Large-scale probabilistic non-linear matrix factorization for drug discovery

Xiangju Qin, Paul Blomstedt, Samuel Kaski
Helsinki Institute for Information Technology HIIT
Department of Computer Science
Aalto University

1 Motivation

Predicting bioactivities between chemical compounds and protein targets or cell lines is a central challenge in drug discovery. The prediction can be done by using machine learning algorithms which learn from chemical (compound descriptions and chemical structures) and genomic (protein structures and gene expression) spaces as well as compound-target bioactivity data. Although there has been a lot of research applying machine learning to solve problems in chemogenomics, many of the techniques have limitations that prevent them from being applied to industry-scale data, which is characteristically very large, extremely sparsely observed, imbalanced and noisy. Standard chemogenomics models depend heavily on a specific type of input data and are designed for single-task learning, while multi-task learning has proven to improve the performance of each individual task by sharing representations among tasks. Additionally, a large proportion of published research has produced black-box models, providing no practical insight on feature significance and confidence estimation for predictions.

2 Matrix factorization

The problem of bioactivity prediction can be interpreted as an instance of collaborative filtering, for which matrix factorization is one of the most widely used approaches, see Figure 2. Bayesian matrix factorization (BMF) [Salakhutdinov and Mnih, 2008] formulates the task as a probabilistic generative model, providing an elegant and flexible framework to address the above concerns (with the exception of scalability). Its many attractive features include the ability to quantify uncertainty, handle missing values and predict with confidence estimation. Besides predicting missing compound-target pairs using relations learned from the observed bioactivity, the BMF framework can be used for:

- Optimizing biological assays or compounds in a desired way by utilizing the obtained latent features in the model.
  - Identifying compounds that promote or suppress the bioacitivity of a target, which may be related to some disease.
  - Discovering compounds sharing similar chemical structures, which could affect a target disease in a similar way.

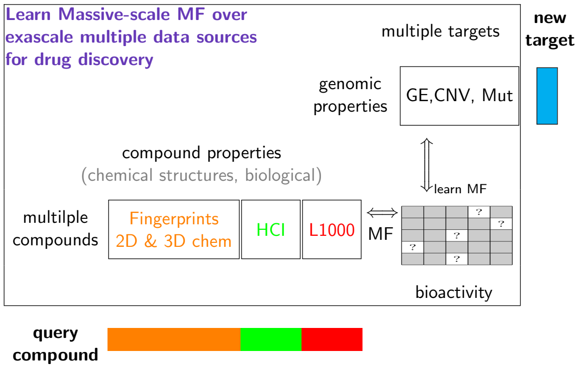

Figure 1: An illustration of drug discovery with multi-view matrix factorization.

- Learning relations between multiple data sources, such as identifying drugs with positive response on targets.

While in many problems, it is reasonable to assume that at least *some* elements have been observed in all rows and columns of the data matrix, a more challenging prediction problem ensues when entire rows (or columns) of the matrix are unobserved. This is referred to as *out-of-matrix* prediction. In such cases, prediction is in general infeasible without additional side-data providing information about the missing rows. The inclusion of side-data can be approached as an instance of multi-view learning [Damianou *et al.*, 2012a; Klami *et al.*, 2015], where the missing information in the target view is complemented with information collected in additional data views, see Figure 1.

3 Large-scale probabilistic non-linear matrix factorization

Although linear models, such as BMF, are often expressive enough to produce reasonable predictions, non-trivial prediction problems may require even more flexible, non-linear models for improved accuracy. Bayesian Gaussian process latent-variable models (GP-LVM) [Titsias and Lawrence, 2010] provide a non-linear generalization of BMF, but they

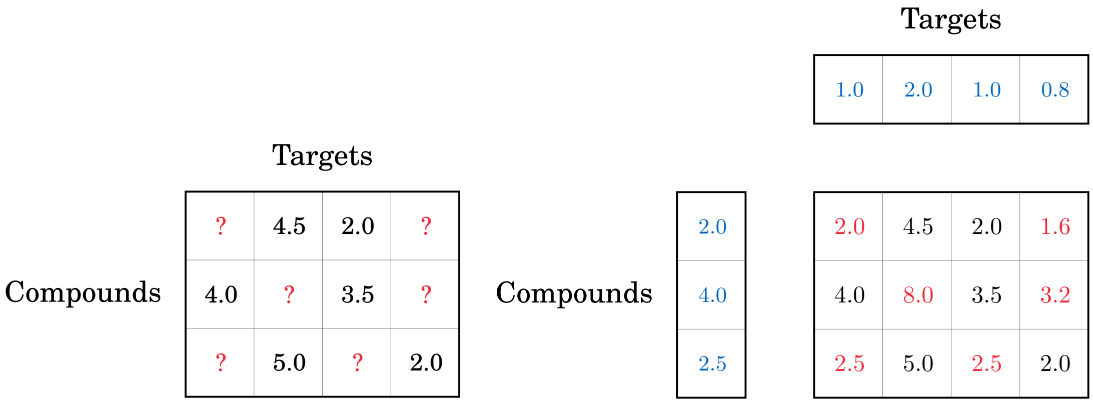

Figure 2: Schematic overview of matrix factorization: **black** numbers are observed bioactivities, red numbers are predicted bioactivities, blue numbers are estimated latent features.

are extremely challenging to scale up. Current distributed solutions to scale up GP-LVM [Dai *et al.*, 2014; Gal *et al.*, 2014] incur a communication overhead, which becomes prohibitive as the degree of parallelism is increased with the problem size. Embarrassingly parallel solutions [Deisenroth and Ng, 2015] would sidestep the problem of communication but are not directly applicable to latent-variable models due to their inherent unidentifiability. In this work, we propose a distributed solution for Bayesian GP-LVM, which borrows elements from a recently proposed hierarchical embarrassingly parallel inference scheme for BMF [Qin *et al.*, 2017]. In our framework, parallel computations are coupled using the output of an initial stage of inference, in which only one data subset is processed. Thus, compared to fully embarrassingly parallel methods, the strategy adds only one extra stage of communication, that of propagating the information from the initial subset to the remaining subsets. The propagation is done using a modified, probabilistic version of incremental learning for GP-LVM [Yao *et al.*, 2011], which is similar in spirit to dynamical GP-LVM [Damianou *et al.*, 2016].

Preliminary results indicate that our distributed framework achieves predictive accuracy close to the full model (on medium-scale data), and has the capability to scale up probabilistic non-linear matrix factorization (Bayesian GP-LVM and its multi-view extension [Damianou *et al.*, 2012b]) to industry-scale problems, such as the ExCAPE-DB chemogenomics data set [Sun *et al.*, 2017].

References

[Dai *et al.*, 2014] Zhenwen Dai, Andreas Damianou, James Hensman, and Neil Lawrence. Gaussian process models with parallelization and GPU acceleration. *arXiv preprint arXiv:1410.4984*, 2014.

[Damianou *et al.*, 2012a] Andreas Damianou, Carl Henrik Ek, Michalis K. Titsias, and Neil D. Lawrence. Manifold relevance determination. In *Proceedings of the 29th International Conference in Machine Learning*, 2012.

[Damianou *et al.*, 2012b] Andreas Damianou, Carl Henrik Ek, Michalis K. Titsias, and Neil D. Lawrence. Manifold relevance determination. In *Proceedings of the 29th International Conference in Machine Learning*, 2012.

[Damianou *et al.*, 2016] Andreas C. Damianou, Michalis K. Titsias, and Neil D. Lawrence. Variational inference for latent variables and uncertain inputs in Gaussian processes. *Journal of Machine Learning Research*, 17(42):1–62, 2016.

[Deisenroth and Ng, 2015] Marc Peter Deisenroth and Jun Wei Ng. Distributed Gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[Gal *et al.*, 2014] Yarin Gal, Mark van der Wilk, and Carl Edward Rasmussen. Distributed variational inference in sparse Gaussian process regression and latent variable models. In *Advances in Neural Information Processing Systems 27*, pages 3257–3265, 2014.

[Klami *et al.*, 2015] Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski. Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9):2136–2147, 2015.

[Qin *et al.*, 2017] Xiangju Qin, Paul Blomstedt, Eemeli Leppäaho, Pekka Parviainen, and Samuel Kaski. Distributed Bayesian matrix factorization with limited communication. *arXiv preprint arXiv:1703.00734*, 2017.

[Salakhutdinov and Mnih, 2008] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning*, pages 880–887, 2008.

[Sun *et al.*, 2017] Jiangming Sun, Nina Jeliazkova, Vladimir Chupakhin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, Nikolay Kochev, Thomas J. Ashby, and Hongming Chen. ExCAPE-DB: an integrated large scale dataset facilitating big data analysis in chemogenomics. *Journal of Cheminformatics*, 9(1):17, 2017.

[Titsias and Lawrence, 2010] Michalis Titsias and Neil Lawrence. Bayesian Gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851. PMLR, 2010.

[Yao *et al.*, 2011] Angela Yao, Juergen Gall, Luc V Gool, and Raquel Urtasun. Learning probabilistic non-linear latent variable models for tracking complex activities. In *Advances in Neural Information Processing Systems 24*, pages 1359–1367, 2011.