Variational Learning for Rectified Factor Analysis*

Markus Harva[1]

Adaptive Informatics Research Centre, Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Espoo, Finland

Ata Kabán

School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK

Abstract

Linear factor models with non-negativity 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. This is achieved by employing a rectification nonlinearity rather than a positively supported prior directly on the latent space. A variational learning procedure is derived for the proposed model and this is contrasted to existing related approaches. Both i.i.d. and first-order AR variants of the proposed model are provided and they are experimentally demonstrated with artificial data. Application to the analysis of galaxy spectra show the benefits of the method in a real world astrophysical problem, where the existing approach is not a viable alternative.

Keywords: positive factor analysis; variational Bayes; source separation

1 Introduction

Factor analysis is a widespread statistical technique, which seeks to relate multivariate observations to typically smaller dimensional vectors of unobserved

* Expanded version of the work presented at IJCNN 2005 [1].

[1] Corresponding author. tel. +358 9 451 3287, fax +358 9 451 3277, email: markus.harva@hut.fi

variables. These unobserved (latent) variables, termed as factors, are hoped to explain the systematic structure inherent in the data. In standard factor analysis [2], the factors may contain both positive and negative elements. However, in many applications negative values are difficult to interpret. Hence, non-negativity often is a desirable constraint, that has received considerable interest in recent years.

Positive matrix factorisation [3], non-negative matrix factorisation [4] and non-negative independent component analysis [5] 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 [6–8] in order to enable a series of advantages such as inference from previously unseen observations and principled model comparison. 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, meaning that the probability mass is concentrated near zero. 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. While in some applications both sparsity and positivity are desirable, in others, as will be seen, 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, obtaining the so called free-form posterior approximation is not immediate, consequently the inference procedure is not as simple as with conjugate priors. However, we show that the free-form variational approximation for the model is still tractable.

Our proposed setup is related to the one by Frey and Hinton [9], where a rectification nonlinearity has been used in connection with nonlinear belief networks. However, our solution differs substantially from that of [9], where a fixed-form Gaussian posterior approximation has been used and the required optimisation has been solved by gradient descent. Our free-form variational algorithm, in addition to being more accurate, does not require any numerical optimisation.

The remainder of the paper is organised as follows: Section 2 reviews existing solutions to the problem of Bayesian positively constrained factor analysis. Section 3 presents the proposed formulation and provides the associated inference procedure. The advantages of the proposed method are first illustrated using artificial data in Section 4. Section 5 demonstrates a real-world application of the proposed method to astrophysical data analysis. Finally we conclude and discuss further directions.

2 Positively Constrained Generative Factor Analysis

Consider a set of N observed variables, each measured across T different instances. We denote by $\mathbf{x}_t \in \mathbb{R}^N$ the t-th instance. The $N \times T$ matrix formed by these vectors is referred to as $\mathbf{X}$ and single elements of this matrix will be denoted by x_{it}. 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 M underlying positive latent components $\mathbf{s}_t \in \mathbb{R}^M$ (factors or hidden causes) through a linear mapping $\mathbf{A} \in \mathbb{R}^{N \times M}$

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

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

2.1 Imposing Positivity as a Distributional Assumption

A straightforward approach to constraining the factors to be non-negative is to formulate a non-negatively supported prior distribution. In doing so, the computationally most convenient alternative is to employ a rectified Gaussian distribution as considered by several authors [6,10,7]. This is defined as

$$\mathcal{N}^R\left(s|m,v\right) = \frac{2}{\mathrm{erfc}(-m/\sqrt{2v})} u(s) \mathcal{N}\left(s|m,v\right),$$

where $u(\cdot)$ is the standard step function, and m and v are the location and scale parameters respectively. 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 m is fixed to zero, effectively making the erfc term vanish. Otherwise, the computations needed

to solve the variational problem are intractable. One option is to resort to an empirical Bayes estimation of the parameters of the rectified Gaussian – which of course would be tractable [8]. However this would lead to over-confident estimates and one would lose a number of further benefits of the more stringent Bayesian approach, e.g. that of making principled model comparisons. Another option is to work with fixed parameters. A common choice in the literature has been to fix the location parameter to zero [6,10,7]. Consequently, due to the use of a zero-location rectified Gaussian prior on the latent variable, sparse [2] positive factors are induced. While this may be desirable in some applications, it is clearly inappropriate in others as will be shown in Section 5.

2.2 Imposing Positivity Through a Rectification Nonlinearity

Let us make the following substitution in (1),

$$\mathbf{s}_t := \mathbf{cut}(\mathbf{r}_t)\,, \tag{2}$$

where **cut** is the component-wise rectification (or cut) function $[\mathbf{cut}(\boldsymbol{\theta})]_i = \mathrm{cut}(\theta_i) = \max(\theta_i, 0)$. This guarantees that the factors $\mathbf{s}_t$ are positive, no matter what the distribution of $\mathbf{r}_t$ is. We employ a Gaussian prior: $r_{jt} \sim \mathcal{N}\left(m_{rj}, \tau_{rj}^{-1}\right)$. The distribution of $s = \mathrm{cut}(r)$ shall be denoted as $\mathcal{R}^N(m, v)$. This is a mixture of a Dirac delta distribution at zero and a rectified Gaussian distribution. Hence, it is genuinely different from the rectified Gaussian distribution $\mathcal{N}^R(m, v)$. [3]

The resulting model is still linear w.r.t. $\mathbf{s}_t$, it satisfies the required positivity constraint due to the cut function and also offers flexibility regarding the location of the probability mass in the latent space. The probabilistic model

[2] In the context of this paper, the term sparse refers to bias towards zero.

[3] Naturally the terminology is a matter of convention — either of these two distributions $\mathcal{N}^R$ or $\mathcal{R}^N$ could be termed as the 'rectified Gaussian'. Indeed, differently from the initial use of this term [11] (where it meant a multivariate distribution over the positive domain), in [12], $\mathcal{R}^N$ has been termed as rectified Gaussian, whereas other, more recent works [6,7], including this paper, refer to $\mathcal{N}^R$ by the same term. In order to avoid confusion, we shall explicitly use two different symbols, $\mathcal{N}^R$ and $\mathcal{R}^N$, to refer to these two distinct distributions.

definition is fully summarised by the following set of equations:

$$
\begin{aligned}
x_{it} &\sim \mathcal{N}\left(\mathbf{a}_i^T \mathbf{cut}(\mathbf{r}_t), \tau_{xi}^{-1}\right) \\
r_{jt} &\sim \mathcal{N}\left(m_{rj}, \tau_{rj}^{-1}\right) \\
m_{rj} &\sim \mathcal{N}\left(0, \sigma_{mr}^2\right) \\
\tau_{xi} &\sim \mathcal{G}\left(\alpha_x, \beta_x\right) \\
\tau_{rj} &\sim \mathcal{G}\left(\alpha_r, \beta_r\right) \\
a_{ij} &\sim \mathcal{N}^R\left(0, 1\right)
\end{aligned}
$$

In the above, $\mathcal{G}(\alpha, \beta)$ denotes the Gamma distribution [4] and the symbols σ_{mr}^2, α_x, β_x, α_r, β_r are constants, whose values should be chosen to match the prior beliefs about the problem in question. In the experiments of this paper we used $\sigma_{mr}^2 = 100$, $\alpha_x = \alpha_r = 1$ and $\beta_x = \beta_r = 10^{-4}$ in order to express vague (but proper) priors for the variables at the top of the hierarchical specification. The prior for the weights a_{ij} of the linear mapping $\mathbf{A}$ was specified as a zero-location rectified Gaussian in order to express a positively constrained mapping. Finally, the rationale behind having chosen the inverse parametrisation for the variances is to enjoy the computational convenience of working with conjugate priors [13] when possible. We refer to the model specified above as Rectified Factor Analysis (RFA).

There is earlier work on the use of rectification nonlinearities. In [14], the rectification nonlinearity for multiple-cause modelling has been considered from the biological plausibility starting point. The authors propose a constrained PCA network which can learn sparsely distributed representations of data sets. However, the model is non-probabilistic and doesn't include flexibility for modelling non-sparse factors. In [15], the rectification is used to refine the expectation maximisation approach to subspace analysis, and obtain an algorithm more suitable for non-negative data.

In the variational setting, the rectification has previously been used within nonlinear belief networks in [9]. However, their variational approximation differs from ours in that they employ a fixed-form Gaussian approximation to the true posterior. In contrast, we develop a free-form approximation, the advantages of which will be detailed in the sequel.

[4] The parametrisation of the Gamma-distribution varies in the literature. In this paper the following is used: $\mathcal{G}(\theta|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \mathrm{e}^{-\beta\theta}$

3 Variational Inference for Rectified Factor Analysis

In this section we derive a variational Bayesian (VB) learning procedure for the model proposed in Section 2.2. Before proceeding, a brief review is given on the variational Bayesian methodology employed.

3.1 Variational Bayes

In Bayesian data modelling, all information is encoded in probability distributions [13]. The exact Bayesian approach starts from constructing a model expressed as the joint probability density function (pdf) $p(\mathbf{X}, \boldsymbol{\theta})$ of the data $\mathbf{X}$ and the parameters $\boldsymbol{\theta}$. Once the model has been specified, the inference simply consists of computing the posterior pdf of the parameters $p(\boldsymbol{\theta}|\mathbf{X})$ i.e. the pdf of the parameters of the model given the observed data. This is done by an invocation of the Bayes' rule. The obtained posterior distribution can be used for making decisions and predictions. The marginal likelihood of the data (called also the model evidence) can be used to compute posterior probabilities of competing models, and therefore it is a tool for comparing different models.

However, although the principle is simple, the required computations are rarely feasible, for one or several of the following reasons.

(1) The normalisation constant of the posterior is intractable. This implies that also the marginal likelihood is intractable.
(2) The marginal densities of the parameters are intractable.
(3) Due to symmetries in the model, the marginal densities computed from the exact posterior are of little interest. This is especially true in factor models since the model is symmetrical w.r.t. permutations of the factors making the posterior pdf have $M!$ equivalent modes (where M was the number of factors).

Hence, often approximations of some form are necessary in practise for Bayesian inference, both due to intractability issues and due to difficulties in interpreting the exact posterior.

Variational Bayesian learning (known also as variational Bayes, variational learning, ensemble learning, and VB for short) (see e.g. [16–18]) is one viable approach where an approximate distribution $q(\boldsymbol{\theta})$ is fitted to the true posterior. This is done by constructing a lower bound for the log evidence, based on

Jensen's inequality:

$$\begin{aligned}\mathcal{B}(q) &:= \langle \log p(\mathbf{X}, \boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta})} - \langle \log q(\boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta})} = \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \, \mathrm{d}\boldsymbol{\theta} \\ &\leq \log \int q(\boldsymbol{\theta}) \frac{p(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \, \mathrm{d}\boldsymbol{\theta} = \log \int p(\mathbf{X}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} = \log p(\mathbf{X})\end{aligned} \tag{3}$$

where $\langle \cdot \rangle_q$ denotes expectation w.r.t. q. The approximate distribution q is found by functional maximisation of the bound $\mathcal{B}(q)$ w.r.t. q. To make the integral tractable, the distribution q needs to have a suitably factorial form. Here a fully-factorial posterior will be employed, meaning that the approximation has the form

$$q(\boldsymbol{\theta}) = \prod_i q(\theta_i) \, . \tag{4}$$

The model estimation algorithm consists of iteratively updating each variable's posterior approximation $q(\theta_i)$ in turn, while keeping all other posterior approximations fixed. Due to the chosen form (4), all updates are local, 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.

Since all variables are integrated over (at least approximately), VB is robust against overfitting — a reoccurring problem when using point estimates. Also, due to the form of the posterior approximation, it will approximate only one of the prominent modes of the exact posterior, therefore the obtained approximation can be hoped to be more easily interpretable. For example, reporting the means of the posterior approximation is perfectly sensible even in cases when it would be completely meaningless with the true posterior. Finally, the bound in (3) can be turned into an approximation of the posterior probabilities over competing models. Thus, the VB methodology equips us with an immediate tool for making inferences, model selections and model comparisons.

The practical utility of employing a fully-factorial approximate posterior is that it leads to a computationally economic learning algorithm. It has empirically been found to work well for a number of models in the literature. However, the rigorous study of the implications of this approximation and how much the obtained estimates differ from those of exact methods is in its infancy, and represents an active topic of research in statistics. There are some general results for the exponential family of models, regarding the asymptotic properties [19] of such approximations. There is also some evidence [20] that the approximate model probabilities provided by VB compare very well even with the best sampling based methods.

If the form of the factors $q(\theta_i)$ in the approximation is not further restricted, by requiring them to belong to a certain family of distributions such as Gaussians, the approximation is said to be free form. If the opposite holds — that assumptions beyond Eq. (4) are made — the approximation is called fixed form.

A free-form variational posterior is not always tractable to compute, which is the very reason why fixed-form approximations are sometimes necessary. In this subsection we show that although the free-form approximation for the RFA model 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 problematic part in the posterior approximation are the factors, since for the other variables in the model, there exist conjugate update rules. Hence, here we will be concentrating on the part of the algorithm that deals with the factors r_{jt}. The computation of the posterior approximation for the rest of the variables in the model is given in Appendix C.

The relevant term in the evidence bound (3) when updating any given factor r_{jt} is

$$-\left\langle \log \frac{q(r_{jt})}{\mathcal{N}\left(a|\,\mathrm{cut}(r_{jt}), b\right)\mathcal{N}\left(r_{jt}|c,d\right)} \right\rangle , \tag{5}$$

where a, b, c, and d are constants w.r.t. $q(r_{jt})$ and can be computed from the Markov blanket of r_{jt}. The exact formulae for them are given in Appendix C (Eqs. (C.5) - (C.8)). Because of the rectification, the likelihood part in the denominator of (5) is no longer Gaussian, and hence no easy conjugate update rule for $q(r_{jt})$ exists.

We now proceed to derive the learning procedure for our model. Tractability of the variational posterior means that analytical expressions can be derived for the following: (i) the relevant part of the evidence bound:[5] $\langle \log p(r|m_r, \tau_r)\rangle - \langle \log q(r)\rangle$, (ii) the posterior mean $\langle r\rangle$ and the variance $\mathrm{var}(r)$ and (iii) the mean $\langle \mathrm{cut}(r)\rangle$ and the variance $\mathrm{var}(\mathrm{cut}(r))$. Here and throughout, $\langle\cdot\rangle$ denote expectations over $q(r)$.

[5] The sub-indices of r are dropped at this point for convenience.

3.2.1 The Form of the Posterior

From (5), an invocation of Gibbs' inequality provides the following free-form solution:

$$q(r) = \frac{1}{Z}\mathcal{N}\left(a|\,\mathrm{cut}(r), b\right)\mathcal{N}\left(r|c, d\right), \tag{6}$$

where Z is the normalising constant, that will be computed shortly. After some manipulations, (6) can be written as $q(r) = q_p(r) + q_n(r)$, where

$$q_p(r) = \frac{w_p}{Z}\mathcal{N}\left(r|\mu_p, \sigma_p^2\right)u(r) \quad \text{and} \quad q_n(r) = \frac{w_n}{Z}\mathcal{N}\left(r|\mu_n, \sigma_n^2\right)u(-r)\,,$$

and for which

$$\begin{aligned} w_p &= \mathcal{N}\left(a|c, b+d\right)\,, \quad w_n = \mathcal{N}\left(a|0, b\right)\,, \\ \sigma_p^2 &= \left(b^{-1} + d^{-1}\right)^{-1}\,, \quad \mu_p = \sigma_p^2\left(a/b + c/d\right), \\ \sigma_n^2 &= d \quad \text{and} \quad \mu_n = c\,. \end{aligned}$$

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

$$\begin{aligned} Z = \int \mathcal{N}\left(a|\,\mathrm{cut}(r), b\right)\mathcal{N}\left(r|c, d\right)\,\mathrm{d}r \\ = \frac{w_n}{2}\,\mathrm{erfc}[\mu_n/\sqrt{2\sigma_n^2}] + \frac{w_p}{2}\,\mathrm{erfc}[-\mu_p/\sqrt{2\sigma_p^2}]. \end{aligned}$$

3.2.2 Relating the Free-Form Approximation to the Fixed-Form Gaussian Approximation

As already mentioned in Section 2.2, a fixed-form variational solution exists for the linear model with rectification nonlinearity, as developed in [9] for nonlinear belief networks. It is therefore interesting to compare this to our free-form approximation. With the fixed-form approximation, the evidence bound can be written analytically [9]. 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.

Consider fitting a fixed-form Gaussian approximation to the true posterior in an example case when the quantities in (5) 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 log-evidence bound 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 log-evidence bound as the function of the parameters of the

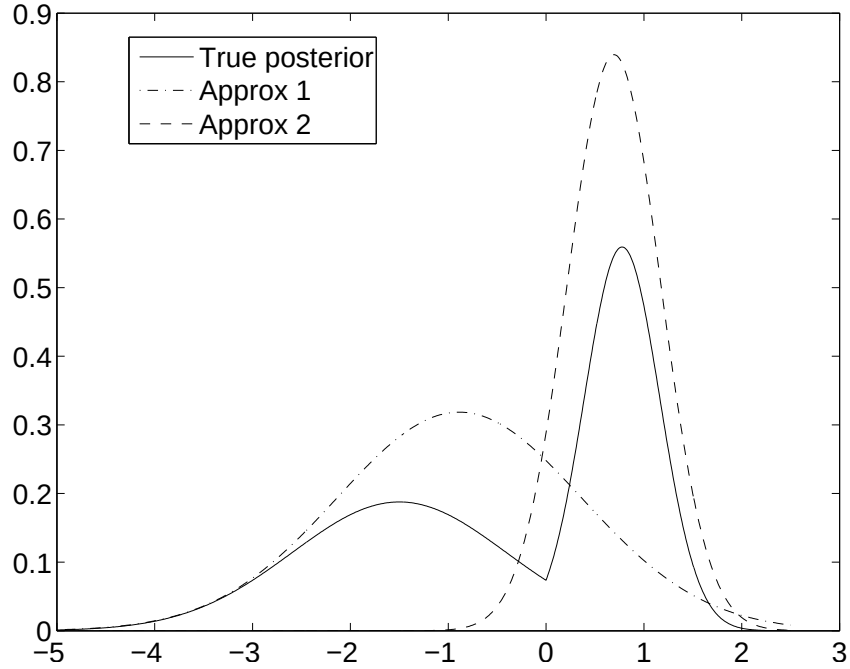


Fig. 1. An example of the true posterior for a factor in RFA and two Gaussian approximations that are locally optimal. The dot-dashed line represents the better optimum.

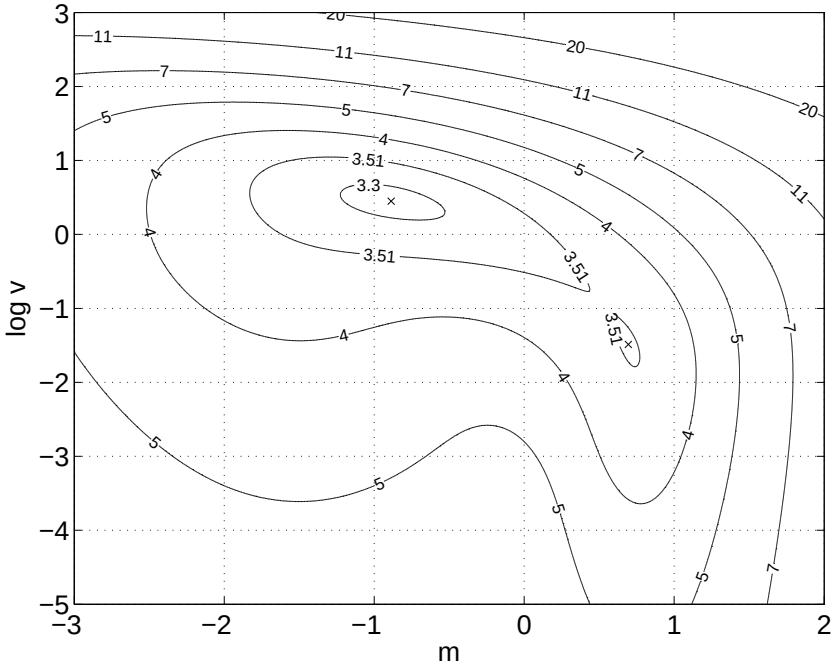


Fig. 2. The negative of the log-evidence bound as a function of the mean m and the log-variance $\log v$ of the Gaussian approximation when it is fitted to the true posterior of Figure 1. The two local optima are marked with crosses.

approximation is shown in Figure 2, where the crosses mark the stable points. This is to demonstrate the complications that can rise when the posterior approximation is further restricted.

3.2.3 Posterior Statistics

In the computation of the required variational posterior statistics and the evidence bound, we can make use of certain moments computed over the approximation.

Define the positive and negative ith order moments as

$$M_p^i = \int r^i \, q_p(r) \, \mathrm{d}r \quad \text{and} \quad M_n^i = \int r^i \, q_n(r) \, \mathrm{d}r \, . \tag{7}$$

It turns out, that we can express the required expectations and the evidence bound using the moments of order 0, 1, and 2. The evaluation of these can be cast back to the evaluation of the equivalent moments of the rectified Gaussian distribution. The expressions are lengthy and hence their presentation is postponed to Appendix B.

The required posterior statistics are now easily obtained using the moments

$$\begin{aligned}
\langle r \rangle &= \int r \, q(r) \, \mathrm{d}r = \int r \, q_p(r) \, \mathrm{d}r + \int r \, q_n(r) \, \mathrm{d}r = M_p^1 + M_n^1 \\
\left\langle r^2 \right\rangle &= \int r^2 \, q(r) \, \mathrm{d}r = \int r^2 \, q_p(r) \, \mathrm{d}r + \int r^2 \, q_n(r) \, \mathrm{d}r = M_p^2 + M_n^2 \\
\langle \mathrm{cut}(r) \rangle &= \int \mathrm{cut}(r) q(r) \, \mathrm{d}r = \int r \, q_p(r) \, \mathrm{d}r = M_p^1 \\
\left\langle \mathrm{cut}^2(r) \right\rangle &= \int \mathrm{cut}^2(r) q(r) \, \mathrm{d}r = \int r^2 \, q_p(r) \, \mathrm{d}r = M_p^2 \, .
\end{aligned}$$

3.2.4 The Evidence Bound

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 $\langle \log p(r|m_r, \tau_r) \rangle$ appearing in the bound is computed as in the case of an ordinary Gaussian variable. See Appendix E for details. The term $\langle \log q(r) \rangle$ in turn is completely different due to the complex form of the posterior:

$$\begin{aligned}
\langle \log q(r) \rangle_{q(r)} &= \int q(r) \log q(r) \, \mathrm{d}r \\
&= \int q_p(r) \log q(r) \, \mathrm{d}r + \int q_n(r) \log q(r) \, \mathrm{d}r \\
&= \int q_p(r) \log q_p(r) \, \mathrm{d}r + \int q_n(r) \log q_n(r) \, \mathrm{d}r \, .
\end{aligned} \tag{8}$$

The two terms in (8) can be expressed using the moments derived above. The first term yields

$$\begin{aligned}
&\int q_p(r) \log q_p(r)\ \mathrm{d}r \\
&\qquad = \int q_p(r) \log \left\{ \frac{w_p}{Z} \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left[-\frac{1}{2\sigma_p^2}(r-\mu_p)^2 \right] \right\}\ \mathrm{d}r \\
&\qquad = \int q_p(r) \left\{ \log \frac{w_p}{Z\sqrt{2\pi\sigma_p^2}} - \frac{\mu_p^2}{2\sigma_p^2} + \frac{\mu_p}{\sigma_p^2} r - \frac{1}{2\sigma_p^2} r^2 \right\}\ \mathrm{d}r \\
&\qquad\qquad = \left(\log \frac{w_p}{Z\sqrt{2\pi\sigma_p^2}} - \frac{\mu_p^2}{2\sigma_p^2} \right) M_p^0 + \frac{\mu_p}{\sigma_p^2} M_p^1 - \frac{1}{2\sigma_p^2} M_p^2 \,. \quad (9)
\end{aligned}$$

Similarly

$$\int q_n(r) \log q_n(r)\ \mathrm{d}r = \left(\log \frac{w_n}{Z\sqrt{2\pi\sigma_n^2}} - \frac{\mu_n^2}{2\sigma_n^2} \right) M_n^0 + \frac{\mu_n}{\sigma_n^2} M_n^1 - \frac{1}{2\sigma_n^2} M_n^2. \quad (10)$$

3.3 Extending the Model to the AR(1) Case

Making the factors follow an AR(1) process can be accomplished by changing their prior to

$$\begin{aligned}
r_{j1} &\sim \mathcal{N}\left(0, \sigma_{r1}^2\right) \\
r_{jt} &\sim \mathcal{N}\left(\mathbf{b}_j^T \mathbf{r}_{t-1} + c_j, \tau_{rj}^{-1}\right) \quad (t > 1)\,.
\end{aligned}$$

This extension doesn't complicate matters terribly, since the form of the posterior approximation does not change. Indeed, since now the likelihood term at index $t+1$ can be combined with the prior at index t (due to the Gaussianity of the prior on $\mathbf{r}_t$), an expression that has exactly the same form as (5) is obtained. The exact update rules are given in Appendix D. The new parameters $\mathbf{B}$ and $\mathbf{c}$ have the priors

$$\begin{aligned}
b_{ij} &\sim \mathcal{N}(0,1) \\
c_j &\sim \mathcal{N}\left(0, \sigma_c^2\right)\,.
\end{aligned}$$

4 Experiments on Artificial Data

In this section experiments with artificial data are presented. We will consider three models. Positive Factor Analysis (PFA) will refer to the method reviewed

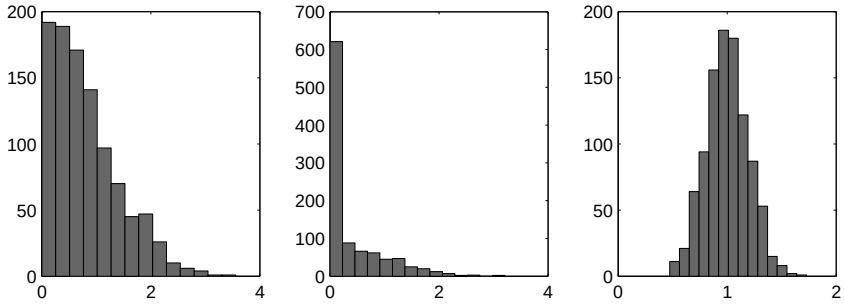


Fig. 3. Histograms of the factors.

in Section 2.1. Rectified Factor Analysis (RFA) and Dynamic Rectified Factor Analysis (DRFA) refer to the model proposed in this paper and its AR variant respectively.

4.1 Static Factors

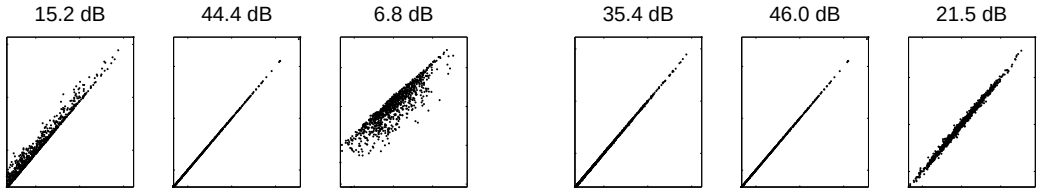
4.1.1 Factor Distributions

First we demonstrate the flexibility of the proposed model in recovering both sparse and non-sparse positive factor distributions, as opposed to the existing alternative approach (PFA), which is restricted to sparse factors.

Let us consider a dataset which is a linear mixture of three underlying factors distributed as $\mathcal{N}^R(0,1)$, $\mathcal{R}^N(0,1)$ and $\mathcal{R}^N(1,0.2^2)$. The histograms of the 1000 samples obtained from these distributions are shown in Figure 3. From the three factors, ten observations were generated by a linear mapping whose weights were sampled from $\mathcal{R}^N(0,1)$. Finally zero mean Gaussian noise with standard deviation (std) 0.01 was added to each observation.

The factors were estimated from the observations using PFA and RFA. For each model the factors were randomly initialised and the learning algorithm was iterated 2000 times. This procedure was repeated ten times and the result with the highest evidence for each of the models was then selected. Figure 4 shows the separation results as scatter plots between the true and the (appropriately permuted) estimated factors. The leftmost three figures are the samples of the PFA factors vs. the ground truth, whereas the rightmost three figures are the RFA factors vs. the same ground truth. The goodness of each can be seen visually by the departure from a straight line. In particular, we see considerable mismatch in the third factor of PFA. In addition, we evaluate the performance as the signal to noise ratio (SNR), defined as

$$\mathrm{SNR} := 10 \log_{10} \frac{\sigma_s^2}{\sigma_e^2}\,, \tag{11}$$



(a) PFA (b) RFA

Fig. 4. Separation results as correlation plots with SNRs printed above. Here both the original and the estimated signals are normalised to the same scale and they are plotted against each other. Hence the optimal result would be a straight thin line. The third factor poses serious difficulties for PFA whereas RFA can model it very well.

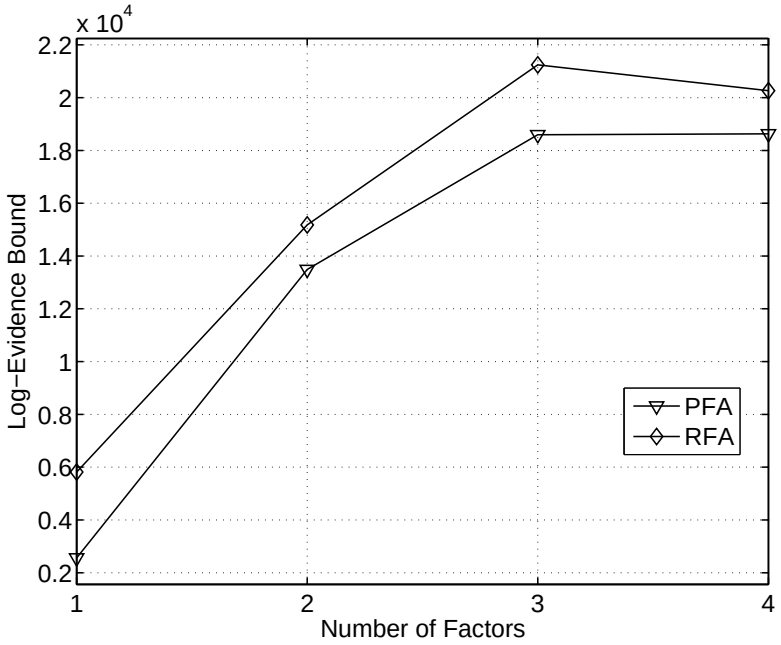


Fig. 5. Log-evidence bounds for PFA and RFA with different number of factors. Clearly, RFA achieves better values than PFA. The maximum for both models is at three factors, which is also the true model order.

where σ_s^2 is the variance of the true factor and σ_e^2 the variance of the error signal $e = s - \hat{s}$, $\hat{s}$ being the estimated factor. The corresponding SNR values are shown in the title lines of the plots. It is the third, non-sparse factor that creates problems for PFA, the SNR being as low as 6.8 dB. The poor estimation of the third factor also affects the estimation of the first factor. With RFA, no such problems occur.

It is also interesting to know whether we could have told that RFA is superior to PFA at modelling this data, just from comparing the model evidences. Figure 5 shows the evidence bounds of the two models with the number of factors varying between one and four. Indeed, the evidence comparison is

clearly in favour of RFA. In addition, Figure 5 shows that both models are able to infer the model order correctly.

4.1.2 Noise Structure

One of the advantages of factor analysis over standard dimensionality reduction methods such as principal component analysis is that the observation noise is not restricted to be isotropic i.e. it is allowed to have different variances for each dimension of the observation. The benefit of RFA over the conceptually simple and therefore popular NMF [21] algorithm stems straightforwardly from the ability of RFA to handle additive noise that is not isotropic.

To test this proposition, we generated 100 different datasets from the following model:

$$\begin{aligned} s_{jt} &\sim \mathcal{R}^N(0,1)\ , \quad j = 1,2 \quad t = 1,\ldots,250 \\ a_{ij} &\sim \mathcal{R}^N(0,1)\ , \quad i = 1,\ldots,10 \\ x_{it} &\sim \mathcal{N}\left(\mathbf{a}_i^T \mathbf{s}_t, 10^{-2\nu_i}\right) \end{aligned}$$

where diverse anisotropic noise was generated by using:

$$\nu_i \sim 0.7\,\delta(2) + 0.2\,\delta(1) + 0.1\,\delta(0)$$

The learning procedure with RFA was similar to that in the previous section. The scheme with NMF was the following. We initialised the factors and their loadings randomly with positive numbers and then iterated the multiplicative update rules of NMF for 1000 iterations. Further, from ten repeats of this procedure we have chosen the model with the best value of the objective function, in order to avoid getting trapped into local optima.

We measured the separation results as the SNR computed between (the appropriately permuted) estimated factors and the original ones. The average SNR values for RFA and NMF were 36.5 and 13.0, respectively. The whole sample is shown in Figure 6. These results clearly show that anisotropic noise poses serious difficulties for NMF, whereas RFA can handle it very well.

4.2 Autoregressive Factors

To demonstrate the auto-regressive (AR) extension of the proposed method, we report experiments with an artificial dataset where the factors have time-structure. A four-dimensional state sequence $\mathbf{r}_t$ was generated according to the first-order AR model below

$$\mathbf{r}_t = \mathbf{B}\mathbf{r}_{t-1}\,.$$

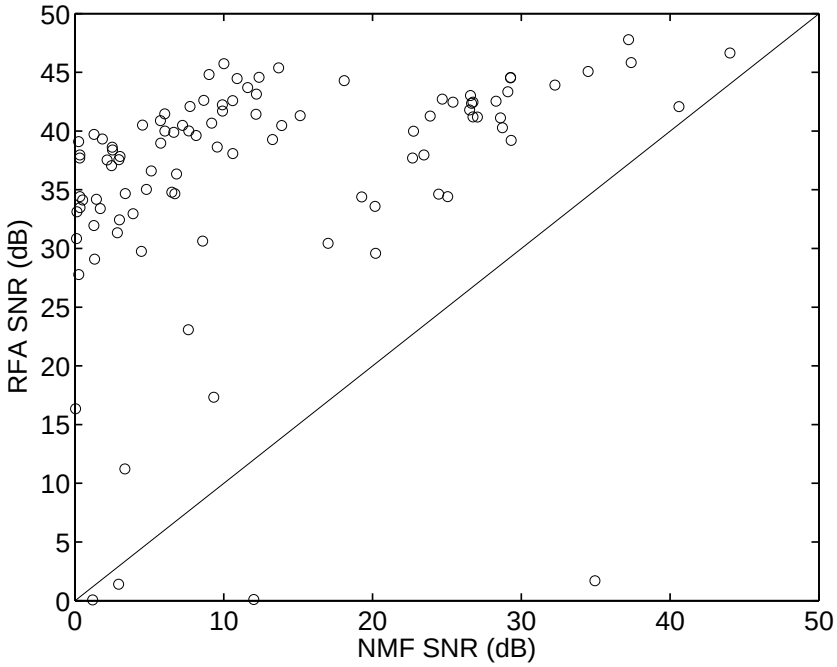


Fig. 6. Separation results for data with anisotropic additive noise. The SNR value obtained with NMF is plotted against that obtained with RFA. The diagonal line marks the points of equal performance. It is evident that RFA achieves better performance than NMF in the vast majority of cases.

The weights of the linear mapping $\mathbf{B}$ were sampled from $\mathcal{N}(0, 1)$ after which the matrix was orthogonalised to produce a well behaved dynamics. The state sequences were shifted by an amount of 0.5 to each direction to make the subsequent rectification preserve a little more of the original process.

Figure 7 shows the pairwise scatter-trajectories of the 4 underlying generator AR sequences $\mathbf{r}_t$. With a total number of 4 sources there are 6 possible pairwise combinations, and these are shown on the plots. The actual AR dynamics is traced with continuous lines. The factors $\mathbf{s}_t$ were obtained by rectification as shown in the figure by the dashed lines. The observations $\mathbf{x}_t$ were then generated in the same way as in the previous experiment, detailed in Section 4.1. Again, the std used for the additive noise was 0.01.

Estimates for the factors were obtained using all the three models: PFA, RFA and DRFA. The learning scheme was the same as in the previous section. Figure 8 shows the SNRs between the estimated and original factors $\mathbf{s}_t$. From the three models RFA and DRFA perform better than PFA and DRFA is slightly better than RFA. The true benefits of DRFA become obvious when examining the SNRs of the underlying state-space sequence $\mathbf{r}_t$, which are shown in Figure 9. Clearly, the DRFA model has been able to learn the dynamics and hence it predicts the state $\mathbf{r}_t$ also in the rectification region. This is confirmed by the Hinton diagrams of the original dynamical mapping versus the (appropriately

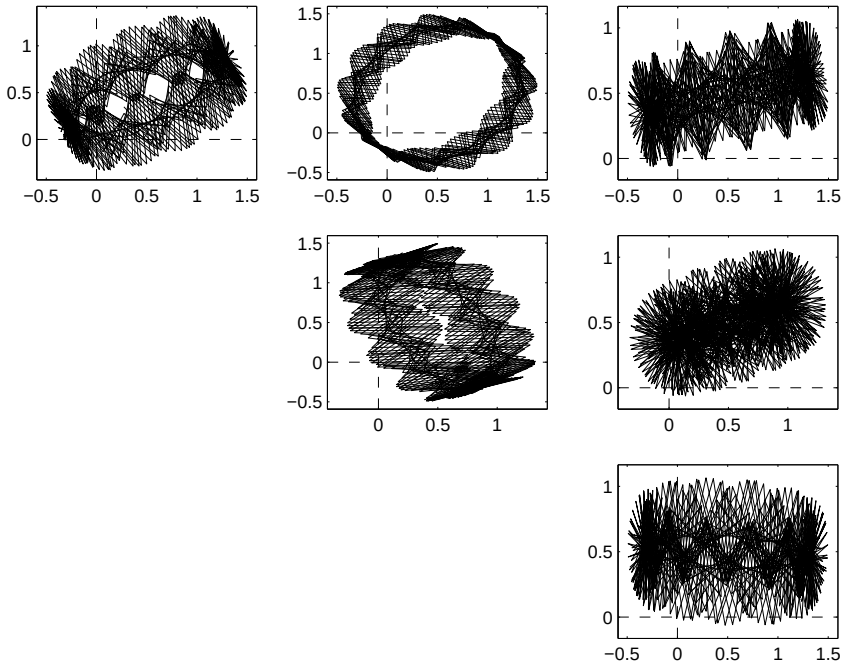


Fig. 7. Pairwise scatter-trajectory-plots of the state sequences following a first order AR model. The actual factors were rectified versions of these. The dashed lines show the rectification thresholds.

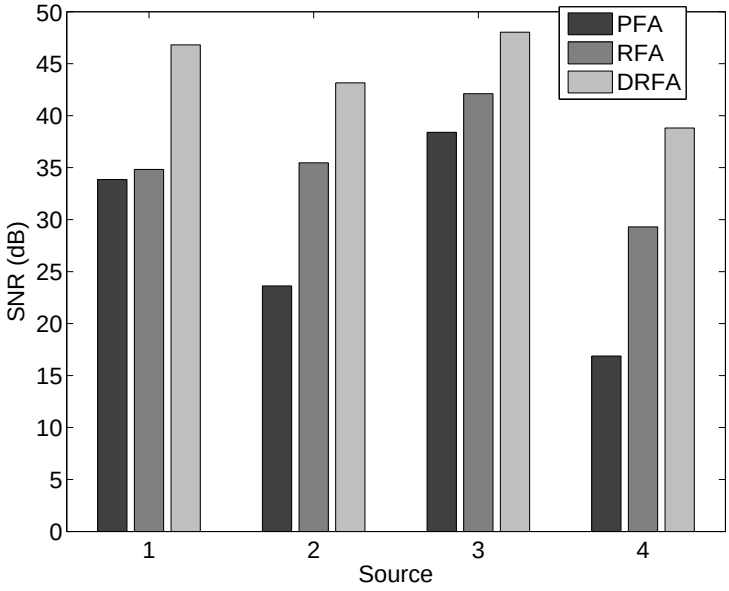


Fig. 8. SNRs of the four estimated factors $\mathbf{s}_t$ for the three models considered. PFA again has difficulties to separate the original factors whereas RFA and DRFA perform very well. DRFA is slightly superior to RFA.

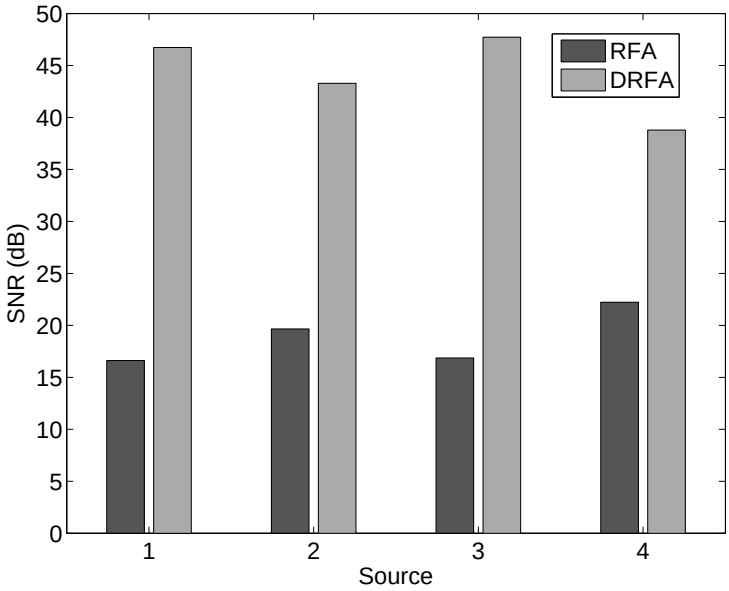
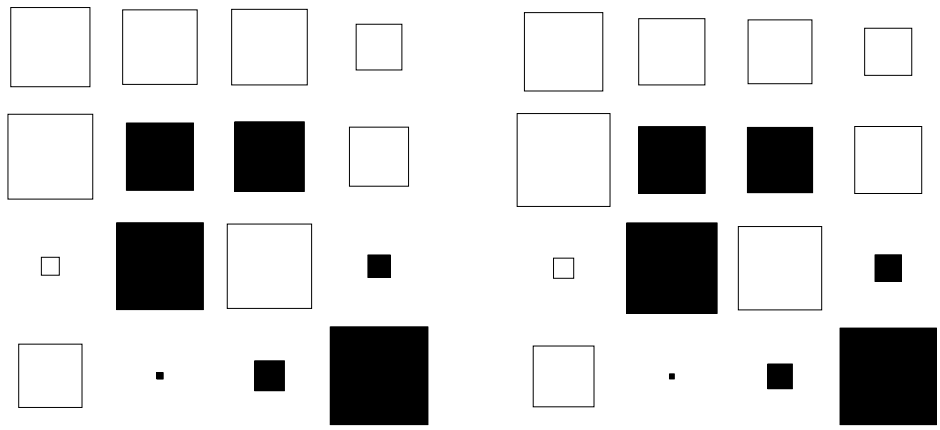


Fig. 9. SNRs of the estimated states $\mathbf{r}_t$. The rectification looses a lot of information from the original states so RFA, having no notion of time structure incorporated, cannot find the original states whereas DRFA provides very good estimates for them meaning that it has been able to learn the underlying dynamical model.



(a) Original (b) Estimated

Fig. 10. The original versus the estimated dynamics mappings $\mathbf{B}$. The similarity is evident.

permuted) estimated mapping which are shown in Figure 10.

In Section 5, the flexible modeling capabilities of RFA and DRFA will further be demonstrated in the context of an astrophysical application. It will also be seen that DRFA is desirable over RFA when it comes to predictive tasks.

5 An Astrophysical Application

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 well studied in physics. The hypothesis that some of these old galactic systems may actually contain young components is relatively new [22,23]. It is therefore of great practical interest 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 but physically meaningful manner. If so, that would allow the development of automated tools to aid more specialised physical analysis for research on galaxy formation and evolution. The positivity constraint is important in this modelling application, as negative values of flux would not be physically interpretable.

5.1 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, that it allows us to handle these in a principled manner. This is achieved simply by making the 'clean' vectors $\mathbf{x}_t$ become hidden variables of the additional error model below

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

Here e_{it} are zero-mean Gaussian noise terms with variances σ^2_{yit} fixed to values that are known from instrumental characteristics and uncertainty in calibration, for each individual measurement $i = 1, \ldots, N, t = 1, \ldots, T$. Handling missing values [24] can also be conveniently implemented in this framework by setting σ^2_{yit} to a large value when the actual measurement is missing.

5.2 Results on Real Data

A number of $N = 21$ real stellar population spectra will be analysed in this subsection. The data [25] 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

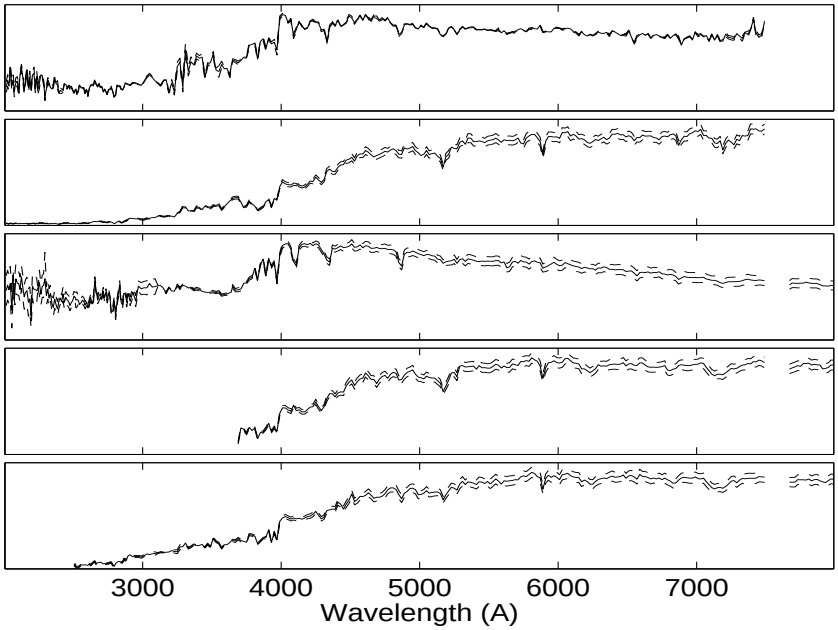


Fig. 11. A sample from the real data of spectral measurements. The dashed lines show the standard deviations of the errors in the data. The blank entries stand for missing values.

units [22,25]) given at a number of $T = 339$ different wavelength bins, ranging between 2005-8000 Ångströms. A part of this data set is shown in Figure 11.

In this section we demonstrate three models in terms of the interpretability of their factor representation created. We have fixed the number of factors to two, as inferring subsequent factors turns out to have no physical interpretation. Also the log-evidence bounds computed for different model orders (see Figure 12) support this decision. We repeated each run ten times with random initialisations drawn from $\mathcal{N}^R(0, 1)$. The model with highest log evidence was then selected. The two factors [6] for each of the models are shown in Figure 13. The shape of the first estimated factor is very similar for all three methods considered. This factor can visually be recognised to correspond to an old and high 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 [25].

Interestingly, the second factor inferred from the data differs more across the models considered. 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

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

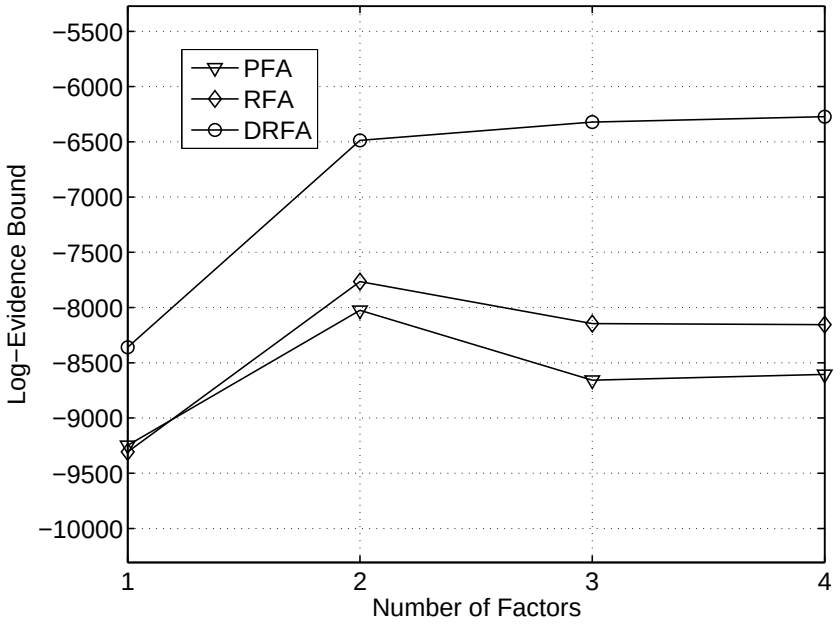


Fig. 12. The log-evidence bound with different models and model orders. With all three models, two factors seem to be the optimal choice although some improvement is gained with DRFA with additional ones.

encode metallicity characteristics of the stellar population.

From astrophysical point of view, the second PFA-component has no clear physical interpretation, as it is too noisy and 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 evidence bound that the variational procedure yields can also be interpreted as a kind of description length [26]. From this perspective it becomes interesting to examine the contribution of the different parts of the models to the overall coding length. These are visualised in Figure 14. The coding length for the data is approximately same for all of the models. The differences arise in the ability of the models to represent the factors. In this sense, RFA is somewhat better than PFA, due to its more flexible prior, and DRFA is even better since it is able to model the correlations between neighbouring wavelengths and hence code the factors most compactly.

5.3 Prediction Results on Synthetic Stellar Population Spectra

Here we employ synthetic spectra in order to assess the predictive performance of the proposed methods in an objective and controlled manner. A random

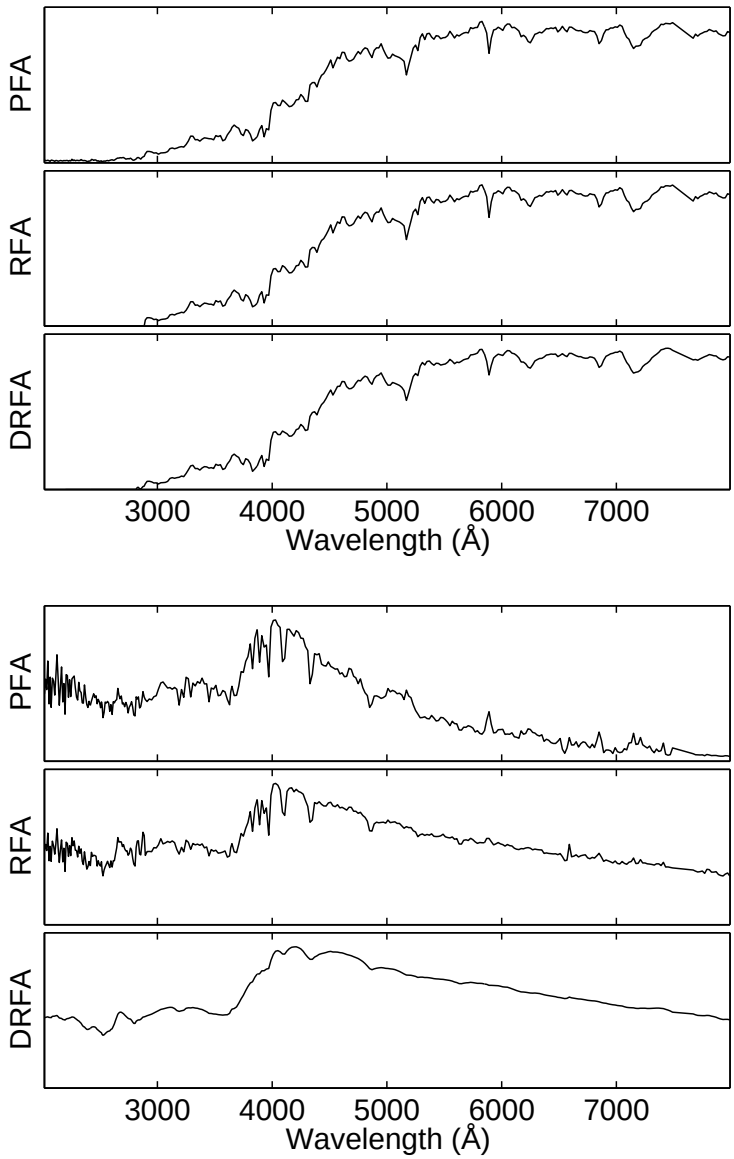


Fig. 13. The first (above) and second (below) factor estimated by the different models. The first factor is similar for all models. The second factor estimated with PFA is distorted towards zero and has no physical interpretation. The second factor estimated using RFA is in turn the most relevant from the physical interpretability point of view. DRFA smooths the second factor excessively.

selection of 100 synthetic composite spectra produced from the stellar population evolutionary synthesis model of Jimenez [25] is utilised. Each of 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 a controlled assessment.

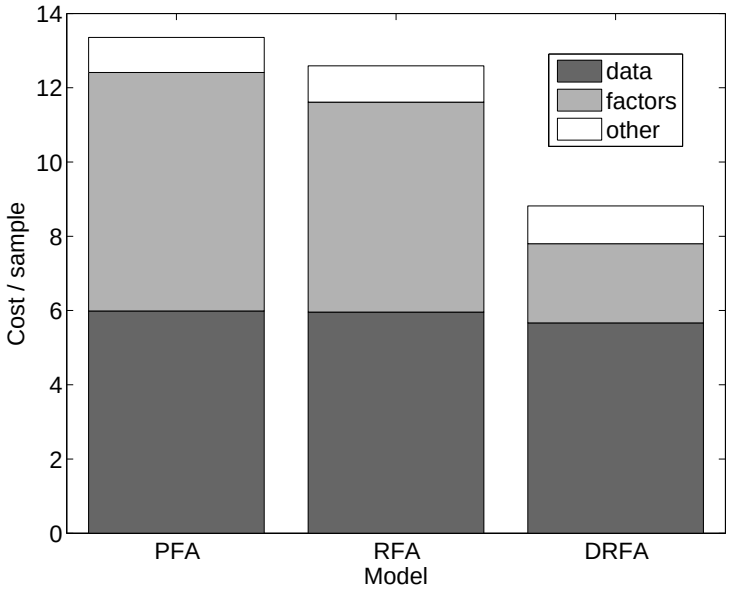


Fig. 14. The coding cost (divided by the number of samples) of the various models considered. Individual parts of the models are highlighted.

We consider an inference task where 50 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 and the percentage of them is varied. The RFA and DRFA models were trained on the same training set and asked to predict the artificially created missing entries in the previously unseen test set. The mean of the predictive distribution can be obtained simply as $\langle \mathbf{A}\mathbf{s}_t \rangle$. This scheme was repeated ten times. Finally, the mean and std of the SNR between the predictions and the true values, for each percentage of missing values in the test set, were computed. The results are shown in Figure 15. As expected, DRFA outperforms RFA in this prediction task when the amount of missing values gets large. The reason for the success of DRFA is that it includes the modelling of the correlations between fluxes at neighbouring wavelength bins. Clearly, this information is very useful when not too many observations are available. With moderate amounts of missing values, RFA is slightly better than DRFA, since no smoothing over wavelengths occurs in its inference.

6 Conclusions

We presented a method for non-negative 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 a learning algorithm for the model, using a factorial free-form posterior approximation. We demonstrated the proposed approach on

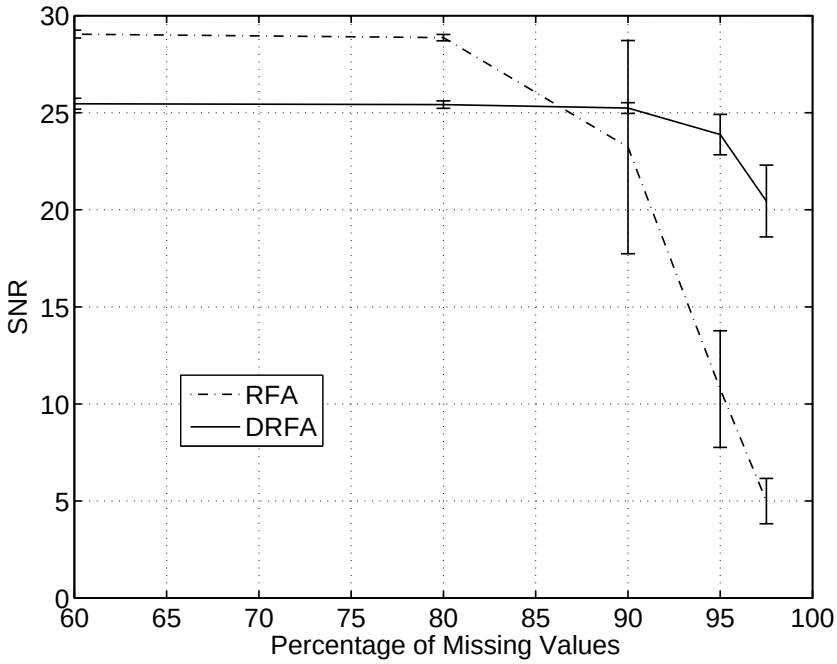


Fig. 15. Prediction of missing entries in out-of-sample wavelength bins with the number of factors being three. DRFA is able to use the information from neighbouring wavelengths and hence continues to perform well even with very high percentage of missing values.

both generated data and an actual application to astrophysical data analysis, for which the existing methods that induce sparse representations were found to be inappropriate. The utility of these results from the astrophysical perspective are detailed elsewhere [27]. The presented approach is applicable in any situation where flexible latent densities over the positive domain are required.

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

Other extensions may also be of interest: In RFA, we used independent priors, i.e. no dependencies between the sources are explicitly modelled. In particular, in the context of our astrophysical application, there is no physical interaction between the galaxy populations that we analyse, therefore the independent prior assumption has been justified. However, where a dependency model is desirable or available from the application domain, it is in principle possible to incorporate it into the prior using a graphical model, as in [28]. Investigating such extensions would constitute valuable further research.

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 from the School of Computer Science of The University of Birmingham, on 'Blind Separation of Information from Galaxy Spectra'. Many thanks to Louisa Nolan and Somak Raychaudhury for sharing their astrophysical expertise and supplying the data.

A Sufficient Statistics of the Rectified Gaussian and Gamma Distributions

The statistics of rectified Gaussian and Gamma distributions that are needed in the learning algorithm are given in this section.

The mean and the mean square suffice for rectified Gaussian distribution $\mathcal{N}^R(\theta|m,v)$:

$$\langle\theta\rangle = m + \sqrt{\frac{2v}{\pi}}\frac{1}{\exp\left((m/\sqrt{2v})^2\right)\operatorname{erfc}\left(-m/\sqrt{2v}\right)} \tag{A.1}$$

$$\left\langle\theta^2\right\rangle = m^2 + v + \sqrt{\frac{2v}{\pi}}\frac{m}{\exp\left((m/\sqrt{2v})^2\right)\operatorname{erfc}\left(-m/\sqrt{2v}\right)} \tag{A.2}$$

The variance is computed applying the familiar formula $\operatorname{var}(\theta) = \langle\theta^2\rangle - \langle\theta\rangle^2$.

From Gamma distribution $\mathcal{G}(\theta|\alpha,\beta)$ the mean and the mean of the logarithm are required:

$$\langle\theta\rangle = \alpha/\beta \tag{A.3}$$

$$\langle\log\theta\rangle = -\log\beta + \frac{\partial\log\Gamma(\alpha)}{\partial\alpha} \tag{A.4}$$

B Posterior Moments for the (D)RFA Model

Here, the expressions for the moments computed over the posterior approximation of the factors are presented. In deriving them, the properties of rectified

Gaussian distribution have been used.

$$M_p^0 = \frac{w_p}{2Z}\,\mathrm{erfc}[-\mu_p/\sqrt{2\sigma_p^2}] \tag{B.1}$$

$$M_p^1 = \frac{w_p}{2Z}\left\{\mathrm{erfc}[-\mu_p/\sqrt{2\sigma_p^2}]\mu_p + \sqrt{\frac{2\sigma_p^2}{\pi}}\frac{1}{\exp(\mu_p^2/2\sigma_p^2)}\right\} \tag{B.2}$$

$$M_p^2 = \frac{w_p}{2Z}\left\{\mathrm{erfc}[-\mu_p/\sqrt{2\sigma_p^2}](\mu_p^2+\sigma_p^2) + \sqrt{\frac{2\sigma_p^2}{\pi}}\frac{\mu_p}{\exp(\mu_p^2/2\sigma_p^2)}\right\} \tag{B.3}$$

$$M_n^0 = \frac{w_n}{2Z}\,\mathrm{erfc}[\mu_n/\sqrt{2\sigma_n^2}] \tag{B.4}$$

$$M_n^1 = \frac{w_n}{2Z}\left\{\mathrm{erfc}[\mu_n/\sqrt{2\sigma_n^2}]\mu_n - \sqrt{\frac{2\sigma_n^2}{\pi}}\frac{1}{\exp(\mu_n^2/2\sigma_n^2)}\right\} \tag{B.5}$$

$$M_n^2 = \frac{w_n}{2Z}\left\{\mathrm{erfc}[\mu_n/\sqrt{2\sigma_n^2}](\mu_n^2+\sigma_n^2) - \sqrt{\frac{2\sigma_n^2}{\pi}}\frac{\mu_n}{\exp(\mu_n^2/2\sigma_n^2)}\right\} \tag{B.6}$$

C Update Rules for RFA

In this section, the update rules for RFA are given. For each variable, first the form of the approximation is shown and then the expressions for its parameters are listed.

$$q(a_{ij}) = \mathcal{N}^R\left(a_{ij}|\mu,\sigma^2\right) \tag{C.1}$$

$$\sigma^2 = \left(1 + \langle\tau_{xi}\rangle\sum_{t=1}^{T}\left\langle \mathrm{cut}^2(r_{jt})\right\rangle\right)^{-1} \tag{C.2}$$

$$\mu = \sigma^2\,\langle\tau_{xi}\rangle\sum_{t=1}^{T}\left(\langle x_{it}\rangle - \sum_{k\neq j}\langle a_{ik}\rangle\,\langle\mathrm{cut}(r_{jt})\rangle\right)\langle\mathrm{cut}(r_{jt})\rangle \tag{C.3}$$

$$q(r_{jt}) = \frac{1}{Z}\Big[w_p \mathcal{N}\left(r_{jt}|\mu_p, \sigma_p^2\right)\ u(r_{jt}) + w_n \mathcal{N}\left(r_{jt}|\mu_n, \sigma_n^2\right)\ u(-r_{jt})\Big] \tag{C.4}$$

$$\sigma_x^2 = \left(\sum_{i=1}^{N} \langle \tau_{xi} \rangle \left\langle a_{ij}^2 \right\rangle\right)^{-1} \tag{C.5}$$

$$\mu_x = \sigma_x^2 \left(\sum_{i=1}^{N} \langle \tau_{xi} \rangle \langle a_{ij} \rangle \left(\langle x_{it} \rangle - \sum_{k \neq j} \langle a_{ik} \rangle \langle \mathrm{cut}(r_{jt}) \rangle\right)\right) \tag{C.6}$$

$$\sigma_n^2 = \langle \tau_{rj} \rangle^{-1} \tag{C.7}$$

$$\mu_n = \langle m_{rj} \rangle \tag{C.8}$$

$$w_p = \mathcal{N}\left(\mu_x|\mu_n, \sigma_x^2 + \sigma_n^2\right) \tag{C.9}$$

$$w_n = \mathcal{N}\left(\mu_x|0, \sigma_x^2\right) \tag{C.10}$$

$$\sigma_p^2 = (\sigma_x^{-2} + \sigma_n^{-2})^{-1} \tag{C.11}$$

$$\mu_p = \sigma_p^2(\mu_x/\sigma_x^2 + \mu_n/\sigma_n^2) \tag{C.12}$$

$$Z = \frac{w_n}{2} \mathrm{erfc}\Big[\mu_n/\sqrt{2\sigma_n^2}\Big] + \frac{w_p}{2} \mathrm{erfc}\Big[-\mu_p/\sqrt{2\sigma_p^2}\Big] \tag{C.13}$$

$$q(\tau_{xi}) = \mathcal{G}\left(\tau_{xi}|\alpha, \beta\right) \tag{C.14}$$

$$\alpha = \frac{T}{2} + \alpha_x \tag{C.15}$$

$$\beta = \frac{1}{2} \sum_{t=1}^{T} \left\langle (x_{it} - \mathbf{a}_i^T \mathbf{cut}(\mathbf{r}_t) - c_i)^2 \right\rangle + \beta_x \tag{C.16}$$

$$q(\tau_{rj}) = \mathcal{G}\left(\tau_{rj}|\alpha, \beta\right) \tag{C.17}$$

$$\alpha = \frac{T}{2} + \alpha_r \tag{C.18}$$

$$\beta = \frac{1}{2} \sum_{t=1}^{T} \left\langle (r_{jt} - m_{rj})^2 \right\rangle + \beta_r \tag{C.19}$$

$$q(x_{it}) = \mathcal{N}\left(x_{it}|\mu, \sigma^2\right) \tag{C.20}$$

$$\sigma^2 = \left(\sigma_{yit}^{-2} + \langle \tau_{xi} \rangle\right)^{-1} \tag{C.21}$$

$$\mu = \sigma^2 \left(\sigma_{yit}^{-2}\, y_{it} + \langle \tau_{xi} \rangle \sum_{j=1}^{M} \langle a_{ij} \rangle \langle \mathrm{cut}(r_{jt}) \rangle\right) \tag{C.22}$$

$$q(m_{rj}) = \mathcal{N}\left(m_{rj}|\mu, \sigma^2\right) \tag{C.23}$$

$$\sigma^2 = \left(\sigma_{mr}^{-2} + T \langle \tau_{rj} \rangle\right)^{-1} \tag{C.24}$$

$$\mu = \sigma^2 \left(\langle \tau_{rj} \rangle \sum_{t=1}^{T} \langle r_{jt} \rangle\right) \tag{C.25}$$

D Update Rules for DRFA

The update rules for the dynamic extension of RFA are given in this section. The rules that do not differ from RFA are not repeated here.

$$q(r_{jt}) = \text{as with RFA, but with} \tag{D.1}$$

$$\sigma_n^2 = \left(\sum_{i=1}^{M} \langle \tau_{ri} \rangle \left\langle b_{ij}^2 \right\rangle + \langle \tau_{rj} \rangle\right)^{-1} \tag{D.2}$$

$$\mu_n = \sigma_n^2 \Big[\sum_{i=1}^{M} \langle \tau_{ri} \rangle \langle b_{ij} \rangle \left(\langle r_{it+1} \rangle - \sum_{k \neq j} \langle b_{ik} \rangle \langle r_{kt} \rangle - \langle c_i \rangle\right) \tag{D.3}$$

$$+ \langle \tau_{rj} \rangle \left\langle \mathbf{b}_j^T \mathbf{r}_{t-1} + c_j \right\rangle\Big] \tag{D.4}$$

The above holds for the cases $1 < t < T$. The boundaries $t = 1$ and $t = T$ are slight modifications of that. For $t = 1$ the quantities σ_n^2 and μ_n are

$$\sigma_n^2 = \left(\sum_{i=1}^{M} \langle \tau_{ri} \rangle \left\langle b_{ij}^2 \right\rangle + \sigma_{r1}^{-2}\right)^{-1} \tag{D.5}$$

$$\mu_n = \sigma_n^2 \Big[\sum_{i=1}^{M} \langle \tau_{ri} \rangle \langle b_{ij} \rangle \left(\langle r_{it+1} \rangle - \sum_{k \neq j} \langle b_{ik} \rangle \langle r_{kt} \rangle - \langle c_i \rangle\right)\Big] \tag{D.6}$$

and for $t = T$ they are

$$\sigma_n^2 = \langle \tau_{rj} \rangle^{-1} \tag{D.7}$$

$$\mu_n = \left\langle \mathbf{b}_j^T \mathbf{r}_{t-1} + c_j \right\rangle \tag{D.8}$$

$$q(b_{ij}) = \mathcal{N}\left(b_{ij} | \mu, \sigma^2\right) \tag{D.9}$$

$$\sigma^2 = \left(1 + \langle \tau_{ri} \rangle \sum_{t=2}^{T} \left\langle r_{jt-1}^2 \right\rangle\right)^{-1} \tag{D.10}$$

$$\mu = \sigma^2 \langle \tau_{ri} \rangle \sum_{t=2}^{T} \left(\langle r_{it} \rangle - \sum_{k \neq j} \langle b_{ik} \rangle \langle r_{kt-1} \rangle - \langle c_i \rangle\right) \langle r_{jt-1} \rangle \tag{D.11}$$

$$q(c_i) = \mathcal{N}\left(c_i | \mu, \sigma^2\right) \tag{D.12}$$

$$\sigma^2 = \left(\sigma_c^{-2} + (T-1) \langle \tau_{ri} \rangle\right)^{-1} \tag{D.13}$$

$$\mu = \sigma^2 \langle \tau_{ri} \rangle \sum_{t=2}^{T} \left(\langle r_{it} \rangle - \left\langle \mathbf{b}_i^T \mathbf{r}_{t-1} \right\rangle\right) \tag{D.14}$$

$$q(\tau_{ri}) = \mathcal{G}\left(\tau_{ri}|\alpha,\beta\right) \tag{D.15}$$

$$\alpha = \frac{T-1}{2} + \alpha_r \tag{D.16}$$

$$\beta = \frac{1}{2}\sum_{t=2}^{T}\left\langle (r_{it} - \mathbf{b}_i^T\mathbf{r}_{t-1} - c_i)^2\right\rangle + \beta_r \tag{D.17}$$

E Evaluating the Evidence Bound

Since a factorial posterior approximation is used, the evidence bound (3) factors correspondingly into terms of the form

$$\left\langle \log p(\theta_i|\,\mathrm{pa}\,\theta_i)\right\rangle_{q(\boldsymbol{\theta})}\ , \tag{E.1}$$

where $\mathrm{pa}\,\theta_i$ stands for the parents of θ_i in the model. Additionally, for latent variables there are terms of the form

$$\left\langle \log q(\theta_i)\right\rangle_{q(\theta_i)}\ . \tag{E.2}$$

The first type of terms (E.1) in the bound are called prior terms and the second (E.2) the approximation terms or P-terms and A-terms for short.

Since we have three types of variables in the models — Gaussian, Gamma and Rectified Gaussian — there are three different types of P-terms:

$$\left\langle \log\mathcal{N}\left(\theta|m,\tau^{-1}\right)\right\rangle = \frac{1}{2}\Big[-\log(2\pi) + \langle\log\tau\rangle - \langle\tau\rangle\left(\left\langle\theta^2\right\rangle - 2\langle\theta\rangle\langle m\rangle + \left\langle m^2\right\rangle\right)\Big] \tag{E.3}$$

$$\langle\log\mathcal{G}\left(\theta|\alpha,\beta\right)\rangle = \alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\langle\log\theta\rangle - \beta\langle\theta\rangle \tag{E.4}$$

$$\left\langle \log\mathcal{N}^R\left(\theta|0,\tau^{-1}\right)\right\rangle = \frac{1}{2}\Big[\log(2) - \log(2\pi) + \langle\log\tau\rangle - \langle\tau\rangle\left\langle\theta^2\right\rangle\Big] \tag{E.5}$$

As can be seen from the update rules in the previous sections, the factors $q(\theta_i)$ in the free-form approximation are of four different forms. Hence the evidence bound has four different A-terms. The computation of the A-term for the factors r_{jt} was detailed in Equations (8)-(10). The rest of the A-terms are given below:

$$\left\langle \log\mathcal{N}\left(\theta|\mu,\sigma^2\right)\right\rangle = -\frac{1}{2}\log\Big(2\pi\,\mathrm{e}\,\sigma^2\Big) \tag{E.6}$$

$$\langle\log\mathcal{G}\left(\theta|\alpha,\beta\right)\rangle = \text{coincides with (E.4)} \tag{E.7}$$

$$\left\langle \log\mathcal{N}^R\left(\theta|\mu,\sigma^2\right)\right\rangle = -\frac{1}{2\sigma^2}\Big(\mathrm{var}(\theta) + (\langle\theta\rangle - \mu)^2\Big) + \frac{1}{2}\log\frac{2}{\pi\sigma^2} - \log\mathrm{erfc}\Big(-\mu/\sqrt{2\sigma^2}\Big) \tag{E.8}$$

Note that in (E.8) $\mu \neq \langle\theta\rangle$ and $\mathrm{var}(\theta) \neq \sigma^2$.

References

[1] M. Harva, A. Kabán, A variational Bayesian method for rectified factor analysis, in: Proc. Int. Joint Conf. on Neural Networks (IJCNN'05), Montreal, Canada, 2005, pp. 185–190.

[2] R. L. Gorsuch, Factor Analysis, 2nd Edition, Lawrence Erlbaum Associates, Hillsdale, NJ, 1983.

[3] P. Paatero, U. Tapper, Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values, Environmetr. 5 (1994) 111–126.

[4] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.

[5] M. Plumbley, E. Oja, A "nonnegative PCA" algorithm for independent component analysis, IEEE Transactions on Neural Networks 15 (1) (2004) 66–76.

[6] J. Miskin, Ensemble learning for independent component analysis, Ph.D. thesis, University of Cambridge, UK (2000).

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

[8] P. Højen-Sørensen, O. Winther, L. K. Hansen, Mean-field approaches to independent component analysis, Neural Computation 14 (4) (2002) 889–918.

[9] B. J. Frey, G. E. Hinton, Variational learning in nonlinear Gaussian belief networks, Neural Computation 11 (1) (1999) 193–214.

[10] J. Winn, C. M. Bishop, Variational message passing, Journal of Machine Learning Research 6 (2005) 661–694.

[11] N. D. Socci, D. D. Lee, H. S. Seung, The rectified Gaussian distribution, in: Advances in Neural Information Processing Systems, Vol. 10, 1998, pp. 350–356.

[12] G. E. Hinton, Z. Ghahramani, Generative models for discovering sparse distributed representations, Philosophical Transactions of the Royal Society of London, B 352 (1997) 1177–1190.

[13] J. M. Bernardo, A. F. M. Smith, Bayesian Theory, J. Wiley, 2000.

[14] D. Charles, C. Fyfe, Modelling multiple cause structure using rectification constraints, Network: Computation in Neural Systems 9 (1998) 167–182.

[15] S. Choi, Sequential EM learning for subspace analysis, Pattern Recognition Letters 25 (14) (2004) 1559–1567.

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

[17] 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.

[18] H. Attias, A variational Bayesian framework for graphical models, in: S. Solla, T. Leen, K.-R. Müller (Eds.), Advances in Neural Information Processing Systems 12, MIT Press, Cambridge, MA, USA, 2000, pp. 209–215.

[19] B. Wang, D. M. Titterington, Convergence and asymptotic normality of variational bayesian approximations for exponential family models with missing values, in: Proc. 20th Conference on Uncertainty in Artificial Intelligence, Banff, Canada, 2004, pp. 577–584.

[20] M. J. Beal, Z. Ghahramani, Variational Bayesian learning of directed graphical models with hidden variables, Bayesian AnalysisTo appear.

[21] D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, in: T. Leen, T. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems 13, MIT Press, Cambridge, MA, USA, 2001, pp. 556–562.

[22] G. Kauffmann, S. White, B. Guideroni, The formation and evolution of galaxies within merging dark matter halos, Monthly Notices of the Royal Astronomical Society 264 (1993) 201–218.

[23] D. S. Madgwick, A. L. Coil, et al., The DEEP2 galaxy redshift survey: Spectral classification of galaxies at $z \sim 1$, The Astrophysical Journal 599 (2) (2003) 997–1005.

[24] Z. Ghahramani, M. Jordan, Learning from incomplete data, Tech. Rep. 108, Center for Biological & Computational Learning, Massachusetts Institute of Technology (1994).

[25] L. Nolan, The star formation history of elliptical galaxies, Ph.D. thesis, The University of Edinburgh, UK (2002).

[26] A. Honkela, H. Valpola, Variational learning and bits-back coding: an information-theoretic view to Bayesian learning, IEEE Transactions on Neural Networks 15 (4) (2004) 800–810.

[27] L. Nolan, M. Harva, A. Kabán, S. Raychaudhury, A data-driven Bayesian approach for finding young stellar populations in early-type galaxies from their UV-optical spectra, Monthly Notices of the Royal Astronomical Society 366 (1) (2006) 321–338.

[28] F. R. Bach, M. I. Jordan, Beyond independent components: trees and clusters, Journal of Machine Learning Research 4 (2003) 1205–1233.