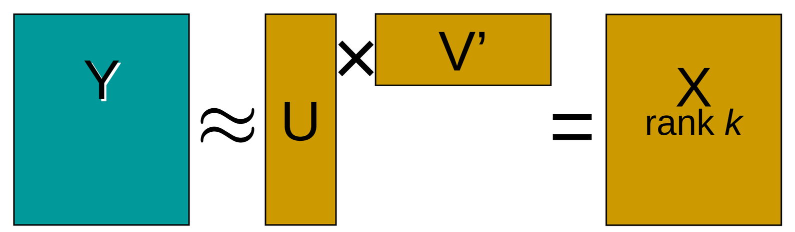
Maximum Margin Matrix Factorization

Nathan Srebro
Department of Computer Science
University of Toronto

Jason Rennie Tommi Jaakkola
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology

Matrix Factorization

Unconstrained: Low Rank Approximation



- Additive Gaussian noise: minimize $|Y-UV'|_{Fro}$ — Unconstrained U,V, fully observed Y → use SVD
- General additive noise
- General conditional models
  - Multiplicative noise, Exponential-PCA [Collins+01], Multinomial (pLSA [Hofman01]), etc
- General loss functions
  - Hinge loss, loss functions appropriate for ratings, etc

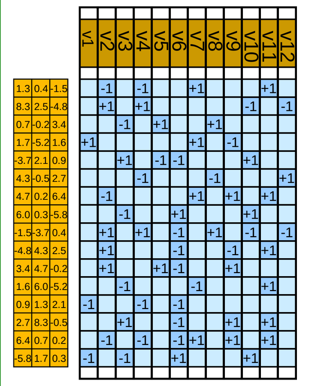
all other situations: non-convex, no explicit solution

Constrained U,V: recover more factors

- Non-Negativity [LeeSeung99]
- Stochasticity (convexity) [LeeSeung97] [Hofman01]
- Sparsity
  - Clustering as an extreme (when rows of U sparse)

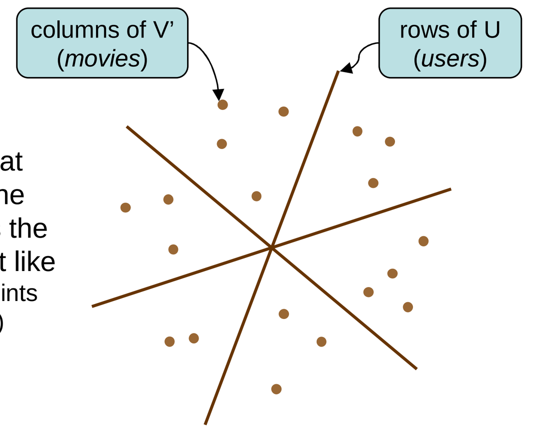
Overall number of factors still constrained
Non-convex optimization problems

Matrix Factorization as Feature Learning



When U is fixed,
each row is a linear classification problem:
- rows of U are feature vectors
- columns of V' are linear classifiers

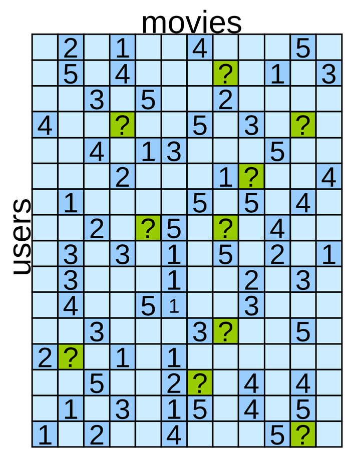
Fitting **U** *and* **V**:
Learning features that work well across all classification problems.



point *and* hyperplane arrangement, such that each user's hyperplane separates the movies the user likes and doesn't like (usually, we are given points and only find hyperplane)

Collaborative Prediction

Based on partially observed matrix
⇒ Predict unobserved entries



"Will user *i* like movie *j*?"

Fit low-rank matrix **X=UV'** to observed entries.

minimize $\sum_{ij\in S} \mathrm{loss}(X_{ij}; Y_{ij})$

observed entries — prediction — observation

Use matrix **X** to predict unobserved entries.

Different matrix factorization methods differ in how they relate real-valued entries in X to the observations (preferences) Y, possibly through a probabilistic model, and in the associated contrast (loss) functions.

Low-rank models of co-occurrence or frequency data

	Multinomial	Independent Binomials	Independent Bernoulli
Mean parameterization $0 \le X_{ij} \le 1$, $E[Y_{ij}\mid X_{ij}]=X_{ij}$	Aspect Model (pLSA) [Hoffman+99] ≡ NMF if $\sum X_{ij}=1$	$Y_{ij}\mid X_{ij} \sim \mathrm{Bin}(N,X_{ij})$ ≈ NMF [Lee+01]	$P(Y_{ij}=1) = X_{ij}$
Natural parameterization unconstrained X_{ij}	SDR [Globerson+02]	$Y_{ij}\mid X_{ij} \sim \mathrm{Bin}(N,g(X_{ij}))$	Logistic Low Rank Approximation [Schein+03]

≈ hinge loss

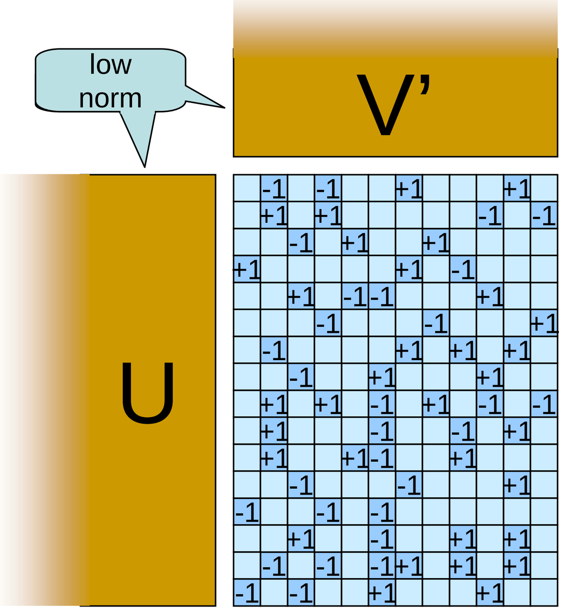
row features most informative about columns

Exponential PCA: [Collins+01] $p(Y_{ij}\mid X_{ij}) \propto \exp(Y_{ij}X_{ij}+F(Y_{ij}))$

$g(x)=1/(1+e^x)$

Max-Margin Matrix Factorization

Instead of bounding dimensionality of **U**,**V**, bound norms of **U**,**V**



bound norms on average:
$(\sum_i |U_i|^2)\,(\sum_j |V_j|^2) \le 1$

bound norms uniformly:
$(\max_i |U_i|^2)\,(\max_j |V_j|^2) \le 1$

For observed $Y_{ij} \in \pm 1$:
$Y_{ij} X_{ij} \ge$ *Margin*
$\langle U_i, V_j\rangle$

U is fixed: each column of **V** is SVM

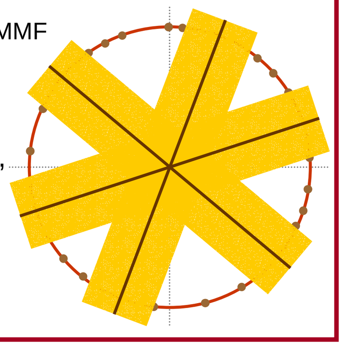
Unlike rank(X) ≤ k, these are convex constraints!

$$\alpha U_1 V_1' + (1-\alpha) U_2 V_2' = \left(\sqrt{\alpha}U_1 \;\; \sqrt{1-\alpha}U_2\right)\begin{pmatrix}\sqrt{\alpha}V_1' \\ \sqrt{1-\alpha}V_2'\end{pmatrix}$$

Geometric Interpretation

for max-norm (uniform) MMMF

point and hyperplane arrangement in infinite dimensional unit sphere, such that each hyperplane separate according to **Y** with large margin.



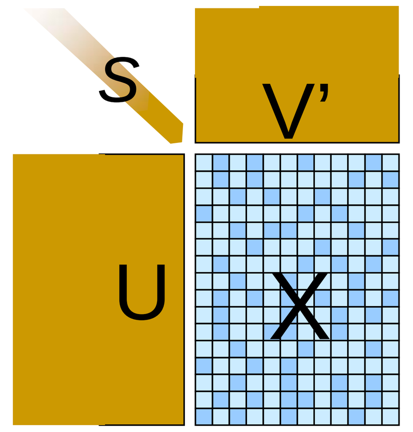
MMMF, Rank and the SVD

rank(**X**) = |singular values|$_0$

trace-norm (average) MMMF minimizes:
$|X|_{tr}$ = |singular values|$_1$

[Fazel Hindi Boyd 2001] suggest $|X|_{tr}$ as convex surrogate to rank(**X**).

Here, we justify it directly through the connection with max-margin classification, and by providing generalization error bounds.



Finding Max-Margin Matrix Factorizations

trace-norm (average) MMMF

maximize M
$Y_{ij} X_{ij} \ge M$
X=UV'
$(\sum_i |U_i|^2)\,(\sum_j |V_j|^2) \le 1$
$|X|_{tr}$

$|X|_{tr} = \sum$ (singular values of X)
$= \min_{X=UV'} \sqrt{(\sum_i |U_i|^2)(\sum_j |V_j|^2)}$
$= \min_{X=UV'} \frac{1}{2}(\sum_i |U_i|^2 + \sum_j |V_j|^2)$
$= \min_{A,B} \frac{1}{2}(\mathrm{tr}(A) + \mathrm{tr}(B))$
s.t. $\begin{pmatrix} A & X \\ X' & B \end{pmatrix}$ p.s.d. [Fazel Hindi Boyd 2001]

primal SDP
minimize $\mathrm{tr}(A)+\mathrm{tr}(B) + c\sum \xi_{ij}$
$Y_{ij} X_{ij} \ge 1 - \xi_{ij}$
$\begin{pmatrix} A & X \\ X' & B \end{pmatrix}$ p.s.d.

dual SDP
maximize $\sum Q_{ij}$
$0 \le Q_{ij} \le c$
$\|Q \otimes Y\|_2 \le 1$
sparse elementwise product (zero for unobserved entries)

max-norm (uniform) MMMF

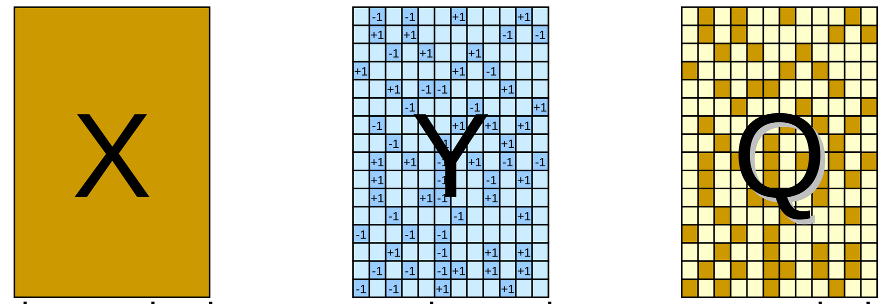
$|X|_{max} = \min_{X=UV'} (\max_i |U_i|)(\max_j |V_j|)$
$= \min_{A,B} \max(\max_i A_{ii}, \max_j B_{jj})$
s.t. $\begin{pmatrix} A & X \\ X' & B \end{pmatrix}$ p.s.d.

maximize M
$Y_{ij} X_{ij} \ge M$
X=UV'
$(\max_i |U_i|)\,(\max_j |V_j|) \le 1$
$|X|_{max}$

dual SDP
maximize $\sum Q_{ij}$
$0 \le Q_{ij} \le c$
$\begin{pmatrix} \Gamma & Q\otimes Y \\ Q'\otimes Y' & \Delta \end{pmatrix}$ p.s.d.
Γ, Δ diagonal; $\mathrm{tr}(\Gamma) + \mathrm{tr}(\Delta) = 1$

primal SDP
minimize $t + c\sum \xi_{ij}$
$Y_{ij} X_{ij} \ge 1 - \xi_{ij}$
$\begin{pmatrix} A & X \\ X' & B \end{pmatrix}$ p.s.d.
$A_{ii} \le t$, $B_{jj} \le t$

Dual variable Q_{ij} for each observed (i,j)



dense primal — sparse observations (constraints) — sparse dual

- Semi-definite program with sparse dual: Limited by number of observations, not size (for both average-norm and max-norm)
- Current implementation: off-the-shelf primal-dual solver, up to 30k observations (e.g. 1000x1000, 3% observed)
- For large-scale problems: updates on dual alone

Reconstructing Primal X* form Dual Q*

X* spanned by Q*⊗Y SVD components of singular value 1

For trace-norm problems without slack, the primal optimal **X*** can be extracted from dual optimal **Q***:

1) Compute the SVD: $Q^*\otimes Y = U\Lambda V'$,
2) Let U*,V* be components of U,V with value 1
3) Primal optimal is of the form $X^* = U^*RR'V'^*$
3) Solve linear equations in RR', with $Q^*_{ij}>0 \Rightarrow X^*_{ij}=Y_{ij}$

Querying Primal X^*_{ij} from Dual Q*

Add constraint $X_{ij}>0$ to primal ⇒ Add variable Q_{ij} to dual
Q* still feasible,
BUT: No optimal solution with $X^*_{ij}>0$ ⇒ Q* not optimal
Q* still optimal ⇒ $X^*_{ij}>0$

To query if sign(X^*_{ij}):
Add $Q^*_{ij}=0$ to Q* with $Y_{ij}=1$ and reoptimize
Add $Q^*_{ij}=0$ to Q* with $Y_{ij}=-1$ and reoptimize

Maximum Margin Matrix Factorization is as an alternative to Low Rank methods:
- Allows an unbounded number of factors
- Convex optimization problem: sparse SDP
- Correspondence with large margin linear classification
- Generalization error bounds
- Applicable in other applications where low-rank approximations are currently used

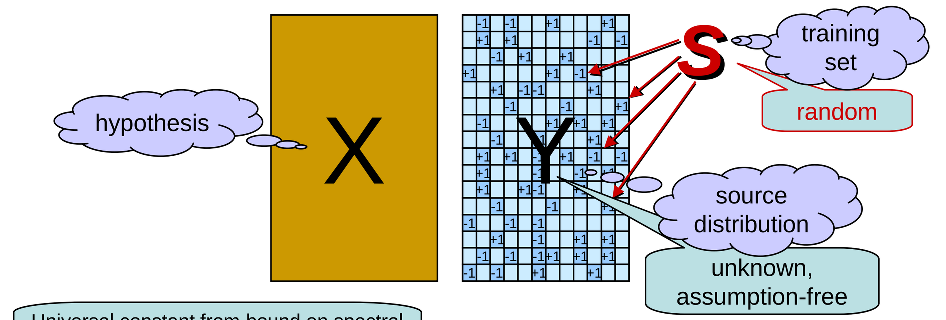
Direct optimization of dual would enable large-scale applications

Generalization Error Bounds

$D(X;Y) = \#_{ij}(X_{ij}\cdot Y_{ij}<0)/nm$ — *generalization error*

$D_S(X;Y) = \#_{ij\in S}(X_{ij}\cdot Y_{ij}<1)/|S|$ — *empirical error*

$$\forall_Y \Pr_S\left(\forall_{\text{rank-}k\ X}\ D(X;Y)<D_S(X;Y)+\varepsilon\right) > 1-\delta$$



Universal constant from bound on spectral norm of random matrix [Seginer00]

$(\sum |U_i|^2/n)(\sum |V_i|^2/m) \le R^2$: $\quad \varepsilon = K\sqrt[4]{\ln m}\sqrt{\frac{R^2(n+m)\log n + \log 1/\delta}{|S|}}$

$(\max |U_i|^2)(\max |V_j|^2) \le R^2$: $\quad \varepsilon = 12\sqrt{\frac{R^2(n+m) + \log 1/\delta}{|S|}}$

Compare with the low-rank bound: [Poster tomorrow!]

rank(X) ≤ k: $\quad \varepsilon = \sqrt{\frac{k(n+m)\log\frac{8em}{k} + \log 1/\delta}{2|S|}}$

Major Assumption: Random Observations

Although we did not make any assumptions about the true preferences **Y**, we made a very strong assumption about the set ***S*** of observed entries: we assumed entries as selected uniformly at random.
For $(\sum |U_i|^2/n)(\sum |V_i|^2/m) \le R^2$, uniformity crucial.
For $(\max |U_i|^2)(\max |V_j|^2) \le R^2$ and rank(**X**)≤k, ***S*** need not be uniform:

$D(X;Y) = E_{ij}[\mathrm{loss}(X_{ij};Y_{ij})]$ $\quad D_S(X;Y) = \sum_{ij\in S} \mathrm{loss}(X_{ij};Y_{ij})/|S|$

same observation distribution

$$\forall_Y \Pr_S\left(\forall_X\ D(X;Y)<D_S(X;Y)+\varepsilon\right) > 1-\delta$$

Not very satisfying: we are guaranteed good generalization only on items the user is likely to observe on its own—not on items we might recommend.

MMMF as a Convex Combination

$\{ X=UV' \mid (\sum |U_i|^2)(\sum |V_i|^2) \le 1 \}$
$= \text{convex-hull}(\{ uv' \mid u \in \mathbb{R}^n, v \in \mathbb{R}^m\ |u|=|v|=1\})$

$\text{conv}(\{ uv' \mid u \in \pm 1^n, v \in \pm 1^m \})$
$\subset \{ X=UV' \mid (\max |U_i|^2)(\max |V_j|^2) \le 1 \}$
$\subset 2\ \text{conv}(\{ uv' \mid u \in \pm 1^n, v \in \pm 1^m \})$

Grothendiek's Inequality

MATLAB code available @ `http://www.cs.toronto.edu/~nati/mmmf`

Experiments

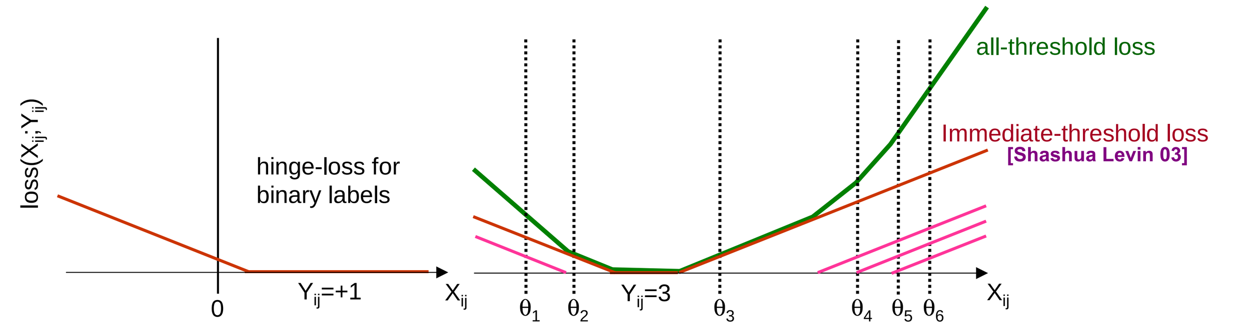
Preliminary experiments on 100 user × 100 movie subset of MovieLens

Two level cross validations:
- Train all variants, with various regularization parameters, on 50% of ratings
- Validate on 25% of data to select best variant and parameters (3-fold CV on 75% of data)
- Evaluate single variant and parameters on held out 25% of data

Compare trace-norm and max-norm MMMF to low-rank approximation minimizing sum-squared error and to K-medians clustering of users.

holdout set	rank agreement error				mean rank difference			
1	rank-2 aprox	0.575	**0.562**	max-norm, c=0.12	2-medians	0.691	**0.677**	max-norm, c=0.12
2	rank-2 aprox	0.562	**0.552**	trace-norm, c=0.24	2-medians	**0.681**	0.683	max-norm, c=0.11
3	rank-1 aprox	0.543	**0.527**	max-norm, c=0.12	2-medians	0.681	**0.646**	max-norm, c=0.12
4	rank-2 aprox	0.558	**0.550**	max-norm, c=0.12	2-medians	0.696	**0.686**	max-norm, c=0.12
average		0.558	**0.548**			0.687	**0.673**	

Generalizations of hinge-loss for ordinal labels



- All-threshold loss is a bound on the absolute rank-difference
- We experimented with both: "all-thresholds" consistently outperformed "immediate-threshold"
- For both loss functions: learn per-user θ's (no extra cost to SDP)