MULTIVARIATE SCALE MIXTURES FOR JOINT SPARSE REGULARIZATION IN MULTI-TASK LEARNING

Ritwik Giri

University of California, San Diego

rgiri@ucsd.edu

Bhaskar D. Rao

University of California, San Diego

brao@ucsd.edu

ABSTRACT

In this paper we address the problem of learning shared sparse representation across several tasks. Assuming that the tasks share a common set of relevant features across all tasks is highly restrictive. This acts as a motivation to look for a generalized model which will be able to learn any correlation structure present between the tasks. We propose a generalized scale mixture distribution, the Multivariate Power Exponential Scale Mixture (M-PESM), as a joint sparsity promoting prior and derive a unified framework which consists of many of the popular Multitask Learning algorithms. Our proposed unified model also has the ability to learn any present correlation structure between tasks which leads to a more robust framework.

1. INTRODUCTION

Consider a linear regression problem, where there are L set of tasks (or measurement vectors) denoted as $\{\mathbf{y}_i\}_{1\dots L}$ where, $\mathbf{y}_i \in \mathbb{R}^{n_i \times 1}$.

$$\mathbf{y_i} = \mathbf{X}_i \mathbf{w_i} + \epsilon_i \tag{1}$$

Where, $\mathbf{X}_i \in \mathbb{R}^{n_i \times m}$ is the data matrix constructed using training data, $\mathbf{w}_i \in \mathbb{R}^{m \times 1}$ is the coefficient vector and $\epsilon_i \in \mathbb{R}^{n_i \times 1}$ could be interpreted as measurement noise. Assuming that the measurement noise is zero mean Gaussian with unknown variance λ, the likelihood function for the coefficient vector $\mathbf{w}_i$ based on the ith task output/target $\mathbf{y}_i$ can be expressed as,

$$p(\mathbf{y}_i|\mathbf{w_i}, \lambda) = (2\pi\lambda)^{-n_i/2} \exp\left(- \frac{||\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i||_2^2}{2\lambda} \right) \tag{2}$$

When the number of features (m) is greater than the number of data points (n_i) in model (1) the problem becomes under-determined [8]. That means there could be infinite number of solutions for the regression coefficients that perfectly explain the data. To obtain a unique solution of regression coefficients we often employ a sparsity promoting regularization, which means only few relevant features will be selected [15]. There has been a lot of interest and work on promoting sparsity using ℓ_1 norm regularization [6, 7, 22]. From a Bayesian perspective supergaussian (i.e. priors with heavier tails than gaussian) distributions have been employed as prior to promote sparsity in the coefficient vector with reasonable success [19, 18]. For Multitask Learning (MTL) or a Multiple Measurement Vector (MMV) sparse recovery problem, notion of joint sparsity has been introduced [25, 1, 26]. Key assumption behind this is that all the tasks will share the same set of relevant features. Joint sparse regularization approach has been used, where we seek row sparsity in the regression coefficient matrix by employing a multivariate supergaussian prior distributions to model joint sparsity, which encourages the entire rows of the coefficient matrix to have zero elements [9, 27, 12]. Joint regularization using ℓ_{2-1} mixed norm is a straightforward extension of LASSO (single task/measurement case), which has been used extensively to solve this problem [17]. In real life applications we often see that all the tasks may not always share the same set of features and some of the tasks could be outliers or could be negatively correlated with other tasks. To model the outlier tasks, recently a Dirty model for MTL has been introduced which uses a combined regularization of ℓ_1/ℓ_∞ to model the joint sparsity and ℓ_1 to model outliers [14]. A probabilistic interpretation of this dirty model has also been proposed in [13]. It has also been discussed in recent literatures [12, 27] that if the model is able to capture the task relatedness, i.e. any present correlation structure, the generalization capability of the model increases significantly. Recently some works [21, 5] have also proposed using Iterative Reweighted Least Square (IRLS) approaches to model joint sparsity from a MTL point of view. In [24] authors have extended the reweighted ℓ_1 minimization [3] approach to model the joint sparsity for MMV recovery problem. In Bayesian based approaches, Multivariate Gaussian Scale mixtures (M-GSM) and Multivariate Laplacian Scale Mixtures (M-LSM) have been used as prior distributions to promote joint sparsity, because of their supergaussian nature. In [23] authors have proposed a new sparse Bayesian multitask learning method based on a GSM prior which also models the correlation structure within tasks.

In our recent work [11], we have introduced a multivariate extension of our recently proposed generalized Scale Mixture framework [10], namely Multivariate Power Exponential Scale Mixtures (M-PESM) as a source prior for a joint blind source separation task. In this paper we present the usefulness of M-PESM to model the joint sparsity and show its application in a multi-task learning framework. This work will primarily focus on the Multivariate Generalized t distribution (M-GT) family of priors, a member of M-PESM, since it has a wide range of tail shapes and includes heavy tailed super gaussian distributions. We also derive a unified MAP estimation framework using M-GT as sparsity inducing prior and show that many of the popular regularization based MTL algorithms falls under our proposed unified framework. Our model also has the flexibility of learning any correlation structure present between tasks which will help us to model any outlier task or task with negative correlation.

The rest of the paper is organized as follows. In Section 2, a generalized scale mixture representation, the Multivariate Power Exponential Scale Mixtures (M-PESM) family, is presented. In Section 3, we derive a unified MAP based inference procedure by employing a joint sparsity promoting prior distribution from the family of M-PESM. In Section 4, we discuss some special cases of the unified framework and show connections with current algorithms in the literature. We present experimental results of the proposed algorithms using both synthetic data and real data in Section 5, in different set-

tings and finally conclusions and some future directions of this work are presented in Section 6.

2. SPARSITY INDUCING PRIOR: SCALE MIXTURES

For joint sparse regularization from a MMV or MTL point of view, multivariate Gaussian scale mixtures and Laplace scale mixtures have been used as sparsity promoting prior. In this section, we discuss a recently proposed [11], more general Multivariate Power Exponential Scale Mixture (M-PESM) distribution, which is a generalization of M-GSM and M-LSM.

2.1. Multivariate Power Exponential (M-PE)

In this work we are concerned with the M-PE distribution, which is also known as Generalized Gaussian Distribution (GGD) and has received lot of attention in the literature. The probability density function of a M-PE is defined by [20],

$$p_{\text{M-PE}}(\mathbf{x}|\mathbf{M},\beta,z) = \frac{1}{|\mathbf{M}|^{1/2}} h_{\beta,z}(\mathbf{x}^T\mathbf{M}^{-1}\mathbf{x}) \tag{3}$$

for any $\mathbf{x} \in \mathbb{R}^{L\times 1}$, where $\mathbf{M}$ is a $L \times L$ symmetric real correlation matrix, and $h()$ is known as the density generator defined by,

$$h_{\beta,z}(y) = \frac{\beta\Gamma(\frac{L}{2})}{\pi^{\frac{L}{2}}\Gamma(\frac{L}{2\beta})z^{\frac{L}{2\beta}}} \exp\left(-\frac{y^\beta}{z}\right) \tag{4}$$

Where, $z > 0$ is the scale parameter and $\beta > 0$ is the shape parameter of the M-PE. It is evident from the above given form, that $\beta = 1$ results in the Multivariate Gaussian distribution, whereas $\beta = 1/2$ connects to the well known Multivariate Double exponential or Laplace distribution.

2.2. Multivariate PESM (M-PESM)

Multivariate PESM family of distributions refer to distributions that can be represented as follows:

$$p_{\mathbf{X}}(\mathbf{x}) = \int p_{\text{M-PE}}(\mathbf{x};\mathbf{M},\beta,z)p_z(z)dz \tag{5}$$

Some special cases of M-PESM includes Multivariate Gaussian Scale Mixtures (M-GSM) when shape parameter $\beta = 1$, Multivariate Laplace Scale Mixtures (M-LSM) when shape parameter $\beta = 1/2$, Multivariate Uniform Scale Mixtures (M-USM) when $\beta \to \infty$. More theoretical details and the properties of M-PESM can be found in [11].

2.3. Multivariate Generalized t Distribution (M-GT)

In this example, we will consider an inverse gamma (IG) distribution as our mixing density $p_z(z) = IG(q,q)$, where $IG(x;a,b) = \frac{b^a}{\Gamma(a)}x^{-a-1}\exp\left(-\frac{b}{x}\right)u(x)$ in the hierarchical representation (5) for the M-PESM family. It leads to a multivariate generalized t distribution [2] which also includes well known supergaussian densities, useful to promote joint sparsity e.g. Multivariate Laplace, Multivariate Student's t distributions, among others. The Multivariate Generalized t Distribution has the form:

$$p_{\text{M-GT}}(\mathbf{x};q,\beta,\mathbf{M}) = \frac{\eta}{(q+s^\beta)^{q+\frac{L}{2\beta}}} \tag{6}$$

Where $s = \mathbf{x}^T\mathbf{M}^{-1}\mathbf{x}$, η is the normalization constant. Interestingly, β and q provide the flexibility to represent different tail behavior using this distribution. In Table 1, we summarize some special cases of Multivariate GT that have been used in literature to promote joint sparsity that arise by different choices of the shape parameters of M-GT, i.e. β and q (With $\mathbf{M} = I$).

3. BAYESIAN INFERENCE

In this section we derive a unified estimation algorithm using M-PESM as the sparse prior. Then we specialize the result using the M-GT as the sparse prior and also show that the generalized algorithm reduces to well known Multi task learning algorithms.

3.1. Unified MAP Estimation

Because of the independence between rows of the coefficient matrix $\mathbf{W}$, every $p(\mathbf{w}_{i,:})$ has an independent scale mixture representation, i.e,

$$p(\mathbf{w}_{i,:}) = \int_0^\infty p(\mathbf{w}_{i,:}|z_i)p(z_i)dz_i \tag{7}$$

For EM algorithm we will treat scale parameters z_i as hidden variables. Hence the complete data log-likelihood can be written as,

$$\log p(\mathbf{Y},\mathbf{W},\mathbf{z}) = \log p(\mathbf{Y}|\mathbf{W}) + \sum_{i=1}^m \log p(\mathbf{w}_{i,:}|z_i) + \sum_{i=1}^m \log p(z_i) \tag{8}$$

To compute the Q function we need the conditional expectation of the complete data log likelihood with respect to the conditional posterior of the hidden variables, i.e, $p(\mathbf{z}|\mathbf{W},\mathbf{Y})$ which reduces to $p(\mathbf{z}|\mathbf{W})$ by virtue of the Markovian property. Now in the M step we will maximize the Q function with respect to $\mathbf{W}$, so we are only interested in the first two terms of the Equation (8). Since only the second has dependencies on the hidden variable $\mathbf{z}$, in the E step we are only concerned with this term, i.e,

$$\begin{aligned}\sum_{i=1}^m \log p(\mathbf{w}_{i,:}|z_i) &= \sum_{i=1}^m \log p_{\text{M-PE}}(\mathbf{w}_{i,:};\mathbf{M}_i,\beta,z_i) \\ &= -\sum_{i=1}^m \frac{(\mathbf{w}_{i,:}\mathbf{M}_i^{-1}\mathbf{w}_{i,:}^T)^\beta}{z_i} + \text{constants}\end{aligned} \tag{9}$$

Hence, the E step essentially becomes computation of the following conditional expectation, $E_{z_i|\mathbf{w}_{i,:}}\left[\frac{1}{z_i}\right]$.

The derivation of the concerned conditional expectation where a M-GT has been employed as the sparsity inducing prior, is given in Appendix, which has been found as,

$$E_{z_i|\mathbf{w}_{i,:}}\left[\frac{1}{z_i}\right] = \frac{q+\frac{L}{2\beta}}{q+E_i^\beta} \tag{10}$$

Where, $E_i = \mathbf{w}_{i,:}\mathbf{M}_i^{-1}\mathbf{w}_{i,:}^T$. Lets define the weights as,

$$v_i = E_{z_i|\mathbf{w}_{i,:}}\left[\frac{1}{z_i}\right] = \frac{q+\frac{L}{2\beta}}{q+E_i^\beta} \tag{11}$$

Hence the M step becomes,

$$\mathbf{W}^{(k+1)} = \arg\min_{\mathbf{W}} \sum_{i=1}^L \frac{1}{2\lambda}\|\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i\|_2^2 + \sum_{i=1}^m v_i^{(k+1)}(\mathbf{w}_{i,:}\mathbf{M}_i^{-1}\mathbf{w}_{i,:}^T)^\beta \tag{12}$$

It's evident from the M step that our proposed unified framework falls under the reweighted schemes where weights of $(k+1)^{th}$ iteration, i.e, $v_i^{(k+1)}$ depend on the coefficients from previous iteration.

3.2. Learning Task Correlation

By incorporating a data adaptive correlation matrix $\mathbf{M}_i$ in our algorithm, we can capture any outlier tasks. It will also help to exploit any present correlation structure in $\mathbf{w}_{i,:}$ through learning $\mathbf{M}_i$ adaptively. In our algorithm we will constrain all the $\mathbf{M}_i = \mathbf{M}$, to prevent overfitting because of the large number of parameters.

Table 1: Variants of Multivariate GT distribution

q	β	Prior Distribution	Penalty Function	SSR Algorithm
$q \to \infty$	1	M-Normal	$\|\|\mathbf{W}\|\|_F$	M-Ridge Regression
$q \to \infty$	1/2	M-Laplacian	$\|\|\mathbf{W}\|\|_{2,1}$	M-LASSO
$q \geq 0$ (degrees of freedom)	1	M-Student t distribution	$\sum_i \log(\epsilon + \|\|\mathbf{w}_{i,:}\|\|_2^2)$	Iterative Reweighted Least Squares
$q \geq 0$ (shape parameter)	1/2	M-Generalized Double Pareto	$\sum_i \log(\epsilon + \|\|\mathbf{w}_{i,:}\|\|_2)$	Reweighted ℓ_1

Revisiting the M step and taking derivative with respect to $\mathbf{M}$ and equating it to zero we get,

$$\mathbf{M}^{(k+1)} = \frac{2\beta}{m} \sum_{i=1}^{m} v_i^{(k+1)} (\mathbf{w}_{i,:} {\mathbf{M}^{(\mathbf{k})}}^{-1} \mathbf{w}_{i,:}^T)^{\beta-1} \mathbf{w}_{i,:}^T \mathbf{w}_{i,:} \quad (13)$$

In real applications we will also add a regularization term to the update of $\mathbf{M}$ to make it robust to the estimation error of $\mathbf{W}$ over the iterations.

$$\mathbf{M}^{(k+1)} \leftarrow \frac{2\beta}{m} \sum_{i=1}^{m} v_i^{(k+1)} (\mathbf{w}_{i,:} {\mathbf{M}^{(\mathbf{k})}}^{-1} \mathbf{w}_{i,:}^T)^{\beta-1} \mathbf{w}_{i,:}^T \mathbf{w}_{i,:} + \alpha I \quad (14)$$

Where, α is a small positive scalar, to maintain the positive definite property of $\mathbf{M}$. We will also normalize $\mathbf{M}$ after every update, i.e, $\hat{\mathbf{M}}^{(k+1)} \leftarrow \mathbf{M}^{(k+1)} / \|\mathbf{M}^{(k+1)}\|_F$. This data adaptive correlation matrix $\mathbf{M}$ can also be interpreted as data adaptive kernel which helps to exploit any structure present among the tasks which is a significant advantage over algorithms that are blind to any correlation structure.

4. SPECIAL CASES OF UNIFIED FRAMEWORK

In this section by choosing specific distributional parameters we will show how our proposed unified framework leads to well known Multitask Learning algorithms.

4.1. ℓ_{2-1} Minimization: Joint Feature Selection

ℓ_{2-1} norm minimization based joint feature selection approach [17] is one of the earliest multitask learning algorithm employing joint sparse regularization. From a Bayesian point of view employing a M-Laplace distribution as the joint sparsity inducing prior over the rows of the coefficient matrix and seeking a MAP estimate will lead to this algorithm. Interestingly we see from Table 1 that for specific values of the shape parameters ($q \to \infty, \beta = 1/2$), a Multivariate GT distribution can be used to represent M-Laplace. Now to relate with the unified MAP estimation framework taking the limit as $q \to \infty$ in Equation (11) we get $v_i = 1$. Hence in the M step we are solving a ℓ_{2-1} norm penalized regression problem where weights are not changing over iteration, showing that ℓ_{2-1} Minimization is a special case of our unified framework.

4.2. Iterative Reweighted ℓ_1 minimization (IRL-1)

In [24, 16] an iterative reweighted ℓ_1 minimization algorithm has been discussed to promote joint sparsity. From a Bayesian point of view, MAP estimation of the coefficient matrix with a M-Generalized double pareto distribution as a prior will lead to the same cost function. Now, substituting the distributional parameters ($q = \epsilon, \beta = 1/2$) from Table 1 in Equation (11) we get weights as, $v_i = \frac{\epsilon + L}{\epsilon + \sqrt{\mathbf{w}_{i,:} \mathbf{w}_{i,:}^T}} = \frac{\epsilon + L}{\epsilon + ||\mathbf{w}_{i,:}||_2}$, same as shown in [16] using MM algorithm. It's evident that this algorithm also falls under our proposed unified framework. On the other hand our framework also allows learning the correlation structure between tasks and leads to correlation aware regularization penalty unlike the algorithm discussed in [16]. We will refer to the context aware version of this algorithm as **C-IRL-1** which involves computing the weights v_i following Equation (11) with $q = \epsilon, \beta = 1/2$, updating the correlation matrix $\mathbf{M}$ using Equation (14) with $\beta = 1/2$ and then solving a weighted ℓ_{2-1} mixed norm minimization problem shown in Equation (12).

4.3. Iterative Reweighted Least Squares (IRLS)

Iterative Reweighted Least Square (IRLS) was first proposed from a single measurement sparse recovery perspective. In recent works [21, 5] it has been extended for joint sparse regularization both from a MMV recovery and Multitask learning point of view. As shown in Table 1, employing a M-student t distribution as a prior and following the MAP estimation route will lead to the same cost function as discussed in [5]. By choosing the specific distributional parameters (from Table 1) and substituting in Equation (11) we get, $v_i = \frac{\epsilon + L/2}{\epsilon + \mathbf{w}_{i,:} \mathbf{w}_{i,:}^T} = \frac{\epsilon + L/2}{\epsilon + ||\mathbf{w}_{i,:}||_2^2}$, which is a straightforward extension of Reweighted ℓ_2 minimization algorithm [4] for MMV case. Since our unified framework allows us to learn the correlation structure, in our proposed correlation aware IRLS (**C-IRLS**) the weights will be computed as, $v_i = \frac{\epsilon + L/2}{\epsilon + \mathbf{w}_{i,:} \mathbf{M}^{-1} \mathbf{w}_{i,:}^T}$ We will also learn the correlation matrix using Equation (14) and then we just need to solve a weighted least squares problem following Equation (12) with $\beta = 1$.

5. EXPERIMENTS

In this section we carry out experiments using both synthetic data and real data to evaluate the empirical performances of the above discussed models.

5.1. Experiments with Synthetic Data

In this case we will assume that same data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ (where, $n = 50, m = 100$) has been used for all the tasks. The entries of the data matrix $\mathbf{X}$ have been sampled from a standard Gaussian distribution with mean zero and standard deviation 1. Lets assume that there are $L = 10$ tasks and all the task share the same set of $K = 22$ relevant features. We will also assume that the first two tasks and the last eight tasks are positively correlated but the two groups are negatively correlated. Thus the nonzero rows of the coefficient matrix $\mathbf{W}_{gen}$ have been sampled from a multivariate Gaussian with mean zero vector and covariance matrix with 1's on the diagonals and either $+\beta$ or $-\beta$ on the off diagonal elements, depending on the locations. Now the target matrix $\mathbf{Y}$ is obtained following $\mathbf{Y} = \mathbf{X}\mathbf{W}_{gen} + \epsilon$. Where the additive noise is gaussian and the variance is chosen such that SNR is 10 dB. The target matrix $\mathbf{Y}$ and data matrix $\mathbf{X}$ are shown to all the competing algorithms and the reconstruction error of model coefficients are measured as: Error $= \frac{\|\hat{\mathbf{W}} - \mathbf{W}_{gen}\|_F}{\|\mathbf{W}_{gen}\|_F}$. The same experiment has been repeated 50 times and the averaged error has been reported in Table 2. We run

the experiments for two values of $\beta = 0.9$ and, 0. In the first case there is a significant correlation structure between tasks, so we hope to see a significant improvement for our proposed correlation aware algorithms. Whereas in the second case there is no correlation structure so we expect to see similar performance of both Correlation aware and correlation unaware algorithms. In Table 2 for $\beta = 0.9$ we see that C-IRLS performs significantly better compared to IRLS whereas C-IRL-1 also shows little improvement over IRL-1.

Table 2: Averaged Reconstruction Error using Synthetic Data

Methods	Error	
	$\beta = 0.9$	$\beta = 0$
ℓ_{2-1}	0.6007	0.4800
M-FOCUSS	0.6321	0.4559
IRL-1	0.3768	0.2712
C-IRL-1 (Proposed)	0.3679	**0.2710**
TMSBL	0.4325	0.3168
IRLS	0.4795	0.3056
C-IRLS (Proposed)	**0.3633**	0.3030
DM	0.6489	0.5629

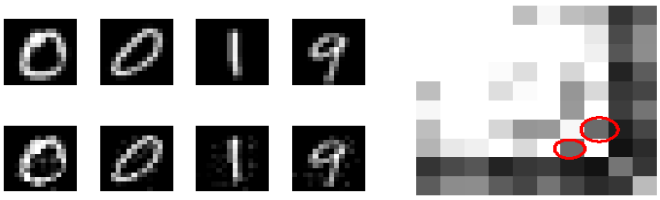
5.2. Experiments with Real Data

In this section we consider the reconstruction of images of hand written digits taken from the popular MNIST dataset. Since for these handwritten digits the background pixels are always zero and most of them share same locations across all the images, joint sparsity could be used here. We downsample the images to 14×14 pixels and vectorize them, where each image is represented using a 196 dimensional vector. We randomly choose 8 images of digit '0' and two randomly chosen images of digit '1' and digit '9'. Last two digits i.e, '1' and '9' can be interpreted as outlier tasks. Now in MTL setup, model coefficients $\mathbf{w}_l$ are the vectorized pixel values. Again we will choose the same data matrix $\mathbf{X} \in \mathbb{R}^{120\times 196}$ for all the tasks and the entries of $\mathbf{X}$ are sampled from a standard Gaussian distribution with mean zero and standard deviation 1. Following the previous section we will generate the target matrix $\mathbf{Y}$ with some additive noise where the SNR is 20 dB. We compare the reconstruction error by several competing algorithms in Table 3. We again see the improvement of performance by correlation aware algorithms, where C-IRLS produces the best reconstruction error.

Table 3: Averaged Reconstruction Error using MNIST

Methods	Error
ℓ_{2-1}	0.3879
M-FOCUSS	0.3218
IRL-1	0.2965
C-IRL-1 (Proposed)	0.2834
TMSBL	0.3039
IRLS	0.3056
C-IRLS (Proposed)	**0.2426**
DM	0.4212

In Figure 1(a) we show true images of two '0's (7th and 8th task) and the outliers '1' and '9' (9th and 10th task) and also the corresponding reconstructed images using C-IRLS. In Figure 1(b) we show the correlation matrix that has been learned by C-IRLS (White corresponds to 1 and black corresponds to 0). Interestingly



(a) (b)

Fig. 1: (a) (Top) True Images, (Bottom) Recon. images using C-IRLS, (b) Correlation between tasks learned by C-IRLS for MNIST

we find out that our model has been able to learn high correlation between the first 8 tasks (images of '0') and also a very low correlation between a true task and last two outlier tasks. Another interesting observation is the correlation learned between 7th and 8th task in Figure 1(b) (Red circled), which is also low, though they belong to the same digit. For sanity check, we can verify from Figure 1(a) that the 7th task and 8th task, i.e., two true images of handwritten '0' are significantly different which leads to a low correlation value captured by **C-IRLS**.

6. CONCLUSION

In this paper we have introduced a new class of multivariate scale mixture prior distribution to model joint sparsity and derived a unified inference framework which covers many of the popular Multi-task learning algorithms. Our proposed correlation aware algorithms provide the flexibility of exploiting any present correlation structure between tasks. Our experimental results over both synthetic data and real data shows improvements of the proposed correlation aware approaches over other competing algorithms.

7. APPENDIX

To compute the concerned expectation we will employ the following trick. Differentiating inside the integral of the marginalized $p(\mathbf{w}_{i,:})$ we get,

$$\begin{aligned} p'(\mathbf{w}_{i,:}) &= \frac{d}{d\mathbf{w}_{i,:}^T}\int_0^\infty p(\mathbf{w}_{i,:}|z_i)p(z_i)dz_i \\ &= -2\beta \times E_i^{\beta-1}\mathbf{M}_i^{-1}\mathbf{w}_{i,:}^T\int_0^\infty \frac{1}{z_i}p(\mathbf{w}_{i,:},z_i)dz_i \end{aligned} \tag{15}$$

Where, $E_i = \mathbf{w}_{i,:}\mathbf{M}_i^{-1}\mathbf{w}_{i,:}^T$.

Now employing the product rule of probability $p(\mathbf{w}_{i,:},z_i) = p(\mathbf{w}_{i,:})p(z_i|\mathbf{w}_{i,:})$ and taking $p(\mathbf{w}_{i,:})$ outside the integral we get,

$$\begin{aligned} p'(\mathbf{w}_{i,:}) &= -2\beta \times E_i^{\beta-1}\mathbf{M}_i^{-1}\mathbf{w}_{i,:}^T p(\mathbf{w}_{i,:})\int_0^\infty \frac{1}{z_i}p(z_i|\mathbf{w_{i,:}})dz_i \\ &= -2\beta \times E_i^{\beta-1}\mathbf{M}_i^{-1}\mathbf{w}_{i,:}^T p(\mathbf{w}_{i,:})E_{z_i|\mathbf{w}_{i,:}}\left[\frac{1}{z_i}\right] \end{aligned} \tag{16}$$

Now lets consider a special case where a Multivariate GT has been employed as a prior, $p(\mathbf{w}_{i,:})$. We can write, $p(\mathbf{w}_{i,:}) = \eta\exp(-f(\mathbf{w}_{i,:}))$, where, $f(\mathbf{w}_{i,:}) = (q+\frac{L}{2\beta})\log\left(q+{E_i}^\beta\right)$.

$$\begin{aligned} p'(\mathbf{w}_{i,:}) &= -p(\mathbf{w}_{i,:})f'(\mathbf{w}_{i,:}) \\ &= -p(\mathbf{w}_{i,:})2\beta \times E_i^{\beta-1}\mathbf{M}_i^{-1}\mathbf{w}_{i,:}^T\frac{q+\frac{L}{2\beta}}{q+E_i^\beta} \end{aligned} \tag{17}$$

Comparing Equation 16 and Equation 17 we get,

$$E_{z_i|\mathbf{w}_{i,:}}\left[\frac{1}{z_i}\right] = \frac{q+\frac{L}{2\beta}}{q+E_i^\beta} \tag{18}$$

8. REFERENCES

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[2] O. Arslan. Family of multivariate generalized t distributions. *Journal of Multivariate Analysis*, 89(2):329–337, 2004.

[3] E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted 1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[4] R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. In *Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on*, pages 3869–3872. IEEE, 2008.

[5] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

[6] D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

[7] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[8] M. Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

[9] M. V. Gerven, B. Cseke, R. Oostenveld, and T. Heskes. Bayesian source localization with the multivariate laplace prior. In *Advances in Neural Information Processing Systems*, pages 1901–1909, 2009.

[10] R. Giri and B. Rao. Type i and type ii bayesian methods for sparse signal recovery using scale mixtures. *IEEE Transactions on Signal Processing*, 64(13):3418–3428, 2016.

[11] R. Giri, B. Rao, and H. Garudadri. Reweighted algorithms for independent vector analysis. *IEEE Signal Processing Letters*, In review, 2016.

[12] D. Hernández-Lobato and J. M. Hernández-Lobato. Learning feature selection dependencies in multi-task learning. In *Advances in Neural Information Processing Systems*, pages 746–754, 2013.

[13] D. Hernández-Lobato, J. M. Hernández-Lobato, and Z. Ghahramani. A probabilistic model for dirty multi-task feature selection.

[14] A. Jalali, S. Sanghavi, C. Ruan, and P. K. Ravikumar. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems*, pages 964–972, 2010.

[15] I. M. Johnstone and D. M. Titterington. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4237–4253, 2009.

[16] Q. Ling, Z. Wen, and W. Yin. Decentralized jointly sparse optimization by reweighted minimization. *Signal Processing, IEEE Transactions on*, 61(5):1165–1170, 2013.

[17] F. Nie, H. Huang, X. Cai, and C. H. Ding. Efficient and robust feature selection via joint 2, 1-norms minimization. In *Advances in neural information processing systems*, pages 1813–1821, 2010.

[18] J. Palmer, K. Kreutz-Delgado, B. D. Rao, and D. P. Wipf. Variational em algorithms for non-gaussian latent variable models. In *Advances in neural information processing systems*, pages 1059–1066, 2005.

[19] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig. Strong sub- and super-gaussianity. In *Latent Variable Analysis and Signal Separation*, pages 303–310. Springer, 2010.

[20] F. Pascal, L. Bombrun, J.-Y. Tourneret, and Y. Berthoumieu. Parameter estimation for multivariate generalized gaussian distributions. *IEEE Transactions on Signal Processing*, 61(23):5960–5971, 2013.

[21] J. Pu, Y.-G. Jiang, J. Wang, and X. Xue. Multiple task learning using iteratively reweighted least square. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1607–1613. AAAI Press, 2013.

[22] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[23] J. Wan, Z. Zhang, J. Yan, T. Li, B. D. Rao, S. Fang, S. Kim, S. L. Risacher, A. J. Saykin, and L. Shen. Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 940–947. IEEE, 2012.

[24] M.-H. Wei, W. R. Scott Jr, and J. H. McClellan. Jointly sparse vector recovery via reweighted 1 minimization. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 3929–3932. IEEE, 2012.

[25] D. P. Wipf and B. D. Rao. An empirical bayesian strategy for solving the simultaneous sparse approximation problem. *Signal Processing, IEEE Transactions on*, 55(7):3704–3716, 2007.

[26] Y. Yang, Y. Yang, Z. Huang, H. T. Shen, and F. Nie. Tag localization with spatial correlations and joint group sparsity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 881–888. IEEE, 2011.

[27] Y. Zhang and J. G. Schneider. Learning multiple tasks with a sparse matrix-normal penalty. In *Advances in Neural Information Processing Systems*, pages 2550–2558, 2010.