# Multi-Task Learning with Gaussian Matrix Generalized Inverse Gaussian Model

**Ming Yang**[†] CAUCHYM@ZJU.EDU.CN
**Yingming Li**[†] LIYMN@ZJU.EDU.CN
**Zhongfei (Mark) Zhang** ZHONGFEI@ZJU.EDU.CN
Department of Information Science and Electronic Engineering, Zhejiang University, China

## Abstract

In this paper, we study the multi-task learning problem with a new perspective of considering the structure of the residue error matrix and the low-rank approximation to the task covariance matrix simultaneously. In particular, we first introduce the Matrix Generalized Inverse Gaussian (MGIG) prior and define a Gaussian Matrix Generalized Inverse Gaussian (GMGIG) model for low-rank approximation to the task covariance matrix. Through combining the GMGIG model with the residual error structure assumption, we propose the GMGIG regression model for multi-task learning. To make the computation tractable, we simultaneously use variational inference and sampling techniques. In particular, we propose two sampling strategies for computing the statistics of the MGIG distribution. Experiments show that this model is superior to the peer methods in regression and prediction.

## 1. Introduction

With the research on multiple task learning for decades (Thrun, 1996; Caruana, 1997; Baxter, 2000), recent years have witnessed the increasing applications of multi-task learning in many fields ranging from classification of protein in bioinformatics to event evolution in cross media due to its capability of transferring the knowledge discovered in one task to the other relevant tasks.

[†]Contributed equally

*Proceedings of the 30$^{th}$ International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. 

An increasing number of efforts on multi-task learning lie in discovering the relationship among the tasks either by directly learning the relatedness of the tasks (Xue et al., 2007; Yu et al., 2007; Jacob et al., 2008) or by mining the common feature structures shared by the tasks (Ando & Zhang, 2005; Zhang et al., 2005; Argyriou et al., 2006; Obozinski et al., 2009; Chen et al., 2009; Rai & Daumé III, 2010), which is equivalent to estimating a matrix model parameter with its rows corresponding to the tasks or its columns corresponding to the features. Therefore, discovering the relationship among tasks corresponds to learning the relationship among the rows and mining the feature structure corresponds to learning the structure of the columns.

Given the variety of configurations of the multi-task learning, it is convenient to describe the multi-task learning problem as a multiple output regression problem as follows where each task produces an output for its corresponding input

$$Y = WX + \mu \mathbf{1}_N^T + \epsilon \tag{1}$$

where $Y = (y_1, \ldots, y_N) \in \mathbb{R}^{d\times N}$ is the correspondence matrix of $N$ samples under $d$ tasks; $X = (x_1, \ldots, x_N) \in \mathbb{R}^{D\times N}$ is the observation matrix of $N$ samples with $D$ features. $W \in \mathbb{R}^{d\times D}$ is the weight matrix or regression matrix. $\mu \in \mathbb{R}^d$ is the offset vector for the tasks, $\mathbf{1}_N$ is an $N$-dimension column vector with all the elements being 1. $\epsilon$ is the residue error matrix with matrix variate normal density (matrix variate Gaussian density) $\mathcal{N}_{d,N}(0, \Sigma_1 \otimes \Sigma_2)$. Herein, $\otimes$ is the Kronecker product of two matrices and we employ the notation from (Gupta & Nagar, 2000) that $\Sigma_1(d \times d)$ and $\Sigma_2(N \times N)$ are positive definite matrices describing the correlations of rows and columns, respectively. Typically, we assume that samples are independent, i.e., $\Sigma_2$ is an identity matrix. $\Sigma_1$, however, is required to be non-trivial and describes the correlation of the tasks in the residue error matrix $\epsilon$.

The regularization method is widely used in multi-task learning to discover the relationship among the tasks (Evgeniou & Pontil, 2004; Argyriou et al., 2007; Agarwal et al., 2010; Jenatton et al., 2011). Argyriou et al. (2006) adopt the $\ell_{2,1}$ matrix norm as the penalty for the weight matrix to learn the feature structure shared across the tasks. Chen et al. (2011) combine the nuclear norm and the $\ell_{p,q}$ norm for $W$ to consider the sparsity of $W$ and the correlation among the tasks simultaneously. The above norm penalties focus on involving the positive (or zero) correlation among the tasks, but fail to establish the negative correlation among the tasks.

To address this issue, Bonilla et al. (2007) model the covariance matrix for the tasks by not only the positive (or zero) correlation but also the negative correlation among the tasks. In this manner, a hierarchical correlated model (Zhang & Yeung, 2010; Zhang & Schneider, 2010) is established for $W$; then the problem of simultaneously discovering the positive and negative correlations among the tasks is reduced to the problem of estimating the task covariance matrix of $W$ under the various constraints.

Despite the advantage of the above models on establishing the complete correlation among the tasks, the computational complexity increases with the scale of the tasks since the covariance matrix of the tasks is often learned in a nonparametric manner. To solve this problem, Bonilla et al. (2007) resort to a low-rank approximation to the task covariance matrix. Archambeau et al. (2011) adopt this approximation by decomposing the weight matrix to the product of the projection matrix and the latent matrix. While this scheme partially addresses the computational concern, it ignores part of the structure to be learned, for example, the residual error structure, which is very important for multiple regression (Breiman & Friedman, 1997; Kim & Xing, 2010; Rothman et al., 2010; Sohn & Kim, 2012).

In this paper, we propose a Bayesian model to study the multi-task learning problem with a new perspective of considering the residual error structure and the low-rank approximation to the task covariance matrix simultaneously. Instead of a nonparametric modeling, we model the task covariance matrix as a random variable with the Matrix Generalized Inverse Gaussian (MGIG) prior. This prior is able to degenerate to a series of common priors, such as Wishart and inverse Wishart prior, either of which is often used as the covariance matrix prior. In particular, a Gaussian Matrix Generalized Inverse Gaussian (GMGIG) model is developed first for the low rank structure of the covariance matrices. Then we combine it with the residual error structure assumption to obtain the GMGIG regression model for multi-task learning. To estimate the parameters in the GMGIG regression model, we propose two sampling methods in the inference for numerical estimation on the statistics of the MGIG distribution. Finally, we report experimental evaluations for the model, and compare it with the peer methods in the literature to demonstrate the effectiveness and promise of the GMGIG regression model for multi-task learning.

## 2. MGIG Prior and GMGIG Model

### 2.1. MGIG prior

The MGIG distribution is introduced from the Generalized Inverse Gaussian (GIG) distribution (Barndorff-Nielsen et al., 1982; Zhang et al., 2012) and is formally proposed by Butler (1998). We denote $\mathbb{S}_+^p$ as the cone of the $p \times p$ positive definite matrices. Let $\Psi,\ \Phi \in \mathbb{S}_+^p$ and $\nu \in \mathbb{R}$.[1] A matrix random variable $G \in \mathbb{S}_+^p$ is MGIG distributed and is denoted as $G \sim \mathcal{MGIG}_p(\Psi, \Phi, \nu)$ if the density of $G$ is

$$\frac{|G|^{\nu-(p+1)/2}}{|\frac{\Psi}{2}|^{\nu} B_\nu(\frac{\Phi}{2}\frac{\Psi}{2})} \operatorname{etr}\left(-\frac{1}{2}\Psi G^{-1} - \frac{1}{2}\Phi G\right) \qquad (2)$$

where $\operatorname{etr}(\cdot) \triangleq \exp \operatorname{Tr}(\cdot)$ is an operator mapping a matrix to the exponent of its trace. $B_\nu(\cdot)$ is the matrix Bessel function defined by (Herz, 1955). The MGIG distribution can easily degenerate to Wishart distribution and inverse Wishart distribution. In 1-dimension case, $B_\nu(\cdot)$ degenerates to Matérn class function[2](Stein, 1999) and the MGIG distribution degenerates to the GIG distribution, further to $\Gamma$ distribution, inverse $\Gamma$ distribution, etc.

In light of the flexibility of the MGIG distribution, we are able to mix a probabilistic model with MGIG prior and obtain various posterior densities, making the regression and prediction more robust.

### 2.2. GMGIG model

We intend to assign the MGIG prior to the covariance matrices of the weight matrix $W$ accommodating data with various characteristics. Hence, we define a statistical model to describe the relationship between the parameters.

[1] $\Psi$ and $\Phi$ can be positive semidefinite according to the value of $\nu$ (Butler, 1998).

[2] For $p = 1$, we have $B_{-\nu}(z^2/4) = 2^{1-\nu} z^\nu K_\nu(z)$. $K_\nu(\cdot)$ is the modified Bessel function of the second kind and the right hand side of the above equation belongs to Matérn class function when $\nu > 0$.

**Definition.** *Define matrices* $W \in \mathbb{R}^{d\times D}$, $V, V_0 \in \mathbb{R}^{d\times K}$, $Z, Z_0 \in \mathbb{R}^{K\times D}$, $\Omega, \Psi_1, \Phi_1 \in \mathbb{S}_+^D$, *and* $\Sigma, \Psi_2, \Phi_2 \in \mathbb{S}_+^d$. *The GMGIG model is a series of dependent random variables satisfying that*

$$\begin{aligned} W &\sim \mathcal{N}_{d,D}(VZ, \Sigma \otimes \Omega) \\ V &\sim \mathcal{N}_{d,K}(V_0, \Sigma \otimes \kappa_1 I_K) \\ Z &\sim \mathcal{N}_{K,D}(Z_0, \kappa_2 I_K \otimes \Omega) \\ \Omega &\sim \mathcal{MGIG}_D(\Psi_1, \Phi_1, \nu_1) \\ \Sigma &\sim \mathcal{MGIG}_d(\Psi_2, \Phi_2, \nu_2) \end{aligned}$$

*where* $\kappa_1, \kappa_2 > 0$ *and* $\nu_1, \nu_2 \in \mathbb{R}$.

In the definition above, $W$ follows matrix variate Gaussian distribution and its covariance matrices follow MGIG distribution; that is why we call it GMGIG model. The mean of $W$ is decomposed into the product of the projection matrix $V$ and the latent matrix $Z$ with $K(< D)$ high relevance directions (Rasmussen & Williams, 2006). Through this decomposition, we are able to obtain a low-rank approximation to the covariance matrices. The GMGIG model can easily degenerate to Gaussian Inverse Wishart (GIW) model (Le & Zidek, 2006) by setting $\kappa_1 = 0$, $\Phi_1 = 0_D$, $\nu_1 < -\frac{D-1}{2}$, and fixing $\Sigma$ to a constant positive definite matrix. This setting is direct if we consider a "non-central" version of the MGIG random matrix $\Sigma - \Sigma_0$ and make $\Psi_2, \Phi_2 \to \infty$; then $\Sigma \to \Sigma_0$. We do not involve such design which may complicate the model; further we set $V_0$ and $Z_0$ to a null matrix for simplicity. From the GMGIG model, we derive the marginal distribution $p(W|V, \Sigma, \Psi_1, \Phi_1, \nu_1)$ as

$$\begin{aligned} &\int \mathcal{N}(W|VZ, \Sigma \otimes \Omega)\mathcal{N}(Z|0, \kappa_2 I_K \otimes \Omega)p(\Omega)dZd\Omega \\ =&\int \mathcal{N}(W|0, \tilde{\Sigma} \otimes \Omega)\mathcal{MGIG}(\Omega|\Psi_1, \Phi_1, \nu_1)d\Omega \\ =&\frac{|\tilde{\Sigma}|^{-D/2}|\Psi_1|^{-d/2}}{\pi^{Dd/2}B_{\nu_1}(\frac{\Phi_1}{2}\frac{\Psi_1}{2})} \cdot |I_D + \Psi_1^{-1}W^T\tilde{\Sigma}^{-1}W|^{\nu_1 - d/2} \\ &\cdot B_{\nu_1 - d/2}\left(\frac{\Phi_1}{2}\frac{\Psi_1 + W^T\tilde{\Sigma}^{-1}W}{2}\right) \end{aligned} \tag{3}$$

where $\tilde{\Sigma} \triangleq \kappa_2 VV^T + \Sigma$. The marginal distribution of $W$ is the Matrix variate Generalized Hyperbolic (MGH) distribution (Butler, 1998). This distribution is mixed by matrix Gaussian distribution and MGIG distribution. It is noted that $\tilde{\Sigma}$ is decomposed as the sum of the original row covariance matrix $\Sigma$ and a low-rank matrix product $VV^T$; hence, $\tilde{\Sigma}$ actually correlates the tasks in the weight matrix $W$ and its low-rank approximation helps identify the high relevance directions for a large number of tasks and further helps reduce the computational complexity in the inference.

The reason to highlight the MGH distribution is that it contains a family of the distributions including matrix variate T distribution, matrix Laplacian distribution, matrix Bessel distribution, and multivariate Pearson type VII distribution. We list two typical degenerations of the MGH distribution:

1. Let $-\nu_1 > \frac{D-1}{2}$ and $\Phi_1 = 0$, then

   $$W|V, \Sigma, \Psi_1, \nu_1 \sim \mathcal{T}_{d,D}(-2\nu_1 - D + 1, 0, \tilde{\Sigma}, \Psi_1)$$

   is matrix variate T distribution according to Section 4.2 of (Gupta & Nagar, 2000).

2. Let $\nu_1 > \frac{D-1}{2}$ and $\Psi_1 = 0$, then

   $$W|V, \Sigma, \Phi_1, \nu_1 \sim \mathcal{MBS}_{d,D}(\nu_1, \tilde{\Sigma}, \Phi_1)$$

   is Matrix variate Bessel (MBS) distribution or matrix variate Variance-Gamma distribution. When $D = 1$, the MBS distribution degenerates to multivariate Bessel distribution as in (Kotz et al., 2001). For MBS, if $\nu_1 - \frac{d}{2} = \frac{D}{2}$ then $W|V, \Sigma, \Phi_1, \nu_1$ is matrix variate Laplacian distribution similar to the degeneration in the multivariate case.

Moreover, we would like to point out that in (Archambeau et al., 2011), the definition of the MGH distribution is derived from the GIW framework and is different from the definition in (3). Archambeau et al. propose the Gaussian scale mixture model: $W = \sqrt{\gamma}X$, where scale factor $\gamma > 0$ follows GIG distribution and $X \sim \mathcal{N}(0, \Sigma \otimes \Omega)$. The MGH conditional distribution for $W$ derived therein is a Matérn class function w.r.t. $\sqrt{\phi + \operatorname{Tr}\Omega^{-1}W^T\Sigma^{-1}W}$ and the covariance matrices $\Omega$ and $\Sigma$ are considered as constant matrices or hyper-parameters; however, in the GIW framework $\Omega$ is inverse Wishart distributed. Hence, the marginal distribution of $W$ is not preserved to be MGH if the matrix $\Omega$ is further integrated out in their model.

Though both definitions are able to degenerate to the multivariate generalized hyperbolic distribution, our definition of MGH is derived from the mixture of MGIG prior, which is a formal matrix prior with a closed form marginal distribution for $W$. We compare our model with theirs in Section 5 and show that our model is better in performance.

## 3. Inference of the GMGIG Regression Model

In this section, we propose a Bayesian model for multi-task learning by which we leverage the residual error

structure based on the GMGIG model. Herein, we make the residual error structure assumption in (1) for $\epsilon$: $\epsilon \sim \mathcal{N}_{d,N}(0, \Sigma \otimes \sigma^2 I_N)$; then we obtain the following statistical dependence on $Y$:

$$Y \sim \mathcal{N}_{d,N}(WX + \mu \mathbf{1}_N^T, \Sigma \otimes \sigma^2 I_N)$$

Notice that we denote $\Sigma$ as the task covariance matrix for $\epsilon$, which is the same as the task covariance matrix for the weight matrix $W$ in the GMGIG model, since we intend to combine the residual error structure with the GMGIG model to arrive at a more stable relationship among the tasks and to make the inference more accurate than the existing literature (Rothman et al., 2010). We define the GMGIG regression model as the graphical model in Figure 1.

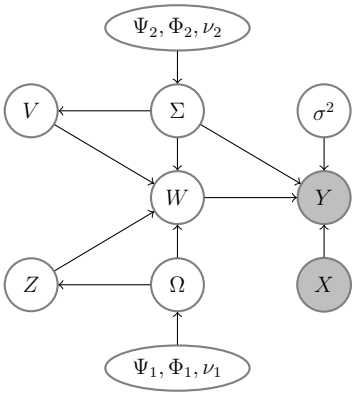


*Figure 1.* Graphical Model of the GMGIG regression Model.

In order to maximize the likelihood of the training data for a precise prediction, we adopt the Expectation Maximization (EM) algorithm under the variational framework to learn the parameters of the GMGIG regression model. Herein, we denote the observed data set as $\Delta = \{X, Y\}$ and the parameter set as $\theta$. Through the variational method, we approximate the marginal likelihood $p(\Delta)$ by

$$\log p(\Delta) \geq \int Q(\theta) \log \frac{p(\Delta, \theta)}{Q(\theta)} d\theta + \mathrm{KL}(Q(\theta)||p(\theta|\Delta))$$

where $Q(\theta)$ is the auxiliary density of the parameters and the inequality is due to the Jensen's inequality and $\mathrm{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence between two distributions, which is nonnegative. Hence, we have $\log p(\Delta) \geq \int Q(\theta) \log \frac{p(\Delta,\theta)}{Q(\theta)} d\theta \triangleq \mathcal{L}(Q)$. Herein, we tend to select an auxiliary distribution $Q(\theta)$ of the parameters $\theta$ to minimize $\mathrm{KL}(Q(\theta)||p(\theta|\Delta))$ in order to approximate $\log p(\Delta)$ by its lower bound $\mathcal{L}(Q)$. The optimal distribution to minimize the KL divergence is (Bishop, 2006):

$$Q_l(\theta_l) = \frac{\exp\langle \log p(\Delta, \theta) \rangle_{l' \neq l}}{\int \exp\langle \log p(\Delta, \theta) \rangle_{l' \neq l} d\theta_l} \tag{4}$$

Herein, $\langle \cdot \rangle_{l' \neq l}$ indicates the expectation of parameter $\theta_l$ under the joint auxiliary density without $\theta_l$.

In the E phase, we have the parameter estimation as :

$$\begin{aligned} W =& (\langle V \rangle \langle Z \rangle \langle \Omega^{-1} \rangle + \sigma^{-2}(Y - \mu \mathbf{1}_N^T) X^T) \Omega_W \\ V =& \langle W \rangle \langle \Omega^{-1} \rangle \langle Z \rangle^T (\kappa_1^{-1} I_K + \langle Z \Omega^{-1} Z^T \rangle)^{-1} \\ Z =& (\kappa_2^{-1} I_K + \langle V^T \Sigma^{-1} V \rangle)^{-1} \langle V \rangle^T \langle \Sigma^{-1} \rangle \langle W \rangle \end{aligned}$$

where $\Omega_W = (\langle \Omega^{-1} \rangle + \sigma^{-2} X X^T)^{-1}$. For parameters $\Omega$ and $\Sigma$, since there is no closed form for the expectation of MGIG distribution, we only obtain their posterior distributions, which are also MGIG distributions, as :

$$\begin{aligned} \Omega \sim& \mathcal{MGIG}_D(\hat{\Psi}_1, \Phi_1, \hat{\nu}_1) \\ \Sigma \sim& \mathcal{MGIG}_d(\hat{\Psi}_2, \Phi_2, \hat{\nu}_2) \end{aligned}$$

In the M phase, the hyperparameters are updated as :

$$\begin{aligned} \hat{\Psi}_1 =& \Psi_1 + \kappa_2^{-1} \langle Z^T Z \rangle + \langle (W - VZ)^T (W - VZ) \rangle \\ \hat{\Psi}_2 =& \Psi_2 + \kappa_1^{-1} \langle VV \rangle^T + \langle (W - VZ) \Omega^{-1} (W - VZ)^T \rangle \\ & + \sigma^{-2} \langle (Y - WX - \mu \mathbf{1}_N^T)(Y - WX - \mu \mathbf{1}_N^T)^T \rangle \\ \hat{\nu}_1 =& \nu_1 - (d + K)/2 \\ \hat{\nu}_2 =& \nu_2 - (d + N + D + K)/2 \end{aligned}$$

## 4. Numerical estimation on the statistics of MGIG distribution

In the previous section, the estimation of the parameters $\Sigma$ and $\Omega$ is obtained in the EM framework. Since there is no closed form for the parameter estimation as far as we know, we intend to offer a numerical estimation. In this section, we first introduce two fundamental propositions; then we present two sampling methods for computing the matrix Bessel function and the corresponding sampling methods for estimating the mean and the reciprocal mean of MGIG distribution. Matrix Bessel function $B_\delta(WZ)$ is defined as an integral over $\mathbb{S}_+^p$:

$$|W|^{-\delta} \int_{\mathbb{S}_+^p} |S|^{-\delta - \frac{p+1}{2}} \mathrm{etr}(-SW - S^{-1}Z) dS \tag{5}$$

where $W, Z \in \mathbb{S}_+^p$ and $\delta \in \mathbb{R}$.[3]

**Proposition 1.** *Assume that $B_\delta(WZ)$ is defined as above. We have*

$$B_\delta(WZ) = |WZ|^{-\delta} B_{-\delta}(ZW).$$

[3] $W$, $Z$ can be positive semidefinite according to the value of $\delta$ (Butler, 1998).

*If* $-\delta > \frac{p-1}{2}$, *we further have*

$$B_\delta(0) = \Gamma_p(-\delta).$$

*where* $\Gamma_p(\cdot)$ *is the multivariate gamma function (Gupta & Nagar, 2000).*

*Proof.* The first equation is from the transformation $S \to S^{-1}$ in the integral (5). The second equation is obtained by setting $Z$ to a null matrix in (5) and using the definition of the multivariate gamma function. □

**Proposition 2.** *If matrix* $G \sim \mathcal{MGIG}_p(\Psi, \Phi, \nu)$, *then* $G^{-1} \sim \mathcal{MGIG}_p(\Phi, \Psi, -\nu)$.

*Proof.* The proof is straightforward by the transformation $G \to G^{-1}$ in (2) and using Proposition 1. □

Computing the matrix Bessel function is an open problem in multivariate statistics; the existing method is the Laplace approximation (Butler & Wood, 2003). Since the approximation is not accurate, we intend to apply the Monte Carlo method to sample the integral. Ideally, we would consider MGIG distribution as the product of Wishart distribution and inverse Wishart distribution. We generate sufficient samples through either distribution and average the evaluations of the samples to estimate the integral. This estimation is valid only if $|\delta| > \frac{p-1}{2}$. For $|\delta| \le \frac{p-1}{2}$, the generation method of the random samples needs to be modified since $\delta$ in this region is not qualified to be the degree of freedom of Wishart distribution or inverse Wishart distribution. We propose two importance sampling methods (Mackay, 2003) in the following.

### 4.1. Estimating the matrix Bessel function

For the case of $|\delta| \le \frac{p-1}{2}$, we first define $t \triangleq \delta - \frac{p-1}{2}$; then we make the importance sampling in "pull" mode or in "push" mode:

- Pull the "degree of freedom"[4] $\delta$ more than $\frac{p-1}{2}$, generate the Wishart random matrices, and average the evaluations of the samples.
- Push the "degree of freedom" $\delta$ less than $-\frac{p-1}{2}$, generate the inverse Wishart random matrices, and average the evaluations of the samples.

For the "pull" mode, we have

$$\begin{aligned}
B_\delta(WZ) =& |W|^{-\delta} \int_{\mathbb{S}_+^p} |S|^{-\delta} \mathrm{etr}(-SW - S^{-1}Z) \frac{dS}{|S|^{(p+1)/2}} \\
=& \int_{\mathbb{S}_+^p} |S|^{-\delta} \mathrm{etr}(-S - W^{T/2}ZW^{1/2}S^{-1}) \frac{dS}{|S|^{(p+1)/2}} \\
=& \int_{\mathbb{S}_+^p} |S|^{(1+\alpha)t-2\delta} \\
& \cdot |S|^{-\delta+2\delta-(1+\alpha)t} \mathrm{etr}(-S - W^{T/2}ZW^{1/2}S^{-1}) \frac{dS}{|S|^{(p+1)/2}} \\
=& \langle |S|^{(1+\alpha)t-2\delta} \mathrm{etr}(-W^{T/2}ZW^{1/2}S^{-1}) \rangle \cdot \Gamma_p(\delta - (1+\alpha)t)
\end{aligned}$$

where $W = W^{1/2}W^{T/2}$; $\alpha > 0$ is a coefficient controlling the surplus of the degree of freedom beyond $\frac{p-1}{2}$ and the samples

$$S \sim Wishart(2I_p, 2(\delta - (1+\alpha)t)). \tag{6}$$

For the "push" mode, we have

$$\begin{aligned}
& B_\delta(WZ) \\
=& |W|^{-\delta} \int_{\mathbb{S}_+^p} |S|^{-\delta} \mathrm{etr}(-SW - S^{-1}Z) \frac{dS}{|S|^{(p+1)/2}} \\
=& \int_{\mathbb{S}_+^p} \frac{1}{|S|^{\delta}} \mathrm{etr}(-S - W^{T/2}ZW^{1/2}S^{-1}) \frac{dS}{|S|^{(p+1)/2}} \\
=& \int_{\mathbb{S}_+^p} |S|^{-(1+\beta)t} \\
& \cdot \frac{1}{|S|^{\delta-(1+\beta)t}} \mathrm{etr}(-S - W^{T/2}ZW^{1/2}S^{-1}) \frac{dS}{|S|^{(p+1)/2}} \\
=& \langle |S|^{-(1+\beta)t} \mathrm{etr}(-S) \rangle \cdot \frac{\Gamma_p(\delta - (1+\beta)t)}{|ZW|^{\delta-(1+\beta)t}}
\end{aligned}$$

where $\beta > 0$ is a coefficient controlling the surplus of the degree of freedom beyond $\frac{p-1}{2}$ and the samples

$$S \sim IWishart(2W^{T/2}ZW^{1/2}, 2(\delta - (1+\beta)t)). \tag{7}$$

The "push-pull" sampling methods are also feasible when $|\delta| > \frac{p-1}{2}$. For $\delta > \frac{p-1}{2}$, we set $\beta = -1$ and take the sampling through (7); for $\delta < -\frac{p-1}{2}$ we use Proposition 1 and the sampling is taken similarly.

### 4.2. Sampling the mean of MGIG

Using the "push-pull" sampling methods above, we have two methods for sampling the mean of the distribution $\mathcal{MGIG}_p(\Psi, \Phi, \nu)$. We first define $t \triangleq \nu - \frac{p-1}{2}$ and for the "pull" mode sampling, we have

$$\begin{aligned}
& \langle G \rangle_{MGIG} \\
=& \int_{\mathbb{S}_+^p} G|G|^{(1+\alpha)t} \frac{|G|^{\nu-(1+\alpha)t}}{|\frac{\Psi}{2}|^{\nu} B_\nu(\frac{\Phi}{2}\frac{\Psi}{2})} \\
& \cdot \mathrm{etr}\left(-\frac{1}{2}G^{-1}\Psi - \frac{1}{2}G\Phi\right) \frac{dG}{|G|^{(p+1)/2}} \\
=& \langle G|G|^{(1+\alpha)t} \mathrm{etr}(-G^{-1}\Psi/2) \rangle \frac{\Gamma_p(\nu - (1+\alpha)t)}{|\frac{\Phi}{2}|^{\nu-(1+\alpha)t} |\frac{\Psi}{2}|^{\nu} B_\nu(\frac{\Phi}{2}\frac{\Psi}{2})}
\end{aligned}$$

[4] Though $\delta$ is not explicitly defined as the degree of freedom for the MGIG distribution, we herein borrow the concept from Wishart distribution.

where $\alpha > 0$ is a coefficient controlling the surplus of the degree of freedom beyond $\frac{p-1}{2}$ and the samples

$$G \sim Wishart(\Phi^{-1}, 2(\nu - (1+\alpha)t)) \tag{8}$$

For the "push" mode, we have

$$\begin{aligned}
&\langle G \rangle_{MGIG} \\
=&\int_{\mathbb{S}_+^p} G|G|^{2\nu-(1+\beta)t} \frac{1}{|G|^{\nu-(1+\beta)t}|\frac{\Psi}{2}|^{\nu} B_\nu(\frac{\Phi}{2}\frac{\Psi}{2})} \\
&\cdot \text{etr}\left(-\frac{1}{2}G^{-1}\Psi - \frac{1}{2}G\Phi\right)\frac{dG}{|G|^{(p+1)/2}} \\
=&\langle G|G|^{2\nu-(1+\beta)t}\text{etr}(-G\Phi/2)\rangle \frac{\Gamma_p(\nu-(1+\beta)t)}{|\frac{\Psi}{2}|^{2\nu-(1+\beta)t}B_\nu(\frac{\Phi}{2}\frac{\Psi}{2})}
\end{aligned}$$

where $\beta > 0$ is a coefficient controlling the surplus of the degree of freedom beyond $\frac{p-1}{2}$ and the samples

$$G \sim IWishart(\Psi, 2(\nu - (1+\beta)t)). \tag{9}$$

The "push-pull" sampling methods are also feasible when $|\nu| > \frac{p-1}{2}$. For $\nu > \frac{p-1}{2}$, we simply set $\alpha = -1$ and take the sampling through (8); for $\nu < -\frac{p-1}{2}$, we set $\beta = 2\nu/t - 1$ and take the sampling through (9).

### 4.3. Sampling the reciprocal mean of MGIG

Similarly, we can sample the reciprocal mean of $\mathcal{MGIG}_p(\Psi, \Phi, \nu)$ by using Proposition 2

$$\begin{aligned}
&\langle G^{-1} \rangle_{MGIG} \\
=&\int_{\mathbb{S}_+^p} G\frac{|G|^{-\nu-(p+1)/2}}{|\frac{\Phi}{2}|^{-\nu}B_{-\nu}(\frac{\Psi}{2}\frac{\Phi}{2})}\text{etr}\left(-\frac{1}{2}G^{-1}\Phi - \frac{1}{2}G\Psi\right)dG
\end{aligned}$$

Hence, we first define $t \triangleq -\nu - \frac{p-1}{2}$ and for the "pull" mode, $\langle G^{-1} \rangle_{MGIG}$ is estimated as

$$\langle G|G|^{(1+\alpha)t}\text{etr}(-G^{-1}\Phi/2)\rangle \frac{\Gamma_p(-\nu-(1+\alpha)t)}{|\frac{\Psi}{2}|^{-(1+\alpha)t}B_\nu(\frac{\Phi}{2}\frac{\Psi}{2})}$$

where $\alpha > 0$ is a coefficient controlling the surplus of the degree of freedom beyond $\frac{p-1}{2}$ and the samples

$$G \sim Wishart(\Psi^{-1}, 2(-\nu - (1+\alpha)t)) \tag{10}$$

For the "push" model, $\langle G^{-1} \rangle_{MGIG}$ is estimated as

$$\langle G|G|^{-2\nu-(1+\beta)t}\text{etr}(-G\Psi/2)\rangle \frac{\Gamma_p(-\nu-(1+\beta)t)}{|\frac{\Phi}{2}|^{-2\nu-(1+\beta)t}B_{-\nu}(\frac{\Psi}{2}\frac{\Phi}{2})}$$

where $\beta > 0$ is a coefficient controlling the surplus of the degree of freedom beyond $\frac{p-1}{2}$ and the samples

$$G \sim IWishart(\Phi, 2(-\nu - (1+\beta)t)) \tag{11}$$

The "push-pull" sampling methods are also feasible when $|\nu| > \frac{p-1}{2}$. For $\nu < -\frac{p-1}{2}$, we set $\alpha = -1$ and take the sampling through (10); for $\nu > \frac{p-1}{2}$, we set $\beta = -2\nu/t - 1$ and take the sampling through (11).

## 5. Experiments

In this section, we report the experimental evaluations on multi-task learning on two datasets: a toy dataset and a real dataset (landmine dataset). In the real data experiment, we compare the GMGIG regression model for multi-task learning (MTL-GMGIG) with the single task learning method, the ridge logistic regression (STL), and the other state-of-the-art multi-task learning methods with least square loss including clustered multi-task learning (MTL-C) (Jacob et al., 2008), multi-task feature learning (MTL-F) (Argyriou et al., 2006; Zhou et al., 2011), multi-task learning with sparse matrix norm (MTL($\Omega$&$\Sigma$)) (Zhang & Schneider, 2010), multi-task relationship learning (MTRL) (Zhang & Yeung, 2010), multiple regression with covariance estimation (MRCE) (Rothman et al., 2010), sparse Bayesian multi-task learning (SBMTL) (Archambeau et al., 2011), and multi-task learning with GIW model (MTL-GIW).

For the hyperparameter configuration in MTL-GMGIG, we set $\Psi_1$ and $\Phi_1$ to infinite matrices and make $\Omega$ approximate to identity matrix $I_D$, $\Psi_2$ and $\Phi_2$ are initiated to $I_d$ and $5I_d$ respectively, $\nu_2$ is initiated to $d+1$, $\sigma$ is set to 10. The hyperparameter configuration for MTL-GIW is the same as that of MTL-GMGIG except that $\Phi_2$ is set to null matrix.

### 5.1. Toy Dataset

Before we apply MTL-GMGIG on the real dataset, we first conduct a proof of concept experiment on a toy dataset. We generate the toy data as follows. We establish three regression tasks according to three regression functions: $Z_1 = 2X_1 + 3Y_1 + 1$, $Z_2 = -2X_2 - 3Y_2 + 2$, and $Z_3 = 1$. For each task, we randomly sample 1000 pairs of points uniformly in the $xOy$ plane $[-5, 5] \times [-5, 5]$. Each function is corrupted by a Gaussian noise process with zero mean and variance equal to 0.1. The data points are plotted in Figure 5.1, with each color (and legend) corresponding to one task. From the coefficients of the regression functions, we expect the correlations $Corr(Z_1, Z_2)$ to approach to $-1$, $Corr(Z_1, Z_3)$ and $Corr(Z_2, Z_3)$ both to approach to 0. After we apply MTL-GMGIG, we obtain the estimated regression functions: $Z_1 = 2.003X_1 + 3.033Y_1 + 1.082$, $Z_2 = -1.964X_2 - 3.007Y_2 + 2.004$, and $Z_3 = -0.0001X_3 - 0.0019Y_3 + 0.9914$. We also obtain the correlation matrix for the three tasks in the left below and for comparison we list the correlation matrix obtained from SBMTL in the right below. Clearly, the task correlations learned herein confirm the expectation that MTL-GMGIG is able to discover the relationships among the tasks for this toy problem

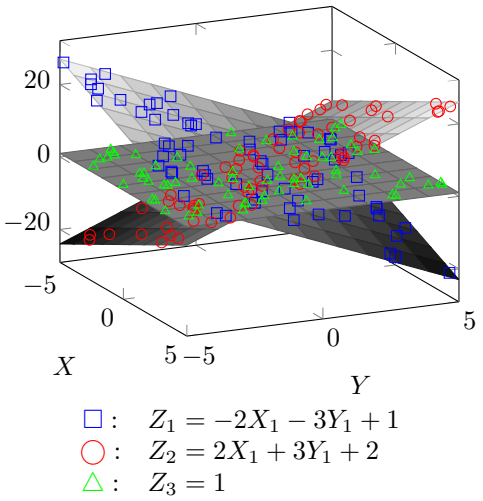


*Figure 2.* Toy dataset

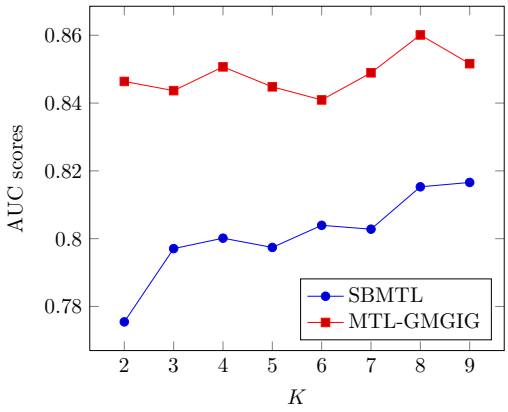


*Figure 3.* The AUC scores under different $K$ values for SBMTL and MTL-GMGIG, respectively. The training size is 30.

with a much better performance than that of SBMTL.

$$\underset{\text{MTL-GMGIG}}{\begin{bmatrix} 1 & -0.875 & -0.017 \\ -0.875 & 1 & 0.048 \\ -0.017 & 0.048 & 1 \end{bmatrix}} \quad \underset{\text{SBMTL}}{\begin{bmatrix} 1 & -0.636 & 0.103 \\ -0.636 & 1 & 0.203 \\ 0.103 & 0.203 & 1 \end{bmatrix}}$$

### 5.2. Landmine Detection Dataset

The landmine detection dataset[5] consists of 14280 examples of 29 tasks collected from various landmine fields. Each example in the dataset is detected by a radar and represented by a 9-dimensional vector describing various features concerned. The landmine detection problem is cast as a binary classification problem to predict landmines (positive class) or clutter (negative class) and we learn the GMGIG regression model for prediction. For a fair comparison with (Xue et al., 2007; Zhang & Schneider, 2010), we also jointly learn the same 19 tasks from landmine fields $1-10$ and $16-24$ in the dataset. As a result, the weight matrix $W$ is $19 \times 10$ matrix corresponding to the 19 tasks and the 10 coefficients (9 features and the intercept) for each task.

We elect to use the average AUC (Area Under the ROC Curve) as the performance measure for the comparison and vary the size of the training set for each task as 30, 40, and 80, respectively. The size of the training set is kept in a small scale since the advantage of multi-task learning would begin to vanish as the training size increases. For each task, the remaining examples are treated as the testing sets. The AUC scores are task-averaged for each run. We report the average AUC scores and their standard errors for 30 runs in Table 1. It is noted that MTL-GMGIG outperforms the other models due to the low-rank approximation to the task covariance matrix and the residual error structure introduced in the model to discover the relationship among the tasks. Consistent with the intuition, like the other methods, the performance of MTL-GMGIG increases when the training size increases. On the other hand, the gain of MTL-GMGIG over all the other methods is more significant when the training size is small where multi-task learning is more appropriate, which indicates that MTL-GMGIG is the best for multi-task learning.

For the dimensionality analysis of the latent relevance $K$ in MTL-GMGIG, we study how the performance varies with different $K$ values. Figure 3 shows that the average performance of MTL-GMGIG on the Landmine detection dataset varies with $K$. For a comparison, we also show the performance variation for SBMTL with different $K$ values. It is noted that both the AUC scores increase, though not monotonically, with the increase of $K$ and the performance of MTL-GMGIG is superior to that of SBMTL.

## 6. Conclusion

In this paper, we study the multi-task learning problem with a new perspective of considering the structure of the residue error matrix and the low-rank approximation to the task covariance matrix simultaneously. For this purpose, we first introduce the MGIG prior and propose the GMGIG model. Combining this model with the residual error structure assumption, we have developed the GMGIG regression model with the variational inference and sampling simultaneously to

[5] http://www.ee.duke.edu/~lcarin/LandmineData.zip

*Table 1.* The average AUC scores in percentage on the landmine detection dataset for $K = 7$ in the form of the mean (standard error).

| | Training Size(average % of the whole size) | | |
|---|---|---|---|
| | 30(6.0%) | 40(8.0%) | 80(16.1%) |
| STL | 63.71(0.91) | 66.72(0.60) | 70.67(0.45) |
| MTL-C | 64.23(1.10) | 69.39(0.87) | 79.75(0.95) |
| MTL-F | 62.77(0.94) | 66.94(1.11) | 70.09(1.26) |
| MTL($\Omega$&$\Sigma$) | 65.46(1.91) | 73.66(1.63) | 83.01(0.61) |
| MTRL | 78.31(0.62) | 80.64(0.65) | 87.02(1.00) |
| MRCE | 75.53(0.53) | 76.86(0.54) | 77.12(0.83) |
| SBMTL | 80.28(0.62) | 82.39(0.61) | 85.84(0.88) |
| MTL-GIW | 82.36(0.65) | 83.63(0.64) | 85.60(0.37) |
| MTL-GMGIG | **84.90(0.44)** | **86.94(0.34)** | **89.00(0.45)** |

make the computation tractable. We have developed two sampling strategies to compute the statistics of the MGIG distribution. Experiments show that this model is superior to the peer methods in regression and prediction. For our future research on multi-task learning problem, we intend to extend the prior of the task covariance martix by introducing the scale mixture model.

## Acknowledgments

We thank Professor Ronald W. Butler for his valuable suggestions. We also thank the area chair and reviewers for their constructive comments. This work is supported in part by the National Basic Research Program of China (2012CB316400), Zhejiang University — Alibaba Financial Joint lab, and Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis. ZZ is also supported in part by US NSF (IIS-0812114, CCF-1017828).

## References

Agarwal, Arvind, Daumé III, Hal, and Gerber, Samuel. Learning multiple tasks using manifold regularization. In *NIPS*, pp. 46–54, 2010.

Ando, Rie Kubota and Zhang, Tong. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

Archambeau, Cédric, Guo, Shengbo, and Zoeter, Onno. Sparse bayesian multi-task learning. In *NIPS*, pp. 1755–1763, 2011.

Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. Multi-task feature learning. In *NIPS*, pp. 41–48. MIT Press, 2006.

Argyriou, Andreas, Micchelli, Charles A., Pontil, Massimiliano, and Ying, Yiming. A spectral regularization framework for multi-task structure learning. In *NIPS*, 2007.

Barndorff-Nielsen, O., Blæsild, P., and Jensen, J. Ledet. Exponential transformation models. *Proceeding of The Royal Society Lond A*, 379(1776):41–65, 1982.

Baxter, Jonathan. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, first edition, 2006.

Bonilla, Edwin V., Chai, Kian Ming Adam, and Williams, Christopher K. I. Multi-task gaussian process prediction. In *NIPS*, 2007.

Breiman, Leo and Friedman, Jerome H. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54, 1997.

Butler, Ronald W. Generalized inverse gaussian distributions and their wishart connections. *Scandinavian Journal of Statistics*, 25(1):69–75, 1998.

Butler, Ronald W. and Wood, Andrew T. A. Laplace approximation for bessel functions of matrix argument. *Journal of Computational and Applied Mathematics*, 155:359–382, 2003.

Caruana, Rich. Multitask learning. In *Machine Learning*, pp. 41–75, 1997.

Chen, Jianhui, Tang, Lei, Liu, Jun, and Ye, Jieping. A convex formulation for learning shared structures from multiple tasks. In *ICML*, pp. 18, 2009.

Chen, Jianhui, Zhou, Jiayu, and Ye, Jieping. Integrating low-rank and group-sparse structures for robust multi-task learning. In *KDD*, pp. 42–50, 2011.

Evgeniou, Theodoros and Pontil, Massimiliano. Regularized multi-task learning. In *KDD*, pp. 109–117, 2004.

Gupta, A. K. and Nagar, D. K. (eds.). *Matrix Variate Distribution*. Chapman & Hall, 2000.

Herz, Carl S. Bessel functions of matrix argument. *Annals of Mathematics*, 61(3):474–523, 1955.

Jacob, Laurent, Bach, Francis, and Vert, Jean-Philippe. Clustered multi-task learning: A convex formulation. In *NIPS*, pp. 745–752, 2008.

Jenatton, Rodolphe, Audibert, Jean-Yves, and Bach, Francis. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

Kim, Seyoung and Xing, Eric P. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, pp. 543–550, 2010.

Kotz, Samuel, Kozubowski, Tomasz J., and Podgórski, Krzysztof. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Birkhäuser, Boston, 2001.

Le, Nhu D. and Zidek, James V. *Statistical Analysis of Environmental Space-Time Processes*. Springer, 2006.

Mackay, David J. C. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

Obozinski, Guillaume, Taskar, Ben, and Jordan, Michael I. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2009.

Rai, Piyush and Daumé III, Hal. Infinite predictor subspace models for multitask learning. *Journal of Machine Learning Research - Proceedings Track*, 9: 613–620, 2010.

Rasmussen, Carl Edward and Williams, Christopher K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Rothman, A. J., Levina, E., and Zhu, J. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, pp. 947–962, 2010.

Sohn, Kyung-Ah and Kim, Seyoung. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. *Journal of Machine Learning Research - Proceedings Track*, 22:1081–1089, 2012.

Stein, Michael L. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

Thrun, Sebastian. Is learning the n-th thing any easier than learning the first? In *NIPS*, pp. 640–646. The MIT Press, 1996.

Xue, Ya, Liao, Xuejun, Carin, Lawrence, and Krishnapuram, Balaji. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8, 2007.

Yu, Shipeng, Tresp, Volker, and Yu, Kai. Robust multi-task learning with $t$-processes. In *ICML*, pp. 1103–1110, 2007.

Zhang, Jian, Ghahramani, Zoubin, and Yang, Yiming. Learning multiple related tasks using latent independent component analysis. In *NIPS*, pp. 1585–1592, 2005.

Zhang, Yi and Schneider, Jeff G. Learning multiple tasks with a sparse matrix-normal penalty. In *NIPS*, pp. 2550–2558. Curran Associates, Inc., 2010.

Zhang, Yu and Yeung, Dit-Yan. A convex formulation for learning task relationships in multi-task learning. In *UAI*, pp. 733–742. AUAI Press, 2010.

Zhang, Zhihua, Wang, Shusen, Liu, Dehua, and I.Jordan, Michael. EP-GIG Priors and Applications in Bayesian Sparse Learning. *Journal of Machine Learning Research*, 13:2031–2061, 2012.

Zhou, J., Chen, J., and Ye, J. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011. URL `http://www.public.asu.edu/~jye02/Software/MALSAR`.