| | | Oracle | 0.886(0.094) | 0.941(0.132) | 0.989(0.233) |

NOTE: Numbers in parentheses are averaged volumes of the ellipsoids. This table is for the case when $(n, p, k) = (200, 2000, 3)$.Best results are bolded (other than Oracle's).

to construct various point estimates, confidence ellipsoids and prediction ellipsoids. Our theoretical results show that, under some mild regularity conditions, the estimates obtained by GMTask are consistent, while the confidence ellipsoids and prediction ellipsoids enjoy correct asymptotic frequentist properties. Numerical results from simulation experiments

**Table 3.** Bias of the estimates of $\boldsymbol{Y}_i$ when $\rho_\Sigma = 0.6$.

| | | $m = 2$ | | $m = 3$ | |
|---|---|---|---|---|---|
| | | MSPE | MSMD | MSPE | MSMD |
| $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **2.036**(0.075) | **1.999**(0.063) | **3.048**(0.103) | **3.008**(0.078) |
| | AICc | 2.225(0.083) | 2.129(0.067) | 3.309(0.114) | 3.156(0.082) |
| | BIC | 2.163(0.080) | 2.085(0.065) | 3.109(0.106) | 3.044(0.079) |
| | V-L0LS-CD | 2.595(0.101) | 2.393(0.077) | 3.122(0.107) | 3.048(0.079) |
| | Oracle | 2.035(0.075) | 1.997(0.063) | 3.045(0.103) | 3.007(0.078) |
| $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **2.037**(0.075) | **2.008**(0.063) | **3.021**(0.103) | **2.984**(0.077) |
| | AICc | 2.223(0.082) | 2.140(0.067) | 3.261(0.112) | 3.124(0.081) |
| | BIC | 2.142(0.079) | 2.082(0.066) | 3.060(0.104) | 3.008(0.078) |
| | V-L0LS-CD | 2.249(0.085) | 2.154(0.068) | 3.057(0.104) | 3.005(0.078) |
| | Oracle | 2.036(0.075) | 2.007(0.063) | 3.020(0.103) | 2.983(0.077) |
| $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **2.028**(0.075) | **2.003**(0.064) | **3.056**(0.104) | **3.011**(0.078) |
| | AICc | 2.202(0.082) | 2.126(0.068) | 3.286(0.114) | 3.145(0.082) |
| | BIC | 2.110(0.078) | 2.061(0.066) | 3.091(0.106) | 3.032(0.078) |
| | V-L0LS-CD | 2.150(0.080) | 2.088(0.067) | 3.089(0.106) | 3.029(0.078) |
| | Oracle | 2.028(0.075) | 2.002(0.064) | 3.056(0.104) | 3.011(0.078) |
| $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **2.051**(0.076) | **2.010**(0.063) | **3.065**(0.104) | **2.995**(0.077) |
| | AICc | 2.229(0.084) | 2.135(0.067) | 3.309(0.114) | 3.138(0.081) |
| | BIC | 2.173(0.082) | 2.095(0.066) | 3.110(0.106) | 3.025(0.078) |
| | V-L0LS-CD | 2.393(0.093) | 2.247(0.072) | 3.110(0.106) | 3.024(0.078) |
| | Oracle | 2.041(0.076) | 2.004(0.063) | 3.046(0.103) | 2.987(0.077) |
| $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **2.011**(0.075) | **1.983**(0.063) | **3.026**(0.102) | **2.981**(0.077) |
| | AICc | 2.188(0.082) | 2.109(0.067) | 3.267(0.112) | 3.119(0.081) |
| | BIC | 2.108(0.079) | 2.052(0.065) | 3.069(0.104) | 3.005(0.078) |
| | V-L0LS-CD | 2.136(0.080) | 2.070(0.066) | 3.063(0.104) | 3.001(0.078) |
| | Oracle | 2.010(0.075) | 1.982(0.063) | 3.026(0.102) | 2.980(0.077) |
| $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **2.035**(0.075) | **2.009**(0.064) | **3.047**(0.103) | **3.012**(0.078) |
| | AICc | 2.206(0.082) | 2.131(0.068) | 3.262(0.112) | 3.143(0.081) |
| | BIC | 2.105(0.078) | 2.059(0.065) | 3.077(0.104) | 3.030(0.078) |
| | V-L0LS-CD | 2.122(0.079) | 2.069(0.066) | 3.072(0.104) | 3.026(0.078) |
| | Oracle | 2.035(0.075) | 2.008(0.064) | 3.046(0.103) | 3.012(0.078) |

NOTE: Numbers in parentheses are standard errors. This table is for the case when $(n, p, k) = (200, 2000, 3)$. Smallest values are bolded (other than Oracle's).

**Table 4.** Bias of the estimates of $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ when $\rho_\Sigma = 0.6$.

| | | $m = 2$ | | $m = 3$ | |
|---|---|---|---|---|---|
| | | MSPE | MSMD | MSPE | MSMD |
| $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.031**(0.002) | **2.124**(0.181) | **0.047**(0.003) | **3.109**(0.106) |
| | AICc | 0.218(0.010) | 57.641(218.803) | 0.296(0.013) | 28.919(3.624) |
| | BIC | 0.154(0.008) | 48.798(218.743) | 0.106(0.007) | 9.334(1.809) |
| | V-L0LS-CD | 0.584(0.032) | 87.937(108.601) | 0.124(0.011) | 10.576(2.510) |
| | Oracle | 0.030(0.002) | 1.939(0.061) | 0.045(0.002) | 3.011(0.078) |
| $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.032**(0.002) | **2.156**(0.115) | **0.047**(0.002) | **3.087**(0.120) |
| | AICc | 0.216(0.010) | 26.482(3.876) | 0.291(0.013) | 29.436(3.963) |
| | BIC | 0.138(0.008) | 16.148(2.722) | 0.085(0.006) | 7.158(1.123) |
| | V-L0LS-CD | 0.245(0.017) | 29.253(5.422) | 0.084(0.007) | 7.023(1.585) |
| | Oracle | 0.031(0.002) | 2.026(0.064) | 0.046(0.002) | 3.018(0.078) |
| $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.031**(0.002) | **2.161**(0.269) | **0.048**(0.002) | **3.075**(0.108) |
| | AICc | 0.207(0.009) | 25.380(4.637) | 0.281(0.013) | 27.724(4.629) |
| | BIC | 0.114(0.007) | 13.813(3.518) | 0.085(0.006) | 7.062(1.159) |
| | V-L0LS-CD | 0.153(0.012) | 19.688(9.590) | 0.083(0.006) | 6.422(1.118) |
| | Oracle | 0.030(0.002) | 2.001(0.063) | 0.048(0.002) | 3.016(0.077) |
| $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.040**(0.003) | **2.477**(0.140) | **0.065**(0.006) | **3.533**(0.139) |
| | AICc | 0.224(0.010) | 29.566(14.757) | 0.308(0.014) | 32.441(6.882) |
| | BIC | 0.165(0.009) | 20.749(7.038) | 0.107(0.007) | 9.355(1.648) |
| | V-L0LS-CD | 0.378(0.024) | 51.046(20.985) | 0.111(0.010) | 9.449(2.323) |
| | Oracle | 0.031(0.002) | 2.006(0.065) | 0.046(0.002) | 2.993(0.078) |
| $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.031**(0.002) | **2.080**(0.141) | **0.047**(0.002) | **3.064**(0.081) |
| | AICc | 0.209(0.009) | 26.972(11.373) | 0.285(0.013) | 30.611(9.020) |
| | BIC | 0.125(0.007) | 14.780(3.026) | 0.086(0.006) | 8.094(4.864) |
| | V-L0LS-CD | 0.159(0.011) | 18.284(3.797) | 0.081(0.006) | 6.845(1.665) |
| | Oracle | 0.030(0.002) | 1.982(0.064) | 0.047(0.002) | 3.046(0.079) |
| $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.031**(0.002) | **2.054**(0.069) | **0.048**(0.002) | **3.111**(0.082) |
| | AICc | 0.202(0.009) | 24.395(4.294) | 0.272(0.013) | 28.593(5.347) |
| | BIC | 0.101(0.006) | 12.278(3.367) | 0.074(0.005) | 6.106(0.904) |
| | V-L0LS-CD | 0.112(0.009) | 13.905(4.509) | 0.070(0.005) | 5.381(0.814) |
| | Oracle | 0.031(0.002) | 2.027(0.065) | 0.047(0.002) | 3.096(0.079) |

NOTE: Numbers in parentheses are standard errors. This table is for the case when $(n, p, k) = (200, 2000, 3)$. Smallest values are bolded (other than Oracle's).

**Table 5.** Empirical coverage rates of the prediction ellipsoids for $\mathbf{Y}_i$ when $\rho_\Sigma = 0.6$.

| | | | 90% | 95% | 99% |
|---|---|---|---|---|---|
| $m = 2$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.897**(13.970) | **0.948**(18.176) | **0.989**(27.941) |
| | | AICc | 0.862(13.144) | 0.923(17.162) | 0.980(26.604) |
| | | BIC | 0.875(13.381) | 0.932(17.472) | 0.983(27.083) |
| | | V-L0LS-CD | 0.768(11.603) | 0.846(15.152) | 0.933(23.495) |
| | | Oracle | 0.897(11.790) | 0.949(15.394) | 0.990(23.860) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.898**(19.144) | **0.949**(24.907) | **0.990**(38.289) |
| | | AICc | 0.862(18.063) | 0.923(23.585) | 0.981(36.560) |
| | | BIC | 0.876(18.408) | 0.932(24.035) | 0.984(37.256) |
| | | V-L0LS-CD | 0.814(17.022) | 0.885(22.226) | 0.959(34.459) |
| | | Oracle | 0.898(11.801) | 0.948(15.408) | 0.990(23.883) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.900**(25.661) | **0.950**(33.385) | **0.988**(51.321) |
| | | AICc | 0.866(24.232) | 0.925(31.640) | 0.980(49.048) |
| | | BIC | 0.877(24.681) | 0.933(32.225) | 0.983(49.950) |
| | | V-L0LS-CD | 0.834(23.306) | 0.899(30.433) | 0.965(47.178) |
| | | Oracle | 0.902(11.867) | 0.951(15.494) | 0.990(24.016) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.896**(13.918) | **0.947**(18.108) | **0.989**(27.837) |
| | | AICc | 0.866(12.932) | 0.926(16.886) | 0.981(26.177) |
| | | BIC | 0.875(13.121) | 0.933(17.132) | 0.984(26.556) |
| | | V-L0LS-CD | 0.808(11.967) | 0.879(15.627) | 0.954(24.228) |
| | | Oracle | 0.900(11.848) | 0.949(15.470) | 0.990(23.978) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.897**(17.911) | **0.948**(23.303) | **0.990**(35.823) |
| | | AICc | 0.869(17.041) | 0.929(22.250) | 0.983(34.492) |
| | | BIC | 0.878(17.281) | 0.936(22.564) | 0.986(34.976) |
| | | V-L0LS-CD | 0.852(16.695) | 0.915(21.799) | 0.975(33.793) |
| | | Oracle | 0.898(11.828) | 0.949(15.443) | 0.990(23.936) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.898**(23.091) | **0.948**(30.043) | **0.989**(46.183) |
| | | AICc | 0.872(22.094) | 0.931(28.849) | 0.983(44.720) |
| | | BIC | 0.881(22.425) | 0.937(29.280) | 0.986(45.386) |
| | | V-L0LS-CD | 0.861(21.853) | 0.921(28.534) | 0.979(44.233) |
| | | Oracle | 0.901(11.806) | 0.949(15.415) | 0.990(23.893) |
| $m = 3$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.896**(48.319) | **0.947**(67.534) | **0.989**(118.128) |
| | | AICc | 0.867(45.922) | 0.928(64.578) | 0.983(114.530) |
| | | BIC | 0.890(47.755) | 0.943(67.148) | 0.988(119.064) |
| | | V-L0LS-CD | 0.886(47.465) | 0.941(66.741) | 0.987(118.344) |
| | | Oracle | 0.898(40.453) | 0.949(56.883) | 0.990(100.866) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.897**(66.940) | **0.948**(93.560) | **0.988**(163.651) |
| | | AICc | 0.868(63.875) | 0.929(89.824) | 0.983(159.303) |
| | | BIC | 0.892(66.388) | 0.945(93.349) | 0.988(165.521) |
| | | V-L0LS-CD | 0.891(66.277) | 0.943(93.192) | 0.987(165.244) |
| | | Oracle | 0.898(40.433) | 0.949(56.854) | 0.990(100.815) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.893**(89.786) | **0.945**(125.492) | **0.988**(219.505) |
| | | AICc | 0.866(85.691) | 0.926(120.502) | 0.982(213.711) |
| | | BIC | 0.889(89.092) | 0.942(125.273) | 0.987(222.125) |
| | | V-L0LS-CD | 0.887(88.857) | 0.941(124.943) | 0.987(221.543) |
| | | Oracle | 0.897(40.243) | 0.947(56.587) | 0.989(100.340) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.897**(48.044) | **0.946**(67.150) | **0.988**(117.456) |
| | | AICc | 0.871(44.791) | 0.930(62.987) | 0.983(111.710) |
| | | BIC | 0.890(46.395) | 0.943(65.237) | 0.987(115.679) |
| | | V-L0LS-CD | 0.887(46.115) | 0.940(64.844) | 0.986(114.984) |
| | | Oracle | 0.897(40.249) | 0.948(56.595) | 0.990(100.355) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.899**(62.619) | **0.948**(87.521) | **0.989**(153.089) |
| | | AICc | 0.877(59.949) | 0.934(84.304) | 0.985(149.515) |
| | | BIC | 0.894(61.722) | 0.946(86.791) | 0.989(153.902) |
| | | V-L0LS-CD | 0.894(61.767) | 0.946(86.853) | 0.989(154.012) |
| | | Oracle | 0.898(40.540) | 0.950(57.005) | 0.990(101.083) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.900**(81.677) | **0.950**(114.158) | **0.989**(199.680) |
| | | AICc | 0.879(78.769) | 0.935(110.769) | 0.985(196.453) |
| | | BIC | 0.895(81.045) | 0.948(113.961) | 0.989(202.081) |
| | | V-L0LS-CD | 0.896(81.134) | 0.948(114.086) | 0.989(202.302) |
| | | Oracle | 0.900(40.419) | 0.949(56.835) | 0.990(100.780) |

NOTE: Numbers in parentheses are averaged volumes of the ellipsoids. This table is for the case when $X_{2001}$ was in the true model. Best results are bolded (other than Oracle's).

confirmed these theoretical findings. To the best of our knowledge, this article is one of the first that provides a systemic solution for quantifying uncertainties in the multi-task learning problem.

**Table 6.** Empirical coverage rates of the prediction ellipsoids for $\mathbf{Y}_i$ when $\rho_\Sigma = 0.6$.

| | | | 90% | 95% | 99% |
|---|---|---|---|---|---|
| $m = 2$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.899**(14.002) | **0.945**(18.218) | **0.986**(28.005) |
| | | AICc | 0.866(13.194) | 0.924(17.228) | 0.979(26.706) |
| | | BIC | 0.877(13.425) | 0.931(17.528) | 0.981(27.170) |
| | | V-L0LS-CD | 0.767(11.556) | 0.842(15.091) | 0.931(23.400) |
| | | Oracle | 0.899(11.841) | 0.949(15.460) | 0.990(23.963) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.897**(19.206) | **0.941**(24.987) | **0.979**(38.411) |
| | | AICc | 0.869(18.139) | 0.922(23.685) | 0.972(36.716) |
| | | BIC | 0.879(18.480) | 0.928(24.129) | 0.974(37.402) |
| | | V-L0LS-CD | 0.827(17.132) | 0.889(22.371) | 0.954(34.682) |
| | | Oracle | 0.898(11.817) | 0.949(15.428) | 0.990(23.914) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.896**(25.534) | **0.936**(33.221) | **0.976**(51.068) |
| | | AICc | 0.872(24.114) | 0.919(31.486) | 0.967(48.809) |
| | | BIC | 0.880(24.566) | 0.925(32.075) | 0.970(49.717) |
| | | V-L0LS-CD | 0.844(23.185) | 0.900(30.273) | 0.956(46.932) |
| | | Oracle | 0.902(11.846) | 0.950(15.467) | 0.989(23.974) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.893**(14.645) | **0.943**(19.054) | **0.983**(29.291) |
| | | AICc | 0.861(13.647) | 0.921(17.820) | 0.973(27.623) |
| | | BIC | 0.873(13.891) | 0.928(18.138) | 0.977(28.114) |
| | | V-L0LS-CD | 0.804(12.666) | 0.874(16.540) | 0.949(25.643) |
| | | Oracle | 0.899(11.843) | 0.950(15.463) | 0.990(23.967) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.888**(20.560) | **0.930**(26.749) | **0.970**(41.120) |
| | | AICc | 0.862(19.375) | 0.913(25.298) | 0.961(39.216) |
| | | BIC | 0.872(19.726) | 0.919(25.756) | 0.964(39.923) |
| | | V-L0LS-CD | 0.838(18.720) | 0.895(24.444) | 0.951(37.894) |
| | | Oracle | 0.898(11.843) | 0.949(15.463) | 0.990(23.967) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.887**(27.890) | **0.928**(36.285) | **0.968**(55.779) |
| | | AICc | 0.866(26.324) | 0.913(34.371) | 0.960(53.281) |
| | | BIC | 0.873(26.825) | 0.918(35.025) | 0.963(54.291) |
| | | V-L0LS-CD | 0.852(25.870) | 0.904(33.779) | 0.954(52.363) |
| | | Oracle | 0.901(11.831) | 0.951(15.447) | 0.990(23.943) |
| $m = 3$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.897**(48.363) | **0.945**(67.596) | **0.986**(118.236) |
| | | AICc | 0.868(45.959) | 0.927(64.630) | 0.980(114.622) |
| | | BIC | 0.891(47.810) | 0.941(67.226) | 0.986(119.201) |
| | | V-L0LS-CD | 0.887(47.519) | 0.939(66.818) | 0.985(118.479) |
| | | Oracle | 0.900(40.568) | 0.950(57.043) | 0.990(101.150) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.891**(66.677) | **0.938**(93.192) | **0.981**(163.008) |
| | | AICc | 0.869(63.635) | 0.923(89.486) | 0.974(158.705) |
| | | BIC | 0.887(66.166) | 0.935(93.037) | 0.980(164.967) |
| | | V-L0LS-CD | 0.886(66.120) | 0.936(92.972) | 0.980(164.852) |
| | | Oracle | 0.898(40.204) | 0.948(56.532) | 0.990(100.243) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.891**(89.973) | **0.935**(125.753) | **0.977**(219.961) |
| | | AICc | 0.870(85.825) | 0.921(120.691) | 0.971(214.046) |
| | | BIC | 0.886(89.167) | 0.933(125.379) | 0.977(222.315) |
| | | V-L0LS-CD | 0.885(88.698) | 0.932(124.721) | 0.976(221.152) |
| | | Oracle | 0.897(40.521) | 0.948(56.978) | 0.990(101.034) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.892**(50.567) | **0.940**(70.677) | **0.982**(123.624) |
| | | AICc | 0.864(47.431) | 0.921(66.699) | 0.975(118.292) |
| | | BIC | 0.886(49.327) | 0.936(69.360) | 0.981(122.985) |
| | | V-L0LS-CD | 0.882(49.044) | 0.933(68.963) | 0.980(122.283) |
| | | Oracle | 0.899(40.432) | 0.949(56.853) | 0.990(100.813) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.886**(71.996) | **0.930**(100.626) | **0.972**(176.011) |
| | | AICc | 0.863(68.598) | 0.916(96.465) | 0.965(171.082) |
| | | BIC | 0.882(71.292) | 0.928(100.245) | 0.972(177.749) |
| | | V-L0LS-CD | 0.881(71.255) | 0.927(100.192) | 0.971(177.656) |
| | | Oracle | 0.897(40.432) | 0.948(56.853) | 0.989(100.812) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.882**(98.359) | **0.926**(137.473) | **0.968**(240.462) |
| | | AICc | 0.864(93.865) | 0.914(131.997) | 0.962(234.098) |
| | | BIC | 0.879(97.522) | 0.924(137.128) | 0.968(243.147) |
| | | V-L0LS-CD | 0.873(95.709) | 0.920(134.581) | 0.966(238.650) |
| | | Oracle | 0.898(40.459) | 0.949(56.891) | 0.990(100.879) |

NOTE: Numbers in parentheses are averaged volumes of the ellipsoids. This table is for the case when $X_{1999}X_{2000}$ was in the true model. Best results are bolded (other than Oracle's).

**Table 7.** Empirical coverage rates of the prediction ellipsoids for $\mathbf{Y}_i$ when $\rho_\Sigma = 0.6$.

| | | | 90% | 95% | 99% |
|---|---|---|---|---|---|
| $m = 2$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.877**(15.907) | **0.926**(20.695) | **0.972**(31.813) |
| | | AICc | 0.848(14.913) | 0.904(19.473) | 0.962(30.186) |
| | | BIC | 0.857(15.175) | 0.911(19.814) | 0.965(30.712) |
| | | V-L0LS-CD | 0.738(12.817) | 0.814(16.738) | 0.904(25.956) |
| | | Oracle | 0.896(11.799) | 0.948(15.406) | 0.989(23.879) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.869**(24.338) | **0.912**(31.664) | **0.958**(48.675) |
| | | AICc | 0.846(22.962) | 0.895(29.982) | 0.949(46.477) |
| | | BIC | 0.853(23.372) | 0.901(30.517) | 0.951(47.302) |
| | | V-L0LS-CD | 0.787(20.996) | 0.848(27.417) | 0.920(42.508) |
| | | Oracle | 0.899(11.888) | 0.949(15.522) | 0.989(24.059) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.864**(34.132) | **0.906**(44.407) | **0.952**(68.264) |
| | | AICc | 0.842(32.205) | 0.889(42.051) | 0.942(65.186) |
| | | BIC | 0.848(32.782) | 0.895(42.802) | 0.946(66.346) |
| | | V-L0LS-CD | 0.794(29.721) | 0.851(38.811) | 0.919(60.173) |
| | | Oracle | 0.895(11.825) | 0.946(15.439) | 0.989(23.931) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.880**(16.179) | **0.928**(21.049) | **0.972**(32.358) |
| | | AICc | 0.848(14.970) | 0.905(19.547) | 0.962(30.301) |
| | | BIC | 0.858(15.228) | 0.913(19.883) | 0.965(30.820) |
| | | V-L0LS-CD | 0.784(13.692) | 0.852(17.879) | 0.931(27.721) |
| | | Oracle | 0.899(11.868) | 0.948(15.496) | 0.990(24.018) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.870**(24.345) | **0.914**(31.674) | **0.958**(48.690) |
| | | AICc | 0.844(22.901) | 0.895(29.903) | 0.949(46.354) |
| | | BIC | 0.853(23.334) | 0.901(30.467) | 0.952(47.225) |
| | | V-L0LS-CD | 0.811(21.786) | 0.868(28.447) | 0.933(44.102) |
| | | Oracle | 0.898(11.870) | 0.949(15.498) | 0.989(24.022) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.867**(34.193) | **0.909**(44.486) | **0.954**(68.386) |
| | | AICc | 0.844(32.230) | 0.893(42.083) | 0.945(65.235) |
| | | BIC | 0.851(32.819) | 0.898(42.851) | 0.948(66.421) |
| | | V-L0LS-CD | 0.819(30.951) | 0.872(40.415) | 0.933(62.655) |
| | | Oracle | 0.898(11.842) | 0.949(15.462) | 0.990(23.966) |
| $m = 3$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.878**(55.940) | **0.926**(78.186) | **0.973**(136.759) |
| | | AICc | 0.852(52.748) | 0.908(74.177) | 0.965(131.553) |
| | | BIC | 0.872(54.857) | 0.923(77.135) | 0.972(136.771) |
| | | V-L0LS-CD | 0.869(54.585) | 0.920(76.753) | 0.970(136.096) |
| | | Oracle | 0.896(40.527) | 0.948(56.987) | 0.989(101.050) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.867**(85.347) | **0.915**(119.287) | **0.963**(208.652) |
| | | AICc | 0.846(81.355) | 0.899(114.405) | 0.954(202.898) |
| | | BIC | 0.863(84.513) | 0.912(118.835) | 0.962(210.711) |
| | | V-L0LS-CD | 0.859(83.770) | 0.910(117.792) | 0.960(208.868) |
| | | Oracle | 0.897(40.463) | 0.949(56.897) | 0.989(100.890) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.864**(120.068) | **0.909**(167.817) | **0.957**(293.537) |
| | | AICc | 0.845(114.340) | 0.894(160.789) | 0.948(285.162) |
| | | BIC | 0.860(118.867) | 0.906(167.141) | 0.955(296.365) |
| | | V-L0LS-CD | 0.823(109.145) | 0.878(153.495) | 0.938(272.270) |
| | | Oracle | 0.896(40.262) | 0.947(56.614) | 0.989(100.388) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.877**(56.435) | **0.926**(78.878) | **0.974**(137.969) |
| | | AICc | 0.852(52.493) | 0.908(73.818) | 0.964(130.917) |
| | | BIC | 0.871(54.679) | 0.923(76.884) | 0.972(136.326) |
| | | V-L0LS-CD | 0.867(54.482) | 0.919(76.608) | 0.971(135.839) |
| | | Oracle | 0.897(40.614) | 0.948(57.109) | 0.990(101.267) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.869**(85.780) | **0.915**(119.892) | **0.962**(209.710) |
| | | AICc | 0.847(81.396) | 0.900(114.463) | 0.953(203.002) |
| | | BIC | 0.865(84.606) | 0.912(118.966) | 0.961(210.944) |
| | | V-L0LS-CD | 0.863(84.212) | 0.912(118.413) | 0.960(209.968) |
| | | Oracle | 0.898(40.498) | 0.949(56.946) | 0.990(100.978) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.863**(119.907) | **0.910**(167.591) | **0.957**(293.143) |
| | | AICc | 0.842(114.088) | 0.894(160.436) | 0.948(284.535) |
| | | BIC | 0.859(118.530) | 0.907(166.667) | 0.956(295.525) |
| | | V-L0LS-CD | 0.828(110.637) | 0.885(155.591) | 0.942(275.973) |
| | | Oracle | 0.897(40.443) | 0.948(56.868) | 0.990(100.839) |

NOTE: Numbers in parentheses are averaged volumes of the ellipsoids. This table is for the case when $X^2_{2000}$ was in the true model. Best results are bolded (other than Oracle's).

**Table 8.** Empirical coverage rates of the prediction ellipsoids for $\mathbf{Y}_i$ when $\rho_\Sigma = 0.6$.

| | | | 90% | 95% | 99% |
|---|---|---|---|---|---|
| $m = 2$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.877**(15.980) | **0.926**(20.791) | **0.971**(31.960) |
| | | AICc | 0.845(14.961) | 0.904(19.535) | 0.962(30.282) |
| | | BIC | 0.856(15.241) | 0.911(19.900) | 0.965(30.845) |
| | | V-L0LS-CD | 0.744(12.951) | 0.817(16.913) | 0.908(26.226) |
| | | Oracle | 0.899(11.837) | 0.949(15.455) | 0.990(23.955) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.871**(24.263) | **0.915**(31.567) | **0.959**(48.526) |
| | | AICc | 0.845(22.859) | 0.895(29.847) | 0.949(46.268) |
| | | BIC | 0.854(23.284) | 0.902(30.402) | 0.952(47.124) |
| | | V-L0LS-CD | 0.785(20.895) | 0.848(27.285) | 0.919(42.304) |
| | | Oracle | 0.899(11.849) | 0.949(15.471) | 0.990(23.981) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.866**(33.970) | **0.908**(44.196) | **0.954**(67.939) |
| | | AICc | 0.844(32.071) | 0.891(41.876) | 0.944(64.914) |
| | | BIC | 0.851(32.693) | 0.897(42.687) | 0.948(66.166) |
| | | V-L0LS-CD | 0.792(29.422) | 0.850(38.420) | 0.918(59.569) |
| | | Oracle | 0.900(11.864) | 0.950(15.491) | 0.989(24.011) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.878**(16.057) | **0.927**(20.891) | **0.973**(32.115) |
| | | AICc | 0.846(14.870) | 0.903(19.416) | 0.962(30.098) |
| | | BIC | 0.855(15.138) | 0.911(19.766) | 0.966(30.638) |
| | | V-L0LS-CD | 0.780(13.580) | 0.850(17.733) | 0.930(27.494) |
| | | Oracle | 0.896(11.827) | 0.948(15.442) | 0.990(23.935) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.866**(24.129) | **0.911**(31.393) | **0.957**(48.258) |
| | | AICc | 0.841(22.656) | 0.893(29.583) | 0.947(45.858) |
| | | BIC | 0.849(23.072) | 0.899(30.125) | 0.951(46.695) |
| | | V-L0LS-CD | 0.814(21.722) | 0.871(28.363) | 0.934(43.971) |
| | | Oracle | 0.897(11.843) | 0.948(15.462) | 0.989(23.967) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.867**(34.029) | **0.909**(44.273) | **0.953**(68.059) |
| | | AICc | 0.843(32.010) | 0.892(41.796) | 0.944(64.790) |
| | | BIC | 0.851(32.595) | 0.898(42.558) | 0.947(65.967) |
| | | V-L0LS-CD | 0.824(30.880) | 0.876(40.321) | 0.933(62.509) |
| | | Oracle | 0.897(11.825) | 0.947(15.440) | 0.989(23.931) |
| $m = 3$ | $b = 1/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.876**(55.576) | **0.926**(77.678) | **0.973**(135.870) |
| | | AICc | 0.851(52.457) | 0.908(73.767) | 0.965(130.827) |
| | | BIC | 0.871(54.536) | 0.923(76.684) | 0.972(135.971) |
| | | V-L0LS-CD | 0.868(54.342) | 0.922(76.411) | 0.971(135.488) |
| | | Oracle | 0.897(40.370) | 0.948(56.766) | 0.990(100.658) |
| | $b = 2/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.867**(85.171) | **0.913**(119.042) | **0.960**(208.223) |
| | | AICc | 0.846(81.072) | 0.898(114.006) | 0.952(202.192) |
| | | BIC | 0.863(84.335) | 0.911(118.584) | 0.959(210.266) |
| | | V-L0LS-CD | 0.861(83.831) | 0.909(117.877) | 0.959(209.017) |
| | | Oracle | 0.899(40.410) | 0.950(56.822) | 0.991(100.758) |
| | $b = 3/\sqrt{3}\ \rho_X = 0$ | GMTask | **0.865**(120.515) | **0.911**(168.441) | **0.958**(294.629) |
| | | AICc | 0.846(114.814) | 0.896(161.456) | 0.949(286.345) |
| | | BIC | 0.861(119.453) | 0.909(167.965) | 0.957(297.824) |
| | | V-L0LS-CD | 0.824(109.771) | 0.880(154.375) | 0.941(273.828) |
| | | Oracle | 0.897(40.525) | 0.949(56.984) | 0.990(101.044) |
| | $b = 1/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.875**(55.923) | **0.925**(78.162) | **0.972**(136.718) |
| | | AICc | 0.849(52.076) | 0.906(73.232) | 0.963(129.878) |
| | | BIC | 0.868(54.216) | 0.921(76.233) | 0.971(135.172) |
| | | V-L0LS-CD | 0.864(54.093) | 0.917(76.061) | 0.969(134.868) |
| | | Oracle | 0.895(40.331) | 0.947(56.712) | 0.989(100.562) |
| | $b = 2/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.868**(85.594) | **0.915**(119.633) | **0.962**(209.256) |
| | | AICc | 0.846(81.065) | 0.900(113.997) | 0.953(202.176) |
| | | BIC | 0.864(84.385) | 0.913(118.654) | 0.961(210.391) |
| | | V-L0LS-CD | 0.862(84.065) | 0.911(118.205) | 0.960(209.598) |
| | | Oracle | 0.898(40.284) | 0.949(56.645) | 0.990(100.444) |
| | $b = 3/\sqrt{3}\ \rho_X = 0.5$ | GMTask | **0.866**(119.389) | **0.911**(166.867) | **0.958**(291.876) |
| | | AICc | 0.845(113.461) | 0.896(159.554) | 0.949(282.971) |
| | | BIC | 0.861(117.913) | 0.908(165.799) | 0.957(293.986) |
| | | V-L0LS-CD | 0.832(115.545) | 0.886(165.584) | 0.943(318.020) |
| | | Oracle | 0.899(40.427) | 0.949(56.845) | 0.990(100.799) |

NOTE: Numbers in parentheses are averaged volumes of the ellipsoids. This table is for the case when $X_1^2$ was in the true model. Best results are bolded (other than Oracle's).

**Table 9.** Empirical coverage rates of the prediction ellipsoids for the polymerase chain reaction dataset. Numbers in parentheses are averaged volumes of the ellipsoids.

| | 90% | 95% | 99% |
|---|---|---|---|
| GMTask | 0.900(8.087) | 0.950(10.521) | 0.983(16.174) |
| AICc | 0.800(4.894) | 0.850(6.462) | 0.900(10.281) |
| BIC | 0.817(5.109) | 0.833(6.740) | 0.933(10.704) |
| V-L0LS-CD | 0.683(4.545) | 0.783(6.023) | 0.833(9.670) |

## Appendix A. Proof of Theorem 1

*Proof.* Since

$$r(M) \propto \Gamma_m(\tfrac{n-|M|-m}{2})|\pi \boldsymbol{S}_M|^{-\frac{n-|M|-m}{2}}[(n-|M|)m]^{\frac{1}{2}} \times q^{|M|},$$

where $q = n^{-\frac{m}{2}}p^{-1}$ is derived by using the MDL principle. Let $k = |M|$ and $k_0 = |M_0|$ for simplicity, thus, we have

$$\frac{r(M)}{r(M_0)} = \exp\{-T_1 - T_2\},$$

where

$$T_1 = \frac{n-k-m}{2}\log\left(\frac{|\hat{\boldsymbol{\Sigma}}_M|}{|\hat{\boldsymbol{\Sigma}}_{M_0}|}\right)$$

and

$$T_2 = \tfrac{k_0-k}{2}\log(|\pi \boldsymbol{S}_{M_0}|) + \log\left\{\Gamma_m(\tfrac{n-k_0-m}{2})/\,\Gamma_m(\tfrac{n-k-m}{2})\right\} + (k_0-k)\log(q) + \tfrac{1}{2}\log\tfrac{n-k_0}{n-k}.$$

As $n \to \infty$, we can see that $\hat{\boldsymbol{\Sigma}}_{M_0} \xrightarrow{P} \boldsymbol{\Sigma}_0$ when $M \notin \mathcal{M}_-$ and $\hat{\boldsymbol{\Sigma}}_M \xrightarrow{P} \boldsymbol{\Sigma}_0^{\frac{1}{2}}\boldsymbol{\Psi}_M\boldsymbol{\Sigma}_0^{\frac{1}{2}} + \boldsymbol{\Sigma}_0$ when $M \in \mathcal{M}_-$, where $\boldsymbol{\Psi}_M = \lim_{n\to\infty}\frac{1}{n}\Delta(M)$. By our assumption, $\boldsymbol{\Psi}_M$ is a positive semidefinite matrix.

**Case 1:** $\forall M \in \mathcal{M}_-$.

We have

$$\frac{T_1}{n} \xrightarrow{P} \frac{1}{2}\log|\boldsymbol{\Sigma}_0^{\frac{1}{2}}\boldsymbol{\Psi}_M\boldsymbol{\Sigma}_0^{\frac{1}{2}} + \boldsymbol{\Sigma}_0| - \frac{1}{2}\log|\boldsymbol{\Sigma}_0| = \frac{1}{2}\log|\boldsymbol{\Psi}_M + \boldsymbol{I}_m| > 0. \tag{20}$$

Since $\boldsymbol{S}_{M_0}$ is distributed as $W_m(n-k_0, \Sigma_0)$, we can derive the log-expectation and log-variance as

$$E\big[\log|\boldsymbol{S}_{M_0}|\big] = \psi_m(\frac{n-k_0}{2}) + m\log 2 + \log|\Sigma_0|$$

and

$$\mathrm{var}\big[\log|\boldsymbol{S}_{M_0}|\big] = \sum_{i=1}^{m}\psi_1\left(\frac{n-k_0+1-i}{2}\right),$$

where $\psi_m$ is the multivariate digamma function; that is, the derivative of the log of the multivariate gamma function, and $\psi_1$ is the trigamma function. Then we have

$$\log|\boldsymbol{S}_{M_0}| = m\log(n-k_0)(1+o_p(1)) = m\log n(1+o_p(1)).$$

By the definition of multivariate gamma function $\Gamma_p$, we have

$$\Gamma_p(a) = \pi^{\frac{p(p-1)}{4}}\prod_{j=1}^{p}\Gamma\left(a + \frac{1-j}{2}\right).$$

According to Stirling's approximation,

$$\log\left\{\Gamma_m(\tfrac{n-k_0-m}{2})/\,\Gamma_m(\tfrac{n-k-m}{2})\right\} = \tfrac{m(k-k_0)}{2}\log n(1+o(1)).$$

Therefore, we have

$$T_2 = \frac{m(k-k_0)}{2}\left\{\log n(o_p(1)) - \log\pi - \frac{\log(q^2)}{m}\right\} + \frac{1}{2}\log\frac{n-k_0}{n-k} \tag{21}$$

and

$$\lim_{n\to\infty}\frac{T_2}{n} = 0. \tag{22}$$

By (20) and (22), for case 1, we have

$$\min_{M\in\mathcal{M}_-} T_1 + T_2 \to \infty.$$

**Case 2:** $\forall M \in \mathcal{M}_+$.

Let $\boldsymbol{V}$ and $\boldsymbol{Z}_M$ be the $m\times m$ and the $k\times m$ matrices defined by

$$\boldsymbol{V} = \frac{1}{\sqrt{n}}(\boldsymbol{U}^T\boldsymbol{U} - n\boldsymbol{I}_m), \boldsymbol{Z}_M = (\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-\frac{1}{2}}\boldsymbol{X}_M^T\boldsymbol{U},$$

where

$$\boldsymbol{U} = (\boldsymbol{Y} - \boldsymbol{X}_0\boldsymbol{B}_0)\boldsymbol{\Sigma}_0^{-\frac{1}{2}}.$$

We know that $\boldsymbol{V}$ has an asymptotic normality as $n \to \infty$ and $\boldsymbol{Z}_M \sim N_{k\times m}(\boldsymbol{0}_{k\times m}, \boldsymbol{I}_{km})$.

Furthermore, since

$$\boldsymbol{\Sigma}_0^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_M\boldsymbol{\Sigma}_0^{-\frac{1}{2}} = \tfrac{1}{n}\boldsymbol{U}^T(\boldsymbol{I}_n - \boldsymbol{H}_M)\boldsymbol{U} = \tfrac{1}{n}(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{Z}_M^T\boldsymbol{Z}_M),$$

we have

$$\boldsymbol{\Sigma}_0^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_M\boldsymbol{\Sigma}_0^{-\frac{1}{2}} = \boldsymbol{I}_m + \frac{1}{\sqrt{n}}\boldsymbol{V} - \frac{1}{n}\boldsymbol{Z}_M^T\boldsymbol{Z}_M. \tag{23}$$

By using (23), $n\log|\hat{\boldsymbol{\Sigma}}_M|$ can be expanded as

$$n\log|\hat{\boldsymbol{\Sigma}}_M| = n\log|\boldsymbol{\Sigma}_0| + \sqrt{n}\mathrm{tr}(\boldsymbol{V}) - \left(\frac{1}{2}\mathrm{tr}(\boldsymbol{V}^2) + \mathrm{tr}(\boldsymbol{Z}_M^T\boldsymbol{Z}_M)\right) + O_p(n^{-\frac{1}{2}}).$$

Then, we derive

$$T_1 = -\tfrac{n-k-m}{2n}\big(\mathrm{tr}(\boldsymbol{Z}_M^T\boldsymbol{Z}_M) - \mathrm{tr}(\boldsymbol{Z}_{M_0}^T\boldsymbol{Z}_{M_0}) + O_p(n^{-\frac{1}{2}})\big).$$

By (21), for all $M \in \mathcal{M}_+$, $\lim_{n\to\infty} T_2 = \infty$. Then for case 2, we have

$$\min_{M\in\mathcal{M}_+} T_1 + T_2 \to \infty.$$

Combing case 1 and case 2, we can show that

$$\max_{M\neq M_0, M\in\mathcal{M}}\frac{r(M)}{r(M_0)} \to 0.$$

Moreover, if condition (A3) holds and $M \neq M_0$, we have

$$\sum_{M\in\mathcal{M}^*}\frac{r(M)}{r(M_0)} \le \sum_{j=1}^{ck_0}\sum_{\mathcal{M}_j^*}\frac{r(M)}{r(M_0)} \le \sum_{j=1}^{ck_0}|\mathcal{M}_j^*|\max_{M\in\mathcal{M}_j^*}\frac{r(M)}{r(M_0)} \to 0.$$

This completes the proof for Theorem 1. □

## Supplementary Materials

The supplementary materials provide additional numerical results. Computing code implementing the proposed method can be obtained from the journal's website.

## Acknowledgments

The authors are most grateful to the reviewers for their most constructive and helpful comments that led to a much improved version of the article.

## Funding

The authors acknowledge the support by the National Science Foundation under grants DMS-1811405, DMS-1811661, DMS-1916125, CCF-1934568, and DMS-2113605.

## Disclosure Statement

The authors report there are no competing interests to declare.

## ORCID

Thomas C. M. Lee https://orcid.org/0000-0001-7067-405X

## References

Bedrick, E. J., and Tsai, C.-L. (1994), "Model Selection for Multivariate Regression in Small Samples," *Biometrics*, 50, 226–231. [5]

Bondell, H. D., and Reich, B. J. (2012), "Consistent High-Dimensional Bayesian Variable Selection via Penalized Credible Regions," *Journal of the American Statistical Association*, 107, 1610–1624. [6]

Caruana, R. (1997), "Multitask LOearning," *Machine Learning*, 28, 41–75. [1]

Dempster, A. P. (2008), "The Dempster–Shafer Calculus for Statisticians," *International Journal of Approximate Reasoning*, 48, 365–377. [2]

Evgeniou, A., and Pontil, M. (2007), "Multi-Task Feature Learning," in *Advances in Neural Information Processing Systems* (Vol. 19), p. 41. [1]

Fan, J., Guo, S., and Hao, N. (2012), "Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression," *Journal of the Royal Statistical Society*, Series B, 74, 37–65. [5]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [3]

Fisher, R. A. (1930), "Inverse Probability," in *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 26), pp. 528–535, Cambridge University Press. [2]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [4]

Hannig, J. (2009), "On Generalized Fiducial Inference," *Statistica Sinica*, 19, 491–544. [2]

Hannig, J., Iyer, H., Lai, R. C., and Lee, T. C. M. (2016), "Generalized Fiducial Inference: A Review and New Results," *Journal of the American Statistical Association*, 111, 1346–1361. [2,4,5]

Hannig, J., and Lee, T. C. M. (2009), "Generalized Fiducial Inference for Wavelet Regression," *Biometrika*, 96, 847–860. [2,3]

Johnson, R. A., and Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis*, Upper Saddle River, NJ: Prentice Hall. [4,6]

Koner, S., and Williams, J. P. (2021), "The EAS Approach to Variable Selection for Multivariate Response Data in High-Dimensional Settings," arXiv preprint arXiv:2107.04873. [2]

Lai, R. C. S., Hannig, J., and Lee, T. C. M. (2015), "Generalized Fiducial Inference for Ultrahigh-Dimensional Regression," *Journal of the American Statistical Association*, 110, 760–772. [2]

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T.-K., Flowers, M. T., Schueler, K. L., Manly, K. F., Williams, R. W., Kendziorski, C., and Attie, A. D. (2006), "Combined Expression Trait Correlations and Expression Quantitative Trait Locus Mapping," *PLoS Genetics*, 2, e6. [6]

Liu, H., Wang, L., and Zhao, T. (2014), "Multivariate Regression with Calibration," in *Advances in Neural Information Processing Systems* (Vol. 27), p. 5630. [1]

Martin, R., and Liu, C. (2015), *Inferential Models: Reasoning with Uncertainty* (Vol. 145), Boca Raton, FL: CRC Press. [2]

Misra, I., Shrivastava, A., Gupta, A., and Hebert, M. (2016), "Cross-Stitch Networks for Multi-Task Learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3994–4003. [1]

Obozinski, G., Taskar, B., and Jordan, M. (2006), "Multi-Task Feature Selection," *Statistics Department, UC Berkeley*, , Technical report 2, 2. [1,4]

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2008), "Union Support Recovery in High-Dimensional Multivariate Regression," in *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, IEEE, pp. 21–26. [5]

Rissanen, J. (1989), *Stochastic Complexity in Statistical Inquiry*, Singapore: World Scientific. [3]

——— (2007), *Information and Complexity in Statistical Modeling*, New York: Springer. [3]

Schweder, T., and Hjort, N. L. (2016), *Confidence, Likelihood, Probability* (Vol. 41), Cambridge: Cambridge University Press. [2]

Seneviratne, A. J., and Solo, V. (2012), "On Vector $l_0$ Penalized Multivariate Regression," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3613–3616. [1,5]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1]

Triantafyllopoulos, K. (2011), "Real-Time Covariance Estimation for the Local Level Model," *Journal of Time Series Analysis*, 32, 93–107. [4]

Weerahandi, S. (1995), "Generalized Confidence Intervals," in *Exact Statistical Methods for Data Analysis*, Springer, pp. 143–168. [2]

Williams, J. P., and Hannig, J. (2019), "Nonpenalized Variable Selection in High-Dimensional Linear Model Settings via Generalized Fiducial Inference," *The Annals of Statistics*, 47, 1723–1753. [2]

Williams, J. P., Xie, Y., and Hannig, J. (2019), "The EAS Approach for Graphical Selection Consistency in Vector Autoregression Models," *Canadian Journal of Statistics* (to appear). [2]

Xie, M.-g., and Singh, K. (2013), "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review," *International Statistical Review*, 81, 3–39. [2]

Yanagihara, H., Wakaki, H., Fujikoshi, Y., et al. (2015), "A Consistency Property of the AIC for Multivariate Linear Models when the Dimension and the Sample Size are Large," *Electronic Journal of Statistics*, 9, 869–897. [4]

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007), "Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression," *Journal of the Royal Statistical Society*, Series B, 69, 329–346. [1]

Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014), "Facial Landmark Detection by Deep Multi-Task Learning," in *European Conference on Computer Vision*, Springer, pp. 94–108. [1]