It is the paper published as:

Author: J. Gao, P. Kwan and Y. Guo
Title: Robust Multivariate L1 Principal Component Analysis and Dimensionality Reduction
Journal: Neurocomputing
ISSN: 0925-2312
Year: 2009
Volume: 72
Issue: 4-Jun
Pages: 1242-1249

Abstract: Further to our recent work on the robust L1 PCA we introduce a new ver-sion of robust PCA model based on the so-called multivariate Laplace distribution(called L1 distribution) proposed in (Eltoft et al., 2006). Due to the heavy tail and high component dependency characteristics of the multivariate L1 distribution, the proposed model is expected to be more robust against data outliers and ¯ttingcomponent dependency. Additionally, we demonstrate how a variational approx-imation scheme enables e®ective inference of key parameters in the probabilistic multivariate L1-PCA model. By doing so, a tractable Bayesian inference can be achieved based on the variational EM-type algorithm.

Author Address: jbgao@csu.edu.au

URL: http://dx.doi.org/10.1016/j.neucom.2008.01.027

http://www.elsevier.com/wps/find/journaldescription.cws_home/505628/description#description

Robust Multivariate L1 Principal Component Analysis and Dimensionality Reduction

Junbin Gao[1]*, Paul W. Kwan[2] and Yi Guo[2]

[1]*School of Computer Science*
Charles Sturt University, Bathurst, NSW 2795, Australia
jbgao@csu.edu.au

[2]*School of Science and Technology*
University of New England, Armidale, NSW 2351, Australia
{yguo4,kwan}@turing.une.edu.au

Revised Version

Abstract

Further to our recent work on the robust L1 PCA we introduce a new version of robust PCA model based on the so-called multivariate Laplace distribution (called L1 distribution) proposed in (Eltoft et al., 2006). Due to the heavy tail and high component dependency characteristics of the multivariate L1 distribution, the proposed model is expected to be more robust against data outliers and fitting component dependency. Additionally, we demonstrate how a variational approximation scheme enables effective inference of key parameters in the probabilistic multivariate L1-PCA model. By doing so, a tractable Bayesian inference can be achieved based on the variational EM-type algorithm.

1 Introduction

Most models favor a Gaussian likelihood distribution, for example, the Gaussian noise model assumption in the probabilistic principal component analysis (PPCA) (Tipping and Bishop, 1999). However if the underlying process generating the data set is not controlled by a Gaussian distribution, then the model that favors a Gaussian estimate may interpret the data set in a misleading way, in particular when there exist data outliers. A robust model that resists against any data outliers or non-Gaussian noises is

*The author to whom all correspondences should be addressed.

always preferred in dealing with real-life data. One solution is to favor an assumption of heavy tailed distributions in the models.

The Student-t distribution is one example of the heavy-tailed distributions. In fact, the Student-t distribution is a heavy-tailed generalization of the Gaussian distribution. Compared with Gaussian distributions, using Student-t distributions in a model significantly increases its robustness. Such kind of work has been done by a number of researchers under different assumptions and different algorithmic implementations. To the authors' best knowledge, the first work regarding robust PCA algorithms was done by Ruymagaart (1981) and an application of robust PCA in computer vision was given by de la Torre and Black (2001). An earlier survey regarding the robust PCA can be found in (de la Torre and Black, 2003). Under the assumption of the Student-t distribution, the related research include a mixture of Student-t models (Peel and McLachlan, 2000), which actually is a generalized mixture of Gaussian models without considering subspace structures, and more recent work such as the robust subspace mixture model (Ridder and Franc, 2003), in which both the likelihood and the hidden variables were supposed to be Student-t distributions and the EM algorithm was applied to the model; robust generative subspace models (Khan and Dellaert, 2004), in which the Student-t distribution is expressed in an infinite superposition of Gaussian distribution; robust projections (Archambeau et al., 2006), in which the authors considered the robust PPCA and PCCA algorithms; and robust Bayesian interpolation and independent component analysis (ICA) (Tipping and Lawrence, 2005). Most recently Archambeau (2005) discussed the robust models in the context of finite mixture models.

Although the Student-t distribution is used in designing a robust statistical model, the same purpose can be achieved by another approach which instead uses the so-called centered Laplacian distribution (or L1 distribution or the least absolute deviance). The L1 distribution is much less sensitive to outliers compared to the Gaussian density. The approach of using the L1 distribution originates from LASSO (Tibshirani, 1996), and has caught some interests in machine learning (Ng, 2004) and statistics.

In early 1996, Baccini et al. (1996) proposed the first L1-PCA model. A PCA based on Gini's mean absolute differences is introduced in order to obtain heuristic estimates of the L1 model. The resulting PCA is connected to the canonical correlation analysis of the original data. By replacing L2-norm (associated with a Gaussian assumption) with L1-norm (associated with the L1 distribution assumption), Ke and Kanade (2005) proposed a matrix factorization algorithm which can handle outliers and missing data. It was shown that the proposed approach outperforms other approaches including IRLS (Iteratively Reweighted Least Squares) (Rubin, 2006) for both synthetic and real data. Similar to (Ke and Kanade, 2005), the so-called R1-norm (a modified L1-norm) was employed in the matrix factorization algorithm in the context of PCA (Ding et al., 2006). More recently one of the authors gave a robust L1 PCA model (Gao, 2008) based on the univariate L1 distribution in which the correlation and dependency between the multivariate data features are supposed to be zero. However when the multivariate components are mutually correlated and have higher-order dependencies, the univariate L1 distribution is far from the best choice for the model. We noted that, in their recent work (Eltoft et al., 2006), the authors introduced a new version of the multivariate Laplace distribution. Due to its heavy tail property it can be used to construct a new

robust L1 PCA model which can resist both outliers and mutual correlation.

This paper is concerned with the generative modelling for multivariate L1-PCA by using Bayesian learning and inference approaches. We noted that the way of defining the multivariate Laplacian distribution offers an approach to handle the multivariate L1 generative modeling by the so-called variational EM-type algorithm.

In the next section, we describe the multivariate Laplacian distribution. In section 3 the corresponding probabilistic multivariate L1-PCA model is introduced. Then, in Section 4, we show how the proposed multivariate L1-PCA model can be solved by the variational Bayesian technique and derive the variational EM-type algorithm for the model. In Section 5, we present the experimental results to evaluate the presented methods. Finally, in Section 6, we present our conclusions.

2 Multivariate Laplace Distribution

The multivariate Laplace distribution was first introduced by Eltoft et al. (2006) as a generalization of the standard univariate Laplace distribution. The standard univariate Laplace distribution is given by the following density function

$$p_{L^1}(y|\lambda) = \frac{1}{2}\sqrt{\frac{2}{\lambda}} \exp\left\{-\sqrt{\frac{2}{\lambda}}|y-\mu|\right\},$$

where μ is the mean and the variance $\sigma^2 = \lambda/2$ gives the variance of the distribution.

It was proved in (Pontil et al., 1998) that the above univariate Laplace distribution can be decomposed into a superposition of an infinite number of Gaussian distributions given by the following relation

$$p_{L^1}(y|\lambda) = \int_0^\infty \frac{1}{\sqrt{2\pi z}} \exp\left\{-\frac{(y-\mu)^2}{2z}\right\} p(z|\lambda)dz \tag{2.1}$$

where the random variable z (the variance of the Gaussian) follows the following Gamma distribution,

$$p(z|\lambda) = \frac{1}{\lambda} \exp\left\{-\frac{z}{\lambda}\right\}. \tag{2.2}$$

Following the relationship (2.1) it is easy to make a multivariate extension for the L1 distribution. Let $\mathbf{y}$ be a D-dimensional datum. Let us define a multivariate Gaussian with the probability density function (PDF) given as, conditioned on the mean and variance,

$$p_{\mathbf{Y}|z}(\mathbf{y}|\boldsymbol{\mu}, \Lambda, z) = \frac{1}{(2\pi z)^{(D/2)}} \exp\left\{-\frac{1}{2z}(\mathbf{y}-\boldsymbol{\mu})^\top \Lambda^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}, \tag{2.3}$$

where the vector $\boldsymbol{\mu}$ and the matrix Λ and z are fixed parameters. Λ satisfying the condition $\det \Lambda = 1$ defines the internal covariance structure of the components of vector $\mathbf{y}$ and z defines the variance scale (the determinant of covariance).

Let us define

$$q(\mathbf{y}) = (\mathbf{y} - \boldsymbol{\mu})^\top \Lambda^{-1} (\mathbf{y} - \boldsymbol{\mu}).$$

It can be shown that

$$\begin{aligned}\int_0^\infty p_{\mathbf{Y}|z}(\mathbf{y}|\boldsymbol{\mu}, \Lambda, z) p(z|\lambda) dz &= \int_0^\infty \frac{1}{(2\pi z)^{(D/2)}} \exp\left\{-\frac{1}{2z} q(\mathbf{y})\right\} p(z|\lambda) dz \\ &= \frac{1}{(2\pi)^{(D/2)}} \frac{2}{\lambda} \frac{K_{(D/2)-1}\left(\sqrt{\frac{2}{\lambda} q(\mathbf{y})}\right)}{\left(\sqrt{\frac{2}{\lambda} q(\mathbf{y})}\right)^{(D/2)-1}} \triangleq p_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\mu}, \Lambda, \lambda) \qquad (2.4)\end{aligned}$$

where $K_m(u)$ denotes the modified Bessel function of the second kind and order m, evaluated at u.

The function $p_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\mu}, \Lambda, \lambda)$ is called the multivariate Laplace distribution. It has three parameters λ, $\boldsymbol{\mu}$ and Λ. $\boldsymbol{\mu}$ is the mean of random vector $\mathbf{y}$ while Λ and λ determine its covariance through hidden variable z. When setting $D = 1$ in (2.4) we restore the relationship (2.1) for the univariate Laplace distribution. In this paper we will consider the newly defined multivariate distribution $p_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\mu}, \Lambda, \lambda)$. Equation (2.4) means the distribution can be decomposed into a superposition of an infinite number of Gaussian distributions, each of which is defined by (2.3), with a blending weight given by the Gamma distribution (2.2) where z is called hidden variable.

From the asymptotic property of the modified Bessel function of the second kind

$$K_m(u) \sim \sqrt{\frac{\pi}{2u}} \exp(-u), \quad \text{when } |u| \to \infty$$

we can see that

$$p_{\mathbf{y}}(\mathbf{y}) \sim \frac{\exp\left(-\sqrt{\frac{2}{\lambda} q(\mathbf{y})}\right)}{q(\mathbf{y})^{(D/2)-(1/2)}} \quad \text{for large } q(\mathbf{y}).$$

Thus the multivariate Laplace distribution is more heavy-tailed than the Gaussian distribution while it is slightly less heavy-tailed than the multivariate Student-t distribution. That means the Student-t distribution is more outlier-resistant than the multivariate Laplace distribution.

3 Multivariate L1-PCA Model

Let $Y = \{\mathbf{y}_i : i = 1, 2, \ldots, N\}$ be N independently and identically distributed random variables with values in $\mathbb{R}^D$. The model we consider assumes that each $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{iD})^\top \in \mathbb{R}^D$ can be additively decomposed as a linear latent variable model and noise given by

$$\mathbf{y}_i = \boldsymbol{\mu} + W\mathbf{x}_i + \boldsymbol{\epsilon}_i \qquad (3.1)$$

where $\boldsymbol{\mu}$ is the mean of the data Y, $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{id})^\top \in \mathbb{R}^d$ is the d-dimensional vector of latent variables, the $D \times d$ matrix W is called the loading factor, and $\boldsymbol{\epsilon}_i$ is a

vector of additive noise which follows the multivariate L1 distribution defined by (2.4) with $\mathrm{E}[\boldsymbol{\epsilon}_i] = 0$ and $\mathrm{Covar}[\boldsymbol{\epsilon}_i]$. That is, we consider the random noise $\boldsymbol{\epsilon}$ $(= \mathbf{y} - \boldsymbol{\mu} - W\mathbf{x})$ follows the special multivariate Laplace distribution $p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}|\mathbf{0}, \Lambda, \lambda)$ with the mean $\mathbf{0}$, see (2.4).

Assuming a normal distribution $\mathcal{N}(\epsilon_j|0, \sigma^2)$ for each of D error components ϵ_j $(j = 1, 2, ..., D)$, the model (3.1) leads to the standard probabilistic PCA (Tipping and Bishop, 1999). A well-known limitation of the Gaussian noise model is that it is not *robust* to the outliers in the observed data, and the accuracy of the linear latent variable model can be significantly compromised. The outliers perhaps represent corrupted observations or genuine samples from a heavy-tailed noise process. To cope with outliers in the data set, many people proposed Bayesian inference procedure by introducing a prior on the precision of the Gaussian, $\beta = \frac{1}{\sigma^2}$, defined by a Gamma distribution

$$p(\beta|a, b) = \Gamma(\beta|a, b) = \frac{b^a}{\Gamma(a)}\beta^{a-1}\exp(-b\beta)$$

whose mean $E(\beta) = a/b$ and the variance $\mathrm{var}(\beta) = a/b^2$.

The above hierarchical specification leads to the so-called Student-t robust PCA models, see (Peel and McLachlan, 2000; Khan and Dellaert, 2004; Tipping and Lawrence, 2005; Archambeau et al., 2006).

When the data contain outliers, a learning algorithm that estimates the model parameters must either eliminate outliers from the data before modelling, or model the outliers explicitly. The Student-t PCA takes the latter approach in which a mixture of infinite number of Gaussian is used to model possible outliers in the data set. In the same sense, the multivariate Laplacian distribution is expected to combat with outliers too due to its heavy-tailed property.

Based on (2.4), the joint distribution of all the components of $\boldsymbol{\epsilon}$ can be written as, in the form of likelihood for data $\mathbf{y}$,

$$p(\mathbf{y}|\boldsymbol{\mu}, W, \mathbf{x}, \Lambda, \lambda) = \frac{1}{(2\pi)^{(D/2)}}\frac{2}{\lambda}\frac{K_{(D/2)-1}\left(\sqrt{\frac{2}{\lambda}q(\mathbf{y})}\right)}{\left(\sqrt{\frac{2}{\lambda}q(\mathbf{y})}\right)^{(D/2)-1}}, \tag{3.2}$$

where $q(\mathbf{y}) = (\mathbf{y} - \boldsymbol{\mu} - W\mathbf{x})^\top\Lambda^{-1}(\mathbf{y} - \boldsymbol{\mu} - W\mathbf{x})$ and $\det\Lambda = 1$.

To develop a generative Bayesian model, we further suppose a prior on the latent variable

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|0, I_d) = \frac{1}{(2\pi)^{d/2}}\exp\left\{-\frac{1}{2}\mathbf{x}^\top\mathbf{x}\right\}. \tag{3.3}$$

For the sake of simplicity, we impose a prior on λ in (3.2). The prior is given by an inverse Gamma distribution

$$p_\lambda(\lambda|a_\lambda, b_\lambda) = \frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)}\lambda^{-a_\lambda-1}\exp\{-\frac{b_\lambda}{\lambda}\} \tag{3.4}$$

The distribution specifications (3.2) - (3.4) are for any general datum $\mathbf{y}$. At each datum $\mathbf{y}_i$ of the given dataset Y, we use the multivariate Laplace distribution given

by (3.2), which through (2.4) can be decomposed into a superposition of Gaussian in terms of the Gamma distribution $p(z_i|\lambda_i)$ as defined in (2.2), with individual covariance determinant z_i but a common normalized covariance Λ satisfying $\det\Lambda = 1$, and each λ_i is a hidden random parameter determined by an inverse Gamma distribution with hyperparameters a_{λ_i} and b_{λ_i} as (3.4). Thus the model offers more flexible covariance at each datum.

For the dataset $Y = \{\mathbf{y}_i\}_{i=1}^N$, we have introduced the hidden variables $Z = \{z_i\}_{i=1}^N$ and the hidden parameters $\boldsymbol{\lambda} = \{\lambda_i\}_{i=1}^N$ via the superposition of scaled Gaussian as defined in (2.4). Denote by $X = \{\mathbf{x}_i\}_{i=1}^N$ the latent variable for the dataset. Then by taking the priors (3.3) and (3.4) into account, we have the following joint distribution of $Y, X, Z, \boldsymbol{\lambda}$

$$\begin{aligned} P(Y,X,Z,\boldsymbol{\lambda}|\boldsymbol{\mu},W,\Lambda) = &\prod_{i=1}^N \frac{1}{(2\pi z_i)^{D/2}} \exp\left\{-\frac{1}{2z_i}(\mathbf{y}_i-\boldsymbol{\mu}-W\mathbf{x}_i)^\top\Lambda^{-1}(\mathbf{y}_i-\boldsymbol{\mu}-W\mathbf{x}_i)\right\} \\ &\frac{1}{\lambda_i}\exp\left\{-\frac{z_i}{\lambda_i}\right\}\frac{1}{(2\pi)^{d/2}}\exp\left\{-\frac{1}{2}\mathbf{x}_i^\top\mathbf{x}_i\right\}p(\lambda_i|a_{\lambda_i},b_{\lambda_i}) \end{aligned} \tag{3.5}$$

In the above model construction, we consider W as normal parameters. In fact, we may view W as random variables as well. The Bayesian methodology requires a suitable choice of prior for W, and then proceeds to treat the parameters as hidden variables as well. A simple choice of prior would be a spherical Gaussian distribution:

$$p(W|\boldsymbol{\nu}) = \prod_{j=1}^D \mathcal{N}(\mathbf{w}_j|0,\nu_j^{-1}I_d), \tag{3.6}$$

where $\mathbf{w}_j$ is the jth row of the matrix W and $\boldsymbol{\nu} = \{\nu_1,\nu_2,...,\nu_D\}$ and each ν_j has a prior given by the Gamma distribution as follows,

$$p(\nu_j) = \Gamma(\nu_j|a_{\nu_j},b_{\nu_j}) = \frac{b_{\nu_j}^{a_{\nu_j}}}{\Gamma(a_{\nu_j})}\nu_j^{a_{\nu_j}-1}\exp\{-b_{\nu_j}\nu_j\}, \tag{3.7}$$

where $j = 1,2,...,D$.

Thus for the given observations Y another generative model can be defined as follows

$$\begin{aligned} P(Y,X,Z,\boldsymbol{\lambda},W|\boldsymbol{\mu},\boldsymbol{\nu},\Lambda) = &\prod_{i=1}^N \frac{1}{(2\pi z_i)^{D/2}} \exp\left\{-\frac{1}{2z_i}(\mathbf{y}_i-\boldsymbol{\mu}-W\mathbf{x}_i)^\top\Lambda^{-1}(\mathbf{y}_i-\boldsymbol{\mu}-W\mathbf{x}_i)\right\} \\ &\frac{1}{\lambda_i}\exp\left\{-\frac{z_i}{\lambda_i}\right\}\frac{1}{(2\pi)^{d/2}}\exp\left\{-\frac{1}{2}\mathbf{x}_i^\top\mathbf{x}_i\right\}p(\lambda_i|a_{\lambda_i},b_{\lambda_i}) \\ &\prod_{j=1}^D \mathcal{N}(\mathbf{w}_j|0,\nu_j^{-1}I_d)p(\nu_j). \end{aligned} \tag{3.8}$$

Unfortunately marginalization of all the hidden/latent variables is intractable, i.e., no analytical forms available for $p(Y)$. If we wish to proceed, we need to turn to an approximate method. We are going to look at the variational Bayesian inference method. But as an example, we only work on the algorithm for the model (3.5). The similar deduction can apply to the model (3.8).

4 Variational Approximation for Multivariate L1-PCA

Consider (3.5). Let $\theta = \{\boldsymbol{\mu}, W, \Lambda\}$ the parameters in the model. For the given observed data Y, maximizing the likelihood $P(Y|\theta)$ as a function of θ is equivalent to maximising the log likelihood:

$$\mathcal{L}(\theta) = \log P(Y|\theta) = \log \int_X \int_Z \int_{\boldsymbol{\lambda}} P(Y, X, Z, \boldsymbol{\lambda}|\boldsymbol{\mu}, W, \Lambda) dX\, dZ\, d\boldsymbol{\lambda}$$

where we call X, Z and $\boldsymbol{\lambda}$ the hidden/latent variables. It can be proved that, for any distributions $Q(X, Z, \boldsymbol{\lambda})$ (called variational distributions), the following inequality is valid,

$$\begin{aligned} \mathcal{L}(\theta) \geq & \int_X \int_Z \int_{\boldsymbol{\lambda}} Q(X, Z, \boldsymbol{\lambda}) \log P(Y, X, Z, \boldsymbol{\lambda}|\theta) dX\, dZ\, d\boldsymbol{\lambda} \\ - & \int_X \int_Z \int_{\boldsymbol{\lambda}} Q(X, Z, \boldsymbol{\lambda}) \log Q(X, Z, \boldsymbol{\lambda}) dX\, dZ\, d\boldsymbol{\lambda} := \mathcal{F}(Q(X, Z, \boldsymbol{\lambda}), \theta) \end{aligned} \tag{4.1}$$

The quantity $\mathcal{F}(Q(X, Z, \boldsymbol{\lambda}), \theta)$ is referred to as the negative free energy in statistical physics and can be considered to be a lower bound of $\mathcal{L}$. The purpose of the variational approximation is to maximize $\mathcal{F}$ with respect to both Q functions and the parameters θ, instead of maximizing $\mathcal{L}$ with respect to θ. This can be done in two-stages procedure: in the E-step, maximize $\mathcal{F}$ with respect to $Q(X, Z, \boldsymbol{\lambda})$ when fixing θ at the current value; and in the M-step, maximize $\mathcal{F}$ with respect to θ when fixing Q at the approximated distribution obtained in the E-step.

In implementation, a basic simple stage is to separate the dependence of the hidden variables X, Z and $\boldsymbol{\lambda}$, that is, to assume that $Q(X, Z, \boldsymbol{\lambda}) = Q(X)Q(Z)Q(\boldsymbol{\lambda})$. For the multivariate L1-PCA model, maximizing the variational functional $\mathcal{F}$ with respect to $Q(X)$, $Q(Z)$ and $Q(\boldsymbol{\lambda})$ as well as all the parameters results in the following approximated individual distributions at each datum and the optimal parameters: (Note: In the following steps, we denote by $\overline{u}$ the mean of u and by $\underline{u}$ the mean value of $\frac{1}{u}$ with respect to the approximated posterior $Q(u)$.)

1. The best $Q(\mathbf{x}_i)$ is a Gaussian given by

   $$\mathcal{N}(\mathbf{x}_i|\overline{\mathbf{x}}_i, \Sigma_i)$$

   where $\Sigma_i = (I_d + \underline{z}_i W^\top \Lambda^{-1} W)^{-1}$ and $\overline{\mathbf{x}}_i = \underline{z}_i \Sigma_i W^\top \Lambda^{-1}(\mathbf{y}_i - \boldsymbol{\mu})$ and both Λ and $\boldsymbol{\mu}$ are the updated optimal value from Steps 5 and 6 respectively.

2. The best $Q(z_i)$ is the Generalized Inverse Gaussian (GIG) distribution given by

   $$Q(z_i) \propto z_i^{(1-D/2)-1} \exp\left\{-\frac{1}{2}\left(\frac{\tilde{q}(\mathbf{y}_i)}{z_i} + 2\underline{\lambda}_i z_i\right)\right\}$$

   where

   $$\tilde{q}(\mathbf{y}_i) = (\mathbf{y}_i - \boldsymbol{\mu} - W\overline{\mathbf{x}}_i)^\top \Lambda^{-1}(\mathbf{y}_i - \boldsymbol{\mu} - W\overline{\mathbf{x}}_i) + \mathrm{tr}(W^\top \Lambda^{-1} W \Sigma_i).$$

3. The best $Q(\lambda_i)$ is still an inverse Gamma distribution but with the updated parameters

$$a_{\lambda_i} \Leftarrow a_{\lambda_i} + 1$$
$$b_{\lambda_i} \Leftarrow b_{\lambda_i} + \overline{z}_i$$

4. In the M-step, the W can be optimized by

$$\max_W \mathcal{L}(W) = -\frac{1}{2}\sum_{i=1}^{N} \underline{z}_i[\tilde{q}(\mathbf{y}_i) + \mathrm{tr}(W^\top \Lambda^{-1} W \Sigma_i)] + \text{const.}$$

The solution of the above optimal problem is given by

$$W = \left(\sum_{i=1}^{N} \underline{z}_i(\mathbf{y}_i - \boldsymbol{\mu})\mathbf{x}_i^\top\right)\left(\sum_{i=1}^{N} \underline{z}_i \mathbf{x}_i \mathbf{x}_i^\top + \underline{z}_i \Sigma_i\right)^{-1}.$$

5. The formula for the optimal value of Λ is

$$\Lambda = \frac{\sum_{i=1}^{N} \underline{z}_i(W\Sigma_i W^\top + \Delta_i)}{\det \sum_{i=1}^{N} \underline{z}_i(W\Sigma_i W^\top + \Delta_i)}$$

where $\Delta_i = (\mathbf{y}_i - \boldsymbol{\mu} - W\overline{\mathbf{x}}_i)(\mathbf{y}_i - \boldsymbol{\mu} - W\overline{\mathbf{x}}_i)^\top$.

6. The updated rule for $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^{N} \underline{z}_i(\mathbf{y}_i - W\overline{\mathbf{x}}_i)}{\sum_{i=1}^{N} \underline{z}_i}.$$

The above procedure has an intuitive interpretation: in the E-step we update one of Q-distributions when keeping others fixed; in M-step we fix all the distributions over the hidden variables and update the W by minimizing robust reconstruction error of the data points.

5 Experiments

In our experiments, for the Laplacian noise process, we choose $a_\lambda = 0.04$ and $b_\lambda = 0.01$, so as to specify a mean of 0.5 for the prior over the standard derivation.

We then cycled through the Q-distribution updates, starting with $Q(\mathbf{x})$. The initial value for W is randomly chosen but according to the scale of the training data in our implementation. The initial value for z is set to 1. For every three sets of updates for all the distributions over the hidden variables, we performed only one update for W in the M-step to speed up the procedure.

We chose to terminate the iterative procedure when there is only little change occuring at each update to the Q-distribution with a tolerance 10^{-6}.

After the estimated W is obtained from the variational procedure, we apply the singular value decomposition (SVD) to $W = USV$ and define the column of U as the principal components of the robust multivariate L1-PCA.

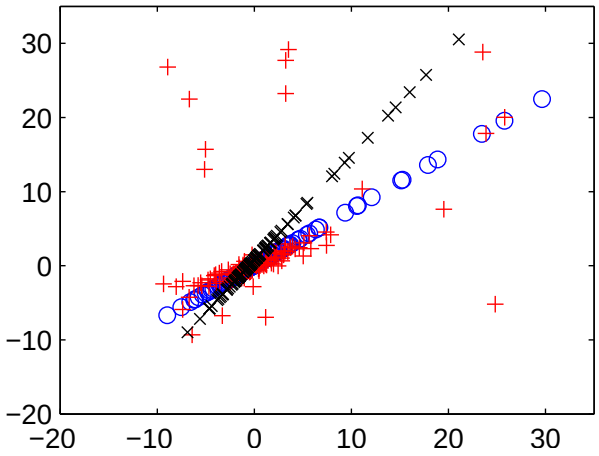

Figure 1: Data set (the red +s) with 20 Outliers (uniform random noise in the interval $(-10\ 30)$ in each direction). The data is drawn from a 2-dimensional Gaussian distribution. Reconstructed Data based on the first Principal Components from two PCA methods. The circled data points are reconstructed from the multivariate L1-PCA with $d = 1$ and the crossed data points are reconstructed from the first principal component picked up from the standard PCA. Due to the presence of the outliers, the standard PCA revealed the incorrect principal components

5.1 Synthetic Data

We first demonstrate the performance of the algorithm on bivariate synthetic data. The data set consists of 120 data, 100 of which were sampled from a Gaussian density with mean $\mu = (0, 0)^\top$ and covariance $\Sigma = \begin{pmatrix} 10 & 5 \\ 5 & 3 \end{pmatrix}$, and the other 20 regarded as corrupted data were sampled from uniform distributions over the range of -10 to 30 along each dimension. The data are shown in Figure 1 as "+".

Ideally we hope the extra 20 data won't have much impact on the model as the majority of data come from a Gaussian. However it is clear from Figure 1 that the standard PCA attempts to model all the dataset including outlier data as a whole and thus incorrectly picks up the first principal component. The data points marked "cross" and "circle" are reconstructed from the first principal components of the standard PCA and multivariate L1-PCA with $d = 1$, respectively. The direction specified by the circled points reflects the principal direction determined by the majority of data except for the outliers while the direction revealed by the standard PCA was misled by the outlier data.

In the experiment, we also noted that the $\overline{z_i}$ (the mean of z_i) associated with the extra outliers is significantly greater than that for the other points while $\underline{z_i}$ (the mean of $1/z_i$) is much smaller than that for the outlier points. Taking $\underline{z_i}$ or $\overline{z_i}$ as an indicator we may drop any possible outliers in the earlier stage of the procedure. However in order to demonstrate the robustness of the model we didn't remove those unveiled outliers in the procedure in all the following experiments.

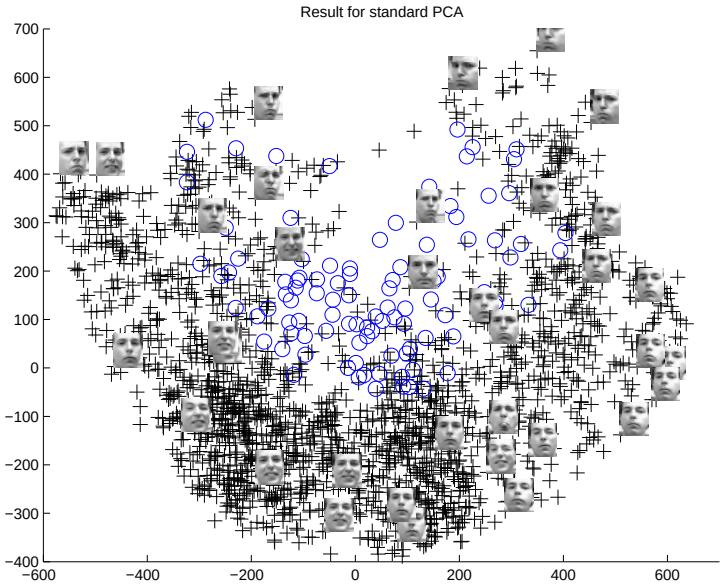

Figure 2: The result of dimensionality reduction given by the standard PCA algorithm: Majority of the corrupted data are not clearly revealed in the projected space because as in the synthetic example the model fits the entire dataset as a whole to find principal space determined by data.

5.2 Dimensionality Reduction for Human Faces Data

To further assess the performance of multivariate L1-PCA on practical data sets, we applied the proposed algorithm to the Frey facial image dataset. The dataset consists of 1965 images (each in 20×28 grayscale pixels) of a single person's face extracted from a digital movie. Many researchers have noted that there are three main intrinsic features among the facial image sequences: head rotation, expression at the mouth and head illumination. Majority of dimensionality reduction algorithms can reveal the intrinsic features. However in this experiment we want to demonstrate that the multivariate L1-PCA can separate outliers while it maintains good exposure of intrinsic features.

To construct outlier images, we randomly chose 100 facial images from the dataset and corrupted them with noises. The noise added is generated from a uniform distribution on the range $[10, 300]$ and the corrupted images are scaled to the standard gray levels from 0 to 255. Then we use the entire dataset and the corrupted images to test the proposed algorithm and compare it with others.

First we run the conventional PCA on the new dataset. We tested the PCA for a latent dimension of $d = 6$ but for the sake of simplicity we plot projection of the facial image data on the first two leading principal components in Figure 2. The circles in the figure correspond to the corrupted facial images. From the figure we can clearly see that the PCA failed to separate the data from the outliers. On the figure we randomly picked 36 projected points and displayed their original facial images. Although the first principal component has revealed the head rotation, it is hard to detect any significant information revealed by the facial expression.

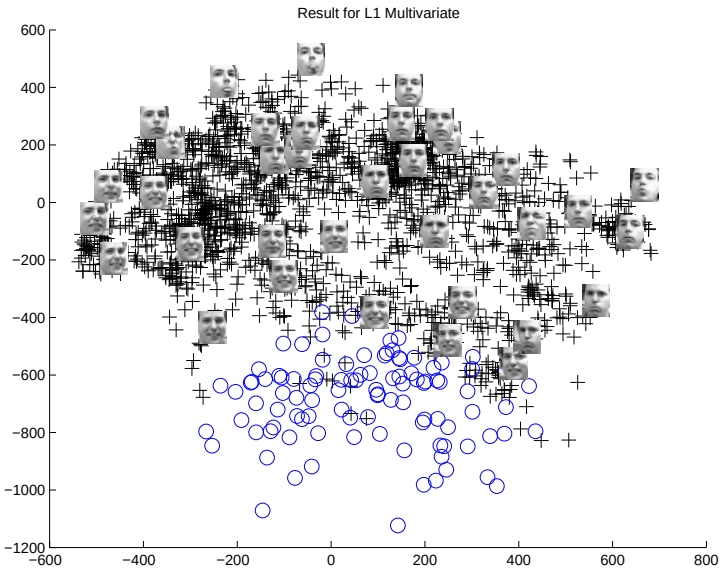

Figure 3: The result of dimensionality reduction given by the the multivariate L1-PCA algorithm

In the experiment of the multivariate L1-PCA for this database, similarly we took the latent dimension $d = 6$ and the initial values for all the parameters were chosen in the same way as that of the experiment for the 2-D synthetic dataset. However for the number of iteration k, we tested a range of values between 20 to 100 and found that $k = 36$ gave a better visual result. In fact, after 35 iterations, most of the parameter values in the Q distributions as well as the likelihood do not alter much.

In Gao (2008) we demonstrated the robust L1-PCA is comparable to the student-t PCA. In terms of dimensionality reduction, we compared the performance of the Student-t PCA, the robust L1-PCA and the multivariate L1-PCA algorithms in the existence of outliers. We run all the three algorithms under the same initial conditions with $d = 6$, that is, to extract 6 components. We visualized the results by showing the first two components in Figures 3 - 5. From these figures, we can see that noised data are significantly separated from the main data. However in this experiment the robust L1-PCA does not perform well, see Figure 4. Of course, a lot of reasons may be behind this. The multivariate L1-PCA models possible covariance between components while both the robust L1-PCA and the Student-t PCA make assumptions that the covariance of approximate Gaussian at each point is diagonal along each dimension and a uniform precision for all dimensions is applied in student-t EM algorithm, see (Khan and Dellaert, 2004). Similarly we can use the values of $\underline{z}_i$ as indicators for outliers in this procedure. In our experiments we observed that the $\underline{z}_i$ values for outlier data have been reduced to less than 10^{-5} after several iterations.

In all the three figures we have randomly chosen 36 points and displayed the corresponding face images on the plot. It is very clear that all three algorithms have revealed the head rotation in their first principal component. And we also noted that the result from the multivariate L1-PCA is more interesting. If we project the first two compo-

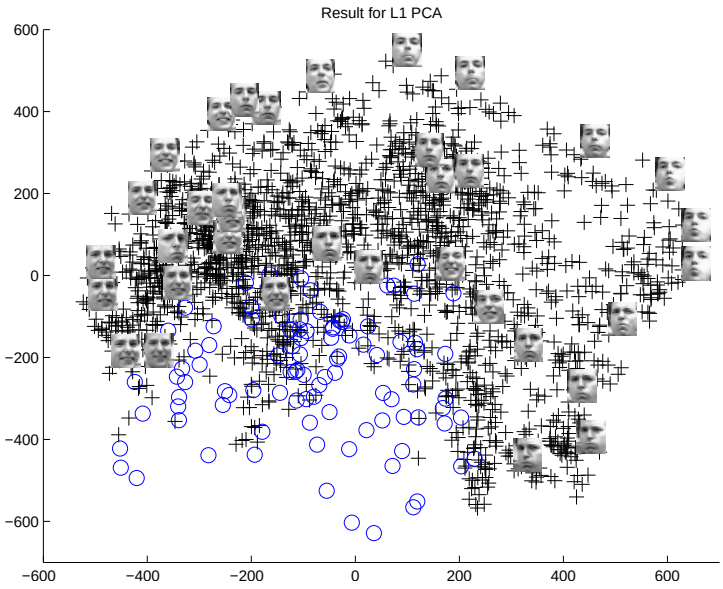

Figure 4: The result of dimensionality reduction given by the the robust L1-PCA algorithm

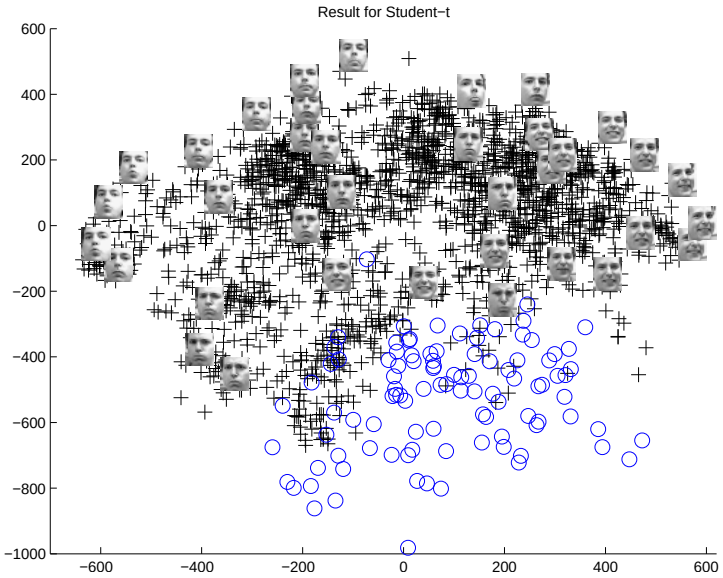

Figure 5: The result of dimensionality reduction given by the the Student-t PCA algorithm

Table 1: Comparison of running time (in seconds) per outer iteration of three methods averaged over 35 iterations of 10 random dataset partitionings

Student-t PCA	Robust L1-PCA	Multivariate L1-PCA
42.14	40.82	43.05

nents onto the direction $e = (1, 1)$ on Figure 3, we can easily observe the change in facial expression from smiling to unhappy. The other two algorithms also demonstrate similar patterns, however the projected components for outlier images given by the other two algorithms are a bit mixed with the components of the major population while the separation from outliers given by the multivariate L1-PCA is much clearer.

The computational complexity of the three algorithms are comparable. The main overhead in three algorithms is finding the matrix inverse of $D \times D$ matrices in inner iterations corresponding to each data points in the outer variational iterative process. Table 1 summarizes the average running time per outer iteration over 35 iterations of 10 random dataset partitionings of the above facial image dataset.

6 Conclusions

Many probabilistic models strongly rely on a Gaussian assumption. In practice, however, this crude assumption may seem unrealistic as the resulting models are very sensitive to non-Gaussian noise processes. A possible approach is to employ kinds of non-Gaussian distribution especially heavy-tailed distributions such as L1 Laplacian densities.

In this paper, we have shown that the robust multivariate L1-PCA can be constructed and solved under the framework of general variational Bayesian learning and inference. Recently, an increasing number of works have used a similar approach in other contexts (Peel and McLachlan, 2000; Archambeau, 2005) where the student-t distribution are used to reduce the effect of outliers. As a result, the lower-dimensional latent subspace is recovered with a higher confidence.

In order to find tractable solutions for the model parameters in multivariate L1-PCA, we use the definition of the multivariate Laplacian distribution as a superposition of infinite number of Gaussian with precisions controlled by another distribution. It enables us to employ the variational approximation to posterior densities of all the uncertainties involved in the model. The algorithm has been designed based on the variational version of EM scheme. The approach works well on several illustrative and practical examples. Of course, the models could be further tested on other real world data sets. Future work includes the extension of the multivariate L1 scheme to our Twin Kernel Embedding algorithm for the purpose of nonlinear dimensionality reduction or embedding of non-vectorial data objects (Guo et al., 2008).

Acknowledgements

The authors are grateful to anonymous reviewers for their constructive suggestions. This work is supported by the National Natural Science Foundation of China (NSFC 60373090), the ARC DP Development Grant from Charles Sturt University.

References

Archambeau, C. (2005). *Probabilistic models in noisy environments and their application to a visual prosthesis for the blind*. Doctoral dissertation, Université Catholique de Louvain, Belgium.

Archambeau, C., N. Delannay, and M. Verleysen (2006). Robust probabilistic projections. In *Proceedings of the 23rd International Conference on Machine Learning*, Puttsburgh, PA.

Baccini, A., P. Besse, and A. deFalguerolles (1996). A L1-norm PCA and a heuristic approach. In E. Diday, Y. Lechevalier, and O. Opitz (Eds.), *Ordinal and Symbolic Data Analysis*, pp. 359–368. Springer.

de la Torre, F. and M. Black (2001). Robust principal component analysis for computer vision. *International Conference on Computer Vision 52*, 362–369.

de la Torre, F. and M. Black (2003). A framework for robust subspace learning. *International Journal of Computer Vision 54*(1), 117 – 142.

Ding, C., D. Zhou, X. He, and H. Zha (2006). R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23 rd International Conference on Machine Learning*, Pittsburgh, PA.

Eltoft, T., T. Kim, and T. Lee (2006). On the multivariate laplace distribution. *IEEE Signal Processing Letters 13*(5), 300–303.

Gao, J. (2008). Robust L1 principal component analysis and its bayesian variational inference. *to appear in Neural Computation* vol 20(2), xxx–xxx.

Guo, Y., J. B. Gao, and P. W. Kwan (2008). Twin kernel embedding. *submitted to IEEE Transaction on PAMI xx*, under third round review.

Ke, Q. and T. Kanade (2005). Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *Proceedings of CVPR*, Volume I, pp. 739–746.

Khan, Z. and F. Dellaert (2004). Robust generative subspace modelling: the subspace t distribution. Technical Report GIT-GVU-04-11, Georgia Institute of Technology.

Ng, A. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of Intl Conf. Machine Learning*.

Peel, D. and G. McLachlan (2000). Robust mixture modelling using the t distribution. *Statistic and Computing 10*, 339–348.

Pontil, M., S. Mukherjee, and F. Girosi (1998). On the noise model of support vector machine regression. A.I. Memo 1651, AI Laboratory, MIT.

Ridder, D. D. and V. Franc (2003). Robust subspace mixture models using t-distributions. In R. Harvey and A. Bangham (Eds.), *BMVC 2003: Proceedings of the 14th British Machine Vision Conference*, London, UK, pp. 319–328. BMVA.

Rubin, D. (2006). *Iteratively reweighted least squares*, Volume 4 of *Encyclopedia of Statistical Sciences*, pp. 272–275. Wiley.

Ruymagaart, F. (1981). A robust principal component analysis. *Journal of Multivariate Analysis 11*, 485–497.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B. 58*, 267–288.

Tipping, M. and C. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B 6*(3), 611–622.

Tipping, M. and N. Lawrence (2005). Variational inference for Student-t models: Robust Bayesian interpolation and generalized component analysis. *NeuroComputing 69*, 123–141.