JMLR: Workshop and Conference Proceedings 39:330–343, 2014 ACML 2014

Variational Gaussian Inference for Bilinear Models of Count Data

Young-Jun Ko YOUNGJUN.KO@EPFL.CH

Mohammad Emtiyaz Khan EMTIYAZ.KHAN@EPFL.CH

École Polytechnique Fédérale de Lausanne, Switzerland

Editor: Dinh Phung and Hang Li

Abstract

Bilinear models of count data with Poisson distribution are popular in applications such as matrix factorization for recommendation systems, modeling of receptive fields of sensory neurons, and modeling of neural-spike trains. Bayesian inference in such models remains challenging due to the product term of two Gaussian random vectors. In this paper, we propose new algorithms for such models based on variational Gaussian (VG) inference. We make two contributions. First, we show that the VG lower bound for these models, previously known to be intractable, is available in closed form under certain non-trivial constraints on the form of the posterior. Second, we show that the lower bound is bi-concave and can be efficiently optimized for mean-field approximations. We also show that bi-concavity generalizes to the larger family of log-concave likelihoods, that subsume the Poisson distribution. We present new inference algorithms based on these results and demonstrate better performance on real-world problems at the cost of a modest increase in computation. Our contributions in this paper, therefore, provide more choices for Bayesian inference in terms of a speed-vs-accuracy tradeoff.

Keywords: Variational Gaussian inference, bilinear models, Poisson likelihood, matrix factorization, latent Gaussian models

1. Introduction

Latent Gaussian factor models, such as probabilistic principal component analysis (PPCA) and factor analysis (FA), are very commonly used density models for continuous-valued data. They are extensively employed in various applications such as latent factor discovery, dimensionality reduction, missing-data imputation, and data fusion.

Such latent factor models can be easily extended to handle non-Gaussian data using the generalized linear model framework (McCullagh and Nelder, 1989) where Gaussian likelihoods are replaced by other distributions (see e.g. Mohamed et al. (2008); Seeger and Bouchard (2012); Khan et al. (2010)). In this paper, we focus on the modeling of count data using latent factor models with Poisson likelihoods. This is motivated by the fact, that counting the occurence of events is a natural mode of observing phenomena. Gathering such data leads to the problem of analysing non-negative and discrete-valued random processes, for which the use of Gaussian likelihoods for computational convenience can lead to inaccuracies, due to undesirable properties, such as symmetry and the distribution of mass over the whole real line. Therefore, latent Gaussian factor models endowed with Pois-

son likelihoods have been successfully used in various real-world applications, examples of which are given in what follows. In neuroscience, (Park and Pillow, 2013) model the receptive field of a sensory neuron by a low-rank latent Gaussian field serving as the intensity function of an inhomogeneous Poisson process, and demonstrate, that Poisson likelihoods model neural spike counts, much more accurately than Gaussians. (Buesing et al., 2012) analyse multi-electrode recordings of neural activity using a similar model for neural spike trains, where the latent Gaussian field captures temporal dependencies.. (Krishnamurthy et al., 2010) analyse counts of newly infected individuals over space and time by describing the spatio-temporal dynamics as a latent Gaussian field. (Seeger and Bouchard, 2012; Zhou et al., 2012) apply such models to the task of imputing incomplete count matrices from different sources, ranging from transportation data to telecommunications to corpora of scientific publications.

Such applications often routinely rely on Bayesian posterior inference for robustness and uncertainty quantification. Bayesian inference, however, is intractable in latent factor models with Poisson likelihoods, due to the non-conjugacy of the Poisson distribution to the Gaussian prior over the latent factors. A variety of approximation methods can be used. One of the most common and computationally attractvive methods, is to approximate the posterior by a maximum-a-posteriori (MAP) point estimate, although being known to be prone to overfitting and thus requiring careful regularization (Welling et al., 2008; Salakhutdinov and Mnih, 2008a). Obtaining exact posterior samples is the goal of Markov-Chain Monte-Carlo (MCMC) (Zhou et al., 2012). But this method can be slow and is notoriously hard to diagnose in terms of convergence (Salakhutdinov and Mnih, 2008a). Naive determinstic approaches such as the Laplace approximation do not capture skewed posteriors well (Kuss and Rasmussen, 2005), while approaches, such as expectation-propagation, can pose numerical challenges, often requiring accurate quadrature methods for convergence (Yu et al., 2006).

In this paper, we focus on a variational Gaussian (VG) approach that assumes the posterior to be Gaussian (Opper and Archambeau, 2009) and therefore extend the work of (Lim and Teh, 2007) by dealing with the complications arising due to non-Gaussian likelihoods.. The Gaussian posterior can be found by optimizing a lower bound on the marginal likelihood. For non-conjugate models, this lower bound is generally intractable, requiring quadrature or other approximation techniques, that can lead to a loss in accuracy (e.g. discussed in Seeger and Bouchard (2012) for the Poisson distribution).

Our first contribution is to show that the lower bound can be, in fact, computed tractably. We obtain an analytical expression for the lower bound and show that it implies non-trivial constraints on posterior parameters. Our second contribution is to show that the lower bound is bi-concave w.r.t. to the posterior distribution of individual latent factors. This allows us to design fast, convergent algorithms. We show empirically, that the proposed approach performs better than other approximation methods at the cost of a modest increase in computation. Our approach, therefore, offers a set of alternatives, that trade off speed versus accuracy, from which practitioners can choose according to their requirements.

2. The Latent Factor Model

We consider the count matrix $\mathbf{Y}$ of size $M \times N$, where M is the *data-dimensionality* and N is the number of *data-observations*. For example, in recommendation systems, N is the number of users and M is the number of songs, available to the users. Each entry of $\mathbf{Y}$, denoted by y_{in}, then represents the number of times the n'th user has listened to the i'th song. We allow for entries of $\mathbf{Y}$ to be missing and denote the set of observed indices by $\mathbb{O} \subseteq \{1, \ldots, M\} \times \{1, \ldots, N\}$, and generally assume that $\mathbf{Y}$ is observed sparsely, i.e. $|\mathbb{O}| \ll MN$. We will denote the count vector for the n'th user by $\mathbf{y}_n = \{y_{in} : (i, n) \in \mathbb{O}\}$, and the count vector for the i'th song by $\mathbf{y}_i$, accordingly. For illustrative purposes, we will sometimes use the terminology of recommendation systems and refer to data-dimensions as *items* and data-examples as *users*. Furthermore, we denote the set of songs, that the user n has listened to, by $\mathbb{O}_n = \{i \in \{1, \ldots, M\} : (i, n) \in \mathbb{O}\}$ and, with a slight abuse of notation, the set of users who have listened to the i'th song by $\mathbb{O}_i = \{n \in \{1, \ldots, N\} : (i, n) \in \mathbb{O}\}$.

Each y_{in} is modeled as a Poisson random variable governed by a latent intensity $\lambda_{in} > 0$. Heavily underdetermined due to the sparse-observation assumption, an information-sharing mechanism needs to be imposed on these intensities, to be able to reason about missing values. This is achieved by postulating a generalized bilinear latent factor model, introducing log-bilinear predictor variables $\eta_{in} = \log(\lambda_{in})$. Specifically, the predictor for each pair (i, n) is modeled by the inner product of two D-dimensional real-valued vectors $\mathbf{w}_i$ and $\mathbf{z}_n$, as shown in Eq. (1). Such a bilinear representation underlies many models for matrix factorization and dimensionality reduction, such as probabilistic PCA/factor analysis (Tipping and Bishop, 1999) or Bayesian Matrix Factorization (Salakhutdinov and Mnih, 2008a). For simplicity, we have ignored the user- and item-bias terms, which can be easily added and were included in our implementation.

Given the vector of linear predictors $\boldsymbol{\eta}_n$ for the n'th user, the data vector $\mathbf{y}_n$ is sampled from a Poisson distribution as shown in Eq. (2), independently for each user.

$$\eta_{in} = \mathbf{w}_i^T \mathbf{z}_n \tag{1}$$

$$p(\mathbf{y}_n | \boldsymbol{\eta}_n) = \prod_{i \in \mathbb{O}_n} \frac{1}{y_{in}!} \exp\left(\eta_{in} y_{in} - e^{\eta_{in}}\right) \tag{2}$$

Note, that λ_{in} is parameterized as $\exp(\eta_{in})$ to fulfill the requirement that $\lambda_{in} > 0$.

For dimensionality-reduction, we assume that $D \ll \min\{M, N\}$ and that $\mathbf{w}_i$ and $\mathbf{z}_n$ follow Gaussian prior distributions, defined below.

$$p(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n | 0, \sigma_z^2 \mathbf{I}_D) \quad , \quad p(\mathbf{w}_i) = \mathcal{N}(\mathbf{w}_i | 0, \sigma_w^2 \mathbf{I}_D) \tag{3}$$

We define matrices $\mathbf{Z} \in \mathbb{R}^{D \times N}$ and $\mathbf{W} \in \mathbb{R}^{D \times M}$, whose columns consist of the latent vectors $\mathbf{z}_n$ and $\mathbf{w}_i$, respectively, and denote the set of parameters by $\theta = \{\sigma_z^2, \sigma_w^2, D\}$.

Bayesian inference, where the goal is to get a handle on objects, such as the posterior distribution (Eq. 4) or the marginal likelihood (Eq. 5), is based on high-dimensional integration, intractable in this context due to non-conjugacy.

$$p(\mathbf{Z}, \mathbf{W} | \mathbf{Y}, \theta) = \frac{p(\mathbf{Y} | \mathbf{W}, \mathbf{Z}) p(\mathbf{W} | \theta) p(\mathbf{Z} | \theta)}{p(\mathbf{Y} | \theta)} \tag{4}$$

$$p(\mathbf{Y} | \theta) = \int_Z \int_W p(\mathbf{Y} | \mathbf{W}, \mathbf{Z}) p(\mathbf{W} | \theta) p(\mathbf{Z} | \theta) d\mathbf{W} d\mathbf{Z} \tag{5}$$

Next, we describe the variational Gaussian approximation method and show the feasibility of the key computational element required to drive this method.

3. Tractable Variational Gaussian (VG) Inference

Dropping the dependence on the parameters θ for notational convenience, let $q(\mathbf{Z}, \mathbf{W})$ denote the approximate posterior. Given this posterior, we can obtain a lower bound on the marginal likelihood by introducing $q(\mathbf{Z}, \mathbf{W})$ as shown on the right hand side of Eq. (6), and then using Jensen's inequality, as shown in Eq. (7). In Eq. (8), it is rewritten in terms of the Kullback-Leibler (KL) divergences between the posterior and prior and sum over posterior expectations of data-log-likelihoods of the observed y_{in}.

$$\log p(\mathbf{Y}) = \log \int_Z \int_W \frac{p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}) p(\mathbf{W}) p(\mathbf{Z})}{q(\mathbf{Z}, \mathbf{W})} \, q(\mathbf{Z}, \mathbf{W}) \, d\mathbf{W} d\mathbf{Z} \tag{6}$$

$$\geq \mathbb{E}_{q(Z,W)} \left[\log \frac{p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}) p(\mathbf{W}) p(\mathbf{Z})}{q(\mathbf{Z}, \mathbf{W})} \right] \tag{7}$$

$$= -D_{KL} \left[q(\mathbf{W}, \mathbf{Z}) || p(\mathbf{W}) p(\mathbf{Z}) \right] + \sum_{n=1}^{N} \sum_{i \in O_n} \mathbb{E}_{q(\mathbf{z}_n, \mathbf{w}_i)} [\log p(y_{in}|\eta_{in})] \tag{8}$$

The VG approximation assumes the variational distribution $q(\mathbf{Z}, \mathbf{W})$ to be Gaussian, rendering the first term tractable, because the KL divergence between two Gaussian distributions has an analytic expression. Additionally, we focus on distributions of the following form.

$$q(\mathbf{Z}, \mathbf{W}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{z}_n | \mathbf{m}_n, \mathbf{V}_n) \prod_{i=1}^{M} \mathcal{N}(\mathbf{w}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{9}$$

Given this factorizaion, the KL term (up to a constant) decomposes in the following way.

$$\tfrac{1}{2} \sum_{n=1}^{N} \left[\log |\mathbf{V}_n| - \frac{1}{\sigma_z^2} \mathrm{Tr}(\mathbf{V}_n) - \mathbf{m}_n^T \mathbf{m}_n \right] + \tfrac{1}{2} \sum_{i=1}^{M} \left[\log |\boldsymbol{\Sigma}_i| - \frac{1}{\sigma_w^2} \mathrm{Tr}(\boldsymbol{\Sigma}_i) - \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i \right] \tag{10}$$

The difficulty with non-conjugate models stems from the likelihood terms in Eq. (8). We now show that, for our model, this term has a closed-form expression. The likelihood terms can be simplified, as shown in Eq. (11), by first substituting the definition of the Poisson distribution from Eq. (2), and then applying the identity $\mathbb{E}[\exp\left(\mathbf{t}^T \mathbf{x}\right)] = \exp(\mathbf{t}^T \mathbf{m} + \frac{1}{2} \mathbf{t}^T \mathbf{V} \mathbf{t})$ for $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{V})$ and a given vector $\mathbf{t}$ (shown in Appendix A), to get Eq. (12).

$$\mathbb{E}_{q(\boldsymbol{z}_n, \boldsymbol{w}_i)} [\log p(y_{in}|\eta_{in})] = \mathbb{E}_{q(\boldsymbol{w}_i)} \left[\mathbb{E}_{q(\boldsymbol{z}_n)} \left(y_{in} \mathbf{w}_i^T \mathbf{z}_n - e^{\mathbf{w}_i^T \mathbf{z}_n} \right) \right] + \text{cnst} \tag{11}$$

$$= \mathbb{E}_{q(\boldsymbol{w}_i)} \left[y_{in} \mathbf{w}_i^T \mathbf{m}_n - e^{\mathbf{m}_n^T \mathbf{w}_i + \frac{1}{2} \mathbf{w}_i^T \mathbf{V}_n \mathbf{w}_i} \right] + \text{cnst} \tag{12}$$

We show in Appendix A, that the expectation in Eq. (12) can be written as

$$y_{in} \boldsymbol{\mu}_i^T \mathbf{m}_n - |\mathbf{S}_{in}|^{-\frac{1}{2}} \exp\left(\tfrac{1}{2} \boldsymbol{\nu}_{in}^T \mathbf{B}_{in}^{-1} \boldsymbol{\nu}_{in} - \tfrac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i \right) \tag{13}$$

where $\mathbf{S}_{in} = \mathbf{I} - \mathbf{\Sigma}_i \mathbf{V}_n$, $\mathbf{B}_{in} = \left(\mathbf{\Sigma}_i^{-1} - \mathbf{V}_n\right)$ and $\boldsymbol{\nu}_{in} = \mathbf{m}_n + \mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i$.

Note, that the expression is symmetric with respect to the posterior of $\mathbf{z}_n$ and $\mathbf{w}_i$, as expected. More importantly, the identity only holds under the constraint that $\mathbf{S}_{in}$ is positive definite, i.e. the expectation is real-valued if $\mathbf{I} - \mathbf{\Sigma}_i \mathbf{V}_n \succ 0$. For illustration, consider the mean-field approximation assuming diagonal $\mathbf{V}_n$ and $\mathbf{\Sigma}_i$. In this case, the constraints imply that $V_{n,dd}\Sigma_{i,dd} < 1$ for all diagonal elements d. This is not a convex set for $(V_{n,dd}, \Sigma_{i,dd})$, but given $V_{n,dd}$, the constraint on $\Sigma_{i,dd}$ is simply a bound constraint, and vice-versa. Similarly, for full covariance matrices, these become positive-definite constraints.

The final lower bound can be written as follows:

$$\phi(\boldsymbol{\xi}) = \tfrac{1}{2}\sum_{n=1}^{N}\left[\log|\mathbf{V}_n| - \frac{1}{\sigma_z^2}\mathrm{Tr}(\mathbf{V}_n) - \mathbf{m}_n^T\mathbf{m}_n\right] + \tfrac{1}{2}\sum_{i=1}^{M}\left[\log|\mathbf{\Sigma}_i| - \frac{1}{\sigma_w^2}\mathrm{Tr}(\mathbf{\Sigma}_i) - \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i\right] \quad (14)$$

$$+ \sum_{n=1}^{N}\sum_{i\in\mathbb{O}_n} y_{in}\boldsymbol{\mu}_i^T\mathbf{m}_n - |\mathbf{S}_{in}|^{-\frac{1}{2}}\exp\left(\tfrac{1}{2}\boldsymbol{\nu}_{in}^T\mathbf{B}_{in}^{-1}\boldsymbol{\nu}_{in} - \tfrac{1}{2}\boldsymbol{\mu}_i^T\mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i\right) \quad (15)$$

where we denote the set of all variational parameters by $\boldsymbol{\xi} = \{\mathbf{m}_n, \mathbf{V}_n, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i\}_{n=1,\ldots,N, i=1,\ldots,M}$.

The variational inference problem is to maximize this lower bound subject to the constrained described previously.

$$\max_{\boldsymbol{\xi}} \quad \phi(\boldsymbol{\xi}) \quad (16)$$

$$\text{s.t. } \mathbf{I} - \mathbf{\Sigma}_i\mathbf{V}_n \succ 0 \quad \forall(i,n) \in \mathbb{O} \quad (17)$$

4. Bi-Concavity of VG Objective

In this section, we discuss concavity of the final lower bound of Eq. (14) and show that this leads to efficient inference for mean-field aproximations. The following theorem establishes the bi-concavity of the lower bound.

Theorem 1 *The lower bound of Eq. (14) is concave with respect to* $\{\mathbf{m}_n, \mathbf{V}_n\}$ *for all* n *given* $\{\boldsymbol{\mu}_i, \mathbf{\Sigma}_i\}$ *for all* i.

Proof The first line of Eq. (14) is concave since the KL divergence is convex w.r.t. its arguments. To prove the convexity of the second term we first note that in Eq. (12) the term inside the expectation is concave w.r.t $\{\mathbf{m}_n, \mathbf{V}_n\}$ given $\mathbf{w}_i$ (Khan et al., 2013). Since convexity is preserved under summation with positive weights, the expectation is also concave, making Eq. (12) concave and proving the concavity of the final lower bound. ■

The above result can be generalized to a bigger class of models, as shown in the following theorem.

Theorem 2 *Let* $\mathbf{T}_n, \mathbf{L}_i$ *denote the Cholesky factors of* $\mathbf{V}_n, \mathbf{\Sigma}_i$, *respectively. If* $\log p(y_{in}|\eta_{in})$ *is concave w.r.t.* η_{in}, *then Eq. (14) is concave with respect to* $\{\mathbf{m}_n, \mathbf{T}_n\}$ *for all* n *given* $\{\boldsymbol{\mu}_i, \mathbf{L}_i\}$ *for all* i.

	MAP	MF	EM	VB
Storage	$O(D(M+N))$	$O(D(M+N))$	$O(DM+D^2N)$	$O(D^2(M+N))$
Computation	$O(DN_{obs})$	$O(DN_{obs})$	$O(D^2N_{obs})$	$O(D^3N_{obs})$

Table 1: Complexity comparison of methods. Complexity increases from left to right. MAP and EM are existing methods, while MF and VB are proposed methods. $N_{obs} = \sum_n |\mathbb{O}_n|$ is the total number of observations.

Proof The proof is analogous to the previous theorem. The KL term is also convex w.r.t. the Cholesky factor. The convexity of the likelihood term follows from the results of Challis and Barber (2011), who show that the following is concave wrt $(\mathbf{m}_n, \mathbf{T}_n)$ for each $\mathbf{w}_i$.

$$\mathbb{E}_{q(\boldsymbol{z}_n)}\left[\log p\left(y_{in}|\mathbf{w}_i^T\mathbf{z}_n\right)\right] \tag{18}$$

Given this, the expectation w.r.t. $q(\mathbf{w}_i)$ is also concave since it is an integral with positive weights. ■

As a consequence, we can write the variational problem in Eq. (16) in terms of $\hat{\boldsymbol{\xi}} = \{\mathbf{m}_n, \mathbf{T}_n, \boldsymbol{\mu}_i, \mathbf{L}_i\}_{n=1,\ldots,N,i=1,\ldots,M}$, which retains the bi-convexity, but simplifies the log-determinant terms due to the prior.

5. Summary of Methods and Implementation Details

We consider two variants of the proposed VG method. The first method is the approximation described in Eq. (9), which we refer to as the variational-Bayes (VB) approximation. The second variant is the mean-field (MF) approximation where the covariances $\mathbf{V}_n$ and $\boldsymbol{\Sigma}_i$ are all assumed to be diagonal.

We compare these methods to two existing methods. The first method is the maximum-a-posterior (MAP) estimate where the posterior distribution is approximated by a dirac-delta distribution (Welling et al., 2008), that places all its mass at the mode of the posterior (or equivalently the mode of the joint distribution $p(\mathbf{Y}, \mathbf{W}, \mathbf{Z})$). Such an estimate is non-Bayesian since no uncertainty is represented. Comparing with this method therefore illustrates the benefits of being Bayesian.

The second method is a "partially" Bayesian method that treats one set of factors as latent variables and the other as parameters. This corresponds to the probabilistic PCA model (Tipping and Bishop, 1999), where the item factors $\mathbf{W}$ are treated as parameters, while $\mathbf{Z}$ is marginalized out. We expect this method to perform better than MAP, but not as good as the VB method. For learning, we use the expectation-maximization (EM) algorithm, similar to (Tipping and Bishop, 1999). We will refer to this approach as EM.

For all other methods, we will use a coordinate ascent approach, that alternates between optimizing $q(\mathbf{Z})$ given $q(\mathbf{W})$ and $q(\mathbf{W})$ given $q(\mathbf{Z})$. Due to their concavity, these subproblems can be solved robustly with reliable convergence diagnostics and fast convergence rates. We use an L-BFGS optimizer in our implementation. Both of our methods have constraints associated with them. In the case of MF, the contraints reduce to simple bound constraints that can be easily incorporated in a quasi-Newton method such as

L-BFGS (Bertsekas, 1999). For VB, the constraints are more complex, so we resort to a simple implementation where we augment the objective with a barrier for infeasible points, and let the line search back away from such solutions. In practice, our line search never encountered such infeasible points. We leave a careful implementation for the future.

We compare these algorithms in terms of their complexity in Table 1. There, the complexity increases from left to right. The storage complexity directly reflects the amount of posterior uncertainty, that is represented. MAP only keeps track of a point estimate of $\mathbf{W}, \mathbf{Z}$. MF additionally represents variances of all variables, effectively doubling the memory requirement. VB represents covariances over all $M + N$ latent factors, requiring $O(D^2)$ memory per factor. EM lies in between MAP and VB, in that covariances are only represented for one set of variables.

As all methods rely on first-order non-linear optimization, computational complexity scales with the number of data-likelihood terms N_{obs}, each of which being involved in the accumulation of the gradient. The methods only differ in the cost per term, which is benign due to the assumption on D being relatively small. For MAP, the contribution to the gradient of a single likelihood term is linear in D. For EM, likelihoods contribute terms like Eq. (12) with $q(\mathbf{w}_i) = \delta(\mathbf{w}_i)$. Thus, the gradient requires dense matrix-vector multiplication, that scales as $O(D^2)$. The fully Bayesian methods deal with terms of the form presented in Eq. (13). For VB, the dependence on matrix inverses leads to cubic scaling in D, while the mean-field approximation greatly simplifies these expressions, reducing the cost back to linear.

Our implementation is based on MATLAB, however we wrote some computational-demanding gradient-computations in C++ using the Armadillo linear algebra library (Sanderson, 2010). For our results, we used a linux server equipped with AMD Opteron 6380 CPUs and 512GB RAM.

6. Results

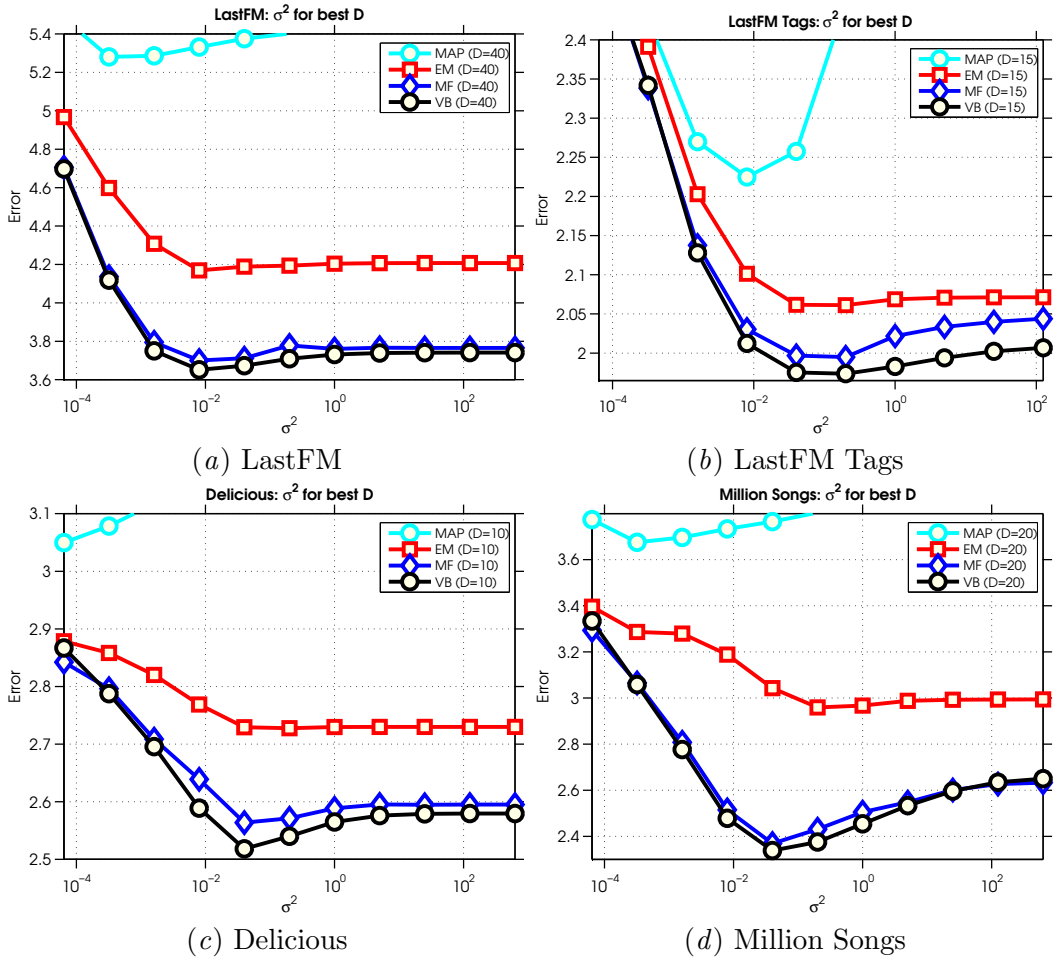
In this section, we compare all methods on real-world datasets. We focus on count datasets arising in recommendation systems. We use the posterior-predictive probability as our performance measure. Specifically, given a test observation y^*_{in}, we compute the (negative) logarithm of the following predictive distribution:

$$-\log p(y^*_{in}|\mathbf{Y}) = -\log \int p(y^*_{in}|\mathbf{z}_n, \mathbf{w}_i) q(\mathbf{z}_n, \mathbf{w}_i|\mathbf{Y}) d\mathbf{z}_n d\mathbf{w}_i \tag{19}$$

Thus, lower values indicate better performances, i.e. the method, that assigns higher probability to the value at a test location, incurs a lower error.

Since the above integral is intractable, we approximate it by a Monte Carlo estimate using 10^5 samples, ensuring a stable estimate. For MAP, the error measure reduces to a simple plug-in estimates. We report the average of this quantity over all the test examples.

We compare methods on four real-world datasets summarized in Table 2. For each data set, we randomly select 8000 observations, of which 25% are held out for testing. The LastFM and Delicious datasets can be downloaded from (grouplens, 2011). The LastFM-Tags and Million-Songs datasets can be obtained from (Lamere, 2008) and (Kaggle, 2012), respectively. To avoid numerical instabilities caused by very large counts, we apply the transformation $y \mapsto \lfloor \sqrt{y} + 0.5 \rfloor$ to LastFM and LastFM Tags.



(*a*) LastFM (*b*) LastFM Tags

(*c*) Delicious (*d*) Million Songs

Figure 1: Effect of prior variance: Strong regularization is necessary for MAP. While EM already enjoys the benefits of the regularization inherent in Bayesian methods, it does not perform on par with the fully Bayesian methods.

Name	M	N	N_{obs}	Description
LastFM	17,632	1,892	92,834	Listening counts of songs per user
LastFM Tags	20,907	100,784	952,707	Counts of tags assigned to artists
Delicious	38,603	1,867	93,210	Counts of webpages bookmarked per user
Million Songs	163,206	110,000	1,450,933	Listening counts of songs per user

Table 2: Details of datasets.

We fix $\sigma_z^2 = 1$ and choose σ_w^2 and D that minimize the error on 1500 validation samples.

We present the effect of σ_w^2 in Fig. 1. The MAP estimate overfits, while Bayesian approaches are more robust. Previous studies have shown similar trends (Salakhutdinov and Mnih, 2008b). This is holds for almost all of the datasets that we studied.

Fig. 2 compares the speed-accuracy trade-off of all the methods. Accuracy is measured using the error defined in Eq. (19), while speed is measured by the running time in seconds.

	MAP	EM	MF	VB
LastFM Tags	2.19 (0.06)	2.02 (0.03)	1.99 (0.02)	1.97 (0.02)
LastFM	5.94 (0.24)	4.65 (0.18)	3.89 (0.08)	3.81 (0.07)
Delicious	3.17 (0.06)	2.62 (0.01)	2.60 (0.01)	2.56 (0.02)
Million Songs	3.54 (0.10)	2.69 (0.07)	2.31 (0.05)	2.28 (0.04)

Table 3: Comparison of methods. We report the error measured by Eq. (19). The error is averaged over all test examples for 10 different train-test splits. The standard error is shown inside brackets. We clearly see that both of our methods, MF and VB, achieve the lowest error values.

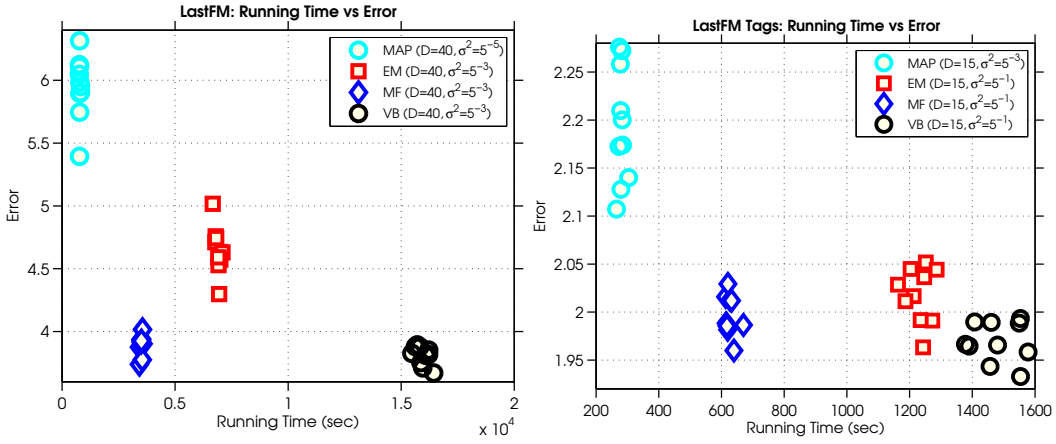
We select σ_w^2 and D by cross-validation, but to make a reasonable comparison of running time, we fix D to be the same for all the methods. The value of D that we choose, gives the same performance as the one chosen by cross-validation. We show the results for 10 different test-train splits in Fig. 2, and summarize them in Table 3 for clarity.

We observe that while VB exhibits the best performance overall, MF is a strong contender due to its speed and competitive performance. It is in the same complexity class as MAP, but slower due to a larger constant factor. Overall, our proposed method not only show improvements in both speed and accuracy, but also present more choices for Bayesian inference in terms of speed-accuracy trade-offs.

7. Conclusion

We showed that the variational Gaussian inference for bilinear latent Gaussian log-concave models can be written as a, potentially constrained, bi-concave maximization problem. For Poisson likelihoods, the objective is available in closed form and can be effectively optimized, rendering fully Bayesian inference feasible. We empirically verified the benefits of the Bayesian approach over non- or partially Bayesian methods.

Potential future directions include the extension to other datatypes and improved scalability by stochastic versions of the algorithms discussed here.



(*a*) LastFM (*b*) LastFM Tags

(*c*) Delicious (*d*) Million Songs

Figure 2: Speed-accuracy comparison. X-axis shows the running time, while Y-axis shows the error. For both, lower is better. We see that for all the datasets MF and VB achieve the lowest error. MF appears to be a good choice since it is also much faster than other Bayesian methods.

References

Dimitri P. Bertsekas. *Nonlinear programming*. Athena Scientific, second edition, 1999.

Lars Buesing, Maneesh Sahani, and Jakob H Macke. Spectral learning of linear dynamics from generalised-linear observations with application to neural population data. In *Advances in neural information processing systems*, pages 1682–1690, 2012.

E. Challis and D. Barber. Concave Gaussian Variational Approximations for Inference in Large-Scale Bayesian Linear Models. In *International conference on Artificial Intelligence and Statistics*, volume 6, page 7, 2011.

grouplens. HetRec2011, 2011. URL http://grouplens.org/datasets/hetrec-2011/.

Kaggle. Million Song Data Set Challenge, 2012. URL https://www.kaggle.com/c/msdchallenge/data.

Mohammad Emtiyaz Khan, Benjamin Marlin, Guillaume Bouchard, and Kevin Murphy. Variational Bounds for Mixed-Data Factor Analysis. In *Advances in Neural Information Processing Systems*, 2010.

Mohammad Emtiyaz Khan, Aleksandr Aravkin, Michael Friedlander, and Matthias Seeger. Fast Dual Variational Inference for Non-Conjugate Latent Gaussian Models. In *International Conference on Machine Learning*, 2013.

Ashok Krishnamurthy, Loren Cobb, Jan Mandel, and Jonathan Beezley. Bayesian Tracking of Emerging Epidemics Using Ensemble Optimal Statistical Interpolation (EnOSI), September 2010. URL http://arxiv.org/abs/1009.4959v1.

M. Kuss and C. Rasmussen. Assessing Approximate Inference for Binary Gaussian Process Classification. *Journal of Machine Learning Research*, 6:1679–1704, 2005.

Paul Lamere. The LastFM ArtistTags2007 Data set, 2008. URL http://musicmachinery.com/2010/11/10/lastfm-artisttags2007.

Yew Jin Lim and Yee Whye Teh. Variational Bayesian Approach to Movie Rating Prediction. In *KDDCup*, 2007.

P. McCullagh and J. Nelder. *Generalized linear models*. Chapman and Hall, 1989. 2nd edition.

S. Mohamed, K. Heller, and Z. Ghahramani. Bayesian Exponential Family PCA. In *Advances in Neural Information Processing Systems*, 2008.

M. Opper and C. Archambeau. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

Mijung Park and Jonathan W. Pillow. Bayesian inference for low rank spatiotemporal neural receptive fields. In *Advances in Neural Information Processing Systems*, pages 2688–2696, 2013.

R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *International Conference on Machine Learning*, pages 880–887. ACM, 2008a.

Ruslan Salakhutdinov and Andriy Mnih. Probabilistic Matrix Factorization. In *Advances in Neural Information Processing Systems*, volume 20, 2008b.

Conrad Sanderson. Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments. Technical report, 2010.

Matthias Seeger and Guillaume Bouchard. Fast variational Bayesian inference for non-conjugate matrix factorization models. In *Proceedings of the 15th international conference on artificial intelligence and statistics*, number EPFL-CONF-174931, 2012.

M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of Royal Statistical Sociecty, Series B*, 21(3):611–622, 1999.

Max Welling, Chaitanya Chemudugunta, and Nathan Sutter. Deterministic Latent Variable Models and their Pitfalls. In *International Conference on Data Mining*, 2008.

Byron M Yu, KV Shenoy, and M Sahani. Expectation propagation for inference in non-linear dynamical models with Poisson observations. In *Nonlinear Statistical Signal Processing Workshop, 2006 IEEE*, pages 83–86. IEEE, 2006.

Mingyuan Zhou, Lauren Hannah, David B. Dunson, and Lawrence Carin. Beta-Negative Binomial Process and Poisson Factor Analysis. In Neil D. Lawrence and Mark Girolami, editors, *AISTATS*, volume 22 of *JMLR Proceedings*, pages 1462–1471. JMLR.org, 2012.

Appendix A. Expected Data Log-Likelihood

First, we show that $\mathbb{E}_{p(\boldsymbol{x})}[\exp\left(\mathbf{t}^T\mathbf{x}\right)] = \exp(\mathbf{t}^T\mathbf{m} + \frac{1}{2}\mathbf{t}^T\mathbf{V}\mathbf{t})$ for $p(\mathbf{x}) = \mathcal{N}(\mathbf{m}, \mathbf{V})$. Expanding the expectation, we see that the mean is shifted by $\mathbf{V}\mathbf{t}$. Completing the square results in an unnormalized Gaussian, for which the integral can be easily computed.

$$\mathbb{E}_{p(\boldsymbol{x})}[e^{\mathbf{t}^T\mathbf{x}}] = \int \mathcal{N}(\mathbf{m}, \mathbf{V}) \exp\left(\mathbf{t}^T\mathbf{x}\right) d\mathbf{x} \tag{20}$$

$$= C \int \exp\left(-\tfrac{1}{2}\left(-2\mathbf{t}^T\mathbf{x} + \mathbf{x}^T\mathbf{V}^{-1}\mathbf{x} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} - 2\mathbf{m}^T\mathbf{V}^{-1}\mathbf{x}\right)\right) d\mathbf{x} \tag{21}$$

$$= C \int \exp\left(-\tfrac{1}{2}\left(\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} - 2\left(\mathbf{m} + \mathbf{V}\mathbf{t}\right)^T\mathbf{V}^{-1}\mathbf{x}\right)\right) d\mathbf{x} \tag{22}$$

$$= \int \mathcal{N}\left(\mathbf{m} + \mathbf{V}\mathbf{t}, \mathbf{V}\right) \exp\left(\tfrac{1}{2}\left((\mathbf{m} + \mathbf{V}\mathbf{t})^T\mathbf{V}^{-1}(\mathbf{m} + \mathbf{V}\mathbf{t}) - \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}\right)\right) d\mathbf{x} \tag{23}$$

$$= \exp(\mathbf{t}^T\mathbf{m} + \tfrac{1}{2}\mathbf{t}^T\mathbf{V}\mathbf{t}) \tag{24}$$

where $C = |2\pi\mathbf{V}|^{-\frac{1}{2}}$.

Next, we show how to evaluate the expectation in Eq. 13, which we restate here.

$$\mathbb{E}_{q(\boldsymbol{z}_n, \boldsymbol{w}_i)}[\log p(y_{in}|\eta_{in})] = \mathbb{E}_{q(\boldsymbol{w}_i)}\left[y_{in}\mathbf{w}_i^T\mathbf{m}_n - e^{\mathbf{m}_n^T\mathbf{w}_i + \frac{1}{2}\mathbf{w}_i^T\mathbf{V}_n\mathbf{w}_i}\right] + \text{cnst} \tag{25}$$

$$= y_{in}\boldsymbol{\mu}_i^T\mathbf{m}_n - \mathbb{E}_{q(\boldsymbol{w}_i)}\left[e^{\mathbf{m}_n^T\mathbf{w}_i + \frac{1}{2}\mathbf{w}_i^T\mathbf{V}_n\mathbf{w}_i}\right] + \text{cnst} \tag{26}$$

Dropping all indices, we expand the expectation and examine the exponent in

$$\mathbb{E}_{q(\boldsymbol{w})}\left[e^{\mathbf{m}^T\mathbf{w} + \frac{1}{2}\mathbf{w}^T\mathbf{V}\mathbf{w}}\right] = \int \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) e^{\mathbf{m}^T\mathbf{w} + \frac{1}{2}\mathbf{w}^T\mathbf{V}\mathbf{w}} d\mathbf{w} \tag{27}$$

which is given and simplified by

$$-\tfrac{1}{2}\left(\left(\mathbf{w} - \boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{w} - \boldsymbol{\mu}\right) - 2\mathbf{m}^T\mathbf{w} - \mathbf{w}^T\mathbf{V}\mathbf{w}\right) \tag{28}$$

$$= -\tfrac{1}{2}\left(\mathbf{w}^T\left(\boldsymbol{\Sigma}^{-1} - \mathbf{V}\right)\mathbf{w} - 2\left(\mathbf{m} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)^T\mathbf{w} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \tag{29}$$

$$= -\tfrac{1}{2}\left(\mathbf{w}^T\mathbf{B}\mathbf{w} - 2\boldsymbol{\nu}^T\mathbf{w} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \tag{30}$$

where we replaced $\mathbf{B} = \left(\boldsymbol{\Sigma}^{-1} - \mathbf{V}\right)$ and $\boldsymbol{\nu} = \mathbf{m} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. With defining $\mathbf{u} = \mathbf{B}^{-1}\boldsymbol{\nu}$, we complete the square

$$= -\tfrac{1}{2}\left(\mathbf{w}^T\mathbf{B}\mathbf{w} - 2\mathbf{u}^T\mathbf{B}\mathbf{w} + \mathbf{u}^T\mathbf{B}\mathbf{u} - \mathbf{u}^T\mathbf{B}\mathbf{u} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \tag{31}$$

$$= -\tfrac{1}{2}\left(\mathbf{w} - \mathbf{u}\right)^T\mathbf{B}\left(\mathbf{w} - \mathbf{u}\right) + \tfrac{1}{2}\boldsymbol{\nu}^T\mathbf{B}^{-1}\boldsymbol{\nu} - \tfrac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{32}$$

The part depending on $\mathbf{w}$ is a valid Gaussian function if $\mathbf{B}$ is positive definite. In that case, we can evaluate the integral

$$\int \exp\left(-\tfrac{1}{2}\left(\mathbf{w} - \mathbf{u}\right)^T\mathbf{B}\left(\mathbf{w} - \mathbf{u}\right)\right) d\mathbf{w} = |2\pi\mathbf{B}^{-1}|^{\frac{1}{2}} \tag{33}$$

Multiplying this with the the normalizing constant of $q(\mathbf{w})$, $|2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}$, gives

$$|2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}|2\pi\mathbf{B}^{-1}|^{\frac{1}{2}} = |\boldsymbol{\Sigma}\mathbf{B}|^{-\frac{1}{2}} = |\mathbf{I} - \boldsymbol{\Sigma}\mathbf{V}|^{-\frac{1}{2}} = |\mathbf{S}|^{-\frac{1}{2}} \tag{34}$$

Note, that this term is real valued when $\mathbf{S}$ is not negative definite, which imposes a constraint on the posterior covariances $\mathbf{V}, \boldsymbol{\Sigma}$.

Putting the remaining terms together yields the formulation of Eq. 13.

$$\mathbb{E}_{q(\boldsymbol{w})}\left[e^{\mathbf{m}^T\mathbf{w}+\frac{1}{2}\mathbf{w}^T\mathbf{V}\mathbf{w}}\right] = |\mathbf{S}|^{-\frac{1}{2}}\exp\left(\tfrac{1}{2}\boldsymbol{\nu}^T\mathbf{B}^{-1}\boldsymbol{\nu} - \tfrac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \tag{35}$$