International Workshop on Statistical-Mechanical Informatics 2009 (IW-SMI 2009) IOP Publishing
Journal of Physics: Conference Series **197** (2009) 012002 doi:10.1088/1742-6596/197/1/012002

Inference algorithms and learning theory for Bayesian sparse factor analysis

Magnus Rattray[1], Oliver Stegle[2,3], Kevin Sharp[1] and John Winn[4]

[1]School of Computer Science, University of Manchester, Manchester M13 9PL, UK
[2]Max-Planck-Institute for Biological Cybernetics, Tübingen, Germany
[3]Max-Planck-Institute for Developmental Biology, Tübingen, Germany
[4]Microsoft Research Cambridge, Roger Needham Building, Cambridge, CB3 0FB, UK

E-mail: magnus.rattray@manchester.ac.uk

Abstract. Bayesian sparse factor analysis has many applications; for example, it has been applied to the problem of inferring a sparse regulatory network from gene expression data. We describe a number of inference algorithms for Bayesian sparse factor analysis using a slab and spike mixture prior. These include well-established Markov chain Monte Carlo (MCMC) and variational Bayes (VB) algorithms as well as a novel hybrid of VB and Expectation Propagation (EP). For the case of a single latent factor we derive a theory for learning performance using the replica method. We compare the MCMC and VB/EP algorithm results with simulated data to the theoretical prediction. The results for MCMC agree closely with the theory as expected. Results for VB/EP are slightly sub-optimal but show that the new algorithm is effective for sparse inference. In large-scale problems MCMC is infeasible due to computational limitations and the VB/EP algorithm then provides a very useful computationally efficient alternative.

1. Introduction

Factor analysis is a classical statistical approach for discovering latent structure in high-dimensional data. Sparse variants of factor analysis have been applied to the problem of uncovering latent variables that influence gene expression through a sparse regulatory network [1, 2]. Bayesian approaches to sparse factor analysis use sparsity-inducing priors to infer sparse posterior distributions over the factor loading matrix [3].

In this paper we describe a number of algorithms for Bayesian inference in sparse factor analysis models. As well as describing well-established Markov chain Monte Carlo (MCMC) and variational Bayes (VB) algorithms we also describe a new message passing algorithm that can be considered a hybrid between VB and Expectation Propagation (VB/EP). We compare the performance of these algorithms to the theoretical performance predicted by a replica analysis for the single factor case with isotropic noise which corresponds to sparse probabilistic principal component analysis (PCA).

2. Bayesian sparse factor analysis

The basic factor analysis model for a data vector $\boldsymbol{y}$ is given by,

$$\boldsymbol{y} \,|\, \boldsymbol{W}, \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu}, \boldsymbol{\Psi}) \;, \qquad \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \;,$$

where $\boldsymbol{\Psi}$ is a diagonal noise covariance matrix. Such models are applied in many settings but in the case of a transcriptional regulation model we might interpret the data $\boldsymbol{y} = [y_j]$ to represent the logged expression level for genes $j = 1 \dots p$ while the latent variables (factors) $\boldsymbol{x} = [x_k]$ represents the levels of regulatory proteins $k = 1 \dots K$ such as transcription factors (TFs). The factor loading matrix $\boldsymbol{W} = [w_{jk}]$ then represents the matrix of regulatory interactions between TFs and genes. We expect this to be sparse in the sense that each gene should be regulated by few TFs. Protein concentration is difficult to measure in a high-throughput manner and TFs are often modified and regulated after transcription so that their expression level may be a poor proxy for the concentration of active protein in the nucleus. Therefore we do not consider them to be observed but instead we treat them as latent variables which have to be integrated out to derive the data likelihood,

$$\boldsymbol{y} \,|\, \boldsymbol{W} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi} + \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}) \ .$$

To simplify the discussion we will assume zero-mean data ($\boldsymbol{\mu} = \boldsymbol{0}$) and we will not explicitly discuss inference of the covariance matrix $\boldsymbol{\Psi}$ although it would typically be inferred along with $\boldsymbol{W}$ in the algorithms that we describe below.

To infer a sparse matrix $\boldsymbol{W}$ we impose the following sparsity-inducing prior on the matrix elements

$$p(\boldsymbol{W} \,|\, \boldsymbol{C}, \lambda) = \prod_{j=1}^{p} \prod_{k=1}^{K} (1 - C_{jk})\delta(w_{jk}) + C_{jk}\mathcal{N}(w_{jk} \,|\, 0, \lambda^{-1}) \ . \tag{1}$$

Here the hyper-parameter $\boldsymbol{C} = [C_{jk}]$ encodes prior knowledge about the probability that there is a regulatory link in the network. The hyper-parameter λ can be learned or more usually is set to a small and uninformative value.

Given a dataset $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots \boldsymbol{y}_n\}$, Bayesian inference can be used to determine the posterior distribution over the loading matrix,

$$p(\boldsymbol{W} \,|\, \boldsymbol{Y}, \boldsymbol{C}, \lambda) \propto p(\boldsymbol{Y} \,|\, \boldsymbol{W}) p(\boldsymbol{W} \,|\, \boldsymbol{C}, \lambda)$$

and hence infer the sparse regulatory network. However, the normalisation of this probability cannot be computed in closed form and approximate inference algorithms are therefore required.

2.1. Markov chain Monte Carlo (MCMC)

The traditional method for carrying out Bayesian inference is to use MCMC. A Markov chain is constructed under which the intractable distribution of interest is invariant. Once convergence is attained, the distribution is approximated by means of a finite set of samples of the states visited.

Gibbs sampling [4] is a variant where each variable is iteratively sampled from its distribution conditioned on the current values of all the others. To construct a Gibbs sampler for this model, a standard way of dealing with the posterior over $\boldsymbol{W}$ induced by the sparsity-inducing mixture prior, is to introduce a binary matrix of indicator variables $\boldsymbol{Z}$ so that:

$$w_{jk} \,|\, z_{jk} = 0 \sim \delta(w_{jk}) \ , \qquad w_{jk} \,|\, z_{jk} = 1 \sim \mathcal{N}(w_{jk} \,|\, 0, \lambda^{-1}) \ ,$$

with independent Bernoulli priors placed over the elements of this matrix:

$$p(\boldsymbol{Z} \,|\, \boldsymbol{C}) = \prod_{j=1}^{p} \prod_{k=1}^{K} (1 - C_{jk})^{(1-z_{jk})} C_{jk}^{z_{jk}} \ . \tag{2}$$

However, although allowing calculation of convenient forms for the conditional distributions of the z_{jk} and w_{jk}, a Gibbs sampler so constructed would mix poorly owing to the high correlation

of these variables. Possible refinements that avoid this impasse are either a *collapsed sampler* or a *soft spike and slab sampler*.

In a *collapsed sampler* the z_{jk} are sampled from their conditional distribution from which the w_{jk} have been marginalised. This may be viewed as a way of sampling from their joint distribution conditional on the current values of the other variables $p(\boldsymbol{Z}, \boldsymbol{W} \,|\, \boldsymbol{\Psi}, \boldsymbol{X}, \boldsymbol{Y})$ (where $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$) by first sampling the elements of $\boldsymbol{Z}$ from $p(\boldsymbol{Z} \,|\, \boldsymbol{\Psi}, \boldsymbol{X}, \boldsymbol{Y})$ followed by those of $\boldsymbol{W}$ from $p(\boldsymbol{W} \,|\, \boldsymbol{\Psi}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z},)$. Provided no other variables are sampled between these steps the posterior remains an invariant distribution of the markov chain. This idea was used by [1] who exploited the conditional independence of the $\boldsymbol{z}_j$ to sample these variables as binary vectors. However, when no constraint is placed on the maximum number of ones in such a vector, normalisation of the resulting multinomial distribution is a combinatorial problem. We avoid this in the same manner as [2] by sampling each z_{jk} independently. Despite this, however, each such step requires inversion of an $s \times s$ matrix where s is the number of ones in the current state of the vector $\boldsymbol{z}_j$

A *soft spike sampler* [5] is a relaxation of the sparsity-inducing mixture prior for the weights, w_{jk}, that approximates $\delta(w_{jk})$ with a narrow Gaussian. Intuitively, the idea is that 'small' values of the w_{jk} are inferred to be 0. The relative widths of the narrow and broader Gaussians determine a trade-off between accuracy and efficient mixing of the chains. Although an approximation, this allows for much cheaper sampling of the z_{jk} as their conditional distributions depend on only the single corresponding w_{jk}, so no costly matrix inversions are required. The sampling steps are essentially the same as for the collapsed sampler, the principal difference being only in the form of the conditional distributions of $\boldsymbol{w}_j$ and z_{jk}.

2.2. Mean-field variational Bayes (VB)

A typically faster alternative to MCMC are deterministic approximation algorithms. Variational Bayes (VB) [6] is a mean-field approximation that can be motivated from statistical physics. The basic idea is to approximate the true posterior $p(\mathbf{W}, \mathbf{X}, \mathbf{Z} \,|\, \mathbf{Y}, \mathbf{C})$ by a simpler, factorised distribution, $q(\mathbf{W}, \mathbf{X}, \mathbf{Z}) = \prod_{j=1}^{p} q(\mathbf{w}_j) \prod_{i=1}^{n} q(\mathbf{x}_i) \prod_{jk} q(z_{jk})$. Here, we explicitly represented the binary indicator variables z_{jk}, choosing between the two mixture components of the sparsity prior (recall equation 2).

Individual variational factors $q(\cdot)$ are updated iteratively. The corresponding update rules can be derived by minimising the VB KL-divergence

$$\mathrm{KL}_{\mathrm{VB}}[q \,||\, p] = \int_{\boldsymbol{\Theta}} q(\boldsymbol{\Theta}) \log \frac{q(\boldsymbol{\Theta})}{p(\boldsymbol{\Theta})} \, \mathrm{d}\boldsymbol{\Theta},$$

where $\boldsymbol{\Theta}$ denotes the set of all model parameters. Free form variational updates, without specifying the functional form of the approximate factors follow then as

$$q(\cdot) \propto \exp\{\langle \ln p(\mathbf{W}, \mathbf{X}, \mathbf{Y}, \mathbf{C} \rangle_{q^{\setminus}}\},$$

where $\langle \rangle_{q^{\setminus}}$ denotes the expectation value with respect to all approximate factors except for the one that is refined.

As an example, we discuss the update for a single weight vector $\mathbf{w}_j$, which follows as

$$\begin{aligned} q(\mathbf{w}_j) &\propto \exp\{ \langle \log p(\mathbf{Y}, \mathbf{C}, \mathbf{W}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\Psi}) \rangle_{q \setminus \mathbf{w}_j} \} \\ &\propto \exp\{ \langle \log p(\mathbf{y}_j \,|\, \mathbf{w}_j, \mathbf{X}, \boldsymbol{\Psi}) + \log p(\mathbf{w}_j \,|\, \mathbf{z}_j) \rangle_{q \setminus \mathbf{w}_j} \} \qquad (3) \\ &\propto \underbrace{\exp\{ \langle \log p(\mathbf{y}_j \,|\, \mathbf{w}_j, \mathbf{X}, \boldsymbol{\Psi}) \rangle_{q \setminus \mathbf{w}_j} \}}_{M_{\mathbf{W} \cdot \mathbf{X} \to \mathbf{w}_j}} \underbrace{\exp\{ \langle \log p(\mathbf{w}_j \,|\, \mathbf{z}_j) \rangle_{q \setminus \mathbf{w}_j} \}}_{M_{\mathbf{W} \,|\, \mathbf{Z} \to \mathbf{w}_j}}. \qquad (4) \end{aligned}$$

The resulting Gaussian overall approximate factor, $q(\mathbf{w}_j)$, can be written as a product of two unnormalised Gaussian terms. The first term represents the evidence coming from the data likelihood, and the second term can be identified with the contribution from the sparsity prior. It is instructive to interpret $M_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}$ as the message sent from the product factor $f_{\mathbf{W}\cdot\mathbf{X}}$ to the weights variables $\mathbf{w}_j$. The parameters of this Gaussian, $M_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j} \propto \mathcal{N}\left(\mathbf{w}_j \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}, \tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}\right)$, can be read off from equation (4).

$$\tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j} = \left(\left\langle \frac{1}{\Psi_{j,j}} \right\rangle \sum_{i=1}^{n} \left\langle \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right\rangle\right)^{-1} \tag{5}$$

$$\tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j} = \tilde{\mathbf{\Sigma}}_{\mathbf{w}_j} \left(\left\langle \frac{1}{\Psi_{j,j}} \right\rangle \sum_{i=1}^{n} \langle \mathbf{x}_i \rangle \langle \mathbf{y}_i \rangle\right). \tag{6}$$

Using the definition of this message, the variational update in equation (3) follows as

$$\begin{aligned} q(\mathbf{w}_j) \propto \exp\Bigg\{ &-\frac{1}{2}\left(\mathbf{w}_j - \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}\right)^T \tilde{\mathbf{\Sigma}}^{-1}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j} \left(\mathbf{w}_j - \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}\right) \\ &-\frac{1}{2}\mathbf{w}_j^{\mathrm{T}} \mathrm{diag}\left(\left\{\sum_{c=0}^{1} q(z_{jk}=c)\lambda_c\right\}_k\right)\mathbf{w}_j\Bigg\}. \end{aligned} \tag{7}$$

Update rules for the responsibilities, $q(z_{jk}=1) = \tilde{C}_{jk}$, can be obtained in the same vein using

$$\begin{aligned} \tilde{C}_{jk} &\propto C_{jk} \exp\left\{\left\langle \log \mathcal{N}\left(w_{jk} \,\middle|\, 0, \lambda_1^{-1}\right)\right\rangle_{q\setminus z_{jk}}\right\} \\ (1-\tilde{C}_{jk}) &\propto (1-C_{jk}) \exp\left\{\left\langle \log \mathcal{N}\left(w_{jk} \,\middle|\, 0, \lambda_0^{-1}\right)\right\rangle_{q\setminus z_{jk}}\right\}. \end{aligned} \tag{8}$$

Note that the mixture component corresponding to an inactive weight, $C_{jk} = 0$, is not a delta spike but has been relaxed to a Gaussian with small variance $\lambda_0^{-1} \ll \lambda_1^{-1}$.

2.3. VB/EP hybrid

The accuracy of the pure mean-field solution, treating indicators $\mathbf{Z}$ and weights $\mathbf{W}$ as factorised variables can be improved, by considering a hybrid algorithm. This algorithms combines the mean-field learning presented in the previous section with Expectation Propagation [7] (EP), an alternative variational approximation based on a KL divergence with swapped arguments

$$\mathrm{KL}_{\mathrm{EP}}[p \,||\, q] = \int_{\mathbf{\Theta}} p(\mathbf{\Theta}) \log \frac{p(\mathbf{\Theta})}{q(\mathbf{\Theta})} \,\mathrm{d}\mathbf{\Theta}.$$

Comparing VB and EP, there is no clear-cut answer as to which approximation is superior, although for a number of problems EP was shown to be more accurate [8, 9]. A drawback of EP is that it is more difficult to apply, can lead to improper messages, and for some models is not tractable at all. In fact, full EP inference in the considered sparse factor-analysis model is not feasible. For EP we need the moments of the product factor $f_{\mathbf{X}\cdot\mathbf{W}}$, which are not available in closed form. Note that for observed factor activations $\mathbf{X}$, the factor analyser reduces to sparse linear regression and inference with EP is possible [10].

The idea of the hybrid scheme is to solve the problem of obtaining a posterior for weights and indicators in EP, while keeping the remainder of the inference within the mean-field framework. The posterior distribution of weights and indicators given the incoming message $M_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}$ is

$$P\left(\mathbf{w}_j, \mathbf{z}_j \,|\, M_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}\right) \propto \mathcal{N}\left(\mathbf{w}_j \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}, \tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}\right) \prod_{k=1}^{K} p(w_{jk} \,|\, z_{jk}) p(z_{jk} \,|\, C_{jk}). \tag{9}$$

As for VB, we choose an approximate form

$$q(\mathbf{w}_j, \mathbf{z}_j) \;\; = s \cdot \mathcal{N}\left(\mathbf{w}_j \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}, \tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}\right) \prod_{k=1}^{K} q(w_{jk})q(z_{jk}), \tag{10}$$

where the factors $q(w_{jk})q(z_{jk})$ are meant to approximate $p(w_{jk} \,|\, z_{jk})p(z_{jk} \,|\, C_{jk})$. The explicit scale of the approximation, s, will be dropped in the following. Choosing factor distributions that match the VB approximation, q-distributions for weights are Gaussian, $q(w_{jk}) = \mathcal{N}\left(w_{jk} \,\middle|\, \tilde{\mu}_{w_{jk}}, \tilde{\sigma}^2_{w_{jk}}\right)$, and factors of indicators are Bernoulli distributed, $q(z_{jk}) = \text{Bernoulli}(z_{jk} \,|\, \tilde{C}_{jk})$. While the overall approximation in equation (10) is fully factorised over indicators z_{jk}, it is multivariate Gaussian in the weights $\mathbf{w}_j$. Writing out the product of the Gaussian prior and the individual Gaussian factors $q(w_{jk})$ yields

$$q(\mathbf{w}_j, \mathbf{z}_j) \propto \mathcal{N}\left(\mathbf{w}_j \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{w}_j}, \tilde{\mathbf{\Sigma}}_{\mathbf{w}_j}\right) \prod_{k=1}^{K} q(z_{jk}). \tag{11}$$

Defining $\tilde{\boldsymbol{\mu}} = \left(\tilde{\mu}_{w_{j1}}, \ldots, \tilde{\mu}_{w_{jK}}\right)$ and $\tilde{\mathbf{\Sigma}} = \text{diag}(1/\tilde{\sigma}^2_{w_{j1}}, \ldots, 1/\tilde{\sigma}^2_{w_{jK}})$, the covariance and the mean of this Gaussian follow as

$$\tilde{\mathbf{\Sigma}}_{\mathbf{w}_j} = \left(\tilde{\mathbf{\Sigma}}^{-1}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j} + \tilde{\mathbf{\Sigma}}\right)^{-1} \qquad \tilde{\mathbf{m}}_{\mathbf{w}_j} = \tilde{\mathbf{\Sigma}}_{\mathbf{w}_j}\left[\tilde{\mathbf{\Sigma}}^{-1}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}\tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j} + \tilde{\mathbf{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}\right]. \tag{12}$$

The idea of EP is to iteratively refine individual pairs of factors for indicators and weights, leaving all other factors fixed. To update the ith pair, $q(w_{ji})q(z_{ji})$, the local KL divergence is

$$\begin{aligned} \text{KL}\Bigg[\quad & \mathcal{N}\left(\mathbf{w}_j \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}, \tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}\right) \prod_{k\neq i} q(w_{jk})q(z_{jk}) \overbrace{p(w_{ji} \,|\, z_{ji})p(z_{ji})}^{\text{exact factor}} \Bigg\| \\ & \mathcal{N}\left(\mathbf{w}_j \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}, \tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}\right) \prod_{k\neq i} q(w_{jk})q(z_{jk}) \underbrace{q(w_{ji})q(z_{ji})}_{\text{approximation}} \Bigg]. \end{aligned} \tag{13}$$

As the arguments of the KL divergence differ only in that ith factor, all other dimensions are marginalised out. This motivates the definition of a cavity distribution

$$\begin{aligned} q_{\setminus i}(w_{ji}) \;\; &= \int_{\mathbf{w}_{j\setminus i}} \mathcal{N}\left(\mathbf{w}_j \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}, \tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_j}\right) \prod_{k\neq i} q(w_{jk}) \, \mathrm{d}\mathbf{w}_{j\setminus i} \\ &= \mathcal{N}\left(w_{ji} \,\middle|\, \tilde{\mu}_{\setminus i}, \tilde{\sigma}^2_{\setminus i}\right). \end{aligned} \tag{14}$$

The cavity distribution $q_{\setminus i}(w_{ji})$ can be calculated efficiently from the current full approximation (equation (11)), dividing out the contribution of the ith factor (see for example [11], chapter 3).

Using this definition, the KL-divergence in equation (13) can be expressed in a compact form

$$\text{KL}\left[q_{\setminus i}(w_{ji}) \overbrace{p(w_{ji} \,|\, z_{ji})p(z_{ji})}^{\text{exact factor}} \,\middle\|\, q_{\setminus i}(w_{ji}) \underbrace{q(w_{ji} \,|\, \tilde{\mu}_{w_{ji}}, \tilde{\sigma}^2_{w_{ji}})q(z_{ji} \,|\, \tilde{C}_{ji})}_{\text{approximation}}\right]. \tag{15}$$

Minimising equation (15) with respect to the parameters of the Gaussian factor $q(w_{ji})$ leads to moment-matching conditions [12]. The new parameters of the approximation $q(w_{ji})$ are set such

that the moments of both arguments of the KL divergence match. The task hence reduces to calculating a set of moments under the exact factor

$$\begin{aligned} F_C &= \int_{w_{ji}} q_{\setminus i}(w_{ji}) \sum_{c=\{0,1\}} p(w_{ji} \,|\, z_{ji}=c) p(z_{ji}=c) \,\mathrm{d}w_{ji} \\ F_\mu &= \frac{1}{F_C} \int_{w_{ji}} q_{\setminus i}(w_{ji}) \sum_{c=\{0,1\}} p(w_{ji} \,|\, z_{ji}=c) p(z_{ji}=c) w_{ji} \,\mathrm{d}w_{ji} \\ F_{\sigma^2} + F_\mu^2 &= \frac{1}{F_C} \int_{w_{ji}} q_{\setminus i}(w_{ji}) \sum_{c=\{0,1\}} p(w_{ji} \,|\, z_{ji}=c) p(z_{ji}=c) w_{ji}^2 \,\mathrm{d}w_{ji}. \end{aligned} \tag{16}$$

Analytic expressions for these moments can be derived considering the moment generating function [13].

In the same vein, optimisation of equation (15) with respect to $\tilde{C}_{ji}$ leads to updates of the posterior over the indicator variables

$$\begin{aligned} \tilde{C}_{ji} &\propto C_{ji} \int_{w_{ji}} q_{\setminus i}(w_{ji}) \mathcal{N}\left(w_{ji} \,\middle|\, 0, \lambda_1^{-1}\right) \,\mathrm{d}w_{ji} \\ (1-\tilde{C}_{ji}) &\propto (1-C_{ji}) \int_{w_{ji}} q_{\setminus i}(w_{ji}) \mathcal{N}\left(w_{ji} \,\middle|\, 0, \lambda_0^{-1}\right) \,\mathrm{d}w_{ji}. \end{aligned} \tag{17}$$

3. Replica theory for a single latent factor

The replica method from statistical mechanics can be used to derive the performance of learning for an idealized situation in which data are produced by a sparse factor analysis generative model. We consider the simplest case of a single latent factor and set the data covariance $\boldsymbol{\Psi} = \boldsymbol{I}$ which corresponds to the probabilistic PCA model [14],

$$\boldsymbol{y} = \boldsymbol{w}x + \boldsymbol{\epsilon} \;, \tag{18}$$

where $x \sim \mathcal{N}(0,1)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Integrating out the latent factor gives the data density under the model,

$$\boldsymbol{y} \,|\, \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I} + \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}) \;, \tag{19}$$

where the model parameters are now in a vector $\boldsymbol{w} = [w_j]$ since there is only a single latent factor. The log-likelihood for dataset $\boldsymbol{Y}$ is given by,

$$\begin{aligned} \ln p(\boldsymbol{Y} \,|\, \boldsymbol{w}) &= -n \ln \sqrt{(2\pi)^p |\boldsymbol{I} + \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}|} - \frac{1}{2} \sum_{i=1}^n \boldsymbol{y}_i^{\mathrm{T}} (\boldsymbol{I} + \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})^{-1} \boldsymbol{y}_i \\ &= -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \ln(1 + ||\boldsymbol{w}||^2) - \frac{1}{2} \sum_{i=1}^n \boldsymbol{y}_i^{\mathrm{T}} \boldsymbol{y}_i + \frac{1}{2} \sum_{i=1}^n \frac{(\boldsymbol{y}_i^{\mathrm{T}} \boldsymbol{w})^2}{1 + ||\boldsymbol{w}||^2} \;. \end{aligned}$$

We use an equal sparsity hyper-parameter $C_j = C$ for each parameter vector component $j = 1, \ldots, p$ in the mixture prior $p(\boldsymbol{w} \,|\, C, \lambda)$ (recall equation (1)).

The marginal likelihood $p(\boldsymbol{Y} \,|\, C, \lambda)$ obtained by integrating out the model parameters $\boldsymbol{w}$ is analogous to a partition function in Statistical Mechanics. The parameter-dependent terms in the log-likelihood can be written,

$$E(\boldsymbol{w}) = \frac{n}{2} \ln(1 + ||\boldsymbol{w}||^2) - \frac{1}{2} \sum_{i=1}^n \frac{(\boldsymbol{y}_i^{\mathrm{T}} \boldsymbol{w})^2}{1 + ||\boldsymbol{w}||^2} \;. \tag{20}$$

Then $p(\boldsymbol{Y} \,|\, C, \lambda) \propto Z$ which is defined,

$$Z = \int \exp\left(-E(\boldsymbol{w})\right) p(\boldsymbol{w} \,|\, C, \lambda) \, \mathrm{d}\boldsymbol{w} \ .$$

To study learning performance we assume that the data are produced by a similar "teacher" distribution to the model in equation (19). The teacher parameters $\boldsymbol{w}^t$ are also generated from a similar prior $p(\boldsymbol{w}^t \,|\, C_t, \lambda_t)$ except that the hyper-parameters λ_t and C_t may differ from those of the model.

The replica calculation has similarities to previous work on non-sparse PCA [15], diluted neural networks [16] and a sparse Bayesian classifier [17]. The replica method makes use of the identity

$$\langle \ln Z \rangle_{\boldsymbol{Y}, \boldsymbol{w}^t} = \lim_{m \to 0} \frac{\partial \ln \langle Z^m \rangle_{\boldsymbol{Y}, \boldsymbol{w}^t}}{\partial m} \ . \tag{21}$$

The left-hand side shows the average we wish to compute but the calculation is intractable. However, we can compute the average over the right-hand side for integer m, then make an analytical continuation to real m and take the limit (see e.g. [18]). Before that we also take the limit $p \to \infty$ in order to use the saddle point method. We find,

$$p^{-1} \langle \ln Z \rangle_{\boldsymbol{Y}, \boldsymbol{w}^t} = \operatorname*{Ext}_{r,k,q,\hat{r},\hat{k},\hat{q}} \alpha G_0(r,q,k) - r\hat{r} + k\hat{k} - \frac{1}{2} q\hat{q} + G_1(\hat{r}, \hat{q}, \hat{k}) \tag{22}$$

where,

$$G_0(r,q,k) = \frac{1}{2}\left(\frac{q + r^2}{1+q} - \ln(1+q) \right) \ , \tag{23}$$

$$G_1(\hat{r}, \hat{q}, \hat{k}) = \left\langle \int_{-\infty}^{\infty} \frac{\mathrm{d}\eta}{\sqrt{2\pi}} \mathrm{e}^{\frac{-\eta^2}{2}} \ \ln \langle \mathrm{e}^{p\hat{r} w w^t - p(\hat{k} + \frac{1}{2}\hat{q}) w^2 + \eta \sqrt{p\hat{q}} w} \rangle_{w|\theta} \right\rangle_{w^t} \ . \tag{24}$$

The angled brackets denote averages with respect to individual components of the vectors $\boldsymbol{w}$ and $\boldsymbol{w}^t$. Notice that although we have taken the limit $p \to \infty$ it still appears in the terms multiplied by w or w^t. This is because individual components of the parameter vectors should be scaled to be $O(1/\sqrt{p})$ so that the length of these vectors remains $O(1)$ as $p \to \infty$. This is achieved by choosing an appropriate scale for the hyper-parameters.

The order parameters r, q and k obtained by solving the saddle point equations represent the following quantities,

$$q = ||\langle \boldsymbol{w} \rangle_{\mathrm{post}}||^2 = ||\boldsymbol{w}^{\mathrm{PM}}||^2 \ , \quad k = \langle ||\boldsymbol{w}||^2 \rangle_{\mathrm{post}} \ , \quad r = \langle \boldsymbol{w} \cdot \boldsymbol{w}^t \rangle_{\mathrm{post}} = \boldsymbol{w}^{\mathrm{PM}} \cdot \boldsymbol{w}^t \ , \tag{25}$$

where averages are over the posterior distribution and $\boldsymbol{w}^{\mathrm{PM}}$ denotes the posterior mean (PM) parameter. These order parameters can be used to assess learning performance with respect to the underlying data generating process.

A useful measure of performance is the cosine-angle, which we will refer to as the overlap, between the parameter estimate and the true parameter. For the posterior mean parameter estimate this can be written in terms of the order parameters as,

$$\rho^{\mathrm{PM}} = \frac{r}{\sqrt{Tq}}$$

where $T = \langle ||\boldsymbol{w}^t||^2 \rangle = C_t / \lambda_t$.

Another relevant measure of performance is the mean log-probability of a test data point under the model with the posterior mean parameter. This can also be written in terms of the order parameters,

$$\mathcal{L}^{\rm PM} = \langle \ln p(\boldsymbol{y}^{\rm test} \,|\, \boldsymbol{w}^{\rm PM}) \rangle_{\boldsymbol{y}^{\rm test}} = \frac{1}{2}\left(\frac{q+r^2}{1+q} - \ln(1+q) \right) + \kappa \;, \tag{26}$$

where $\kappa = -p\ln(2\pi)/2 - (1+T)/2$ is a constant term that does not depend on the model parameters.

For Bayesian learning a better prediction of test data may be obtained by averaging over the posterior distribution. In this case we have to average the log of the predictive distribution over test data,

$$\mathcal{L}^{\rm bayes} = \langle \ln \langle p(\boldsymbol{y}^{\rm test} \,|\, \boldsymbol{w}) \rangle_{\rm post} \rangle_{\boldsymbol{y}^{\rm test}} = \left\langle \ln \left\langle \frac{1}{\sqrt{1+||\boldsymbol{w}||^2}} \exp\left(\frac{(\boldsymbol{y}^{\rm T}\boldsymbol{w})^2}{2(1+||\boldsymbol{w}||^2)} \right) \right\rangle_{\rm post} \right\rangle_{\boldsymbol{y}|\boldsymbol{w}^t} + \kappa \;. \tag{27}$$

For large p we assume that the central limit theorem can be applied to the sum $\boldsymbol{y}^{\rm T}\boldsymbol{w} = \sum_j y_j w_j$ in which case it has a Gaussian distribution for large p with mean and variance,

$$\begin{aligned} E_{\rm post}[\boldsymbol{y}^{\rm T}\boldsymbol{w}] &= rx + \boldsymbol{\epsilon}^{\rm T}\boldsymbol{w}^{\rm PM} \;, \\ {\rm Var}_{\rm post}[\boldsymbol{y}^{\rm T}\boldsymbol{w}] &= \langle \boldsymbol{w}^{\rm T}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\rm T}\boldsymbol{w} \rangle_{\rm post} - \langle \boldsymbol{w}^{\rm T} \rangle_{\rm post} \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\rm T} \langle \boldsymbol{w} \rangle_{\rm post} + O(1/p) \\ &= k - q + O(1/p) \;, \end{aligned}$$

where x and $\boldsymbol{\epsilon}$ are defined after (18). The variance is self-averaging for large p and this allows us to take the average with respect to the posterior distribution. We write $\boldsymbol{y}^{\rm T}\boldsymbol{w} = rx + \boldsymbol{\epsilon}^{\rm T}\boldsymbol{w}^{\rm PM} + \nu\sqrt{k-q}$ with $\nu \sim \mathcal{N}(0,1)$. The other terms in the log-likelihood only involve $||\boldsymbol{w}||^2$, which is also self-averaging for large p, and we therefore obtain,

$$\begin{aligned} \mathcal{L}^{\rm bayes} &= \left\langle \ln \left\langle \frac{1}{\sqrt{1+k}} \exp\left(\frac{(rx + \boldsymbol{\epsilon}^{\rm T}\boldsymbol{w}^{\rm PM} + \nu\sqrt{k-q})^2}{2(1+k)} \right) \right\rangle_{\nu} \right\rangle_{z,\boldsymbol{\epsilon}} + \kappa \\ &= \frac{1}{2}\left(\frac{r^2+q}{1+q} - \ln(1+q) \right) + \kappa = \mathcal{L}^{\rm PM} \quad \text{as } p \to \infty \;. \end{aligned} \tag{28}$$

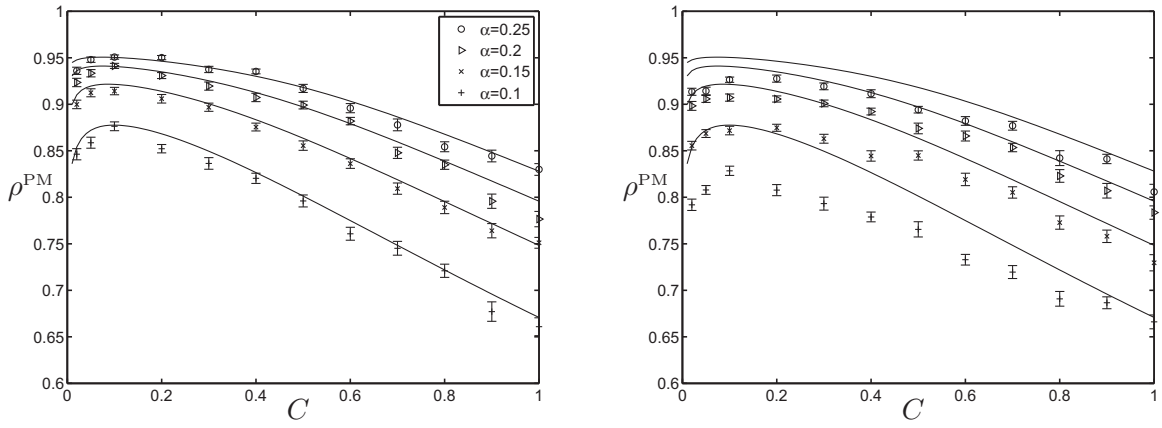
So we see that the performance of full Bayesian inference is equivalent to using the posterior mean parameter in the large p limit considered here.

4. Simulation results

In figure 1 we compare the replica theory with results from MCMC and the VB/EP algorithm applied to simulated data. MCMC can be considered a gold standard given sufficient computation time and we see that the results agree closely with the theoretical performance. As expected, optimal performance is achieved when the sparsity of the model matches the sparsity in the data ($C = C_t = 0.1$). The results for VB/EP show lower accuracy but still demonstrate that sparse inference provides a significant benefit over non-sparse PCA ($C = 1$ is the non-sparse result). Also, we observe that the optimal performance of VB/EP is obtained when the model and data sparsity are matched. Similar results are obtained for the predictive likelihood.

5. Conclusion

We have presented a number of MCMC and deterministic inference algorithms for Bayesian sparse factor analysis, including a novel VB/EP hybrid algorithm. We compared the empirical



(a) MCMC versus theory (b) VB/EP versus theory

Figure 1. The theory for sparse PCA is compared to (a) MCMC results and (b) VB/EP results for data with 90% sparsity ($C_t = 0.1$). We show the overlap ρ^{PM} between the true underlying vector and the posterior mean parameter as a function of the sparsity hyper-parameter C with $C = 1$ corresponding to standard non-sparse PCA. Results are shown for $n = 100$ with different values of $\alpha = n/p$ and $\lambda = \lambda_t = 0.01$. For the MCMC results we generated 500 Gibbs samples after discarding 500 as burn-in. Results are averaged over 50 replicate datasets.

performance of MCMC and the VB/EP algorithm with results from a replica theory of learning performance. The MCMC results confirmed the accuracy of the theory and results for the VB/EP algorithm are encouraging. Our ongoing work shows that this algorithm can provide very significant gains over MCMC in large-scale applications where MCMC has no hope of converging in a reasonable time.

Acknowledgments

We thank Michalis Titsias for useful discussions. OS gratefully acknowledges financial support from the Cambridge Gates Trust. MR thanks Martin Weigt and Andrea Pagnani for hosting visits to ISI Turin where the theoretical work was initiated.

References

[1] Sabatti C and James G M 2006 *Bioinformatics* **22** 739–46
[2] Pournara I and Wernisch L 2007 *BMC Bioinformatics* **8** 61
[3] West M 2003 *Bayesian Statistics* vol 7 (Oxford University Press) pp 723–32
[4] Geman S and Geman D 1984 *IEEE Trans. Pattern. Anal.* **6** 721741
[5] George E and McCulloch R 1993 *J. Am. Stat. Assoc.* **88** 881–9
[6] Jordan M, Ghahramani Z, Jaakkola T S and Saul L K 1999 *Mach. Learn.* **37** 183–233
[7] Minka T P 2001 *UAI '01: Proc. of the 17th Conf. in Uncertainty in Artificial Intelligence* vol 17 (San Fransisco: Morgan Kaufmann Publishers) pp 362–9
[8] Nickisch H and Rasmussen C E 2008 *J. Mach. Learn. Res.* **9** 2035–78
[9] Frey B J, Patrascu R, Jaakkola T and Moran J 2000 *Advances in Neural Information Processing Systems* vol 12 pp 493–9
[10] Seeger M W 2008 *J. Mach. Learn. Res.* **9** 759–813
[11] Rasmussen C E and Williams C K I 2006 *Gaussian Processes for Machine Learning* (MIT Press)
[12] Minka T P 2001 *A family of algorithms for approximate Bayesian inference* Ph.D. thesis Massachusetts Institute of Technology
[13] DeGroot M and Schervish M 2002 *Probability and Statistics* (Addison-Wesley Boston)
[14] Tipping M E and Bishop C M 1999 *J. R. Stat. Soc.* B **61** 611–22
[15] Reimann P, Van den Broeck C and Bex G J 1996 *J. Phys. A: Math. Gen.* **29** 3521–35

[16] Kuhlmann P and Müller K R 1994 *J. Phys. A: Math. Gen.* **27** 3759–74
[17] Uda S and Kabashima Y 2005 *J. Phys. Soc. Japan* **74** 2233–42
[18] Engel A and Van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge University Press)