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

Bence Bolgár and Péter Antal

Budapest University of Technology and Economics
Department of Measurement and Information Systems

September 6, 2016

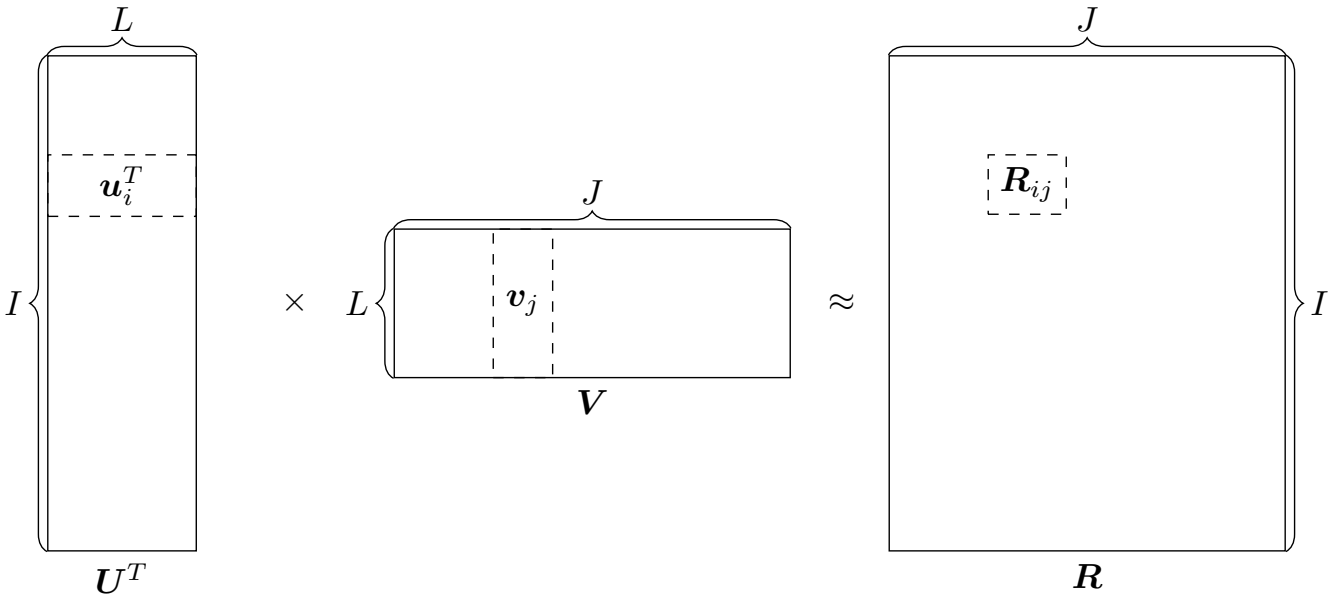
Goals

From known drug–target interaction measurements, estimate binding affinities for other drug–target pairs.

Problems:

1. Incorporating **entity-wise** "side information", *e.g.* molecular structures, side-effect profiles *etc.*
2. Incorporating other estimates of **pairwise** interaction data, *e.g.* molecular docking simulations.
3. Measurement data are highly incomplete, *i.e.* most of the drug–target pairs are not measured or kept in secret. We aim to exploit the information hidden in this "missingness pattern".

Matrix factorization



Find **complete** factors $\boldsymbol{U} \in \mathbb{R}^{L \times I}$ and $\boldsymbol{V} \in \mathbb{R}^{L \times J}$, such that $\boldsymbol{U}^T\boldsymbol{V} \approx \boldsymbol{R}$.

- $\boldsymbol{u}_i \sim i$th drug,
- $\boldsymbol{v}_j \sim j$th target,
- $\boldsymbol{R}_{ij} \sim$ their binding affinity,
- $L \ll I, J$ free parameter (rank).

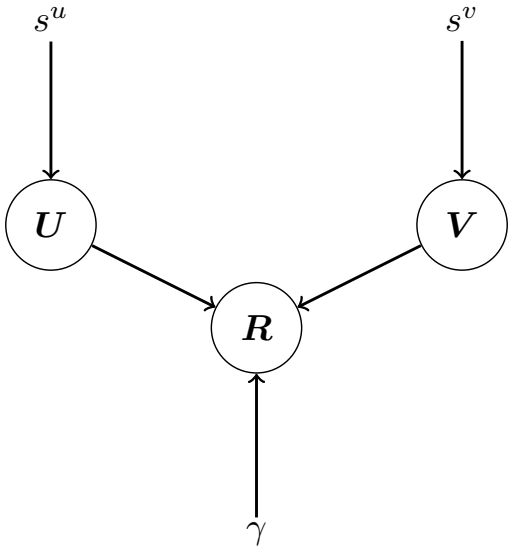
Singular Value Decomposition

$$\arg\min_{U,V} \left\| \boldsymbol{R} - \boldsymbol{U}^T \boldsymbol{V} \right\|_F$$

Solve for $\boldsymbol{U}$ and $\boldsymbol{V}$ using SVD and compose $\boldsymbol{U}, \boldsymbol{V}$ from the vectors corresponding to the L largest singular values. However:

- Does not handle missing entries,
- $\boldsymbol{U}, \boldsymbol{V}$ can have arbitrarily large values $\Rightarrow$ overfitting.

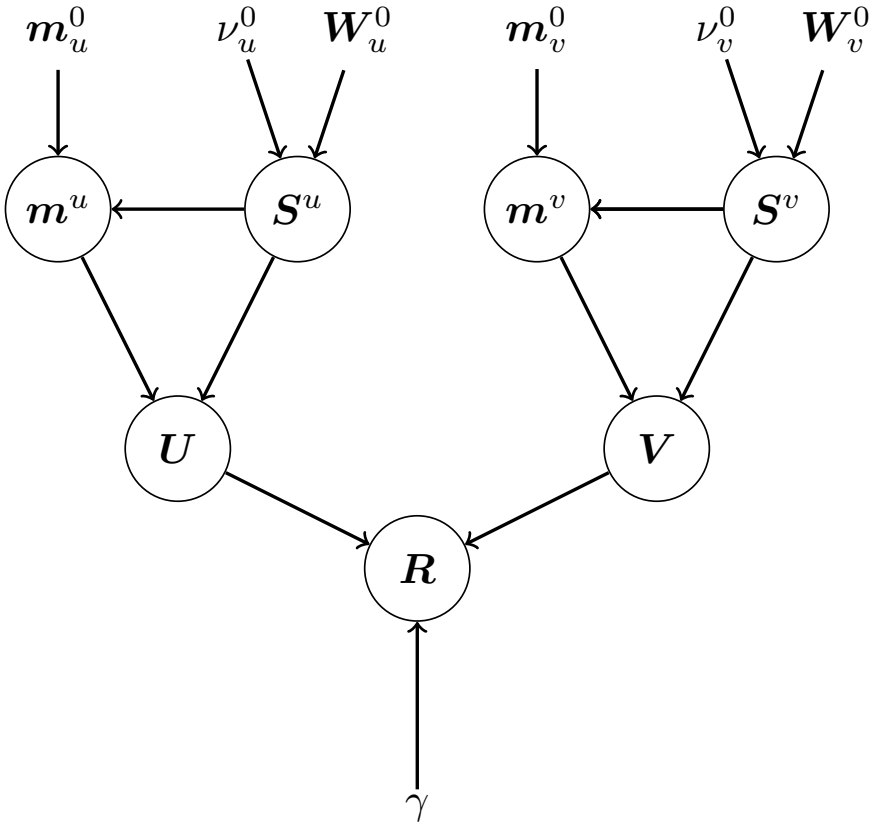
Probabilistic Matrix Factorization (Salakhutdinov *et al.*, 2008.)



$$p(\boldsymbol{R}|\boldsymbol{U},\boldsymbol{V},\gamma) = \prod_i \prod_j \left[\mathcal{N}(\boldsymbol{R}_{ij}|\boldsymbol{u}_i^T\boldsymbol{v}_j,\gamma^{-1})\right]^{I_{ij}}$$

$$p(\boldsymbol{U}|s^u) = \prod_i \mathcal{N}(\boldsymbol{u}_i|\boldsymbol{0},s^u\boldsymbol{I})$$

Bayesian Probabilistic Matrix Factorization (Salakhutdinov *et al.*, 2008.)



$$p(\boldsymbol{U}|\boldsymbol{m}^u, \boldsymbol{S}^u) = \prod_{i=1}^{I} \mathcal{N}(\boldsymbol{u}_i|\boldsymbol{m}_i^u, \boldsymbol{S}_i^{u\,-1}),$$

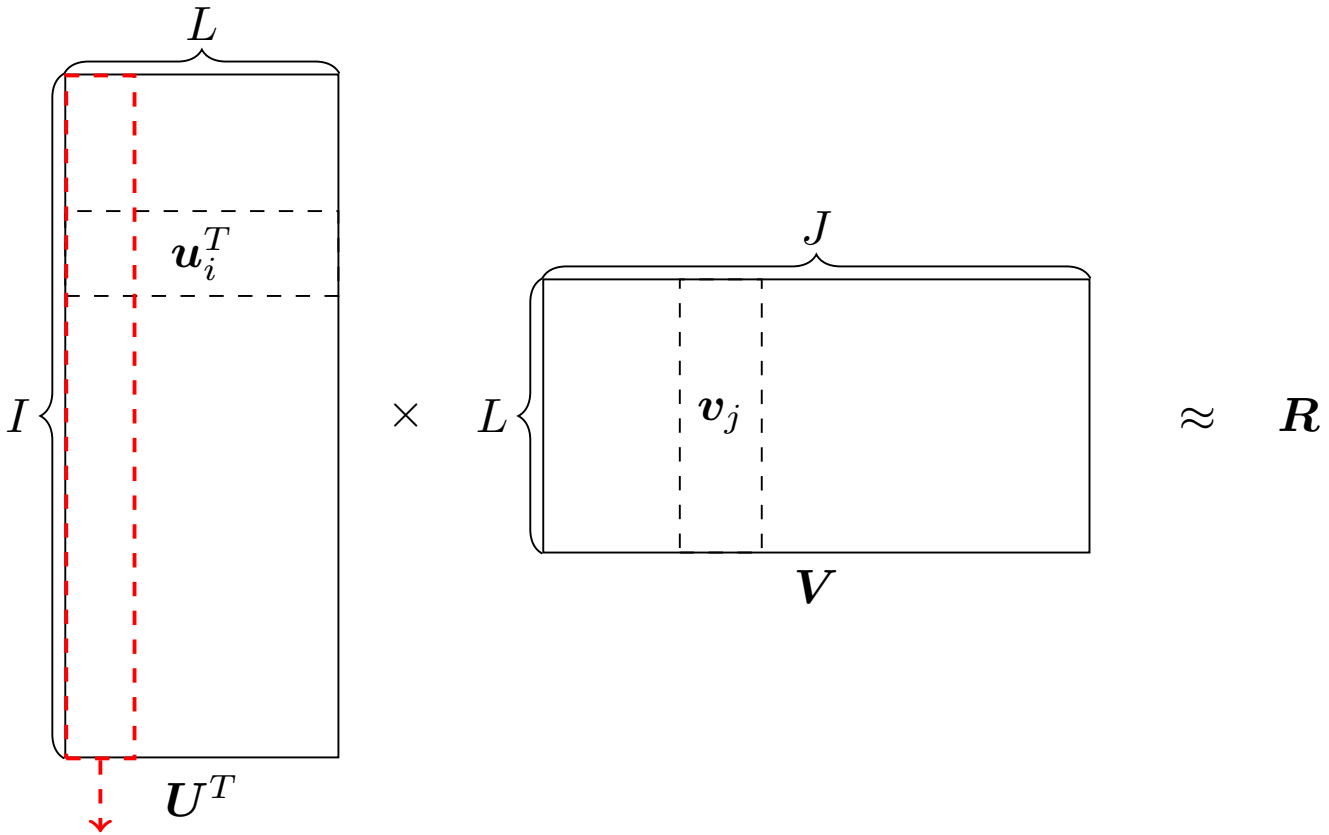
$$p(\boldsymbol{m}^u, \boldsymbol{S}^u|\boldsymbol{m}_u^0, \nu_u^0, \boldsymbol{W}_u^0) = \mathcal{NW}(\boldsymbol{m}^u, \boldsymbol{S}^u|\boldsymbol{m}_u^0, \kappa, \boldsymbol{W}_u^0, \nu_u^0),$$

Incorporating side information

- In chemoinformatics, side information usually come in the form of high-dimensional real vectors encoding chemical structure ("fingerprints").
- Very often, similarity matrices are computed ("Similar Property Principle") and used in prioritization algorithms ("Virtual Screening").
- With a suitable choice of similarity measure(s), these matrices are symmetric and PD.

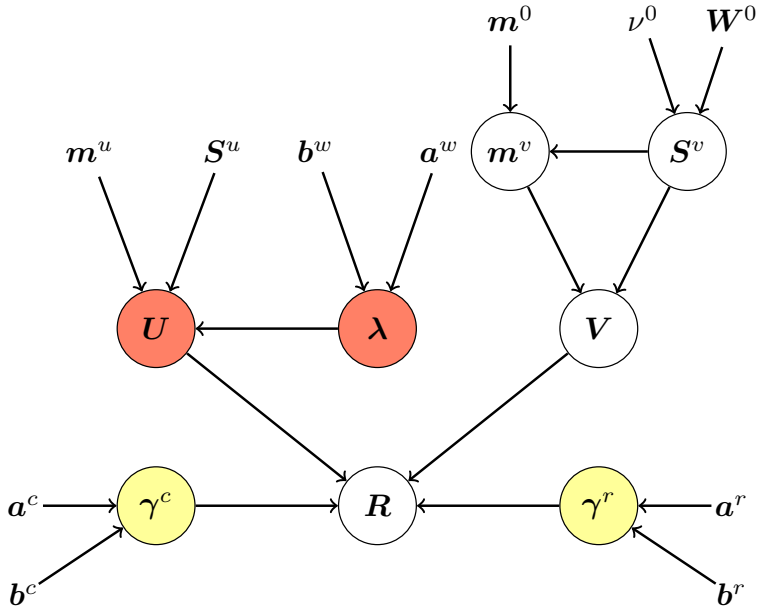
Let's use them as a covariance matrices of L independent Gaussian Processes over the **rows** of $\boldsymbol{U}$, enforcing similarities over $\boldsymbol{u}_i$'s (Zhou *et al.*, 2012).

Incorporating side information with Gaussian Processes



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

Model so far

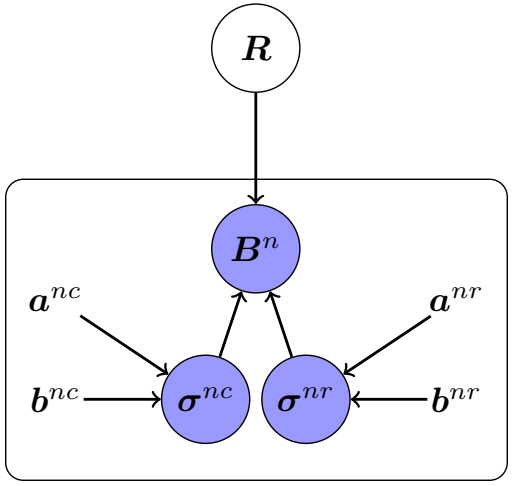


$$p(\boldsymbol{\lambda}|\boldsymbol{a}^w, \boldsymbol{b}^w) = \prod_{l=1}^{L} \mathcal{IG}(\lambda_l|a_l^w, b_l^w), \qquad (\bullet \text{ weighted features})$$

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

$$p(\boldsymbol{\gamma}^c|\boldsymbol{a}^c, \boldsymbol{b}^c) = \prod_{i=1}^{I} \mathcal{G}a(\gamma_i^c|a_i^c, b_i^c). \qquad (\bullet \textbf{ per-entity} \text{ precison})$$

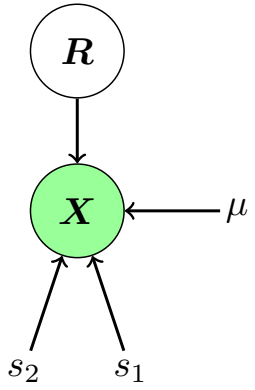
Incorporating background knowledge



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

$$p(\boldsymbol{\sigma}^{nc}|\boldsymbol{a}^{nc}, \boldsymbol{b}^{nc}) = \prod_{i=1}^{I}\mathcal{G}a(\sigma_i^{nc}|a_i^{nc}, b_i^{nc}), \qquad (\bullet \textbf{ per-entity} \text{ precision})$$

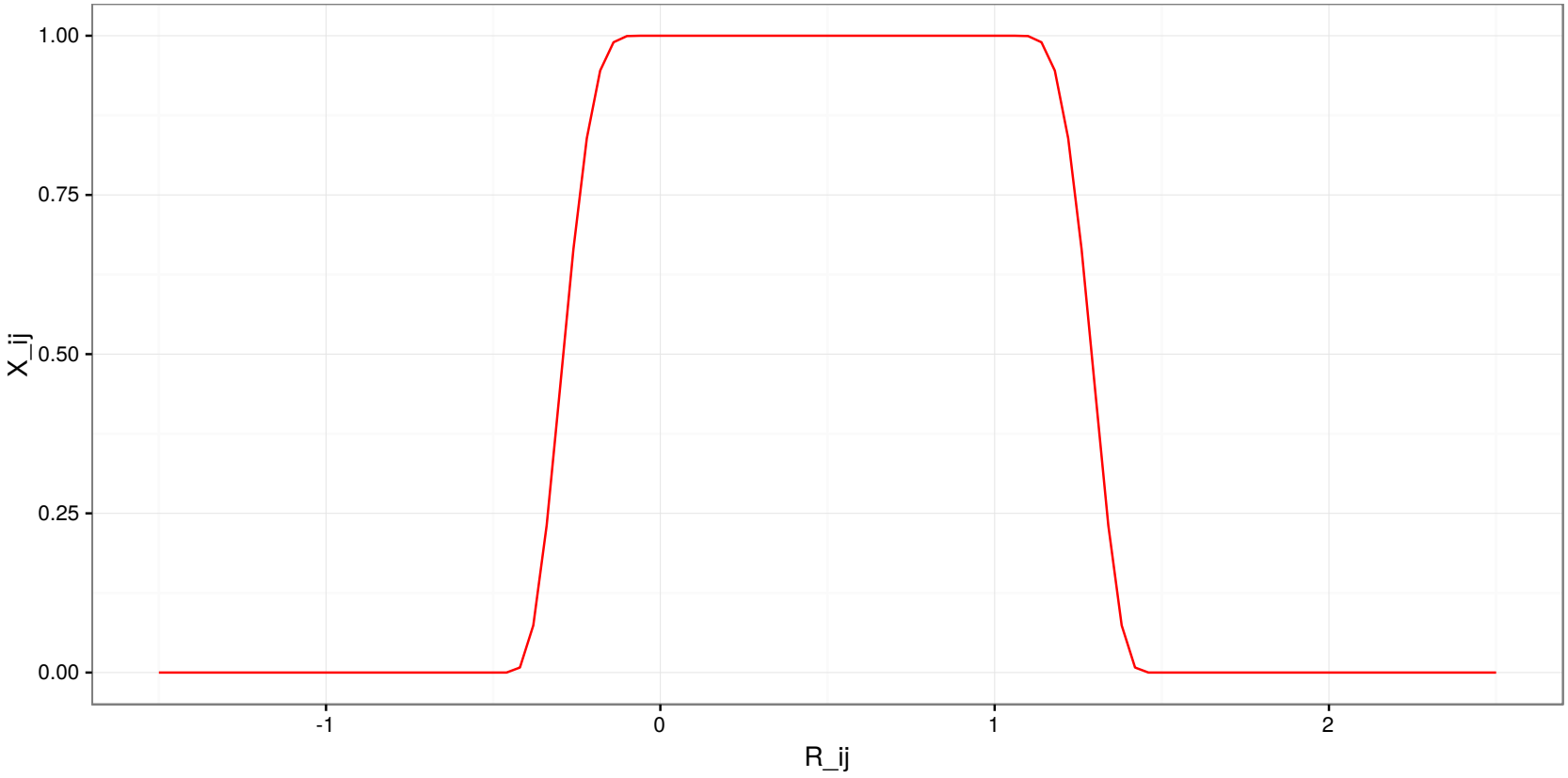
Handling missing data



$$p(\boldsymbol{X}|\boldsymbol{R}, s_1, s_2, \mu) = \prod_i \prod_j \mathcal{B}\left(\boldsymbol{X}_{ij} | f(\boldsymbol{R}_{ij}, s_1, s_2, \mu)\right), \qquad (\bullet \text{ missingness})$$

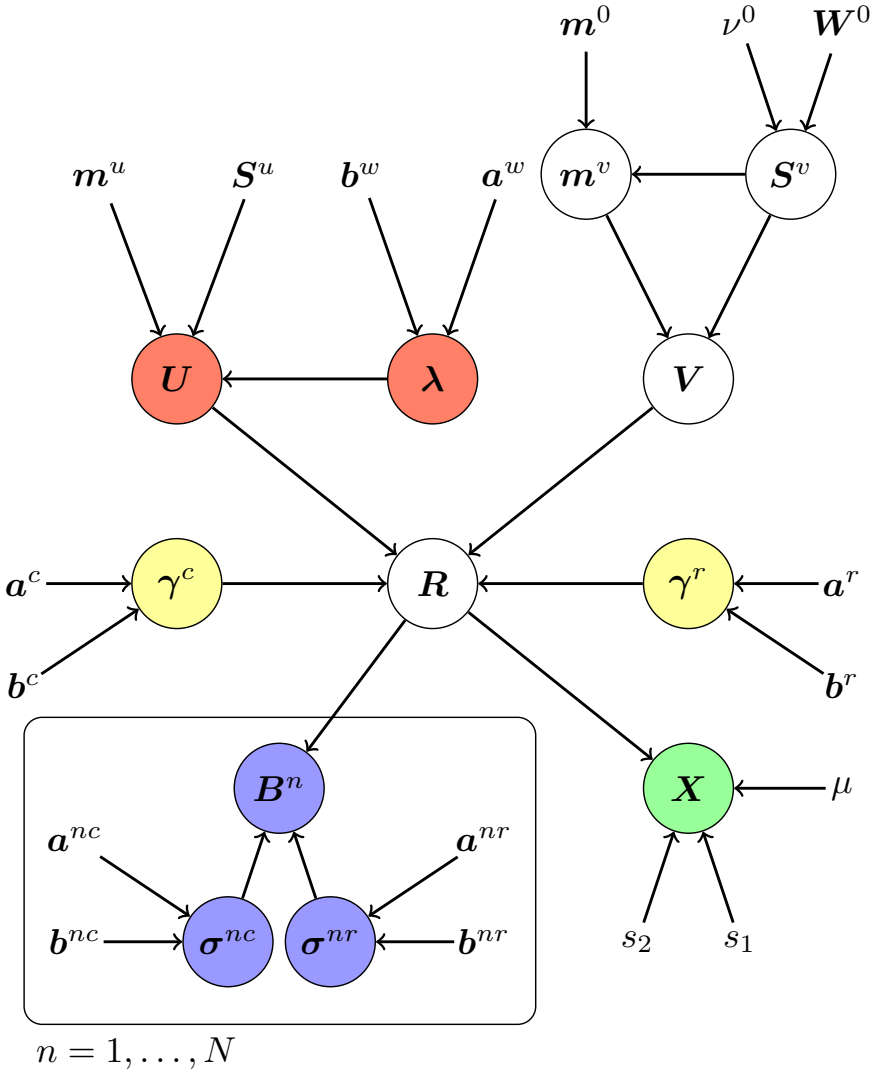
$$f(x, s_1, s_2, \mu) = \begin{cases} 1, & \text{if } |x-\mu| < s_1 \\ 0, & \text{if } |x-\mu| \geq s_2 \\ \sigma\left(-\frac{s_1^2+s_2^2-2(x-\mu)^2}{((x-\mu)^2-s_1^2)\cdot((x-\mu)^2-s_2^2)}\right) & \text{otherwise.} \end{cases}$$

Bump function



“irrelevant” “public” “secret”

Complete model



Gibbs sampling

This choice of conjugate priors makes the derivation of conditionals trivial for almost all variables, except[1]:

- Sampling $\boldsymbol{U}$. Still Gaussian, mean vector and covariance matrix still efficiently computable with BLAS.
- Sampling $\boldsymbol{\lambda}$. Still $\mathcal{IG}$, looks very much like the usual update equation with a slightly different quadratic term in the second parameter.
- Sampling $\boldsymbol{R}$. We have not found the correct normalization coefficient yet, moreover, the conditional is in general not log-concave. Therefore we utilize slice sampling for this step.

[1] Proofs included in the Appendix of the article.

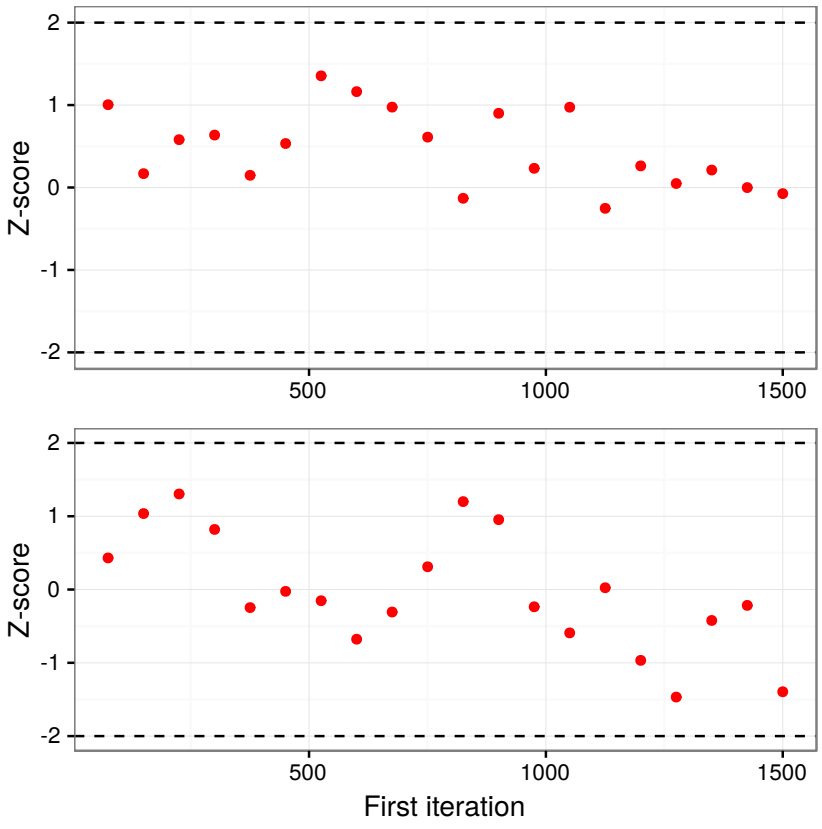
Root mean squared error

	2K+MDM	HuTolt 2K	1K	0K	Macau	BPMF
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

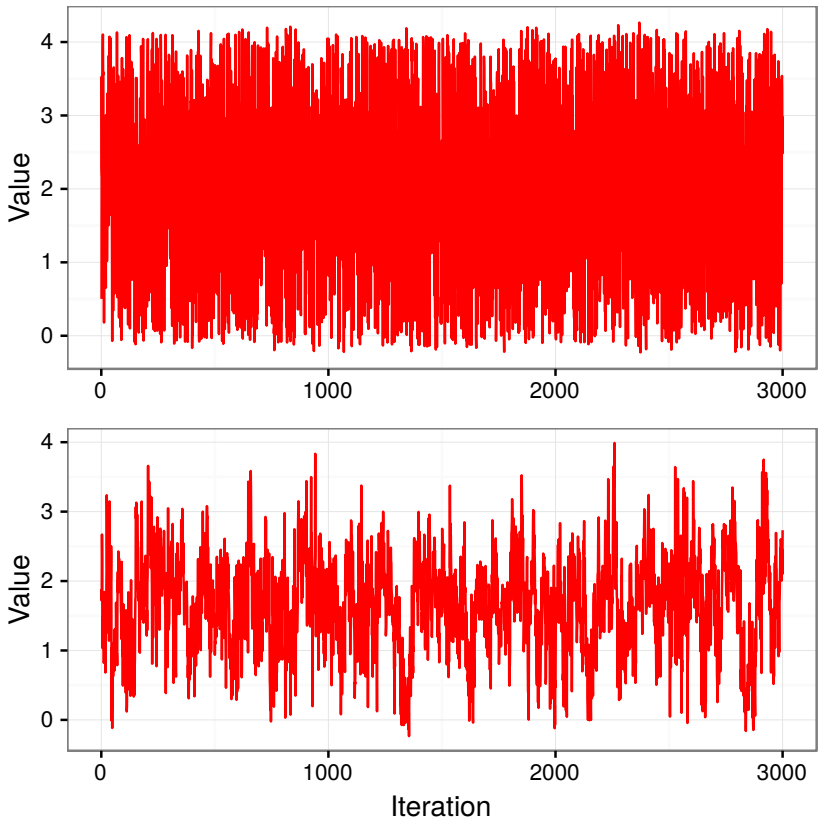
Settings:

- 37 psychiatric drugs from the `N06*` ATC class with 82 targets.
- 446 binding affinities from the ChEMBL database (14.7% completeness).
- Klekota–Roth and MACCS fingerprints with the Tanimoto similarity measure.
- For a fair comparison, the background knowledge module was not utilized.
- 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 similarity).
- Evaluated with 100-fold $80\% - 20\%$ cross-validation, compared to
  - BPMF (Salakhutdinov *et al.*, 2008).
  - Macau (Simm *et al.*, 2016).

Convergence for high and low affinities

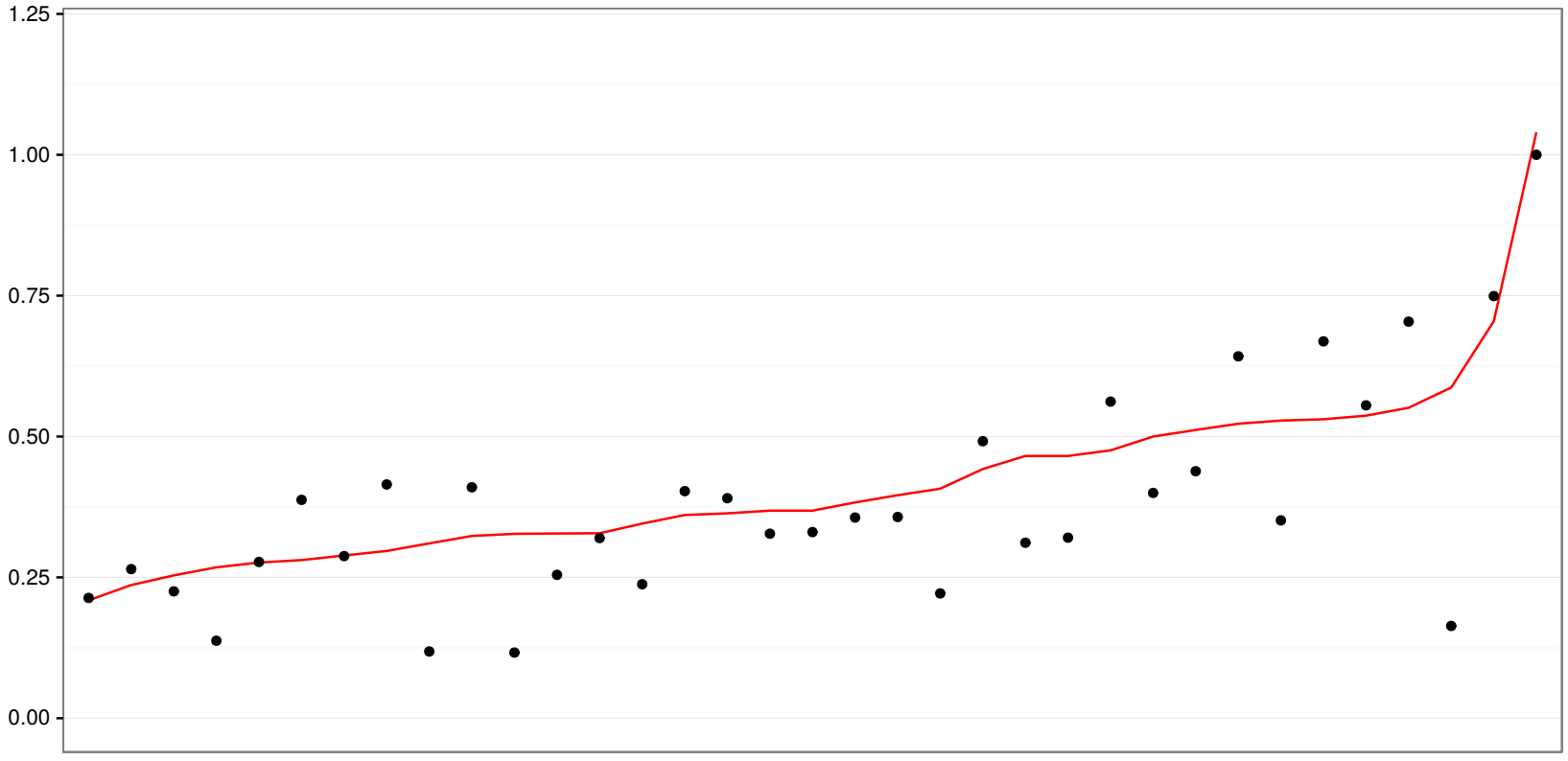


Geweke–Brooks



Trace plot

Correlation between fingerprint and factor similarities



Future work

- Investigating the detailed effects of the modules by systematically evaluating their combinations.
- Investigating the detailed effects of the hyperparameters.
- Scaling up using parallel implementations (GPGPU), alternative MCMC methods, low-rank approximation.
- Handling multiple interaction scores in a multitask fashion.

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.