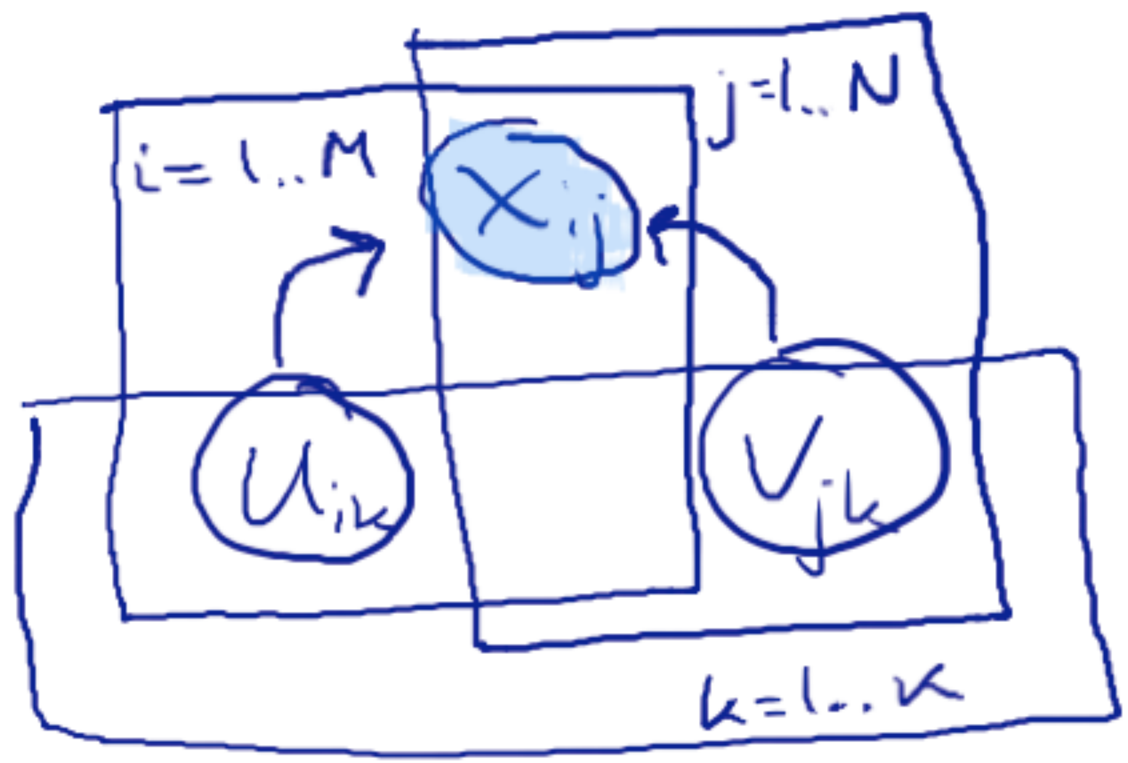
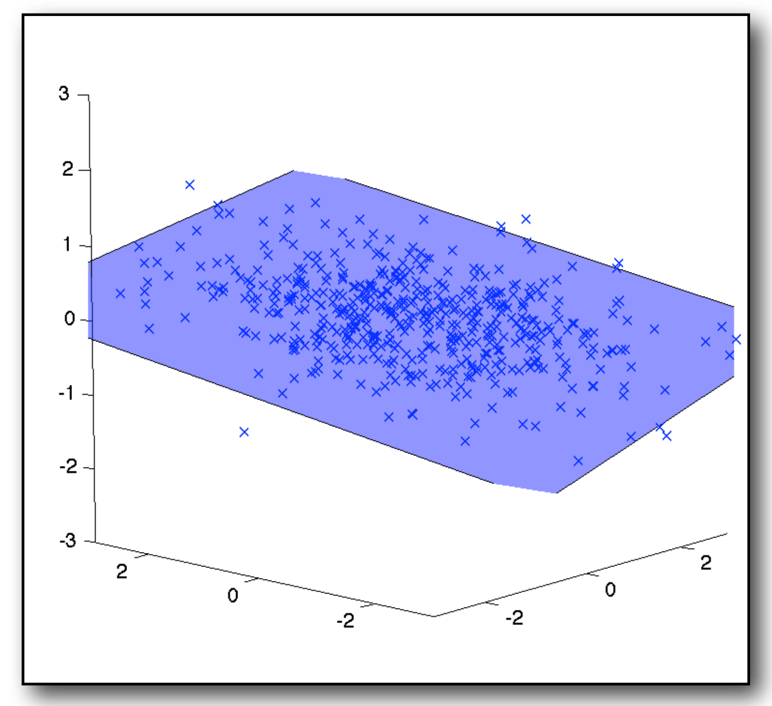
Review

- Gibbs sampling
  - MH with proposal
    - $Q(\mathbf{x} \mid \mathbf{x}') = P(\mathbf{x}_{B(i)} \mid \mathbf{x}_{\neg B(i)})\ I(\mathbf{x}_{\neg B(i)} = \mathbf{x}'_{\neg B(i)})\ /\ \#B$
  - failure mode: "lock-down"
- Relational learning (properties of ***sets*** of entities)
  - document clustering, recommender systems, eigenfaces

Review



- Latent-variable models
- PCA, pPCA, Bayesian PCA
  - everything Gaussian
  - $E(X \mid U,V) = UV^T$
  - MLE: use SVD

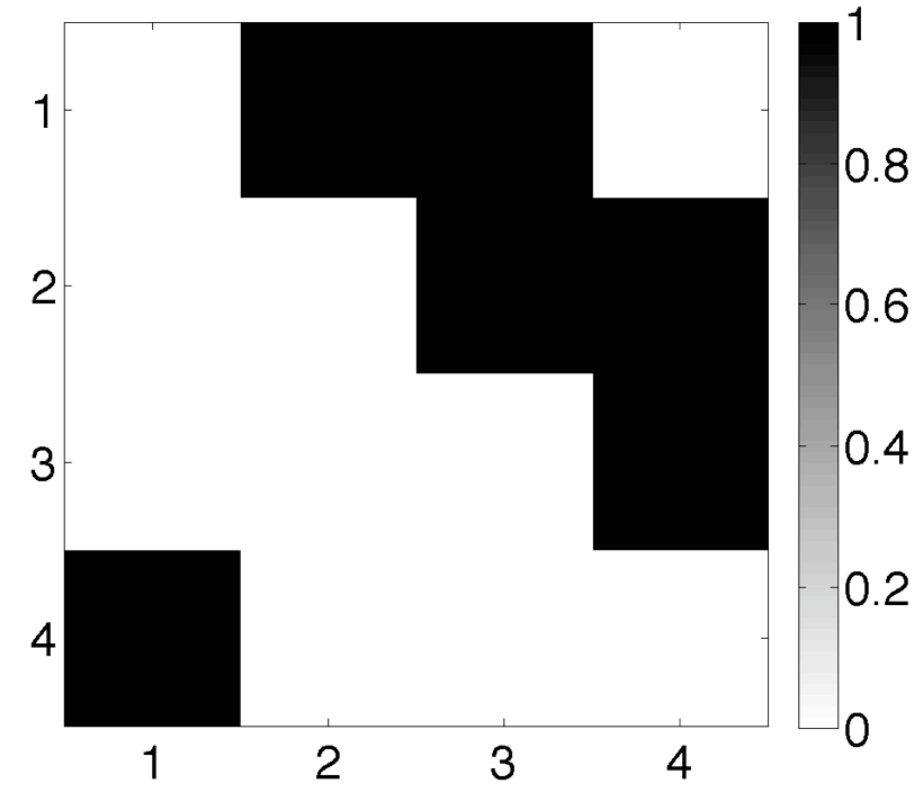


- Mean subtraction, example weights

PageRank

- SVD is pretty useful: turns out to be main computational step in other models too
- A famous one: PageRank
  - Given: web graph (V, E)
  - Predict: which pages are important

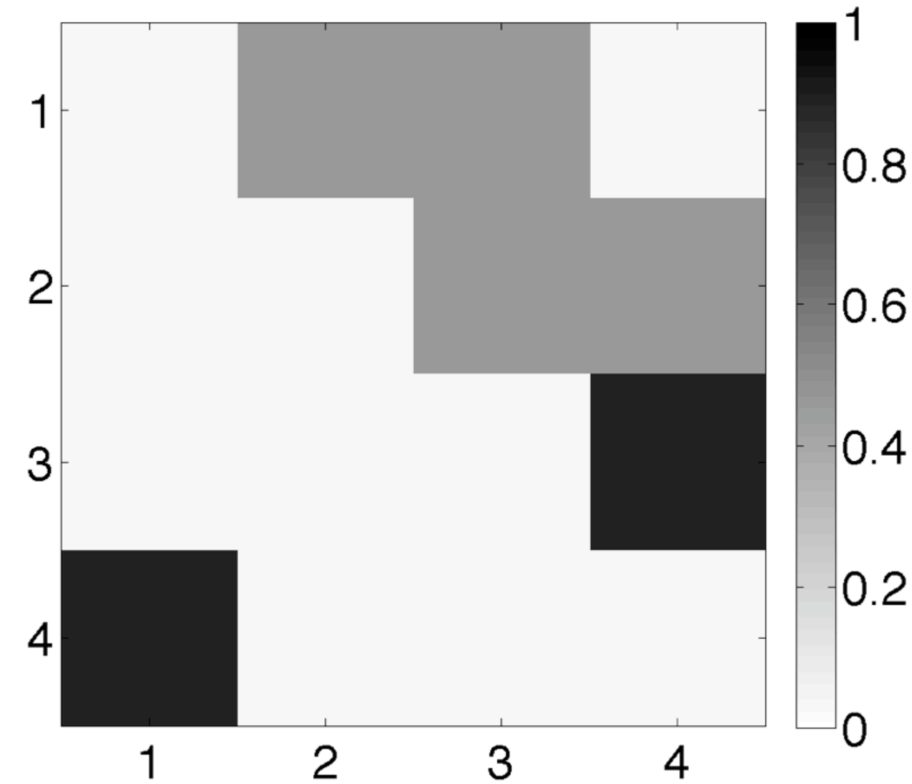
PageRank: adjacency matrix



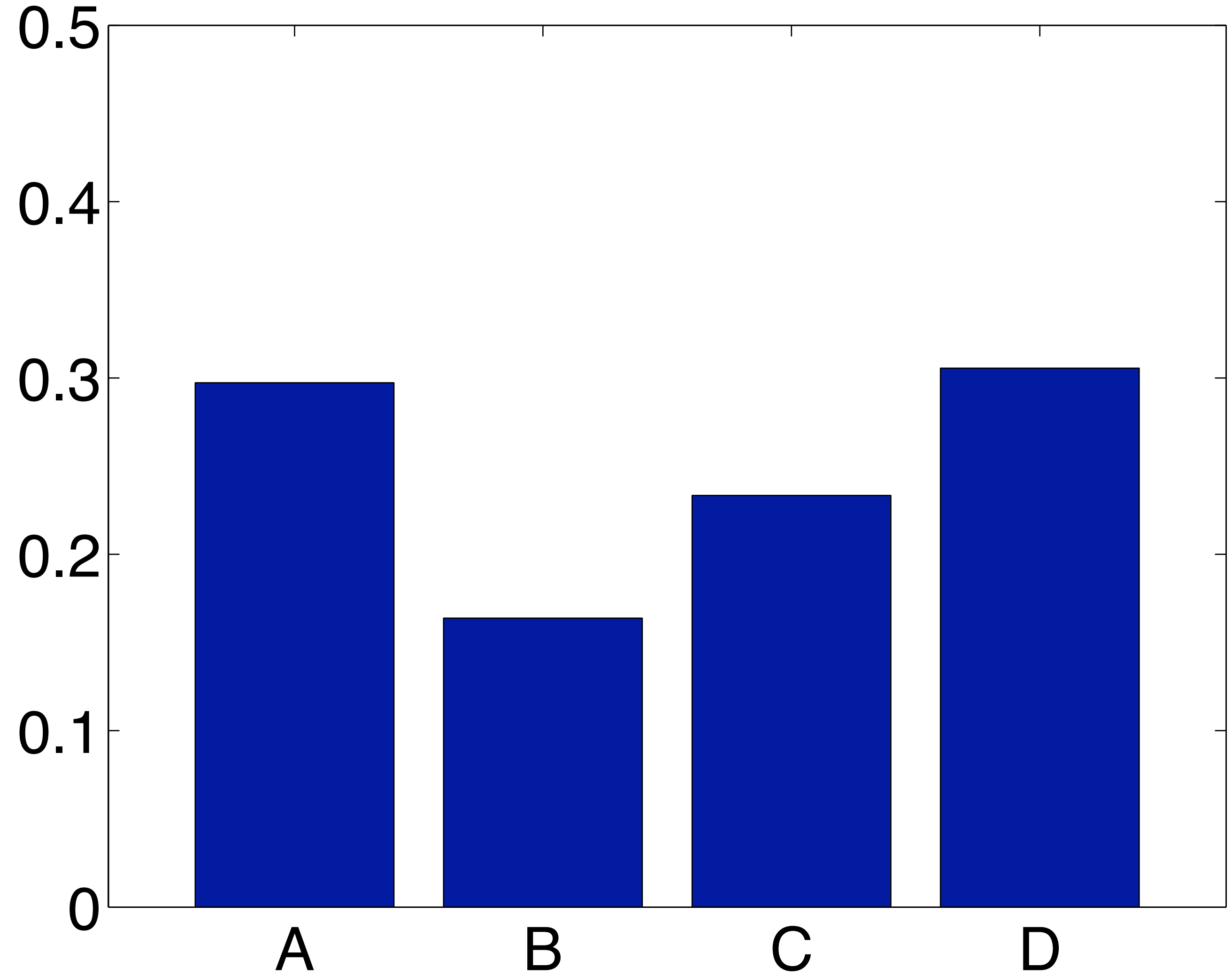
Random surfer model

- W. p. α:
- W. p. (1–α):

- Intuition: page is important if a random surfer is likely to land there



Stationary distribution



Thought experiment

- What if A is symmetric?
  - note: we're going to stop distinguishing A, A'

- So, stationary dist'n for symmetric A is:
- What do people do instead?

Spectral embedding

- Another famous model: spectral embedding (and its cousin, spectral clustering)
- Embedding: assign low-D coordinates to vertices (e.g., web pages) so that similar nodes in graph $\Rightarrow$ nearby coordinates
  - A, B similar = random surfer tends to reach the same places when starting from A or B

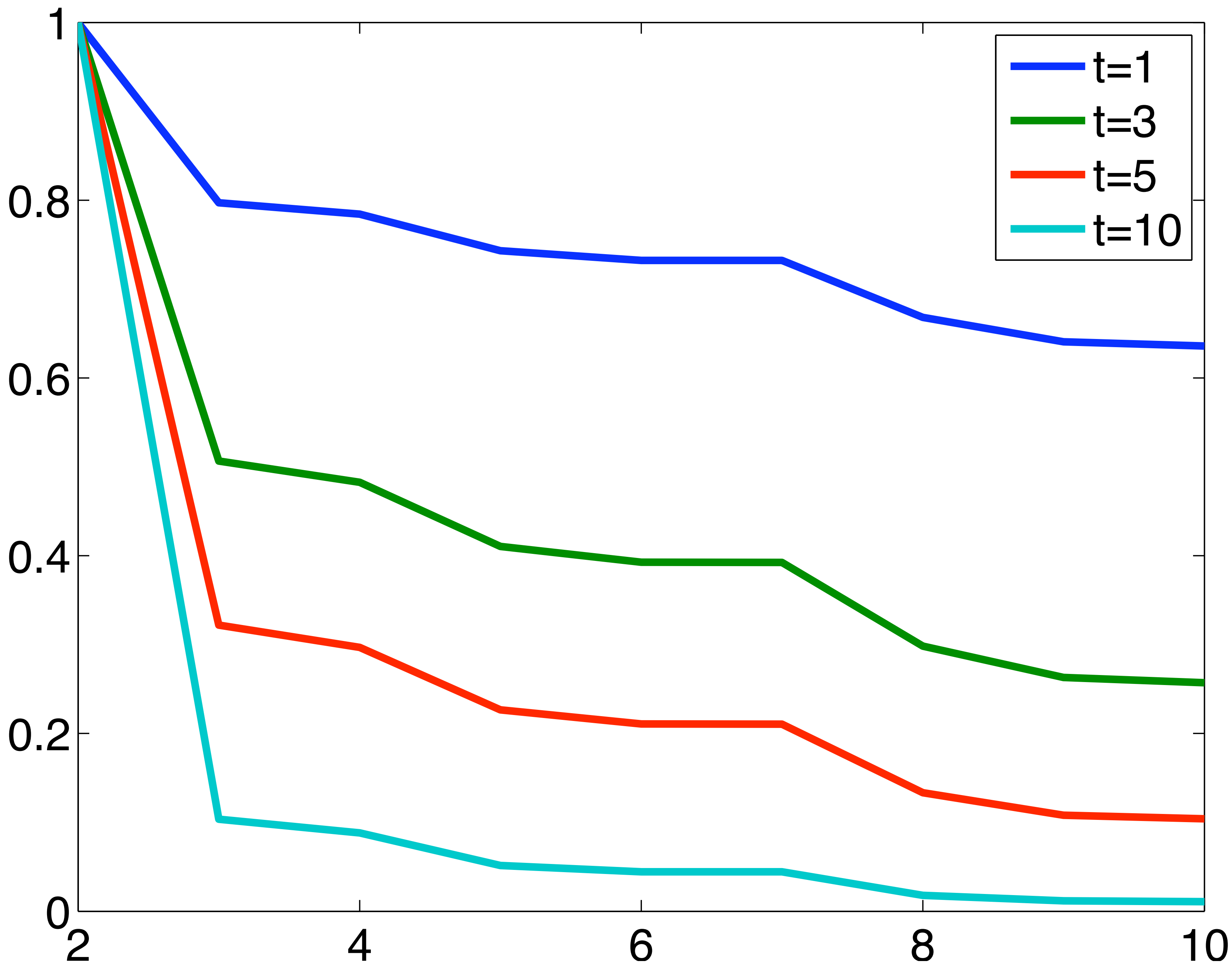
Where does random surfer reach?

- Given graph:
- Start from distribution π
  - after 1 step: P(k | π, 1-step) =
  - after 2 steps: P(k | π, 2-step) =
  - after t steps:

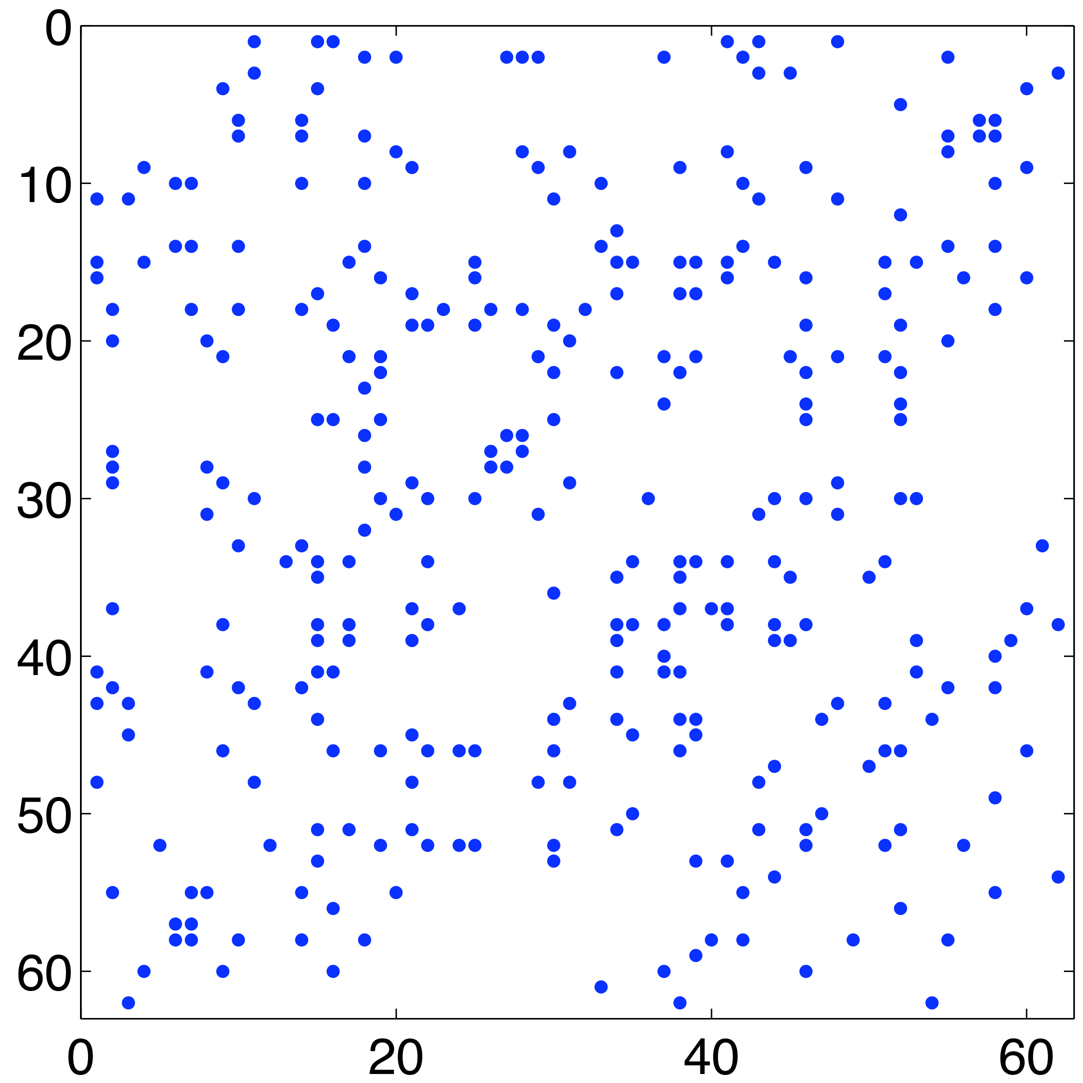
Similarity

- A, B similar = random surfer tends to reach the same places when starting from A or B
- $P(k \mid \pi, \text{t-step})$ =
  - If π has all mass on i:
  - Compare i & j:
  - Role of Σ^t:

Role of Σ^t (real data)



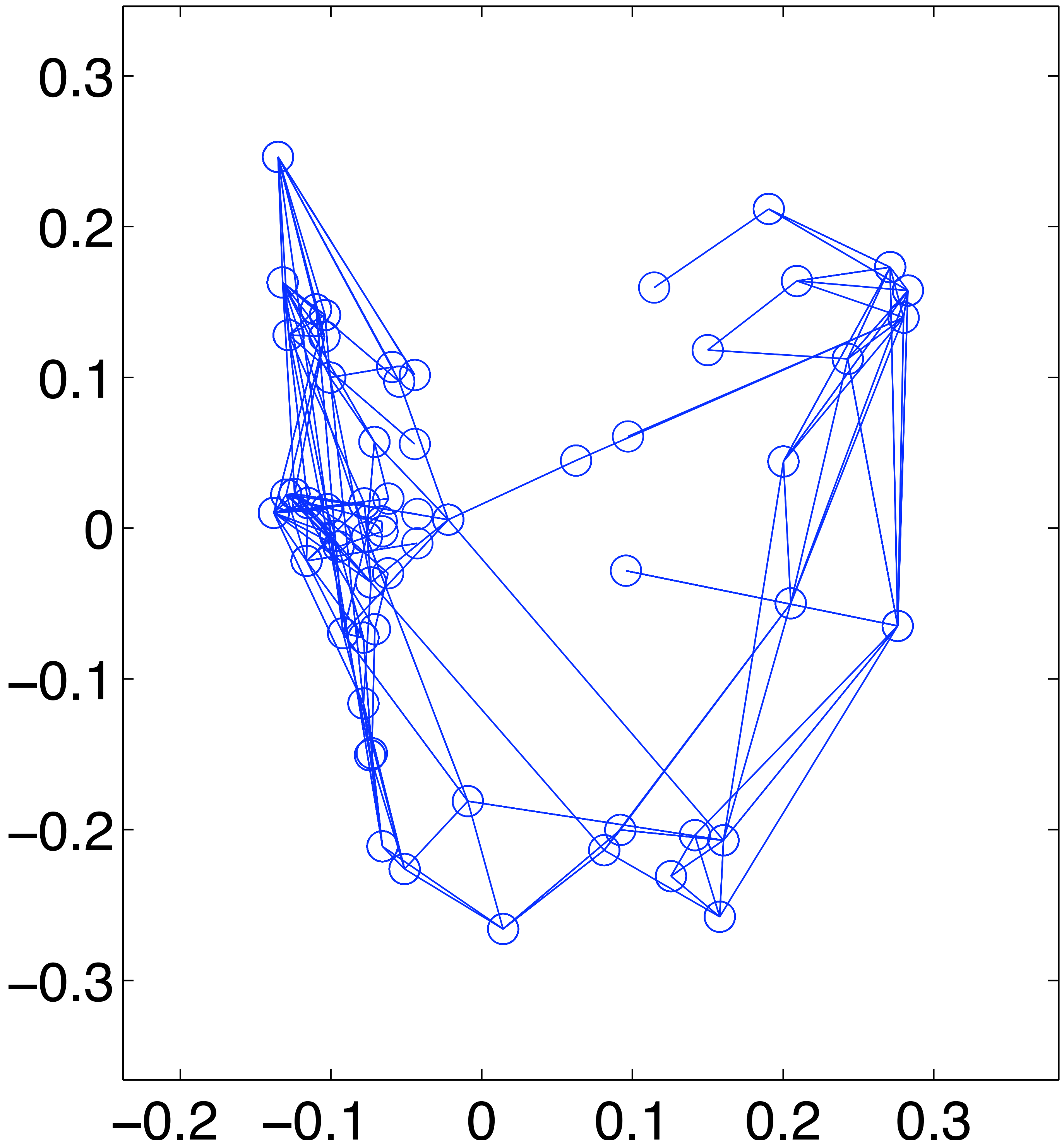
Example: dolphins



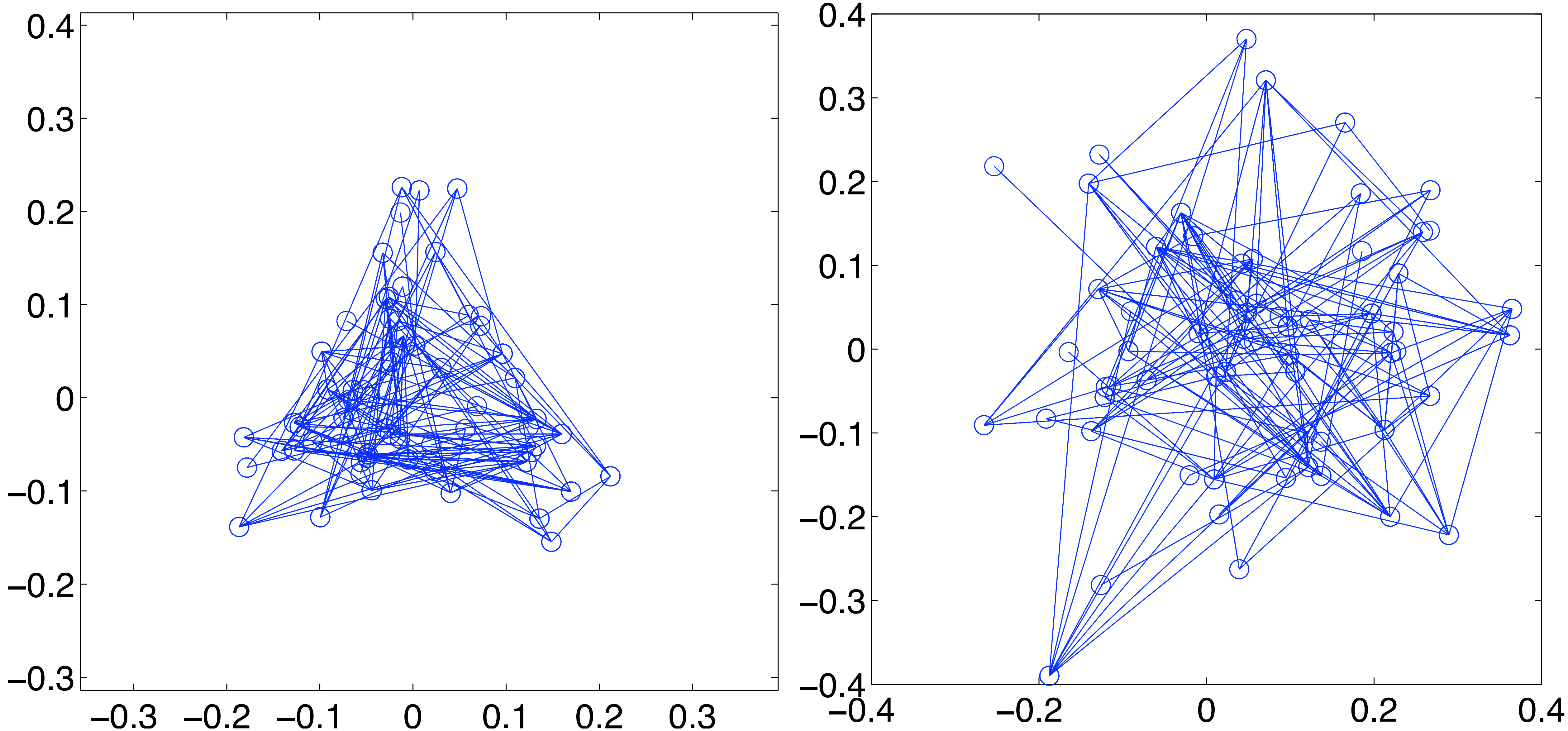
- 62-dolphin social network near Doubtful Sound, New Zealand
  - A_{ij} = 1 if dolphin i friends dolphin j

(Lusseau et al., 2003)

Dolphin network



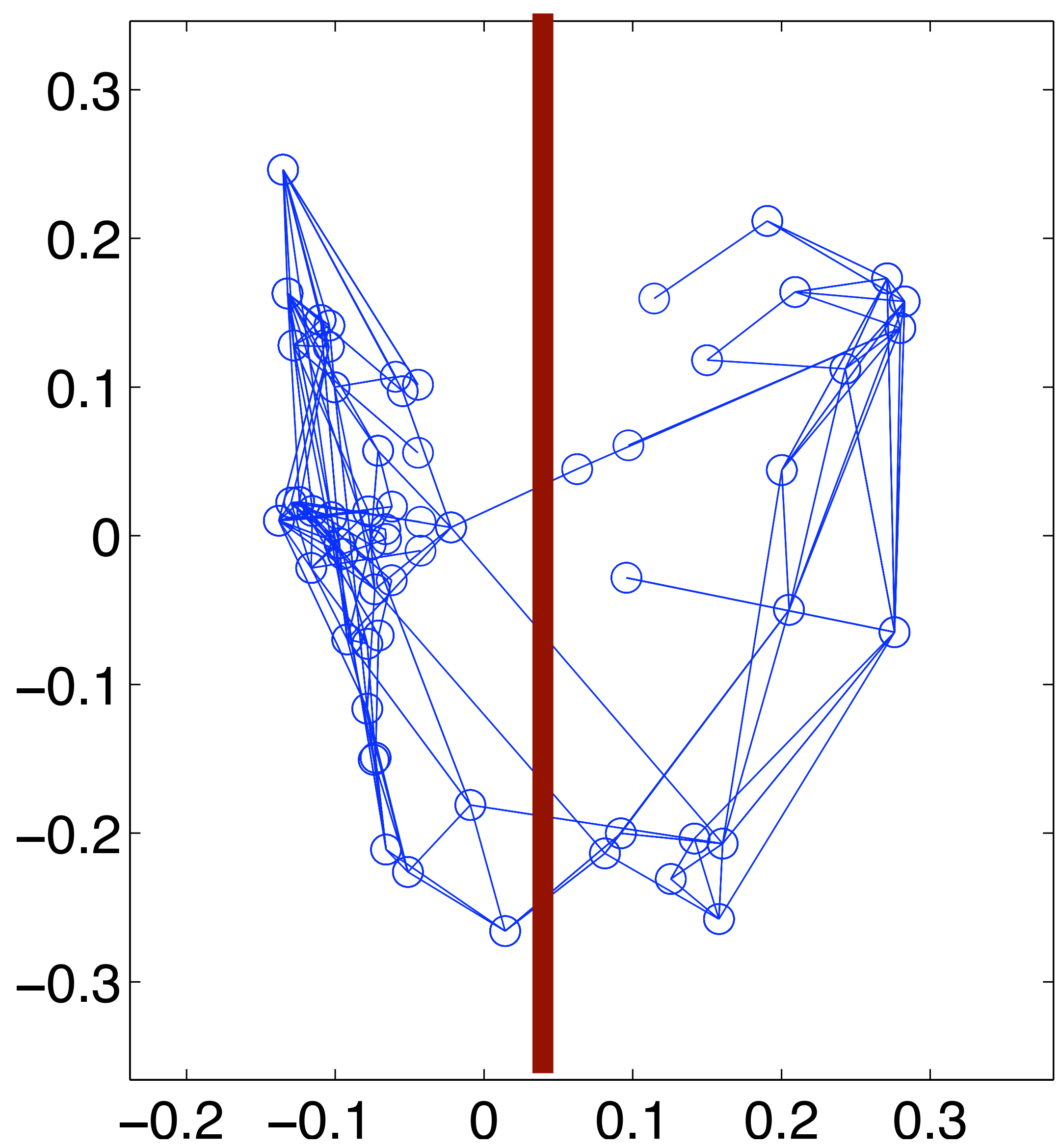
Comparisons



spectral embedding of random data

random embedding of dolphin data

Spectral clustering



- Use your favorite clustering algorithm on coordinates from spectral embedding

PCA: the good, the bad, and the ugly

- The good: simple, successful
- The bad: linear, Gaussian
  - $E(X) = UV^T$
  - X, U, V ~ Gaussian
- The ugly: failure to generalize to new entities

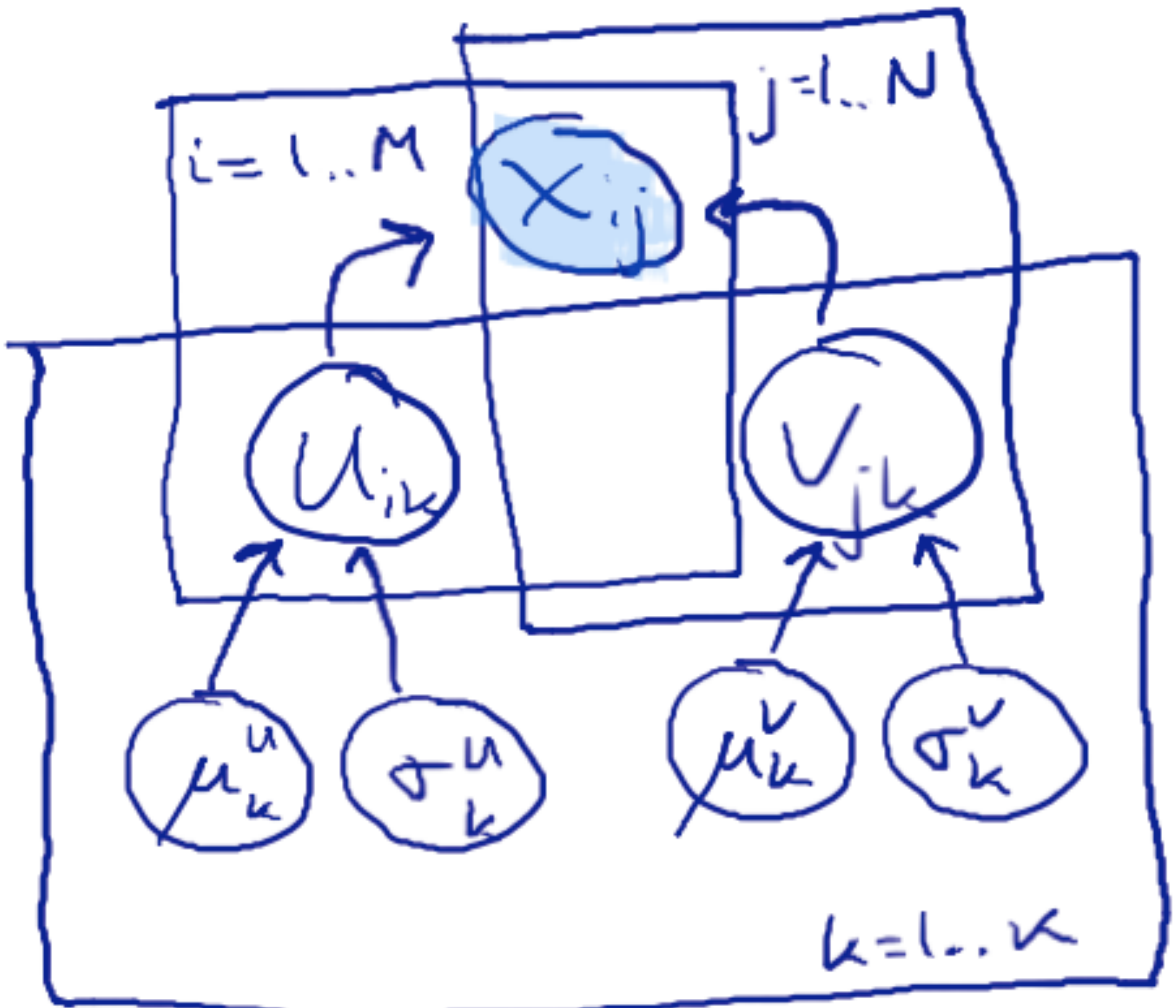
Consistency

- Linear & logistic regression are ***consistent***
- What would consistency mean for PCA?
  - forget about row/col means for now
- Consistency:
  - #users, #movies, #ratings (= nnz(W))
  - numel(U), numel(V)
  - consistency =

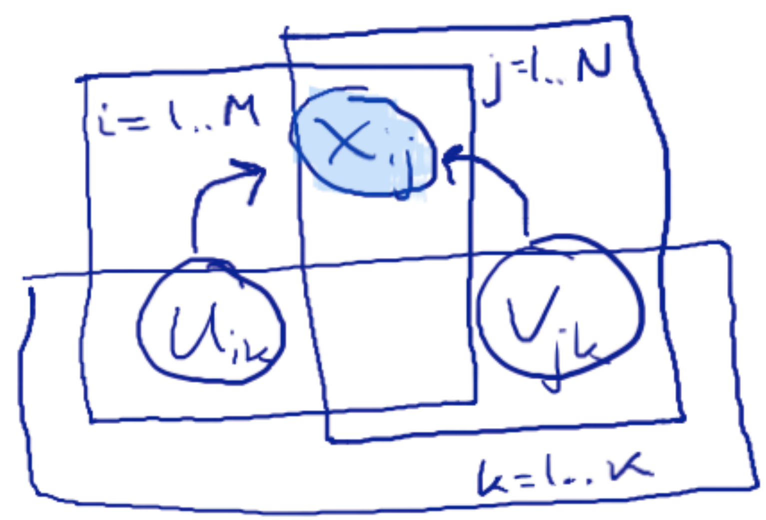
Failure to generalize

- What does this mean for generalization?
  - new user's rating of movie_j: only info is
  - new movie rated by user_i: only info is
  - all our carefully-learned factors give us:
- Generalization is:

Hierarchical model



old, non-hierarchical model



Benefit of hierarchy

- Now: only k μ^U latents, k μ^V latents (and corresponding σs)
  - can get consistency for these if we observe more and more X_{ij}
- For a new user or movie:

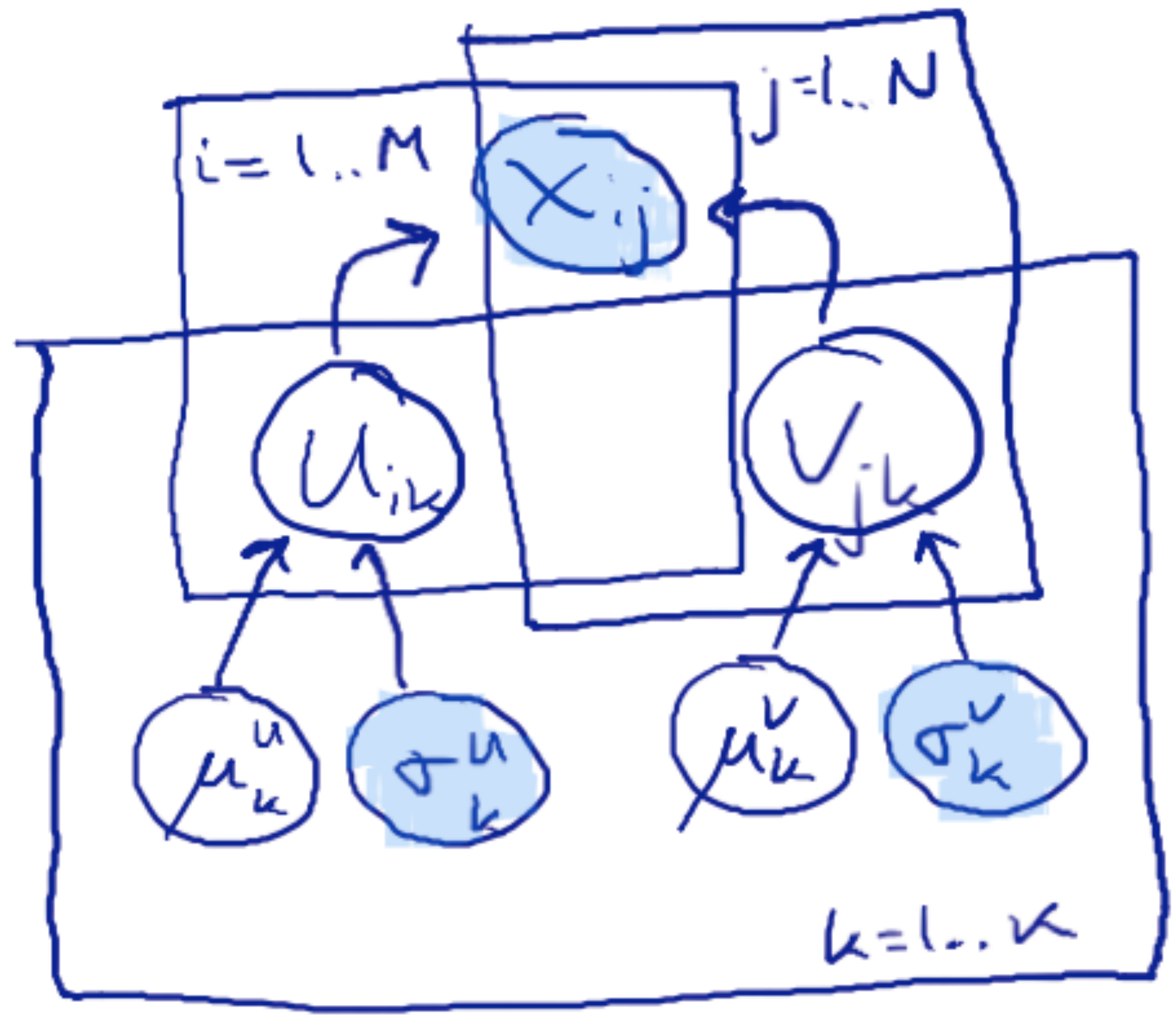
Mean subtraction

- Can now see that mean subtraction is a special case of our hierarchical model
  - Fix $V_{j1} = 1$ for all j; then U_{i1} =
  - Fix $U_{i2} = 1$ for all i; then V_{j2} =
  - global mean:

What about the second rating for a new user?

- Estimating U_i from one rating:
  - knowing μ^U:
  - result:
- How should we fix?
- Note: often we have only a few ratings per user

MCMC for PCA



- Can do Bayesian inference by Gibbs sampling—for simplicity, assume σs known

Recognizing a Gaussian

- Suppose $X \sim N(X \mid \mu, \sigma^2)$
- $L = -\log P(X=x \mid \mu, \sigma^2) =$
  - $dL/dx =$
  - $d^2L/dx^2 =$
- So: if we see $d^2L/dx^2 = a$, $dL/dx = a(x - b)$
  - $\mu =$ $\sigma^2 =$

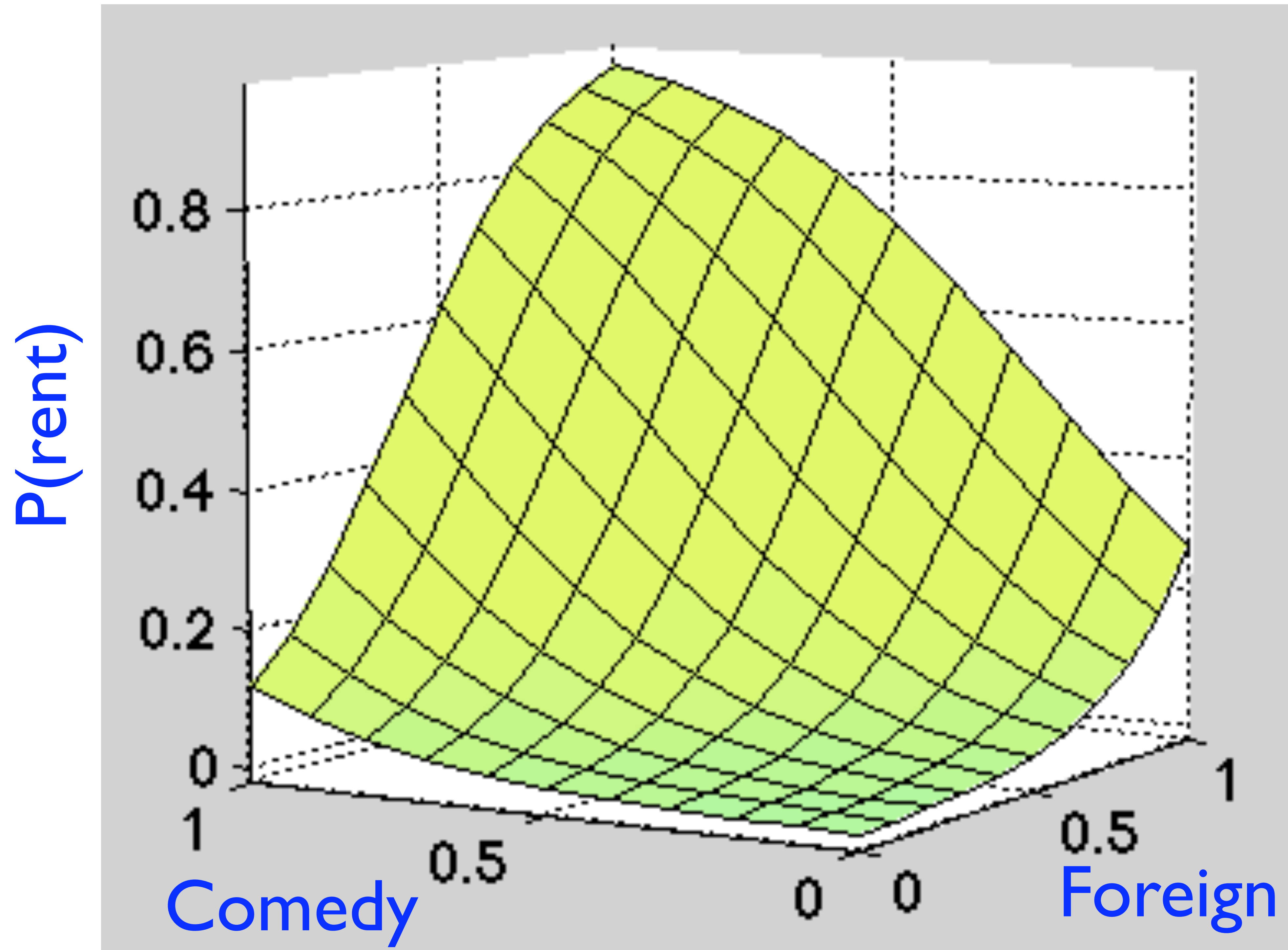
Gibbs step for an element of μ^U

Gibbs step for an element of U

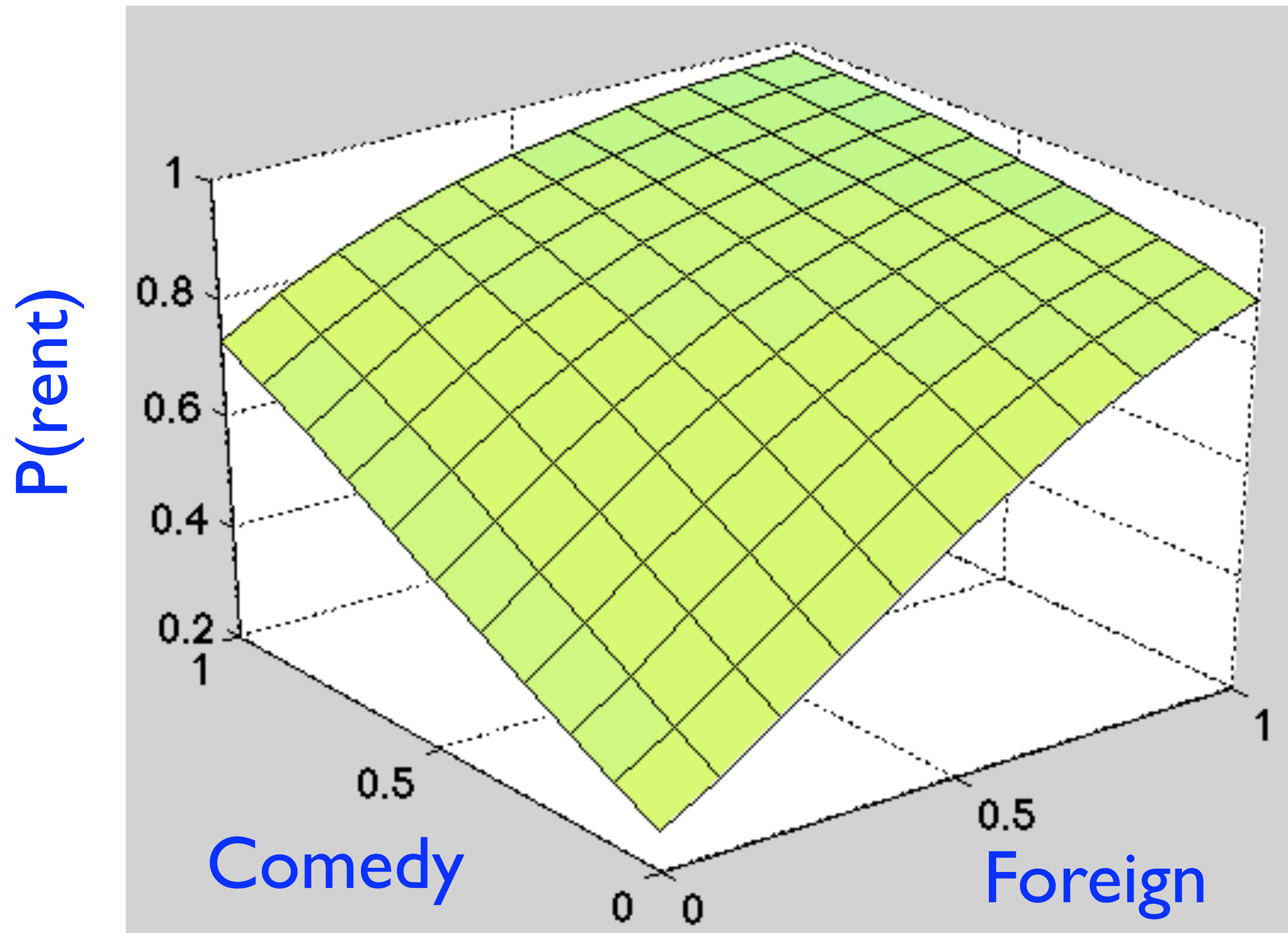
In reality

- We'd do ***blocked Gibbs*** instead
- Blocks contain entire rows of U or V
  - take gradient, Hessian to get mean, covariance
  - formulas look a lot like linear regression (normal equations)
- And, we'd fit σ^U, σ^V too
  - sample $1/\sigma^2$ from a ***Gamma*** (or Σ^{-1} from a ***Wishart***) distribution

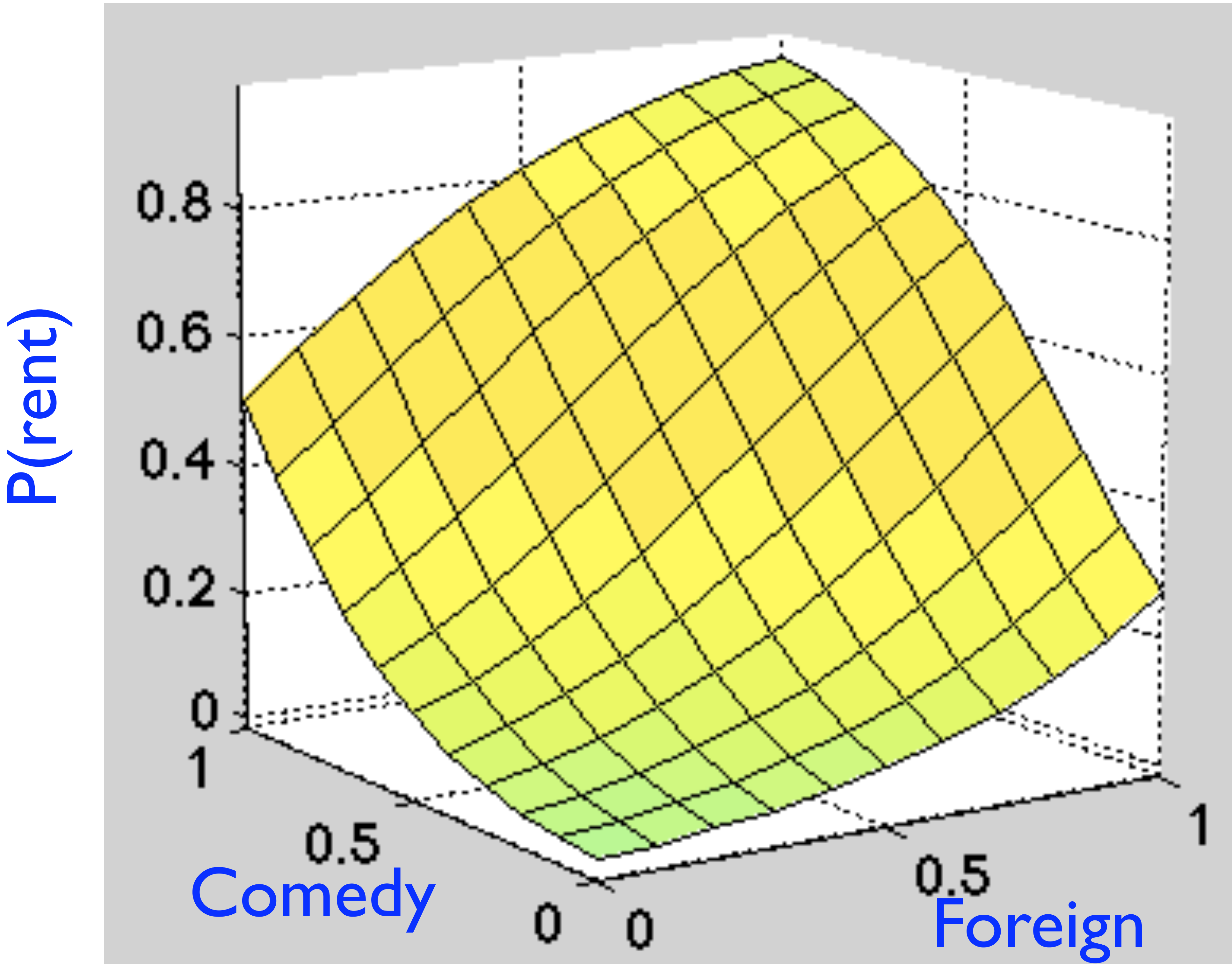
Nonlinearity: conjunctive features



Disjunctive features



“Other”



Non-Gaussian

- X, U, and V could each be non-Gaussian
  - e.g., binary!
    - rents(U, M), comedy(M), female(U)
- For X: predicting –0.1 instead of 0 is only as bad as predicting +0.1 instead of 0
- For U, V: might infer –17% comedy or 32% female

Logistic PCA

- Regular PCA: $X_{ij} \sim N(U_i \cdot V_j, \sigma^2)$
- Logistic PCA:

More generally…

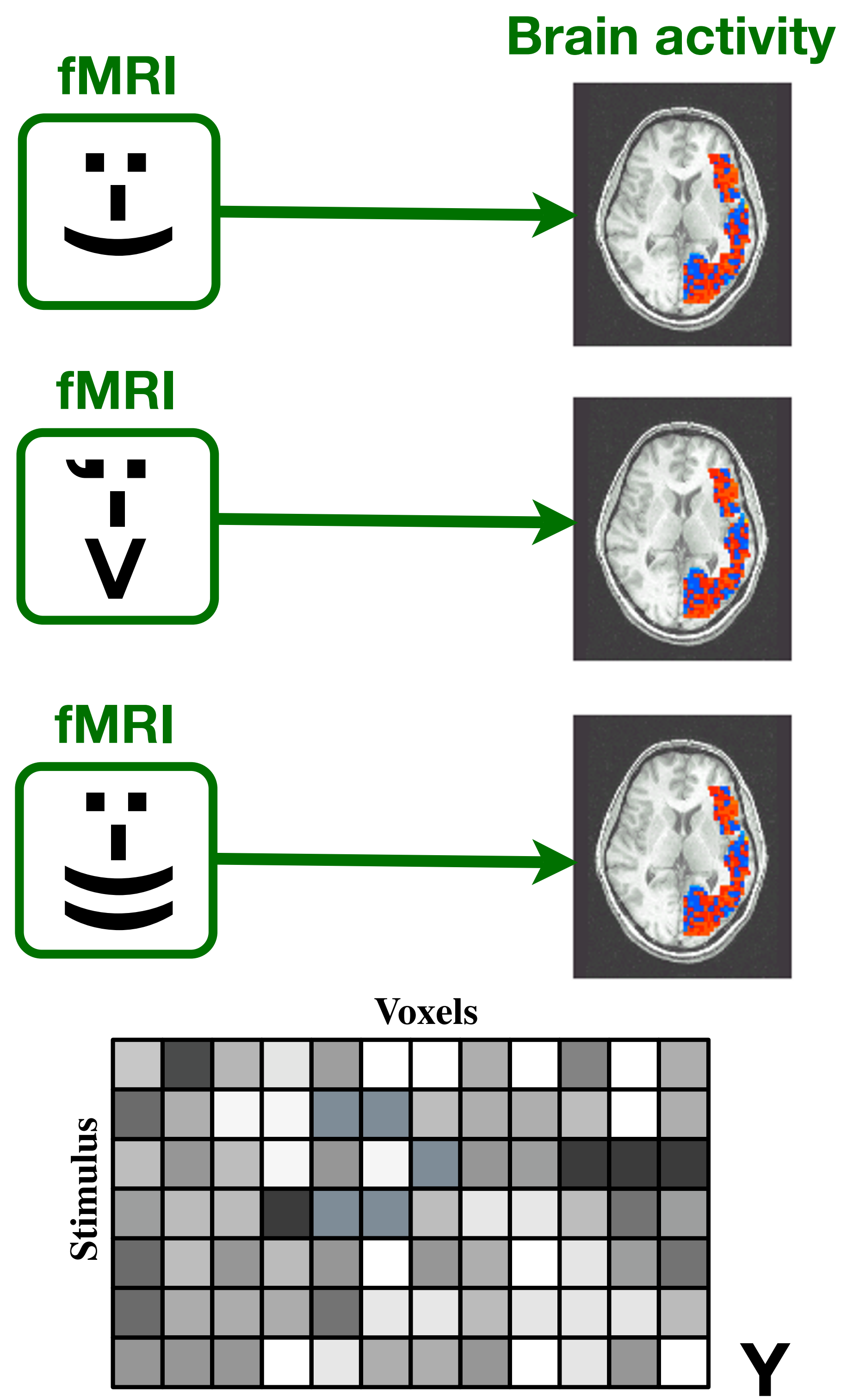
- Can have
  - $X_{ij} \sim \text{Poisson}(\mu_{ij}), \mu_{ij} = \exp(U_i \cdot V_j)$
  - $X_{ij} \sim \text{Bernoulli}(\mu_{ij}), \mu_{ij} = \sigma(U_i \cdot V_j)$
  - …
- Called ***exponential family PCA***
- Might expect optimization to be difficult

Application: fMRI

stimulus: “dog”

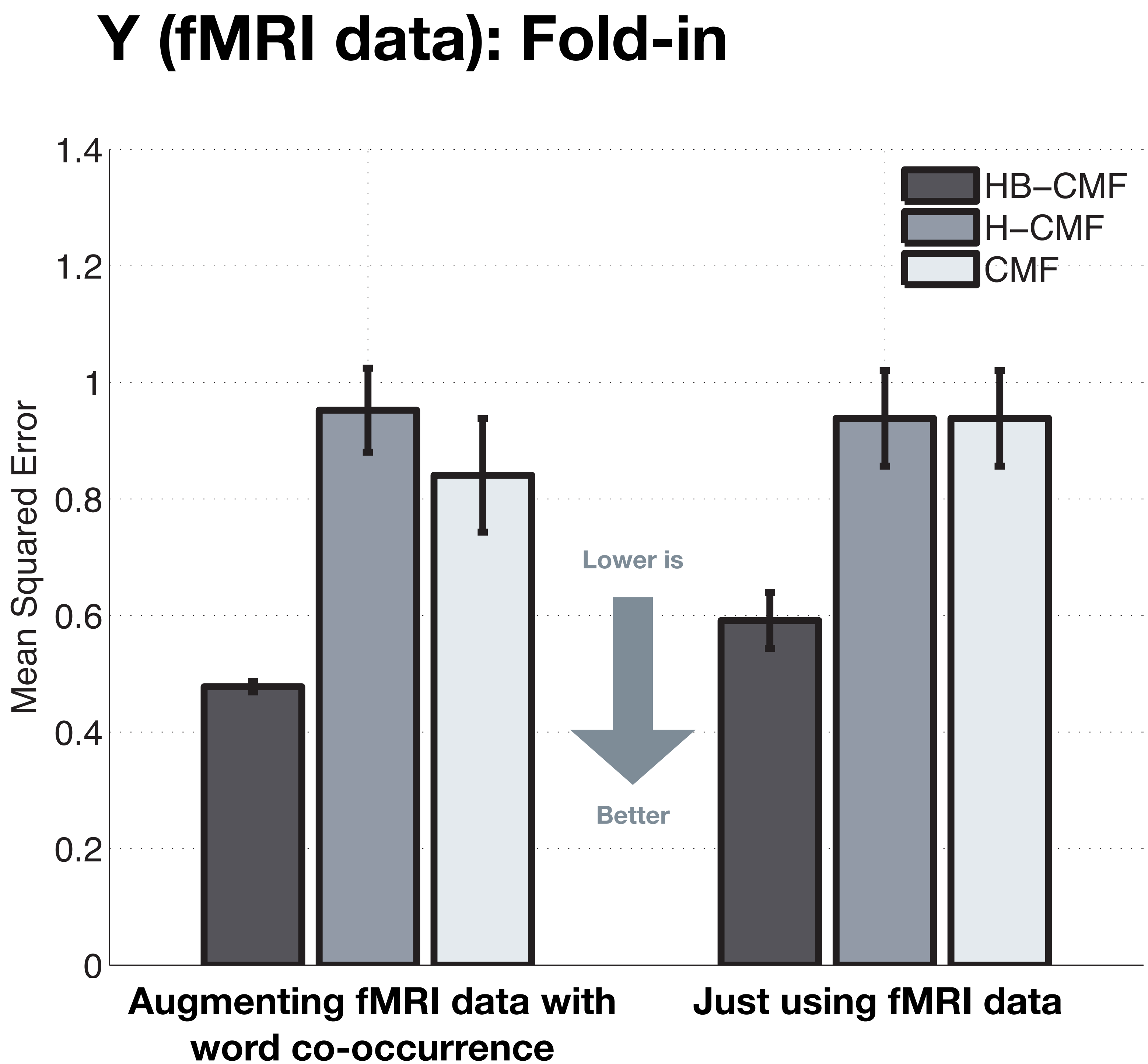
stimulus: “cat”

stimulus: “hammer”



credit: Ajit Singh

Results (logistic PCA)



credit: Ajit Singh