SMURFF: a High-Performance Framework for Matrix Factorization Methods (extended abstract)

Tom Vander Aa[1], Imen Chakroun[1], Thomas J. Ashby[1], Jaak Simm[2], Adam Arany[2], Yves Moreau[2], Thanh Le Van[3], José Felipe Golib Dzib[3], Jörg Wegner[3], Vladimir Chupakhin[3], Hugo Ceulemans[3], Roel Wuyts[1], and Wilfried Verachtert[1]

[1] ExaScience Life Lab at imec, Leuven, Belgium tom.vanderaa@imec.be
[2] ESAT-STADIUS, KU Leuven, Leuven, Belgium
[3] Janssen Pharmaceutica, Beerse, Belgium

Abstract This is a 2-page condensed abstract of the paper published at [1].

Bayesian Matrix Factorization Matrix factorization is a common machine learning technique for recommender systems, like books for Amazon or movies for Netflix [2]. The Bayesian Matrix Factorization (BMF) variant is especially powerful because it produces good results and is relatively robust against overfitting. As sketched in Figure 1, the idea of these methods is to approximate the user-movie rating matrix R as a product of two low-rank matrices U and V such that $R \approx U \times V$. In this way U and V are constructed from the known ratings in R, which is usually very sparsely filled. The recommendations can be made from the approximation $U \times V$ which is dense.

High-Performance Matrix Factorization with SMURFF While BMF is powerful, it is computationally very intensive and thus more challenging to implement for large datasets. In this work we present SMURFF a high-performance feature-rich framework to compose and construct different Bayesian matrix-factorization methods, namely:

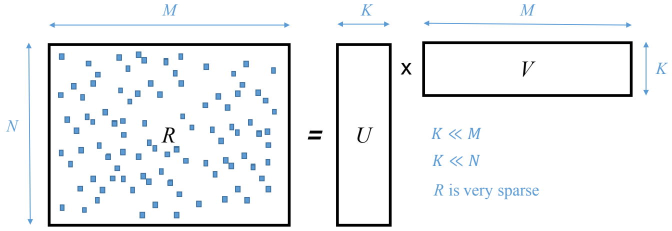

Fig. 1. Low-rank Matrix Factorization

- BPMF, the basic version [4];
- Macau, adding support for high-dimensional side information to the factorization [7];
- GFA, doing Group Factor Analysis [6].

Applications of SMURFF The framework has been used very successfully in drug discovery. Here a key problem is the identification of candidate molecules that affect proteins associated with diseases [3]. SMURFF with Macau has been used to predict compound-on-protein activity on a matrix with more than one million compounds (rows) and several thousand proteins (columns) [8]. Chemical fingerprints were used for the compounds, in both dense and sparse formats. This has led to important new insights and potential new compounds to be used in drug discovery.

Getting SMURFF SMURFF is available as open-source [5] and can be used both on a supercomputer and on a desktop or laptop machine. Documentation and several examples are provided as Jupyter notebooks using SMURFF's high-level Python API.

Acknowledgments This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 801051.

References

1. T. Vander Aa, et.al, "SMURFF: a High-Performance Framework for Matrix Factorization," *2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pp. 304–308, 2019. [Online]. Available: https://arxiv.org/abs/1904.02514
2. C. A. Gomez-Uribe and N. Hunt, "The netflix recommender system: Algorithms, business value, and innovation," *ACM Trans. Manage. Inf. Syst.*, vol. 6, no. 4, pp. 13:1–13:19, Dec. 2015. [Online]. Available: http://doi.acm.org/10.1145/2843948
3. M. Bredel and E. Jacoby, "Chemogenomics: an emerging strategy for rapid target and drug discovery," *Nature Reviews Genetics*, vol. 5, p. 262, Apr. 2004. [Online]. Available: http://dx.doi.org/10.1038/nrg1317
4. R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo," in *Proceedings of the International Conference on Machine Learning*, vol. 25, 2008.
5. "SMURFF matrix factorization source code," https://github.com/ExaScience/smurff.
6. S. Virtanen, A. Klami, S. A. Khan, and S. Kaski, "Bayesian group factor analysis." in *AISTATS*, 2012, pp. 1269–1277.
7. J. Simm, A. Arany, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, H. Ceulemans, and Y. Moreau, "Macau: Scalable bayesian factorization with high-dimensional side information using mcmc," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2017, pp. 1–6.
8. The ExCAPE Consortium. ExCAPE: Exascale Compound Activity Prediction Engine. http://excape-h2020.eu/. Retrieved: June 2017.