ICML | Atlanta

International Conference on Machine Learning

Fast Max-Margin Matrix Factorization with Data Augmentation

Minjie Xu, Jun Zhu & Bo Zhang

Tsinghua University

Matrix Factorization and M³F (I)

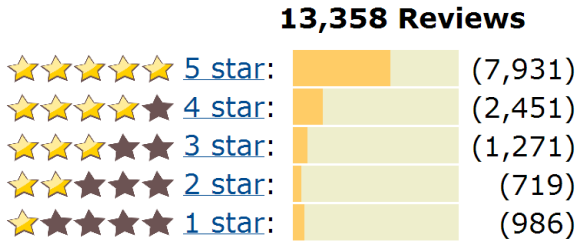
- Setting: fit a partially observed matrix $Y \in \mathbb{R}^{N \times M}$ with X subject to certain constraints
- Examples
  - Singular Value Decomposition (SVD):
    when Y is fully observed, approximate it with the K leading components (hence $\mathrm{rank}(X) = K$ and X minimizes ℓ_2-loss)
  - Probabilistic Matrix Factorization (PMF):
    assume $X = UV^\top$ with Gaussian prior and likelihood
    (equivalent to ℓ_2-loss minimization with F-norm regularizer)
  - Max-Margin Matrix Factorization (M³F):
    hinge loss minimization with nuclear norm regularizer on X
    (or equivalently, F-norm regularizer on U and V)

(I) $\min_X \|X\|_* + C \sum_{ij \in \mathcal{I}} h\left(Y_{ij} X_{ij}\right)$ **(II)** $\min_{U,V} \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) + C \sum_{ij \in \mathcal{I}} h\left(Y_{ij} U_i V_j^\top\right)$

observed entries

Matrix Factorization and M³F (II)

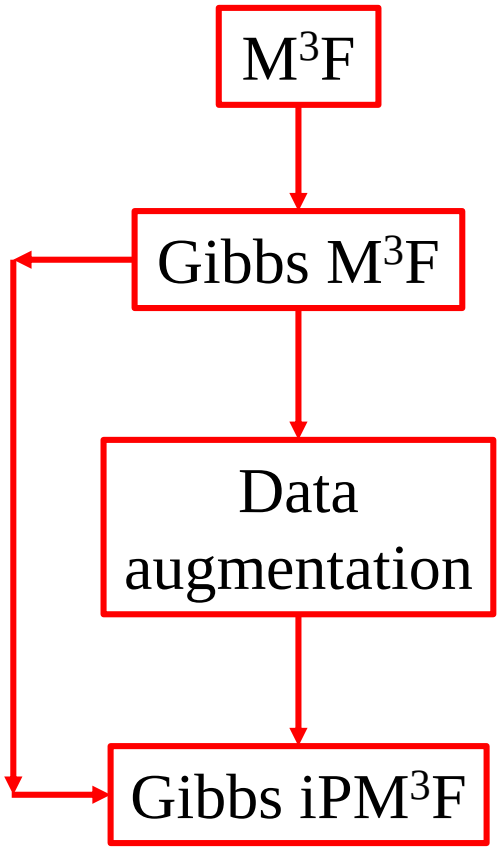
- Benefits of M^{3}F
  - max-margin approach, more applicable to binary, ordinal or categorical data (e.g. *ratings*)
  - the nuclear norm regularizer (I) allows *flexible* latent dimensionality



- Limitations
  - **scalability vs. flexibility:** SDP solvers for (I) scale poorly; while the more scalable (II) requires a pre-specified fixed finite K
  - **efficiency vs. approximation:** gradient descent solvers for (II) require a smooth hinge; while bilinear SVM solvers can be time-consuming
- Motivations: to build a M^{3}F model that is both scalable, flexible and admits highly efficient solvers.

Roadmap

$$Y \simeq UV^\top$$



$$\min_{U,V} \Omega(U,V) + C\mathcal{R}(U,V;Y) \qquad U \in \mathbb{R}^{N\times K}, V \in \mathbb{R}^{M\times K}$$

$$q(U,V) \propto \exp\{-\Omega(U,V) - C\mathcal{R}(U,V;Y)\} \quad U \in \mathbb{R}^{N\times K}, V \in \mathbb{R}^{M\times K}$$

$$\max_{U,V} q(U,V)$$

$$q(U,V) = \int q(U,V,\boldsymbol{\lambda}) d\boldsymbol{\lambda}$$

$$q(Z,V) \propto q_0(Z,V)\exp\{-C\mathcal{R}(Z,V;Y)\} \quad Z \in \{0,1\}^{N\times\infty}, V \in \mathbb{R}^{M\times\infty}$$

RRM as MAP, A New Look (I)

- Setting: fit training data $\mathcal{X} = \{\mathcal{X}_n\}_{n=1}^N$ with model $\mathcal{M}$
- Regularized Risk Minimization (RRM):

$$\mathcal{X}_n = (\mathbf{x}_n, y_n)$$

feature **label**

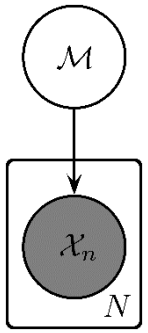
$$\min_{\mathcal{M}} \Omega(\mathcal{M}) + C \sum_{n=1}^{N} \mathcal{R}(\mathcal{M}; \mathcal{X}_n)$$

regularizer **empirical risk**

- Maximum *a Posteriori* (MAP):

$$\max_{\mathcal{M}} p(\mathcal{M}|\mathcal{X}) \propto p_0(\mathcal{M}) \prod_{n=1}^{N} \mathcal{L}(\mathcal{M}|\mathcal{X}_n)$$

posterior **prior** **likelihood**



- For discriminative models

$$\mathcal{R}(\mathcal{M}; \mathcal{X}_n) = L(y_n, f(\mathcal{M}; \mathbf{x}_n)) \qquad \mathcal{L}(\mathcal{M}|\mathcal{X}_n) \triangleq p(y_n|\mathcal{M}, \mathbf{x}_n)$$

loss function **discriminant function**

RRM as MAP, A New Look (II)

- Bridge RRM and MAP via *delegate* prior (likelihood)
  - jointly intact: $(p_0, \mathcal{L})$ and $(\dot{p}_0, \dot{\mathcal{L}})$ induce exactly the same joint distribution (and thus the same posterior)

$$p_0(\mathcal{M}) \prod_{n=1}^{N} \mathcal{L}(\mathcal{M}|\mathcal{X}_n) \propto \dot{p}_0(\mathcal{M}) \prod_{n=1}^{N} \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_n)$$

  - singly relaxed: free from the normalization constraints (and thus no longer probability densities)

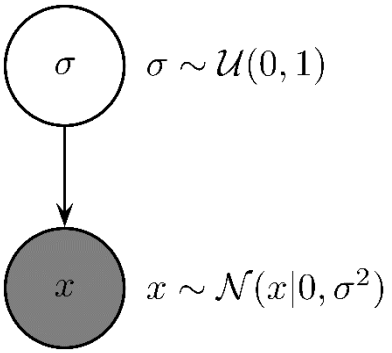
$$\dot{p}_0(\mathcal{M}) \propto p_0(\mathcal{M}) / \prod_{n=1}^{N} \zeta_n(\mathcal{M}), \ \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_n) \propto \zeta_n(\mathcal{M}) \mathcal{L}(\mathcal{M}|\mathcal{X}_n)$$

- The transition:

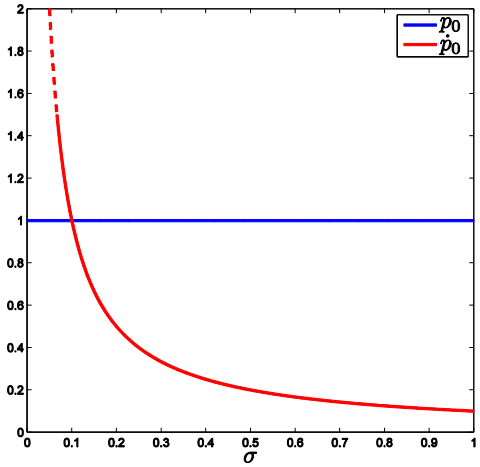
$$\dot{p}_0(\mathcal{M}) = e^{-\Omega(\mathcal{M})}, \ \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_n) = e^{-C\mathcal{R}(\mathcal{M};\mathcal{X}_n)}$$

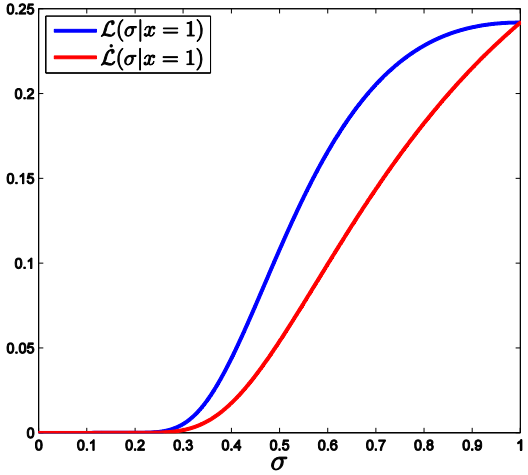
Delegate prior & likelihood

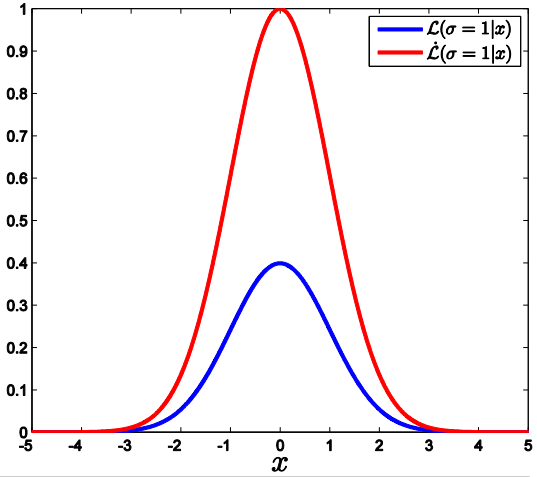
- Consider a simplest case: $\mathcal{M} = \sigma,\ \mathcal{X} = \{x\}$
  - genuine pair: $(p_0, \mathcal{L}) = \Big(\mathcal{U}(0,1), \mathcal{N}(x|0,\sigma^2)\Big)$
  - delegate pair: $(\dot{p}_0, \dot{\mathcal{L}}) = \left(\frac{\mathbb{I}_{\sigma\in(0,1)}}{\sigma}, e^{-\frac{x^2}{2\sigma^2}}\right)$



  - $(\dot{p}_0, \dot{\mathcal{L}})$ can be completely different from $(p_0, \mathcal{L})$ when viewed as functions of the model $\mathcal{M}$







Both $\dot{p}_0$ and $\dot{\mathcal{L}}$ are scaled (up to a constant) for better visualization.

M³F as MAP: the full model

- We consider M³F for ordinal ratings $Y_{ij} \in \{1, 2, \ldots, L\}$
- Risk: introduce thresholds $\theta_{i1} \leq \cdots \leq \theta_{i(L-1)}$ and sum over the $L-1$ binary M³F losses for each θ_{ir}

$$\mathbf{s} = f(U, V, \boldsymbol{\theta}; (i,j)) = \boldsymbol{\theta}_i - (U_i V_j^\top)\mathbf{1}_{L-1},\ L(Y_{ij}, \mathbf{s}) = \sum_{r=1}^{L-1} h_\ell(T_{ij}^r s_r)$$

$$\text{where } T_{ij}^r \triangleq \begin{cases} +1 & \text{for } r \geq Y_{ij} \\ -1 & \text{for } r < Y_{ij} \end{cases},\ h_\ell(x) \triangleq \max(0, \ell - x)$$

- Regularizer: $\Omega(U, V) + \Omega(\boldsymbol{\theta})$, where

$$\Omega(\boldsymbol{\theta}) = \frac{1}{2\varsigma^2} \sum_{i=1}^{N} \|\boldsymbol{\theta}_i - \boldsymbol{\rho}\|_2^2 \qquad (\rho_1 < \cdots < \rho_{L-1})$$

- MAP: $\mathcal{M} = (U, V, \boldsymbol{\theta})$ with hyper-parameters $\{\sigma, \boldsymbol{\rho}, \varsigma, C, \ell, \ ?\}$

$$\dot{p}_0(U, V, \boldsymbol{\theta}) = \prod_{i=1}^{N} \mathcal{N}(U_i | \mathbf{0}, \sigma^2 I) \mathcal{N}(\boldsymbol{\theta}_i | \boldsymbol{\rho}, \varsigma^2 I) \cdot \prod_{j=1}^{M} \mathcal{N}(V_j | \mathbf{0}, \sigma^2 I)$$

$$\dot{\mathcal{L}}(U, V, \boldsymbol{\theta} | (i,j), Y_{ij}) = \prod_{r=1}^{L} e^{-2\max(\Delta_{ij}^r, 0)}, \text{ where } \Delta_{ij}^r \triangleq \frac{C}{2}(\ell - T_{ij}^r(\theta_{ir} - U_i V_j^\top))$$

Data Augmentation for M³F (I)

- Data augmentation in general
  - introduce auxiliary variables to facilitate Bayesian inference on the original variables of interest
  - inject *independence*:
    e.g. EM algorithm (joint);
    stick-breaking construction (conditional)
  - exchange for a much *simpler* conditional representation:
    e.g. slice-sampling; data augmentation strategy for logistic models and that for SVMs
- Lemma (location-scale mixture of Gaussians):

$$e^{-2\max(u,0)} = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{(u+\lambda)^2}{2\lambda}} d\lambda = \int_0^\infty \underline{\phi(u|-\lambda,\lambda)d\lambda}$$

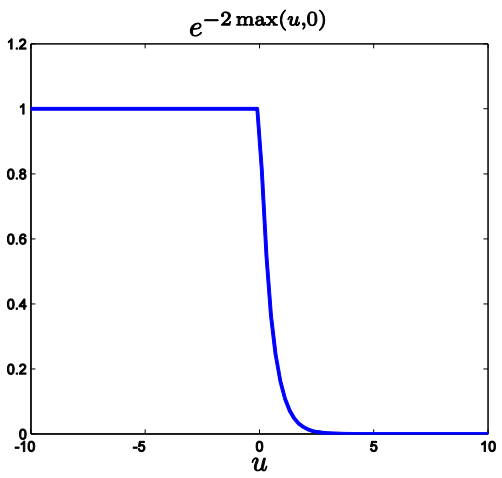
Gaussian density function

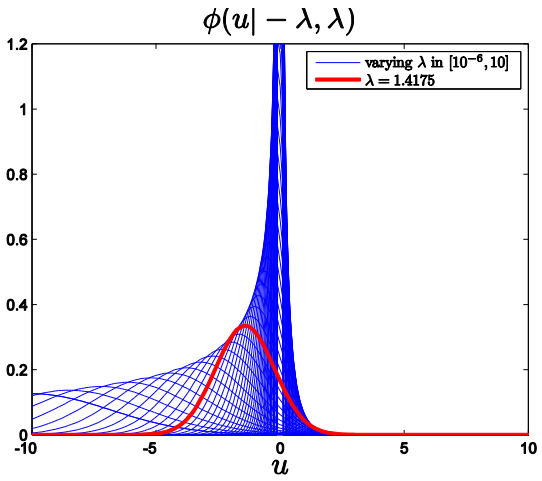
Data Augmentation for M³F (II)

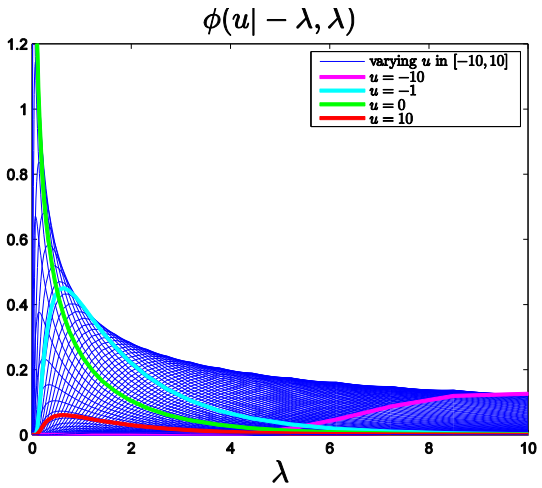
- Benefit of the augmented representation $\phi(u|-\lambda,\lambda)$
  - $u|\lambda$: Gaussian $\mathcal{N}(u|-\lambda,\lambda)$, "conjugate" to Gaussian "prior"
  - $\lambda|u$: Generalized inverse Gaussian $\mathcal{GIG}(\lambda|1/2,1,u^2)$

$$\mathcal{GIG}(\lambda|p,a,b) \propto \lambda^{p-1}e^{-\frac{1}{2}\left(a\lambda+\frac{b}{\lambda}\right)}$$

  - $\lambda^{-1}|u$: inverse Gaussian $\mathcal{IG}(\lambda^{-1}||u|^{-1},1)$







Data Augmentation for M³F (III)

- M³F before augmentation:

$$p(\mathcal{M}|\mathcal{X}) \propto \dot{p}_0(\mathcal{M}) \prod_{ij\in\mathcal{I}} \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij})$$

where $\dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}) = \prod_{r=1}^{L} e^{-2\max(\Delta_{ij}^r, 0)} = \int_{\mathbb{R}_+^{L-1}} \phi(\boldsymbol{\Delta}_{ij}| - \boldsymbol{\lambda}_{ij}, \mathrm{diag}(\boldsymbol{\lambda}_{ij}))d\boldsymbol{\lambda}_{ij}$

and $\boldsymbol{\Delta}_{ij} = (\Delta_{ij}^1, \ldots, \Delta_{ij}^{L-1})^\top$, $\boldsymbol{\lambda}_{ij} = (\lambda_{ij1}, \ldots, \lambda_{ij(L-1)})^\top$

- M³F after augmentation (auxiliary variables $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{ij\in\mathcal{I}}$):

$$p(\mathcal{M}, \boldsymbol{\lambda}|\mathcal{X}) \propto \dot{p}_0(\mathcal{M}) \prod_{ij\in\mathcal{I}} \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij})$$

where $\dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij}) \triangleq \phi(\boldsymbol{\Delta}_{ij}| - \boldsymbol{\lambda}_{ij}, \mathrm{diag}(\boldsymbol{\lambda}_{ij}))$

Data Augmentation for M³F (IV)

- Posterior inference via Gibbs sampling
  - Draw λ_{ijr}^{-1} from $\mathcal{IG}(|\Delta_{ij}^r|^{-1}, 1)$ for $ij \in \mathcal{I}, r = 1, \ldots, L$
  - Draw V_j from $\mathcal{N}(b_j, B_j)$ for $j = 1, \ldots, M$
  - Draw U_i likewise for $i = 1, \ldots, N$
  - Draw θ_{ir} from $\mathcal{N}(a_{ir}, A_{ir})$ for $i = 1, \ldots, N, r = 1, \ldots, L$
  - For details, please refer to our paper

Step	Asymptotic complexity
Sample $\boldsymbol{\lambda}$	$O(\lvert\mathcal{I}\rvert LK)$
Sample V (and U likewise)	$O(\lvert\mathcal{I}\rvert(L+K^2)) + O(MK^3)$
Calculate $\{B_j^{-1}\}_{j=1}^M$	$O(\lvert\mathcal{I}\rvert(L+K^2))$
Cholesky Decomposition	$O(MK^3)$
Calculate $\{b_j\}_{j=1}^M$	$O(\lvert\mathcal{I}\rvert(L+K)) + O(MK^2)$
Draw $\{V_j\}_{j=1}^M$ from $\mathcal{N}$	$O(MK^2)$
Sample $\boldsymbol{\theta}$	$O(\lvert\mathcal{I}\rvert(L+K)) + O(NL)$

Nonparametric M³F (I)

cross validation

- We want to automatically infer from data the latent dimensionality K in an elegant way
- The Indian buffet process
  - induces a distribution on binary matrices with an unbounded number of columns
  - follows a culinary metaphor
  - e.g. $\alpha = 3$

behavioral pattern of the *i*th customer:

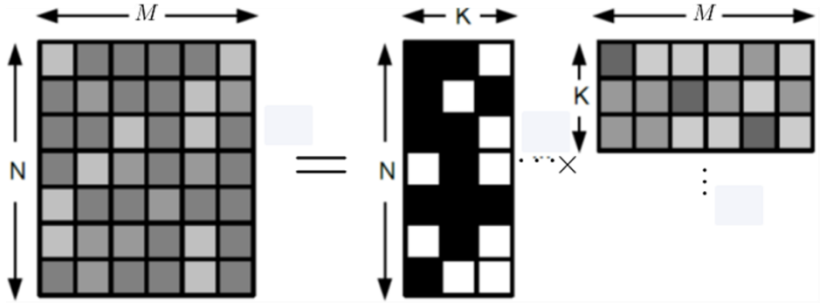
- for *k*th ***sampled*** dish: sample according to popularity m_k/i
- then sample a $K_1^{(i)} = \text{Poisson}(\alpha/i)$ number of ***new*** dishes

Nonparametric M³F (II)

- IBP enjoys several nice properties
  - favors sparse matrices
  - finite columns for finite customers (with probability one)
  - exchangeability → Gibbs sampling would be easy
- We replace U with Z and change the delegate prior

$$\dot{p}_0(Z, V, \boldsymbol{\theta}) = \mathrm{IBP}(Z|\alpha) \cdot \prod_{j=1}^{M} \mathcal{N}(V_j|\mathbf{0}, \sigma^2 I) \cdot \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{\theta}_i|\boldsymbol{\rho}, \varsigma^2 I)$$

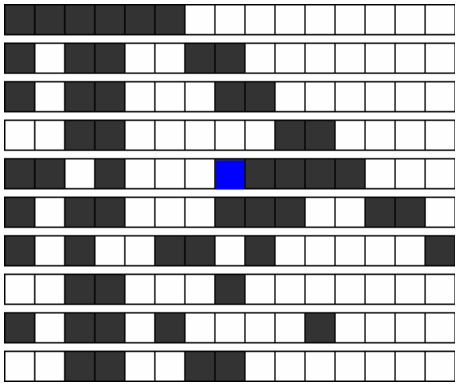
- $\mathcal{M} = (Z, V, \boldsymbol{\theta})$ with hyper-parameters $\{\sigma, \boldsymbol{\rho}, \varsigma, C, \ell, \alpha\}$



Nonparametric M³F (III)

- Inference via Gibbs sampling
  - Draw λ_{ijr}^{-1} from $\mathcal{IG}(|\Delta_{ij}^r|^{-1}, 1)$
  - Draw Z_{ik} from

$$\text{Bernoulli}(Z_{ik}|\sum_{j\neq i} Z_{jk}/N) \prod_{j|ij\in\mathcal{I}} \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij})$$



  - Draw $Z_i^\nu = \mathbf{1}_{k_i}^\top$ from $\text{Poisson}(k_i|\alpha/N) \prod_{j|ij\in\mathcal{I}} \frac{|\Sigma_{ijk_i}|^{1/2}}{\sigma^{k_i}} e^{\frac{1}{2}\boldsymbol{\omega}_{ijk_i}^\top \Sigma_{ijk_i}^{-1} \boldsymbol{\omega}_{ijk_i}}$ where

$$\Sigma_{ijk_i}^{-1} = \frac{1}{\sigma^2} I_{k_i\times k_i} + \sum_{r=1}^{L-1} \frac{C^2}{4\lambda_{ijr}} \cdot \mathbf{1}_{k_i\times k_i}$$

$$\boldsymbol{\omega}_{ijk_i} = -\frac{C}{2} \sum_{r=1}^{L-1} T_{ij}^r \left(1 + \frac{\Delta_{ij}^r}{\lambda_{ijr}}\right) \cdot \Sigma_{ijk_i} \mathbf{1}_{k_i}$$

  - Draw $V^{i\nu}$ from $\mathcal{N}(\boldsymbol{\omega}_{ijk_i}, \Sigma_{ijk_i})$
  - Draw V_j and θ_{ir}

Sampler for	Asymptotic complexity
$\boldsymbol{\lambda}, V, \boldsymbol{\theta}$	same as Gibbs M³F
$\{Z_{ik}\}_{i=1,k=1}^{N,K}$	$O(\lvert\mathcal{I}\rvert(L+K)) + O(NK)$
$\{Z_i^\nu\}_{i=1}^N : \{k_i\}_{i=1}^N$	$O(\lvert\mathcal{I}\rvert\kappa) + O(N\kappa)$
$\{V^{i\nu}\}_{i=1}^N$	tight: $O(\sum_{ij\in\mathcal{I}} k_i^3) + O(M\sum_i k_i)$ loose: $O(\lvert\mathcal{I}\rvert\kappa^3) + O(MN\kappa)$

Experiments and Discussions

- Datasets: MovieLens 1M & EachMovie
- Test error (NMAE):

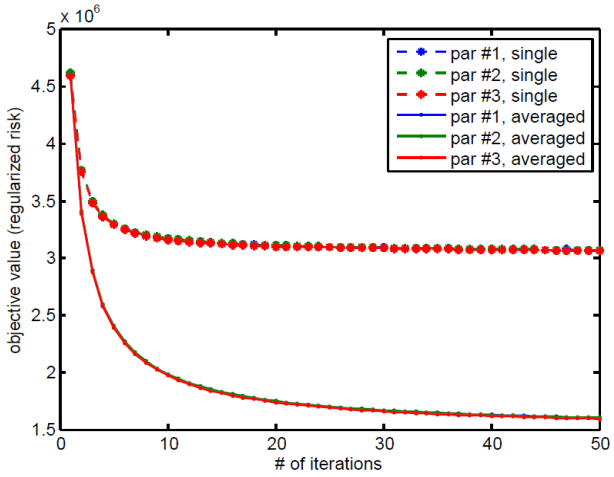
	MovieLens		**EachMovie**	
Algorithm	**weak**	**strong**	**weak**	**strong**
M^3F	$.4156 \pm .0037$	$.4203 \pm .0138$	$.4397 \pm .0006$	$.4341 \pm .0025$
bcd M^3F	$.4176 \pm .0016$	$.4227 \pm .0072$	$.4348 \pm .0023$	$.4301 \pm .0034$
Gibbs M^3F	$.4037 \pm .0005$	$.4040 \pm .0055$	$.4134 \pm .0017$	$.4142 \pm .0059$
iPM^3F	$.4031 \pm .0030$	$.4135 \pm .0109$	$.4211 \pm .0019$	$.4224 \pm .0051$
Gibbs iPM^3F	$.4080 \pm .0013$	$.4201 \pm .0053$	$.4220 \pm .0003$	$.4331 \pm .0057$

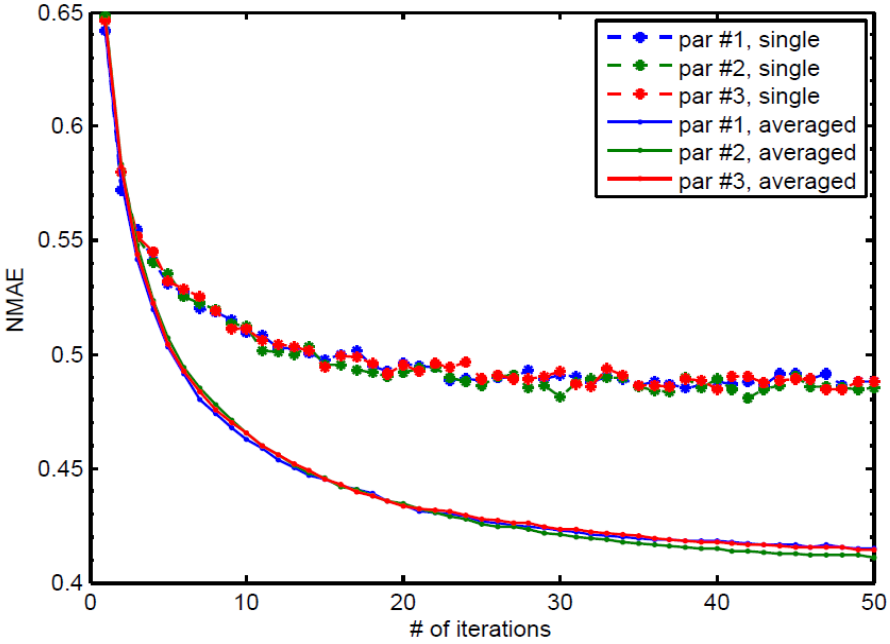
- Training time:

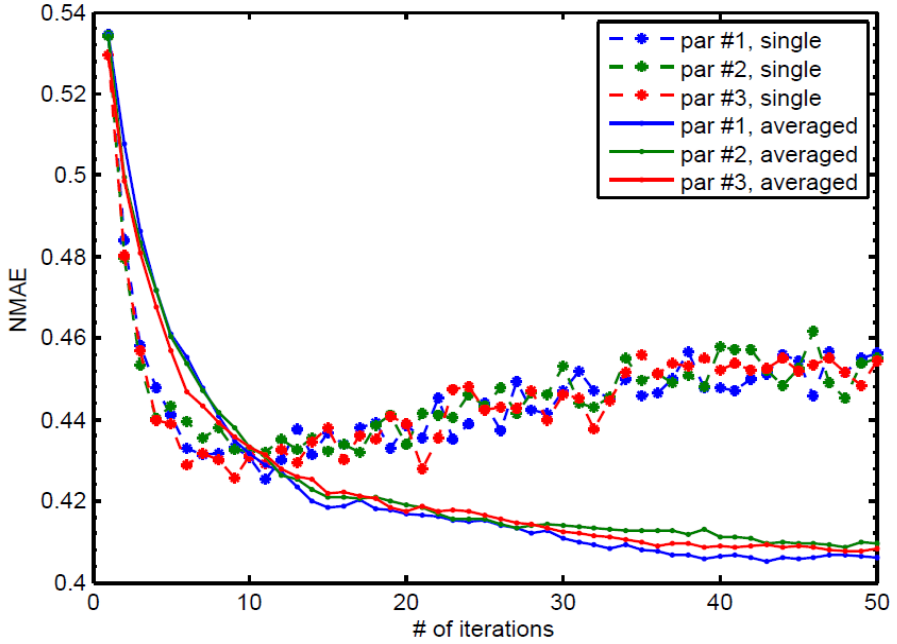
Algorithm	**MovieLens**	**EachMovie**	**Iters**
M^3F	5h	15h	100
bcd M^3F	4h	10h	50
Gibbs M^3F	0.11h	0.35h	50
iPM^3F	4.6h	5.5h	50
Gibbs iPM^3F	0.68h	0.70h	50

Experiments and Discussions

- Convergence:
  - single samples vs. averaged samples
  - RRM objective
  - Validation error (NMAE)

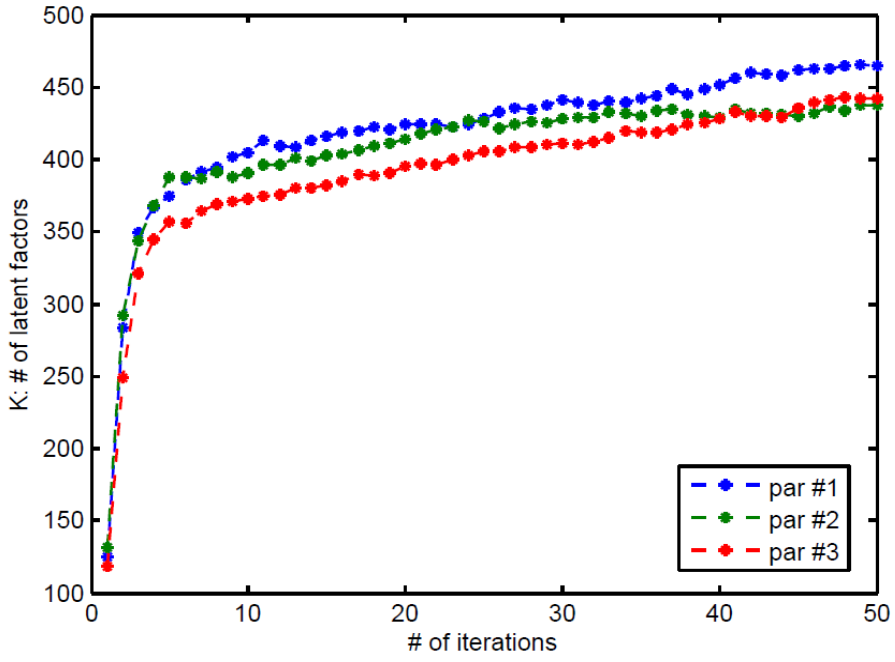


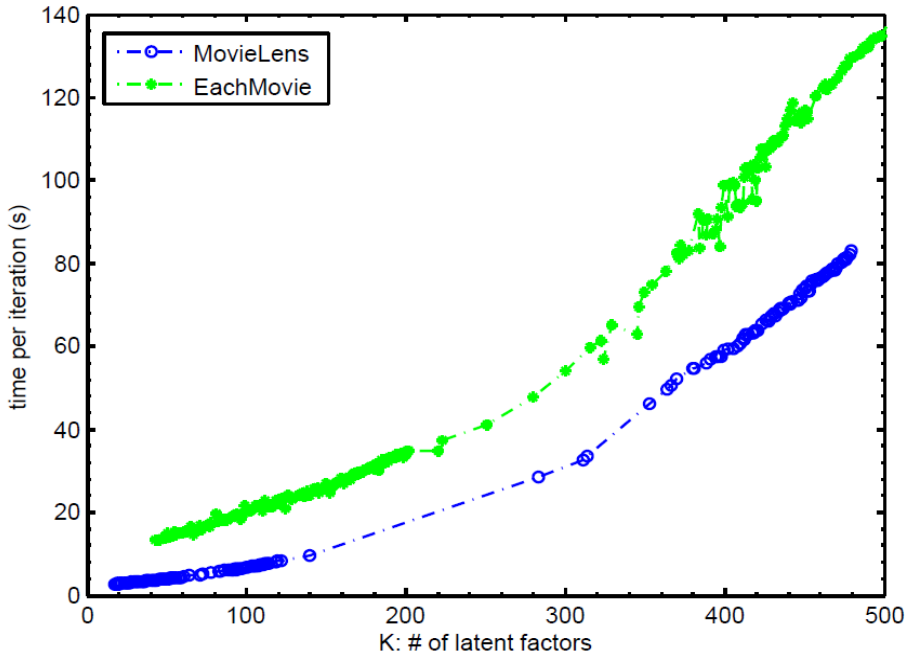




Experiments and Discussions

- Latent dimensionality:





Thanks!