
Variational inference for neural network matrix factorization and its application to stochastic blockmodeling

Onno Kampman¹ Creighton Heaukulani²

Abstract

We consider the probabilistic analogue to neural network matrix factorization (Dziugaite & Roy, 2015), which we construct with Bayesian neural networks and fit with variational inference. We find that a linear model fit with variational inference can attain equivalent predictive performance to the regular neural network variants on the Movielens data sets. We discuss the implications of this result, which include some suggestions on the pros and cons of using the neural network construction, as well as the variational approach to inference. Such a probabilistic approach is required, however, when considering the important class of stochastic block models. We describe a variational inference algorithm for a neural network matrix factorization model with nonparametric block structure and evaluate its performance on the NIPS co-authorship data set.

1. Introduction

Matrix factorization models are an important class of machine learning methods, playing a prominent role in dimensionality reduction, with applications to product recommendations in commerce, among others. For example, $X_{n,m}$ could represent the amount of item $m \leq M$ purchased by user $n \leq N$. A classic approach to factorizing the $N \times M$ matrix X would assume a linear model such as

$$X_{n,m} = U_n^T V_m = \sum_{k=1}^K U_{n,k} V_{m,k}, \quad n \leq N, m \leq M, \quad (1)$$

for some (relatively small) number of factors $K \ll N, M$, and where the parameter vectors U_n and V_m are to be inferred with a procedure such as singular value decomposition. Dziugaite & Roy (2015) consider a *neural network*

matrix factorization alternative that replaces the linear function in Eq. (1) with a feed-forward neural network (with inputs U_n and V_m), which improves predictive performance when predicting missing entries of the matrix.

Here we take a probabilistic approach by using a *Bayesian neural network*, and we fit the parameters of the model with variational inference. While *probabilistic matrix factorization* (Mnih & Salakhutdinov, 2008) has shown improvements (for linear models) over its non-Bayesian counterpart, we find only a small improvement for this Bayesian variant of neural network matrix factorization (fit via variational inference, anyway) upon the predictive performance of the neural network on the Movielens 100K and 1M data sets. However, we do find that variational inference can get a linear model to match the performance of the neural network, and that the neural network structure provides further improvements when side information (such as the genre of the film) is included. In light of this (rather surprising) result, we provide a discussion on the pros and cons of using neural network structures and/or variational inference in these contexts.

Finally, one case when a probabilistic approach is *required* for tractable inference is in the important class of stochastic block models. We present a variant of neural network matrix factorization applied to network models (i.e., the matrix X is symmetric in this case) that captures nonparametric block structure, similar in spirit to the *infinite relational model* (Kemp et al., 2006). We derive the variational inference procedure for such a model, and we show that its predictive performance improves upon its linear analogues when applied to the NIPS co-authorship data set.

2. Neural network matrix factorization

Following Dziugaite & Roy (2015), model the entries of X as the outputs from a neural network f_θ with parameters θ , whose inputs are (unobserved) *features* of the users and items. In particular, for every $n \leq N$ and $m \leq M$, let

$$X_{n,m} = f_\theta([U_n, V_m, U'_{n,1} \circ V'_{m,1}, \dots, U'_{n,D} \circ V'_{m,D}]),$$

where the parameters have the following shapes: $U_n, V_m \in \mathbb{R}^K$, and $U'_{n,d}, V'_{m,d} \in \mathbb{R}^{K'}$, $d \leq D$, for some selected K, K' ,

¹Goldman Sachs, Hong Kong ²No affiliation. Correspondence to: Onno Kampman <onno.kampman@gs.com>, Creighton Heaukulani <c.k.heaukulani@gmail.com>.

K' , and D . The notation \circ here denotes the element-wise product, and $[a, b, \dots]$ denotes the *vectorization* function, i.e., the vectors a, b, \dots are concatenated into a single vector. Note that this neural network has $2K + K'D$ inputs and a univariate output.

Classic, linear constructions of the matrix factorization model can be recovered by restricting f_θ to be a linear function. The vectors $U'_{n,d} \circ V'_{m,d}$ play an analogous role to the traditional *bilinear* terms in the linear variants of matrix factorization, and the terms U_n and V_m play the role of the user- and item-specific *bias* terms in modeling variants such as those presented by Koren et al. (2009).

Inference in this model could then minimize the following regularized squared error loss function

$$\begin{aligned} & \sum_{(n,m) \in \mathcal{O}} (X_{n,m} - \hat{X}_{n,m})^2 + \lambda \cdot \left[\sum_n \|U_n\|_2^2 \right. \\ & \quad \left. + \sum_m \|V_m\|_2^2 + \sum_n \|U'_n\|_F^2 + \sum_m \|V'_m\|_F^2 \right], \quad (2) \\ & \hat{X}_{n,m} = f_\theta([U_n, V_m, U'_{n,1} \circ V'_{m,1}, \dots, U'_{n,D} \circ V'_{m,D}]), \end{aligned}$$

where \mathcal{O} denotes the set of observed edges, $\|A\|_F$ denotes the Frobenius norm for an array A , and $\lambda > 0$ is a regularization parameter.

3. Stochastic variational inference

We consider letting f_θ be a Bayesian neural network and elect a mean-field variational approach to inference. In the Bayesian perspective, the likelihood of the parameters given the data is conditionally Gaussian

$$\begin{aligned} X_{n,m} | \mu_{n,m} & \sim \mathcal{N}(\mu_{n,m}, \sigma^2) \quad (3) \\ \mu_{n,m} & = f_\theta([U_n, V_m, U'_{n,1} \circ V'_{m,1}, \dots, U'_{n,D} \circ V'_{m,D}]), \end{aligned}$$

for every $n \leq N$, $m \leq M$ and some additional noise parameter $\sigma > 0$. The components of the input arrays U , V , U' , and V' are all given independent mean zero Gaussian prior distributions (with array-specific, shared variance parameters), as are the weights and biases in θ .

We follow Salimans & Knowles (2013); Kingma & Welling (2014) to implement a gradient-based variational inference routine, where minibatches are subsampled from the observed edges in the graph, and where the required gradients are estimated by low-variance Monte-Carlo approximation routines. This technique is applied to both the neural network parameters θ and the inputs U, V, U', V' , which are updated in alternating steps during the gradient descent routine. This has become a common practice for variational inference with Bayesian neural networks, and so we defer the reader to the references for technical details.

3.1. Exploration of the linear model

We ran experiments on the Movielens 100K and Movielens 1M data sets (Harper & Konstan, 2016), which contain $N = 943$ users and $M = 1,682$ items (with 100,000 observations) and $N = 6,040$ users and $M = 3,706$ items (with 1,000,209 observations), respectively. Following the experimental setup of Dziugaite & Roy (2015), we create five random training/testing splits of the data sets, where 10% of the data set is held out as a test set in each instance. The root mean squared error (RMSE) is displayed for various models in Table 1.

The results from Dziugaite & Roy (2015) using a neural network for f_θ with hidden layers, each with 50 sigmoidal units, are reported as NN(3) and NN(4), and the models fit with variational inference as VI(0) and VI(3). In all of these variants, $K = 10$, $D' = 60$, and $K' = 1$. The VI models adapted the learning rates using Adam (Kingma & Ba, 2015), with an initial learning rate of 0.001. Batch learning (i.e., no minibatches) was used for all models. Due to memory constraints, we used training minibatches of 30,000 for the 1M data set. For reference, we have also included a singular value decomposition (SVD) baseline (truncated at 60 singular values), and the *biased matrix factorization* (Bias-MF) model (Koren et al., 2009).

Rather surprisingly, with variational inference we were able to get a linear model to match the performance of the neural network architecture. One possible conclusion is that variational inference is simply better at model selection than even a fine grid search. A Bayesian neural network fit with mean-field variational inference has the interpretation of placing a separate L2 regularization parameter (associated with the variance parameters of the Gaussian distributed components of the variational distribution) on each weight (and possibly bias) parameter of the function f_θ . This is rarely done in the non-Bayesian approaches to training neural neural networks, where typically a single or very few such regularization parameters are shared across the weights of the network. Moreover, with variational inference, these (possibly very many) weight regularization parameters are fit during gradient descent, whereas in non-Bayesian approaches they are typically selected by grid searching across multiple inference runs, which are easy to implement in parallel with the appropriate computing infrastructure, though can be a bit cumbersome to do so systematically. We note that Dziugaite & Roy (2015) did not regularize the parameters of f_θ in their experiments. However, it's still a useful (if unsurprising) lesson to see that within a single run of the inference procedure, variational inference is able to seamlessly do an otherwise piecemeal computational task. There is a slightly larger computational burden associated with variational inference, however, since the number of parameters to fit during inference doubles. Computations also

Table 1. RMSE scores for the Movielens data sets. The results for Bias-MF, NN(3), and NN(4) are taken from Dziugaite & Roy (2015).

DATA SET	SVD	BIAS-MF	NN(3)	NN(4)	VI(0)	VI(3)	VI(0)+S	VI(3)+S
MOVIELENS 100K	0.987	0.911	0.907	0.903	0.903	0.902	0.900	0.898
MOVIELENS 1M	0.917	0.852	0.846	0.843	0.839	0.836	-	-

increase linearly with the number of Monte Carlo samples used to approximate the required gradients (see Salimans & Knowles (2013) and Kingma & Welling (2014)), though this number can usually be very small (often one).

Viewed alternatively, the performance of the neural network suggests that by using its expressive power along with modern techniques in gradient-based inference, a user may largely ignore careful model selection on the weights of the neural network, or exhaustively fine grid searches over the regularization parameter λ .

3.2. Incorporating side information

For the Movielens 100K experiments, we also included the genre of each film as side information into the model, concatenated to the movie embedding V_m in the form of a one-hot vector. There are 19 different genres. The results are presented in Table 1 as VI(0)+S for the linear model and VI(3)+S for a neural network with 3 hidden layers of 50 units each. We can see that the performance of both models improves, perhaps suggesting that the nonlinear structure of the neural network is advantageous when handling (observed) side information.

4. Stochastic block models for network data

In this section, we will restrict our attention to the special case of network data sets, where the rows and columns of an $N \times N$ data matrix X correspond to the same set of N users, and an entry $X_{i,j} = 1$ if there is a “link” between users i and j and $X_{i,j} = 0$ otherwise. Such models are appropriate for social networks, where links represent friendships between individuals. We further assume the matrix X is symmetric (i.e., $X_{i,j} = X_{j,i}$), and we do not allow self-links (i.e., the diagonal elements of X are meaningless).

In the previous section, we considered some pros and cons of optionally using a Bayesian neural network f_θ . However, one scenario where a Bayesian approach is *required* for tractable inference is with *stochastic block models* (Kemp et al., 2006; Airoldi et al., 2008). In this important class of “community detection” models for network data, the users are clustered into groups, and the parameters of the model are shared amongst the members of a group in order to capture a well-observed phenomenon known as *homogeneity*. For example, clusters in a social network could represent shared interests of the users, or geographic location, both of

which presumably increase the likelihood that those users will be linked.

We take a nonparametric, Bayesian approach to stochastic blockmodeling, in a similar spirit to the *infinite relational model* by Kemp et al. (2006), which uses the Dirichlet process to model a potentially unbounded number of clusters that is inferred from the data. For every $i \leq N$, let $Z_i \in \{1, 2, \dots\}$ denote the (random) assignment of user i to one of an unbounded number of groups. For every $c = 1, 2, \dots$, let $U_c \in \mathbb{R}^K$ and $U'_{c,d} \in \mathbb{R}^{K'}$, $d \leq D$, denote the shared input features for the users in group c . Then for every $i < j \leq N$, let

$$X_{i,j} | p_{i,j} \sim \text{Bernoulli}(p_{i,j}) \quad (4)$$

$$p_{i,j} = f_\theta([U_{Z_i}, U_{Z_j}, U'_{Z_i,1} \circ U'_{Z_j,1}, \dots, U'_{Z_i,D} \circ U'_{Z_j,D}]),$$

where the neural network f_θ is now specified so that its output layer is pushed through a mapping to $(0, 1)$, such as the logistic sigmoid function.

The distribution on the assignment variables $Z := (Z_1, \dots, Z_n)$ is given by the (assignments under a) *Dirichlet process mixture model*, which we may describe via the *stick-breaking construction* for the Dirichlet process (Sethuraman, 1994). Independently for every $i \leq N$, let $Z_i | \pi \sim \pi$ be a sample in $\{1, 2, \dots\}$ according to the (infinite dimensional) probability vector $\pi := (\pi_1, \pi_2, \dots)$ defined as follows

$$\pi_i = V_i \prod_{j=1}^{i-1} (1 - V_j), \quad i = 1, 2, \dots, \quad (5)$$

$$V_i \sim \text{beta}(1, \alpha), \quad i = 1, 2, \dots, \quad (6)$$

where $\sum_{i=1}^{\infty} \pi_i = 1$, almost surely (as required), and $p(Z_n | \pi) = \pi_{Z_n}$, for every $n \leq N$, and $\alpha > 0$ is some *concentration parameter*.

The likelihood of the parameters given the data is then

$$\begin{aligned} \mathcal{L} &= p(\theta) \prod_{i=1}^{\infty} \text{beta}(V_i; 1, \alpha) \prod_{n=1}^N \left[p(Z_n | \pi) p(U_n) p(U'_n) \right] \\ &\times \prod_{i < j \leq N} \text{Bernoulli}(X_{i,j}; p_{i,j}), \end{aligned} \quad (7)$$

where $p(U_n)$, $p(U'_n)$, and $p(\theta)$ are the usual component-wise Gaussian densities for the inputs and neural network parameters specified in Section 3.

Table 2. RMSE and AUC scores on the NIPS co-authorship data set. Bias-MF is non-Bayesian. VI and SBM are linear, since additional layers did not improve results. Note that SBM has significantly fewer parameters than the other models.

METRIC	SVD	BIAS-MF	VI	SBM
RMSE	0.136	0.125	0.120	0.128
AUC	0.707	0.839	0.844	0.718

4.1. Gradient-based variational inference

We follow Blei & Jordan (2006) and take a mean-field variational approach to inference with this model, in which the discrete variables Z are integrated out, turning an intractable inference task into an optimization of continuous variables. The number of groups is also automatically inferred during this process. In particular, the variational approximation introduces a *truncation level* as the maximum number of components of the Dirichlet process. In practice, this truncation is selected to be large enough so that the algorithm does not “exhaust” all available components. Let

$$q(Z, V) = \prod_{i=1}^N \text{Mult}(Z_i; \eta_i) \prod_{c=1}^T \text{beta}(V_c; \rho_{c,1}, \rho_{c,2}) \quad (8)$$

denote the mean-field variational approximation, where η_i is a T -dimensional probability vector for some selected truncation level T , and $\rho_{c,1}, \rho_{c,2} > 0$.

The parameters η_i may be updated analytically following derivations similar to those by Blei & Jordan (2006) as follows. For every $i \leq N$ and $t \leq T$,

$$\eta_{i,t} \propto \exp \left\{ \mathbb{E}_q[\log V_t] + \sum_{\ell=1}^{t-1} \mathbb{E}_q[\log(1 - V_\ell)] \right. \\ \left. + \sum_{j: (i,j) \in \mathcal{O}} \mathbb{E}_q[\log \text{Bernoulli}(X_{i,j} | p_{i,j})] \right\}, \quad (9)$$

where $\mathbb{E}_q[\log V_t] = \psi(\rho_{t,1}) - \psi(\rho_{t,1} + \rho_{t,2})$ and $\mathbb{E}_q[\log(1 - V_t)] = \psi(\rho_{t,2}) - \psi(\rho_{t,1} + \rho_{t,2})$, with $\psi(a) := \Gamma'(a)/\Gamma(a)$ denoting the digamma function, and where the term $\mathbb{E}_q[\log \text{Bernoulli}(X_{i,j} | p_{i,j})]$ is approximated with a Monte-Carlo estimate.

The variational parameters $\rho_{c,1}, \rho_{c,2}$ also have analytic updates, however, we found it more successful to infer them with gradient-based updates. The concentration parameter α is optimized directly with gradient-based updates (i.e., type-I maximum likelihood). Finally, the inputs U and U' and the neural network parameters θ are inferred in the usual way (specified in Section 3). The parameter update schedule we followed is shown in Algorithm 1.

Algorithm 1 Stochastic variational inference for the stochastic block model

Data: $N \times M$ matrix X .

repeat

- 1 Sample a minibatch of the edges $\mathcal{O}^b \subset \mathcal{O}$.
- 2 For every node n present in (an edge in) the minibatch \mathcal{O}^b , update η_n according to Eq. (9) with gradients approximated on \mathcal{O}^b .
- 3 Update $q(V)$ and α .
- 4 Update $q(\theta)$ with gradients approximated on \mathcal{O}^b .
- 5 For every node n present in the minibatch \mathcal{O}^b , update $q(U_n)$ and $q(U'_n)$ with gradients approximated on \mathcal{O}^b .

until *Convergence*;

4.2. Exploring the NIPS co-authorship dataset

We ran experiments on the NIPS 1–17 co-authorship data set (Chechik & Globerson, 2007), consisting of authors that had published at least nine papers at NIPS between 1988 and 2003 (resulting in $N = 234$ authors). A link occurs between two authors if they co-authored at least one paper. A truncation level of $T = 7$ was used in the variational approximation to the Dirichlet process, and we note that these did not appear to be “exhausted” in our experiments. The experimental setup (five randomly held out test sets) and hyperparameter settings are otherwise identical to those in Section 3.1. The RMSE and AUC scores (averaged over the training runs and test sets) are reported in Table 2. Note that the (non-Bayesian) neural network matrix factorization model with no hidden layers is equivalent to the biased matrix factorization model “Bias-MF”, and so we use that name here. Bias-MF and its Bayesian analogue (fit with variational inference) “VI” only slightly best the linear variant of the stochastic block model “SBM”, which is remarkable since the stochastic block model has significantly fewer features. In particular, note SBM uses $T * (K + K' * D)$ input parameters, whereas Bias-MF and VI use $N * (K + K' * D)$. This difference is perhaps more pronounced, since the properties of the Dirichlet process attempt to effectively “pinch out” some of these features. Additional layers did not improve results here.

5. Future directions

On one hand, our results suggest investigation into models constructed from neural networks on whether their success depends on increasing model capacity/complexity. On the other hand, conventional wisdom has always suggested that more parsimonious models generalize better to new data, though that does not seem to be a hindrance to the neural network models in our experiments. Finally, the apparent advantages of the neural network when incorporating side information should be further explored.

References

- Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.
- Blei, D. M. and Jordan, M. I. Variational inference for Dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.
- Chechik, G. and Globerson, A. Nips 1–17 data, 2007. URL <https://chechiklab.biu.ac.il/gal/data.html>.
- Dziugaite, G. K. and Roy, D. M. Neural network matrix factorization. *arXiv preprint 1511.06443*, 2015.
- Harper, F. M. and Konstan, J. A. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent systems*, 5(4), 2016.
- Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. Learning systems of concepts with an infinite relational model. In *AAAI*, 2006.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *ICLR*, 2014.
- Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 42: 30–37, 2009.
- Mnih, A. and Salakhutdinov, R. R. Probabilistic matrix factorization. In *NIPS*, 2008.
- Salimans, T. and Knowles, D. A. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.
- Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica sinica*, pp. 639–650, 1994.