JMLR: Workshop and Conference Proceedings vol 52, 25-36, 2016 PGM 2016

Bayesian Matrix Factorization with Non-Random Missing Data using Informative Gaussian Process Priors and Soft Evidences

Bence Bolgár BOLGAR@MIT.BME.HU
Péter Antal ANTAL@MIT.BME.HU
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest (Hungary)

Abstract

We propose an extended Bayesian matrix factorization method, which can incorporate multiple sources of side information, combine multiple *a priori* estimates for the missing data and integrates a flexible missing not at random submodel. The model is formalized as probabilistic graphical model and a corresponding Gibbs sampling scheme is derived to perform unrestricted inference. We discuss the application of the method for completing drug–target interaction matrices, also discussing specialties in this domain. Using real-world drug–target interaction data, the performance of the method is compared against both a general Bayesian matrix factorization method and a specific one developed for drug–target interaction prediction. Results demonstrate the advantages of the extended model.

Keywords: Bayesian matrix factorization; missing not at random; drug–target interaction prediction.

1. Introduction

Analyzing incomplete dyadic data (*i.e.* an incomplete set of pairwise interaction scores) has drawn considerable interest in recent years. An early benchmark challenge of such problems is the movie recommendation problem, which aims to model and predict movie ratings provided by a large number of users. Matrix factorization methods, particularly their Bayesian extension became the most widely applied approaches to cope with overfitting in these problems (Salakhutdinov and Mnih, 2008b). However, the incorporation of heterogeneous side information about the entities and modelling the potentially informative, complex dependency patterns of missing data are still open challenges (Hernández-Lobato et al., 2014).

The matrix completion problem is also abundant in the life sciences, such as in gene function prediction (Zitnik and Zupan, 2014), gene prioritization (Zakeri et al., 2015) and in drug–target interaction prediction (Gönen et al., 2013; Yang et al., 2014; Buza, 2016). Drug–target interaction prediction is particularly important, as large-scale screening results are curated into publicly available repositories, thus containing unprecedented amount of high quality bioactivity measurements (Williams et al., 2012; Jupp et al., 2014). Additionally, the inclusion of rich side information (*e.g.* molecular similarities) and the development of a refined model of missing data are especially promising directions in this domain.

Utilizing the advantages of the Bayesian statistical and PGM frameworks for structured data and knowledge fusion, we present an extended Bayesian matrix factorization scheme formalized as a probabilistic graphical model and derive a Gibbs sampling scheme to perform unrestricted inference. The proposed method can incorporate multiple kernels as side information and *a priori*

potentially incomplete estimates for interaction scores, and also contains a novel missing not at random (MNAR) data submodel. We evaluate its predictive performance using real-world drug–target interaction data and compare it against general and drug-target specific Bayesian matrix factorization methods.

2. Earlier Works

In machine learning, matrix factorization-based methods have become a well-established and powerful approach to analyze dyadic data. The main idea is to find a low-rank approximation for a matrix of observations $\boldsymbol{R} \in \mathbb{R}^{I \times J}$ as a product of factors $\boldsymbol{U} \in \mathbb{R}^{L \times I}$ and $\boldsymbol{V} \in \mathbb{R}^{L \times J}$, such that

$$\boldsymbol{R} \approx \boldsymbol{U}^T \boldsymbol{V},$$

where $L \ll \operatorname{rank}(\boldsymbol{R})$. The usual interpretation is to think of the columns of $\boldsymbol{U}$ as some representation of I "row-entities" (*e.g.* users), the columns of $\boldsymbol{V}$ as some representation of J "column-entities" (*e.g.* movies), and $\boldsymbol{R}$ contains the $I \times J$ interaction scores (*e.g.* movie ratings). In most works, the Frobenius norm is employed in the loss function:

$$\min_{\boldsymbol{U},\boldsymbol{V}} \quad \left\| \boldsymbol{R} - \boldsymbol{U}^T \boldsymbol{V} \right\|_F^2 . \tag{1}$$

Singular value decomposition (SVD) provides a unique and optimal solution for this problem, easily tweaked to handle missing observations by re-formulating (1) in an element-wise manner and keeping only the terms where $\boldsymbol{R}_{ij}$ is known. Matrix completion can then be achieved by predicting the missing values of $\boldsymbol{R}$ by simply multiplying the resulting factors.

This basic model faces multiple challenges. The SVD solution is prone to overfitting as the factors can get arbitrarily large. This was remedied by introducing a generative model, coined PMF for probabilistic matrix factorization, which places a Gaussian noise model on $\boldsymbol{R}$ and treats the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ as zero-mean multivariate normal variables (Salakhutdinov and Mnih, 2008b). PMF was later extended to full Bayesian inference (BPMF) by putting Normal-Wishart hyperpriors onto the parameters of PMF (Salakhutdinov and Mnih, 2008a). PMF and BPMF are illustrated on Figure 1.

A significant amount of research was conducted to find ways to incorporate "side information" (*i.e.* prior representations of entities) into the matrix factorization framework. The traditional approach to harness this extra information is collaborative filtering (CF). CF computes similarities between entities (*e.g.* user–user or movie–movie similarities) using prior descriptions and uses this extra information to take values for "similar" entities into account (for a detailed review, see (Ekstrand et al., 2011)). More recently, other schemes were proposed, including models using Gaussian Processes (Adams et al., 2010; Zhou et al., 2012), linear regression models on the latent vectors (Agarwal and Chen, 2009; Park et al., 2013; Simm et al., 2015) and joint decomposition models (Yoo and Choi, 2011; Singh and Gordon, 2012).

The "missingness" pattern of $\boldsymbol{R}$ in matrix factorization methods is usually treated as missing at random (MAR). However, in practice, the MAR assumption often fails, for example, instead of giving a bad movie a low score, users tend not to rate it at all, *i.e.* the missingness pattern depends on the values of $\boldsymbol{R}$. There were only a handful of works in this field which addressed this issue. A proposed solution is to build a separate matrix factorization model for the binarized observation matrix (Hernández-Lobato et al., 2014).

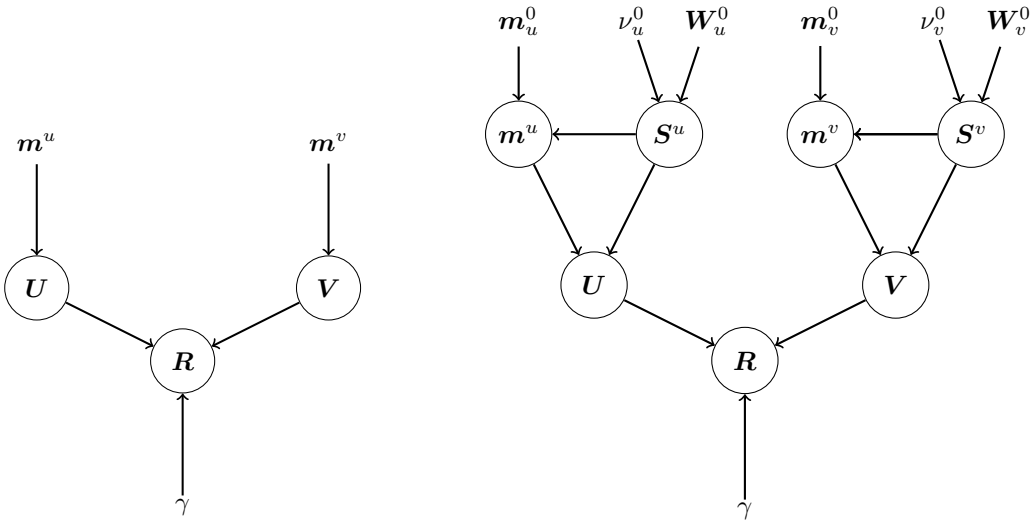

Figure 1: Graphical models corresponding to PMF (left) and BPMF (right). Notation was adapted to this work.

3. The Extended Bayesian Factorization Model

These challenges also arise in drug–target interaction prediction. As side information, a wide range of molecular representations ("fingerprints") and induced similarities are available. However, a distinctive feature of this application area is the availability of analytic models (*e.g.* docking simulations) for the interaction scores, which may provide valuable estimates. Finally, the set of existing interaction data is much larger than the shared data located in publicly available repositories, as a consequence of obvious irrelevance, difficulties of measurements or policies of the pharmaceutical industry – a clear violation of the MAR condition.

Hence, in this work, we propose an extended Bayesian matrix factorization method, HuTolt, with the following properties:

- It can incorporate multiple sources of side information through Gaussian Process priors, which enables the fusion of heterogeneous data about the entities through multiple kernel learning (Gönen et al., 2013).
- It contains a novel missing not at random (MNAR) model instead of the currently prevailing missing at random assumption.
- It can incorporate multiple estimates for missing interaction scores.

3.1 Matrix Factorization with Gaussian Processes

We follow the notation employed in (Salakhutdinov and Mnih, 2008a). Let us denote the matrix to be approximated by $\boldsymbol{R} \in \mathbb{R}^{I \times J}$. Our goal is to find a low-rank approximation of $\boldsymbol{R}$ with factors $\boldsymbol{U} \in \mathbb{R}^{L \times I}$ and $\boldsymbol{V} \in \mathbb{R}^{L \times J}$, such that

$$\boldsymbol{R} \approx \boldsymbol{U}^T \boldsymbol{V},$$

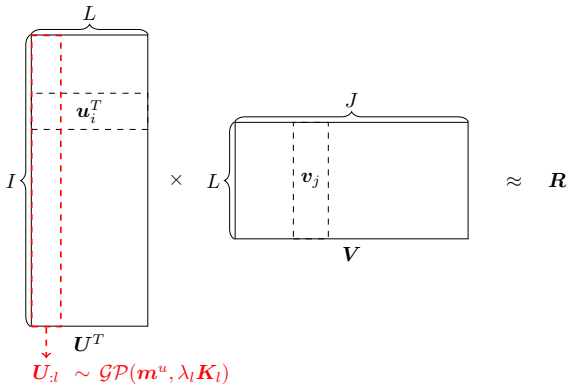

Figure 2: Matrix factorization with Gaussian Processes. Each *row* of $\boldsymbol{U}$ is governed by a (weighted) Gaussian Process. Their covariance matrices enforce similarities over the *columns* of $\boldsymbol{U}$.

where L is the number of latent dimensions. Each column $\boldsymbol{u}_i$ of $\boldsymbol{U}$ can be thought of as a low-dimensional representation of the entity corresponding to the ith row of $\boldsymbol{R}$.

Now let us assume that we also have access to multiple symmetric positive definite kernel matrices containing inner products of some high-dimensional representations of these same entities. Following Zhou et al. (2012), we require the distributions of the *rows* of $\boldsymbol{U} \in \mathbb{R}^{L \times I}$ to be governed by L independent Gaussian Processes, each using a weighted kernel as covariance matrix:

$$p(\boldsymbol{U}|\boldsymbol{m}^u, \boldsymbol{K}, \boldsymbol{\lambda}) = \prod_{l=1}^{L} \mathcal{N}(\boldsymbol{U}_{l:}|\boldsymbol{m}^u_{l:}, \lambda_l \boldsymbol{S}_l^{u-1}),$$

where $\boldsymbol{m}^u_{l:}$ is the mean vector for the lth process, $\boldsymbol{S}^u_l$ is the inverse of the lth kernel matrix $\boldsymbol{K}_l$ and λ_l is the associated kernel weight. This ensures that in lth row of $\boldsymbol{U}$, values respect the inner products specified by the kernel $\boldsymbol{K}_l$ and the columns of $\boldsymbol{U}$, in general, share similarities governed by the kernels (Figure 2 illustrates this idea). Moreover, the "importance" of the latent dimensions are automatically established through learning the kernel weights.

The distributions of the *columns* of $\boldsymbol{V} \in \mathbb{R}^{L \times J}$ are, as usual,

$$p(\boldsymbol{V}|\boldsymbol{m}^v, \boldsymbol{S}^v) = \prod_{j=1}^{J} \mathcal{N}(\boldsymbol{v}_j|\boldsymbol{m}^v_j, \boldsymbol{S}_j^{v-1}),$$
$$p(\boldsymbol{m}^v, \boldsymbol{S}^v|\boldsymbol{m}^0, \nu^0, \boldsymbol{W}^0) = \mathcal{NW}(\boldsymbol{m}^v, \boldsymbol{S}^v|\boldsymbol{m}^0, \kappa^0, \boldsymbol{W}^0, \nu^0),$$

where $\boldsymbol{v}_j$ is the jth column of $\boldsymbol{V}$ and we put a Normal–Wishart prior on $\boldsymbol{m}^v$ and $\boldsymbol{S}^v$. Note that this can easily be replaced with a Gaussian process prior such as in the case of $\boldsymbol{U}$. In this work, we keep

the standard Normal–Wishart prior. The distribution of the incomplete matrix $\boldsymbol{R}$ is given by

$$p(\boldsymbol{R}|\boldsymbol{U},\boldsymbol{V},\gamma^c,\gamma^r) = \prod_{i=1}^{I}\prod_{j=1}^{J}\left[\mathcal{N}(\boldsymbol{R}_{ij}|\boldsymbol{u}_i^T\boldsymbol{v}_j,(\gamma_i^c\gamma_j^r)^{-1})\right]^{I_{ij}},$$

$$p(\boldsymbol{\gamma}^c|\boldsymbol{a}^c,\boldsymbol{b}^c) = \prod_{i=1}^{I}\mathcal{G}a(\gamma_i^c|a_i^c,b_i^c),$$

$$p(\boldsymbol{\gamma}^r|\boldsymbol{a}^r,\boldsymbol{b}^r) = \prod_{j=1}^{J}\mathcal{G}a(\gamma_j^r|a_j^r,b_j^r),$$

where I_{ij} is a binary indicator variable denoting the existence of R_{ij} and $\mathcal{G}a$ is the Gamma distribution. The kernel weights $\boldsymbol{\lambda}$ follow the Inverse Gamma distribution

$$p(\boldsymbol{\lambda}|\boldsymbol{a}^w,\boldsymbol{b}^w) = \prod_{l=1}^{L}\mathcal{IG}(\lambda_l|a_l^w,b_l^w).$$

3.2 Background Knowledge Model

Let us assume that, at least for some (i,j), we have access to additional information regarding $\boldsymbol{R}_{ij}$ (*e.g.* values from other prediction schemes or expert opinions). To account for this in the model, we specify the nth background knowledge model as

$$p(\boldsymbol{B}^n|\boldsymbol{R},\sigma^c,\sigma^r) = \prod_{i=1}^{I}\prod_{j=1}^{J}\left[\mathcal{N}(\boldsymbol{B}_{ij}^n|\boldsymbol{R}_{ij},(\sigma_i^{nc}\sigma_j^{nr})^{-1})\right]^{I_{ij}},$$

$$p(\boldsymbol{\sigma}^{nc}|\boldsymbol{a}^{nc},\boldsymbol{b}^{nc}) = \prod_{i=1}^{I}\mathcal{G}a(\sigma_i^{nc}|a_i^{nc},b_i^{nc}),$$

$$p(\boldsymbol{\sigma}^{nr}|\boldsymbol{a}^{nr},\boldsymbol{b}^{nr}) = \prod_{j=1}^{J}\mathcal{G}a(\sigma_j^{nr}|a_j^{nr},b_j^{nr}),$$

where $\boldsymbol{B} \in \mathbb{R}^{I\times J}$. By estimating the precision parameters $\boldsymbol{\sigma}^{nc}$ and $\boldsymbol{\sigma}^{nr}$, the accuracy of this extra information with respect to each entity (*i.e.* row and column) will be automatically taken into account when performing Bayesian inference.

3.3 Missing Data Model

The missing data model is built around *a priori* specified intervals which influence whether a given $\boldsymbol{R}_{ij}$ is accessible or not (*i.e.* a NMAR approach). We give $\boldsymbol{X} \in \{0,1\}^{I\times J}$ a Bernoulli distribution:

$$p(\boldsymbol{X}|\boldsymbol{R},s_1,s_2,\mu) = \prod_i\prod_j f(\boldsymbol{R}_{ij},s_1,s_2,\mu)^{\boldsymbol{X}_{ij}}(1-f(\boldsymbol{R}_{ij},s_1,s_2,\mu))^{1-\boldsymbol{X}_{ij}}.$$

Here we utilize the "bump" function

$$f(x,s_1,s_2,\mu) = \begin{cases} 1, & \text{if } |x-\mu| < s_1 \\ 0, & \text{if } |x-\mu| \geq s_2 \\ \sigma\left(-\frac{s_1^2+s_2^2-2(x-\mu)^2}{((x-\mu)^2-s_1^2)\cdot((x-\mu)^2-s_2^2)}\right) & \text{otherwise,} \end{cases}$$

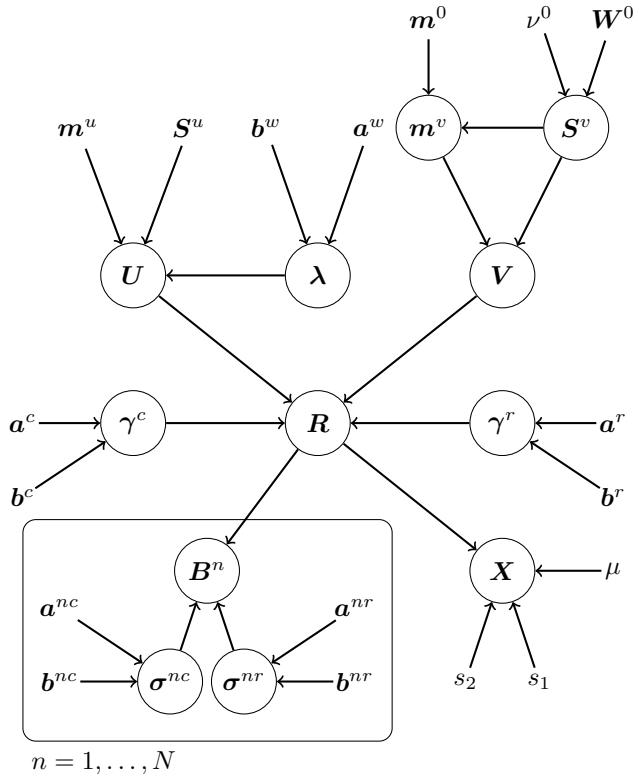

Figure 3: HuTolt: the extended matrix factorization model. $\boldsymbol{R}$ and all its ancestors constitute the matrix factorization model. $\boldsymbol{B}_n$ denotes the nth background knowledge model and $\boldsymbol{X}$ is the missing data model.

where $\sigma(\cdot)$ is the logistic sigmoid function, μ is the mean parameter and s_1, s_2 specify the "intermediate" region on the sides. When $\boldsymbol{R}_{ij}$ is outside the support of f, it will be considered to be missing with probability 1. This lets us

- constrain the unknown (sampled) values of $\boldsymbol{R}_{ij}$ to meaningful intervals,
- exploit the information carried by missing data points.

This model can be refined by multiplying several bump functions to define more intervals or by incorporating a smoothing parameter ρ to soften the probability 1 assumption; in both cases, the function will remain in $\mathcal{C}^\infty$. Another interesting extension is the application of a binomial distribution, modelling how many times a particular interaction has been measured.

3.4 Inference using Gibbs Sampling

We use Gibbs sampling to perform Bayesian inference. Using conjugate priors makes the inference easy, the only non-trivial parts being sampling the columns of $\boldsymbol{U}$ and the missing entries of $\boldsymbol{R}$.

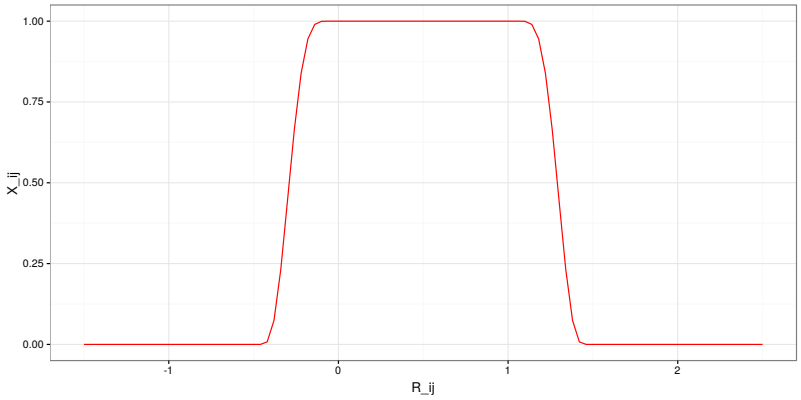

Figure 4: "Bump" function with $s_1 = 0.5$, $s_2 = 1$, $\mu = 0.5$.

Let $\boldsymbol{\Theta}^{\setminus \boldsymbol{U}}$ denote all variables except $\boldsymbol{U}$. A fairly easy derivation shows that the conditional is

$$
\begin{aligned}
p(\boldsymbol{U}|\boldsymbol{\Theta}^{\setminus \boldsymbol{U}}) &= \prod_i \mathcal{N}(\boldsymbol{u}_i|\boldsymbol{\psi}_i, \boldsymbol{\Lambda}_i^{-1}), \\
\boldsymbol{\Lambda}_i &= \gamma_i^c \sum_j \gamma_j^r \left(\boldsymbol{v}_j \boldsymbol{v}_j^T\right) I_{ij} + \mathrm{diag}_l(\lambda_l^{-1} \boldsymbol{S}_{lii}^u), \\
\boldsymbol{\psi}_i &= \boldsymbol{\Lambda}_i^{-1} \left[\gamma_i^c \sum_j \gamma_j^r \left(\boldsymbol{R}_{ij} \boldsymbol{v}_j\right)^{I_{ij}} - \mathrm{vec}_l \left(\lambda_l^{-1} \sum_{n \neq i} \boldsymbol{S}_{lin}^u \left(\boldsymbol{U}_{ln} - \boldsymbol{m}_{ln}^u\right) \right) \right],
\end{aligned}
$$

where $\mathrm{vec}_l \left\{x_l\right\}_{l=1}^L := [x_1, x_2, \ldots, x_L]^T$. A detailed proof can be found in Appendix A. The joint distribution with respect to $\boldsymbol{R}_{ij}$ is

$$
p(\boldsymbol{R}_{ij}, \boldsymbol{\Theta}^{\setminus \boldsymbol{R}_{ij}}) = \mathcal{N}(\boldsymbol{R}_{ij}|\boldsymbol{u}_i^T \boldsymbol{v}_j, (\gamma_i^c \gamma_j^r)^{-1}) \prod_{n=1}^N \mathcal{N}(\boldsymbol{B}_{ij}^n|\boldsymbol{R}_{ij}, (\sigma_i^{nc} \sigma_j^{nr})^{-1}) \mathcal{B}(\boldsymbol{X}_{ij}|f(\boldsymbol{R}_{ij}, s_1, s_2, \mu))
$$

where $\mathcal{B}$ is the Bernoulli distribution. It is easy to see that the conditional is, in general, not log-concave, hence we had to resort to a slice sampling step within the Gibbs sampling. The convolution of the first two terms gives a normal distribution on $\boldsymbol{R}_{ij}$, which can be calculated analytically. Then we separate the unimodular and bimodular cases corresponding to $\boldsymbol{X}_{ij} = 0$ and $\boldsymbol{X}_{ij} = 1$ and apply a simple slice sampling scheme.

4. Experiments

4.1 Drug–Target Interaction Prediction

We evaluated the HuTolt method in a drug–target interaction prediction task. Binding affinities were collected from the public ChEMBL database (Bento et al., 2014). We restricted our attention to psychiatric drugs belonging to the ATC class N06*. K_i values were logarithmically transformed, multiple measurements were aggregated by their median values and inexact measurements were

	HuTolt				Macau	BPMF
	2K+MDM	2K	1K	0K		
Mean	**0.669**	0.698	0.733	0.767	0.749	0.817
StDev	0.041	0.017	0.032	0.075	0.058	0.132
Diff	0.126	0.050	0.087	0.176	0.159	0.392

Table 1: RMSE values in the drug–target interaction prediction task in a 80%-20% cross-validation setting. "nK" denotes the number of kernels used (in the case $n = 0$, the identity matrix was used). "MDM" denotes the utilization of the missing data model.

discarded, resulting in an incomplete matrix with 37 drugs and 82 targets. The matrix contained 446 entries within the interval $[0, 4.328]$ with mean 0.986.

Chemical fingerprints were computed using the CDK library (Steinbeck et al., 2003). Specifically, MACCS and Klekota-Roth fingerprints were used. Kernels were computed using the Tanimoto similarity measure, which is the gold standard similarity measure in chemoinformatics (Eckert and Bajorath, 2007).

The HuTolt algorithm was compared to the general BPMF method (Salakhutdinov and Mnih, 2008a) and the drug–target specific Macau tool (Simm et al., 2015). RMSE values were computed using a 80%-20% cross-validation scheme. Due to space limitations, here we present a cumulative evaluation of the extensions. In this experiment, we used $\mathcal{NW}(\mathbf{0}, 1000, \boldsymbol{I}, L)$ for the prior of $\boldsymbol{V}$, $\mathcal{N}(\mathbf{0}, \boldsymbol{S}^u)$ for $\boldsymbol{U}$, Gamma priors were parameterized with $a = 10$, $b = 1$, Inverse Gammas with $a = 1$, $b = 2$ and 8 latent factors were utilized (4 for each kernel). Since BPMF cannot incorporate a background knowledge model, we also omitted their use for a fair comparison (results from molecular docking simulations would be a reasonable choice here). The missing data model was parameterized to include the interval described above.

Results are shown in Table 1. HuTolt is on par with Macau using only MACCS fingerprints and both outperform BPMF which does not utilize side information. It is worth mentioning that HuTolt outperforms BPMF even with no side information (*i.e.* using the identity matrix in the Gaussian Process priors) which highlights the benefits of the more sophisticated noise model of HuTolt, utilizing a product of per-entity noise variables $\boldsymbol{\gamma}^c$ and $\boldsymbol{\gamma}^r$ instead of a single Gaussian noise variable employed in BPMF. The second kernel brings an additional decrease in RMSE which indicates the advantages of data fusion and Multiple Kernel Learning. Finally, using the missing data model yields best results which demonstrate the applicability of NMAR-type models in drug–target interaction prediction.

We used 500 burn-in steps, which was sufficient for convergence (Figure 5). We also examined the resulting low-dimensional representations of the drugs, *i.e.* the columns of $\boldsymbol{U}$. Figure 6 illustrates the correlation between column–column similarities and corresponding kernel values. The number of latent factors was increased to 15 for this experiment.

5. Conclusion and Future Work

The extended Bayesian matrix factorization method, HuTolt, allows the synergistic use of three extensions. First, it allows the incorporation of multiple similarities both over row-entities and column-entities, performing adaptive multiple kernel learning driven by the whole system. Second,

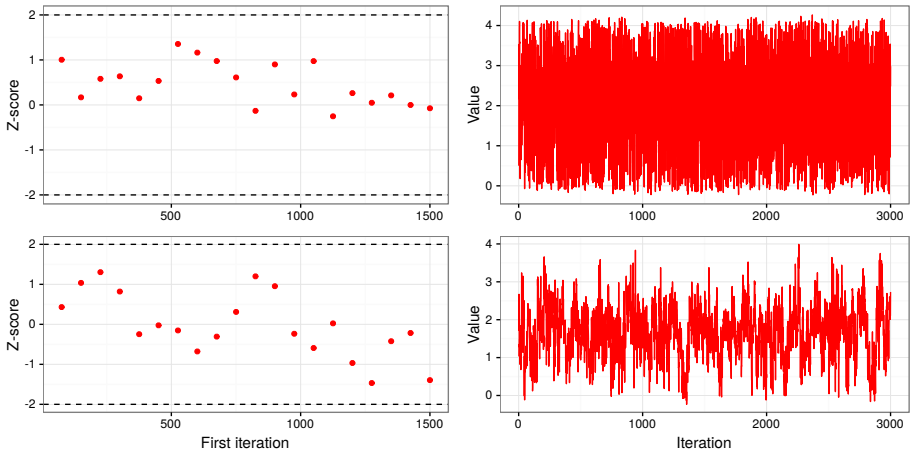

Figure 5: Geweke–Brooks plots demonstrating the convergence of the Gibbs inference for high and low affinities in $\boldsymbol{R}$. On the right, corresponding trace plots are shown.

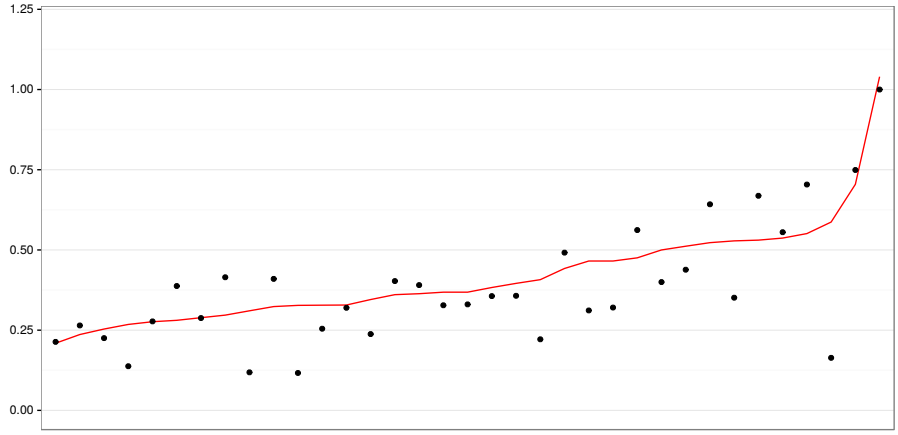

Figure 6: Correlation between $\boldsymbol{U}$ and the inner product values of $\boldsymbol{K}$ after a run with 15 latent factors and one kernel. Similarities of $\boldsymbol{u}_i$ to the fixed column $\boldsymbol{u}_1$ for all i are illustrated with black dots. The corresponding kernel values $\boldsymbol{K}_{1i}$ are denoted with a red line. The pairs $\{(sim(\boldsymbol{u}_1, \boldsymbol{u}_i), \boldsymbol{K}_{1i})\}_{i=1}^{I}$ were sorted w.r.t. $\boldsymbol{K}_{1i}$.

it supports the fusion of analytic estimates or expert hints for missing data, again, automatically establishing their importance w.r.t. the whole system. Third, it extracts information from the missingness pattern, whose effect is integrated consistently into the overall posterior. In summary, the proposed system defines a consistent, encompassing, Bayesian fusion of entity similarities, interaction data, interaction estimates and missingness status. The derived Gibbs sampling based inference offers an efficient inference scheme in a practical dimension (for less than 1000 row-entities and column-entities), but scaling up can be crucial in many domains. In general, this is a difficult question due to the memory complexity of storing the kernels and the notoriously hard problem

of scaling up Gibbs sampling. We plan to investigate low-rank kernel approximations, alternative MCMC schemes and GPU-based implementations.

Whereas the proposed method can be applied in multiple domains, drug–target interaction prediction expectedly remains a central challenge. In this domain important open issues are as follows: the availability of multiple interaction scores (*i.e.* multiple bioactivity data), the selection of row-entities and column-entities (*i.e.* compounds and targets) using prioritization methods and the interpretation of the results (*i.e.* finding the interesting predictions using linked open data and semantic technologies (Williams et al., 2012)).

Acknowledgments

This work has been supported by OTKA 112915, the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (P. Antal) and Richter Témapályázat 2014.

Appendix A.

The conditional for $\boldsymbol{U}$ can be computed as follows.

$$\begin{aligned}
\ln p(\boldsymbol{U}, \boldsymbol{\Theta}^{\backslash \boldsymbol{U}}) &\propto \left[\sum_i \sum_j -\frac{1}{2} \gamma_i^c \gamma_j^r \left(\boldsymbol{R}_{ij} - \boldsymbol{u}_i^T \boldsymbol{v}_j \right)^2 I_{ij} \right] + \left[\sum_l -\frac{1}{2\lambda_l} \left(\boldsymbol{U}_{l:} - \boldsymbol{m}_{l:}^u \right)^T \boldsymbol{S}_l^u \left(\boldsymbol{U}_{l:} - \boldsymbol{m}_{l:}^u \right) \right] \\
&= \left[\sum_i \sum_j -\frac{1}{2} \gamma_i^c \gamma_j^r \left(\boldsymbol{R}_{ij} - \boldsymbol{u}_i^T \boldsymbol{v}_j \right)^2 I_{ij} \right] + \left[\sum_l -\frac{1}{2\lambda_l} \sum_i \sum_n \left(\boldsymbol{U}_{li} - \boldsymbol{m}_{li}^u \right) \boldsymbol{S}_{l_{in}}^u \left(\boldsymbol{U}_{ln} - \boldsymbol{m}_{ln}^u \right) \right] \\
&= \sum_i -\frac{1}{2} \left[\gamma_i^c \sum_j \gamma_j^r \left(\boldsymbol{R}_{ij} - \boldsymbol{u}_i^T \boldsymbol{v}_j \right)^2 I_{ij} + \sum_l \lambda_l^{-1} \sum_n \left(\boldsymbol{U}_{li} - \boldsymbol{m}_{li}^u \right) \boldsymbol{S}_{l_{in}}^u \left(\boldsymbol{U}_{ln} - \boldsymbol{m}_{ln}^u \right) \right] \\
&\propto \sum_i -\frac{1}{2} \left[\boldsymbol{u}_i^T \left(\sum_j \gamma_j^r \boldsymbol{v}_j \boldsymbol{v}_j^T I_{ij} + \mathrm{diag}_l(\lambda_l^{-1} \boldsymbol{S}_{l_{ii}}^u) \right) \boldsymbol{u}_i \right. \\
&\qquad \left. + 2 \cdot \left(-\gamma_i^c \sum_j \gamma_j^r \boldsymbol{R}_{ij} \boldsymbol{v}_j I_{ij} + \mathrm{vec}_l \left(\lambda_l^{-1} \sum_{n \neq i} \boldsymbol{S}_{l_{in}}^u \left(\boldsymbol{U}_{ln} - \boldsymbol{m}_{ln}^u \right) \right) \right)^T \boldsymbol{u}_i \right],
\end{aligned}$$

so by completing the square $\boldsymbol{u}_i^T \boldsymbol{A} \boldsymbol{u}_i + 2 \boldsymbol{b}^T \boldsymbol{u}_i \propto (\boldsymbol{u}_i - (-\boldsymbol{A}^{-1} \boldsymbol{b}))^T \boldsymbol{A} (\boldsymbol{u}_i - (-\boldsymbol{A}^{-1} \boldsymbol{b}))$

$$\begin{aligned}
p(\boldsymbol{U} | \boldsymbol{\Theta}^{\backslash \boldsymbol{U}}) &= \prod_i \mathcal{N}(\boldsymbol{u}_i | \boldsymbol{\psi}_i, \boldsymbol{\Lambda}_i^{-1}), \\
\boldsymbol{\Lambda}_i &= \gamma_i^c \sum_j \gamma_j^r \left(\boldsymbol{v}_j \boldsymbol{v}_j^T \right) I_{ij} + \mathrm{diag}_l(\lambda_l^{-1} \boldsymbol{S}_{l_{ii}}^u), \\
\boldsymbol{\psi}_i &= -\boldsymbol{\Lambda}_i^{-1} \left[-\gamma_i^c \sum_j \gamma_j^r \left(\boldsymbol{R}_{ij} \boldsymbol{v}_j \right) I_{ij} + \mathrm{vec}_l \left(\lambda_l^{-1} \sum_{n \neq i} \boldsymbol{S}_{l_{in}}^u \left(\boldsymbol{U}_{ln} - \boldsymbol{m}_{ln}^u \right) \right) \right].
\end{aligned}$$

The conditionals for $\boldsymbol{V}$ can be found in (Salakhutdinov and Mnih, 2008b). The conditionals for $\boldsymbol{\gamma}^c$, $\boldsymbol{\gamma}^r$, $\boldsymbol{\sigma}^{nc}$ and $\boldsymbol{\sigma}^{nr}$ follow from the standard expressions for the posterior of conjugate models using

the Gamma prior. For $\boldsymbol{\lambda}$, using the conjugacy of the Inverse Gamma prior yields

$$\begin{aligned} p(\boldsymbol{\lambda}, \boldsymbol{\Theta}^{\setminus \boldsymbol{\lambda}}) \propto & \left[\prod_{l=1}^{L} \frac{b_l^{w a_l^w}}{\Gamma(a_l^w)} \lambda_l^{-a_l^w - 1} \exp\left\{ -\frac{b_l^w}{\lambda_l} \right\} \right] \\ & \times \left[\prod_{l=1}^{L} \left((2\pi)^I \lambda_l \left| \boldsymbol{S}_l^{u-1} \right| \right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\lambda_l} (\boldsymbol{U}_{l:} - \boldsymbol{m}_{l:}^u)^T \boldsymbol{S}_l^u (\boldsymbol{U}_{l:} - \boldsymbol{m}_{l:}^u) \right\} \right] \\ \propto & \prod_{l=1}^{L} \lambda_l^{-(a_l^w + \frac{1}{2}) - 1} \exp\left\{ -\frac{b_l^w + 1/2 \cdot (\boldsymbol{U}_{l:} - \boldsymbol{m}_{l:}^u)^T \boldsymbol{S}_l^u (\boldsymbol{U}_{l:} - \boldsymbol{m}_{l:}^u)}{\lambda_l} \right\}, \end{aligned}$$

i.e.

$$\begin{aligned} p(\boldsymbol{\lambda} | \boldsymbol{\Theta}^{\setminus \boldsymbol{\lambda}}) &= \prod_{l=1}^{L} \mathcal{IG}(\lambda_l | \alpha_l^w, \beta_l^w), \\ \alpha_l^w &= a_l^w + \frac{1}{2}, \\ \beta_l^w &= b_l^w + \frac{1}{2} (\boldsymbol{U}_{l:} - \boldsymbol{m}_{l:}^u)^T \boldsymbol{S}_l^u (\boldsymbol{U}_{l:} - \boldsymbol{m}_{l:}^u). \end{aligned}$$

References

R. P. Adams, G. E. Dahl, and I. Murray. Incorporating side information in probabilistic matrix factorization with Gaussian Processes. In P. Grünwald and P. Spirtes, editors, *UAI*, pages 1–9. AUAI Press, 2010. ISBN 978-0-9749039-6-5.

D. Agarwal and B.-C. Chen. Regression-based latent factor models. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 19–28, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: http://doi.acm.org/10.1145/1557019.1557029. URL `http://doi.acm.org/10.1145/1557019.1557029`.

A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Kruger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, and J. P. Overington. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, 42:D1083–1090, Jan 2014.

K. Buza. Drug-target interaction prediction with hubness-aware machine learning. In *11th IEEE International Symposium on Applied Computational Intelligence and Informatics*, 2016.

H. Eckert and J. Bajorath. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today*, 12(5-6):225–233, Mar 2007.

M. D. Ekstrand, J. T. Riedl, and J. A. Konstan. Collaborative filtering recommender systems. *Found. Trends Hum.-Comput. Interact.*, 4(2):81–173, Feb. 2011. ISSN 1551-3955. doi: 10.1561/1100000009. URL `http://dx.doi.org/10.1561/1100000009`.

M. Gönen, S. A. Khan, and S. Kaski. Kernelized Bayesian matrix factorization. In *Proceedings of ICML 2013, the 30th International Conference on Machine Learning*, volume 28 of *JMLR W&CP*, pages 864–872. JMLR, 2013. Implementations in Matlab are available at http://research.ics.aalto.fi/mi/software/kbmf/.

J. M. Hernández-Lobato, N. Houlsby, and Z. Ghahramani. Probabilistic matrix factorization with non-random missing data. In *ICML*, volume 32 of *JMLR Proceedings*, pages 1512–1520. JMLR.org, 2014.

S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. Le Novere, H. Parkinson, E. Birney, and A. M. Jenkinson. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, 30(9): 1338–1339, May 2014.

S. Park, Y.-D. Kim, and S. Choi. Hierarchical Bayesian matrix factorization with side information. In F. Rossi, editor, *IJCAI*. IJCAI/AAAI, 2013. ISBN 978-1-57735-633-2.

R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov Chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 880–887, New York, NY, USA, 2008a. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390267. URL `http://doi.acm.org/10.1145/1390156.1390267`.

R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20, 2008b.

J. Simm, A. Arany, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, H. Ceulemans, and Y. Moreau. Macau: Scalable Bayesian Multi-relational Factorization with Side Information using MCMC. *ArXiv e-prints*, Sept. 2015.

A. P. Singh and G. J. Gordon. A Bayesian matrix factorization model for relational data. *CoRR*, abs/1203.3517, 2012.

C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The chemistry development kit (cdk): an open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2):493–500, 2003. doi: 10.1021/ci025584y.

A. J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E. L. Willighagen, C. T. Evelo, N. Blomberg, G. Ecker, C. Goble, and B. Mons. Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today*, 17(21-22):1188–1198, Nov 2012.

J. Yang, Z. Li, X. Fan, and Y. Cheng. Drug-disease association and drug-repositioning predictions in complex diseases using causal inference-probabilistic matrix factorization. *J Chem Inf Model*, 54(9):2562–2569, Sep 2014.

J. Yoo and S. Choi. Bayesian matrix co-factorization: Variational algorithm and cramer-rao bound. In *Proceedings of the ECML/PKDD 2011*, 2011.

P. Zakeri, J. Simm, A. Arany, S. Elshal, and Y. Moreau. Gene prioritization through Bayesian matrix factorization. In *Proc. of the ISMB/ECCB 2015 Conference*, 2015.

T. Zhou, H. Shan, A. Banerjee, and G. Sapiro. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *SDM*, pages 403–414. SIAM / Omnipress, 2012. ISBN 978-1-61197-282-5.

M. Zitnik and B. Zupan. Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold. *Pac Symp Biocomput*, pages 400–411, 2014.