Efficient Matrix Completion with Gaussian Models

Flavien Léger
CMLA, ENS Cachan, France

Guoshen Yu and Guillermo Sapiro
ECE, University of Minnesota

Abstract

A general framework based on Gaussian models and a MAP-EM algorithm is introduced in this work for solving matrix/table completion problems. The numerical experiments with the standard and challenging movie ratings data show that the proposed approach, based on probably one of the simplest probabilistic models, leads to the results in the same ballpark as the state-of-the-art, at a lower computational cost.

Matrix Completion

Example: movie ranking

	Star Wars	Titanic	Spider-Man	Harry Potter	...	Transformers	Legion
User 1	5	?	?	?	?	?	?
User 2	?	?	?	?	?	3	?
User 3	?	?	4	?	2	?	?
⋮	?	?	?	1	?	?	?
User N	?	?	?	?	?	?	4

4% rankings available, 96% to estimate.

Formulation

$$\mathbf{Y} = \mathbf{H} \bullet \mathbf{F} + \mathbf{W}$$

$\mathbf{F}$: true signal. $\mathbf{H}$: binary mask. $\mathbf{W}$: noise. $\mathbf{Y}$: observed signal. Objective: estimate $\mathbf{F}$ from $\mathbf{Y}$.

Gaussian Models

- Rewrite the matrix row by row (or column by column).

$$\mathbf{y}_i = \mathbf{U}_i \mathbf{f}_i + \mathbf{w}_i$$

where $\mathbf{U}_i$ is a masking operator on the i-th row.

- Assume each row follows a Gaussian distribution.

$$\mathbf{f}_i \sim \mathcal{N}(\mu, \Sigma)$$

- Assume Gaussian noise $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{w}})$

MAP (Maximum a Posteriori) Estimate

- The optimal estimate that minimizes the mean squared error in the Gaussian setting.

$$\begin{aligned} \tilde{\mathbf{f}}_i &= \arg\max_{\mathbf{f}_i} \log p(\mathbf{f}_i | \mathbf{y}_i, \mu, \Sigma) \\ &= \mu + \Sigma \mathbf{U}_i^T (\mathbf{U}_i \Sigma \mathbf{U}_i^T + \Sigma_{\mathbf{w}})^{-1} (\mathbf{y}_i - \mathbf{U}_i \mu) \end{aligned}$$

Matlab code: *http://www.cmap.polytechnique.fr/∼yu/research/MC/demo.html*

EM-MAP Algorithm

- The Gaussian parameters (μ, Σ) are unknown. Estimate through an iterative EM-MAP algorithm.
- E-step: assume $(\tilde{\mu}, \tilde{\Sigma})$ known, estimate $\tilde{\mathbf{f}}_i$ with the MAP estimate.

$$\tilde{\mathbf{f}}_i = \tilde{\mu} + \tilde{\Sigma} \mathbf{U}_i^T (\mathbf{U}_i \tilde{\Sigma} \mathbf{U}_i^T + \tilde{\Sigma_{\mathbf{w}}})^{-1} (\mathbf{y}_i - \mathbf{U}_i \tilde{\mu})$$

- M-step: $\tilde{\mathbf{f}}_i$ assume known, estimate $(\tilde{\mu}, \tilde{\Sigma})$ with the ML (maximum likelihood) estimate.

$$\tilde{\mu} = \frac{1}{M} \sum_{i=1}^{M} \tilde{\mathbf{f}}_i \qquad \text{and} \qquad \tilde{\Sigma} = \frac{1}{M} \sum_{i=1}^{M} (\tilde{\mathbf{f}}_i - \tilde{\mu})(\tilde{\mathbf{f}}_i - \tilde{\mu})^T$$

Numerical Experiments

Benchmark Datasets

- EachMovie: 1,648 movies, 74,424 users, 2.8 million ratings (4.3% available).
- 1M MovieLens: 3,900 movies, 6,040 users, 1 million ratings (4.7% available).

Standard Protocols

- Weak generalization: measure the ability of a method to generalize to other items rated by the same users used for training the method.
- Strong generalization: measure the ability of the method to generalize to some items rated by novel users that have not been used for training.

Results

- Error measure: normalized mean absolute error (NMAE) – random guessing produces a score of 1.
- The proposed approach, based on the simplest Gaussian models, produces results in the same ballpark as the state-of-the-art, at a lower computational cost.

EachMovie	Weak NMAE	Strong NMAE
URP [Marlin, 04]	0.4422	0.4557
Attitude [Marlin, 04]	0.4520	0.4550
MMMF [Rennie et al., 05]	0.4397	0.4341
IPCF [Park et al., 05]	0.4382	0.4365
E-MMMF [Decoste et al., 06]	0.4287	0.4301
GPLVM [Lawrence et al., 09]	0.4179	0.4134
M^3F [Mackey et al., 10]	0.4293	n/a
NBMC [Zhou et al., 10]	**0.4109**	**0.4091**
GM (proposed)	0.4164	0.4163

1M MovieLens	Weak NMAE	Strong NMAE
URP [Marlin, 04]	0.4341	0.4444
Attitude [Marlin, 04]	0.4320	0.4375
MMMF [Rennie et al., 05]	0.4156	0.4203
IPCF [Park et al., 05]	0.4096	0.4113
E-MMMF [Decoste et al., 06]	0.4029	0.4071
GPLVM [Lawrence et al., 09]	0.4026	0.3994
NBMC [Zhou et al., 10]	**0.3916**	0.3992
GM (proposed)	0.3959	**0.3928**

References on Gaussian models

Matrix completion: *F. Léger, G. Yu and G. Sapiro. Efficient Matrix Completion with Gaussian Models, Proc. ICASSP, 2011, Prague.*

Imaging inverse problems: *G.Yu, G. Sapiro, and S. Mallat, Solving Inverse Problems with Piecewise Linear Estimators: From Gaussian Mixture Models to Structured Sparsity, submitted, arxiv.org/abs/1006.3056, 2010.*

Compressed sensing: *G.Yu and G. Sapiro, Statistical Compressive Sensing of Gaussian Mixture Models, submitted, arxiv.org/abs/1101.5785, 2011. (Lecture: Thursday, May 26, 13:45 - 14:05)*