Bayesian Robust PCA for Incomplete Data

Jaakko Luttinen, Alexander Ilin, and Juha Karhunen

Helsinki University of Technology TKK
Department of Information and Computer Science
P.O. Box 5400, FI-02015 TKK, Espoo, Finland

Abstract. We present a probabilistic model for robust principal component analysis (PCA) in which the observation noise is modelled by Student-t distributions that are independent for different data dimensions. A heavy-tailed noise distribution is used to reduce the negative effect of outliers. Intractability of posterior evaluation is solved using variational Bayesian approximation methods. We show experimentally that the proposed model can be a useful tool for PCA preprocessing for incomplete noisy data. We also demonstrate that the assumed noise model can yield more accurate reconstructions of missing values: Corrupted dimensions of a "bad" sample may be reconstructed well from other dimensions of the same data vector. The model was motivated by a real-world weather dataset which was used for comparison of the proposed technique to relevant probabilistic PCA models.

1 Introduction

Principal component analysis (PCA) is a widely used method for data preprocessing (see, e.g., [1–3]). In independent component analysis (ICA) and source separation problems, PCA is used for reducing the dimensionality of the data to avoid overlearning, to suppress additive noise, and for prewhitening needed in several ICA algorithms [2, 4]. PCA is based on the quadratic criteria of variance maximisation and minimisation of the mean-square representation error, and therefore it can be sensitive to outliers in the data. Robust PCA techniques have been introduced to cope this problem, see, for example, [4] and the references therein. The basic idea in robust PCA methods is to replace quadratic criteria leading to standard PCA by more slowly growing criteria.

PCA has a probabilistic interpretation as maximum likelihood estimation of a latent variable model called probabilistic PCA (PPCA) [5]. While PPCA is a rather simplistic model based on Gaussion assumptions, it can be used as a basis for building probabilistic extensions of classical PCA. Probabilistic models provide a principled way to cope with the overifitting problem, to do model comparison and to handle missing values. Probabilistic models for robust PCA have been introduced recently [6–8]. They treat possible outliers by using heavy-tailed distributions, such as Student-t or Laplacian, for describing the noise.

In this paper, we present a new robust PCA model based on the Student-t distribution and show how it can be identified for incomplete data, that is,

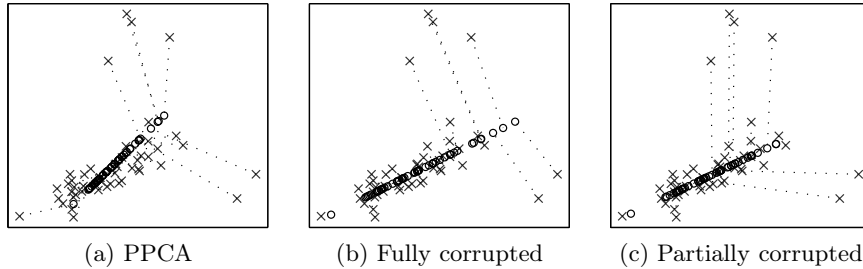(a) PPCA   (b) Fully corrupted   (c) Partially corrupted

Fig. 1: Principal subspace estimation using (a) probabilistic PCA [5], (b) robust PCA assuming fully corrupted outliers [7] and (c) robust PCA assuming partially corrupted outliers. The crosses represent data points and the circles show their projections onto the found principal subspace.

datasets with missing values. We assume that the outliers can arise independently in each sensor (i.e. for each dimension of a data vector). This is different to the previously introduced techniques [6, 7] which assume that all elements of an outlier data vector are corrupted. This work was inspired by our intention to apply a semi-blind source separation technique, called denoising source separation (DSS) to a weather dataset which is too much corrupted by outliers and missing values. We have earlier successfully applied DSS to exploratory analysis of global climate data [9].

Our modelling assumption can be more realistic for some datasets and therefore they can improve the quality of the principal subspace estimation and achieve better reconstructions of the missing values. The model can also be used to remove outliers by estimating the true values of their corrupted components from the uncorrupted ones. This is illustrated in Fig. 1 using an artificial two-dimensional data with a prominent principal direction and a few outliers. The subspace found by the simplest PCA model is affected by outliers, whereas robust techniques are able to find the right principal subspace. However, the reconstruction of the data is quite different depending on whether one assumes fully corrupted or partially corrupted outliers: Fully corrupted outliers can be reconstructed by projecting orthogonally onto the subspace, while improbable values of partially corrupted samples can be ignored and reconstructed based on the uncorrupted dimensions.

2   Model

Let us denote by $\{\boldsymbol{y}_n\}_{n=1}^N$ a set of M-dimensional observations $\boldsymbol{y}_n$. The data are assumed to be generated from hidden D-dimensional states $\{\boldsymbol{x}_n\}_{n=1}^N$ using the transformation:

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

where $\boldsymbol{W}$ is a $M \times D$ loading matrix, $\boldsymbol{\mu}$ is a bias term and $\boldsymbol{\epsilon}_n$ is noise. Usually the dimensions fulfil $D < M < N$. The prior models for the latent variables are the same as in PPCA and we use conjugate prior for $\boldsymbol{\mu}$ and hierarchical prior for $\boldsymbol{W}$ as in [10] to diminish overfitting [11]:

$$p(\boldsymbol{X}) = \prod_{m=1}^{M} \prod_{n=1}^{N} \mathcal{N}(x_{mn}|0,1),$$

$$p(\boldsymbol{W}|\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{d=1}^{D} \mathcal{N}(w_{md}|0,\alpha_d^{-1}),$$

$$p(\boldsymbol{\mu}) = \prod_{m=1}^{M} \mathcal{N}(\mu_m|0,\beta^{-1}),$$

$$p(\boldsymbol{\alpha}) = \prod_{d=1}^{D} \mathcal{G}(\alpha_d|a_{\boldsymbol{\alpha}},b_{\boldsymbol{\alpha}}).$$

Hyperparameters $a_{\boldsymbol{\alpha}}$, $b_{\boldsymbol{\alpha}}$, and β are fixed to some proper values.

The noise term $\boldsymbol{\epsilon}_n$ is modelled using independent Student-t distributions for its elements. This is achieved by using a hierarchical model with extra variables u_{mn}:

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

which yields a product of Student-t distributions $\mathcal{S}(y_{mn}|\boldsymbol{w}_m^{\mathrm{T}}\boldsymbol{x}_n + \mu_m, \tfrac{1}{\tau_m}, \nu_m)$ with degrees of freedom ν_m when $\boldsymbol{U}$ is marginalised out [12]. Here, $\mathcal{O}_{mn}$ denotes such indices that the corresponding y_{mn} is actually observed and $\boldsymbol{w}_m^{\mathrm{T}}$ is the m-th row of $\boldsymbol{W}$. Precision τ_m defines a scaling variable which is assigned a conjugate prior

$$p(\boldsymbol{\tau}) = \prod_{m=1}^{M} \mathcal{G}(\tau_m|a_{\boldsymbol{\tau}},b_{\boldsymbol{\tau}}),$$

with $a_{\boldsymbol{\tau}}$ and $b_{\boldsymbol{\tau}}$ set to proper values. Separate τ_m and ν_m are used for each dimension but with simple modifications the dimensions can have a common value. Especially for the precision $\boldsymbol{\tau}$, common modelling may prevent bad local minima. For the degrees of freedom $\boldsymbol{\nu}$ we set a uniform prior.

3   Posterior Approximation

Bayesian inference is done by evaluating the posterior distribution of the unknown variables given the observations. We use variational Bayesian approach to cope with the problem of intractability of the joint posterior distribution (see, e.g., [3, ch.10] for more details). The approximate distribution q is factorised with
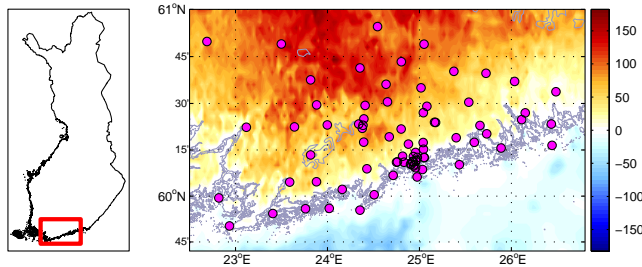
Fig. 2: The weather stations are shown as purple dots on the topographical map of the studied area. The colour represents the altitude above sea level in meters.

respect to the unknown variables as

$$\prod_{n=1}^{N} q(\boldsymbol{x}_n) \prod_{m=1}^{M} q(\boldsymbol{w}_m) \prod_{m=1}^{M} q(\mu_m) \prod_{m=1}^{M} q(\tau_m) \prod_{m=1}^{M} \prod_{n=1}^{N} q(u_{mn}) \prod_{d=1}^{D} q(\alpha_d)$$

and each factor $q(\boldsymbol{\theta}_i)$ is updated assuming the other factors are fixed. This is done by minimising the Kullback-Leibler divergence cost function. Using conjugate priors yields simple update rules presented in the appendix.

4 Experiments with real-world data

The proposed model was largely motivated by the analysis of real-world weather data from the Helsinki Testbed research project of mesoscale meteorology. The data consists of temperature measurements in Southern Finland over a period of almost two years with an interval of ten minutes, resulting in 89 000 time instances. 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 79 stations are shown in Fig. 2.

The quality of the dataset was partly poor. Approximately 35% of the data was missing and a large number of measurements were corrupted. Fig. 3 shows representative examples of measurements from four stations. The quality of the dataset can be summarised as follows: Half of the stations were relatively good, having no outstanding outliers and only short periods missing. More than 10 stations had a few outliers, similarly to the first signal from Fig. 3. Five stations had a large number of outliers, see the second signal in Fig. 3. The quality of the data from the rest of the stations was somewhat poor: The signals contained a small number of measurements and were corrupted by outliers, see the two signals at the bottom of Fig. 3.

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 as a
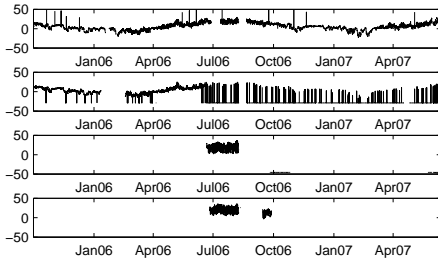
Fig. 3: Temperature data from four stations from the Helsinki Testbed dataset.

preprocessing step which automatically solves the problems of outlier removal, dimensionality reduction and infilling missing values. To keep the preprocessing step simple, we did not take into account the temporal structure of the data.

In the presented experiment, we estimated the four-dimensional principal subspace of the data using the following models: probabilistic PCA [5], robust PPCA (RPCA-s) [7] and the robust model presented in this paper (RPCA-d). For RPCA-d, the degrees of freedom $\{\nu_m\}_{m=1}^M$ were modelled separately for each station whereas the precision $\tau_m = \tau$ was set to be common. Broad priors were obtained by setting $a_{\boldsymbol{\alpha}} = b_{\boldsymbol{\alpha}} = \beta = a_{\boldsymbol{\tau}} = b_{\boldsymbol{\tau}} = 10^{-3}$.

Fig. 4 presents the reconstruction of the missing data for the four signals from Fig. 3 using the compared techniques. The reconstructions obtained by PPCA and RPCA-s are clearly bad. Both models are over-fitted 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 RPCA-d 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. 4c (the analysed spatial area is small and the annual cycle is obviously the dominant pattern), the reconstructed signals look very plausible.

The loading matrix $\boldsymbol{W}$ obtained with the different techniques is also visualised in Fig. 4. Each column of $\boldsymbol{W}$ is a collection of weights showing the contribution of one principal component in reconstructing data in different spatial locations. The patterns shown in Fig. 4 are interpolations of the weights over the map of Southern Finland. The patterns produced by PPCA and RPCA-s clearly contain lots of artefacts: the components are over-fitted to the outliers registered in some weather stations. On the contrary, the components found by RPCA-d are much more meaningful (though they contain some artefacts due to problematic stations in the central area): The first component explains the dominant yearly and daily oscillations and the patterns associated with the rest of the principal components are very typical for PCA applied to spatially distributed data. Since the investigated area is rather small, the first principal component has similar loading for all weather stations. Note a clear coast line pattern in the second and the third components.
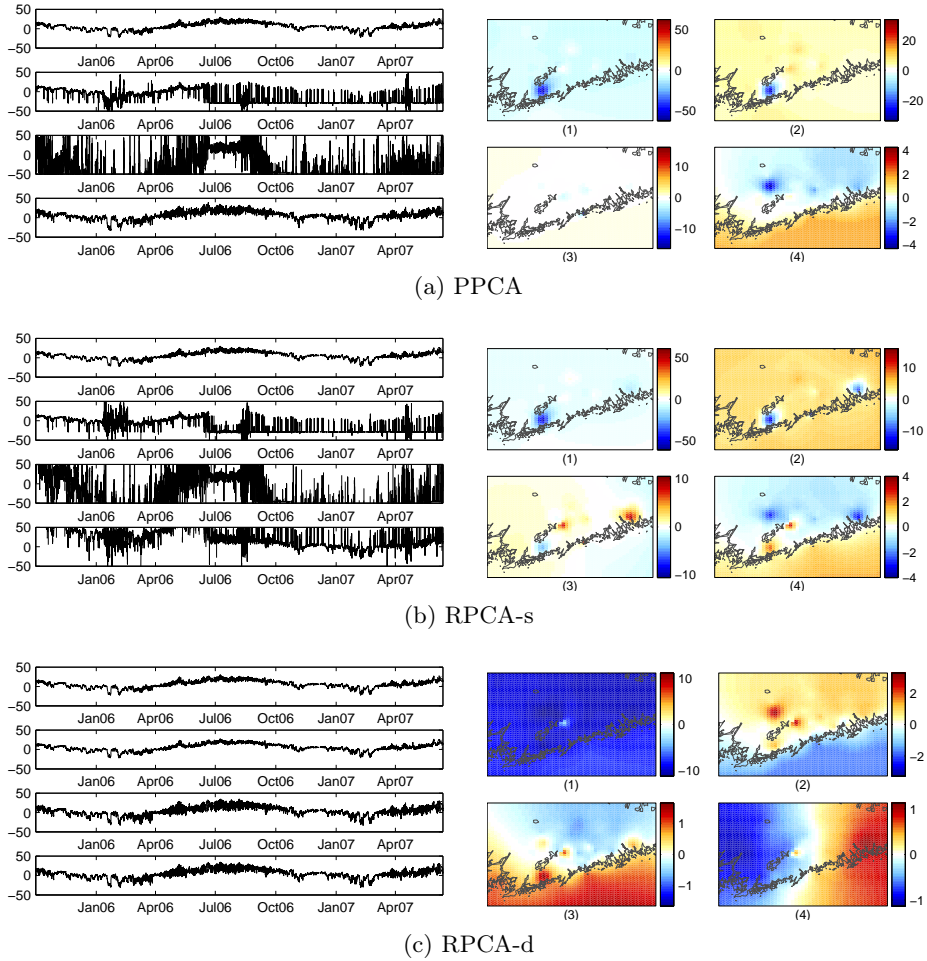
(a) PPCA

(b) RPCA-s

(c) RPCA-d

Fig. 4: Experimental results obtained for the Helsinki Testbed dataset with different models. Left: The reconstructions of the signals shown in Fig. 3. Right: The principal component loadings interpolated over the map of Southern Finland.

5   Conclusions

In this paper, we presented a probabilistic model for robust PCA which can be a useful tool for preprocessing incomplete data with outliers. The effect of outliers is diminished by using the Student-t distribution for modelling the observation noise. We showed that using a model with independent elements of the noise vector can be more appropriate for some real-world datasets. We tested the proposed method on a real-world weather dataset and compared our approach with the probabilistic PCA model [5] and robust PPCA assuming fully

corrupted outlier vectors [7]. The experiment showed the superior performance of the presented model, which found meaningful spatial patterns for the principal components and provided reasonable reconstruction in the place of missing data.

The proposed algorithm is based on a probabilistic model and therefore it provides information about the uncertainty of the estimated parameters. The uncertainty information can be taken into account, for example, when the principal components are ordered according to the amount of explained data variance [11]. The model can easily be extended, for example, by taking into account the temporal structure of the data. This would result in better performance in the tasks of missing value reconstruction and outlier removal.

In our work, we use the proposed technique as a preprocessing step for further exploratory analysis of data. For example, one can investigate a principal subspace found for weather data in order to find meaningful weather patterns or to extract features which might be useful for statistical weather forecasts. This can be done, for example, by using rotation techniques closely related to ICA. We have earlier used this approach for analysis of global climate data [9].

References

1. Jolliffe, I.T.: Principal Component Analysis. 2nd edn. Springer, New York (2002)
2. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. J. Wiley (2001)
3. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
4. Cichocki, A., Amari, S.: Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. John Wiley & Sons, Inc., New York, NY, USA (2002)
5. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society Series B **61**(3) (1999) 611–622
6. Zhao, J., Jiang, Q.: Probabilistic PCA for t distributions. Neurocomputing **69** (October 2006) 2217–2226
7. Archambeau, C., Delannay, N., Verleysen, M.: Robust probabilistic projections. In: Proceedings of the 23rd international conference on machine learning (ICML'2006), New York, NY, USA, ACM (2006) 33–40
8. Gao, J.: Robust L1 principal component analysis and its Bayesian variational inference. Neural Computation **20**(2) (2008) 555–572
9. Ilin, A., Valpola, H., Oja, E.: Exploratory analysis of climate data using source separation methods. Neural Networks **19**(2) (2006) 155–167
10. Bishop, C.M.: Variational principal components. In: Proceedings of the ninth international conference on artificial neural networks (ICANN'99). Volume 1. (1999) 509–514
11. Ilin, A., Raiko, T.: Practical approaches to principal component analysis in the presence of missing values. Technical Report TKK-ICS-R6, Helsinki University of Technology, Espoo, Finland (2008) Available at http://www.cis.hut.fi/alexilin/.
12. Liu, C., Rubin, D.B.: ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statistica Sinica **5** (1995) 19–9

Appendix: Update rules

$q(\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{x}_n | \overline{\boldsymbol{x}}_n, \boldsymbol{\Sigma}_{\boldsymbol{x}_n})$, $q(\boldsymbol{w}_m) = \mathcal{N}(\boldsymbol{w}_m | \overline{\boldsymbol{w}}_m, \boldsymbol{\Sigma}_{\boldsymbol{w}_m})$ and $q(\boldsymbol{\mu}) = \mathcal{N}(\mu_m | \overline{\mu}_m, \widetilde{\mu}_m)$ are Gaussian density functions updated as follows:

$$\boldsymbol{\Sigma}_{\boldsymbol{x}_n}^{-1} = \boldsymbol{I} + \sum_{m|\mathcal{O}_{mn}} \langle \tau_m \rangle \langle u_{mn} \rangle (\overline{\boldsymbol{w}}_m \overline{\boldsymbol{w}}_m^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{w}_m})$$

$$\overline{\boldsymbol{x}}_n = \boldsymbol{\Sigma}_{\boldsymbol{x}_n} \sum_{m|\mathcal{O}_{mn}} \langle \tau_m \rangle \langle u_{mn} \rangle \overline{\boldsymbol{w}}_m (y_{mn} - \overline{\mu}_m)$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w}_m}^{-1} = \mathrm{diag} \langle \boldsymbol{\alpha} \rangle + \langle \tau_m \rangle \sum_{n|\mathcal{O}_{mn}} \langle u_{mn} \rangle (\overline{\boldsymbol{x}}_n \overline{\boldsymbol{x}}_n^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{x}_n})$$

$$\overline{\boldsymbol{w}}_m = \boldsymbol{\Sigma}_{\boldsymbol{w}_m} \langle \tau_m \rangle \sum_{n|\mathcal{O}_{mn}} \langle u_{mn} \rangle \overline{\boldsymbol{x}}_n (y_{mn} - \overline{\mu}_m)$$

$$\widetilde{\mu}_m^{-1} = \beta + \langle \tau_m \rangle \sum_{n|\mathcal{O}_{mn}} \langle u_{mn} \rangle$$

$$\overline{\mu}_m = \widetilde{\mu}_m \langle \tau_m \rangle \sum_{n|\mathcal{O}_{mn}} \langle u_{mn} \rangle \left(y_{mn} - \overline{\boldsymbol{w}}_m^{\mathrm{T}} \overline{\boldsymbol{x}}_n \right)$$

where $\langle \cdot \rangle$ denotes expectations over the approximate distribution.

Approximate $q(\tau_m) = \mathcal{G}(\tau_m | \breve{a}_{\tau_m}, \breve{b}_{\tau_m})$, $q(u_{mn}) = \mathcal{G}(u_{mn} | \breve{a}_{u_{mn}}, \breve{b}_{u_{mn}})$ and $q(\alpha_d) = \mathcal{G}(\alpha_d | \breve{a}_{\boldsymbol{\alpha}}, \breve{b}_{\alpha_d})$ are Gamma density functions updated as follows:

$$\breve{a}_{\tau_m} = a_{\boldsymbol{\tau}} + \tfrac{N_m}{2} \qquad\qquad \breve{b}_{\tau_m} = b_{\boldsymbol{\tau}} + \tfrac{1}{2} \sum_{n|\mathcal{O}_{mn}} \langle u_{mn} \rangle (e_{mn}^2 + \widetilde{\mu}_m + \widetilde{\xi}_{mn})$$

$$\breve{a}_{u_{mn}} = \tfrac{\nu_m}{2} + \tfrac{1}{2} \qquad\qquad \breve{b}_{u_{mn}} = \tfrac{\nu_m}{2} + \tfrac{1}{2} \langle \tau_m \rangle (e_{mn}^2 + \widetilde{\mu}_m + \widetilde{\xi}_{mn})$$

$$\breve{a}_{\boldsymbol{\alpha}} = a_{\boldsymbol{\alpha}} + \tfrac{M}{2} \qquad\qquad \breve{b}_{\alpha_d} = b_{\boldsymbol{\alpha}} + \tfrac{1}{2} \sum_{m=1}^{M} \langle w_{md}^2 \rangle$$

where $\breve{a}_{u_{mn}}$ and $\breve{b}_{u_{mn}}$ are estimated only for observed y_{mn}, N_m denotes the number of observed values in the set $\{y_{mn}\}_{n=1}^N$, while e_{mn} and $\widetilde{\xi}_{mn}$ are shorthand notations for

$$e_{mn} = y_{mn} - \overline{\boldsymbol{w}}_m^{\mathrm{T}} \overline{\boldsymbol{x}}_n - \overline{\mu}_m$$

$$\widetilde{\xi}_{mn} = \overline{\boldsymbol{w}}_m^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{x}_n} \overline{\boldsymbol{w}}_m + \overline{\boldsymbol{x}}_n^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{w}_m} \overline{\boldsymbol{x}}_n + \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{w}_m} \boldsymbol{\Sigma}_{\boldsymbol{x}_n}) \, .$$

The degrees of freedom $\boldsymbol{\nu}$ are point-estimated in order to keep the posterior approximation analytically tractable. The maximum likelihood estimate is found by maximising the lower bound of the model loglikelihood. This yields

$$1 + \log\left(\tfrac{\nu_m}{2}\right) - \psi\left(\tfrac{\nu_m}{2}\right) + \tfrac{1}{N_m} \sum_{n|\mathcal{O}_{mn}} \left(\langle \log u_{mn} \rangle - \langle u_{mn} \rangle \right) = 0 \, ,$$

which can be solved using line search methods. One may try to start updating the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\nu}$ after the iteration has already run for some time if the algorithm seems to converge to bad local optimum.