A Variational Bayesian Method for Rectified Factor Analysis

Markus Harva
Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 5400, FI-02015 HUT, Espoo, Finland

Ata Kaban
School of Computer Science
The University of Birmingham
Birmingham B15 2TT, UK

Abstract— **Linear factor models with nonnegativity constraints have received a great deal of interest in a number of problem domains. In existing approaches, positivity has often been associated with sparsity. In this paper we argue that sparsity of the factors is not always a desirable option, but certainly a technical limitation of the currently existing solutions. We then reformulate the problem in order to relax the sparsity constraint while retaining positivity. A variational inference procedure is derived and this is contrasted to existing related approaches. Both i.i.d. and first-order AR variants of the proposed model are provided and these are experimentally demonstrated in a real-world astrophysical application.**

I. Introduction

Factor analysis is a widespread statistical technique, which seeks to relate multivariate observations to typically smaller dimensional vectors of unobserved variables. These unobserved (latent) variables, termed as factors, are hoped to explain the systematic structure inherent in the data. In standard factor analysis [1], the factors may contain both positive and negative elements. However, in many applications negative values are difficult to interpret. Hence, nonnegativity often is a desirable constraint, that has received considerable interest in recent years.

Positive matrix factorisation [2], nonnegative matrix factorisation [3] and nonnegative independent component analysis [4] are methods that perform a factorisation into positively constrained components. These methods are relatively fast and stable under reasonably mild assumptions, however, they lack a clear probabilistic generative semantics. Bayesian formulations of similar ideas have also been studied [5], [6], [7] in order to enable a series of advantages such as a principled model comparison and inference from previously unseen observations. In these works, positivity of the factors is achieved by formulating a prior that has zero probability mass on the negative axis, such as the exponential, the rectified Gaussian, or mixtures of these. The rectified Gaussian distribution is particularly convenient, as it is conjugate to the Gaussian likelihood and hence it yields a rectified Gaussian posterior distribution.

Unfortunately, all these existing solutions have a serious technical limitation: they hard-wire the assumption that the latent factors are sparse. This is because the likelihood for the location parameter of the latent prior is very awkward and makes it technically impossible to handle a hierarchical prior over it [5]. However, while in some applications both sparsity and positivity are desirable, in others sparsity is inappropriate.

In this paper we provide a different formulation of the positivity constraint in linear factor analysis, which gets round of the mentioned problems. This is achieved by employing a rectification nonlinearity as part of the model. An ordinary Gaussian prior is then employed for the argument of the rectification function, which can further have hierarchical priors for both its location and scale parameter. In this setup, the posterior is no longer of any convenient form, consequently the inference procedure is not as simple as with conjugate priors. However, we show that the free-form variational approximation for the factors is still tractable.

The remainder of the paper is organised as follows: Section II reviews existing solutions to the problem of Bayesian positively constrained factor analysis. Section III presents the proposed formulation and provides the associated inference procedure. Section IV demonstrates a real-world application of the proposed method to astrophysical data analysis. Finally we conclude and discuss further directions.

II. Positively Constrained Generative Factor Analysis

Consider a set of N observed variables, each measured across T different instances. We denote by $\boldsymbol{x}_t \in \mathbb{R}^N$ the t-th instance. The $N \times T$ matrix formed by these vectors is referred to as $\boldsymbol{X}$ and single elements of this matrix will be denoted by x_{nt}. Similar notational convention will also apply to other variables.

As in linear factor analysis, the modelling hypothesis made is that the N observations can be explained as a superposition of $K < N$ underlying latent components $\boldsymbol{s}_t \in \mathbb{R}^K$ (factors or hidden causes) through a linear mapping $\boldsymbol{A} \in \mathbb{R}^{N \times K}$

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{n}_t\,. \tag{1}$$

The noise term $\boldsymbol{n}_t$ is assumed to be zero-mean i.i.d. Gaussian, to account for the notion that all dependencies that exist in $\boldsymbol{x}_t$ should be explained by the underlying hidden components.

A. Imposing Positivity as a Distributional Assumption

A straightforward approach to constraining the factors to be nonnegative is to formulate a nonnegatively supported prior distribution. In doing so, the computationally most convenient alternative is to employ a rectified Gaussian distribution as considered by several authors [5], [6], [7]. It is defined as

$$\mathcal{N}^R(s_k|\bar{s}_k, \tilde{s}_k) = \frac{2}{\mathrm{erfc}(\bar{s}_k/\sqrt{2\tilde{s}_k})} u(s_k)\mathcal{N}(s_k|\bar{s}_k, \tilde{s}_k),$$

where $u(\cdot)$ is the standard step function. It is easy to see that the rectified Gaussian prior is conjugate to a Gaussian likelihood and the posterior can be computed in exactly same manner as with an ordinary Gaussian distribution.

However, as also noted in these works, the computations with the rectified Gaussian prior are only possible if the location parameter $\bar{s}_k$ is fixed to zero, effectively making the erfc term vanish. In all other cases, computations needed to solve the variational problem are intractable.

Consequently, due to the use of a zero-location rectified Gaussian prior on the latent variable, sparse positive factors are induced. While this may be desirable in some applications, it is clearly inappropriate in others as will be shown in Section IV.

B. Imposing Positivity Through a Rectification Nonlinearity

Let us make the following substitution in (1),

$$\boldsymbol{s}_t := f(\boldsymbol{r}_t)\,, \tag{2}$$

where $f : \mathbb{R}^K \to \mathbb{R}^K$ is the component-wise rectification function such that $f_k(\boldsymbol{r}_t) = \max(r_{kt}, 0)$. This guarantees that the factors s_{kt} are positive, no matter what the distribution of r_{kt} is. We employ a Gaussian prior: $r_{kt} \sim \mathcal{N}(m_{rk}, \exp(-v_{rk}))$.

This rectification nonlinearity has previously been used within nonlinear belief networks in [8]. A variational solution is developed in the mentioned work by employing a fixed form Gaussian approximation to the true posterior. By doing so, the cost function can be written analytically [8]. However, the stable points cannot be analytically solved, but require numerical optimisation. Note that finding the global optimum is not trivial due to the existence of multiple stable points. These issues will be illustrated in the next section, where we develop a free-form variational posterior approximation for positively constrained factor analysis.

III. Variational Bayesian Rectified Factor Analysis

In this section we propose a linear factor model that satisfies the positivity constraint by employing the rectification nonlinearity. We refer to this model as Rectified Factor Analysis (RFA).

Once the substitution (2) has been made in (1), a Gaussian prior is then employed over $\boldsymbol{r}$. The resulting model is still linear w.r.t. $\boldsymbol{s}_t$, it satisfies the required positivity constraint due to $f(\cdot)$ and also offers flexibility regarding the location of the probability mass in the latent space. The model can be summarised by the following set of equations:

$$\begin{aligned}
\boldsymbol{x}_t &\sim \mathcal{N}(\boldsymbol{A}f(\boldsymbol{r}_t), \mathrm{diag}(\exp(-\boldsymbol{v}_x)))\\
r_{kt} &\sim \mathcal{N}(m_{rk}, \exp(-v_{rk}))\\
a_{nk} &\sim \mathcal{N}(0, 1)\,.
\end{aligned}$$

To obtain a truly nonnegative model, the weights of the linear mapping need to be constrained to be positive too. This can be achieved by putting a rectified Gaussian prior on them. Vague hierarchical priors are formulated for the rest of the variables.

To make the notation concise, we will refer to the latent variables by $\boldsymbol{\theta}$ and to the data by $\boldsymbol{X}$. Handling the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ is intractable and hence we resort to a variational scheme [9], [10], [11], where an approximative distribution $q(\boldsymbol{\theta})$ is fitted to the true posterior. This is done by constructing a lower bound of the log evidence, based on Jensen's inequality:

$$\begin{aligned}
\log p(\boldsymbol{X}) &= \log \int p(\boldsymbol{X}, \boldsymbol{\theta})\, d\boldsymbol{\theta}\\
&\geq \langle \log p(\boldsymbol{X}, \boldsymbol{\theta})\rangle_{q(\boldsymbol{\theta})} - \langle \log q(\boldsymbol{\theta})\rangle_{q(\boldsymbol{\theta})}\,,
\end{aligned} \tag{3}$$

where $\langle . \rangle_q$ denotes expectation w.r.t. q.

The variational approach to be tractable, the distribution q needs to have suitably factorial form. Here a fully-factorial posterior [12], [10], [11], [13], [14], [6] will be employed.

The model estimation algorithm consists of iteratively updating each variable's posterior approximation in turn, while keeping all other posterior approximations fixed. It can be shown that due to the fully-factorial posterior approximation, all updates are local, i.e. requiring posterior statistics of the so called Markov blanket only. That is, for updating any of the variable nodes, the posterior statistics of its children, parents and co-parents are needed only. This has been exploited in the Bayes Blocks framework [13], [15], [16] which is also used in this work. The scaling of the resulting variational Bayesian algorithm is thus multi-linear in N, T and K, giving the theoretical computational complexity of $O(NTK)$ per iteration.

A. Free-form Posterior Approximation

The fixed form approximation employed in [8] essentially fixes $q(r_{kt})$ to a Gaussian. In this subsection we show that although the free-form approximation of the posterior has a non-standard form, it can be handled analytically, it is more accurate compared to the fixed form approximation and it is also computationally more convenient.

The relevant term of the cost function when updating any given factor r_{kt} is

$$\left\langle \log \frac{q(r_{kt})}{\mathcal{N}(a|f(r_{kt}), b)\mathcal{N}(r_{kt}|c, d)} \right\rangle\,, \tag{4}$$

where a, b, c and d are constants w.r.t. $q(r_{kt})$ and can be computed from the Markov blanket of r_{kt}. Because of the rectification f, the likelihood part in the denominator of (4) is no longer Gaussian, and hence no easy conjugate update rule for $q(r_{kt})$ exists.

Before proceeding to derive the update rule for $q(r_{kt})$, it is worth noticing that once this is completed, the same methodology will apply if a first order AR prior

$$\boldsymbol{r}_t \sim \mathcal{N}\left(\boldsymbol{B}\boldsymbol{r}_{t-1}, \mathrm{diag}[\exp(-\boldsymbol{v}_r)]\right)$$

is considered. Indeed, since the likelihood term at index $t+1$ can be combined with the prior at index $t-1$ (due to the Gaussianity of the prior on $\boldsymbol{r}_t$), an expression that has exactly the same form as (4) is obtained.

We now proceed to deriving the required inference procedure for our model. Tractability of the variational posterior means that analytical expressions can be derived for the following: (i) the cost function:[1]

$$\mathcal{C} = \mathcal{C}_q + \mathcal{C}_p = \langle \log q(r) \rangle - \langle \log p(r|m_r, v_r) \rangle \,,$$

(ii) the posterior mean $\langle r \rangle$ and the variance $\mathrm{Var}\{r\}$ and (iii) the mean $\langle f(r) \rangle$ and the variance $\mathrm{Var}\{f(r)\}$. Here and throughout, $\langle \cdot \rangle$ denote expectations over $q(r)$.

1) The Form of the Posterior: From (4), an invocation of Gibbs' inequality immediately gives us the free form solution:

$$q(r) = \frac{1}{Z} \mathcal{N}\left(a|f(r), b\right) \mathcal{N}\left(r|c, d\right) , \tag{5}$$

where Z is the scaling constant, that will be computed shortly. After some manipulations, (5) can be written as

$$\begin{aligned} q(r) = q_p(r) + q_n(r) = \frac{w_p}{Z} \mathcal{N}\left(r|m_p, v_p\right) u(r) \\ + \frac{w_n}{Z} \mathcal{N}\left(r|m_n, v_n\right) u(-r), \end{aligned}$$

where

$$\begin{aligned} & w_p = \mathcal{N}\left(a|c, b+d\right) \,, \quad w_n = \mathcal{N}\left(a|0, b\right) \,, \\ & v_p = \left(b^{-1} + d^{-1}\right)^{-1} \,, \quad m_p = v_p\left(a/b + c/d\right) , \\ & v_n = d \quad \text{and} \quad m_n = c \,. \end{aligned}$$

Thus, it turns out that the free form posterior approximation is a mixture of two rectified Gaussians. One of these has all its probability mass on the positive real axis whereas the other on the negative axis. The normalisation constant Z of the posterior is then the following:

$$\begin{aligned} Z &= \int \mathcal{N}\left(a|f(r), b\right) \mathcal{N}\left(r|c, d\right) \, dr \\ &= \frac{w_n}{2} \operatorname{erfc}[m_n/\sqrt{2v_n}] + \frac{w_p}{2} \operatorname{erfc}[-m_p/\sqrt{2v_p}]. \end{aligned}$$

2) Relating the Free-Form Approximation to the Fixed-Form Gaussian Approximation: Now, consider fitting the fixed form Gaussian posterior to the true one, e.g. when the quantities in (4) are $a = 1.1$, $b = 0.17$, $c = -1.5$ and $d = 1.2$. The free-form posterior is shown in Figure 1. Looking at its form it should not be surprising that the cost function (which essentially measures the misfit between the approximate and the true posterior) has two stable points. These are shown in Figure 1 in dashed and dot-dashed lines. The dot-dashed line represents the global minimum whereas the dashed line is just a local minimum. The cost function is shown in Figure 2, where the crosses mark the stable points. It is thus clear, that an inference procedure that is able to handle the free-form posterior is preferable to an inference based on the fixed-form Gaussian approximation.

[1]The sub-indexes of r are dropped at this point for convenience.

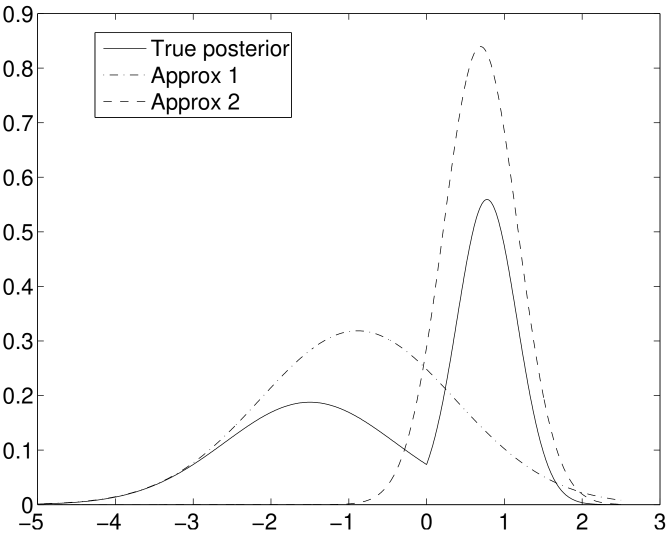

Fig. 1. The true posterior and two Gaussian approximations that are locally optimal.

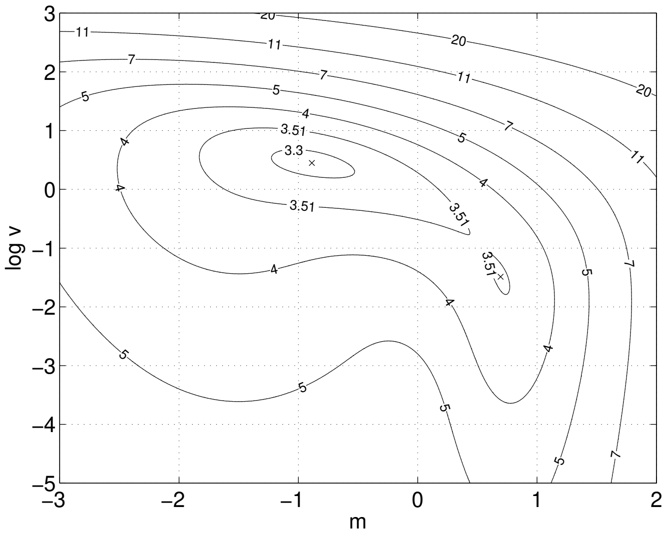

Fig. 2. The cost as a function of the mean m and the log-variance $\log v$ of the Gaussian approximation.

3) Posterior Statistics: Before proceeding to derive the required variational posterior statistics and the cost function, a set of moments are computed. Using these, the expectations as well as the $\mathcal{C}_q$ term of the cost function can be easily expressed.

We define the positive and negative ith order moments as

$$M_p^i = \int r^i \, q_p(r) \, dr \quad \text{and} \quad M_n^i = \int r^i \, q_n(r) \, dr \,. \tag{6}$$

It turns out, that we can express the required expectations and the cost function using the moments of order 0, 1, and 2. The evaluation of these can be cast back to evaluation of the equivalent moments of the rectified Gaussian distribution. The derivations are lengthy and are omitted.

The required posterior statistics are now easily obtained using these moments

$$\begin{aligned}\langle r\rangle &= \int r\, q(r)\, dr = \int r\, q_p(r)\, dr + \int r\, q_n(r)\, dr\\ &= M_p^1 + M_n^1\\ \langle r^2\rangle &= \int r^2\, q(r)\, dr = \int r^2\, q_p(r)\, dr + \int r^2\, q_n(r)\, dr\\ &= M_p^2 + M_n^2\\ \langle f(r)\rangle &= \int f(r) q(r)\, dr = \int r\, q_p(r)\, dr = M_p^1\\ \langle f^2(r)\rangle &= \int f^2(r) q(r)\, dr = \int r^2\, q_p(r)\, dr = M_p^2\,.\end{aligned}$$

The variances are computed using the familiar formula $\mathrm{Var}\{x\} = \langle x^2\rangle - \langle x\rangle^2$.

4) Cost Function: The cost function, which is the negative of the log evidence bound (3), can be used both for monitoring the convergence of the algorithm and more importantly, for comparing different solutions and models.

The term $\mathcal{C}_p$ of the cost function is computed as in the case of ordinary Gaussian variable, see [16] for details. The $\mathcal{C}_q$ term in turn is completely different due to the complex form of the posterior:

$$\begin{aligned}\mathcal{C}_q = \langle \log q(r)\rangle_{q(r)} &= \int q(r) \log q(r)\, dr\\ &= \int q_p(r) \log q(r)\, dr + \int q_n(r) \log q(r)\, dr\\ &= \int q_p(r) \log q_p(r)\, dr + \int q_n(r) \log q_n(r)\, dr\,. \quad (7)\end{aligned}$$

The two terms in (7) can be expressed using the moments derived above. The first term yields

$$\begin{aligned}&\int q_p(r) \log q_p(r)\, dr\\ &= \int q_p(r) \log\left\{\frac{w_p}{Z}\frac{1}{\sqrt{2\pi v_p}} \exp\left[-\frac{1}{2v_p}(r-m_p)^2\right]\right\} dr\\ &= \int q_p(r)\left\{\log\frac{w_p}{Z\sqrt{2\pi v_p}} - \frac{m_p^2}{2v_p} + \frac{m_p}{v_p} r - \frac{1}{2v_p} r^2\right\} dr\\ &= \left(\log\frac{w_p}{Z\sqrt{2\pi v_p}} - \frac{m_p^2}{2v_p}\right) M_p^0 + \frac{m_p}{v_p} M_p^1 - \frac{1}{2v_p} M_p^2\,.\end{aligned}$$

Similarly

$$\begin{aligned}&\int q_n(r) \log q_n(r)\, dr\\ &= \left(\log\frac{w_n}{Z\sqrt{2\pi v_n}} - \frac{m_n^2}{2v_n}\right) M_n^0 + \frac{m_n}{v_n} M_n^1 - \frac{1}{2v_n} M_n^2.\end{aligned}$$

IV. EXPERIMENTS

In this section we present an application of the proposed model to astrophysical data analysis. Experiments have been conducted on both real and synthetic stellar population spectra of elliptical galaxies, addressing both the physical interpretability of the representations created and the predictive capabilities of the models. Ellipticals are the oldest galactic systems in the local Universe and are relatively well understood in physics. The hypothesis that some of these old galactic systems may actually contain young components is relatively new [17]. It is therefore interesting to investigate whether a set of stellar population spectra can be decomposed and explained in terms of a small set of unobserved spectral prototypes in a data driven manner. The positivity constraint is important here, as negative values of flux would not be physically interpretable. The mixing proportions also need to be positive, hence standard rectified Gaussians are employed for the weights s.t. $a_{nk} \sim \mathcal{N}^R(a_{nk}|0,1)$.

A. Missing Values and Measurements Errors

Classical non-probabilistic approaches do not offer the flexibility for taking known measurement errors into account. It is an important practical advantage of the probabilistic framework adopted, that it allows us to handle them in a principled manner. This is achieved simply by making the 'clean' vectors $\boldsymbol{x}_t$ become hidden variables of the additional error model below

$$y_{nt} = x_{nt} + e_{nt}\,.$$

Here e_{nt} is a zero-mean Gaussian noise term with variance v_{ynt} fixed to values that are known from the properties of the physical instrument, for each individual measurement $n = 1:N, t = 1:T$. Missing values can also be handled in this framework by setting v_{ynt} to a very large value.

B. Results on Real Data

A number of $N = 21$ real stellar population spectra will be analysed in this subsection. The data [18] was collected from real elliptical galaxies, along with known measurement uncertainties, given as individual standard deviations on each spectrum & wavelength pair. The data also contains missing entries.

Each of these 21 spectra is characterised by flux values (measured in arbitrary units [18]) given at a number of $T = 339$ different wavelength bins, ranging between 2005-8000 Angstroms. A part of this data set is shown in Figure 3.

In this section we demonstrate three models in terms of the interpretability of their factor representation created. Positive Factor Analysis (PFA) will refer to the method reviewed in Section II-A. Rectified Factor Analysis (RFA) and Dynamic Rectified Factor Analysis (DRFA) refer to the model proposed in this paper and its AR variant respectively. We have fixed the number of factors to two, as inferring subsequent components turns out to have no physical interpretation. We repeated each run ten times with random initialisations drawn from $\mathcal{N}^R(0,1)$. The model with smallest cost function value was then selected. The two components[2] for each of the models are shown in Figure 4. The shape of the first component is very similar for all three methods considered. This component can visually be recognised to correspond to an old and high

[2]The order of the components is of course arbitrary, we have manually grouped them for the ease of visual inspection.

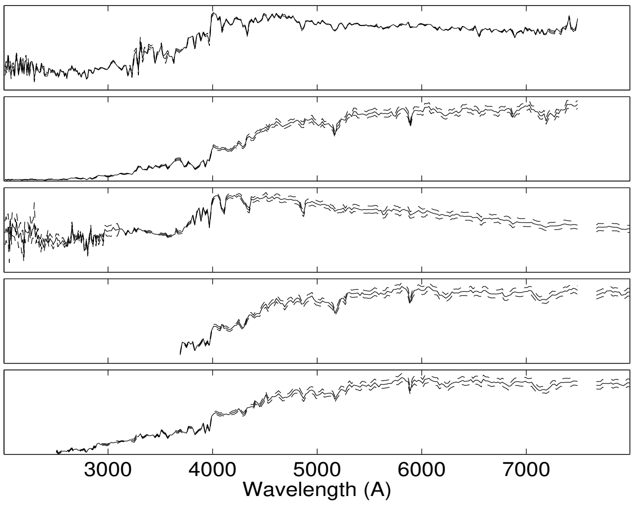

Fig. 3. A sample from the real data of spectral measurements. The dashed lines show the standard deviations of the errors in the data. The blank entries denote missing values.

metallicity stellar population. This kind of component in elliptical stellar populations has been known to physicists for a long time. In turn, the existence of a second component is a relatively recent finding in astrophysics [18].

Interestingly, the second component inferred from the data differs more across the models considered. However, the RFA second component turned out to be physically interpretable, as it exhibits many of the characteristic features of a young and low metallicity stellar population spectrum. The second component from DRFA is similar in its main shape, providing an indication for the age of this stellar population component, however it lacks some of the wiggles that encode metallicity characteristics of the stellar population.

From astrophysical point of view, the second component of PFA has no clear physical interpretation, as its distribution is biased toward zero. This is most likely due the fact that the location parameter for the rectified Gaussian distribution is required to be zero and hence small values are favoured. This results in a poor match with any known physical model. The sparsity constraint of PFA is clearly inappropriate in this application.

The values of the cost function at $K = 2$ are detailed in Figure 5. The contributions of the various individual terms of the overall cost associated to these methods are also shown in this figure. DRFA provides the lowest overall cost, since it is able to code the factors most compactly [19]. However, as we can also read from the figure, the error term corresponding to the data reconstruction accuracy is a little larger (compared to the other two methods considered here). RFA has the smallest cost for the data reconstruction term as well, implying an accurate reconstruction of the data details.

C. Prediction Results on Synthetic Data

Here we employ synthetic spectra in order to assess the predictive performance of the proposed methods in an ob-

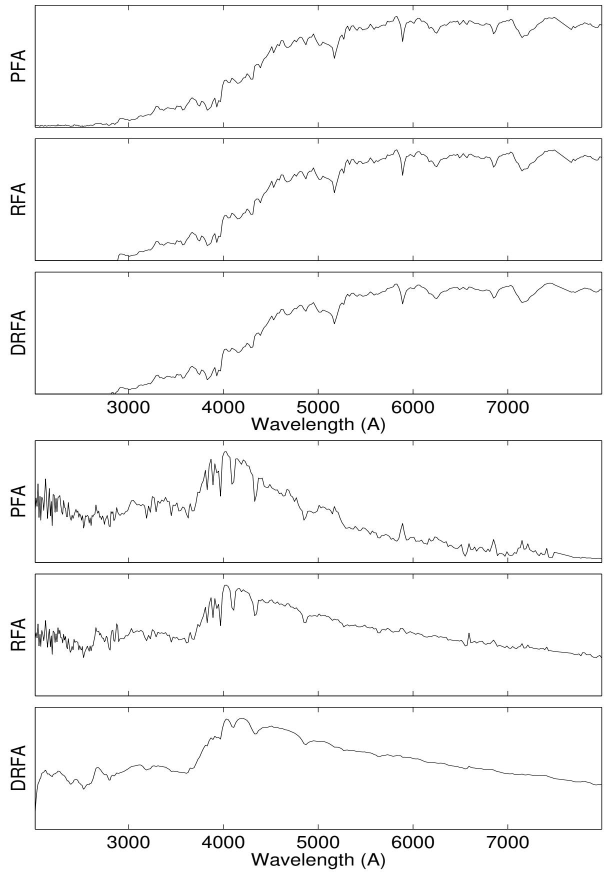

Fig. 4. The first (above) and second (below) components for the different models.

jective and controlled manner. A random selection of 100 synthetic composite spectra produced from the stellar population evolutionary synthesis model of Jimenez [18] is utilised. Each of the these may contain the superposition of two stellar population spectra with varying parameters (age, metallicity and proportion). The wavelength coverage as well as the binning of these spectra is identical to those described for the real data. The mixing proportions depend on the masses of the component stellar populations in a physically realistic manner. There are no missing entries or measurement errors in this data set, making it suitable for an controlled assessment.

We consider an inference task where half of the flux values at a random selection of wavelength bins are held out as a test set and used for evaluation purpose only. Missing values are artificially created at random in the test set. The RFA and DRFA models are trained on the same training set and asked to predict the artificially created missing entries in the previously unseen test set. The prediction can be obtained simply from the posterior mean of $q(x_{nt})$. The number of factors has been $K = 3$, determined from the evidence bound, for both models in this experiment. The SNR between the predictions and the true values, when varying the percentage of missing values in the test set, are shown in Figure 6. Clearly, as expected, DRFA outperforms RFA in this prediction task, especially

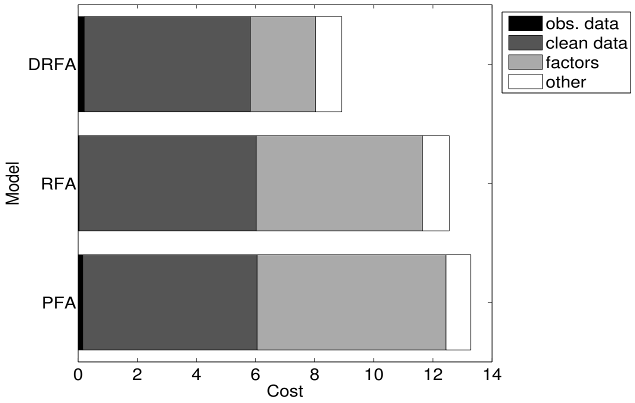

Fig. 5. The cost (divided by the number of samples) of the various models considered. Individual terms of the overall cost are highlighted for each model. Obs. data stands for $\boldsymbol{y}_t$, clean data for $\boldsymbol{x}_t$, factors for $\boldsymbol{r}_t$ (or $\boldsymbol{s}_t$ in case of PFA) and other for other variables in the models such as the noise variances v_{xn}.

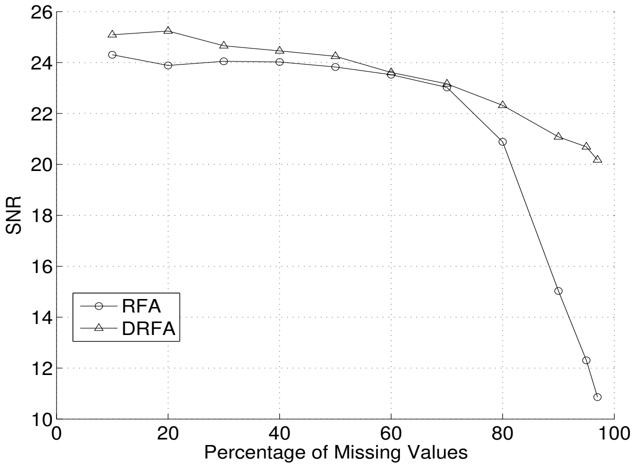

Fig. 6. Prediction of missing entries in out-of-sample wavelength bins.

when the amount of missing values gets large. The reason for this is that DRFA includes the modelling of the correlations between fluxes at neighbouring wavelength bins. Evidently this information turns out to be useful in the prediction task considered.

V. Discussion

We presented a method for nonnegative factor analysis, based on variational Bayesian learning. The proposed solution gets round of the shortcomings of approaches that impose a positively supported prior directly on the latent space. We derived analytical expressions for performing inference in this model, using a factorial free-form approximation for the factors. We demonstrated the proposed model in an astrophysical data analysis application, where approaches that induce sparse representations are inappropriate. The presented approach is applicable in any situation where flexible latent densities over the positive domain are required.

We note that the methodology developed and employed here can straightforwardly be extended e.g. to include multiple rectification. Also, Gaussian mixture priors for the argument of this function could be employed in place of the single Gaussian utilised here, in order to further enhance flexibility.

Acknowledgements

This research has been funded by the Finnish Centre of Excellence Programme (2000-2005) under the project New Information Processing Principles and a Paul & Yuanbi Ramsay research award at the School of Computer Science of The University of Birmingham. Many thanks to Louisa Nolan and Somak Raychaudhury for sharing their astrophysical expertise and supplying the data.

References

[1] R. L. Gorsuch, *Factor Analysis*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

[2] P. Paatero and U. Tapper, "Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values," *Environmetr.*, vol. 5, pp. 111–126, 1994.

[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[4] M. Plumbley and E. Oja, "A "nonnegative PCA" algorithm for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 15, no. 1, pp. 66–76, 2004.

[5] J. Miskin, "Ensemble learning for independent component analysis," Ph.D. dissertation, University of Cambridge, UK, 2000.

[6] J. Winn and C. M. Bishop, "Variational message passing," *Journal of Machine Learning Research*, 2004, submitted.

[7] M. Harva, "Hierarchical variance models of image sequences," Master's thesis, Helsinki University of Technology, Espoo, 2004.

[8] B. J. Frey and G. E. Hinton, "Variational learning in nonlinear Gaussian belief networks," *Neural Computation*, vol. 11, no. 1, pp. 193–214, 1999.

[9] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. Jordan, Ed. Cambridge, MA, USA: The MIT Press, 1999, pp. 105–161.

[10] H. Lappalainen, "Ensemble learning for independent component analysis," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, Aussois, France, 1999, pp. 7–12.

[11] H. Attias, "A variational Bayesian framework for graphical models," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K.-R. Müller, Eds., vol. 12. MIT Press, 2000, pp. 209–215.

[12] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[13] H. Valpola, T. Raiko, and J. Karhunen, "Building blocks for hierarchical latent variable models," in *Proc. 3rd Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, USA, 2001, pp. 710–715.

[14] C. M. Bishop, D. Spiegelhalter, and J. Winn, "VIBES: A variational inference engine for Bayesian networks," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermayer, Eds., vol. 15. MIT Press, 2003, pp. 793–800.

[15] H. Valpola, A. Honkela, M. Harva, A. Ilin, T. Raiko, and T. Östman, "Bayes blocks software library," *http://www.cis.hut.fi/projects/bayes/software/*, 2003.

[16] T. Raiko, H. Valpola, M. Harva, and J. Karhunen, "Building blocks for variational Bayesian learning of latent variable models," *Journal of Machine Learning Research*, 2004, submitted.

[17] D. S. Madgwick, A. L. Coil, C. J. Conselice, M. C. Cooper, M. Davis, R. S. Ellis, S. M. Faber, D. P. Finkbeiner, B. Gerke, P. Guhathakurta, N. Kaiser, D. C. Koo, J. A. Newman, A. C. Phillips, C. C. Steidel, B. J. Weiner, C. N. A. Willmer, and R. Yan, "The DEEP2 galaxy redshift survey: Spectral classification of galaxies at $z \sim 1$," *The Astrophysical Journal*, vol. 599, no. 2, pp. 997–1005, 2003.

[18] L. Nolan, "The star formation history of elliptical galaxies," Ph.D. dissertation, The University of Edinburgh, UK, 2002.

[19] A. Honkela and H. Valpola, "Variational learning and bits-back coding: an information-theoretic view to Bayesian learning," *IEEE Transactions on Neural Networks*, vol. 15, no. 4, pp. 800–810, 2004.