Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

A Comparison of Inference in Sparse Factor Analysis

Oliver Stegle[⋆] OLIVER.STEGLE@TUEBINGEN.MPG.DE
Max Planck Institutes Tübingen
Spemannstrasse 38, 72076 Tübingen, 72076, Germany

Kevin Sharp[⋆] KEVINSHARP@BTINTERNET.COM
School of Computer Science
University of Manchester
Manchester M13 9PL, UK

Kevin Sharp KEVINSHARP@BTINTERNET.COM
School of Computer Science
University of Manchester

John Winn JWINN@MICROSOFT.COM
Microsoft Research
Cambridge, UK

[⋆] *These authors contributed equally*

Editor:

Abstract

Comparison of inference techniques on models with wide-ranging applications is important and instructive. In this paper we investigate approximate inference algorithms applied to the sparse factor analyser where rich prior information about the sparsity pattern is available. Sparse factor analysis is used in diverse areas including vision, signal analysis and computational biology. We explore four different approximate Bayesian inference methods, based on sampling techniques and deterministic approximations. These include a novel hybrid Expectation Propagation/Variational algorithm, which achieves encouraging results particularly when taking the trade-off between accuracy and computational efficiency into account.

Keywords: Approximate inference, sparse factor analysis, MCMC

1. Introduction

Factor analysis is a general purpose technique for dimension reduction that linearly maps high dimensional data samples onto points in a lower dimensional subspace. Variations of factor analysis and mixtures of factor analysers have been extensively applied in a number of important application domains. For example, in computer vision linear dimension reduction techniques are used to model facial expressions (Turk and Pentland, 1991). In collaborative filtering, factor analysis-type models are widely used to explain the similarity of user tastes or related items (Srebro and Jaakkola, 2003). In computational biology, factor analysis has been used to understand the variation of high-dimensional gene expression profiles

across individuals or samples by mapping them onto lower dimensional transcription factor activations (Liao et al., 2003; Sabatti and James, 2006; Pournara and Wernisch, 2007).

Sparse extensions of the factor analysis model are of considerable interest (West, 2003). In some cases sparsity is employed simply as a regularisation mechanism to prevent overfitting in overparameterised models, but in others it reflects a genuine belief in the underlying structure of the model. For example, in the context of the computational biology application mentioned above the use of a sparse model is motivated by a strong *a priori* belief about the connectivity structure of the network of regulatory relationships between transcription factors and genes. These *gene regulatory networks* are known to be sparsely connected (Hartemink, 2005) and, for model organisms such as yeast, empirically derived beliefs about their connectivity structure are available in public databases (e.g. Teixeira et al., 2006).

In a Bayesian context, a prior belief in sparsity is modelled by a sparsity inducing prior distribution on the elements of the mixing matrix. So called *zero-norm* priors assign finite probability mass to sparse solutions and Markov chain Monte Carlo (MCMC) techniques are typically used to solve the resulting intractable inference problem (Mitchell and Beauchamp, 1988; West, 2003; Carvalho et al., 2008, 2009). An alternative approach is to use so called *shrinkage priors* which are continuous heavy-tailed densities which favour sparse solutions. The use of shrinkage priors is more closely related to non-Bayesian sparse estimation techniques. The canonical example is the Laplace distribution which leads to L1 or LASSO regularisation under Maximum a Posteriori (MAP) parameter estimation (Tibshirani, 1996; Williams, 1995). LASSO regularisation has been used for the closely related problem of sparse principal component analysis (Zou et al., 2006; Sigg and Buhmann, 2008). Shrinkage priors with heavier tailed densities have been applied using variational Bayesian inference in another closely related model (Archambeau and Bach, 2009).

Shrinkage priors offer considerable computational advantages over zero-norm priors because they transform an inference problem over discrete parameters into a continuous problem which is more easily addressed using standard deterministic approximate inference methods (Seeger, 2008; Archambeau and Bach, 2009). However, although Maximum a Posteriori (MAP) parameter estimates obtained with Shrinkage priors are sparse, samples from the posterior distribution will not be truly sparse. This is a significant drawback if one is interested in characterising the uncertainty about whether or not a parameter is exactly zero. This is especially problematic in applications such as the transcriptional regulatory network example which are believed to be genuinely sparse and where factors have a clear physical interpretation. Another problem with shrinkage priors is that it is not obvious how to incorporate specific prior knowledge about sparse structure when it is available.

In this work we focus on zero-norm priors which do assign finite probability mass to sparse solutions. These priors better characterise a prior belief in sparsity and should therefore lead to more meaningful posterior beliefs. A natural implementation of a zero-norm sparsity prior in this context is a spike and slab prior. This is a mixture prior on the entries of the mixing matrix, where one mixture component drives the weight to zero ("no link") while the other mixture component allows for non-zero entries ("link"). This prior, suggested by West (2003), not only assigns finite probability mass to truly sparse solutions, but also allows available information about the sparse structure to be included in a natural

and interpretable manner: prior probabilities over specific entries in the mixing matrix can be used to adjust the relative weights of the corresponding mixture components.

Unfortunately, it is challenging to perform Bayesian inference under this model, particularly in high dimensions. The likelihood contour is not log-concave and hence MAP approaches as well as mean-field variational Bayes approximations are prone to trapping in local optima. Similarly, MCMC methods tend to mix slowly due to the multimodal nature of the posterior distribution. Consequently, the spike and slab prior remains unpopular for the kind of large-scale inference problems seen in many of today's Machine Learning applications. Nevertheless, given its appealing properties it is important to explore and to characterise the comparative performance of different approaches to inference in the face of these challenges.

In this work we investigate and compare the performance of four alternative inference methods applied to the sparse factor analysis model incorporating a spike and slab mixture prior. Two of them are deterministic approximations. One is based on a standard variational mean-field approach; the other is a novel extension implementing parts of the inference using Expectation Propagation (EP) while keeping the remainder of the inference within the variational framework. These deterministic methods are contrasted with two alternative implementations of a Gibbs sampler. In addition to a refinement of the collapsed Gibbs sampler used by Sabatti and James (2006) we consider a sampler based on a softened slab and spike model. We evaluate all four algorithms in the context of a gene regulatory network inference problem. We highlight practical and theoretical challenges in this context and we demonstrate the utility of explicitly accounting for the label-switching problem that occurs under an informative connectivity prior because factors are no longer exchangeable.

The paper is organized as follows. In Section 2 we introduce the sparse factor analysis model. Section 3 addresses the principal challenges for inference in this model, in particular the need of explicitly addressing the label-switching problem. In Section 4 we describe the details of the two Gibbs samplers, followed by the two deterministic approximations which are developed in Section 5, including the novel hybrid approach. Experiments on simulated data (Section 6) highlight different efficiency-accuracy trade-offs depending on the size and difficulty of the problem. We conclude with a large-scale case study on a real world dataset from computational biology.

2. Sparse Factor Analysis

We introduce the sparse factor analysis model in the context of a gene regulatory network inference problem where specific prior information on the sparse network structure is available. Here, the goal is to explain an expression data matrix $\mathbf{Y}$, covering G gene expression levels for each of J individuals or samples. These high-dimensional expression profiles are linearly mapped into a lower-dimensional representation of K regulatory transcription factors with activations $\mathbf{X}$. The $G \times K$ mixing matrix $\mathbf{W}$ weights the contribution from factor activations to gene expression. The sparsity of this mixing matrix reflects the stucture of the regulatory network.

First, we start by reviewing standard factor analysis. The generative model can be cast as an inner product between weights $\mathbf{W}$ and factor activations $\mathbf{X}$. The expression profile

for a single individual j is then given by

$$\mathbf{y}_j = \mathbf{W} \cdot \mathbf{x}_j + \boldsymbol{\psi}_j, \tag{1}$$

where $\boldsymbol{\psi}_j$ represents observation noise. Assuming independent samples and Gaussian noise with precisions τ_g for each gene, the likelihood for the expression matrix follows as

$$P(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}, \boldsymbol{\tau}) = \prod_{j=1}^{J} \mathcal{N}\left(\mathbf{y}_j \,\middle|\, \mathbf{W} \cdot \mathbf{x}_j, \mathrm{diag}\{\tau_g^{-1}\}\right). \tag{2}$$

A natural approach to achieve sparsity in this model is a mixture prior on the entries of the mixing matrix (West, 2003), with one mixture component driving the weight to zero ("no link") and a second component allowing for non-zero entries ("link"). These two distinct meanings of the mixture components render this prior interpretable, unlike other sparsity priors in common usage, and allow the inclusion of additional information about the connectivity structure. Choosing a Gaussian prior for the active weights, a single weight $w_{g,k}$ is distributed as

$$P(w_{g,k}) = \pi_{g,k}\,\mathcal{N}\left(w_{g,k} \,\middle|\, 0, \sigma_1^2\right) + (1 - \pi_{g,k})\,\delta\left(w_{g,k}\right), \tag{3}$$

where the coefficient $\pi_{g,k}$ denotes the prior probability that factor k regulates gene g and $\delta(x)$ is the Dirac delta function, with non-zero density only for $x = 0$. In the absence of strong prior information about the scale of the $w_{g,k}$, the hyperparameter σ_1 can be set to some large value or it can be learned. This prior accurately models the idea that a relationship between a factor and a gene either exists or does not exist. It has been used in the context of both strong (Sabatti and James, 2006) and weak (West, 2003) prior information on each $\pi_{g,k}$ and is the exact model used in the collapsed Gibbs sampler described in Section 4.2.

A practical relaxation of this prior is to replace the delta function by a second Gaussian component,

$$P(w_{g,k}) = \pi_{g,k}\,\mathcal{N}\left(w_{g,k} \,\middle|\, 0, \sigma_1^2\right) + (1 - \pi_{g,k})\,\mathcal{N}\left(w_{g,k} \,\middle|\, 0, \sigma_0^2\right), \tag{4}$$

where $\sigma_0^2 \ll \sigma_1^2$ thereby forcing the corresponding weight to take near-zero values. Inference using this relaxed prior can be easier (Section 3), and for the limiting case $\sigma_0 \to 0$ this relaxation turns into the original form in Equation 3.

To incorporate prior information on the connectivity structure, we condition on an indicator variable $z_{g,k}$ that chooses between the two mixture components

$$\begin{aligned} P(w_{g,k} \mid z_{g,k} = 0) &= \mathcal{N}\left(w_{g,k} \,\middle|\, 0, \sigma_0^2\right) \\ P(w_{g,k} \mid z_{g,k} = 1) &= \mathcal{N}\left(w_{g,k} \,\middle|\, 0, \sigma_1^2\right). \end{aligned} \tag{5}$$

Existing binary knowledge about the regulatory network structure can then be encoded as a Bernoulli prior on the indicator variables $z_{g,k}$

$$\pi_{g,k} = P(z_{g,k} = 1) = \begin{cases} \eta_0 & \text{no link} \\ 1 - \eta_1 & \text{link} \end{cases}, \tag{6}$$

where η_0 can be identified as the false negative rate (FNR) and η_1 as the false positive rate (FPR) of the observed prior network structure. In the most general setting, the probability of a link $P(z_{g,k} = 1)$ can also be set individually for every link, reflecting the available prior knowledge.

The hyperparameter σ_1^2 can either be set to a fixed, large value or learned. For this inference we put a gamma prior on the inverse variance

$$P(\sigma_1^2) = \Gamma\left(\frac{1}{\sigma_1^2} \,|\, a_{\sigma_1}, b_{\sigma_1}\right). \tag{7}$$

The specification of prior probabilities for factor activations

$$P(\mathbf{X}) = \prod_{k=1}^{K} \prod_{j=1}^{J} \mathcal{N}\left(x_{k,j} \,|\, 0, 1\right) \tag{8}$$

and the noise precisions

$$P(\boldsymbol{\tau}) = \prod_{g=1}^{G} \Gamma(\tau_g \,|\, a_\tau, b_\tau) \tag{9}$$

complete the definition of the model. The corresponding graphical model representation of this sparse factor analyser is shown in Figure 1.

3. Inference challenges for the Sparse Factor Analysis model

The most significant challenge for inference posed by the posterior distribution of the sparse factor analysis model defined in the previous section is that it is highly *multimodal*. When these modes are not equivalent, inference is challenging: greedy deterministic methods become trapped in local optima and MCMC samplers suffering from slow mixing.

Here, multimodality arises from two sources. First, the sparsity-inducing mixture prior (Equation (3)), being itself multimodal, induces large numbers of non-equivalent modes in the posterior. The novel hybrid of VB and EP described in section 5.3, the collapsed Gibbs sampler (Section 4.2) and the relaxation of the sparsity inducing prior, equation (5), are all measures aimed at improving inference in the face of the multiple modes that arise in this manner. However, the problem is exacerbated by multimodality arising from the inherent symmetries of factor analysis models in general. This identifiability problem needs to be addressed separately.

3.1 Factor model symmetry

The inherent symmetry of factor analysis models arises from the fact that factorisation of the data matrix into weights and activations is not uniquely defined: the product, $\mathbf{W} \cdot \mathbf{X}$, is invariant to simultaneous orthogonal transformations of $\mathbf{W}$ and $\mathbf{X}$ by an arbitrary non-singular matrix $\mathbf{R}$ and its inverse:

$$\mathbf{Y} = \mathbf{W}\left(\mathbf{R}\mathbf{R}^T\right)\mathbf{X} = \left(\mathbf{W}\mathbf{R}\right)\left(\mathbf{R}^T\mathbf{X}\right) = \tilde{\mathbf{W}}\tilde{\mathbf{X}}. \tag{10}$$

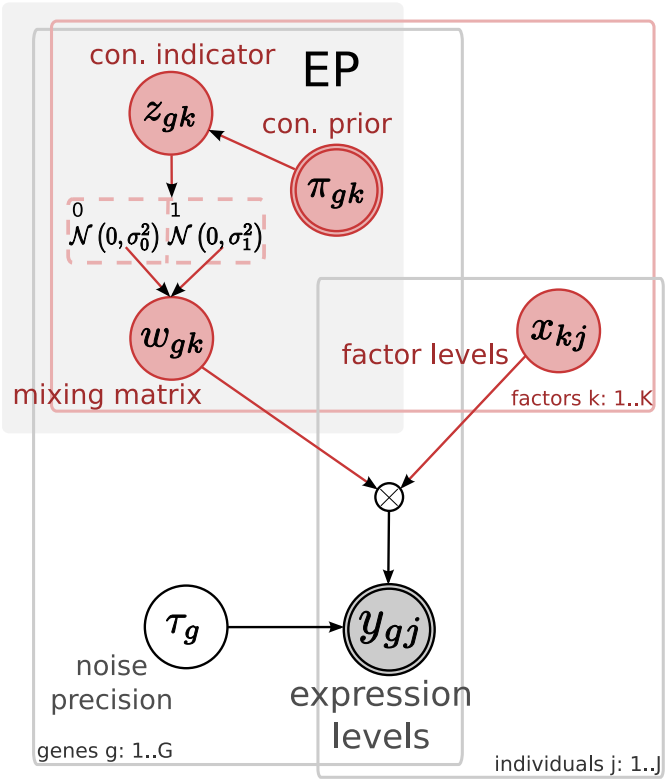


Figure 1: **The Bayesian network of the sparse factor analysis model.** Observed data $y_{g,j}$ for genes $g \in \{1, \ldots, G\}$ in individuals $j \in \{1, \ldots, J\}$ are modelled by the product between unobserved factor activations $\mathbf{x}_j$ and weights $\mathbf{w}_g$, and Gaussian observation noise. The indicator variables $z_{g,k}$ determine the state of the gate, either switching the corresponding mixing weight off ($w_{g,k} \sim \mathcal{N}(0, \sigma_0^2)$) or on ($w_{g,k} \sim \mathcal{N}(0, \sigma_1^2)$). *A priori* knowledge about the connectivity structure is introduced as a prior on the Bernoulli distribution parameter $\pi_{g,k}$. For the hybrid algorithm VB/EP, Expectation Propagation is used for inference in the submodel enclosed in the grey shaded area "EP".

These transformations $\mathbf{R}$ include arbitrary rotations, rescaling of the elements of $\mathbf{W}$ and $\mathbf{X}$, sign flips of the elements of any column of $\mathbf{W}$ and the corresponding row of $\mathbf{X}$ and arbitrary *permutations* of the factor labels. The sign ambiguity can be resolved *post-hoc* subject to application of suitable domain-specific knowledge. However, to render the model identifiable in the face of the remaining symmetries, further constraints are required.

In a Bayesian context, these constraints are provided through the specification of informative prior distributions. For example, we address the scaling ambiguity in a standard way by fixing the scale of the prior distribution over the latent factors to 1 (Equation (8)). To break the rotation and permutation symmetries, one might expect that the necessary constraints would be provided by the informative, sparsity inducing prior (Equation (5). Unfortunately, however, it turns out that this prior is only weakly effective in breaking the permutation symmetry: the magnitude of the symmetry breaking is small compared to the

magnitude of the density associated with modes corresponding to different permutations of the factor labels. Furthermore, for permutations involving the labels of factors with very different sparsity patterns, these modes can be separated by regions of very low probability density. This causes problems for both greedy, deterministic methods and Gibbs samplers: both methods can become stuck in a single mode. In addition, when these modes are relatively close, samplers can also suffer from the *label-switching problem* (Jasra et al., 2005) rendering ergodic averages meaningless.

Finding a single mode would not be a problem if the factor labels did not map directly to real entities and all permutations of the labels were equivalent. This is often the case in machine learning where the latent factors are introduced primarily as a device to facilitate the inference process. However, we are interested in situations where the meaning and identity of the factors is important. Consequently, only modes corresponding to a correct labelling of the factors are of interest.

To solve this problem, we exploit the informative prior in a more direct manner to differentiate between the inequivalent modes by incorporating an additional step within both deterministic and stochastic inference methods. The details of these steps are explained in Sections 4.4 and 5.6 respectively. The crucial importance of such a step is investigated in Section 6.1, where the reconstruction performance of the binary indicator matrix $\mathbf{Z}$ is compared for different algorithms with and without a permutation step.

4. Gibbs sampling

Gibbs sampling (Geman and Geman, 1984) is a versatile, stochastic inference method for approximating an intractable distribution by means of a finite set of samples. Each variable, or variable group, is sampled iteratively from its distribution conditioned on the current values of all other variables. This procedure constitutes the transition kernel of a markov chain under which the full joint distribution is invariant. Consequently, provided convergence to the equilibrium distribution is attained, it can be thought of as providing a "gold-standard" for the purpose of comparison. For a comprehensive introduction to such MCMC techniques see Robert and Casella (2004).

In the remainder of this section we outline the details of the Gibbs samplers.

4.1 Sampler implementation

The conditional distributions are derived by considering the functional dependence of the joint distribution on each variable or variable group. The choice of conjugate priors leads to conditional distributions of known form.

From equations (2),(3), (5) and (8) the full joint posterior for the sparse factor analysis model can be written as

$$\begin{aligned} P\left(\mathbf{W},\mathbf{X},\mathbf{Z},\boldsymbol{\tau},\sigma_1^2\,|\,\mathbf{Y},\boldsymbol{\pi}\right) &\propto P\left(\mathbf{Y}\,|\,\mathbf{W},\mathbf{X},\mathbf{Z},\boldsymbol{\tau}\right)P\left(\mathbf{X}\right)P\left(\mathbf{W}\,|\,\mathbf{Z}\right)P\left(\mathbf{Z}\,|\,\boldsymbol{\pi}\right)P\left(\boldsymbol{\tau}\right)P\left(\sigma_1^2\right) \\ &= \prod_{g=1}^{G}\prod_{j=1}^{J}\mathcal{N}\left(y_{g,j}\,\middle|\,\sum_{k=1}^{K}w_{g,k}x_{k,j},\tau_g\right)\prod_{k=1}^{K}\prod_{j=1}^{J}\mathcal{N}\left(x_{k,j}\,|\,0,1\right) \\ &\times \prod_{g=1}^{G}\prod_{k=1}^{K}\mathcal{N}\left(w_{g,k}\,|\,0,\sigma_1^2\right)^{z_{g,k}}\delta\left(w_{g,k}\right)^{1-z_{g,k}}\prod_{g=1}^{G}\prod_{k=1}^{K}\pi_{g,k}^{z_{g,k}}\left(1-\pi_{g,k}\right)^{1-z_{g,k}} \\ &\times \prod_{g=1}^{G}\frac{1}{\Gamma\left(\alpha_\tau\right)}\beta_\tau^{\alpha_\tau}\tau_g^{\alpha_\tau-1}\times\exp\left(-\beta_\tau\tau\right) \\ &\times \frac{1}{\Gamma\left(\alpha_\sigma\right)}\beta_\sigma^{\alpha_\sigma}\left(\frac{1}{\sigma_1^2}\right)^{\alpha_\sigma-1}\times\exp\left(-\frac{\beta_\sigma}{\sigma_1^2}\right). \end{aligned} \tag{11}$$

Under this model, the elements of the matrix $\boldsymbol{\pi}$ representing the prior network structure are all fixed and the variance of the prior Gaussian distribution over the latent factors is set to 1 to aid identifiability. With α_τ, β_τ, α_σ and β_σ set to uninformative values, the width of the 'slab', σ_1^2, is the only hyperparameter that is learned. The choice of a conjugate gamma prior distribution for the inverse width, $\frac{1}{\sigma_1^2}$, ensures that its conditional distribution is of known form (see Appendix A).

To derive conditional distributions for the four groups of parameters, $\mathbf{W}$, $\mathbf{X}$, $\mathbf{Z}$ and $\boldsymbol{\tau}$, it is important to note the conditional independence structure of the model: $\mathbf{z}_{g_1,.}\perp\mathbf{z}_{g_2,.}$ for $g_1 \neq g_2$, $\mathbf{w}_{g_1,.}\perp\mathbf{w}_{g_2,.}$ for $g_1 \neq g_2$, $\mathbf{x}_{.,j_1}\perp\mathbf{x}_{.,j_2}$ for $j_1 \neq j_2$ and $\tau_{g_1}\perp\tau_{g_2}$ for $g_1 \neq g_2$. Wherever possible, we sample conditionally dependent variables in groups. Derivations of the conditional distributions for the $\mathbf{X}$ and $\boldsymbol{\tau}$ groups are straight-forward and lead to standard distributions (see Appendix A) which may be easily sampled. However, due to the strong mutual dependencies of $\mathbf{W}$ and $\mathbf{Z}$, sampling of these parameter groups requires careful design of the sampler.

Clearly, pairs of parameters $z_{g,k}$ and $w_{g,k}$ are highly correlated: the probability of sampling a value of 1 for an indicator variable, $z_{g,k}$, depends heavily on the value of the associated $w_{g,k}$. However, while the $z_{g,k}$ remains zero, the associated $w_{g,k}$ cannot change and hence remains uninfluenced by the observed data. Consequently, standard sampling of these parameter groups leads to poor mixing: even if the data supports a non-zero value for the weight $w_{g,k}$ (and hence the indicator $z_{g,k}$) the algorithm will be slow to discover this. To address this problem we employ two alternative approaches: a *collapsed sampler* which samples from the true model (Equation (3)) and a *soft spike and slab* sampler which samples from the relaxed sparsity prior (Equation (5)).

4.2 Collapsed Gibbs sampler

A collapsed Gibbs sampler represents an attempt to improve mixing in the case of correlated variable groups by sampling one of them after first marginalising out the other. Such a scheme has previously been used by Sabatti and James (2006) and Pournara and Wernisch (2007) in the context of a similar sparse factor analysis model.

In this case, the $z_{g,k}$ are sampled after first marginalising out the $w_{g,k}$. This approach may be viewed as a means of sampling from the joint conditional distribution of the $z_{g,k}$ and $w_{g,k}$ by first sampling the elements of $\mathbf{Z}$ from $P(\mathbf{Z} \,|\, \mathbf{Y}, \boldsymbol{\pi}, \mathbf{X}, \boldsymbol{\tau})$ followed by those of $\mathbf{W}$ from $P(\mathbf{W} \,|\, \mathbf{Z}, \mathbf{Y}, \boldsymbol{\pi}, \mathbf{X}, \boldsymbol{\tau})$. Provided no other variables are sampled between these steps the posterior remains an invariant distribution of the markov chain.

Extracting from equation (11) those terms that depend on the elements of $\mathbf{Z}$ and $\mathbf{W}$ we find an expression for the joint conditional posterior of these variables:

$$P(\mathbf{W}, \mathbf{Z} \,|\, .) \propto \prod_{g=1}^{G} \left\{ \exp\left(-\frac{1}{2} \left(\left(\boldsymbol{y}_g - \mathbf{X}^{\mathrm{T}} \boldsymbol{\chi}_g \boldsymbol{w}_g \right)^{\mathrm{T}} \tau_g \mathbf{I} \left(\boldsymbol{y}_g - \mathbf{X}^{\mathrm{T}} \boldsymbol{\chi}_g \boldsymbol{w}_g \right) + \frac{1}{\sigma_1^2} \left(\left(\boldsymbol{\chi}_g \boldsymbol{w}_g \right)^{\mathrm{T}} \left(\boldsymbol{\chi}_g \boldsymbol{w}_g \right) \right) \right) \right) \right.$$
$$\left. \times \left(\frac{1}{\sigma_1^2} \right)^{|\boldsymbol{z}_g|} \prod_{k=1}^{K} \left\{ \pi_{g,k}^{z_{g,k}} \left(1 - \pi_{g,k} \right)^{1 - z_{g,k}} \right\} \right\} \tag{12}$$

Here $\boldsymbol{y}_g$ is a J-dimensional column vector corresponding to a column of the data matrix, $\mathbf{Y}$; $\mathbf{w}_g$ is a K-dimensional column vector corresponding to a row of the matrix $\mathbf{W}$; and $|\boldsymbol{z}_g|$ stands for the cardinality of the g^{th} row of $\mathbf{Z}$. Multiplication of $\boldsymbol{w}_g$ by the square matrix $\boldsymbol{\chi}_g = \mathrm{diag}(\mathbf{z}_g)$ gives a parameter vector with the sparsity pattern indicated by $\boldsymbol{z}_g$.

Noticing that the exponent is a quadratic form in $\boldsymbol{w}_g$, we may express the exponential term as the product of an unnormalised Gaussian and terms that depends on $\boldsymbol{z}_g$ but not $\boldsymbol{w}_g$:

$$P(\mathbf{W}, \mathbf{Z} \,|\, .) \propto \prod_{g=1}^{G} \left\{ \left(\frac{1}{\sigma_1^2} \right)^{|\boldsymbol{z}_g|} \exp\left\{ \frac{1}{2} \mathbf{m}_g^{\mathrm{T}} \boldsymbol{\Sigma}_g^{-1} \mathbf{m}_g \right\} \exp\left\{ -\frac{1}{2} \left(\mathbf{w}_g - \mathbf{m}_g \right)^{\mathrm{T}} \boldsymbol{\Sigma}_g^{-1} \left(\mathbf{w}_g - \mathbf{m}_g \right) \right\} \right\}$$
$$\times \prod_{g=1}^{G} \prod_{k=1}^{K} \pi_{g,k}^{z_{g,k}} \left(1 - \pi_{g,k} \right)^{1 - z_{g,k}} \ ,$$

where the mean, $\mathbf{m}_g$, and covariance, $\boldsymbol{\Sigma}_g$, of the unnormalised Gaussian are given by:

$$\boldsymbol{\Sigma}_g = \left(\tau_g \boldsymbol{\chi}_g \mathbf{X} \mathbf{X}^{\mathrm{T}} \boldsymbol{\chi}_g + \sigma_1^{-2} \boldsymbol{\chi}_g \right)^{-1} \qquad \mathbf{m}_g = \tau_g \boldsymbol{\Sigma}_g \boldsymbol{\chi}_g \mathbf{X} \mathbf{y}_g, \tag{13}$$

This distribution factorises over the G dimensions of the data vector, so we consider each dimension, g, separately. In the joint distribution, $P(\mathbf{w}_g, \mathbf{z}_g \,|\, .)$, only the unnormalised Gaussian term depends on $\mathbf{w}_g$. Consequently, marginalising over $\mathbf{w}_g$ immediately yields the following expression for the conditional distribution over $\mathbf{z}_g$:

$$P(\mathbf{z}_g \,|\, .) \propto \left(\frac{1}{\sigma_1^2} \right)^{|\boldsymbol{z}_g|} \exp\left\{ \frac{1}{2} \mathbf{m}_g^{\mathrm{T}} \boldsymbol{\Sigma}_g^{-1} \mathbf{m}_g \right\} \det |\boldsymbol{\Sigma}_g|^{1/2} \prod_{k=1}^{K} \pi_{g,k}^{z_{g,k}} \left(1 - \pi_{g,k} \right)^{1 - z_{g,k}} . \tag{14}$$

However, when no constraint is placed on the maximum number of ones in a row of indicators, $\mathbf{z}_g$, normalisation of the resulting multinomial distribution scales exponentially with the number of latent factors. To render the problem tractable, Sabatti and James (2006) considered a simplified problem in which large numbers of the indicators $z_{g,k}$ were

fixed to be exactly 0 or 1. We avoid imposing such a constraint by sampling each $z_{g,k}$ individually from its distribution conditioned on the current values of all other $z_{g,k}$, a procedure previously suggested by Pournara and Wernisch (2007):

$$P\left(z_{g,k} \,|\, .\right) = \frac{1}{N_{g,k}} \left\{ \pi_{g,k} \left(\frac{1}{\sigma_1^2}\right)^{z_{g,k}} \left[\det |\mathbf{\Sigma_g}|^{1/2} \exp\left\{\frac{1}{2}\mathbf{m}_g^{\mathrm{T}} \mathbf{\Sigma}_g^{-1} \mathbf{m}_g\right\}\right] \delta\left(z_{g,k}-1\right) \right.$$
$$\left. + \left(1-\pi_{g,k}\right) \left(\frac{1}{\sigma_1^2}\right)^{z_{g,k}} \left[\det |\mathbf{\Sigma}_g|^{1/2} \exp\left\{\frac{1}{2}\mathbf{m}_g^{\mathrm{T}} \mathbf{\Sigma}_g^{-1} \mathbf{m}_g\right\}\right] \delta\left(z_{g,k}\right) \right\}. \quad (15)$$

The normalising term, $N_{g,k}$, is given by

$$N_{g,k} = \pi_{g,k} \left(\frac{1}{\sigma_1^2}\right) \left[\det |\mathbf{\Sigma}_g|^{1/2} \exp\left\{\frac{1}{2}\mathbf{m}_g^{\mathrm{T}} \mathbf{\Sigma}_g^{-1} \mathbf{m}_g\right\}\right]_{z_{g,k}=1}$$
$$+ \left(1-\pi_{g,k}\right) \left[\det |\mathbf{\Sigma}_g|^{1/2} \exp\left\{\frac{1}{2}\mathbf{m}_g^{\mathrm{T}} \mathbf{\Sigma}_g^{-1} \mathbf{m}_g\right\}\right]_{z_{g,k}=0}, \quad (16)$$

where $[.]_{z_{g,k}=1}$ indicates that the quantity within square brackets is evaluated for $z_{g,k}=1$.

Once the K elements of the gth row of $\mathbf{Z}$ have been sampled, the collapsing step is completed by sampling the corresponding row of $\mathbf{W}$ from its conditional distribution:

$$P\left(\mathbf{w}_g \,|\, .\right) = \mathcal{N}\left(\mathbf{w}_g \,|\, \mathbf{m}_g, \mathbf{\Sigma}_g\right). \quad (17)$$

Unfortunately, this procedure entails a number of comparatively costly operations. For K latent factors, computing each normalising term, $N_{g,k}$ requires the inversion of a matrix with dimensionality equal to the number of non-zero elements in the gth row of $\mathbf{Z}$. In the worst case this could entail inversion of a $K \times K$ matrix for each such variable, leading to a time complexity scaling as GK^4. Fortunately, provided the problem is inherently sparse, the data will support non-zero values for only a small number, K_1, of the elements in a row of $\mathbf{Z}$. Typically K_1 is substantially smaller than K. In particular in applications such as gene regulatory networks, this number is thought to increase very slowly with K. Consequently, in practice the average time complexity of this sampling step $\mathcal{O}\left(GKK_1^3\right)$ may scale approximately linearly with the number of factors.

Nevertheless, this step constitutes a bottleneck. The soft spike sampler discussed below allows for faster sampling steps of the indicator variables $z_{g,k}$.

4.3 Soft Spike Gibbs Sampler

A soft spike sampler was previously introduced by George and McCulloch (1993) as a means of stochastic variable selection for regression. This algorithm samples from the relaxation of the slab and spike prior (Equation (5)).

The relative widths of the narrow and broader Gaussians determine a trade-off between the magnitude of $w_{g,k}$ necessary to infer $z_{g,k}=1$ and the mixing efficiency of the Markov chains. Larger ratios are expected to result in better mixing, but poorer inference. The sampling steps are the same as for the collapsed sampler with the exception of a different form for the conditional distributions of $\mathbf{w}_g$ and $z_{g,k}$

$$P\left(\mathbf{w_g} \,|\, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}, \right) = \mathcal{N}\left(\mathbf{w}_g \,\middle|\, \boldsymbol{\mu}_g, \mathbf{\Lambda}_g\right) \quad (18)$$

$$P\left(z_{g,k} \,|\, \pi_{g,k}, w_{g,k}\right) \propto \left(\pi_{g,k} e^{-w_{g,k}^2/2\sigma_1^2}\right)^{z_{g,k}} \left(\left(1-\pi_{g,k}\right) e^{-w_{g,k}^2/2\sigma_0^2}\right)^{1-z_{g,k}}, \quad (19)$$

where

$$\boldsymbol{\mu}_g = \tau_g \boldsymbol{\Lambda}_g \mathbf{X} \mathbf{y}_g \tag{20}$$

$$\boldsymbol{\Lambda}_g^{-1} = \boldsymbol{\tau}_g \mathbf{X}\mathbf{X}^T + \left(\sigma_1^{-2} \boldsymbol{\chi}_g + \sigma_0^{-2} \left(1 - \boldsymbol{\chi}_g\right)\right) \tag{21}$$

$$\boldsymbol{\chi}_g = \operatorname{diag}\left(\mathbf{z}_g\right). \tag{22}$$

This approach results in considerably cheaper sampling of the indicator variables $z_{g,k}$. As these variables are all conditionally independent in this model, their conditional distributions depend on only the single corresponding weight $w_{g,k}$ (Equation (19)). As a result, there is no costly matrix inversion required in order to sample from these weights. Instead, computation of the full covariance matrix for the multivariate normal distribution of each vector $\mathbf{w}_g$ requires a single matrix inversion. If this is done by computing the Cholesky decomposition of the matrix in Equation (21), then the required 'square-root' of the covariance is already available for the sampling step. Consequently, sampling of one complete group of $\mathbf{z}_g$ and $\mathbf{w}_g$ requires only a single Cholesky decomposition of a $K \times K$ matrix and so each Gibbs step has time complexity $\mathcal{O}\left(GK^3\right)$.

4.4 Permutation step

A common approach to addressing the problem of multiple modes due to label-switching (first proposed by Diebolt and Robert (1994)) is to enforce a hard identifiability constraint: at every iteration the labels are permuted so as to satisfy the constraint. However, aside from the problems of determining a suitable constraint in a multivariate problem, Fruhwirth-Schnatter (1998) demonstrated that an inappropriate choice can lead to a significant bias towards the constraint. Our solution to this problem is to use a local, permutation step that exploits the informative prior on the indicators $z_{g,k}$ to enforce a *soft* identifiability constraint. The "softness" of the constraint arises as a result of the stochastic nature of the step. It is local, in that it considers only pairwise swaps of factor labels. Locality is essential for the algorithm to scale to the numbers of factors in realistic problems.

The local permutation step is integrated within the Gibbs sampling scheme, by treating the configurations of pairs of factor labels as random variables. A single step consists of sampling from the posterior distribution of the labelling of a given pair, conditioned on the current values of all other variables, including all other factor labels. In each step we sample a labelling, either (k, m) or (m, k), for a pair of factors from $P\left(k, m \,|\, \mathbf{Y}, \mathbf{W}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\pi}\right)$. The form of this distribution is easily derived by retaining only those terms in the expression for the posterior that are *not* invariant to a permutation of the factor labels:

$$\begin{aligned} P\left(k, m \,|\, \mathbf{Y}, \mathbf{W}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\pi}\right) &\propto P\left(\mathbf{Y}, \mathbf{W}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\pi}\right) \\ &\propto P\left(\mathbf{Y} \,|\, \mathbf{W}, \mathbf{X}, \boldsymbol{\tau}\right) P\left(\mathbf{W} \,|\, \mathbf{Z}\right) P\left(\mathbf{X}\right) P\left(\mathbf{Z} \,|\, \boldsymbol{\pi}\right) P\left(\boldsymbol{\tau}\right) \\ &\propto P\left(\mathbf{Z} \,|\, \boldsymbol{\pi}\right) \\ &\propto P\left(\mathbf{z}_k, \mathbf{z}_m \,|\, \mathbf{Z}_{\setminus k,m}, \boldsymbol{\pi}_k, \boldsymbol{\pi}_m, \boldsymbol{\pi}_{\setminus k,m}\right). \end{aligned} \tag{23}$$

Here, $\mathbf{z_k}$ and $\boldsymbol{\pi}_k$ represent the kth columns of the binary matrix $\mathbf{Z}$ and the matrix of parameters, $\boldsymbol{\pi}$ respectively; $\mathbf{Z}_{\setminus \mathbf{k},\mathbf{m}}$ and $\boldsymbol{\pi}_{\setminus k,m}$ represent the remaining columns of these two matrices.

The simplification in the third line of Equation (23) arises from two observations. Firstly, we note that $P(\boldsymbol{\tau})$ does not depend on the factor labels. Secondly, we observe that $P(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}, \boldsymbol{\tau})$ is a product of Gaussian distributions in this model, each of which depends on the factor labels only through scalar products of $\mathbf{w}_g$ and $\mathbf{x}_j$. As the operation of taking the scalar product is invariant to a permutation of these labels, $P(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}, \boldsymbol{\tau})$ is therefore also invariant to such a permutation. Similar reasoning applies to $P(\mathbf{W} \mid \mathbf{Z})$ and $P(\mathbf{X})$. The fourth line follows simply because, in each step, we consider swapping the labels of only one pair of factors with all other labels remaining fixed.

Thus we find that the conditional distribution for the configuration of a given label pair is a Bernoulli distribution with parameter, μ_{km} given by

$$\mu_{km} = \frac{\prod_{g=1}^{G} \pi_{im}^{z_{ik}} \left(1 - \pi_{im}\right)^{(1-z_{ik})} \pi_{ik}^{z_{im}} \left(1 - \pi_{ik}\right)^{(1-z_{im})}}{\sum_{j,l=(k,m)} \prod_{g=1}^{G} \pi_{im}^{z_{ij}} \left(1 - \pi_{im}\right)^{(1-z_{ij})} \pi_{ik}^{z_{il}} \left(1 - \pi_{ik}\right)^{(1-z_{il})}}, \tag{24}$$

The normalising sum in the denominator is over the two possible configurations of the indices j, ℓ, i.e., $j = k, \ell = m$, or $j = m, \ell = k$. Sampling from this distribution is straightforward.

It is clear from Equation (24) that this move 'coaxes' the markov chain to sample predominantly from permutations of labels that have a high probability under the prior. Consequently, it relies on the prior being informative.

The individual sampling steps of a Gibbs algorithm may be combined in a variety of ways that all preserve the properties of the markov chain: in fixed order, random order or palindromically. We implement the permutation steps in a fixed order, cycling through all possible pairwise combinations of labels. This step has time complexity $\mathcal{O}\left(GK^2\right)$. However, it is likely that more efficient, though problem-specific, adaptive schedules could be devised to exploit the fact that factors with larger numbers of links are learned more quickly than those with few.

This completes the description of the samplers. The steps of both algorithms are summarised below.

Sampling scheme:
repeat
  sample σ_1^2 from $P\left(\sigma_1^2 \,|\, \mathbf{Z}, \mathbf{W}, \alpha_\sigma, \beta_\sigma\right)$
  for $g = 1$ to G **do**
    if Collapsed Gibbs **then**
      for $k = 1$ to K **do**
        sample $z_{g,k}$ from $P\left(z_{g,k} \,|\, \{\mathbf{Y}, \mathbf{X}, \mathbf{z_g} \setminus z_{g,k}\}, \boldsymbol{\tau}, \boldsymbol{\pi}_g\right)$
    else
      for $k = 1$ to K **do**
        sample $z_{g,k}$ from $P\left(z_{g,k} \,|\, \mathbf{Y}, w_{g,k}, \mathbf{X}, \{\mathbf{z_g} \setminus z_{g,k}\}, \boldsymbol{\tau}, \pi_{g,k},\right)$
    sample $\mathbf{w_g}$ from $P\left(\mathbf{w_g} \,|\, \mathbf{Y}, \mathbf{X}, \mathbf{z_g}, \boldsymbol{\tau},\right)$
  for $j = 1$ to J **do**
    sample $\mathbf{x}_j$ from $P\left(\mathbf{x}_j \,|\, \mathbf{Y}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\tau}\right)$
  for $g = 1$ to G **do**
    sample τ_g from $P\left(\tau_g \,|\, \mathbf{Y}, \mathbf{W}, \mathbf{Z}\right)$
  for $k = 1$ to $K - 1$ **do**
    for $m = k + 1$ to K **do**
      sample a pairwise labelling k, m from $P\left(k, m \,|\, \mathbf{z_k}, \mathbf{z_m}, \boldsymbol{\pi}_k, \boldsymbol{\pi}_m\right)$
until convergence

5. Deterministic approximate inference

Deterministic approximations are appealing mainly because they typically converge faster than sampling schemes. In the following we begin by reviewing variational Bayesian (VB) learning and then discuss a novel, hybrid scheme. This combines VB with Expectation Propagation (EP) for inference in the sparse factor analysis model. Both of these algorithms are based on the relaxed sparsity prior (Equation (5)).

5.1 Variational Bayesian learning

Variational Bayesian (VB) learning is a general purpose technique for deterministic approximate inference in a wide range of probabilistic models (**?**). In this mean field approach the exact posterior distribution $P(\mathbf{H} \,|\, \mathbf{V})$ is approximated by a factorised distribution $Q(\mathbf{H} \,|\, \mathbf{V}) = \prod_i Q(\mathbf{H}_i)$, where $\mathbf{V}$ denotes the set of all visible (observed) variables and $\mathbf{H}$ are hidden variables – including both latent variables and model parameters. The variational approximation is fitted by minimising the *exclusive* KL divergence, $\mathrm{KL}[Q \,||\, P]$, between the approximate Q-distribution and the true posterior. Functional minimisation of this KL divergence leads to update rules for individual factors $Q(\mathbf{H}_i)$, which require the calculation of the log average likelihood under the current state of all other Q-distributions:

$$\tilde{Q}(\mathbf{H}_i) \propto \exp\Big\{\langle \log P(\mathbf{V}, \mathbf{H}) \rangle_{Q \setminus \mathbf{H}_i}\Big\}. \tag{25}$$

These updates are iterated in turn for all factors $Q(\mathbf{H}_i)$ until convergence to a fixed point solution is reached.

To derive a variational algorithm for the sparse factor analyser, we choose a factorisation of the form

$$Q(\mathbf{W}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}) = \prod_{g=1}^{G} \left[Q(\mathbf{w}_g) Q(\tau_g) \prod_{k=1}^{K} Q(z_{g,k}) \right] \prod_{j=1}^{J} Q(\mathbf{x}_j). \tag{26}$$

The explicit update equations for individual factors $Q(\cdot)$ follow from Equation (25). In the following a detailed treatment of the update rules of the mixing weights $Q(\mathbf{w}_g)$ and the indicators $Q(z_{g,k})$ is provided. The remaining updates for factor activations $Q(\mathbf{x}_j)$ and the noise levels $Q(\tau_g)$ are fairly standard and are given in Appendix B. Using Equation (25), the update for the mixture weights of one gene, $Q(\mathbf{w}_g)$, is

$$\begin{aligned} Q(\mathbf{w}_g) &\propto \exp\left\{ \langle \log P(\mathbf{Y}, \mathbf{W}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}) \rangle_{Q \setminus \mathbf{w}_g} \right\} \\ &\propto \exp\left\{ \langle \log P(\mathbf{y}_g \,|\, \mathbf{w}_g, \mathbf{X}, \tau_g) + \log P(\mathbf{w}_g \,|\, \mathbf{z}_g) \rangle_{Q \setminus \mathbf{w}_g} \right\} \qquad (27) \\ &\propto \underbrace{\exp\left\{ \langle \log P(\mathbf{y}_g \,|\, \mathbf{w}_g, \mathbf{X}, \tau_g) \rangle_{Q \setminus \mathbf{w}_g} \right\}}_{M_{\mathbf{W} \cdot \mathbf{X} \to \mathbf{w}_g}} \underbrace{\exp\left\{ \langle \log P(\mathbf{w}_g \,|\, \mathbf{z}_g) \rangle_{Q \setminus \mathbf{w}_g} \right\}}_{M_{\mathbf{W} \,|\, \mathbf{Z} \to \mathbf{w}_g}}, \qquad (28) \end{aligned}$$

where the expectations are with respect to all Q-distributions except the one that is being refined. The resulting Gaussian approximate factor, $Q(\mathbf{w}_g)$, 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.

In a message-passing scheme (Winn and Bishop, 2006), we interpret $M_{\mathbf{W} \cdot \mathbf{X} \to \mathbf{w}_g}$ as the message sent from the product factor $f_{\mathbf{W} \cdot \mathbf{X}}$ to the variable $\mathbf{w}_g$. The parameters of this Gaussian $M_{\mathbf{W} \cdot \mathbf{X} \to \mathbf{w}_g} \propto \mathcal{N}\left(\mathbf{w}_g \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W} \cdot \mathbf{X} \to \mathbf{w}_g}, \tilde{\boldsymbol{\Sigma}}_{\mathbf{W} \cdot \mathbf{X} \to \mathbf{w}_g} \right)$ are:

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{W} \cdot \mathbf{X} \to \mathbf{w}_g} = \left(\langle \tau_g \rangle \sum_{j=1}^{J} \left\langle \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}} \right\rangle \right)^{-1} \tag{29}$$

$$\tilde{\mathbf{m}}_{\mathbf{W} \cdot \mathbf{X} \to \mathbf{w}_g} = \tilde{\boldsymbol{\Sigma}}_{\mathbf{w}_g} \left(\langle \tau_g \rangle \sum_{j=1}^{J} \langle \mathbf{x}_j \rangle \left(\mathbf{y}_j \right) \right). \tag{30}$$

These follow directly from Equation (28), by writing out the individual likelihood terms. In the same way we identify $M_{\mathbf{W} \,|\, \mathbf{Z} \to \mathbf{w}_g}$ as a message from the mixture prior to $\mathbf{w}_g$.

To facilitate the derivation of the VB/EP hybrid in Section 5.3, we split the sparse factor analysis model into two models, treating $\mathbf{W}$ as a shared variable (see Figure 1). In the following we will refer to the model defined over the observed data $\mathbf{Y}$, noise precisions $\boldsymbol{\tau}$, factor activations $\mathbf{X}$ and the shared weights $\mathbf{W}$ as the core factor analysis. The smaller model, defined over the shared weights $\mathbf{W}$, the mixture indicators $\mathbf{Z}$ and prior mixing coefficients $\boldsymbol{\pi}$ will be referred to as the sparsity submodel.

5.2 VB for the sparsity submodel

Making the conditioning on the incoming messages $\left\{M_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right\}_g$ explicit, the joint probability over weights and indicators in the submodel is

$$P\left(\mathbf{W},\mathbf{Z}\,|\,\left\{M_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right\}_g\right)\propto \prod_{g=1}^{G}\left[\mathcal{N}\left(\mathbf{w}_g\,\middle|\,\tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g},\tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right)\prod_{k=1}^{K}P(w_{g,k}\,|\,z_{g,k})P(z_{g,k})\right], \tag{31}$$

which factorises over genes g. If we again choose Variational Bayes for approximate inference, we obtain update rules for the approximate factor $Q(\mathbf{w}_g)$ that are consistent with those derived earlier in Equation (27),

$$Q(\mathbf{w}_g)\propto\mathcal{N}\left(\mathbf{w}_g\,\middle|\,\tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g},\tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right)\exp\left\{\langle\log P(\mathbf{w}_g\,|\,\mathbf{z}_g)\rangle_{Q\setminus\mathbf{w}_g}\right\}. \tag{32}$$

Writing out both terms, we obtain

$$Q(\mathbf{w}_g)\propto\exp\Bigg\{-\frac{1}{2}\left(\mathbf{w}_g-\tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right)^T\tilde{\mathbf{\Sigma}}^{-1}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\left(\mathbf{w}_g-\tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right) -\frac{1}{2}\mathbf{w}_g^{\mathrm{T}}\mathrm{diag}\left(\left\{\sum_{c=0}^{1}Q(z_{g,k}=c)\frac{1}{\sigma_c^2}\right\}_k\right)\mathbf{w}_g\Bigg\}, \tag{33}$$

and hence the explicit parameters of the Gaussian factor $Q(\mathbf{w}_g)=\mathcal{N}\left(\mathbf{w}_g\,\middle|\,\tilde{\mathbf{m}}_{\mathbf{w}_g},\tilde{\mathbf{\Sigma}}_{\mathbf{w}_g}\right)$ follow as

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_{\mathbf{w}_g}&=\left[\tilde{\mathbf{\Sigma}}^{-1}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}+\mathrm{diag}\left(\left\{\sum_{c=0}^{1}Q(z_{g,k}=c)\frac{1}{\sigma_c^2}\right\}_k\right)\right]^{-1}\\ \tilde{\mathbf{m}}_{\mathbf{w}_g}&=\tilde{\mathbf{\Sigma}}_{\mathbf{w}_g}\tilde{\mathbf{\Sigma}}^{-1}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}.\end{aligned} \tag{34}$$

Update rules for the responsibilities, $Q(z_{g,k}=1)=\tilde{\pi}_{g,k}$, can be obtained in the same vein using

$$\begin{aligned}\tilde{\pi}_{g,k}&\propto\pi_{g,k}\exp\left\{\left\langle\log\mathcal{N}\left(w_{g,k}\,\middle|\,0,\sigma_1^2\right)\right\rangle_{Q\setminus z_{g,k}}\right\}\\ (1-\tilde{\pi}_{g,k})&\propto(1-\pi_{g,k})\exp\left\{\left\langle\log\mathcal{N}\left(w_{g,k}\,\middle|\,0,\sigma_0^2\right)\right\rangle_{Q\setminus z_{g,k}}\right\}.\end{aligned} \tag{35}$$

With these updates for weights and indicator variables, the description of VB learning in the sparse factor analyser is completed. It is important to reemphasise that, for VB, the split of the model does not alter the inference and leads to identical effective update rules.

5.3 VB/EP hybrid inference

As an alternative to Variational Bayesian learning, we now consider Expectation Propagation (Minka, 2001b). As with VB, EP is based on the minimisation of a KL divergence,

however with swapped arguments, $\mathrm{KL}_{\mathrm{EP}} = \mathrm{KL}\left[P \,||\, Q\right]$, which leads to an approximation with rather different properties. In settings where the true posterior is multimodal, as it is the case for the sparsity prior (Section 3), the VB KL divergence favours the approximation to lock onto a single mode. In contrast, EP averages over the set of modes (see Discussion in Minka (2005)). A comprehensive introduction to EP can be found in Minka (2001b) and Bishop (2006, Chapter 10).

In general, there is no clear-cut answer as to which of VB and EP provides a better approximation (for a number of problems where EP has been shown to be more accurate see (e.g. Nickisch and Rasmussen, 2008; Frey et al., 2000)). 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 (Seeger, 2008).

Applying EP to the model considered here, we follow an alternative route and connect VB with EP. As it turns out the resulting hybrid inference algorithm combines the stability and efficiency of VB for the core factor analysis with the improved accuracy of EP for the sparsity mixture. For the sparsity mixture prior, EP is more accurate due to it's mode averaging behavior, retaining more uncertainty in the estimates of the indicator variables. Note that the benefits of this mode averaging are linked to the strong prior information that breaks the symmetry for the mixture components of the sparsity prior. When multiple modes of the posterior are equivalent, for example in symmetric mixture models, VB is often found to yield practical and accurate answers despite locking onto single mode (Paquet, 2008). However, in situations with strong prior knowledge, mode averaging of EP is likely to yield fewer "false positives" in the sense of overconfident decisions regarding the state of the indicator variables; see also the discussion in Minka (2005).

From a theoretical perspective this hybrid algorithm can be understood as choosing alternative divergence measures for different parts of the graphical model (see Figure 1). Minka has noted earlier (Minka, 2005) that combinations of EP and VB (in fact a more general class of α-divergences) are possible. However, there are few applications where hybrid inference schemes have been used in practise, probably because of the difficulty of implementing such algorithms[1]

5.4 EP for the sparsity submodel

As for inference using VB (Section 5.2), the factorisation of the incoming messages, $\left\{M_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right\}_g$, induces a factorisation over genes. Hence, we consider inference for a single gene only. Conditioned on the incoming message, the joint probability over a vector of weights and corresponding indicators for a gene g is

$$P\left(\mathbf{w}_g, \mathbf{z}_g \,|\, M_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right) \propto \mathcal{N}\left(\mathbf{w}_g \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}, \tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right) \prod_{k=1}^{K} P(w_{g,k} \,|\, z_{g,k}) P(z_{g,k}). \quad (36)$$

1. One example is Welling et al. (2008), who study a hybrid of Gibbs sampling and VB in another context.

As for VB, we choose an approximate form:

$$Q(\mathbf{w}_g, \mathbf{z}_g) = s \cdot \mathcal{N}\left(\mathbf{w}_g \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}, \tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right) \prod_{k=1}^{K} q(w_{g,k})q(z_{g,k}), \tag{37}$$

where the factors $q(w_{g,k})q(z_{g,k})$ are meant to approximate $P(w_{g,k} \mid z_{g,k})P(z_{g,k})$. The explicit scale of the approximation, s can be used to obtain estimates of the marginal likelihood within the EP framework (Minka, 2005). As we aim to connect this model with VB we drop this scale in the following. Choosing factor distributions that match the VB approximation, Q-distributions for weights are Gaussian, $q(w_{g,k}) = \mathcal{N}\left(w_{g,k} \,\middle|\, \tilde{\mu}_{w_{g,k}}, \tilde{\sigma}^2_{w_{g,k}}\right)$, and factors of indicators are Bernoulli distributed, $q(z_{g,k}) = \text{Bernoulli}(z_{g,k} \mid \tilde{\pi}_{g,k})$. While the overall approximation in Equation (37) is fully factorised over indicators $z_{g,k}$, it is multivariate Gaussian in the weights $\mathbf{w}_g$. Writing out the product of the Gaussian prior and the individual Gaussian factors $q(w_{g,k})$ yields

$$Q(\mathbf{w}_g, \mathbf{z}_g) \propto \mathcal{N}\left(\mathbf{w}_g \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{w}_g}, \tilde{\mathbf{\Sigma}}_{\mathbf{w}_g}\right) \prod_{k=1}^{K} q(z_{g,k}). \tag{38}$$

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

$$\tilde{\mathbf{\Sigma}}_{\mathbf{w}_g} = \left(\tilde{\mathbf{\Sigma}}^{-1}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g} + \tilde{\mathbf{\Sigma}}\right)^{-1} \qquad \tilde{\mathbf{m}}_{\mathbf{w}_g} = \tilde{\mathbf{\Sigma}}_{\mathbf{w}_g}\left[\tilde{\mathbf{\Sigma}}^{-1}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g} + \tilde{\mathbf{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}\right]. \tag{39}$$

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, $q(w_{g,i})q(z_{g,i})$, the local KL divergence to be minimised is

$$\begin{aligned}\text{KL}\Bigg[\mathcal{N}\left(\mathbf{w}_g \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}, \tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right) \prod_{k\neq i} q(w_{g,k})q(z_{g,k}) \overbrace{P(w_{g,i} \mid z_{g,i})P(z_{g,i})}^{\text{exact factor}} \Bigg| \Bigg| \\ \mathcal{N}\left(\mathbf{w}_g \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}, \tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right) \prod_{k\neq i} q(w_{g,k})q(z_{g,k}) \underbrace{q(w_{g,i})q(z_{g,i})}_{\text{approximation}}\Bigg].\end{aligned} \tag{40}$$

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:

$$\begin{aligned}q_{\setminus i}(w_{g,i}) &= \int_{\mathbf{w}_{g,\setminus i}} \mathcal{N}\left(\mathbf{w}_g \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}, \tilde{\mathbf{\Sigma}}_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}\right) \prod_{k\neq i} q(w_{g,k})\, \mathrm{d}\mathbf{w}_{g,\setminus i} \\ &= \mathcal{N}\left(w_{g,i} \,\middle|\, \tilde{\mu}_{\setminus i}, \tilde{\sigma}^2_{\setminus i}\right).\end{aligned} \tag{41}$$

As the $z_{g,k}$ are independent in the approximation, marginalisation of the $q(z_{g,k})$ is trivial and factors other than $q(w_{g,i})$ can be dropped. The cavity distribution $q_{\setminus i}(w_{g,i})$ can be calculated efficiently from the current full approximation (Equation (38)),by dividing out the contribution of the ith factor (for an instructive tutorial on how to handle cavity distributions efficiently, see Rasmussen and Williams (2006) chapter 3.6). Using the definition of

the cavity distribution, the KL-divergence in Equation (40) can be expressed in a compact form:

$$\mathrm{KL}\Bigg[q_{\setminus i}(w_{g,i})\overbrace{P(w_{g,i}\,|\,z_{g,i})P(z_{g,i})}^{\text{exact factor}}\;\Bigg|\Bigg|\;q_{\setminus i}(w_{g,i})\underbrace{q(w_{g,i}\,|\,\tilde{\mu}_{w_{g,i}},\tilde{\sigma}^2_{w_{g,i}})q(z_{g,i}\,|\,\tilde{\pi}_{g,i})}_{\text{approximation}}\Bigg]. \tag{42}$$

Minimising Equation (42) with respect to the parameters of the Gaussian factor $q(w_{g,i})$ leads to moment-matching conditions (Minka, 2001a). As an exponential family member, the new parameters of the approximate factor $q(w_{g,i})$ 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:

$$\begin{aligned}
F_C &= \int_{w_{g,i}} q_{\setminus i}(w_{g,i}) \sum_{c=\{0,1\}} P(w_{g,i}\,|\,z_{g,i}=c)P(z_{g,i}=c)\,\mathrm{d}w_{g,i}\\
F_\mu &= \frac{1}{F_C}\int_{w_{g,i}} q_{\setminus i}(w_{g,i}) \sum_{c=\{0,1\}} P(w_{g,i}\,|\,z_{g,i}=c)P(z_{g,i}=c)w_{g,i}\,\mathrm{d}w_{g,i}\\
F_{\sigma^2}+F_\mu^2 &= \frac{1}{F_C}\int_{w_{g,i}} q_{\setminus i}(w_{g,i}) \sum_{c=\{0,1\}} P(w_{g,i}\,|\,z_{g,i}=c)P(z_{g,i}=c)w_{g,i}^2\,\mathrm{d}w_{g,i}.
\end{aligned} \tag{43}$$

Analytic expressions for these moments are derived in Appendix C. In the same vein, optimisation of Equation (42) with respect to $\tilde{\pi}_{g,i}$ leads to updates of the posterior over the indicator variables

$$\begin{aligned}
\tilde{\pi}_{g,i} &\propto \pi_{g,i}\int_{w_{g,i}} q_{\setminus i}(w_{g,i})\mathcal{N}\left(w_{g,i}\,\middle|\,0,\sigma_1^2\right)\,\mathrm{d}w_{g,i}\\
1-\tilde{\pi}_{g,i} &\propto (1-\pi_{g,i})\int_{w_{g,i}} q_{\setminus i}(w_{g,i})\mathcal{N}\left(w_{g,i}\,\middle|\,0,\sigma_0^2\right)\,\mathrm{d}w_{g,i}.
\end{aligned} \tag{44}$$

5.5 Connecting the EP submodel with VB

Having established EP-inference in the sparsity submodel, the remaining task is to connect both models. In a joint inference schedule, first a VB iteration of the core factor analysis is performed, updating hidden activations $Q(\mathbf{X})$ and noise estimates $Q(\boldsymbol{\tau})$ and calculating the messages $M_{\mathbf{W}\cdot\mathbf{X}\to\mathbf{w}_g}$. Subsequently, EP is applied to infer approximate marginals $Q(\mathbf{W})$ and $Q(\mathbf{Z})$. These factors enter consecutive VB updates in a manner analogous to standard VB factors.

For EP updates in the submodel, we need to choose a schedule for individual factor updates, i.e. an order and the number of sweeps through all factors k for the local KL updates in Equation (42). In the experiments, a single sweep in a randomised order is used. Empirically we found that the results differ very little when performing additional cycles. Most likely this is due to the fact that the speed of convergence of the VB model is comparably slow and hence an approximate solution from a single EP iteration is sufficient.

An aspect left for future work is to investigate how to retain an approximation to the model evidence within the hybrid inference scheme. Both VB and EP alone yield an approximation to the model evidence, and in principle it should be possible to estimate the evidence in the combined model as well (Minka, 2005).

5.6 Label-switching move

The weak symmetry breaking discussed in Section 3.1 also affects deterministic inference methods. The respective sparsity sub-models can be readily extended by representing the permutation of factors explicitly

$$P(\mathbf{Z} \mid \boldsymbol{\pi}) = \prod_{k=1}^{K} P(\mathbf{Z}_k \mid \boldsymbol{\pi}_{q(k)}). \tag{45}$$

In EM-type manner, label switching can be accounted for by optimising over the permutation $q(k)$. This is implemented greedily, optimising the permutation assignment for one factor at a time.

6. Experiments

We compared the performance of the deterministic inference methods (VB, VB/EP) with the two Gibbs samplers (collapsed Gibbs, soft spike and slab Gibbs) on simulated problems of two different sizes and on real gene expression data from the baker's yeast *Saccharomyces cerevisiae*. For the real data we used a connectivity prior for the network empirically determined from genome-wide Chromatin Immunoprecipitation (ChIP-chip) data. The details of these experiments are described below.

6.1 Simulated networks

First, we considered two simulated datasets of different sizes (*small* and *large*). The elements of the matrices, $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{Z}$ and $\mathbf{X}$ were drawn from the model – Equations (2), (3) and (8). The simulation parameters for both datasets are summarised in Table 1.

	Small	**Large**
Dimension (genes)	486	1000
No. latent factors	20	60
No. of individuals	20	100
FNR (η_0)	0.1	0.05
FPR (η_1)	0.25	0.25
Sparsity	0.15	0.09
Noise hyperparameter (α_τ)	1.0	1.0
Noise hyperparameter (β_τ)	0.01	0.01
Slab width σ_1^2	1.0	1.0

Table 1: Summary of simulation parameter for both synthetic datasets.

Sampling yields "gold standard"

When the Gibbs samplers reach convergence the inferred posterior can be regarded as a "gold standard". Sampler convergence was monitored by means of well established diagnostics: the Rhat test of Gelman and Rubin (Gelman and Rubin, 1992) applied to five

independent markov chains. Figure 2d summarises the results of applying this test on the small example, monitoring the fraction of parameters remaining unconverged as a function of the runtime. From these data it is evident that both samplers reached convergence within the allowed runtime, where the soft slab and spike sampler converged slightly quicker than the collapsed sampler.

Performance as a function of CPU time

To study the trade-off between performance and computational cost, we examined the performance of the different methods as a function of CPU time under three different measures: predictive accuracy for the network reconstruction task, the mean log probability of the true network under the posterior, and the mean squared error in estimating the real valued parameters. The mean squared error was computed using the products of individual elements of $\mathbf{W}$ and $\mathbf{X}$, as these quantities are invariant to ambiguities of sign and scale.

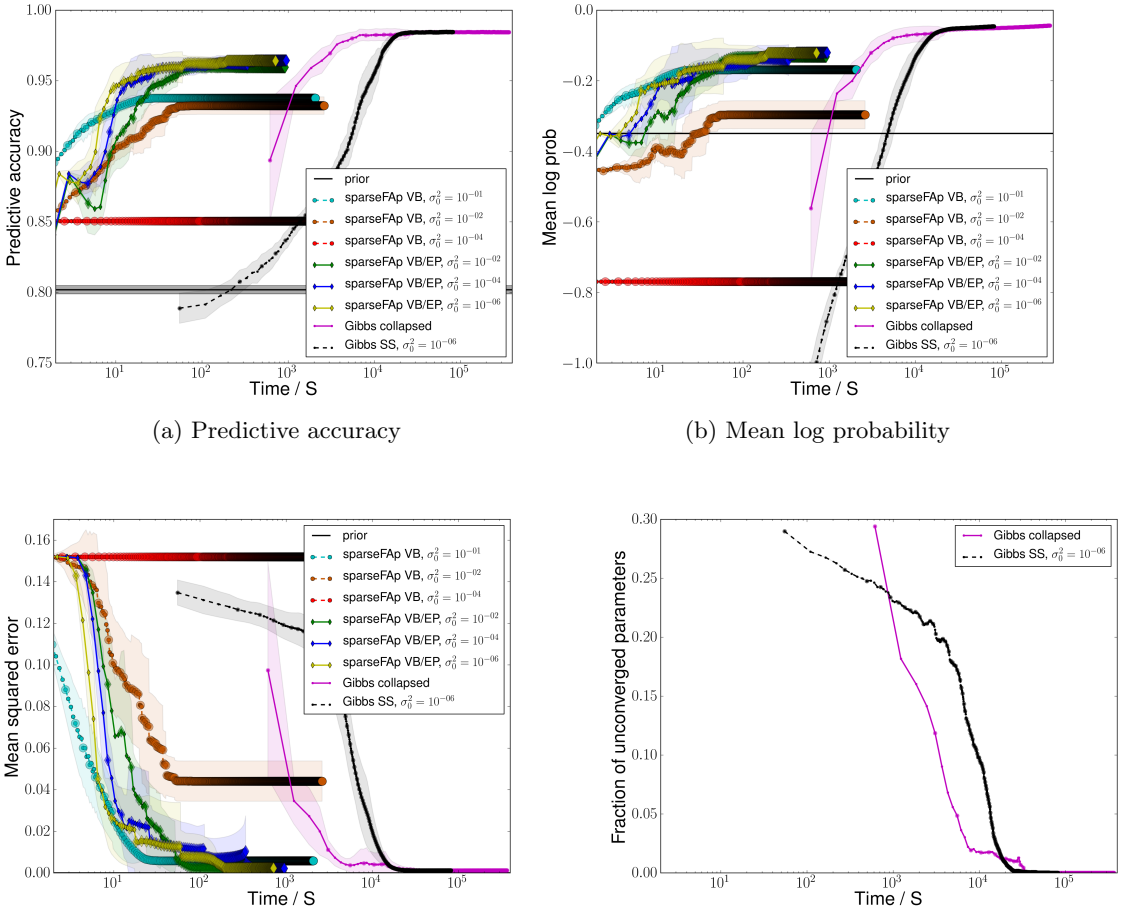
Figure 2 shows the comparative performance of the different algorithms on the *small* network example. The figure also illustrates the the effect of varying σ_0^2, the hyperparameter determining the variance of the narrow Gaussian. The corresponding results for the *large* network are shown in Figure 3.

Two comparisons are particularly worthy of note. First, it can be seen from figure 2 that the best deterministic method, the VB/EP hybrid, suffers only a small reduction in performance under all three measures when compared to the "gold standard" provided by the samplers. At the same time, it achieves a considerable saving in computation time: the approximate, deterministic methods converged at a rate more than two orders of magnitude faster than the samplers. This result was qualitatively similar for the larger example as shown in figure 3, suggesting that the VB/EP hybrid might provide a useful approximation on real data sets where the convergence of samplers is likely to become prohibitively slow.

The second important comparison is between the performances of the two different approximate inference algorithms. Under all three measures, the VB/EP hybrid, performing EP on the sparsity submodel, consistently outperformed pure variational bayes. The results also illustrate that the limit $\sigma_0 \to 0$ is not accessible in pure VB where performance degrades for small σ_0. This behaviour can be explained by the fully factorised form of the variational approximation (Section 5.2), that in combination with the greedy behaviour of VB leads to inflexible solutions. In contrast, the "mode averaging" behaviour of EP appeared to be more robust: The approximation is well-behaved in the sense that the limit $\sigma_0 \to 0$ is accessible in practice and the model performance increased as the true simulation prior was approached.

Utility of the permutation move

Next, we investigated the utility of the permutation move, addressing the weak symmetry breaking property of the model (Section 3.1). Figure 4 illustrates that, for both families of methods – deterministic approximations and sampling methods – the permutation move improved the accuracy of the inferred network by a significant margin. It is notable that the problem is particularly severe for the sampling methods where, as explained in section



(a) Predictive accuracy

(b) Mean log probability

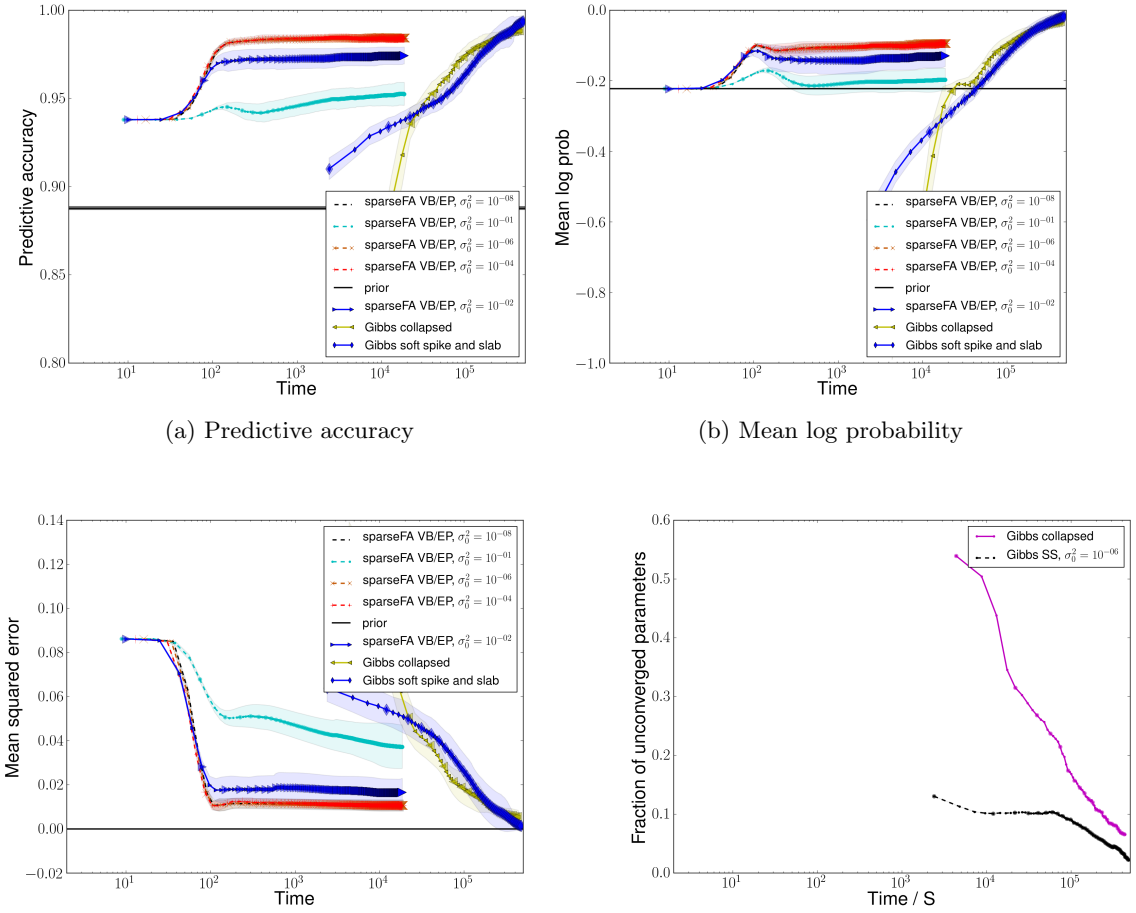
(c) Mean Squared error

(d) Unconverged parameters (Gibbs)

Figure 2: **Small: Performance on simulated dataset for different algorithms and error measures.** The performance of each model is plotted as a function of the CPU time for alternative performance measures. **(a)** Predictive accuracy of the inferred network structure. **(b)** Mean log probability of the true network under the marginal predictive distribution. **(c)** Prediction error of weights and activation profiles, evaluated as the root mean squared error of individual product terms, $w_{g,k}\, x_{j,k}$. **(d)** Convergence of samplers, monitored as the fraction of unconverged parameters using 5 Gibbs chains. Empirical error bars of plus or minus one standard deviation (**a,b,c**) are from 5 random restarts or Gibbs chains respectively.

3.1, chains can easily become stuck in modes corresponding to an incorrect permutation of the factor labels.

An explanation for this observation is the local nature of the sampling updates for each iteration of the Gibbs sampler. Compared to these small moves, the deterministic



(a) Predictive accuracy

(b) Mean log probability

(c) Mean Squared error

(d) Unconverged parameters (Gibbs)

Figure 3: **Large: Performance on simulated dataset for different algorithms and error measures.** The performance of each model is plotted as a function of the CPU time for alternative performance measures. **(a)** Predictive accuracy of the inferred network structure. **(b)** Mean log probability of the true network under the marginal predictive distribution. **(c)** Prediction error of weights and activation profiles, evaluated as the root mean squared error of individual product terms, $w_{g,k}\, x_{j,k}$. **(d)** Convergence of samplers, monitored as the fraction of unconverged parameters using 5 Gibbs chains. Empirical error bars of plus or minus one standard deviation (**a,b,c**) are from 5 random restarts or Gibbs chains respectively.

approximations progress faster and hence are able to overcome local optima boundaries; however, at the price of being more greedy.

Comparison of marginal distributions for indicators **Z**

The analysis of the number of converged parameters (Section 6.1) suggests that the Gibbs samplers converged to the true equilibrium distribution on the small dataset. Hence, it is

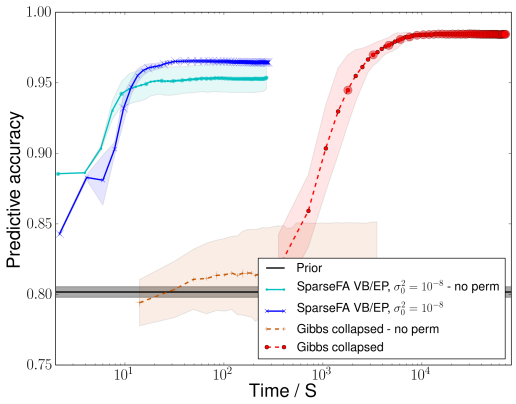


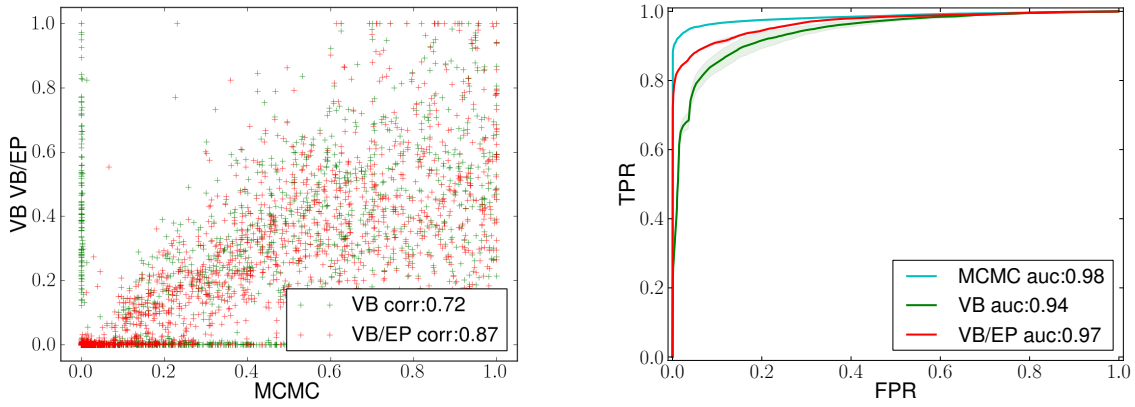
Figure 4: **Small: Accuracy of inferred network structure for different algorithms with and without permutation move.** Comparative results for the deterministic algorithm VB/EB and the collapsed Gibbs sampler with and without a permutation move addressing the weak symmetry breaking property. For samplers the move is essential; for VB/EP it also leads to a significant gain in performance and reduced variance over different initialisations.

interesting to compare the posterior marginal distributions inferred by the samplers and the deterministic approximations. Figure 5a regresses the "gold standard" marginal probabilities from the converged collapsed sampler against those inferred by both the pure VB and the VB/EP hybrid. The scatter plot and correlation coefficients suggest that the posterior inferred by the VB/EP approach provides a significantly better approximation than that inferred by pure VB.

The accuracy of the posterior indicators at convergence are compared in Figure 5b which shows ROC curves for the collapsed sampler, VB and VB/EP. This indicates that VB/EP achieves the most significant performance increase over VB for indicator variables that are highly ranked.

6.2 Real data

Having demonstrated the potential utility of the VB/EP hybrid inference method on larger scale problems, we also carried out a performance comparison on a real biological network infernece problem. This problem concerned inference of the parameters of the transcription network of the baker's yeast *Saccharomyces cerevisiae*. The data were combined from two different sources. The data matrix $\mathbf{Y}$ consisted of microarray measurements of gene expression for 6217 genes under 205 varied experimental conditions from the study of Mnaimneh et al. (2004). These data were normalised to have zero mean and unit variance across all genes. The informative prior over the elements of the connectivity matrix Z was constructed, as described in Section D, from the results of high throughput *ChIP-chip* experiments by C. T. Harbison *et al* (2004). In total, this prior network consisted of 203 latent factors (transcription factors).



(a) Regression against MCMC

(b) ROC at convergence

Figure 5: **Comparison of the marginal posterior distribution of $\mathbf{W} \cdot \mathbf{Z}$ sampler versus VB and VB/EP. a)** Regression of VB and VB/EP against MCMC marginals. **b)** ROC curves of marginal $\mathbf{Z}$ MCMC,VB and VB/EP at convergence. The results indicate that VB/EP inference results are in greater accordance with sampling estimates than those obtained from the pure VB model.

As there is no reliable gold-standard for the true network structure available, we compared alternative methods by means of a *fill-in test*.

Fill-in Test

In this predictive test we trained each candidate model on the full set of genes and 95% of the experimental conditions. For the 5% fo the conditions not used for training, we removed a fraction ρ of the expression measurements and applied the trained model to fill-in these missing values. The motivation behind this experiment is that models which better capture the true network structure will be able to better predict the missing expression levels.

Figure 6 shows the mean squared error of the fill-in task for different fractions missing data ρ and alternative methods. We compared the two best deterministic models on simulated data and the collapsed Gibbs sampler with either 750 samples (runtime: XX, fraction of unconverged parameters: YY) or ZZZ samples (runtime: XXX, fraction of unconverged parameters: YY). The results show, that on this real-world sized problems, the deterministic approaches achieve a significantly better fill-in performance at a fraction of the computational costs.

Strictly, the posterior over the model parameters inferred from the training runs should be used as a prior for the test runs so that their distributions could also be influenced by the test data. However, such an approach is difficult to implement for the samplers. Instead, the distributions of these parameters were assumed to be known from the training runs and were simply sampled uniformly from the samples collected in the training run.

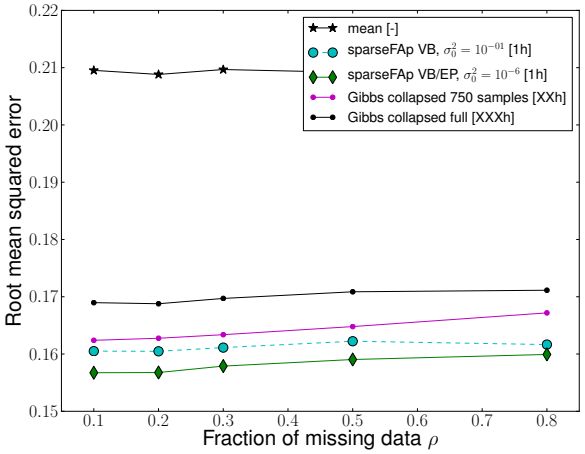


Figure 6: **Fill-in results on yeast dataset.** The prediction error for filling-in a fraction of missing values on an independent test dataset. Mean denotes the error when solely fitting the mean effect.

7. Discussion

In this work, we investigated alternative inference approaches for sparse factor analysis in the context of strong prior information. We considered both – approaches based on MCMC sampling and deterministic approximate inference techniques.

The empirical investigation on simulated and real datasets shows a tradeoff between accuracy and efficiency. While on small problems, sampling is feasible and yields gold-standard accuracy, MCMC fails short on larger problems. In this regime, deterministic approximations reach useful solutions in a fraction of the CPU runtime, allowing for real-world applications such as the regulatory network inference problem considered in Section 6.2.

References

C. Archambeau and F. Bach. Sparse probabilistic projections. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 73–80. 2009.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

C. T. Harbison *et al*. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431 (7004):99–104, 2004.

C.M. Carvalho, J. Chang, J.E. Lucas, J.R. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

C.M. Carvalho, N.G. Polson, and J.G. Scott. Handling sparsity via the horseshoe. *Journal of Machine Learning Research*, W&CP 5:73–80, 2009.

J Diebolt and C P Robert. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society Series B*, 56:363–376, 1994.

B. J. Frey, R. Patrascu, T. Jaakkola, and J. Moran. Sequentially fitting inclusive trees for inference in noisy-OR networks. In *Advances in Neural Information Processing Systems*, volume 12, pages 493–499, 2000.

S. Fruhwirth-Schnatter. MCMC estimation of classical and dynamic switching and mixture models. working paper. *Journal of the American Statistical Association*, 96(493):194–209, 1998.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472, 1992.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721741, 1984.

E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal-Amerial Statistical Association*, 88:881–881, 1993.

A. J. Hartemink. Reverse engineering gene regulatory networks. *Nat Biotech*, 23:554–555, May 2005.

A. Jasra, C. C. Holmes, and D. A. Stephens. Mcmc and the label switching problem in bayesian mixture models. *Statistical Science*, 20:50–67, 2005.

M. Kuss, T. Pfingsten, L. Csato, and C. E. Rasmussen. Approximate inference for robust Gaussian process regression. Technical report, Max Planck Institute for Biological Cybernetics, Tubingen, 2005.

J. C. Liao, R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury. Network Component Analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 100 (26), 2003.

T. Minka. Divergence measures and message passing. Technical report, Microsoft Research, 2005.

T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001a.

T. P. Minka. Expectation Propagation for approximate bayesian inference. In *Uncertainty in Artificial Intelligence*, volume 17, pages 362–369, 2001b.

T.J. Mitchell and J.J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032, 1988.

S. Mnaimneh, A. P. Davierwala, J. Haynes, J. Moffat, W. T. Peng, W. Zhang, X. Yang, J. Pootoolal, G. Chua, and A. Lopez. Exploration of essential gene functions via titratable promoter alleles. *Cell*, 118:31–44, 2004.

H. Nickisch and C. E. Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, 2008.

U. Paquet. *Bayesian inference for latent variable models*. PhD thesis, University of Cambridge, 2008.

I. Pournara and L. Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *feedback*, 2007.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, December 2006.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

C. Sabatti and G. M. James. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22(6):739–746, 2006.

M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *The Journal of Machine Learning Research*, 9:759–813, 2008.

C. Sigg and J. Buhmann. Expectation-maximization for sparse and non-negative PCA. In Andrew McCallum and Sam Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 960–967. 2008.

N. Srebro and T. Jaakkola. Weighted low-rank approximations. In *Machine Learning-International Workshop then Conference-*, volume 20, page 720, 2003.

M. C. Teixeira, P. Monteiro, P. Jain, S. Tenreiro, A. R. Fernandes, N. P. Mira, M. Alenquer, A. T. Freitas, A. L. Oliveira, and I. Sá-Correia. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Research*, 34:D3–D5, 2006.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591, 1991.

M. Welling, Y. W. Teh, and B. Kappen. Hybrid Variational/Gibbs collapsed inference in topic models. In *Uncertainty in Artificial Intelligence*, 2008.

M. West. Bayesian factor regression models in the 'Large p, Small n' paradigm. *Bayesian Statistics*, 7:723–732, 2003.

P.M. Williams. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7:117–143, 1995.

J. Winn and C. M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6(1):661, 2006.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 1:265, 2006.

Appendix A. Conditional Distribution of X,$\boldsymbol{\tau}$ and σ_1^2

For X we find that the posterior factorises into a product of Gaussian distributions:

$$P(X \mid .) \propto \prod_{j=1}^{J} \mathcal{N}\left(\mathbf{x}_j \mid \mathbf{M}_{x_j}, \mathbf{\Sigma}_{x_j}\right) \tag{46}$$

where $\mathbf{M}_{x_j}$ and $\mathbf{\Sigma}_{x_j}$, the mean and covariance of the distribution, are given by:

$$\mathbf{\Sigma}_{x_j} = \left(\tau_g \sum_{g=1}^{G} \mathcal{X}_g \mathbf{w}_g \mathbf{w}_g^{\mathrm{T}} \mathcal{X}_g + \mathbf{I}\right)^{-1} \tag{47}$$

$$\mathbf{M}_{x_j} = \mathbf{\Sigma}_{x_j} \left(\tau_g \sum_{g=1}^{G} y_{gj} \mathcal{X}_g \mathbf{w}_g\right), \tag{48}$$

where $\mathcal{X}_g = \mathrm{diag}(\mathbf{z}_g)$. This ensures that only those $w_{g,k}$ for which the current state of the corresponding $z_{g,k}$ is 1 contribute to the calculation.

For each $\boldsymbol{\tau}_g$ we find that the posterior is gamma distributed:

$$\boldsymbol{\tau}_g \sim \Gamma\left(\boldsymbol{\tau}_g \mid \alpha_g^*, \beta_g^*\right) \tag{49}$$

with

$$\begin{aligned} \alpha_g^* &= \frac{J}{2} + \alpha_g \\ \beta_g^* &= \beta_g + \frac{1}{2} \sum_{j=1}^{J} \left(y_{gj} - \sum_{k=1}^{K} z_{g,k} w_{g,k} x_{kj}\right)^2 \end{aligned} \tag{50}$$

With a vague, conjugate gamma prior over the precision σ_1^{-2}, the conditional posterior is also gamma distributed:

$$\frac{1}{\sigma_1^2} \sim \Gamma\left(\frac{1}{\sigma_1^2} \mid \alpha_\sigma^*, \beta_\sigma^*\right)$$

with

$$\begin{aligned} \alpha_\sigma^* &= \frac{|Z|}{2} + \alpha_\sigma \\ \beta_\sigma^* &= \beta_\sigma + \frac{1}{2} \sum_{g=1}^{G} \sum_{k=1}^{K} \left(z_{g,k} w_{g,k}\right)^2 \end{aligned} \tag{51}$$

where $|Z|$ represents the cardinality of the set of nono-zero indicator variables; ($\alpha_\sigma = 0.7$,$\beta_\sigma = 0.3$).

Appendix B. Full update rules for VB in sparseFA model

This section provides the full update rules of standard VB learning of factor analysis. Starting form the chosen factorisation of the approximation (Equation 26), update can be obtained by substituting VB factors into Equation (25).

This leads to the following functional forms and update rules of the approximate factors:

(Approximate distributions)

$$Q(\mathbf{X}) = \prod_{j=1}^{J} \mathcal{N}\left(\mathbf{x}_j \,\middle|\, \tilde{\mathbf{m}}_{\mathbf{x}_j}, \tilde{\Sigma}_{\mathbf{x}_j}\right) \tag{52}$$

$$Q(\boldsymbol{\tau}) = \prod_{g=1}^{G} \Gamma(\tau_g \,|\, \tilde{a}_{\tau_g}, \tilde{b}_{\tau_g}) \tag{53}$$

(Update rules)

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_j} &= \left(\mathbf{I} + \left\langle \mathbf{W}^T \mathrm{diag}\left(\boldsymbol{\tau}\right) \mathbf{W} \right\rangle\right)^{-1} \\ \tilde{\mathbf{m}}_{\mathbf{x}_j} &= \tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_j} \left[\left\langle \mathbf{W}^{\mathrm{T}} \right\rangle \mathrm{diag}\left\langle \boldsymbol{\tau} \right\rangle \left(\mathbf{y}_j\right)\right] \end{aligned} \tag{54}$$

$$\begin{aligned} \tilde{a}_{\tau_g} &= a_\tau + \frac{1}{2} \sum_{j=1}^{J} \left\langle \left(y_{g,j} - \mathbf{w}_g \mathbf{x}_j\right)^2 \right\rangle \\ \tilde{b}_{\tau_g} &= b_\tau + \frac{J}{2}. \end{aligned} \tag{55}$$

Appendix C. Moments for EP model

The required moments for the EP updates correspond to moment matching equations of a mixture model with two mixing components ($C = 1$).

$$\begin{aligned} F_C &= \sum_{c=0}^{C} \pi_c \int_{f_i} \mathcal{N}\left(f_i \,\middle|\, \tilde{\mu}_{\setminus i}, \tilde{\nu}^2_{\setminus i}\right) \mathcal{N}\left(f_i \,\middle|\, t_i, \sigma_c^2\right) \, \mathrm{d}f_i \\ &= \sum_{c=0}^{C} \pi_c \underbrace{\mathcal{N}\left(\tilde{\mu}_{\setminus i} \,\middle|\, t_i, \tilde{\nu}^2_{\setminus i} + \sigma_c^2\right)}_{Z_c}. \end{aligned} \tag{56}$$

The first moment is

$$\begin{aligned} F_\mu &= \frac{1}{F_C}\sum_{c=0}^{C}\pi_c\int_{f_i}\mathcal{N}\left(f_i\,\middle|\,\tilde{\mu}_{\setminus i},\tilde{\nu}^2_{\setminus i}\right)\mathcal{N}\left(f_i\,|\,t_i,\sigma_c^2\right)f_i\,\mathrm{d}f_i \\ &= \frac{1}{F_C}\sum_{c=0}^{C}\pi_c Z_c\int_{f_i}\mathcal{N}\left(f_i\,|\,u_c,v_c^2\right)f_i\,\mathrm{d}f_i \\ &= \frac{1}{F_C}\sum_{c=0}^{C}\pi_c Z_c u_c, \end{aligned} \tag{57}$$

and similarly

$$\begin{aligned} F_{\sigma^2}+F_\mu^2 &= \frac{1}{F_C}\sum_{c=0}^{C}\pi_c\int_{f_i}Z_c\mathcal{N}\left(f_i\,|\,u_c,v_c^2\right)f_i^2\,\mathrm{d}f_i \\ &= \frac{1}{F_C}\sum_{c=0}^{C}\pi_c Z_c\left[u_c^2+v_c^2\right]. \end{aligned} \tag{58}$$

We defined $Z_c = \mathcal{N}\left(\tilde{\mu}_{\setminus i}\,\middle|\,t_i,\tilde{\nu}^2_{\setminus i}+\sigma_c^2\right)$ and the relations $v_c^2 = \left(\tilde{\nu}^{-2}_{\setminus i}+\sigma_c^{-2}\right)^{-1}$ and $u_c = v^2\left(\frac{\tilde{\mu}_{\setminus i}}{\tilde{\nu}^2_{\setminus i}}+\frac{t_i}{\sigma_c^2}\right)$.

An alternative derivation of the moment equations can be found in Kuss et al. (2005), who derived these moment matching equations in the context of robust Gaussian process regression.

Appendix D. Connectivity prior for yeast data

The connectivity prior used for inference on the real data set in section 6.2 was derived from the ChIP-chip study of C. T. Harbison *et al* (2004). Based on a combination of more restricted, follow-up experiments and literature mining the following FPR and FNR estimates were provided:

$$\text{Estimated FP rate} - \Pr\left(\tilde{z}_{g,k}=1|z_{g,k}=0\right) \sim 3.7/6500 \tag{59}$$

$$\text{Estimated FN rate} - \Pr\left(\tilde{z}_{g,k}=0|z_{g,k}=1\right) \sim 0.3 \tag{60}$$

$$\text{Confirmatory FP rate} - \Pr\left(z_{g,k}=0|\tilde{z}_{g,k}=1\right)) \sim 0.06 \tag{61}$$

where $z_{g,k}$ represents the truth about the presence or absence of a regulatory link and $\tilde{z}_{g,k}$ represents whether or not such a link was 'observed' based on thresholded ChIP data.

These can be used to determine the required parameters, $\pi_{g,k}$:

$$\Pr\left(z_{g,k}=1|\tilde{z}_{g,k}=1\right) = 1-\Pr\left(z_{g,k}=0|\tilde{z}_{g,k}=1\right)) \tag{62}$$

$$\begin{aligned} \Pr\left(z_{g,k}=1|\tilde{z}_{g,k}=0\right) &= \Pr\left(\tilde{z}_{g,k}=0|z_{g,k}=1\right)\Pr\left(z_{g,k}=1\right)/\Pr\left(\tilde{z}_{g,k}=0\right) \\ &= \frac{1}{1+\left(1-P_1\right)P_3\left(1-P_2\right)/P_2P_1\left(1-P_3\right)} \end{aligned} \tag{63}$$

where P_1, P_2 and P_3 are given by 59, 60 and 61 respectively. This leads to the following estimates:

$$\begin{aligned} \Pr\left(z_{g,k}=1|\tilde{z}_{g,k}=1\right) &= 0.94 \\ \Pr\left(z_{g,k}=1|\tilde{z}_{g,k}=0\right) &= 0.0038 \end{aligned} \tag{64}$$

which were consistent with estimates of the marginal probabilities $\Pr\left(\tilde{z}_{g,k}=0\right)$ and $\Pr\left(\tilde{z}_{g,k}=1\right)$ based on the observed frequencies of binding relationships in the experiment.