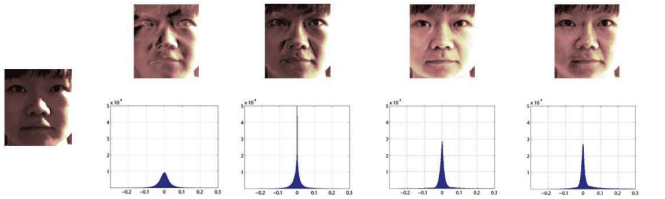
Bayesian Adaptive Matrix Factorization with Automatic Model Selection

Peixian Chen, Naiyan Wang, Nevin L. Zhang, Dit-Yan Yeung
Department of Computer Science and Engineering, The Hong Kong University of Science and Technology

Low-rank matrix factorization has long been recognized as a fundamental problem in many computer vision applications. It approximates a given data matrix by the product of a basis matrix and a coefficient matrix under some criteria. If the underlying rank R of the two factor matrices is lower than that of the original data matrix, matrix factorization is an effective way to reveal the low-dimensional structure of the data.

Nevertheless, the reliability of existing matrix factorization methods is often hard to guarantee due to challenges brought by two model selection issues. The first one is selection of the noise model which affects how well each entry of the matrix can be represented by the model, and the second one is the selection of the capacity of the two factor matrices related to the expressive power of the model and it is generally difficult to estimate the underlying rank accurately. In Fig. 1, we empirically show the effect of these two problems in a face shadow removal application. As we can see, the noise model indeed affects the performance greatly while the rank controls the degree of abstraction.



(a) Original (b) Rank 8, 1 Gaussian (c) Rank 8, 1 Laplace (d) Rank 4, 32 Gaussian(Ours) (e) Rank 8, 32 Gaussian(Ours)

Figure 1: A face shadow removal example demonstrating the effect of some model selection issues. The original image is shown in (a) and the recovery results of different models are shown in (b)-(e). In the results, the first row shows the recovered images and the second row shows the corresponding noise distributions of the matrices.

For the first problem, we propose to use a *Dirichlet process Gaussian mixture model* (DP-GMM) [5] as the noise model. On one hand, we can take advantage of the fact that GMM is a universal approximator for any continuous distribution [3] and thus able to fit various types of noise. On the other hand, we can infer the number of Gaussian components needed from data, instead of doing heuristic pruning or trying ungrounded guesses. For the second problem, we choose an *automatic relevance determination* (ARD) [2, 4] prior for the factor matrices. ARD has long been recognized as an effective technique for detecting the relevant components of the input, so that we can automatically infer the optimal rank by pruning other irrelevant ones. In this paper, we propose our novel non-parametric full Bayesian model for *adaptive matrix factorization* (AMF). AMF for the first time makes full use of the flexibility and adaptiveness of DP-GMM as noise model, and is completed by ARD for automatic rank selection. It is also designed to be capable of handling input with missing data. For model inference, we devise an efficient variational method based on the stick-breaking representation of DP.

We now combine the desirable features of ARD and DP-GMM to define the AMF model. The graphical model of AMF is depicted in Fig. 2 and the generative process is given as follows:

1. Draw component mixing proportions $\theta \sim \mathbf{GEM}(\alpha)$.
2. For each cluster k of noise:
   - Draw variance $\sigma_k \sim \mathbf{IG}(a_0, b_0)$.
3. For each dimension r of $\mathbf{U}$ and $\mathbf{V}$ (i.e., each column of $\mathbf{U}$ and $\mathbf{V}$):
   - Draw variance $\lambda_r \sim \mathbf{IG}(a_1, b_1)$.
4. For each element in $\mathbf{U}$ and $\mathbf{V}$:
   - Draw $u_{mr}, v_{nr} \sim \mathcal{N}(0, \lambda_r)$.
5. For each data element y_{mn}:
   - Draw noise cluster label $z_{mn} \sim \mathbf{Mult}(\theta)$;
   - Draw observation $y_{mn} \sim \mathcal{N}(\mathbf{u}_{m\cdot}\mathbf{v}_{n\cdot}^T, \sigma_{z_{mn}})$.

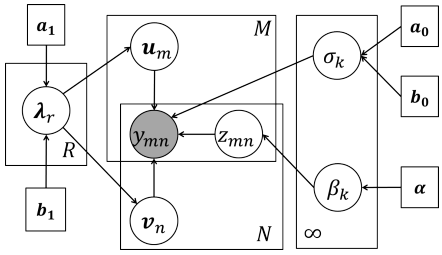


Figure 2: Graphical model of AMF

Here $\theta_k \triangleq \beta_k \prod_{l=1}^{k-1}(1-\beta_l)$ and β_k is drawn independently from stick-breaking construction of DP. Based on the generative process, the joint distribution can be expressed as:

$$\begin{aligned}&\mathbf{p}(\mathbf{U},\mathbf{V},\mathbf{Y},\mathbf{z},\sigma,\lambda,\beta \mid a_0,b_0,a_1,b_1,\alpha)\\ =&\mathbf{p}(\mathbf{Y} \mid \mathbf{U},\mathbf{V},\mathbf{z},\sigma)\,\mathbf{p}(\mathbf{U} \mid \lambda)\,\mathbf{p}(\mathbf{V} \mid \lambda)\,\mathbf{p}(\lambda \mid a_1,b_1)\\ &\mathbf{p}(\sigma \mid a_0,b_0)\,\mathbf{p}(\mathbf{z} \mid \beta)\,\mathbf{p}(\beta \mid \alpha).\end{aligned}$$

We devise an efficient inference algorithm based on mean-field variational method for AMF, details can be found in the full paper. For experimental validation, AMF is tested on both synthetic and real-world data. We use text removal and face shadow removal tasks to demonstrate the effectiveness of the automatic model selection capabilities of AMF. The results show that AMF consistently achieves better or comparable performance among state-of-the art algorithms included. It has high potential to handle a wide range of applications with automatic model selection.

[1] F. De La Torre and J.M. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1-3):117–142, 2003.

[2] David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3): 415–447, 1992.

[3] V. Maz'ya and G. Schmidt. On approximate approximations using Gaussian kernels. *IMA Journal of Numerical Analysis*, 16(1):13–29, 1996.

[4] Radford M Neal. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995.

[5] C.E. Rasmussen. The infinite Gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.

[6] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[7] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma. Face recognition with contiguous occlusion using Markov random fields. In *ICCV*, pages 1050–1057, 2009.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.