Closed-form Marginal Likelihood in Gamma-Poisson Matrix Factorization

Louis Filstroff, Alberto Lumbreras, Cédric Févotte

IRIT, Univ. Toulouse, CNRS

July 12, 2018

Non-negative Matrix Factorization (NMF)

- Find the best approximation of a non-negative matrix $\mathbf{V}$ as the product of two non-negative matrices :

$$\underset{F \times N}{\mathbf{V}} \simeq \underset{\substack{F \times K \\ \text{Dictionary}}}{\mathbf{W}} \times \underset{\substack{K \times N \\ \text{Activations}}}{\mathbf{H}}$$

Non-negative Matrix Factorization (NMF)

- Find the best approximation of a non-negative matrix $\mathbf{V}$ as the product of two non-negative matrices :

$$\underset{F \times N}{\mathbf{V}} \simeq \underset{\substack{F \times K \\ \text{Dictionary}}}{\mathbf{W}} \times \underset{\substack{K \times N \\ \text{Activations}}}{\mathbf{H}}$$

- Problem traditionally solved by minimizing a cost function :

$$\min_{\mathbf{W},\mathbf{H} \geq 0} D(\mathbf{V}|\mathbf{WH}) \tag{1}$$

Non-negative Matrix Factorization (NMF)

- Find the best approximation of a non-negative matrix $\mathbf{V}$ as the product of two non-negative matrices :

$$\underset{F \times N}{\mathbf{V}} \simeq \underset{\substack{F \times K \\ \text{Dictionary}}}{\mathbf{W}} \times \underset{\substack{K \times N \\ \text{Activations}}}{\mathbf{H}}$$

- Problem traditionally solved by minimizing a cost function :

$$\min_{\mathbf{W},\mathbf{H} \geq 0} D(\mathbf{V}|\mathbf{WH}) \quad (1)$$

- A popular cost function is the generalized Kullback-Leibler divergence (KL-NMF)
[Lee and Seung, 2000, Févotte and Idier, 2011] :

$$D_{\mathrm{KL}}(\mathbf{V}|\mathbf{WH}) = \sum_{f,n} \left(v_{fn} \log \frac{v_{fn}}{[\mathbf{WH}]_{fn}} - v_{fn} + [\mathbf{WH}]_{fn} \right) \quad (2)$$

KL-NMF and probabilistic equivalence

- Minimizing the KL divergence w.r.t. $\mathbf{W}$ and $\mathbf{H}$ is equivalent to the joint maximum likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ in the following observation model :

$$v_{fn} \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{3}$$

KL-NMF and probabilistic equivalence

- Minimizing the KL divergence w.r.t. $\mathbf{W}$ and $\mathbf{H}$ is equivalent to the joint maximum likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ in the following observation model :

$$v_{fn} \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{3}$$

- In other words :

$$\min_{\mathbf{W},\mathbf{H}\geq 0} D_{KL}(\mathbf{V}|\mathbf{WH}) \Leftrightarrow \max_{\mathbf{W},\mathbf{H}} p(\mathbf{V}|\mathbf{W},\mathbf{H}) \tag{4}$$

KL-NMF and probabilistic equivalence

- Minimizing the KL divergence w.r.t. **W** and **H** is equivalent to the joint maximum likelihood estimation of **W** and **H** in the following observation model :

$$v_{fn} \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{3}$$

- In other words :

$$\min_{\mathbf{W},\mathbf{H} \geq 0} D_{KL}(\mathbf{V}|\mathbf{WH}) \Leftrightarrow \max_{\mathbf{W},\mathbf{H}} p(\mathbf{V}|\mathbf{W},\mathbf{H}) \tag{4}$$

- Can be questioned from a statistical point of view : the number of estimated parameters, $FK + KN$, grows with the number of samples, N

Maximum Marginal Likelihood Estimation (MMLE)

- To overcome this problem, it was proposed to treat $\mathbf{H}$ as latent variables with a prior distribution, $p(\mathbf{H})$

Maximum Marginal Likelihood Estimation (MMLE)

- To overcome this problem, it was proposed to treat $\mathbf{H}$ as latent variables with a prior distribution, $p(\mathbf{H})$
- $\mathbf{W}$ is treated as a deterministic parameter

Maximum Marginal Likelihood Estimation (MMLE)

- To overcome this problem, it was proposed to treat $\mathbf{H}$ as latent variables with a prior distribution, $p(\mathbf{H})$
- $\mathbf{W}$ is treated as a deterministic parameter
- We wish instead to maximize the marginal likelihood :

$$\max_{\mathbf{W}} p(\mathbf{V}|\mathbf{W}) = \int_{\mathbf{H}} p(\mathbf{V}|\mathbf{W},\mathbf{H})p(\mathbf{H})\mathrm{d}\mathbf{H} \tag{5}$$

Maximum Marginal Likelihood Estimation (MMLE)

- To overcome this problem, it was proposed to treat $\mathbf{H}$ as latent variables with a prior distribution, $p(\mathbf{H})$
- $\mathbf{W}$ is treated as a deterministic parameter
- We wish instead to maximize the marginal likelihood :

$$\max_{\mathbf{W}} p(\mathbf{V}|\mathbf{W}) = \int_{\mathbf{H}} p(\mathbf{V}|\mathbf{W}, \mathbf{H}) p(\mathbf{H}) \mathrm{d}\mathbf{H} \tag{5}$$

- Previous work in this model [Dikmen and Févotte, 2012] showed an empirical "self-regularization" phenomenon on the columns of $\mathbf{W}$, which was left unexplained

The Gamma-Poisson (GaP) model

- Poisson observation model on $\mathbf{V}$ + Gamma prior on $\mathbf{H}$ = "Gamma-Poisson" (GaP) model [Canny, 2004, Buntine and Jakulin, 2006]

The Gamma-Poisson (GaP) model

- Poisson observation model on $\mathbf{V}$ + Gamma prior on $\mathbf{H}$ = "Gamma-Poisson" (GaP) model [Canny, 2004, Buntine and Jakulin, 2006]
- Generative model as follows :

$$h_{kn} \sim \text{Gamma}(\alpha_k, \beta_k) \tag{6}$$

$$v_{fn}|\mathbf{h}_n \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{7}$$

The Gamma-Poisson (GaP) model

- Poisson observation model on $\mathbf{V}$ + Gamma prior on $\mathbf{H}$ = "Gamma-Poisson" (GaP) model [Canny, 2004, Buntine and Jakulin, 2006]
- Generative model as follows :

$$h_{kn} \sim \text{Gamma}(\alpha_k, \beta_k) \tag{6}$$

$$v_{fn}|\mathbf{h}_n \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{7}$$

- In our work α_k and β_k are fixed hyperparameters

The Gamma-Poisson (GaP) model

- Poisson observation model on $\mathbf{V}$ + Gamma prior on $\mathbf{H}$ = "Gamma-Poisson" (GaP) model [Canny, 2004, Buntine and Jakulin, 2006]
- Generative model as follows :

$$h_{kn} \sim \text{Gamma}(\alpha_k, \beta_k) \tag{6}$$

$$v_{fn}|\mathbf{h}_n \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{7}$$

- In our work α_k and β_k are fixed hyperparameters
- Can be rewritten with auxiliary variables $\mathbf{C}$:

$$h_{kn} \sim \text{Gamma}(\alpha_k, \beta_k) \tag{8}$$

$$c_{fkn}|h_{kn} \sim \text{Poisson}(w_{fk} h_{kn}) \tag{9}$$

$$v_{fn} = \sum_k c_{fkn} \tag{10}$$

Contributions (1/2)

- **H** can be integrated out in GaP, i.e. we are able to rewrite the generative model free of **H**

$$\mathbf{c}_{kn} \sim \mathrm{NM}\left(\alpha_k, \left[\frac{w_{1k}}{\sum_f w_{fk} + \beta_k}, \ldots, \frac{w_{Fk}}{\sum_f w_{fk} + \beta_k}\right]\right)$$

$$\mathbf{v}_n = \sum_k \mathbf{c}_{kn}$$

where NM denotes the Negative Multinomial distribution

Contributions (1/2)

- $\mathbf{H}$ can be integrated out in GaP, i.e. we are able to rewrite the generative model free of $\mathbf{H}$

$$\mathbf{c}_{kn} \sim \mathrm{NM}\left(\alpha_k, \left[\frac{w_{1k}}{\sum_f w_{fk} + \beta_k}, \ldots, \frac{w_{Fk}}{\sum_f w_{fk} + \beta_k}\right]\right)$$
$$\mathbf{v}_n = \sum_k \mathbf{c}_{kn}$$

where NM denotes the Negative Multinomial distribution

- This new model leads to a closed form for $p(\mathbf{V}|\mathbf{W})$:

$$p(\mathbf{V}|\mathbf{W}) = \sum_{\mathbf{C}\in\mathcal{C}_\mathbf{V}} p(\mathbf{C}|\mathbf{W}) = \sum_{\mathbf{C}\in\mathcal{C}_\mathbf{V}} \prod_k \prod_n \underbrace{p(\mathbf{c}_{kn}|\mathbf{W})}_{\mathrm{NM}} \tag{11}$$

where $\mathcal{C}_\mathbf{V} = \{\mathbf{C} \in \mathbb{N}^{F\times K\times N} \,|\, \forall(f,n), \sum_k c_{fkn} = v_{fn}\}$.

Contributions (2/2)

- After computation :

$$-\frac{1}{N}\log p(\mathbf{V}|\mathbf{W}) = \overbrace{-\frac{1}{N}\log\left(\sum_{\mathbf{C}\in\mathcal{C}_{\mathbf{V}}} f_{\alpha}(\mathbf{W};\mathbf{C})\right)}^{\text{“data-fitting term”}} \tag{12}$$

$$+ \underbrace{\sum_k \alpha_k \log(||\mathbf{w}_k||_1 + \beta_k)}_{\text{“regularization term”}} + \text{cst.} \tag{13}$$

Contributions (2/2)

- After computation :

$$-\frac{1}{N}\log p(\mathbf{V}|\mathbf{W}) = \overbrace{-\frac{1}{N}\log\left(\sum_{\mathbf{C}\in\mathcal{C}_{\mathbf{V}}} f_{\alpha}(\mathbf{W};\mathbf{C})\right)}^{\text{“data-fitting term”}} \tag{12}$$

$$+\underbrace{\sum_k \alpha_k \log(||\mathbf{w}_k||_1 + \beta_k)}_{\text{“regularization term”}} + \text{cst.} \tag{13}$$

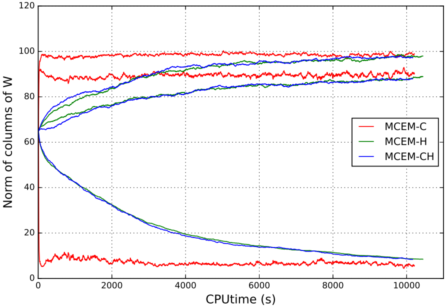
- Regularization terms of the form $\sum_i \log(x_i + \epsilon)$ are known to be sparsity-inducing [Candès et al., 2008]

Optimization of the likelihood

- Problem with observed and latent variables : (Monte Carlo) EM algorithm [Wei and Tanner, 1990]. Comparison of three algorithms based on the three possible choices for the set of latent variables : $\{\mathbf{C},\mathbf{H}\}$, $\{\mathbf{H}\}$, $\{\mathbf{C}\}$

Optimization of the likelihood

- Problem with observed and latent variables : (Monte Carlo) EM algorithm [Wei and Tanner, 1990]. Comparison of three algorithms based on the three possible choices for the set of latent variables : $\{\mathbf{C},\mathbf{H}\}$, $\{\mathbf{H}\}$, $\{\mathbf{C}\}$



- The algorithm based on $\{\mathbf{C}\}$ has a tendency to converge faster

Take-home messages

- Closed-form expression of the marginal likelihood in the Gamma-Poisson model, a probabilistic matrix factorization model for count data

Take-home messages

- Closed-form expression of the marginal likelihood in the Gamma-Poisson model, a probabilistic matrix factorization model for count data
- Reveals a regularization term on the columns of $\mathbf{W}$, explaining the ability of MMLE to automatically prune columns of $\mathbf{W}$

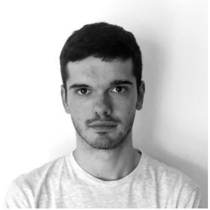
Take-home messages

- Closed-form expression of the marginal likelihood in the Gamma-Poisson model, a probabilistic matrix factorization model for count data
- Reveals a regularization term on the columns of $\mathbf{W}$, explaining the ability of MMLE to automatically prune columns of $\mathbf{W}$
- The marginalization of $\mathbf{H}$ leads to an EM algorithm with favorable properties (as observed in experiments)

Thank you for your attention

Questions ?

Come see us at poster #55 tonight !



L. Filstroff

A. Lumbreras

C. Févotte

Buntine, W. and Jakulin, A. (2006).
Discrete component analysis.
In *Subspace, Latent Structure and Feature Selection*, pages 1–33. Springer.

Candès, E. J., Wakin, M. B., and Boyd, S. P. (2008).
Enhancing sparsity by reweighted l1 minimization.
Journal of Fourier analysis and applications, 14(5):877–905.

Canny, J. (2004).
GaP: a factor model for discrete data.
In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129.

Dikmen, O. and Févotte, C. (2012).
Maximum marginal likelihood estimation for nonnegative dictionary learning in the gamma-poisson model.
IEEE Transactions on Signal Processing, 60(10):5163–5175.

Févotte, C. and Idier, J. (2011).
Algorithms for nonnegative matrix factorization with the β-divergence.
Neural computation, 23(9):2421–2456.

Lee, D. D. and Seung, H. S. (2000).
Algorithms for non-negative matrix factorization.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 556–562.

Wei, G. C. and Tanner, M. A. (1990).
A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms.
Journal of the American statistical Association, 85(411):699–704.