Neural Processing Letters manuscript No.
(will be inserted by the editor)

Bayesian robust PCA of incomplete data

Jaakko Luttinen · Alexander Ilin · Juha Karhunen

Received: date / Accepted: date

Abstract We present a probabilistic model for robust factor analysis (FA) and principal component analysis (PCA) in which the observation noise is modeled by Student-*t* distributions in order to reduce the negative effect of outliers. The Student-*t* distributions are modeled independently for each data dimensions, which is different from previous works using multivariate Student-*t* distributions. We give a unifying comparison on using the proposed noise distribution, the multivariate Student-*t* and the Laplace distribution. Intractability of evaluating the posterior probability density is solved using variational Bayesian approximation methods. We demonstrate that the assumed noise model can yield accurate reconstructions because corrupted dimensions of a bad quality sample can be reconstructed using the other dimensions of the same data vector. Experiments on an artificial dataset and a weather dataset show that the dimensional independency and the flexibility of the proposed Student-*t* noise model can make it superior in some applications.

Keywords Variational Bayesian methods · Principal component analysis · Factor analysis · Robustness · Outliers · Missing values

1 Introduction

Principal component analysis (PCA) [4, 7, 10, 12] is a classical data analysis method which is optimal for compressing data in the mean squared error sense. Dropping the dimensionality of the data using PCA is useful in many cases for avoiding overlearning, suppressing noise, and decreasing the computational load of subsequent processing.

PCA can be derived as an optimal solution to the quadratic criteria of variance maximization and minimization of mean squared error, therefore it can be sensitive

E-mail: firstname.lastname@aalto.fi
Fax: +358-9-470-23277
Aalto University School of Science
Department of Information and Computer Science
PO Box 15400
FI-00076 Aalto
Finland

to outliers in the data. Robust PCA techniques have been introduced to cope with this problem, see, for example, [7] and the references therein. The basic idea in these robust PCA methods is to replace quadratic criteria of standard PCA by more slowly growing criteria. In [5, 6, 19], the data matrix is decomposed into a sum of a low-rank and a sparse matrix, which are computed using convex optimization methods.

PCA can be interpreted probabilistically as a latent variable model [3, 17, 18]. While it is a rather simplistic model based on Gaussian assumptions, it can be used as a basis for building probabilistic extensions of classical PCA. The popularity of probabilistic models is due to their principled way to cope with overfitting problems, to do model comparison and to handle missing values and uncertainty.

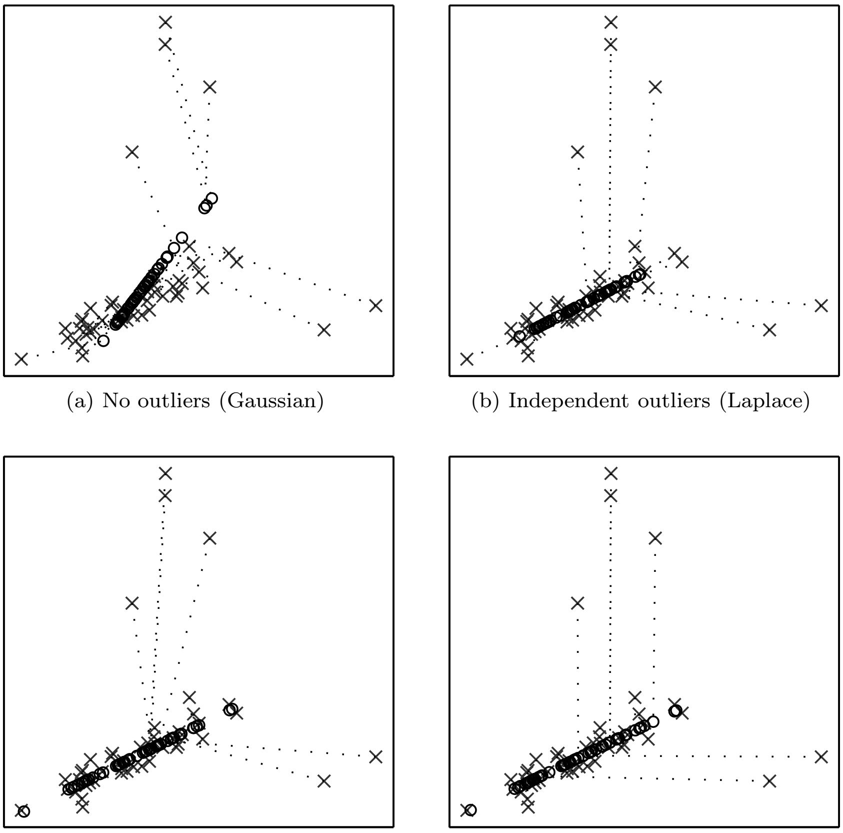
Probabilistic models for robust PCA have been introduced recently [1, 9, 13, 20]. They treat possible outliers by using heavy-tailed distributions (instead of the Gaussian distribution) for describing the noise. The Laplace distribution was used in [9], while the multivariate Student-t distribution was used in [1, 13, 20]. Instead of heavy-tailed distributions, [8] models outliers as a sparse matrix with the non-zero elements from a Gaussian distribution with a large variance.

This paper presents a new robust PCA and factor analysis (FA) model using the Student-t distribution for describing the noise. One of the important assumptions of our model is that the outliers can arise independently in each sensor, that is, for each component of a data vector. This is different to the previously introduced Student-t models [1, 13, 20] which assume that all elements of an outlier data vector are corrupted. The Student-t distribution also includes a parameter that allows to vary the "degree of robustness", which provides more flexibility compared to the models based on the Laplace distribution [9].

The new model was motivated by our analysis of a Southern Finland weather data set, which had a large amount of missing and corrupted data. The corrupted measurements arise independently at each weather stations instead of all the stations being corrupted simultaneously, thus the existing models using Student-t distribution were unrealistic for this application. The proposed method is able to model the independency of the outliers between the stations, which improves the results remarkably.

Fig. 1 illustrates the difference in the principal subspace estimation and sample reconstruction assuming different noise models. There we use a two-dimensional dataset with a single principal direction and a few outliers. The principal subspace found by assuming Gaussian noise is affected by outliers (Fig. 1a), whereas robust techniques are able to find the correct principal subspace. However, the reconstructions are quite different depending on the noise distribution. The multivariate Student-t model (Fig. 1c) assumes fully corrupted outliers, which give no information about the true values of the vectors, thus the outliers are reconstructed close to the mean. The independent Student-t (Fig. 1d) and Laplace models (Fig. 1b) assume partially corrupted data vectors, which makes it possible to ignore the corrupted dimension and reconstruct it based on the uncorrupted dimension. However, because the Laplace distribution has only one parameter to control the general noise level and the noise level of the outliers, the outliers cause the reconstruction of non-corrupted samples to be regularized too much.

We apply the proposed model to an artificial dataset and a badly corrupted real-world weather dataset. A comparison to other related models shows that both the element-wise independency of the outliers and the flexibility of the Student-t



(a) No outliers (Gaussian)

(b) Independent outliers (Laplace)

(c) Multivariate outliers (Student-t)

(d) Independent outliers (Student-t)

Fig. 1: Principal subspace estimation using PCA with different assumptions about outliers in the data: (a) no outliers, (b) partially (or fully) corrupted outliers, (c) fully corrupted outliers, and (d) partially corrupted outliers. The crosses represent data points and the circles show their projections onto the found principal subspace.

distribution may be necessary properties for the model in order to achieve good performance. We show how to learn the proposed model from datasets with missing values and illustrate that the proposed model gives reasonable reconstructions of both missing values and outliers. This work extends our previous works [11] by handling both outliers and missing values simultaneously.

The paper is organized as follows. Section 2 presents the existing probabilistic model for factor analysis and principal component analysis. The novel noise model is presented in Section 3 along with the existing noise models. Section 4 shows the equations for the variational Bayesian inference. Section 5 compares the models using an artificial dataset and a badly corrupted real-world weather dataset. The paper ends with conclusions in Section 6.

2 Factor analysis

Denote by $\{\mathbf{y}_n\}_{n=1}^N$ the set of M-dimensional data vectors (observations) $\mathbf{y}_n$. We assume that the data vectors are generated from N hidden D-dimensional latent variable vectors $\{\mathbf{x}_n\}_{n=1}^N$ using the transformation

$$\mathbf{y}_n = \mathbf{W}\mathbf{x}_n + \boldsymbol{\mu} + \boldsymbol{\epsilon}_n .$$

where $\mathbf{W}$ is an $M \times D$ loading matrix, $\boldsymbol{\mu}$ is a bias term and $\boldsymbol{\epsilon}_n$ is a noise vector. Usually, the dimensions fulfil $D < M < N$.

The noise is modeled with independent Gaussian distributions for each element of $\boldsymbol{\epsilon}_n$. Assuming that some of the elements of $\mathbf{y}_n$ are missing, this can be written as

$$p(\boldsymbol{Y}|\mathbf{W},\mathbf{X},\boldsymbol{\mu},\boldsymbol{\tau}) = \prod_{mn\in\mathcal{O}} \mathcal{N}(y_{mn}|\mathbf{w}_m^{\mathrm{T}}\mathbf{x}_n + \mu_m, \tau_m^{-1}) ,$$

where $\mathcal{N}(y|\mu,\sigma^2)$ denotes the Gaussian probability density function (pdf) with mean μ and (co)variance σ^2, $\mathcal{O}$ is the set of indices mn for which y_{mn} are observed and $\mathbf{w}_m^{\mathrm{T}}$ is the m-th row of $\mathbf{W}$. Here we denote by $\boldsymbol{Y} = \{y_{mn}|mn \in \mathcal{O}\}$ all observed data, $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_N]$ is the matrix containing the latent variables $\mathbf{x}_n$, and $\boldsymbol{\tau}$, $\boldsymbol{\mu}$ are vectors with elements τ_m, μ_m respectively.

The prior models for $\mathbf{X}$, $\mathbf{W}$ and $\boldsymbol{\mu}$ are similar to the extension of probabilistic PCA from [3]:

$$\begin{aligned} p(\mathbf{X}) &= \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\mathbf{0},\mathbf{I}), \\ p(\mathbf{W}|\boldsymbol{\tau},\boldsymbol{\alpha}) &= \prod_{m=1}^{M}\prod_{d=1}^{D} \mathcal{N}(w_{md}|0,\tau_m^{-1}\alpha_d^{-1}), \\ p(\boldsymbol{\mu}|\boldsymbol{\tau},\beta) &= \prod_{m=1}^{M} \mathcal{N}(\mu_m|0,\tau_m^{-1}\beta^{-1}). \end{aligned}$$

Having $\boldsymbol{\alpha} = [\alpha_1,\ldots,\alpha_D]$ in the hierarchical prior of $\mathbf{W}$ diminishes overfitting and helps finding automatically the dimensionality of the principal subspace [11]. Parameters τ_m are used in the prior of $\mathbf{W}$ and $\boldsymbol{\mu}$ as discussed in [21]. It would be possible to use more structured and complex priors for $\mathbf{W}$ and $\mathbf{X}$ (see, e.g., [2, 15]), but because the focus of this paper is in the noise distribution and not in $\mathbf{W}$ nor $\mathbf{X}$, we settle for this simple factor analysis model.

The noise parameter $\boldsymbol{\tau}$ and the hyperparameters $\boldsymbol{\alpha}$ and β are assigned conjugate priors

$$\begin{aligned} p(\boldsymbol{\tau}) &= \prod_{m=1}^{M} \mathcal{G}(\tau_m|a_{\boldsymbol{\tau}},b_{\boldsymbol{\tau}}), \\ p(\boldsymbol{\alpha}) &= \prod_{d=1}^{D} \mathcal{G}(\alpha_d|a_{\boldsymbol{\alpha}},b_{\boldsymbol{\alpha}}), \\ p(\beta) &= \mathcal{G}(\beta|a_\beta,b_\beta), \end{aligned}$$

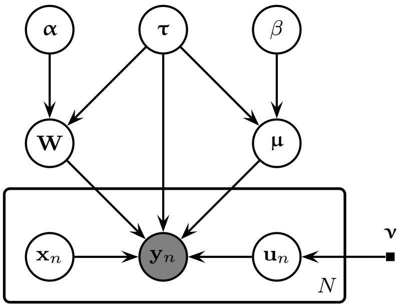


Fig. 2: A graphical model of the robust probabilistic PCA.

where $\mathcal{G}(\tau|a,b)$ is the gamma pdf with shape a and inverse scale b. One can use, for instance, $a_{\boldsymbol{\tau}} = b_{\boldsymbol{\tau}} = a_{\boldsymbol{\alpha}} = b_{\boldsymbol{\alpha}} = a_\beta = b_\beta = 10^{-5}$ in order to obtain broad distributions. In addition, it is possible to use a common noise level $\tau_m = \tau$ in order to obtain PCA model with isotropic noise.

3 Robust factor analysis

In order to make the factor analysis model robust to outliers, one can use heavy-tailed distributions to model the noise. Previous works include using the multivariate Student-t and the Laplace distributions [1, 9, 13, 20]. We propose to model the noise $\boldsymbol{\epsilon}_n$ with independent Student-t distributions for each element of the vector. This can be written as

$$p(\boldsymbol{Y}|\mathbf{W},\mathbf{X},\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\nu}) = \prod_{mn\in\mathcal{O}} \mathcal{S}(y_{mn}|\mathbf{w}_m^{\mathrm{T}}\mathbf{x}_n + \mu_m, \tau_m^{-1}, \nu_m)\,, \tag{1}$$

where $\mathcal{S}(y|\mu,\sigma^2,\nu)$ denotes the Student-t pdf with location parameter μ, scale parameter σ and degrees of freedom ν. The vector $\boldsymbol{\nu}$ contains the elements ν_m. The assumption of independently corrupting dimensions is realistic if the elements of $\mathbf{y}_n$ are observed independently, for instance, each element is a measurement from a different sensor.

The Student-t distribution can be constructed hierarchically by using a Gaussian distribution with extra latent variables u_{mn}

$$p(\boldsymbol{Y}|\mathbf{W},\mathbf{X},\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{U}) = \prod_{mn\in\mathcal{O}} \mathcal{N}\left(y_{mn}|\mathbf{w}_m^{\mathrm{T}}\mathbf{x}_n + \mu_m, \tau_m^{-1}u_{mn}^{-1}\right), \tag{2}$$

and giving u_{mn} gamma prior

$$p(\boldsymbol{U}) = \prod_{mn\in\mathcal{O}} \mathcal{G}(u_{mn}|\tfrac{\nu_m}{2},\tfrac{\nu_m}{2}). \tag{3}$$

This construction is equivalent to (1) when $\boldsymbol{U}$ is marginalized out [14]. Thus, the Student-t distribution can be interpreted as an infinite mixture of Gaussian distributions, and u_{mn} controls the noise level of each individual observation y_{mn}. The variable $\boldsymbol{U}$ denotes the set $\{u_{mn}|mn \in \mathcal{O}\}$ and the model is reduced to factor analysis with Gaussian noise if one fixes $u_{mn} = 1$ (i.e., $\nu_m = \infty$). Here, separate

ν_m are used for each dimension but the dimensions can have a common value $\nu_m = \nu$. The graphical model is shown in Fig. 2.

This robust model is closely related to the multivariate Student-t robust models [1, 13, 20]. Multivariate Student-t is obtained by setting $u_{mn} = u_n$ and $\nu_m = \nu$. This implies that if one element of $\mathbf{y}_n$ is badly corrupted, all the elements of the vector are considered unreliable as was shown in Fig. 1c. The existing methods [1, 13, 20] used u_n also in the prior of $\mathbf{x}_n$, which causes the outliers to be projected orthogonally to the latent subspace, thus, the outliers are being reproduced. However, our construction aims at removing the outliers and reconstructing them closer to the data mean with high uncertainty. The former approach is better if the outliers are interesting rare observations whereas the latter approach is better if the outliers are noise. Compared to the independent Student-t distribution, the multivariate Student-t distribution is realistic when the outliers are such that the whole vector $\mathbf{y}_n$ is corrupted. For instance, if $\mathbf{y}_n$ are observed hand-written digits, a digit 5 or a picture of pure noise would be multivariate outliers in a dataset of digits 7.

In [9], noise is modeled by the Laplace distribution, that is,

$$p(\boldsymbol{Y}|\mathbf{W}, \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\tau}) = \prod_{mn \in \mathcal{O}} \mathcal{L}\left(y_{mn} \middle| \mathbf{w}_m^{\mathrm{T}} \mathbf{x}_n + \mu_m, \tau_m^{-\frac{1}{2}} \right),$$

where $\mathcal{L}(x|\mu, \sigma)$ is the Laplace pdf with location μ and scale σ. This can be obtained by using the same likelihood as in (2) but changing the prior of $\boldsymbol{U}$ in (3) to

$$p(\boldsymbol{U}) = \prod_{mn} \mathcal{IG}\left(u_{mn} \middle| 1, \frac{1}{2} \right),$$

where $\mathcal{IG}(u|\alpha, \beta)$ is the inverse-gamma pdf with shape α and scale β. Note that Laplace distribution has no parameter similar to the degrees of freedom $\boldsymbol{\nu}$, which would control the amount of probability mass in the tail areas. Thus, badly corrupted observations may cause the noise scale parameter to increase so much that non-corrupted observations are regularized too much by a very large noise level as was shown in Fig. 1b. In addition, the 1-norm in the Laplace pdf implies that it is equivalent to corrupt a single element by x units or N elements by x/N units, thus, a large error in one observation can be equivalently considered as small errors in several observations. This might be undesirable if outliers are rare and they have large errors.

We emphasize that the different robust models are based on slightly different assumptions about the outliers, thus, they do not replace each other but they are all useful in different problems. Thus, the choice on the noise distribution should be made based on the application.

4 Posterior Approximation

Bayesian inference is done by evaluating the posterior distribution of the unknown variables given the observations. We use variational Bayesian approach (see, e.g., [4]) to cope with the problem of intractability of the joint posterior distribution. The key idea in variational Bayesian methods is to fit to the true posterior distribution $p(\boldsymbol{\Theta}|\boldsymbol{Y})$ a simpler approximate distribution $q(\boldsymbol{\Theta})$ using a cost function derived from the Kullback-Leibler divergence $\mathrm{KL}(q(\boldsymbol{\Theta}) \| p(\boldsymbol{\Theta}|\boldsymbol{Y}))$, which measures the difference between the two probability distributions.

In our case, an approximate posterior distribution $q(\mathbf{X}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{U}, \boldsymbol{\alpha}, \beta)$ is constructed by factorizing it with respect to the variables as

$$p(\mathbf{X}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{U}, \boldsymbol{\alpha}, \beta|\boldsymbol{Y}) \approx q(\mathbf{X}) q(\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\tau}) q(\boldsymbol{U}) q(\boldsymbol{\alpha}) q(\beta). \tag{4}$$

We update each factor in its turn keeping the other factors fixed. This is done by minimizing the relevant parts of the cost function, which results in simple update rules when conjugate priors are used. However, we estimate the degrees of freedom $\boldsymbol{\nu}$ using maximum likelihood methodology in order to keep the update rules analytically tractable.

In the following update rules, we denote by $\langle \cdot \rangle$ the expectations over the approximate posterior q. Some useful expectations for a Gaussian variable $\mathbf{x} \sim \mathcal{N}(\overline{\mathbf{x}}, \boldsymbol{\Sigma})$ and a Gamma variable $u \sim \mathcal{G}(a, b)$ are given below:

$$\langle \mathbf{x} \rangle = \overline{\mathbf{x}}, \qquad \left\langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \right\rangle = \overline{\mathbf{x}}\overline{\mathbf{x}}^{\mathrm{T}} + \boldsymbol{\Sigma}, \qquad \langle u \rangle = \frac{a}{b}.$$

4.1 Posterior approximation of $\mathbf{X}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}$

The factors in (4) are updated in turn using the following formulas. The optimal $q(\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\tau})$ is

$$\begin{aligned} q(\mathbf{W}, \boldsymbol{\mu}|\boldsymbol{\tau}) &= \prod_{m=1}^{M} \mathcal{N}\left(\begin{bmatrix} \mathbf{w}_m \\ \mu_m \end{bmatrix} \middle\| \begin{bmatrix} \overline{\mathbf{w}}_m \\ \overline{\mu}_m \end{bmatrix}, \tau_m^{-1} \boldsymbol{\Sigma}_m \right), \\ q(\boldsymbol{\tau}) &= \prod_{m=1}^{M} \mathcal{G}(\tau_m | \breve{a}_{\tau_m}, \breve{b}_{\tau_m}), \end{aligned}$$

and its parameters are updated as

$$\begin{aligned} \boldsymbol{\Sigma}_m^{-1} &= \operatorname{diag} \begin{bmatrix} \langle \boldsymbol{\alpha} \rangle \\ \langle \beta \rangle \end{bmatrix} + \sum_{n \in \mathcal{O}_{m:}} \langle u_{mn} \rangle \begin{bmatrix} \left\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right\rangle & \langle \mathbf{x}_n \rangle \\ \langle \mathbf{x}_n \rangle^{\mathrm{T}} & 1 \end{bmatrix}, \\ \begin{bmatrix} \overline{\mathbf{w}}_m \\ \overline{\mu}_m \end{bmatrix} &= \boldsymbol{\Sigma}_m \sum_{n \in \mathcal{O}_{m:}} \langle u_{mn} \rangle y_{mn} \begin{bmatrix} \langle \mathbf{x}_n \rangle \\ 1 \end{bmatrix}, \end{aligned}$$

$$\breve{a}_{\tau_m} = a_{\boldsymbol{\tau}} + \tfrac{1}{2} N_m, \tag{5}$$

$$\breve{b}_{\tau_m} = b_{\boldsymbol{\tau}} + \tfrac{1}{2} \sum_{n \in \mathcal{O}_{m:}} \langle u_{mn} \rangle y_{mn}^2 - \tfrac{1}{2} \begin{bmatrix} \overline{\mathbf{w}}_m \\ \overline{\mu}_m \end{bmatrix}^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1} \begin{bmatrix} \overline{\mathbf{w}}_m \\ \overline{\mu}_m \end{bmatrix}, \tag{6}$$

where N_m is the number of observed y_{mn} for given m and $\mathcal{O}_{m:}$ is the set of indices n for which y_{mn} is observed, that is, $\mathcal{O}_{m:} = \{n | mn \in \mathcal{O}\}$. This distribution has the following expectations:

$$
\begin{aligned}
\langle \tau_m \mathbf{w}_m \rangle &= \langle \tau_m \rangle \overline{\mathbf{w}}_m, \\
\langle \tau_m \mu_m \rangle &= \langle \tau_m \rangle \overline{\mu}_m, \\
\left\langle \tau_m \mathbf{w}_m \mathbf{w}_m^{\mathrm{T}} \right\rangle &= \langle \tau_m \rangle \overline{\mathbf{w}}_m \overline{\mathbf{w}}_m^{\mathrm{T}} + \boldsymbol{\Sigma}_m(\mathbf{w}_m, \mathbf{w}_m), \\
\left\langle \tau_m \mu_m^2 \right\rangle &= \langle \tau_m \rangle \overline{\mu}_m \overline{\mu}_m + \boldsymbol{\Sigma}_m(\mu_m, \mu_m), \\
\langle \tau_m \mathbf{w}_m \mu_m \rangle &= \langle \tau_m \rangle \overline{\mathbf{w}}_m \overline{\mu}_m + \boldsymbol{\Sigma}_m(\mathbf{w}_m, \mu_m),
\end{aligned}
$$

where $\boldsymbol{\Sigma}_m(\mathbf{x}, \mathbf{y})$ is that part of $\boldsymbol{\Sigma}_m$ which corresponds to the covariance of $\mathbf{x}$ and $\mathbf{y}$. Isotropic noise ($\tau_m = \tau$) can be obtained by summing over m appropriately in the update rules (5) and (6). The distribution $q(\mathbf{X})$ is updated as:

$$
\begin{aligned}
q(\mathbf{X}) &= \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n | \overline{\mathbf{x}}_n, \boldsymbol{\Sigma}_{\mathbf{x}_n}), \\
\boldsymbol{\Sigma}_{\mathbf{x}_n}^{-1} &= \mathbf{I} + \sum_{m \in \mathcal{O}_{:n}} \langle u_{mn} \rangle \left\langle \tau_m \mathbf{w}_m \mathbf{w}_m^{\mathrm{T}} \right\rangle, \\
\overline{\mathbf{x}}_n &= \boldsymbol{\Sigma}_{\mathbf{x}_n} \sum_{m \in \mathcal{O}_{:n}} \langle u_{mn} \rangle (y_{mn} \langle \tau_m \mathbf{w}_m \rangle - \langle \tau_m \mathbf{w}_m \mu_m \rangle),
\end{aligned}
$$

where $\mathcal{O}_{:n}$ is the set of indices m for which y_{mn} is observed, that is, $\mathcal{O}_{:n} = \{m | mn \in \mathcal{O}\}$. The update rule for factor $q(\boldsymbol{\alpha})$ is

$$
\begin{aligned}
q(\boldsymbol{\alpha}) &= \prod_{d=1}^{D} \mathcal{G}(\alpha_d | \breve{a}_{\boldsymbol{\alpha}}, \breve{b}_{\alpha_d}), \\
\breve{a}_{\boldsymbol{\alpha}} &= a_{\boldsymbol{\alpha}} + \tfrac{M}{2}, \\
\breve{b}_{\alpha_d} &= b_{\boldsymbol{\alpha}} + \tfrac{1}{2} \sum_{m=1}^{M} \left\langle \tau_m w_{md}^2 \right\rangle,
\end{aligned}
$$

and for factor $q(\beta)$

$$
\begin{aligned}
q(\beta) &= \mathcal{G}(\beta | \breve{a}_{\beta}, \breve{b}_{\beta}), \\
\breve{a}_{\beta} &= a_{\beta} + \tfrac{M}{2}, \\
\breve{b}_{\beta} &= b_{\beta} + \tfrac{1}{2} \sum_{m=1}^{M} \left\langle \tau_m \mu_m^2 \right\rangle.
\end{aligned}
$$

In order to speed up the learning, it is possible to move the bias from $\mathbf{X}$ to $\boldsymbol{\mu}$ and rotate the latent subspace such that

$$
\frac{1}{N} \sum_{n=1}^{N} \left\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right\rangle = \mathbf{I}, \qquad \sum_{m=1}^{M} \left\langle \tau_m \mathbf{w}_m \mathbf{w}_m^{\mathrm{T}} \right\rangle = \text{diagonal},
$$

as discussed in [16]. The update rules presented above are the same for the different robust noise distributions.

4.2 Independent Student-t noise model

For the independent Student-t noise model, the distribution $q(\boldsymbol{U})$ is updated as follows:

$$
\begin{aligned}
q(\boldsymbol{U}) &= \prod_{mn\in\mathcal{O}} \mathcal{G}(u_{mn}|\breve{a}_{u_{mn}}, \breve{b}_{u_{mn}}), \\
\breve{a}_{u_{mn}} &= \tfrac{\nu_m}{2} + \tfrac{1}{2}, \\
\breve{b}_{u_{mn}} &= \tfrac{\nu_m}{2} + \tfrac{1}{2}\psi_{mn}
\end{aligned}
$$

where

$$
\begin{aligned}
\psi_{mn} &= \left\langle \tau_m (y_{mn} - \mathbf{w}_m^{\mathrm{T}} \mathbf{x}_n - \mu_m)^2 \right\rangle \qquad (7) \\
&= \langle \tau_m \rangle y_{mn}^2 + \operatorname{tr}\left(\left\langle \tau_m \mathbf{w}_m \mathbf{w}_m^{\mathrm{T}} \right\rangle \left\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right\rangle \right) + \left\langle \tau_m \mu_m^2 \right\rangle - \\
&\qquad 2 y_{mn} \langle \tau_m \mathbf{w}_m \rangle^{\mathrm{T}} \langle \mathbf{x}_n \rangle - 2 y_{mn} \langle \tau_m \mu_m \rangle + 2 \langle \tau_m \mathbf{w}_m \mu_m \rangle^{\mathrm{T}} \langle \mathbf{x}_n \rangle .
\end{aligned}
$$

The hyperparameters $\{\nu_m\}_{m=1}^{M}$ can be updated before $q(\boldsymbol{U})$ by maximizing the following term with respect to ν_m:

$$
\sum_{n\in\mathcal{O}_{m:}} \log \mathcal{S}_1 \left(\sqrt{\psi_{mn}} \middle| \nu_m \right), \qquad (8)
$$

where

$$
\log \mathcal{S}_M(\mathbf{x}|\nu) = \log \mathit{\Gamma}((\nu + M)/2) - \log \mathit{\Gamma}(\nu/2) - \tfrac{M}{2}\log(\pi\nu) - \tfrac{\nu+M}{2} \log\left(1 + \tfrac{\mathbf{x}^{\mathrm{T}}\mathbf{x}}{\nu}\right)
$$

is the log pdf of the M-dimensional Student-t distribution with location parameter zero, unit scale and degrees of freedom ν. In order to obtain common degrees of freedom $\nu_m = \nu$, one should also sum over m in (8).

4.3 Multivariate Student-t noise model

In order to study the effect of the noise model more carefully, in this paper we also present a model with Gaussian variables $\mathbf{x}_n$ and multivariate Student-t distribution for the noise $\boldsymbol{\epsilon}_n$. This model is obtained by using $u_{mn} = u_n$ and $\nu_m = \nu$. Then, $q(\boldsymbol{U})$ is updated as

$$
\begin{aligned}
q(\boldsymbol{U}) &= \prod_{n=1}^{N} \mathcal{G}(u_n|\breve{a}_{u_n}, \breve{b}_{u_n}), \\
\breve{a}_{u_n} &= \tfrac{\nu}{2} + \tfrac{1}{2} M_n, \\
\breve{b}_{u_n} &= \tfrac{\nu}{2} + \tfrac{1}{2} \sum_{m\in\mathcal{O}_{:n}} \psi_{mn},
\end{aligned}
$$

where ψ_{mn} is given in (7). The hyperparameter ν is found by maximizing

$$
\sum_{n=1}^{N} \log \mathcal{S}_{M_n} \left(\sqrt{\sum_{m\in\mathcal{O}_{:n}} \psi_{mn}} \middle| \nu \right),
$$

where M_n is the number of observed y_{mn} for given n.

Table 1: RMSE of reconstruction shown separately for those function values which were corrupted by outliers. The values were averaged over 10 artificial datasets.

Noise model	Non-outliers	Outliers
Gaussian (GPCA)	0.996	10.560
Laplace (LPCA)	0.894	1.205
Multivariate Student-t (SPCA-m)	0.829	1.970
Independent Student-t (SPCA-i)	0.687	0.815

4.4 Laplace noise model

For completeness, we show the update formula for the Laplace noise model [9]. The approximate posterior of the latent variable $\boldsymbol{U}$ is

$$q(\boldsymbol{U}) = \prod_{mn \in \mathcal{O}} \mathcal{IN} \left(u_{mn} \middle| \frac{1}{\sqrt{\psi_{mn}}}, 1 \right),$$

where $\mathcal{IN}(u|\mu, \lambda)$ is the inverse Gaussian distribution with mean μ and shape λ, and ψ_{mn} is given in (7). The mean with respect to this distribution is

$$\langle u_{mn} \rangle = \frac{1}{\sqrt{\psi_{mn}}},$$

which is needed for updating the other variables. For this noise model, $\langle \log u_{mn} \rangle$ is intractable, thus, it is not possible to analytically compute the log-likelihood lower bound, which could be used for model comparison or monitoring the convergence of the algorithm.

5 Experiments

5.1 Artificial dataset

We generated ten artificial datasets to compare the different noise models. Each dataset consisted of $N = 100$ noisy vector samples with $M = 10$ dimensions from a Gaussian distribution and some added outliers. The true values were generated from a Gaussian distribution using a 4-dimensional subspace with standard deviations $4, 3, 2, 1$. These values were first corrupted by isotropic noise with standard deviation 1 and then outliers were generated by replacing each y_{mn} with probability 0.02 by a draw from a uniform distribution over $[-30, 30]$.

The datasets were used to learn the PCA model with different noise distributions: Gaussian (GPCA), Laplace (LPCA), multivariate Student-t (SPCA-m) and dimensionally independent Student-t (SPCA-i). The models used 9-dimensional latent space and isotropic noise $\tau_m = \tau$. SPCA-i used a pooled degrees-of-freedom parameter $\nu_m = \nu$. The estimated posterior means of the variables were used for reconstructing the noiseless true values.

Table 1 shows the root-mean-square error (RMSE) of the reconstructions computed separately for values that were corrupted by Gaussian noise and outliers, averaged over the ten datasets. Obviously, each method reconstructs the non-outliers

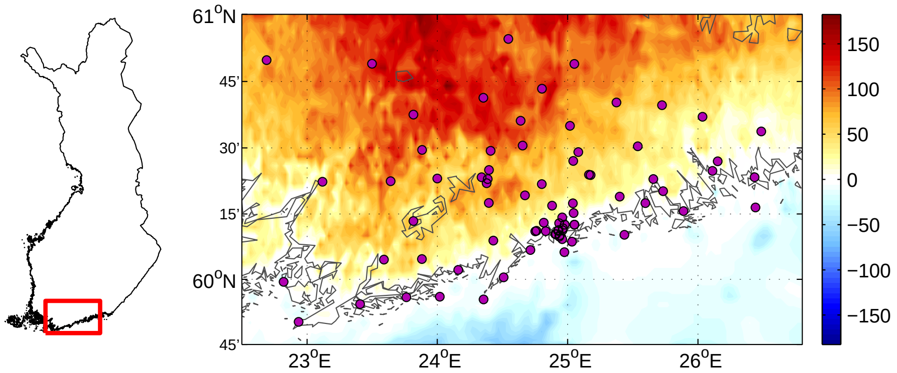


Fig. 3: The weather stations are shown as purple dots on the topographical map of Southern Finland. The color represents the altitude above sea level in meters.

better than the badly corrupted outliers. The novel model SPCA-i reconstructs both sets better than the others. Interestingly, SPCA-i reconstructs even the outliers more accurately than the other models reconstruct the non-outliers. This happens because the other models are not able to identify the outliers well, thus the estimation of the principal subspace and the reconstructions are corrupted. LPCA reconstructs outliers better than SPCA-m because it correctly assumes independently corrupted dimensions. However, SPCA-m is better than LPCA in reconstructing the non-outliers because the distribution resembles more the Gaussian distribution which was used for generating the noise. GPCA, on the other hand, has severe problems with the outliers as expected. The total CPU time for each method was approximately 10 seconds.

5.2 Meteorological recordings from weather stations

The proposed model was largely motivated by an analysis of a real-world weather data set from the Helsinki Testbed research project of mesoscale meteorology (see http://testbed.fmi.fi/). A straightforward analysis was impossible because of the large amount of missing and corrupted data, thus we needed a principled way to preprocess the data in order to reconstruct the missing values and to remove the outliers for further analysis. The data consists of temperature measurements in Southern Finland over a period of almost two years with an interval of ten minutes, resulting in $N = 89\,202$ samples for each weather station. Some parts of the data were discarded: stations with no observations were removed, and we used only the measurements taken in the lowest altitude in each location. The locations of the remaining $M = 79$ weather stations are shown in Fig. 3.

The quality of the dataset was partly poor. Approximately 35% of the data were missing and a large number of measurements were corrupted by outliers. Fig. 4 shows representative examples of measurements from five stations. The quality of the dataset can be summarized with the six example signals in Figure 4 as follows: Half of the stations were relatively good, having no outstanding outliers

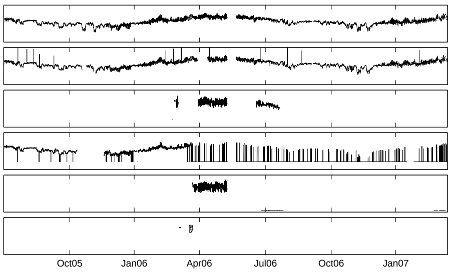


Fig. 4: Temperature data from five stations from the Helsinki Testbed dataset. The scale is from −50°C to 50°C.

Table 2: The predictive log-densities for the weather dataset using different noise distributions.

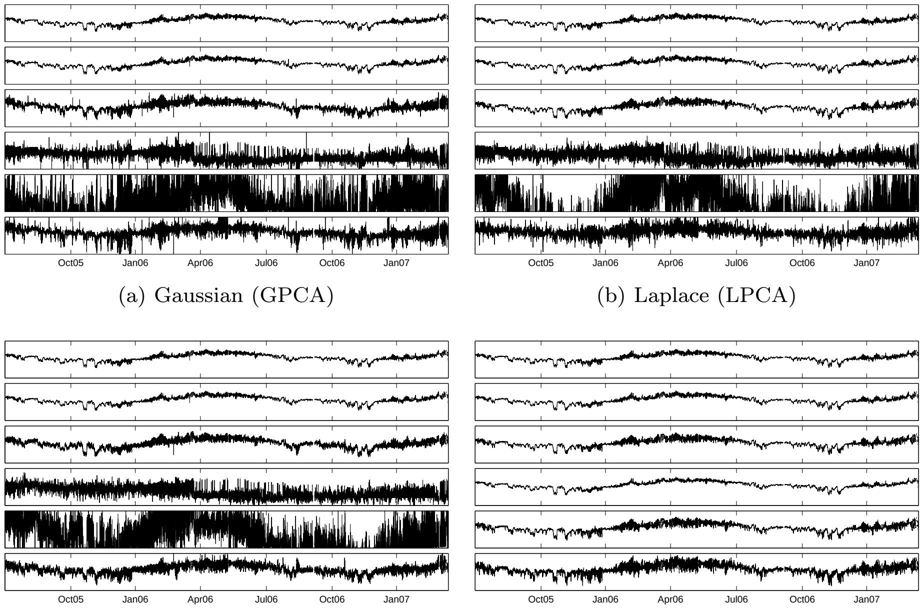
Noise distribution	PLD
Gaussian (GPCA)	$-10.26 \cdot 10^6$
Laplace (LPCA)	$-1.63 \cdot 10^6$
Multivariate Student-t (SPCA-m)	$-2.82 \cdot 10^6$
Independent Student-t (SPCA-i)	$-1.12 \cdot 10^6$

and only short periods missing (similarly to the first signal). More than 10 stations had a few outliers (similarly to the second signal). More than 20 stations had large amount of data missing (similarly to the third signal). Five stations had a large number of outliers compared to the number of uncorrupted observations (similarly to the last three signals).

Although the outliers may sometimes be easily distinguished from the data, removing them by hand requires a tedious procedure which turned out to be non-trivial in some cases. Therefore, we used the proposed robust PCA method to automatically solve the problems of outlier removal, dimensionality reduction, and filling in the missing values. Because the weather stations may corrupt measurements independently of each other, the robust noise distribution using independent Student-t distributions seems reasonable.

In the presented experiment, we estimated a 30-dimensional principal subspace of the data using models with Gaussian components $\mathbf{x}_n$ and different noise models. We used the same noise models as in the artificial experiment: GPCA, LPCA, SPCA-m and SPCA-i. Note that the hierarchical prior on the loadings $\mathbf{W}$ should eliminate irrelevant components. We used a pooled precision parameter $\tau_m = \tau$. For SPCA-i, the degrees of freedom $\{\nu_m\}_{m=1}^{M}$ were modeled separately for each station. The bias term $\boldsymbol{\mu}$ was ignored because the temperature mean is close to zero in the research area. Although the data are spatio-temporal, none of the models utilized the timestamps or spatial coordinates of the observations, that is, the rows and columns of the data matrix Y are exchangeable.

The four models were trained using 80% of the data by discarding elements randomly from the $M \times N$ data matrix. The remaining 20% was used as a test set



(a) Gaussian (GPCA)

(b) Laplace (LPCA)

(c) Multivariate Student-t (SPCA-m)

(d) Independent Student-t (SPCA-i)

Fig. 5: Reconstructions of the corrupted signals using different noise models.

for computing predictive log-densities. The predictive log-densities (PLDs)

$$\log p(\boldsymbol{Y}_{\text{test}}|\boldsymbol{Y}_{\text{train}}) = \log \int p(\boldsymbol{Y}_{\text{test}}|\mathbf{W},\mathbf{X},\boldsymbol{\tau},\boldsymbol{U})q(\mathbf{X})q(\mathbf{W},\boldsymbol{\tau})q(\boldsymbol{U})d\mathbf{X}d\mathbf{W}d\boldsymbol{\tau}d\boldsymbol{U}$$

were computed by integrating over $q(\boldsymbol{U})$ analytically and using samples from $q(\mathbf{X})q(\mathbf{W},\boldsymbol{\tau})$ to compute the required integral. Table 2 shows that using at least some robust noise distribution improves the predictive performance significantly. In this case, the dimensionally independent robust noise distributions LPCA and SPCA-i are clearly better than multivariate SPCA-m because the outliers arise independently at each station. SPCA-i models the noise better than LPCA probably because it can adjust the degrees of freedom for each station. However, in addition to this quantitative measure, it is important to compare the results qualitatively.

Fig. 5 presents the reconstructions for the five signals from Fig. 4 using the compared techniques. The reconstructions obtained by GPCA (Fig. 5a), LPCA (Fig. 5b) and SPCA-m (Fig. 5c) are clearly bad. These models are overfitted to outliers and to spontaneous correlations observed in scarce measurements from problematic stations. The methods reproduce accurately some outliers and generate new outliers in the place of missing values. In contrast, the results by SPCA-i are clearly much better: the outliers are removed and reasonable reconstructions of the missing values are obtained. Although the signals look rather similar in Fig. 5d because the analyzed spatial area is small and the annual cycle is obviously the dominant pattern, the reconstructed signals look very plausible. The CPU time for the methods on a regular desktop computer varied between 90–150 seconds per iteration step and they required approximately 100 steps to converge, thus each algorithm ran for 2–4 CPU hours, where GPCA was the fastest, SPCA-i the slowest, and LPCA and SPCA-m in between.

6 Conclusions and remarks

This paper presented a probabilistic model for robust FA and PCA which can be a useful tool for preprocessing, analyzing or modeling incomplete data with outliers. The effect of outliers is diminished by using the Student-t distribution for modeling the observation noise. Existing approaches have used the multivariate Student-t [1, 13, 20], but we showed that modeling the elements of the noise vector independently can be more appropriate for some datasets. In addition, the Student-t distribution provides more flexibility than the existing method using the Laplace distribution [9]. We gave a unifying comparison of the different robust noise distributions by showing the minor differences in the model details and providing the equations for the variational Bayesian learning algorithms.

The proposed method was tested on an artificial dataset and a real-world weather dataset by comparing it with PCA model using Gaussian noise and robust PCA models using the multivariate Student-t and the Laplace distribution. Our experiments demonstrated the superior performance of the presented model, which provided good predictive measures and reasonable reconstructions of missing data and outliers.

It is straightforward to extend the presented robust FA model in different directions. For instance, the temporal structure could be modeled with linear Gaussian state-space models [2] or the spatio-temporal structure by Gaussian-process FA [15], which might improve further the experimental results with the tested real-world weather dataset. In addition, the modularity of probabilistic modeling enables one to use the proposed noise distribution with other models than the factor analysis model.

Acknowledgement

This work was supported by the Academy of Finland project 'Novel machine learning techniques for studying climate variability and its impacts'. The authors would like to thank the Finnish Meteorological Institute for providing the data.

References

1. Archambeau C, Delannay N, Verleysen M (2006) Robust probabilistic projections. In: Proc. of the 23rd Int. Conf. on Machine Learning (ICML2006), pp 33–40
2. Beal MJ (2003) Variational algorithms for approximate Bayesian inference. PhD thesis, Gatsby Computational Neuroscience Unit, University College London

3. Bishop C (1999) Variational principal components. In: Proc. of the 9th Int. Conf. on Artificial Neural Networks (ICANN'99), vol 1, pp 509–514
4. Bishop C (2006) Pattern Recognition and Machine Learning. Springer-Verlag, New York, USA
5. Candès EJ, Li X, Ma Y, Wright J (2011) Robust principal component analysis? Journal of the ACM 58:37
6. Chandrasekaran V, Sanghavi S, Parrilo PA, Willsky AS (2009) Sparse and low-rank matrix decomposition. In: IFAC Symposium on System Identification
7. Cichocki A, Amari SI (2002) Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. Wiley, New York, USA
8. Ding X, He L, Carin L (2011) Bayesian robust principal component analysis. IEEE Transactions on Image Processing PP(99):1
9. Gao J (2008) Robust L1 principal component analysis and its Bayesian variational inference. Neural Computation 20(2):555–572
10. Hyvärinen A, Karhunen J, Oja E (2001) Independent Component Analysis. J. Wiley, New York, USA
11. Ilin A, Raiko T (2010) Practical approaches to principal component analysis in the presence of missing values. Journal of Machine Learning Research 11:1957–2000
12. Jolliffe I (2002) Principal Component Analysis, 2nd edn. Springer-Verlag, New York
13. Khan Z, Dellaert F (2004) Robust generative subspace modeling: The subspace t distribution. Tech. rep., GVU Center, College of Computing, Georgia
14. Liu C, Rubin D (1995) ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statistica Sinica pp 19–9
15. Luttinen J, Ilin A (2009) Variational Gaussian-process factor analysis for modeling spatio-temporal data. In: Advances in Neural Information Processing Systems 22, MIT Press, Cambridge, MA, USA, pp 1177–1185
16. Luttinen J, Ilin A (2010) Transformations in variational Bayesian factor analysis to speed up learning. Neurocomputing 73(7–9):1093–1102
17. Roweis S (1998) EM algorithms for PCA and SPCA. In: Jordan M, Kearns M, Solla S (eds) Advances in Neural Information Processing Systems, MIT Press, vol 10, pp 626–632
18. Tipping M, Bishop C (1999) Probabilistic principal component analysis. J of the Royal Statistical Society Series B 61(3):611–622
19. Wright J, Peng Y, Ma Y, Ganesh A, Rao S (2009) Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization. In: Advances in Neural Information Processing Systems 22, MIT Press, Cambridge, MA
20. Zhao J, Jiang Q (2006) Probabilistic PCA for t distributions. Neurocomputing 69:2217–2226
21. Zhao Jh, Yu PLH (2009) A note on variational Bayesian factor analysis. Neural Networks 22:988–997