Learning with Matrix Factorizations

Nati Srebro

Department of Electrical Engineering and Computer Science

Massachusetts Institute of Technology

Adviser: Tommi Jaakkola (MIT)

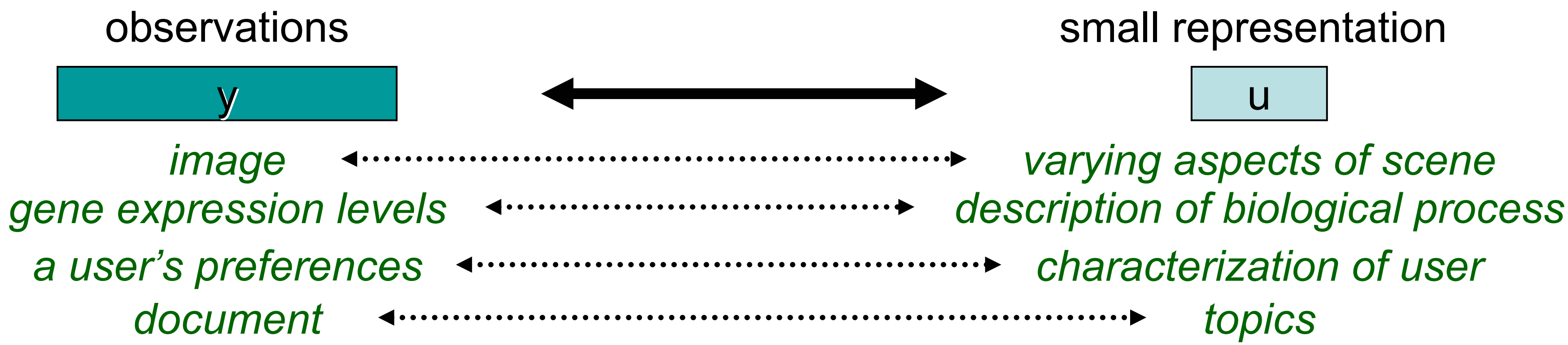
Committee: Alan Willsky (MIT), Tali Tishby (HUJI), Josh Tenenbaum (MIT)

Joint work with Shie Mannor (McGill), Noga Alon (TAU) and Jason Rennie (MIT)

Lots of help from John Barnett, Karen Livescu, Adi Shraibman, Michael Collins, Erik Demaine, Michel Goemans, David Karger and Dmitry Panchenko

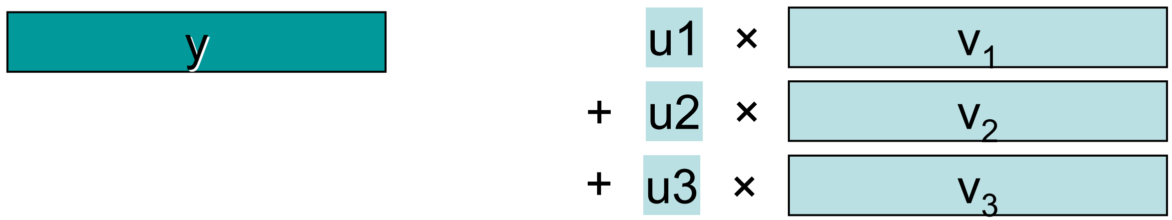
Dimensionality Reduction:

Low dimensional representation capturing important aspects of high dimensional data

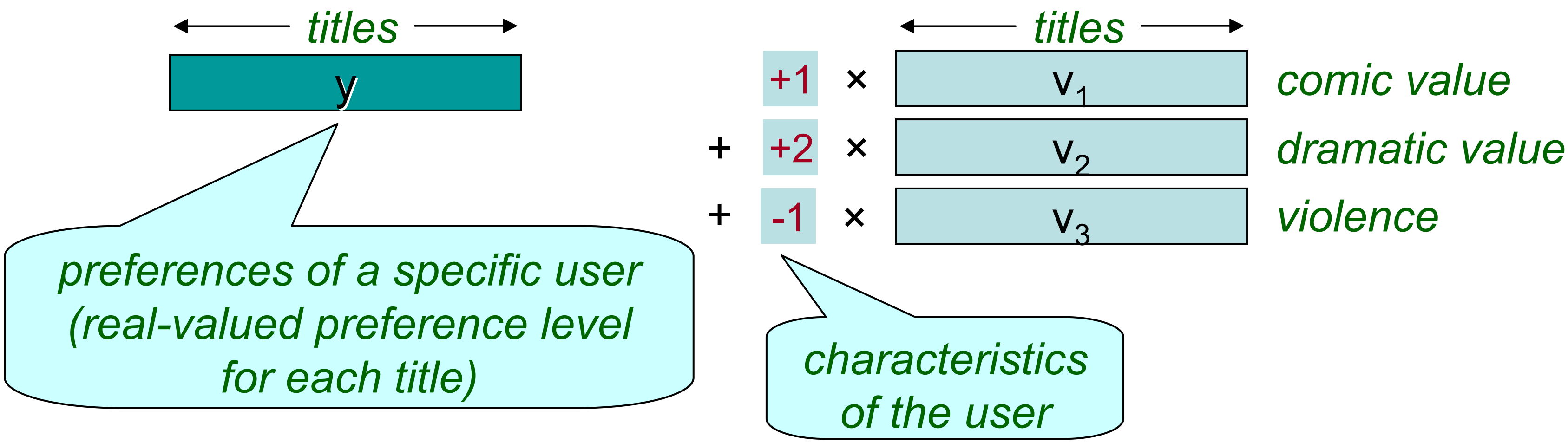


- Compression (mostly to reduce processing time)
- Reconstructing latent signal
  - biological processes through gene expression
- Capturing structure in a corpus
  - documents, images, etc
- Prediction: collaborative filtering

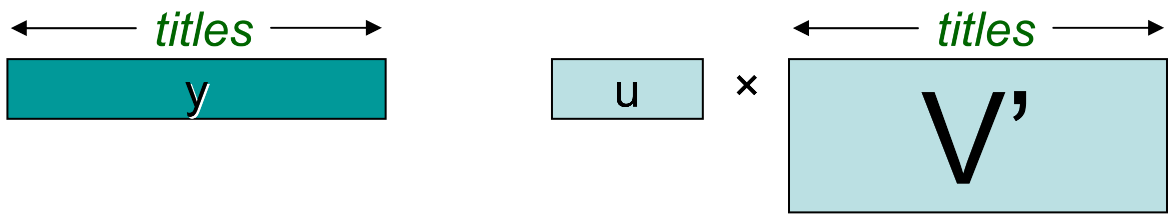
Linear Dimensionality Reduction



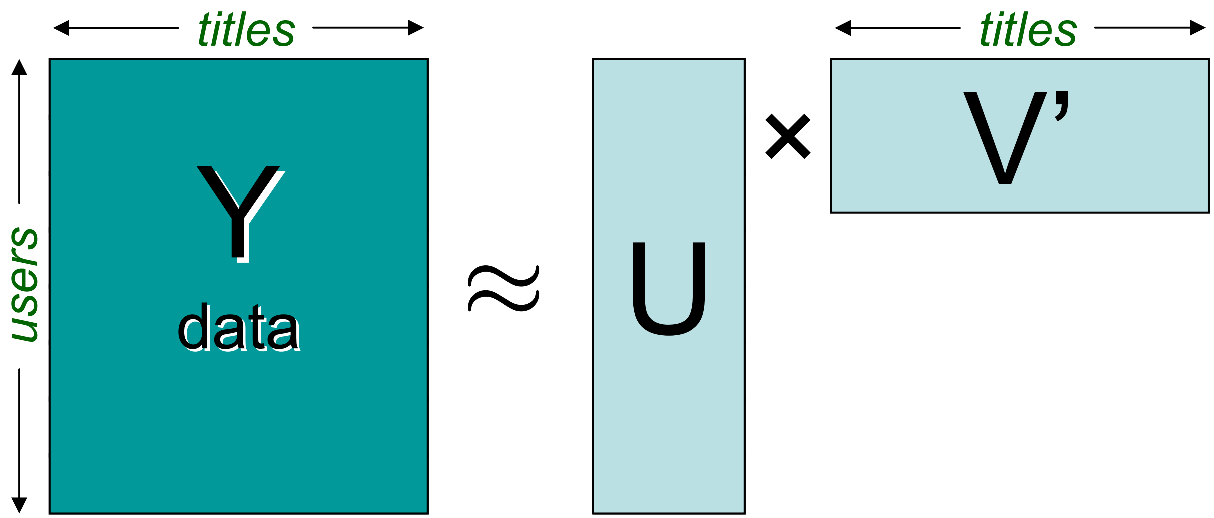
Linear Dimensionality Reduction



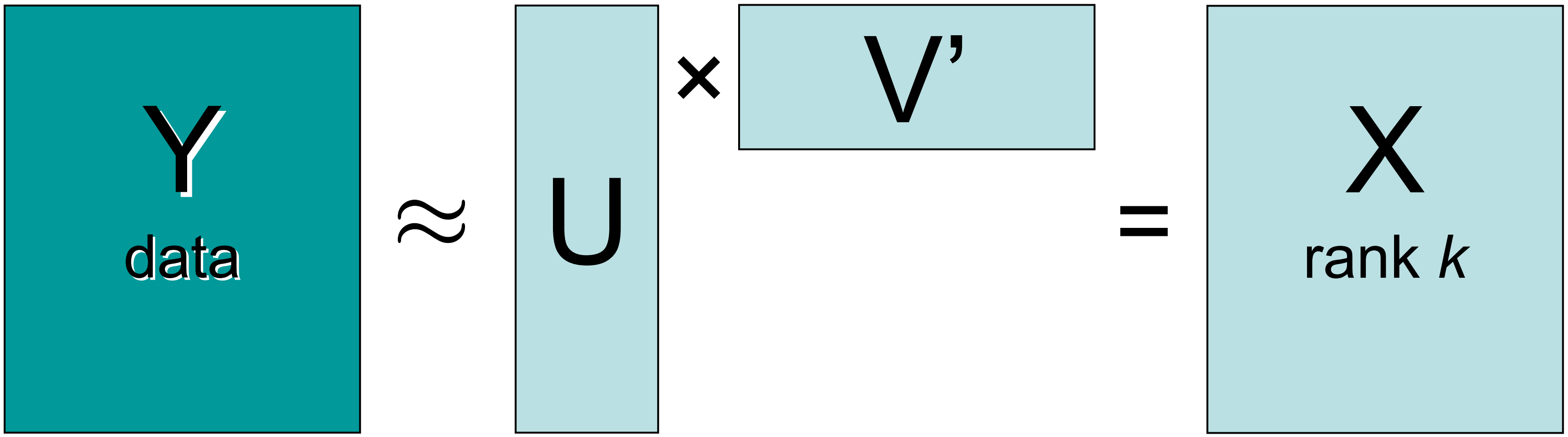
Linear Dimensionality Reduction



Linear Dimensionality Reduction



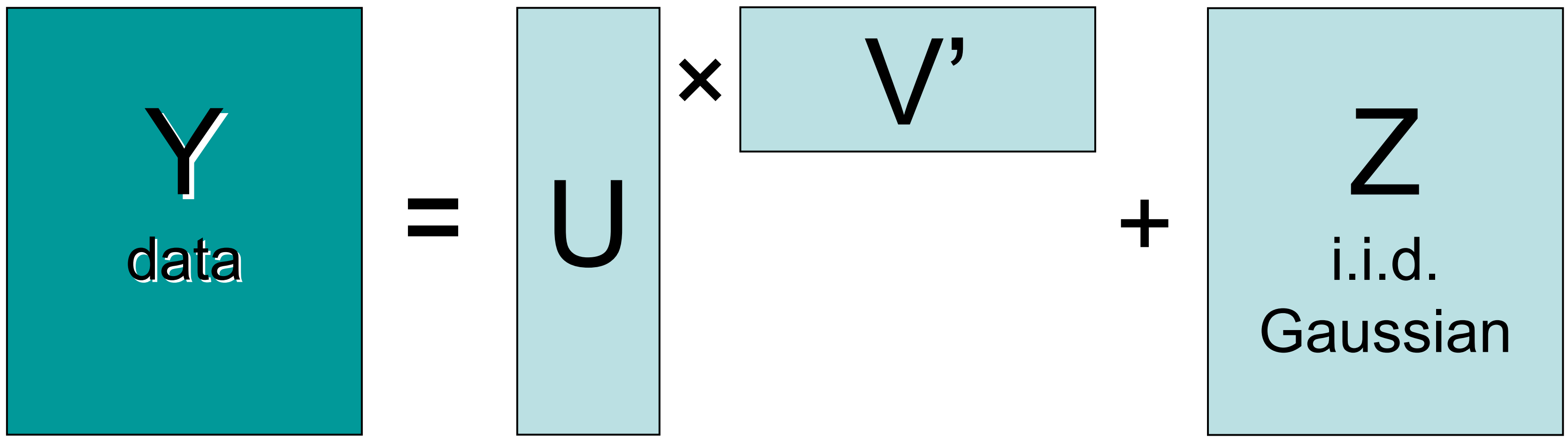
Matrix Factorization



- Non-Negativity [LeeSeung99]
- Stochasticity (convexity) [LeeSeung97][Barnett04]
- Sparsity
  - Clustering as an extreme (when rows of U sparse)

- Unconstrained: Low Rank Approximation

Matrix Factorization

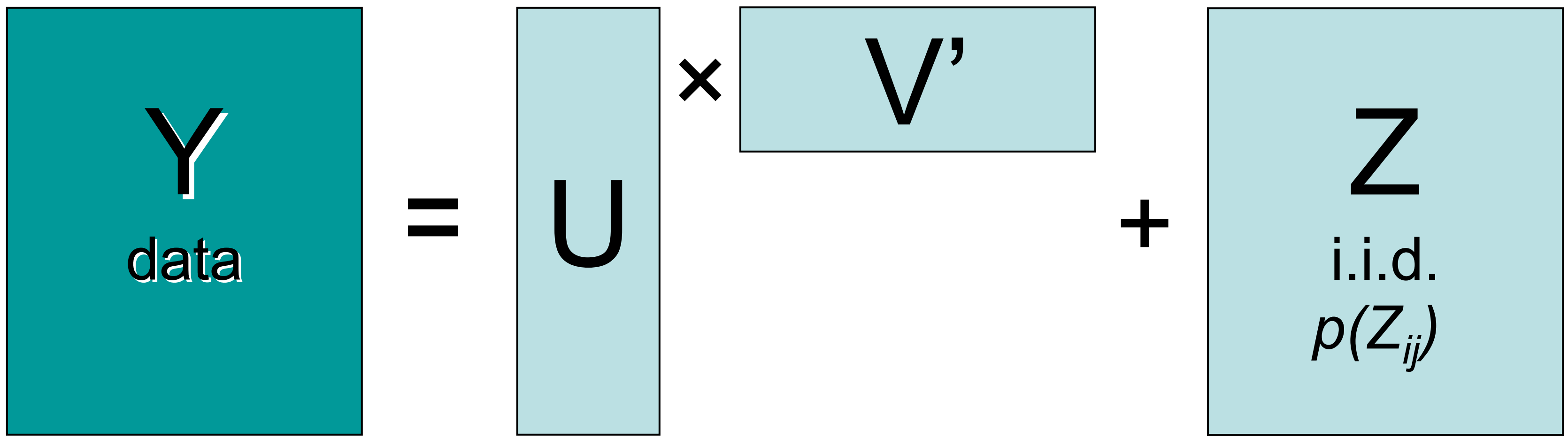


- Additive Gaussian noise

minimize $|Y\text{-}UV'|_{Fro}$

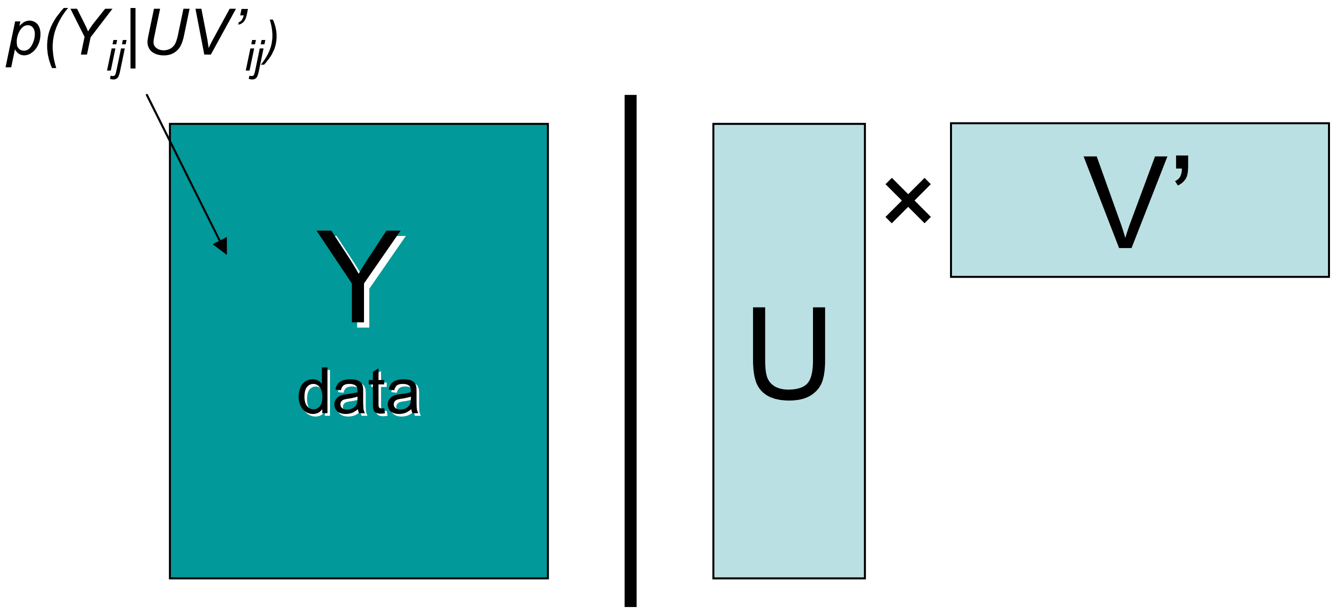
$$\log L\left(UV';Y\right) = \sum_{ij} \log P(Y_{ij} \mid UV'_{ij}) = \tfrac{-1}{2\sigma^2} \sum_{ij} (Y_{ij} - UV'_{ij})^2 + const$$

Matrix Factorization



- Additive Gaussian noise — minimize $|Y-UV'|_{Fro}$
- General additive noise — minimize $\sum -\log p(Y_{ij}-UV'_{ij})$

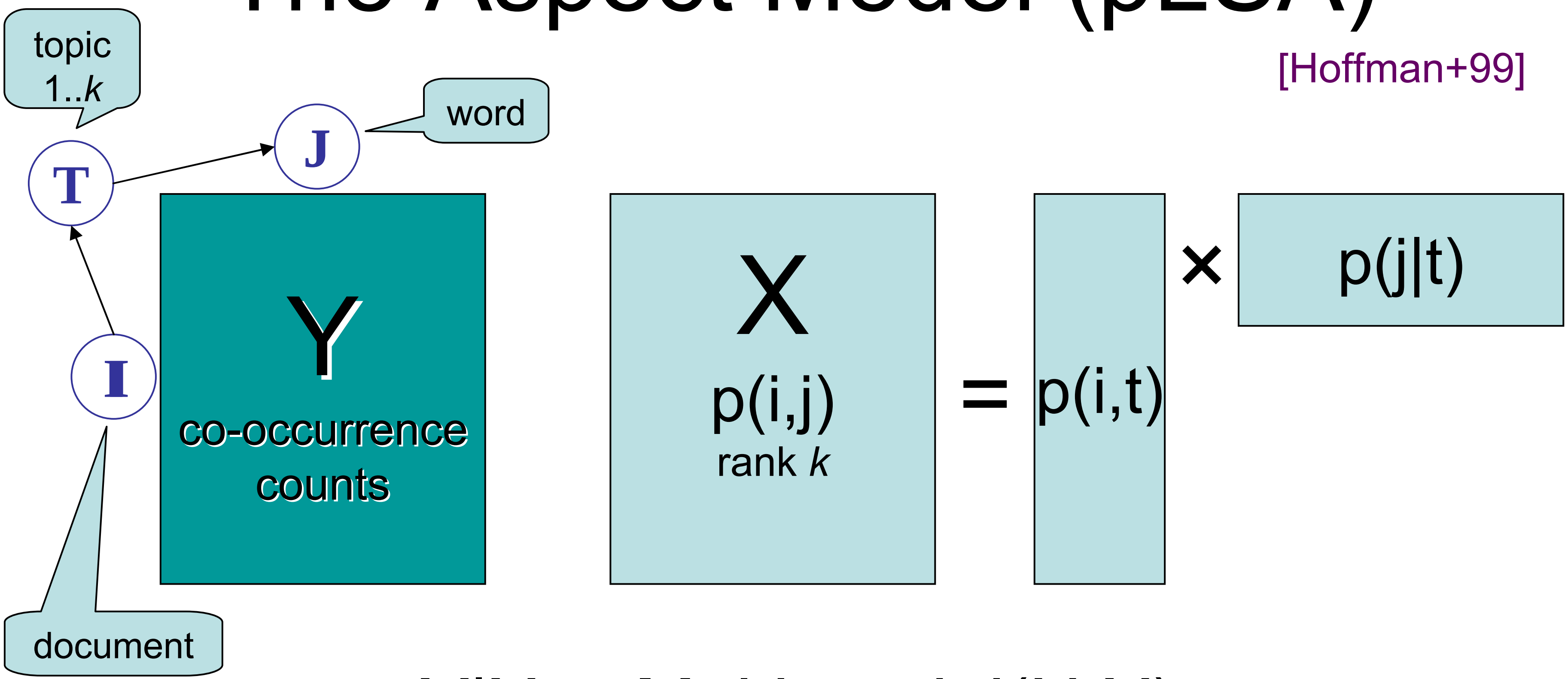
Matrix Factorization



- Additive Gaussian noise — minimize $|Y-UV'|_{Fro}$
- General additive noise — minimize $\sum$ -log $p(Y_{ij}-UV'_{ij})$
- General conditional models — minimize $\sum$ -log $p(Y_{ij}|UV'_{ij})$
  - Multiplicative noise
  - Binary observations: Logistic LRA
  - Exponential PCA [Collins+01]
  - Multinomial (pLSA [Hofman01])
- Other loss functions [Gordon02] — minimize $\sum \text{loss}(UV'_{ij};Y_{ij})$

The Aspect Model (pLSA)

[Hoffman+99]



$$Y|X \sim \text{Multinomial}(N,X)$$

$$Y_{ij}|X_{ij} \sim \text{Binomial}(N,X_{ij})$$

$N=\sum Y_{ij}$

Low-Rank Models for Matrices of Counts, Occurrences or Frequencies

	Multinomial	Independent Binomials	Independent Bernoulli
Mean parameterization $0 \leq X_{ij} \leq 1$ $E[Y_{ij}\|X_{ij}]=X_{ij}$	Aspect Model (pLSA) [Hoffman+99] $\equiv$ NMF if $\sum X_{ij}=1$	$Y_{ij}\|X_{ij}$~$Bin(N,X_{ij})$ $\approx$ NMF [Lee+01]	$P(Y_{ij}=1) = X_{ij}$
Natural parameterization unconstrained X_{ij}	Sufficient Dimensionality Reduction [Globerson+02]	$Y_{ij}\|X_{ij}$~$Bin(N,g(X_{ij}))$	Logistic Low Rank Approximation [Schein+03]

$\approx$ hinge loss

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

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

Outline

- Finding Low Rank Approximations
  - **W**eight **L**ow **R**ank **A**pprox: minimize $\sum_{ij} \mathbf{W}_{ij}(\mathbf{Y}_{ij} - \mathbf{X}_{ij})^2$
  - Use WLRA Basis for other losses / conditional models

- Consistency of Low Rank Approximation

  When more data is available, do we converge to correct solution? Not always…

- Matrix Factorization for Collaborative Filtering
  - **M**aximum **M**argin **M**atrix **F**actorization
  - Generalization Error Bounds (Low Rank & MMMF)

Finding Low Rank Approximation

Find rank-*k* X minimizing $\sum loss(X_{ij};Y_{ij})$ (=log p(Y|X))

- Non-convex:
  - "X is rank-*k*" is a non-convex constraint
  - $\sum loss((UV)_{ij};Y_{ij})$ not convex in U,V
- rank-*k* X minimizing $\sum(Y_{ij}-X_{ij})^2$:
  - non-convex, but no (non-global) local minima
  - solution: leading components of SVD
- For other loss functions, or with missing data:
  cannot use SVD, local minima, difficult problem
- Weighted Low Rank Approximation:
  minimize $\sum W_{ij}(Y_{ij}-X_{ij})^2$

Arbitrarily specified weights (part of input)

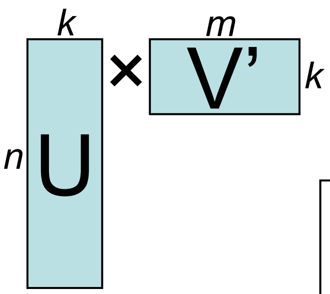
WLRA: Optimization

$$J(UV') = \sum_{ij} W_{ij} (Y - UV')^2_{ij}$$

For fixed *V*, find optimal *U*

For fixed *U*, find optimal V

$$J^*(V) = \min_U J(UV')$$



$$\tfrac{\partial}{\partial V} J^*(V') = 2U^{*\prime}((U^* V' - Y) \otimes W)$$

Conjugate gradient descent on *J**

Optimize *km* parameters instead of *k(n+m)*

EM approach:

elementwise product

$$X \leftarrow \mathrm{LowRankApprox}\big(W \otimes Y + (1 - W) \otimes X\big)$$

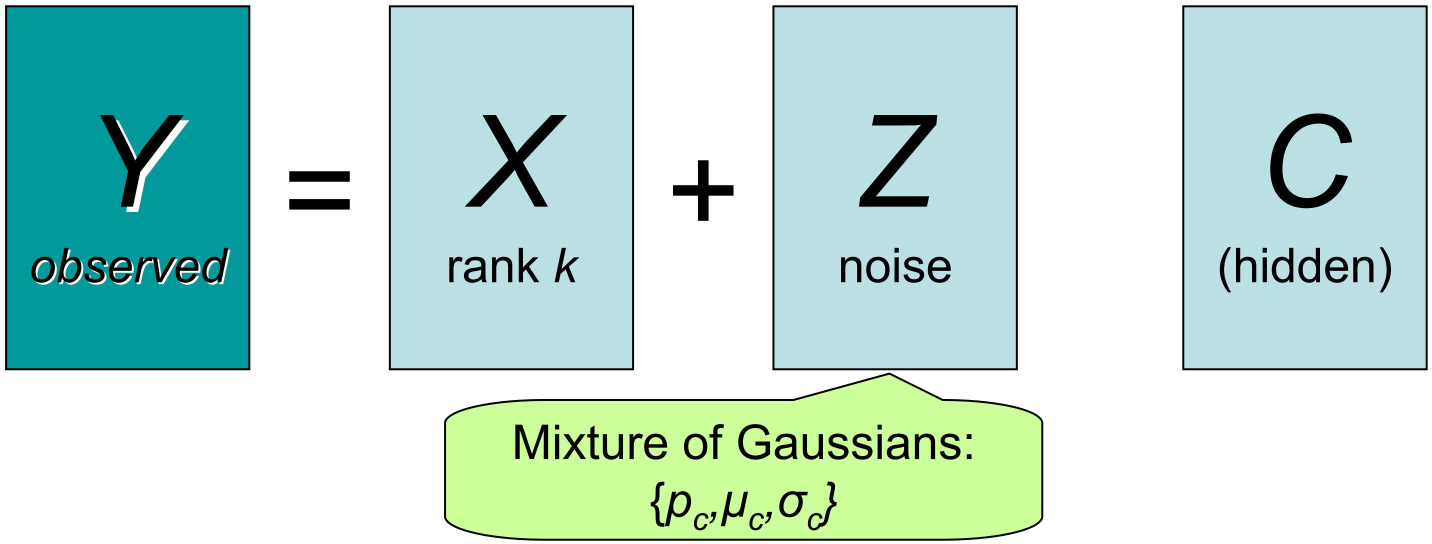
Newton Optimization for Non-Quadratic Loss Functions

minimize $\sum_{ij} \text{loss}(X_{ij};Y_{ij})$

$\text{loss}(X_{ij};Y_{ij})$ convex in X_{ij}

- Iteratively optimize quadratic approximations of objective
- Each such quadratic optimization is a weighted low rank approximation

Maximum Likelihood Estimation with Gaussian Mixture Noise



E step: calculate posteriors of *C*

M step: WLRA with

$$W_{ij} = \sum_c \frac{\Pr(C_{ij} = c)}{\sigma_C^2} \qquad A_{ij} = Y_{ij} + \sum_c \frac{\Pr(C_{ij} = c)\mu_C}{\sigma_C^2} \Big/ W_{ij}$$

Outline

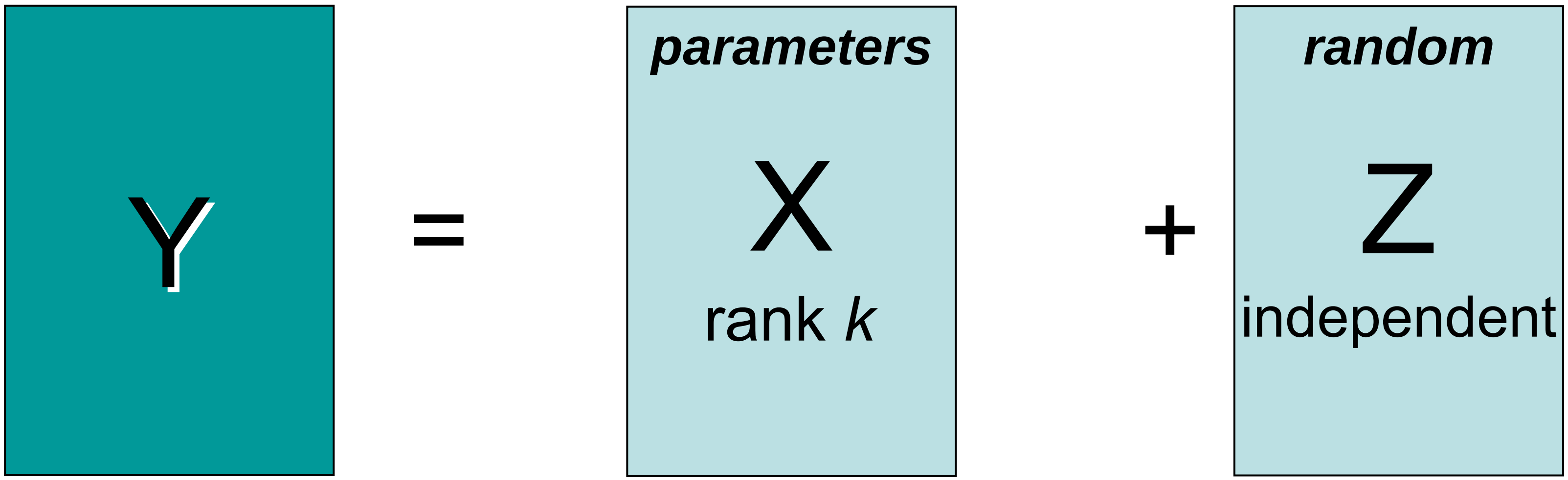
- Finding Low Rank Approximations
  - **W**eight **L**ow **R**ank **A**pprox: minimize $\sum_{ij} \mathbf{W}_{ij}(\mathbf{Y}_{ij}-\mathbf{X}_{ij})^2$
  - Use WLRA Basis for other losses / conditional models

➔ Consistency of Low Rank Approximation

When more data is available, do we converge to correct solution? Not always…

- Matrix Factorization for Collaborative Filtering
  - **M**aximum **M**argin **M**atrix **F**actorization
  - Generalization Error Bounds (Low Rank & MMMF)

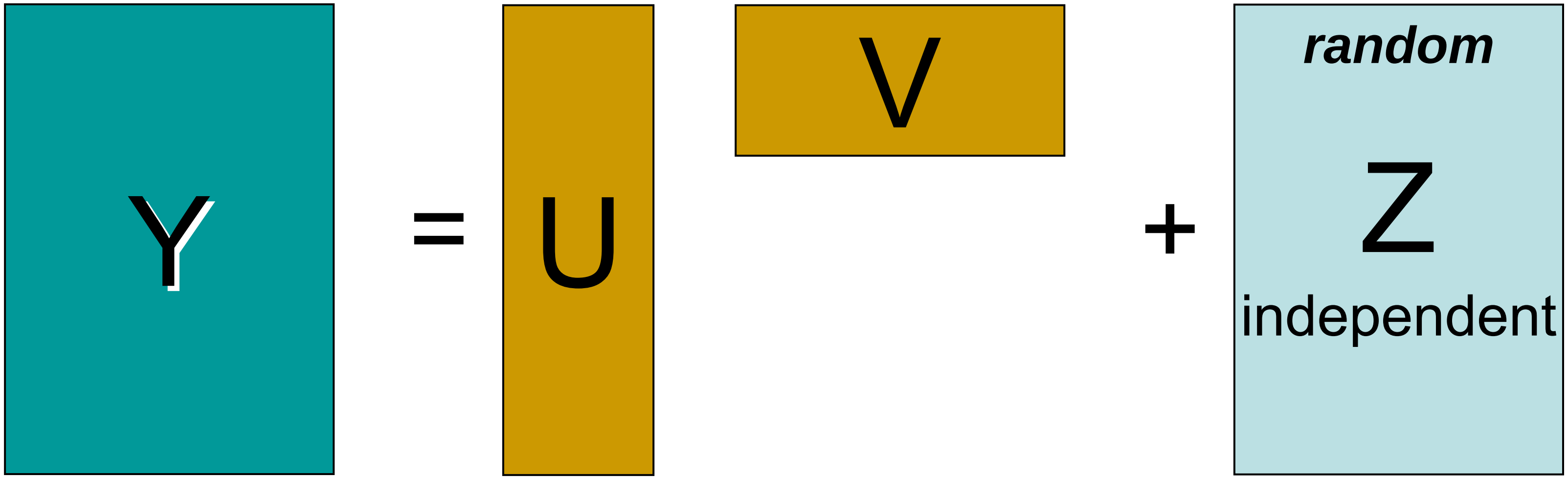
Asymptotic Behavior of Low Rank Approximation



Single observation,
Number of parameters is linear in number of observables

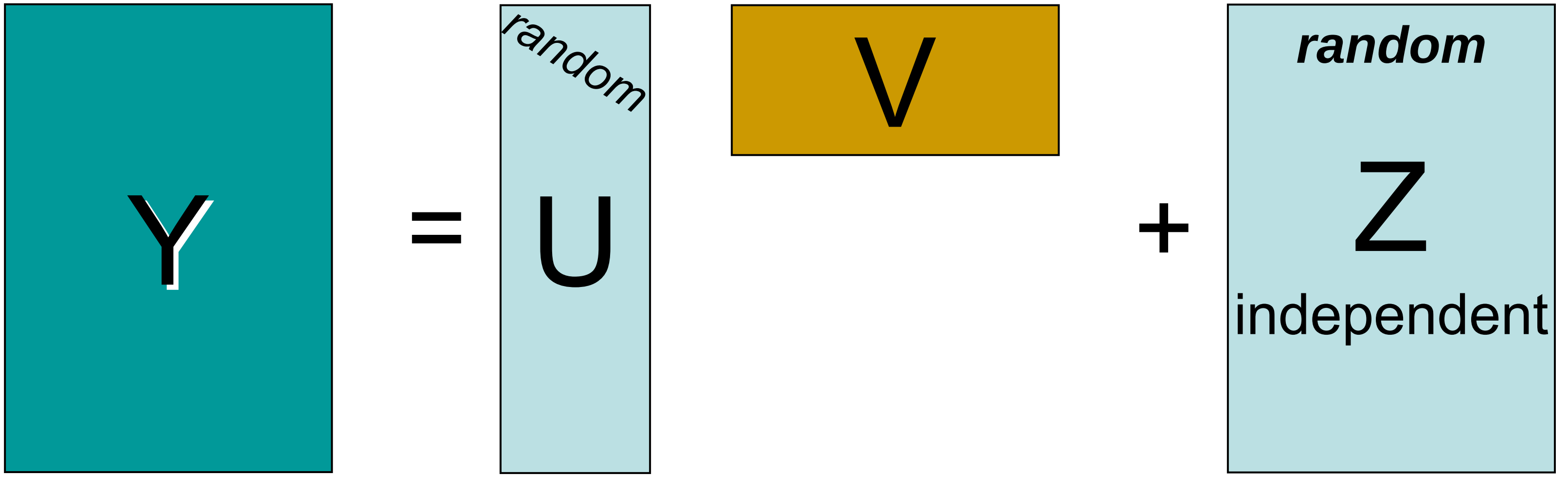
Can never approach correct estimation of parameters

What *can* be estimated is row-space of X

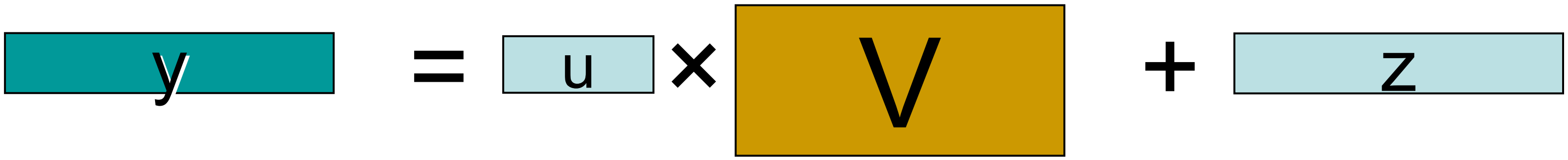


Number of parameters is linear in number of observables

What *can* be estimated is row-space of X



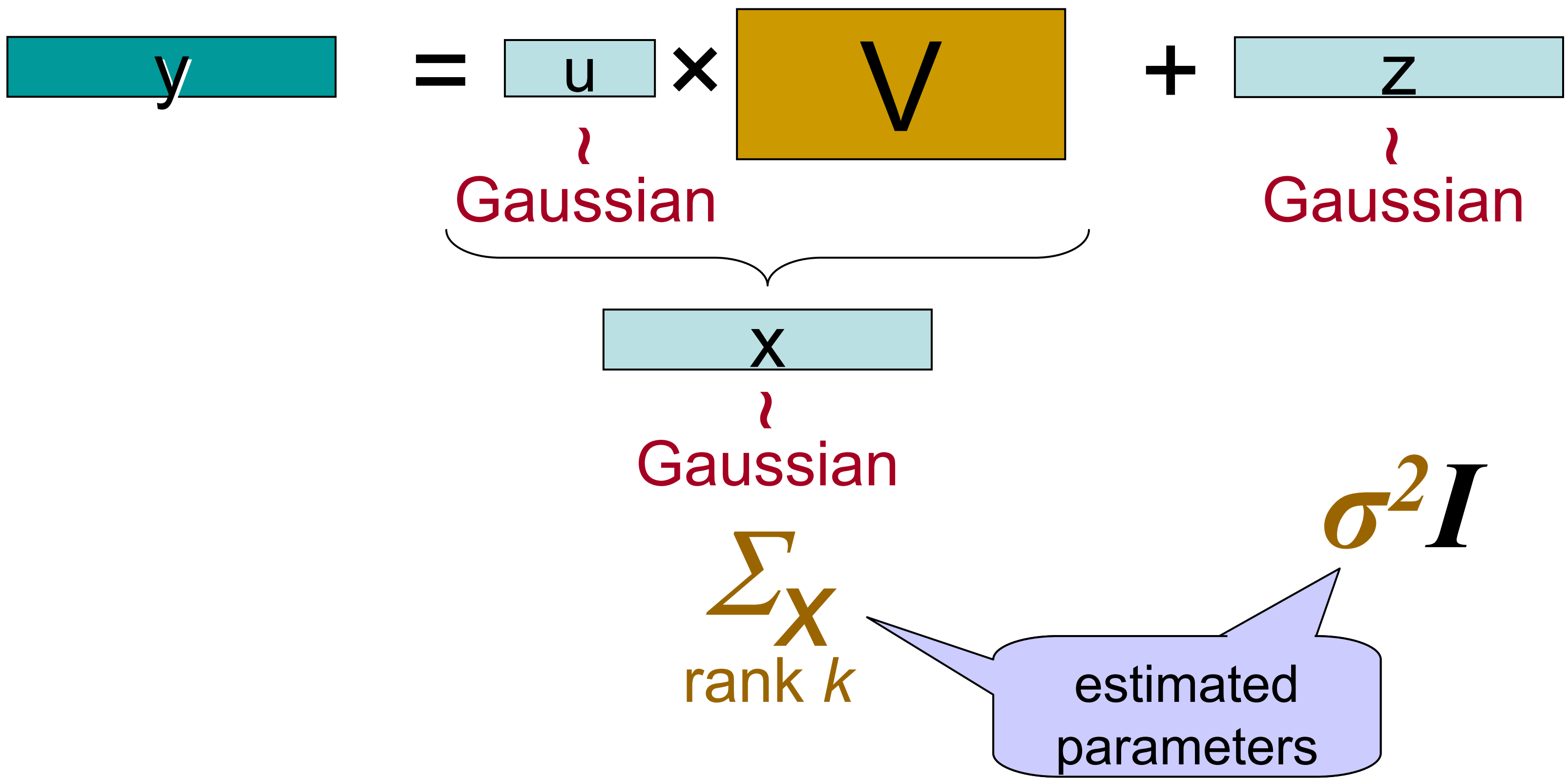
What *can* be estimated is row-space of X



Multiple samples of random variable y

What *can* be estimated is row-space of X

Probabilistic PCA



Maximum likelihood ≡ PCA

[Tipping Bishop 97]

Probabilistic PCA

[Tipping Bishop 97]

$y = u \times V + z$

$u \sim$ Gaussian

$z \sim$ Gaussian($\sigma^2 I$)

estimated parameters

Latent Dirichlet Allocation

[Blei Ng Jordan 03]

$y \sim$ Multinomial(N, $u \times V$)

$u \sim$ Dirichlet(α)

estimated parameters

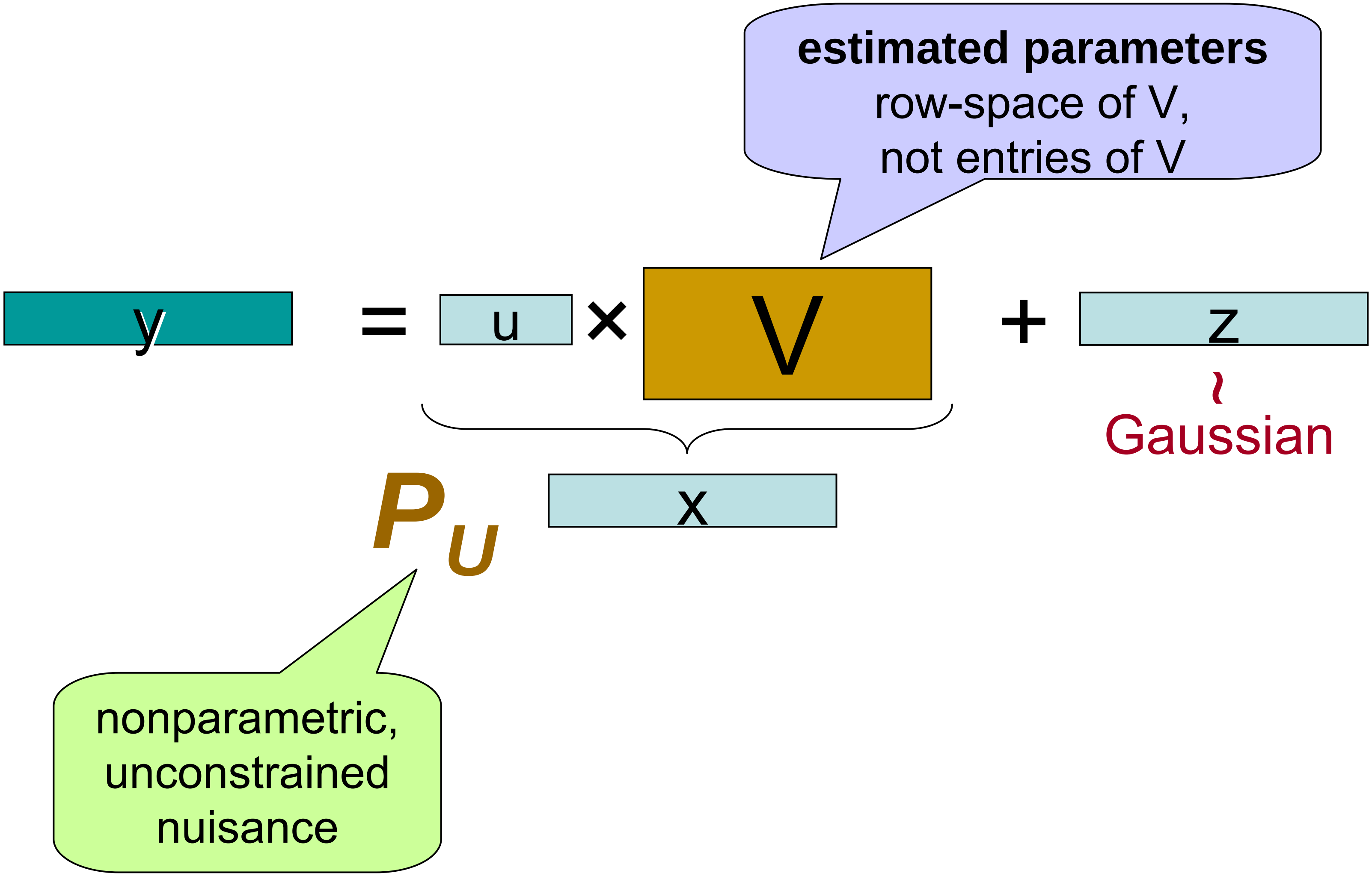
Generative and Non-Generative Low Rank Models

Y=**X**+Gaussian ⟷ "Probabilistic PCA"

pLSA,
Y~Binom(N,**X**) ⟷ Latent Dirichlet Allocatoin

Non-parametric generative models

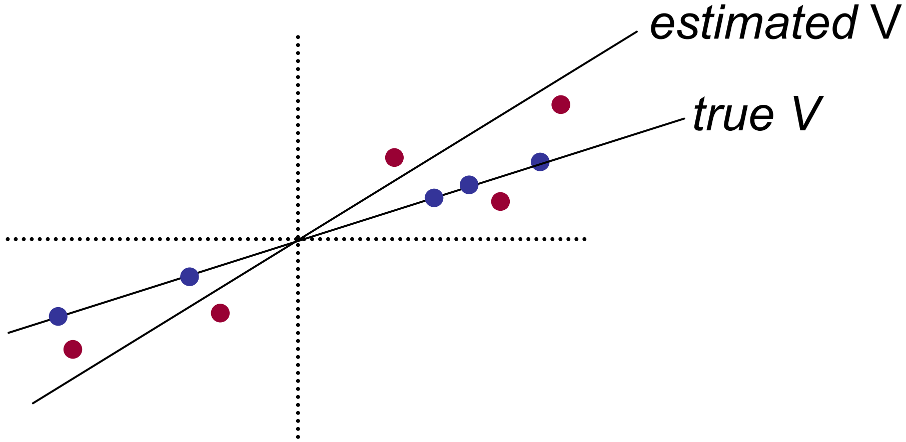
Parametric generative models:
Consistency of Maximum Likelihood estimation guaranteed
if model assumptions hold
(both on **Y|X** *and* on **U**)



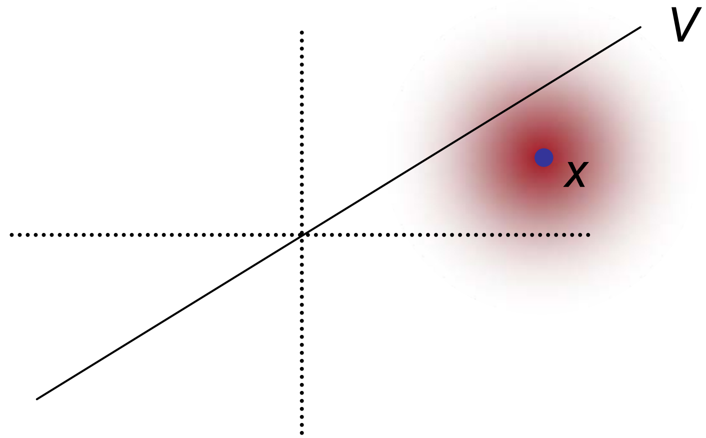
Non-parametric model,
estimation of a parametric part of the model

Maximum Likelihood ≡ “Single Observed Y”

Consistency of ML Estimation



Consistency of ML Estimation



expected contribution of *x* to likelihood of *V*

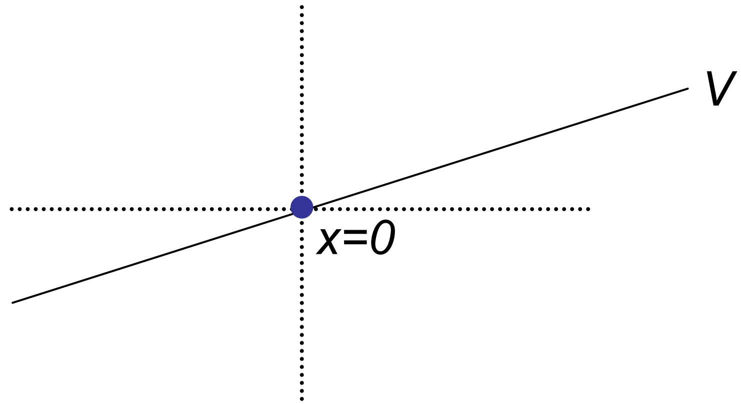
$$\Psi(V;x) = E_z\left[\max_u \log p_Z((x+z) - uV)\right]$$

ML estimator is consistent for any P_u	$\rightleftarrows$	for all x, V maximizes $\Psi(V;x)$ iff V spans x

When *Z*~Gaussian, $\Psi(V;x) =$ **E**[L_2 distance of x+z from *V*]

For iid Gaussian noise, ML estimation (PCA) of the low-rank sub-space is consistent

Consistency of ML Estimation



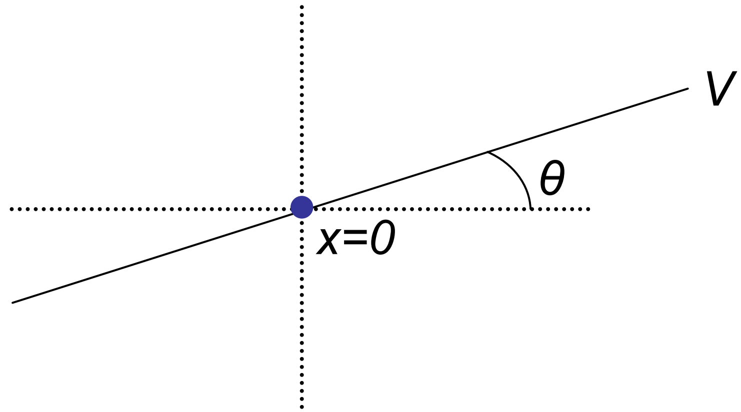
expected contribution of *x* to likelihood of *V*

$$\Psi(V;x) = E_z\left[\max_u \log p_Z((x+z) - uV)\right]$$

ML estimator is consistent for any P_u $\Leftrightarrow$ for all x, V maximizes $\Psi(V;x)$ iff V spans x

ML estimator is consistent for any P_u $\Rightarrow$ $\Psi(V;0)$ is constant for all V

Consistency of ML Estimation

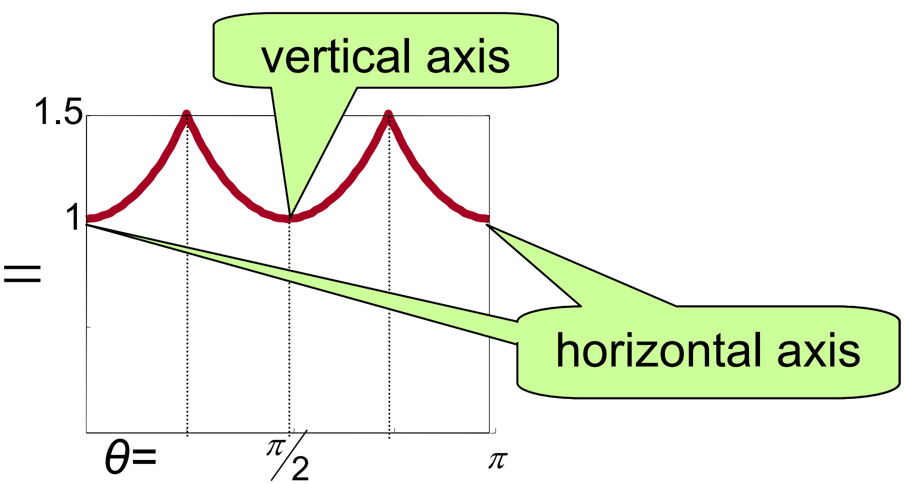


expected contribution of *x* to likelihood of *V*

$$\Psi(V;x) = E_z\left[\max_u \log p_Z((x+z)-uV)\right]$$

Laplace: $p_Z(z[i]) = \frac{1}{2}e^{-|z[i]|}$

$$-\Psi(V;0) =$$



Consistency of ML Estimation

X=UV'

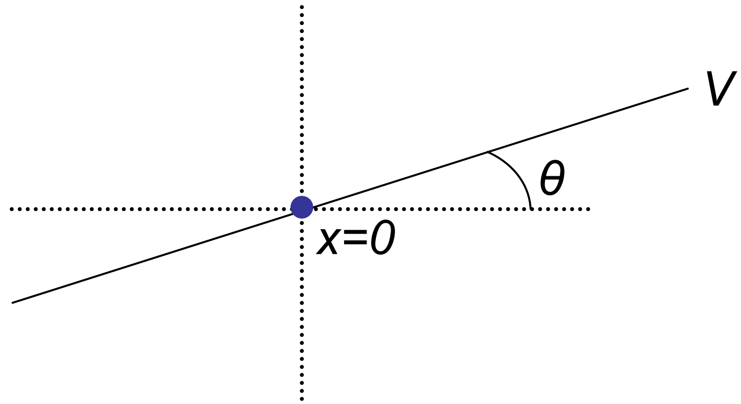
General conditional model for Y|X

expected contribution of *x* to likelihood of *V*

$$\Psi(V;x) = E_{Y|X}\left\lfloor \max_u \log p_{Y|X}(Y \mid uV) \mid x \right\rfloor$$

ML estimator is consistent for any P_u	$\rightleftarrows$	for all x, V maximizes $\Psi(V;x)$ iff V spans x
ML estimator is consistent for any P_u	$\Longrightarrow$	$\Psi(V;0)$ is constant for all V

Consistency of ML Estimation

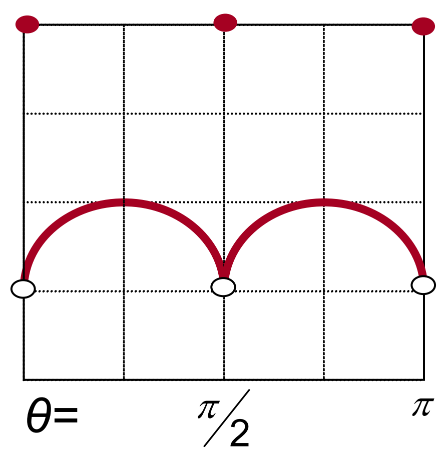


expected contribution of *x* to likelihood of *V*

$$\Psi(V;x) = E_{Y|X}\left[\max_u \log p_{Y|X}(Y \mid uV) \mid x\right]$$

Logistic: $p_{Y|X}(Y[i]=1 \mid x[i]) = \dfrac{1}{1-e^{-x[i]}}$

$$-\Psi(V;0) =$$



Consistent Estimation

- Additive i.i.d. noise Y=X+Z:
  - Maximum Likelihood generally not consistent
  - PCA is consistent (for any noise distribution)

Span of k leading eigenvectors of $\hat{\Sigma}_{\mathbf{Y}} \rightarrow \Sigma_{\mathbf{X}} + \sigma^2 \boldsymbol{I}$

rank k

Consistent Estimation

- Additive i.i.d. noise Y=X+Z:
  - Maximum Likelihood generally not consistent
  - PCA is consistent (for any noise distribution)

$s_1, s_2, \ldots, s_k, 0, 0, \ldots, 0$

$$\hat{\Sigma}_Y \rightarrow \Sigma_Y = \Sigma_X + \Sigma_Z = \Sigma_X + \sigma^2 I$$

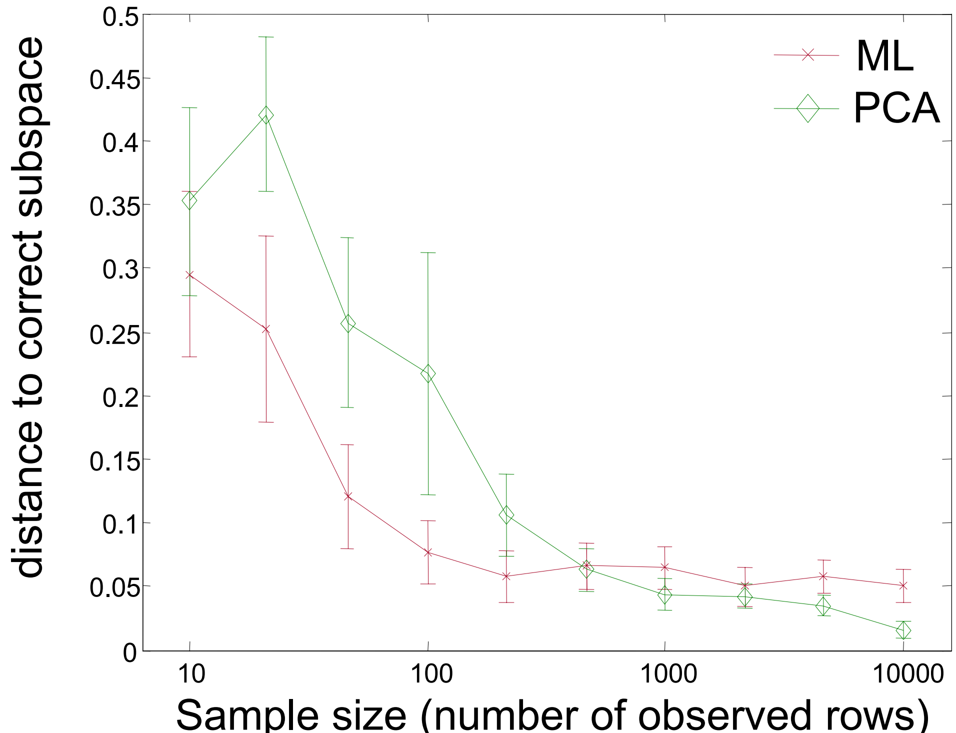
$s_1+\sigma^2, s_2+\sigma^2, \ldots, s_k+\sigma^2, \sigma^2, \sigma^2, \ldots, \sigma^2$

Consistent Estimation

- Additive i.i.d. noise Y=X+Z:

  Maximum Likelihood generally not consistent

  PCA is consistent (for any noise distribution)



Noise:
0.99 *N*(0,1) + 0.01 *N*(0,100)

Consistent Estimation

- Additive i.i.d. noise Y=X+Z:
  Maximum Likelihood generally not consistent
  PCA is consistent (for any noise distribution)

- Unbiased noise E[Y|X]=X*:*
  Maximum Likelihood generally not consistent
  PCA not consistent
  -- Can correct by ignoring diagonal of covariance

- Exponential PCA (X are *natural* parameters)
  Maximum Likelihood generally not consistent
  Covariance methods not consistent
  ???

Outline

- Finding Low Rank Approximations
  - **W**eight **L**ow **R**ank **A**pprox: minimize $\sum_{ij} \mathbf{W}_{ij}(\mathbf{Y}_{ij}\text{-}\mathbf{X}_{ij})^2$
  - Use WLRA Basis for other losses / conditional models
- Consistency of Low Rank Approximation

  When more data is available, do we converge to correct solution? Not always…

- → Matrix Factorization for Collaborative Filtering
  - **M**aximum **M**argin **M**atrix **F**actorization
  - Generalization Error Bounds (Low Rank & MMMF)

Collaborative Filtering

Based on preferences so far, and preferences of others:

$\Rightarrow$ Predict further preferences

movies

users

	2		1			4				5	
	5		4						1		3
		3		5			2				
4						5		3			
		4		1	3				5		
			2				1				4
	1					5		5		4	
		2			5				4		
	3		3		1		5		2		1
	3				1			2		3	
	4			5	1			3			
		3				3				5	
2			1		1						
		5			2			4		4	
	1		3		1	5		4		5	
1		2			4				5		

Implicit or explicit preferences?

Type of queries.

Matrix Completion

Based on partially observed matrix:

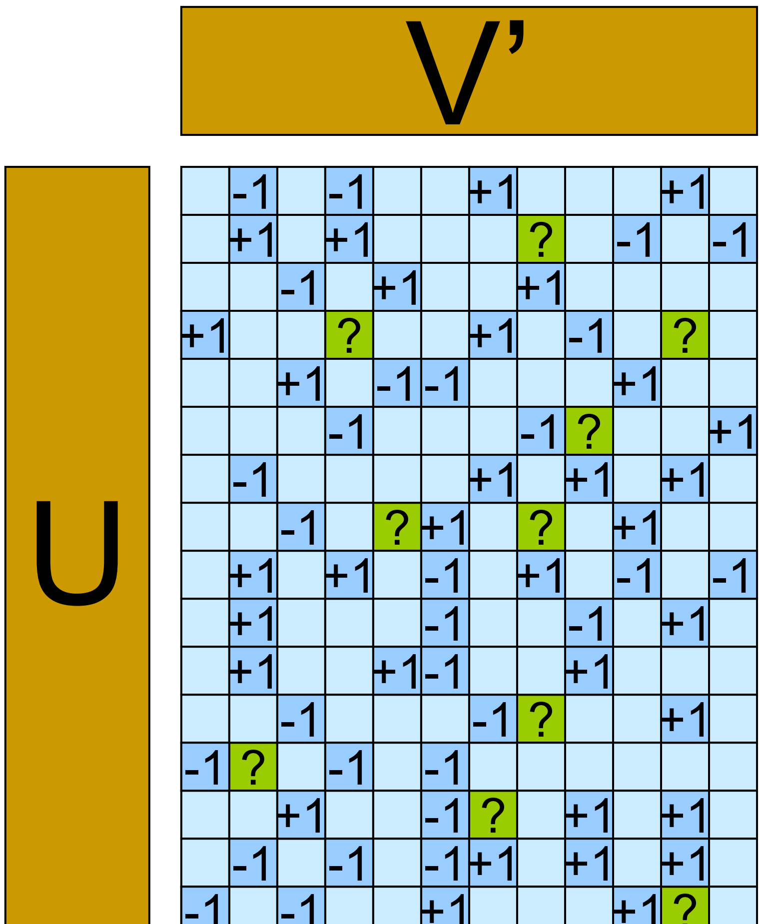
⇒ Predict unobserved entries "Will user *i* like movie *j*?"

movies

users

	2		1			4				5	
	5		4				?		1		3
		3		5			2				
4			?			5		3		?	
		4		1	3				5		
			2				1	?			4
	1					5		5		4	
		2		?	5		?		4		
	3		3		1		5		2		1
	3				1			2		3	
	4			5	1			3			
		3				3	?			5	
2	?		1		1						
		5			2	?		4		4	
	1		3		1	5		4		5	
1		2			4				5	?	

Matrix Completion with Matrix Factorization



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

minimize $\Sigma \text{loss}(X_{ij};Y_{ij})$

prediction

observation

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

[Sarwar+00] [Azar+01] [Hoffman04] [Marlin+04]

Matrix Completion with Matrix Factorization

1.3	0.4	-1.5
8.3	2.5	-4.8
0.7	-0.2	3.4
1.7	-5.2	1.6
-3.7	2.1	0.9
4.3	-0.5	2.7
4.7	0.2	6.4
6.0	0.3	-5.8
-1.5	-3.7	0.4
-4.8	4.3	2.5
3.4	4.7	-0.2
1.6	6.0	-5.2
0.9	1.3	2.1
2.7	8.3	-0.5
6.4	0.7	0.2
-5.8	1.7	0.3

v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12
	-1		-1			+1				+1	
	+1		+1						-1		-1
		-1		+1			+1				
+1						+1		-1			
		+1		-1	-1				+1		
			-1				-1				+1
	-1					+1		+1		+1	
		-1			+1				+1		
	+1		+1		-1		+1		-1		-1
	+1				-1			-1		+1	
	+1			+1	-1			+1			
		-1				-1				+1	
-1			-1		-1						
		+1			-1			+1		+1	
	-1		-1		-1	+1		+1		+1	
-1		-1			+1				+1		

When U is fixed,

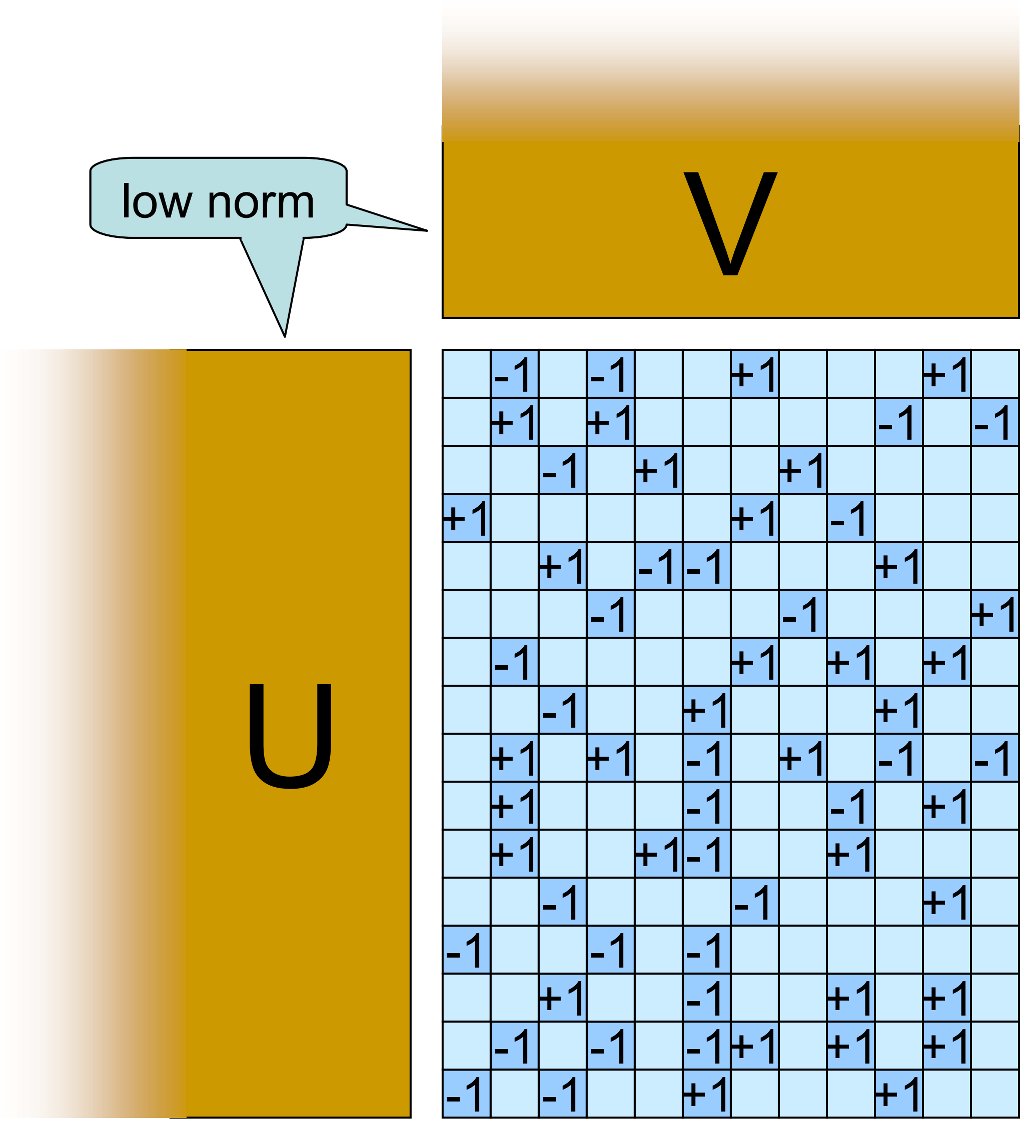
each row is a linear classification problem:

- 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.

Max-Margin Matrix Factorization



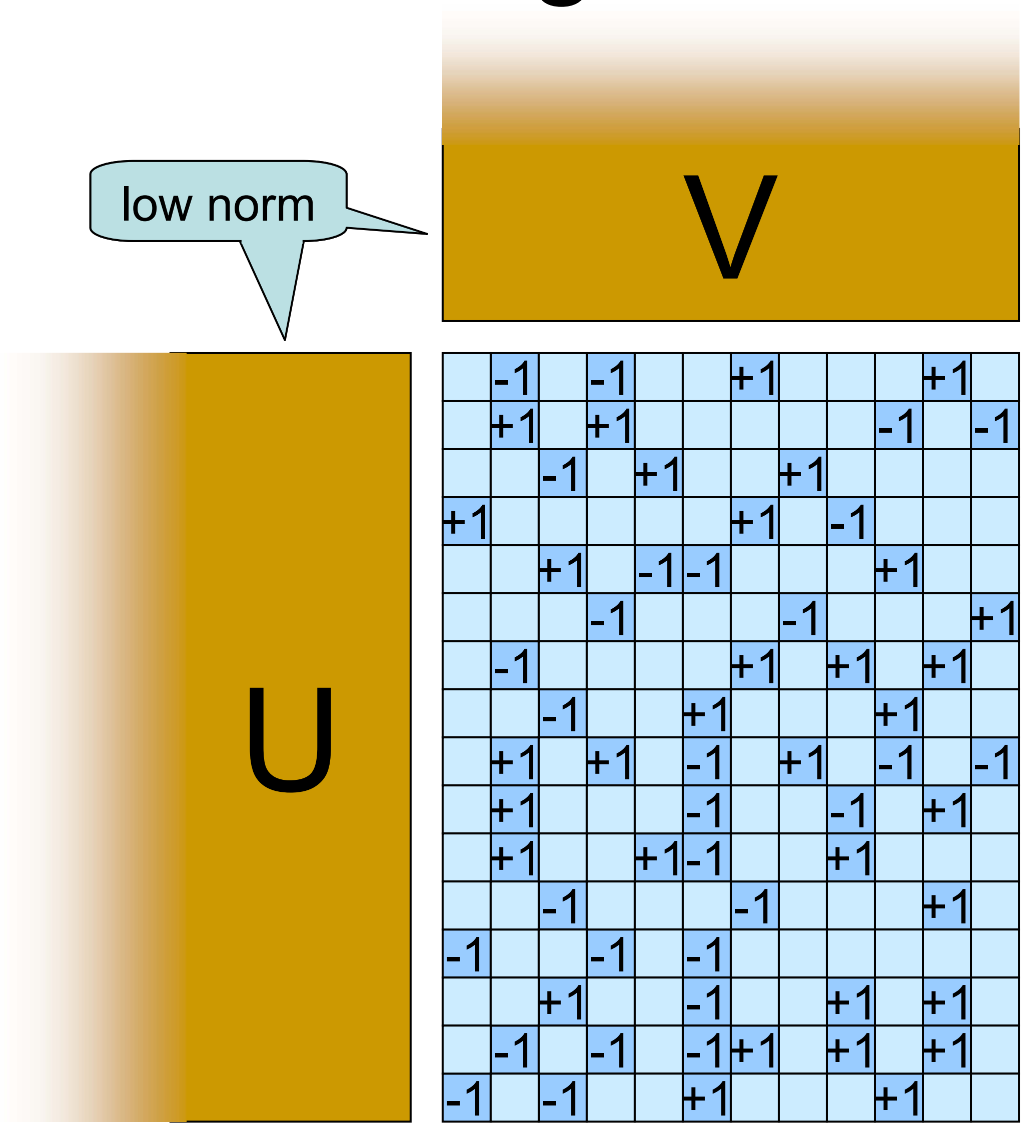
Instead of bounding dimensionality of U,V, bound norms of U,V

For observed $Y_{ij} \in \pm 1$:

$$Y_{ij} X_{ij} \geq \textit{Margin}$$

$\langle U_i, V_j \rangle$

Max-Margin Matrix Factorization



bound norms on average:

$(\sum_i |U_i|^2)\ (\sum_j |V_j|^2) \leq 1$

bound norms uniformly:

$(\max_i |U_i|^2)\ (\max_j |V_j|^2) \leq 1$

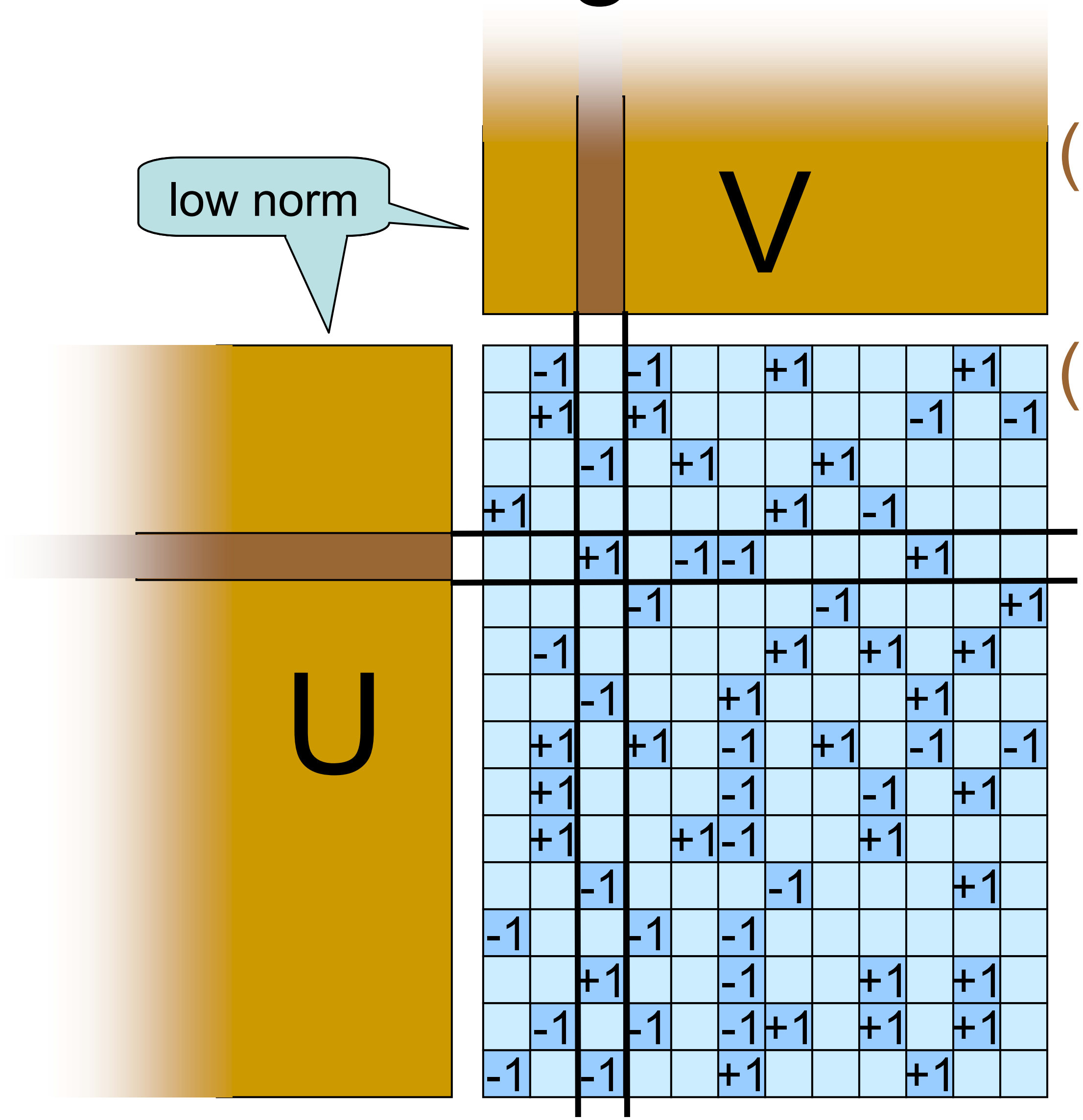
For observed $Y_{ij} \in \pm 1$:

$Y_{ij}\ X_{ij} \geq$ *Margin*

$\langle U_i, V_j \rangle$

U is fixed:
each column of V is SVM

Max-Margin Matrix Factorization



bound norms on average:

$(\sum_i |U_i|^2)\ (\sum_j |V_j|^2) \leq 1$

bound norms uniformly:

$(\max_i |U_i|^2)\ (\max_j |V_j|^2) \leq 1$

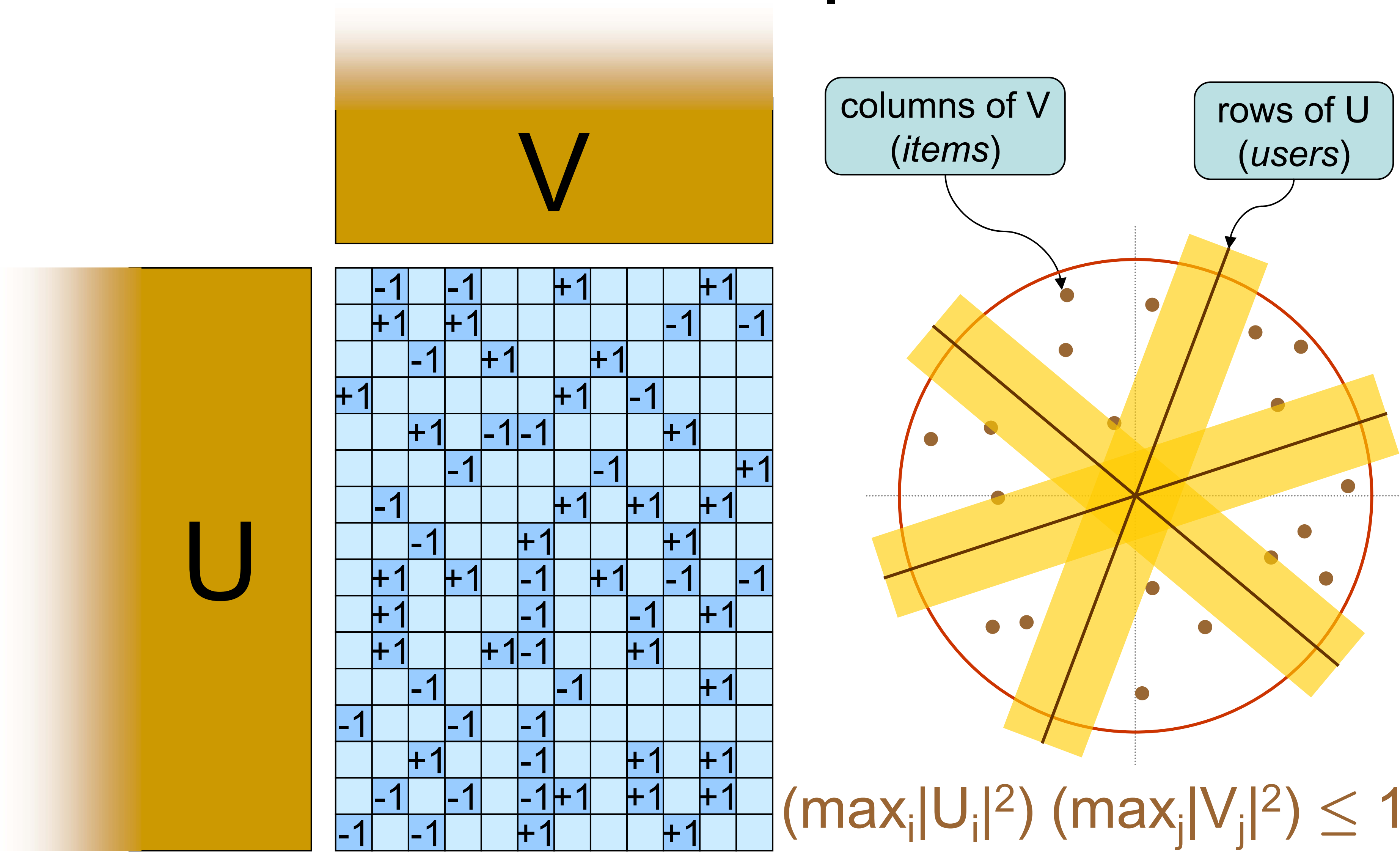
For observed $Y_{ij} \in \pm 1$:

$Y_{ij}\ X_{ij} \geq$ *Margin*

$\langle U_i, V_j \rangle$

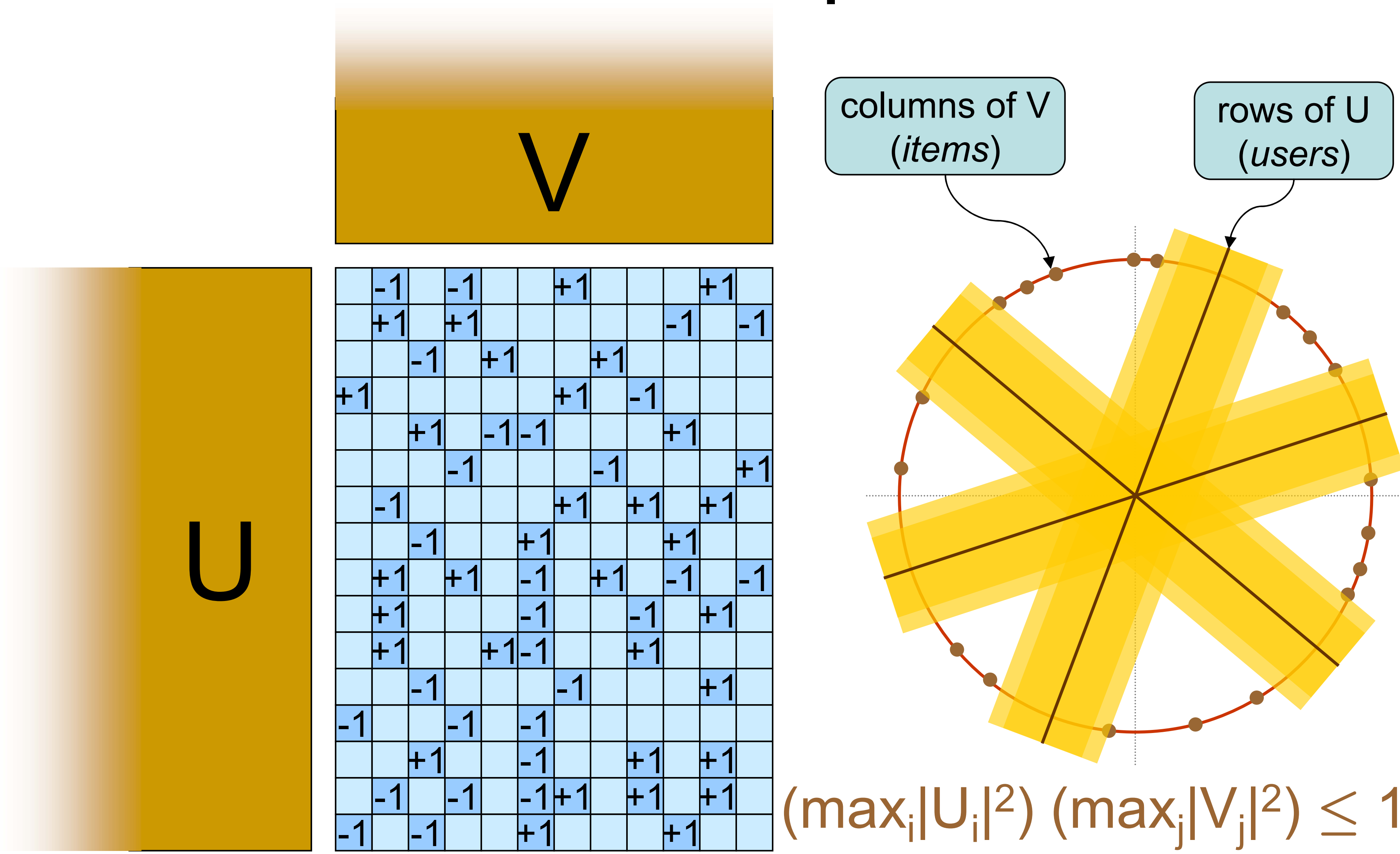
U is fixed:
each column of V is SVM

Geometric Interpretation



$$(\max_i |U_i|^2)\ (\max_j |V_j|^2) \leq 1$$

Geometric Interpretation



$$(\max_i |U_i|^2)\ (\max_j |V_j|^2) \leq 1$$

Finding Max-Margin Matrix Factorizations

maximize M

$Y_{ij} X_{ij} \geq M$

$X = UV$

$(\sum_i |U_i|^2)\ (\sum_j |V_j|^2) \leq 1$

maximize M

$Y_{ij} X_{ij} \geq M$

$X = UV$

$(\max_i |U_i|^2)\ (\max_j |V_j|^2) \leq 1$

Unlike rank($\boldsymbol{X}$) $\leq k$, these are convex constraints!

Finding Max-Margin Matrix Factorizations

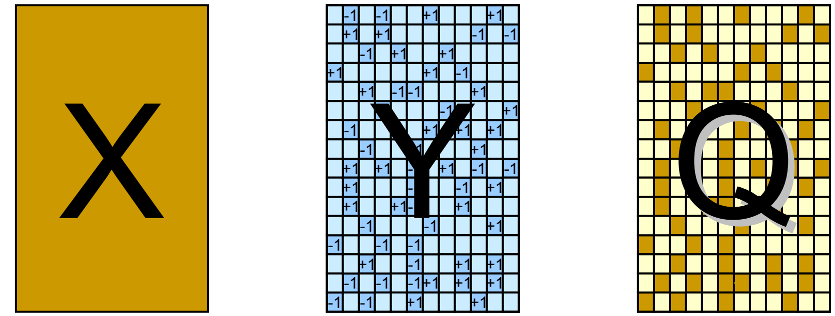
maximize M

$Y_{ij} X_{ij} \geq M$

$X = UV$

$(\sum_i |U_i|^2)\,(\sum_j |V_j|^2) \leq 1$

$|X|_{tr} = \sum$ (singular values of X)



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

minimize $\mathrm{tr}(A) + \mathrm{tr}(B) + c \sum \xi_{ij}$

$Y_{ij} X_{ij} \geq 1 - \xi_{ij}$

$\begin{pmatrix} A & X \\ X' & B \end{pmatrix}$ p.s.d.

maximize $\sum Q_{ij}$

$0 \leq Q_{ij} \leq c$

$||Q \otimes Y||_2 \leq 1$

sparse elementwise product
(zero for unobserved entries)

Finding Max-Margin Matrix Factorizations

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

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

minimize $\mathrm{tr}(A)+\mathrm{tr}(B)+ c \sum \xi_{ij}$

$Y_{ij} X_{ij} \geq 1 - \xi_{ij}$

$\begin{pmatrix} A & X \\ X' & B \end{pmatrix}$ p.s.d.

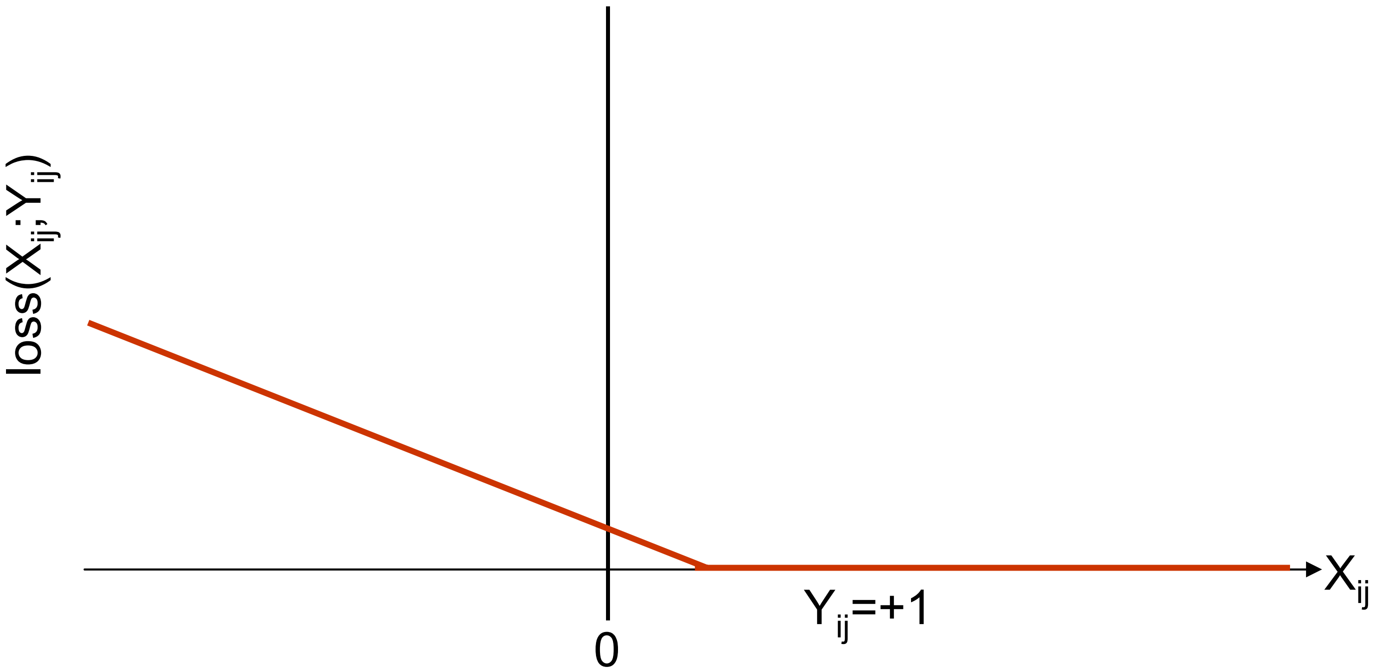
maximize $\sum Q_{ij}$

$0 \leq Q_{ij} \leq c$

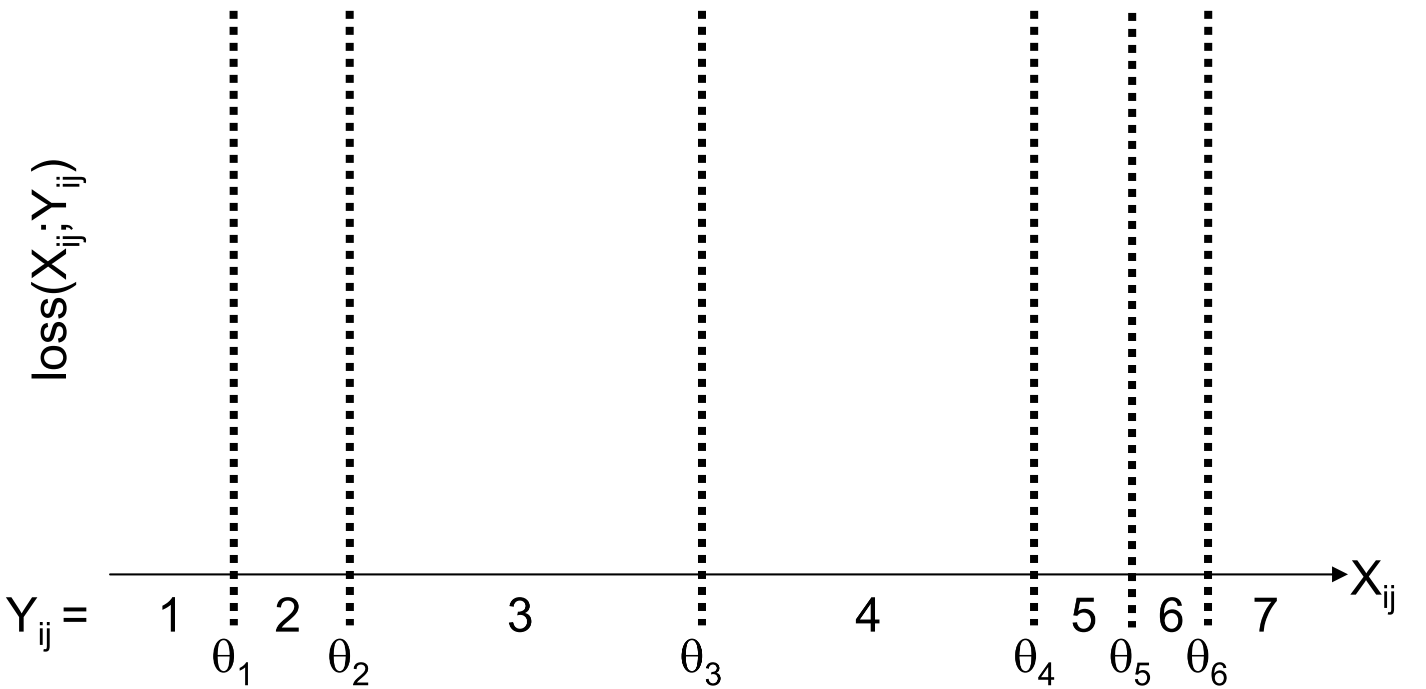
$||Q \otimes Y||_2 \leq 1$

sparse elementwise product
(zero for unobserved entries)

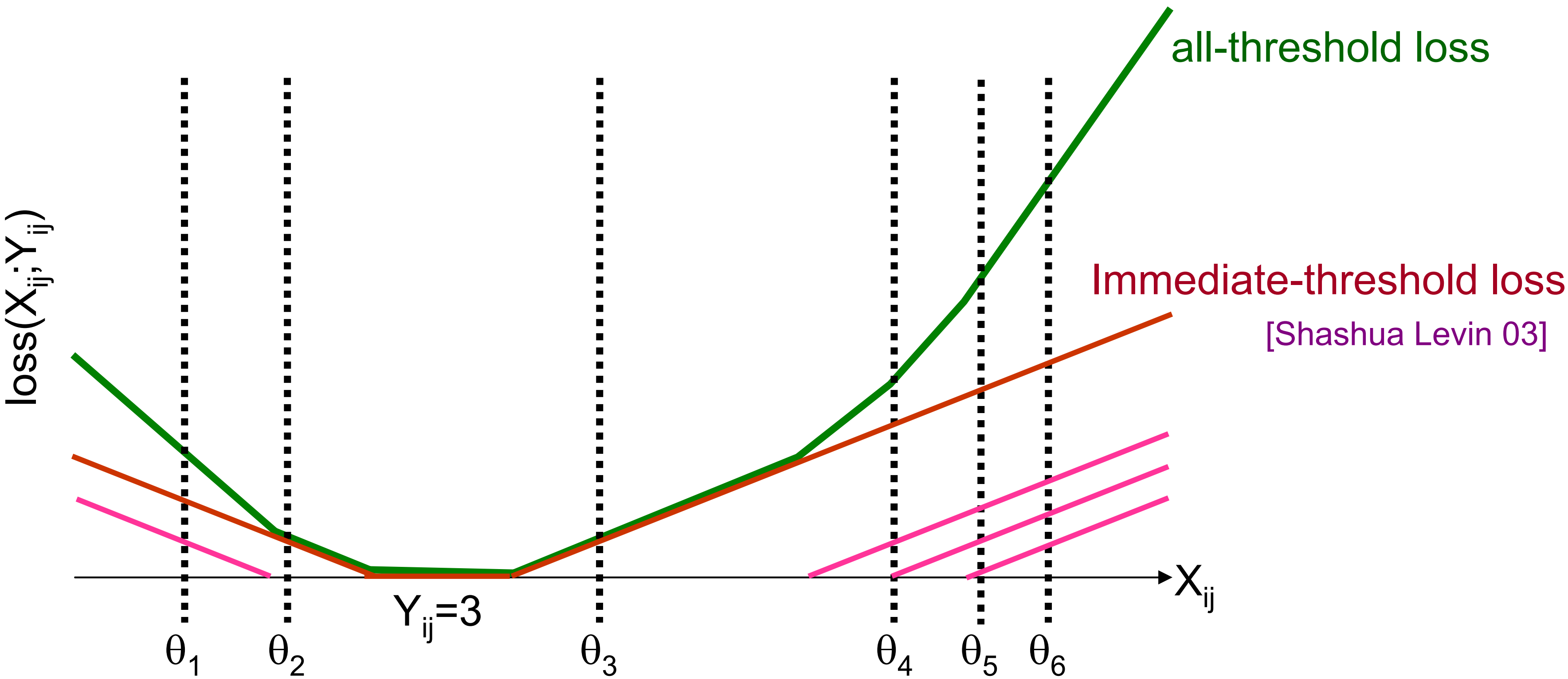
Loss Functions for Rankings



Loss Functions for Rankings



Loss Functions for Rankings



- All-threshold loss is a bound on the absolute rank-difference
- For both loss functions: learn per-user θ's

Experimental Results on MovieLens Subset

	all threshold MMMF	immediate threshold MMMF	K-medians K=2	Rank-1	Rank-2
Rank Difference	**0.670**	0.715	0.674	0.698	0.714
Zero One Error	0.553	**0.542**	0.558	0.559	0.553

100 users $\times$ 100 movies subset of MovieLens, 3515 training ratings, 3515 test ratings

Outline

- Finding Low Rank Approximations
  - **W**eight **L**ow **R**ank **A**pprox: minimize $\sum_{ij}\mathbf{W}_{ij}(\mathbf{Y}_{ij}-\mathbf{X}_{ij})^2$
  - Use WLRA Basis for other losses / conditional models
- Consistency of Low Rank Approximation

  When more data is available, do we converge to correct solution? Not always…
- Matrix Factorization for Collaborative Filtering
  - **M**aximum **M**argin **M**atrix **F**actorization
  - → Generalization Error Bounds (Low Rank & MMMF)

Generalization Error Bounds

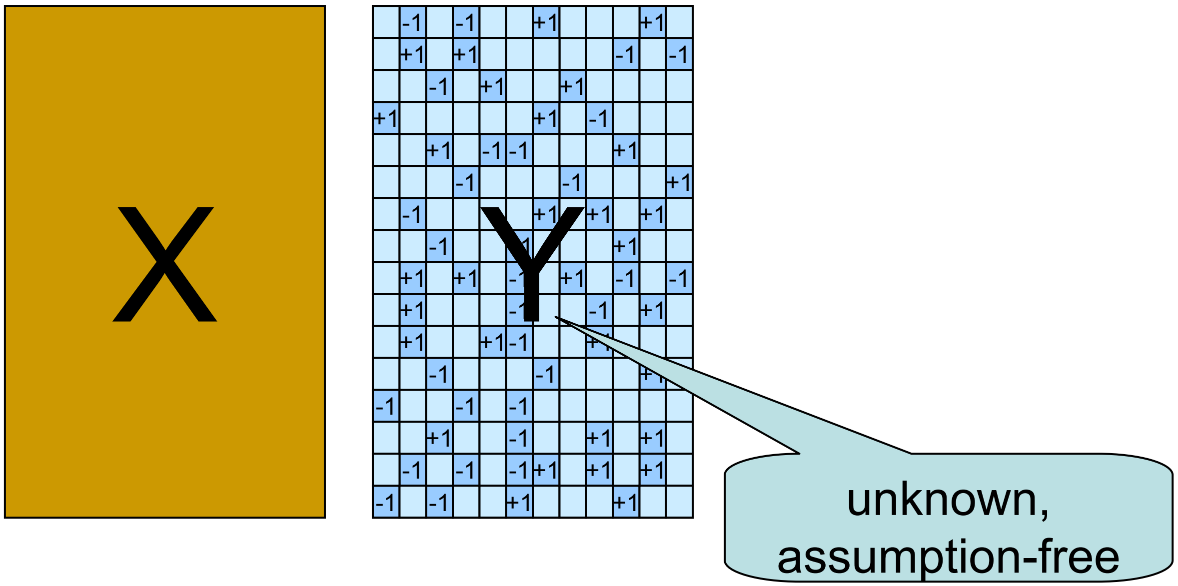
$D(\mathbf{X};\mathbf{Y}) = \sum_{ij} \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})$

generalization error

Assuming a low-rank structure (eigengap):

Asymptotic behavior [Azar+01]

Sample complexity, query strategy [Drineas+02]



Generalization Error Bounds

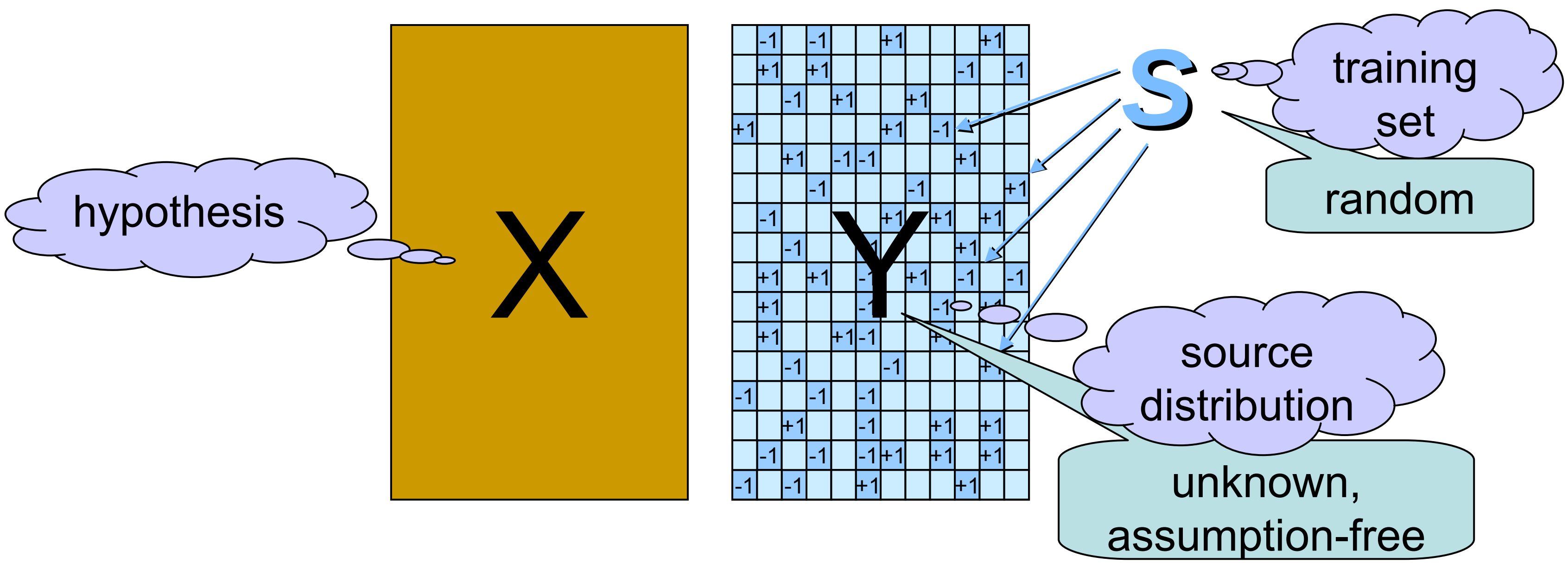
$D(\mathbf{X};\mathbf{Y}) = \sum_{ij} \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})$

generalization error

$D_{\mathbf{S}}(\mathbf{X};\mathbf{Y}) = \sum_{ij\in \mathbf{S}} \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})$

empirical error

$$\forall_{\mathbf{Y}} \Pr_{\mathbf{S}} \left(\forall_{\text{rank-}k\ \mathbf{X}}\ D(\mathbf{X};\mathbf{Y}) < D_{\mathbf{S}}(\mathbf{X};\mathbf{Y}) + \varepsilon \right) > 1-\delta$$



Generalization Error Bounds

0/1 loss: $\text{loss}(X_{ij};Y_{ij}) = 1$ when $\text{sign}(X_{ij}) \neq Y_{ij}$

$D(\mathbf{X};\mathbf{Y}) = \sum_{ij} \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})/nm$
generalization error

$D_{\mathbf{S}}(\mathbf{X},\mathbf{Y}) = \sum_{ij\in \mathbf{S}} \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})/|\mathbf{S}|$
empirical error

For particular **X**,**Y**: $\text{loss}(X_{ij};Y_{ij}) \sim \text{Bernoulli}(D(X;Y))$

random

$\Pr(\, D_{\mathbf{S}}(\mathbf{X};\mathbf{Y}) < D(\mathbf{X};\mathbf{Y}) - \varepsilon \,) < e^{-2|\mathbf{S}|\varepsilon^2}$

random

Union bound over all possible **X**s:

$$\forall_{\mathbf{Y}} \Pr_{\mathbf{S}} (\, \forall_{\mathbf{X}}\, D(\mathbf{X},\mathbf{Y}) < D_{\mathbf{S}}(\mathbf{X},\mathbf{Y}) + \varepsilon \,) > 1-\delta$$

$$\varepsilon = \sqrt{\frac{\text{log(\# of possible }\mathbf{X}\text{s)} + \log \frac{1}{\delta}}{2|S|}}$$

Number of Sign Configurations of Rank-*k* Matrices

V

1.11	-0.81	[illegible]	2.24	0.57
0.22	-0.58	[illegible]	-0.52	0.00

U

-0.46	0.65
-0.20	0.70
0.02	0.74
-0.61	-0.59
-0.50	0.33
-0.38	0.31
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
1.27	-0.52
0.44	0.17
-0.06	0.09
-0.66	-0.29
-1.65	0.09
0.58	0.16

X

-0.36	-0.00	-0.98	-1.36	-0.26
-0.07	-0.24	-1.14	-0.82	-0.12
0.19	-0.44	-1.27	-0.34	0.01
-0.81	0.84	1.17	-1.07	-0.35
-0.48	0.21	-0.43	-1.28	-0.28
-0.35	0.13	-0.43	-1.02	-0.22
-0.54	0.68	[illegible]	-0.43	-0.20
-0.22	0.40	[illegible]	0.10	-0.05
1.13	-1.06	[illegible]	1.72	0.51
1.30	-0.73	0.55	3.12	0.72
0.53	-0.45	-0.40	0.90	0.25
-0.04	-0.00	-0.14	-0.17	-0.03
-0.80	0.70	0.67	-1.32	-0.37
-1.81	1.28	0.28	-3.73	-0.93
0.68	-0.56	-0.43	1.21	0.33

nk variables

mk variables

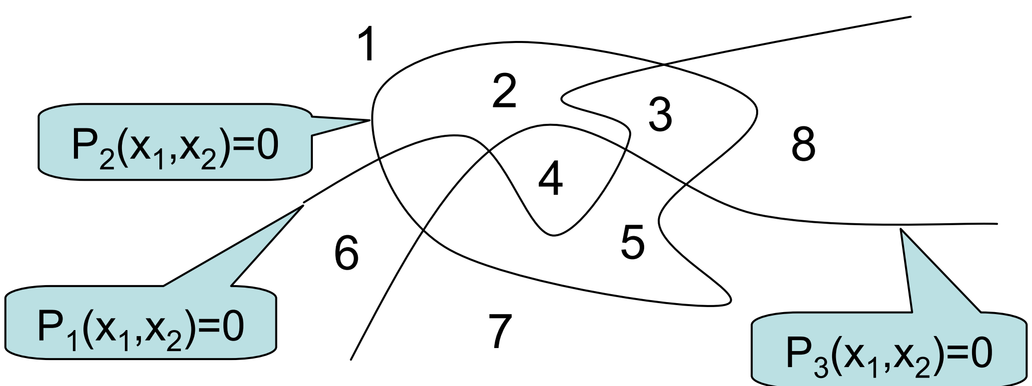
$$X_{i,j} = \sum_r U_{i,r} V_{r,j}$$

nm polynomials of degree 2

polynomials

Warren (1968): The number of connected components of $\{ \underline{x} \mid \forall_i \; P_i(\underline{x}) \neq 0 \}$

is at most $\left(\frac{4\, e\, (\text{degree})\, (\#\text{polys})}{(\#\text{ variables})} \right)^{(\#\text{ variables})}$



Based on [Alon95]

Number of Sign Configurations of Rank-*k* Matrices

V

1.11	-0.81	[illegible]	2.24	0.57
0.22	-0.58	[illegible]	-0.52	0.00

U

-0.46	0.65
-0.20	0.70
0.02	0.74
-0.61	-0.59
-0.50	0.33
-0.38	0.31
-0.35	-0.69
[illegible]	-0.56
[illegible]	[illegible]
1.27	-0.52
0.44	0.17
-0.06	0.09
-0.66	-0.29
-1.65	0.09
0.58	0.16

X

-0.36	-0.00	-0.98	-1.36	-0.26
-0.07	-0.24	-1.14	-0.82	-0.12
0.19	-0.44	-1.27	-0.34	0.01
-0.81	0.84	1.17	-1.07	-0.35
-0.48	0.21	-0.43	-1.28	-0.28
-0.35	0.13	-0.43	-1.02	-0.22
-0.54	0.68	[illegible]	-0.43	-0.20
-0.22	0.40	[illegible]	0.10	-0.05
1.13	-1.06	[illegible]	1.72	0.51
1.30	-0.73	0.55	3.12	0.72
0.53	-0.45	-0.40	0.90	0.25
-0.04	-0.00	-0.14	-0.17	-0.03
-0.80	0.70	0.67	-1.32	-0.37
-1.81	1.28	0.28	-3.73	-0.93
0.68	-0.56	-0.43	1.21	0.33

nk variables

mk variables

$$X_{i,j} = \sum_r U_{i,r} V_{r,j}$$

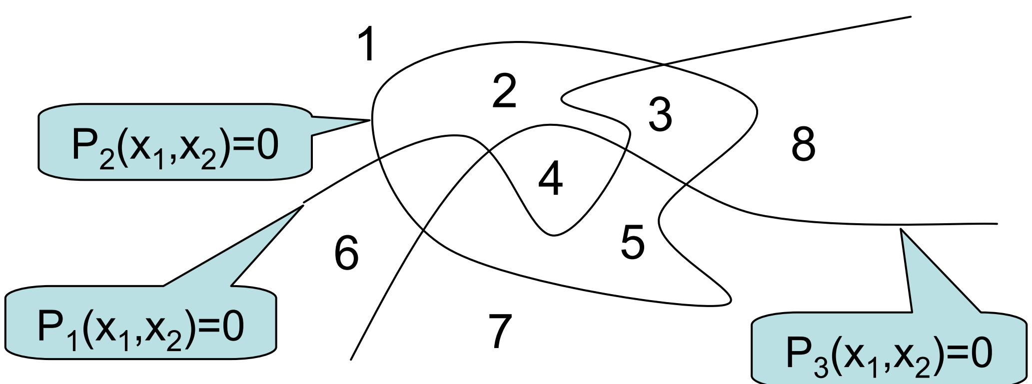
nm polynomials of degree 2

polynomials

Warren (1968): The number of connected components of $\{ \underline{x} \mid \forall_i\ P_i(\underline{x}) \neq 0 \}$

is at most

$$\left(\frac{4e \cdot 2 \cdot nm}{k(n+m)} \right)^{k(n+m)}$$



Based on [Alon95]

Generalization Error Bounds: Low Rank Matrix Factorization

$D(\mathbf{X};\mathbf{Y}) = \sum_{ij} \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})/nm$

generalization error

$D_S(\mathbf{X},\mathbf{Y}) = \sum_{ij\in S} \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})/|\mathbf{S}|$

empirical error

$\forall_{\mathbf{Y}} \Pr_S (\forall_{\mathbf{X} \text{ of rank-}k}\ D(\mathbf{X},\mathbf{Y}) < D_S(\mathbf{X},\mathbf{Y}) + \varepsilon) > 1-\delta$

0/1 loss: $\text{loss}(X_{ij};Y_{ij}) = \text{sign}(X_{ij}Y_{ij})$

$$\varepsilon = \sqrt{\frac{k(n+m)\log\frac{8em}{k} + \log \frac{1}{\delta}}{2|S|}}$$

$\text{loss}(X_{ij};Y_{ij}) \leq 1$:
(by bounding the psudodimension)

$$\varepsilon = 6\sqrt{\frac{k(n+m)\log\frac{8em}{k}\log\frac{|S|}{k(n+m)} + \log \frac{1}{\delta}}{|S|}}$$

Generalization Error Bounds: Large Margin Matrix Factorization

$D(\mathbf{X};\mathbf{Y}) = \sum_{ij} \mathrm{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})/nm$

generalization error

$\mathrm{loss}(X_{ij};Y_{ij}) = \mathrm{sign}(X_{ij}Y_{ij})$

$D_S(\mathbf{X};\mathbf{Y}) = \sum_{ij\in S} \mathrm{loss}^1(\mathbf{X}_{ij};\mathbf{Y}_{ij})/|S|$

empirical error

$\mathrm{loss}^1(X_{ij};Y_{ij}) = \mathrm{sign}(X_{ij}Y_{ij}-1)$

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

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

$(\sum |U_i|^2/n)(\sum |V_i|^2/m) \le R^2$:

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

$(\max |U_i|^2)(\max |V_j|^2) \le R^2$:

$$\varepsilon = 12\sqrt{\frac{R^2(n+m) + \log 1/\delta}{|S|}}$$

Maximum Margin Matrix Factorization as a Convex Combination of Classifiers

$\{ UV \mid (\sum |U_i|^2)(\sum |V_i|^2) \leq 1 \}$

$= \text{convex-hull}(\{ uv' \mid u \in R^n, v \in R^m \ |u|=|v|=1\})$

$\text{conv}(\{ uv' \mid u \in \pm 1^n, v \in \pm 1^m\})$

$\subset \{ UV \mid (\max |U_i|^2)(\max |V_j|^2) \leq 1 \}$

$\subset 2\ \text{conv}(\{ uv' \mid u \in \pm 1^n, v \in \pm 1^m\})$

Grothendiek's Inequality

Summary

- Finding Low Rank Approximations
  - Weighted Low Rank Approximations
  - Basis for other loss function: Newton, Gaussian mixtures
- Consistency of Low Rank Approximation
  - ML for popular low-rank models is not consistent!
  - PCA consistent for additive noise; diagonal ignoring for unbiased
  - Efficient estimators?
  - Consistent estimators for Exponential-PCA?
- Maximum Margin Matrix Factorization
  - Correspondence with large margin linear classification
  - Sparse SDPs for both avarage-norm and max-norm formulations
  - Direct optimization of dual would enable large-scale applications
- Generalization Error Bounds for Collaborative Prediction
  - First “assumption free” bounds for matrix completion
  - Both for Low-Rank and for Max-Margin
  - Observation process?

Average February Temperature

(centigrade)