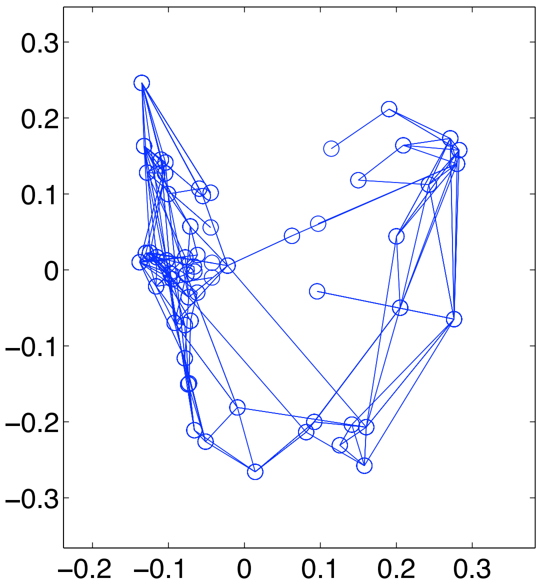
Review

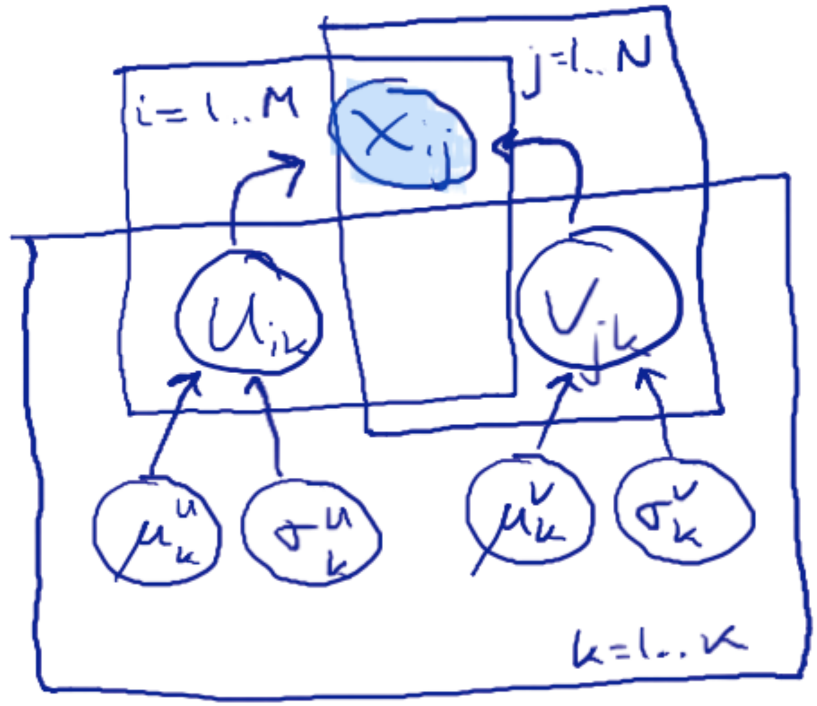
- Models that use SVD or eigen-analysis
  - PageRank: eigen-analysis of ***random surfer*** transition matrix
    - usually uses only *first* eigenvector
  - Spectral embedding: eigen-analysis (or equivalently SVD) of random surfer model in ***symmetric*** graph
    - usually uses 2nd–Kth EVs (small K)
    - first EV is boring
  - Spectral clustering = spectral embedding followed by clustering

dolphin friendships



Review: PCA

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

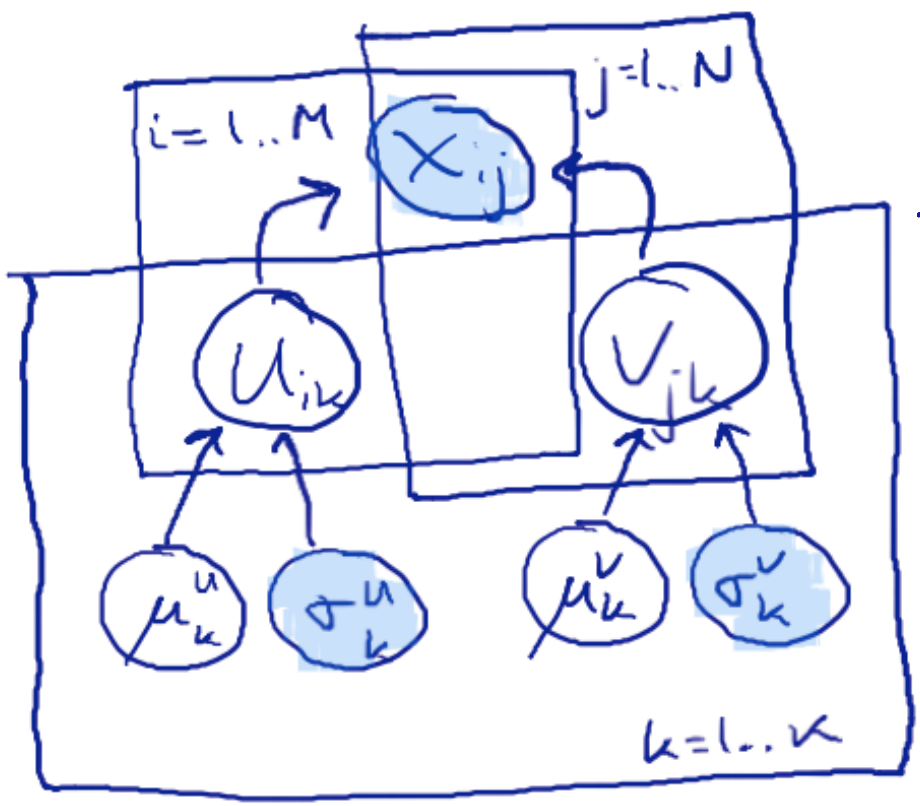


What about the second rating for a new user?

- MLE/MAP of $U_{i\cdot}$ from one rating: overfit
    - knowing μ^U: helps — but doesn't fix
    - result: confidently wrong
- How should we fix? be Bayesian
- Note: often have only a few ratings per user

so, significant posterior uncertainty $\Rightarrow$ Bayes is important no matter how much data we have

MCMC for PCA



$X_{\cdot j}$

$X_{i\cdot}$

Need:

$$P(U_{ik} \mid V, \mu^u_k, \text{rest of } U_{i\cdot})$$

$$P(V_{jk} \mid U, \mu^v_k, \text{rest of } V_{j\cdot})$$

$$P(\mu^u_k \mid U_{\cdot k})$$

$$P(\mu^v_k \mid V_{\cdot k})$$

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

Recognizing a Gaussian

- Suppose $X \sim N(X \mid \mu, \sigma^2)$
- $L = -\log P(X{=}x \mid \mu, \sigma^2) =$ $\log \sqrt{2\pi}\,\sigma + \frac{1}{2\sigma^2}(x-\mu)^2$
  - $dL/dx =$ $\frac{1}{\sigma^2}(x-\mu)$
  - $d^2L/dx^2 =$ $1/\sigma^2$
- So: if we see $d^2L/dx^2 = a$, $dL/dx = a(x - b)$
  - $\mu =$ b $\qquad \sigma^2 =$ $1/a$

Gibbs step for an element of ~~μU~~ U_{ik}

$$\frac{dL}{dU_{ik}} = p^2 U_{ik} - \sum_j \frac{V_{jk}}{\sigma^2} E_{ij} - \frac{\mu^u_k}{(\sigma^u_k)^2}$$

$$E_{ij} = X_{ij} - \sum_{k' \neq k} U_{ik'} V_{jk'}$$

(*)

- L = $\text{const.} + \sum_{ij} \left(X_{ij} - \sum_{k'} U_{ik'} V_{jk'}\right)^2 / 2\sigma^2$

$$+ \sum_{ik} (U_{ik} - \mu^u_k)^2 / 2(\sigma^u_k)^2 + \sum_{jk} (V_{jk} - \mu^v_k)^2 / 2(\sigma^v_k)^2$$

see (*)

$$\frac{dL}{dU_{ik}} = \sum_j \left(X_{ij} - \sum_{k'} U_{ik'} V_{jk'}\right) / \sigma^2 \, (-V_{jk}) + (U_{ik} - \mu^u_k) / (\sigma^u_k)^2 = p^2 U_{ik} - \sum_j \frac{V_{jk}}{\sigma^2} \underbrace{\left(X_{ij} - \sum_{k' \neq k} U_{ik'} V_{jk'}\right)}_{E_{ij}} - \mu^u_k / (\sigma^u_k)^2$$

$$\frac{d^2 L}{dU_{ik}^2} = V_{jk}^2 / \sigma^2 + \frac{1}{(\sigma^u_k)^2} = p^2$$

post. var. (of U_{ik}) $= 1/p^2$ post. mean $\dfrac{\sum_j \frac{V_{jk}}{\sigma^2} E_{ij} + \frac{1}{(\sigma^u_k)^2} \mu^u_k}{p^2}$

Gibbs: element of ~~U~~ μ^u_k

- L = $\text{const.} + \sum_{ij}\left(X_{ij} - \sum_{k'} U_{ik'} V_{jk'}\right)^2 / 2\sigma^2$

$$+ \sum_{ik} \left(U_{ik} - \mu^u_k\right)^2 / 2(\sigma^u_k)^2 + \sum_{jk} \left(V_{jk} - \mu^v_k\right)^2 / 2(\sigma^v_k)^2$$

- ~~$dL / dU_{ik} =$~~ $dL / d\mu^u_k = \sum_i \left(U_{ik} - \mu^u_k\right) / {\sigma^u_k}^2 \, (-1)$

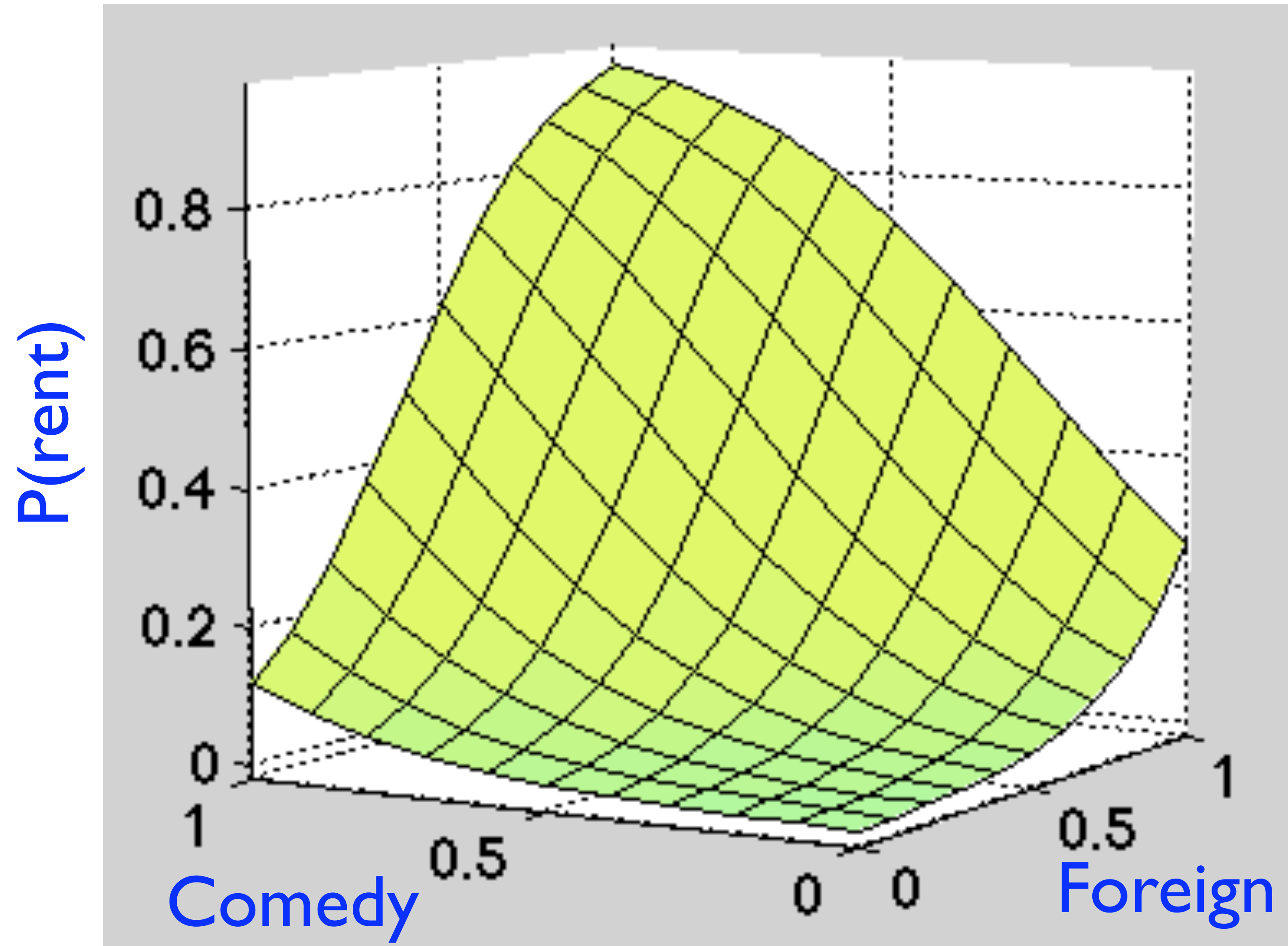
$$d^2 L / d{\mu^u_k}^2 = M / {\sigma^u_k}^2$$

- ~~$dL^2 / (dU_{ik})^2 =$~~
  - post. mean = $\dfrac{\sum_i U_{ik}}{M}$ post. var. = $\dfrac{{\sigma^u_k}^2}{M}$

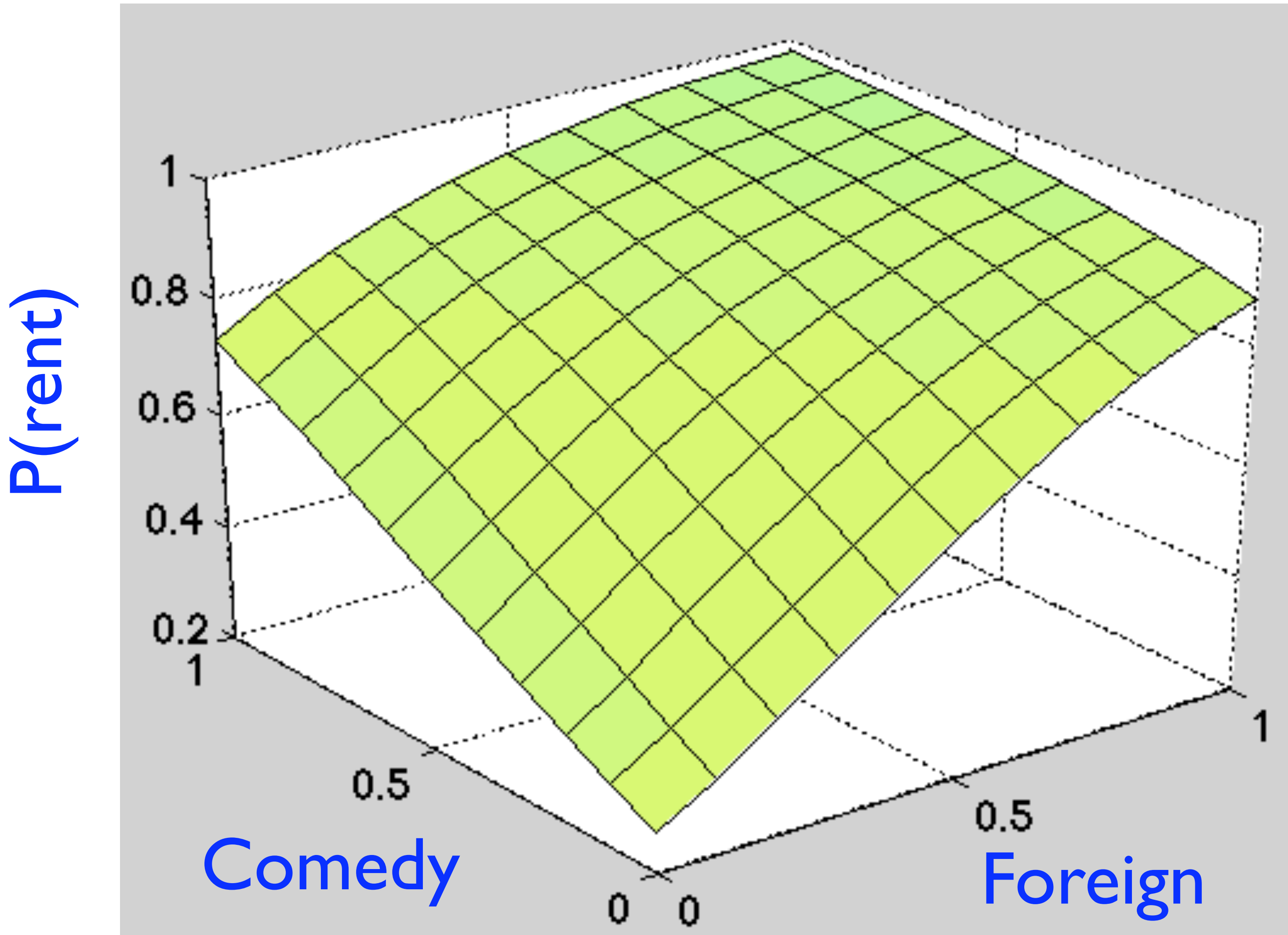
In reality

- Above, blocks are single elements of U or V
- Better: blocks are entire rows of U or V
  - take gradient, Hessian to get mean, covariance
  - formulas look a lot like linear regression (normal equations)
- And, want to 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

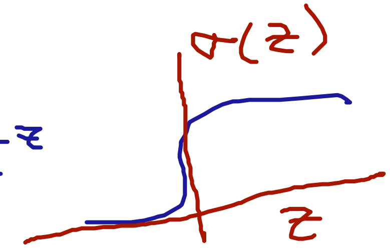


Non-Gaussian

- 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

$\sigma(z) = 1/(1+e^{-z})$



$U_i = U_{i\cdot}$: i^{th} row

$V_j = V_{j\cdot}$: j^{th} row

- Regular PCA: $X_{ij} \sim N(U_i \cdot V_j, \sigma^2)$
- Logistic PCA: $P(X_{ij} = 1 \mid U, V) = \sigma(U_i \cdot V_j)$

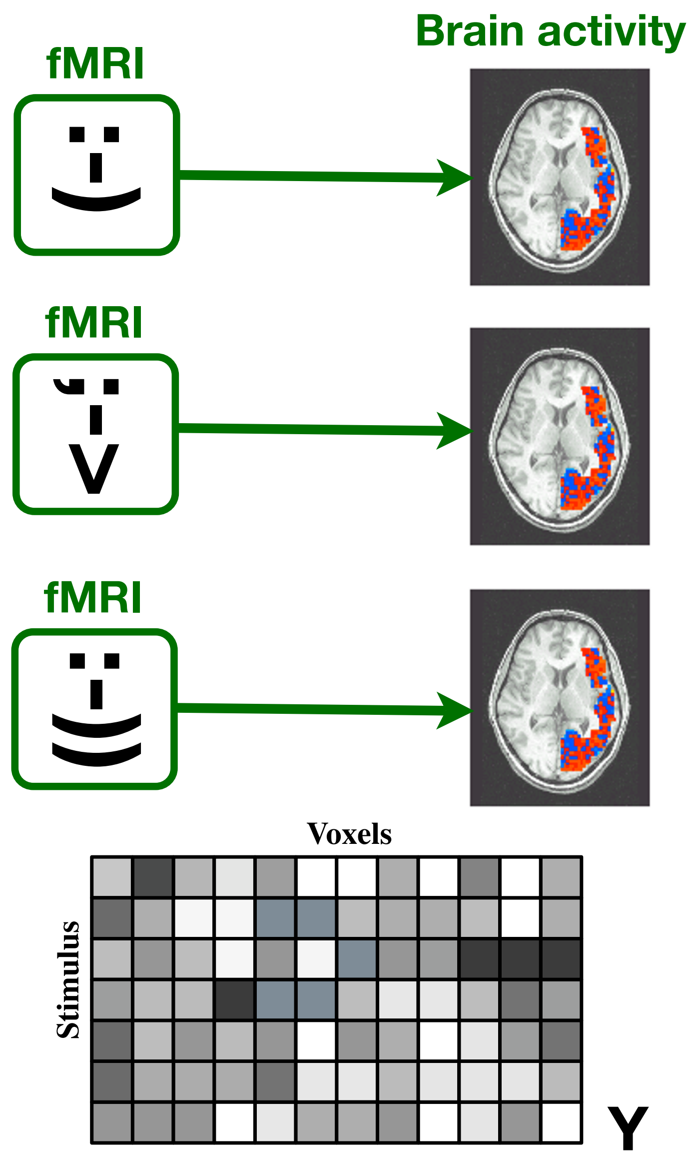
- Might expect learning, inference to be hard
  - but, MH works well, using $dL/d\theta$, $d^2L/d\theta^2$
- Generalization: ***exponential family PCA***
  - w/ optional hierarchy, Bayesianism

Application: fMRI

stimulus: "dog"

stimulus: "cat"

stimulus: "hammer"



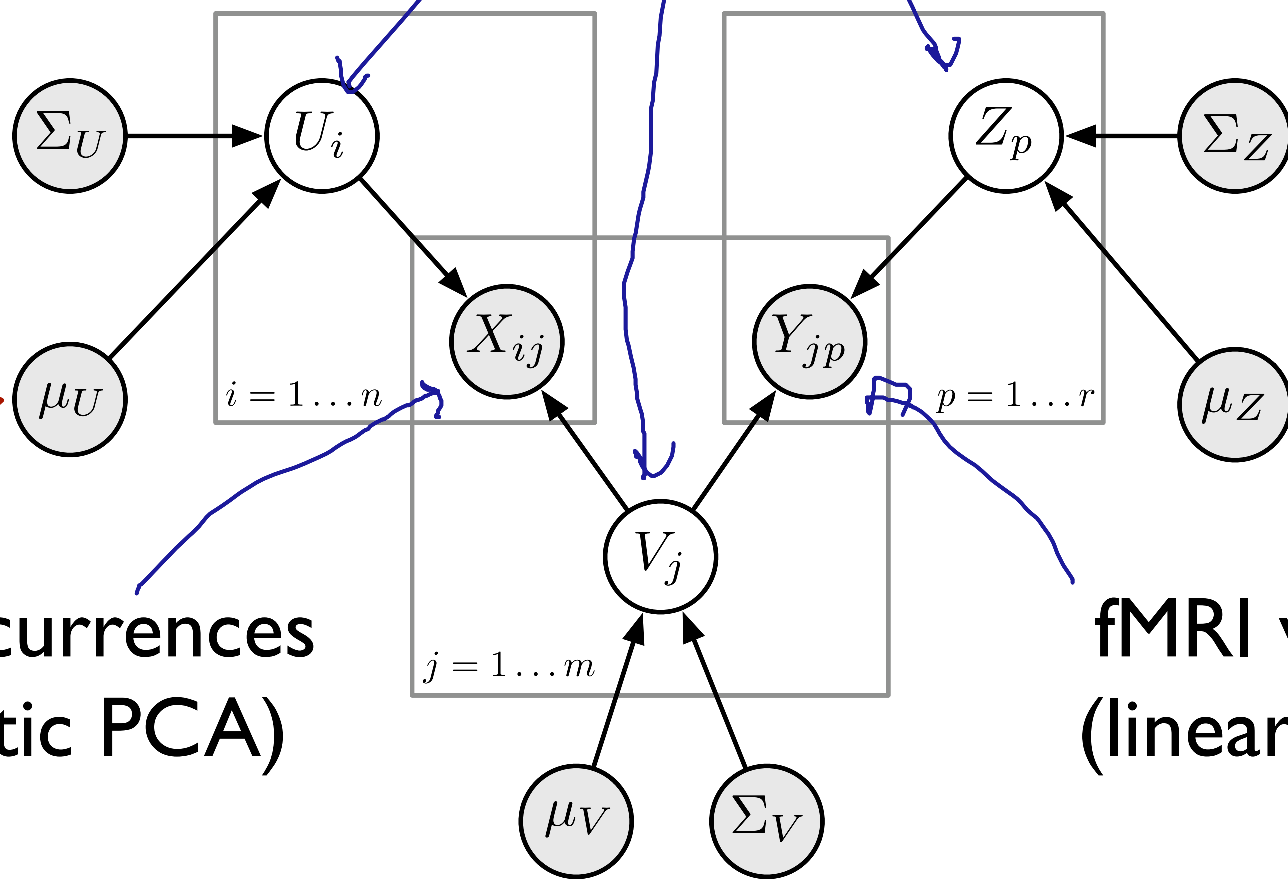
credit: Ajit Singh

2-matrix model

vector (length k)

k x k matrix

k-vector



co-occurrences
(logistic PCA)

fMRI voxels
(linear PCA)

linear PCA &

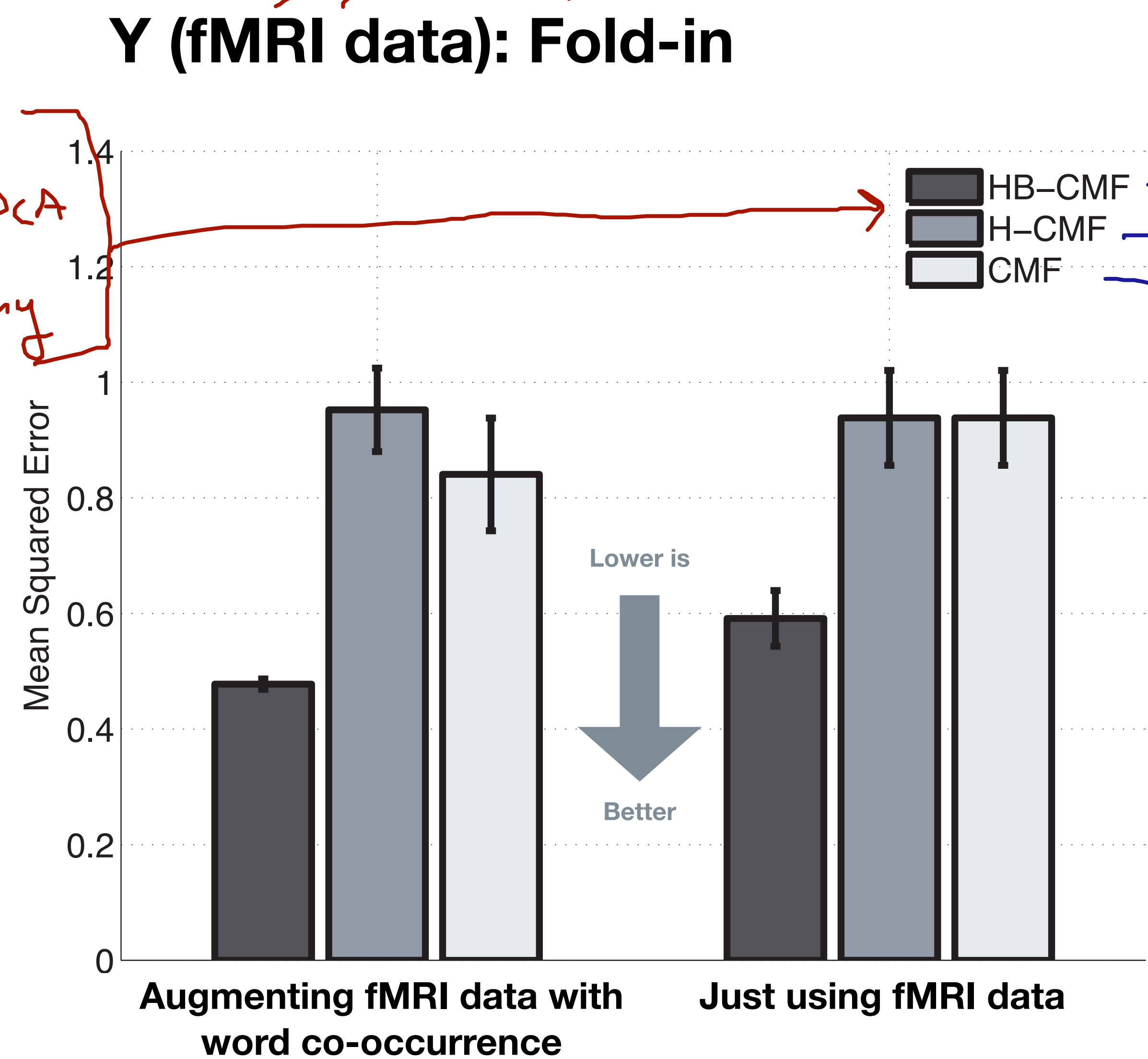
Results (logistic PCA)

linear PCA side of model

CMF = 2-matrix PCA

H = hierarchy

B = Bayes



MH

MAP

MLE (no hierarchy)

credit: Ajit Singh