Multiple Cause Vector Quantization

David Ross & Richard Zemel

University of Toronto

October 22, 2002

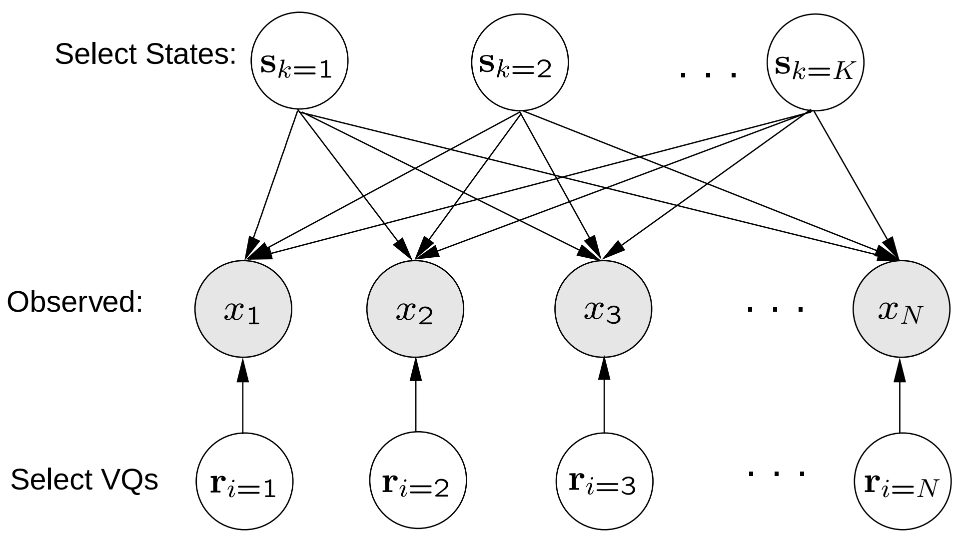
Factorial Learning

- data was generated by the actions of a (small) number of independent unobserved variables
- Eg. 1: pixels of a natural image
  $\longrightarrow$ which objects are present, where they are located in the scene
- Eg. 2: an individual's ratings for various movies
  $\longrightarrow$ genre, which actors are present
- Goal: learn a model that captures these underlying causes, infer the values of the unobserved variables for a new example

Learning a Composite Sketch

- Goal: learn a parts-based representation of data vectors
- Motivating Assumptions:
  1. data dimensions separable into disjoint subsets *(Multiple Causes)*
  2. each cause has a small number of discrete states *(Vector Quantizer)*
  3. causes take on states independently of each other
- Example: on face image data,
  causes could be *eyes*, *nose*, and *mouth*
  states could be different appearances of each part
- Win: combinatorial power
  - VQ with N states represents N items
  - MCVQ with j states per N/j VQs represents $j^{N/j}$ items

Generating an Example $\mathbf{x}$



1. select one state of each VQ k
   $s_{jk} = 1 \Leftrightarrow$ state j of VQ k is active
2. select one VQ for each data dim. i
   $r_{ik} = 1 \;\Leftrightarrow\;$ VQ k relevant for x_i
3. value of x_i depends on params of selected state of selected VQ

Learning & Inference

- $\mathbf{x} \in \mathbb{R}^N$ data vector
- $\mathbf{R} = \{\mathbf{r}_i\}$ K-dim. indicator vectors, select one VQ per data dimension
- $\mathbf{S} = \{\mathbf{s}_k\}$ J-dim. indicator vectors, select one state per VQ
- $\theta = \{\mu_{ijk}, \sigma_{ijk}\}$ parameters of dimension i, from j^{th} state of k^{th} VQ
- $\mathbf{a}_i$'s and $\mathbf{b}_k$'s prior distribution over $\mathbf{r}$'s and $\mathbf{s}$'s

$$
\begin{aligned}
P(\mathbf{x}, R, S|\theta) &= P(R|\theta)P(S|\theta)P(X|R, S, \theta) \\
&= \prod_{i,k,j \in k} a_{ik}^{r_{ik}} \quad b_{jk}^{s_{jk}} \quad \mathcal{N}(x_i\,;\,\theta)^{r_{ik}s_{jk}}
\end{aligned}
$$

- E-Step: compute $P(R, S|\mathbf{x}, \theta)$
- Variational E-Step: approximate posterior with

$$Q(R, S|\mathbf{x}, \theta) = \prod_{i,k} g_{ik}^{r_{ik}} \prod_{k,j \in k} m_{jk}^{s_{jk}}$$

$$\begin{aligned} \mathcal{F}(Q, \theta) &= E_Q\Big[-\log P(\mathbf{x}, R, S|\theta) + \log Q(R, S|\mathbf{x}, \theta) \Big] \\ &= \sum_{k,j \in k} m_{jk} \log m_{jk} + \sum_{i,k} g_{ik} \log g_{ik} + \sum_{i,k,j} g_{ik} m_{jk} d_{ijk} \end{aligned}$$

where $d_{ijk} = \log \sigma_{ijk} + \frac{(x_i - \mu_{ijk})^2}{2\sigma_{ijk}^2}$

further constraint: $\{g_{ik}^c\}$ consistent for any observation $X^c \rightarrow$ favours distributions over $\{\mathbf{r}_i\}$ that are consistent with other observed data vectors

EM Updates

E Step

$$m^c_{jk} = \exp\Big(-\sum_i g_{ik}\, d^c_{ijk}\Big) \Big/ \sum_{\alpha=1}^{J} \exp\Big(-\sum_i g_{ik}\, d^c_{i\alpha k}\Big)$$

M Step

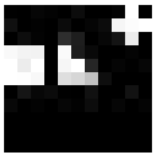
$$g_{ik} = \exp\Big(-\sum_{c,j} m^c_{jk}\, d^c_{ijk}\Big) \Big/ \sum_{\beta=1}^{K} \exp\Big(-\sum_{c,j} m^c_{j\beta}\, d^c_{ij\beta}\Big)$$

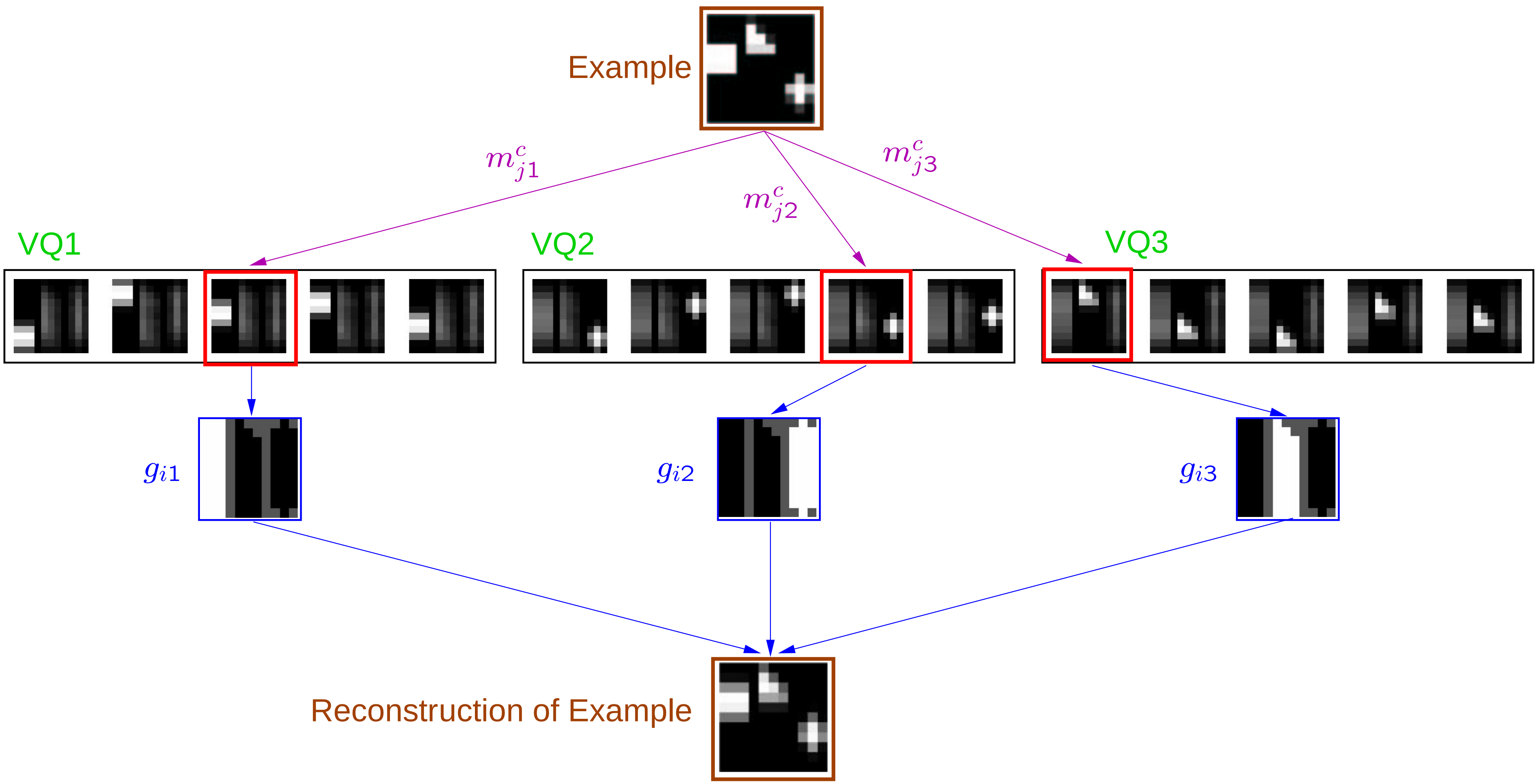
$$\mu_{ijk} = \sum_c m^c_{jk}\, x^c_i \,/\, \sum_c m^c_{jk} \qquad \sigma^2_{ijk} = \sum_c m^c_{jk}(x^c_i - \mu_{ijk})^2 \,/\, \sum_c m^c_{jk}$$

Intuition: one state per VQ, choose one VQ per pixel, that matches input

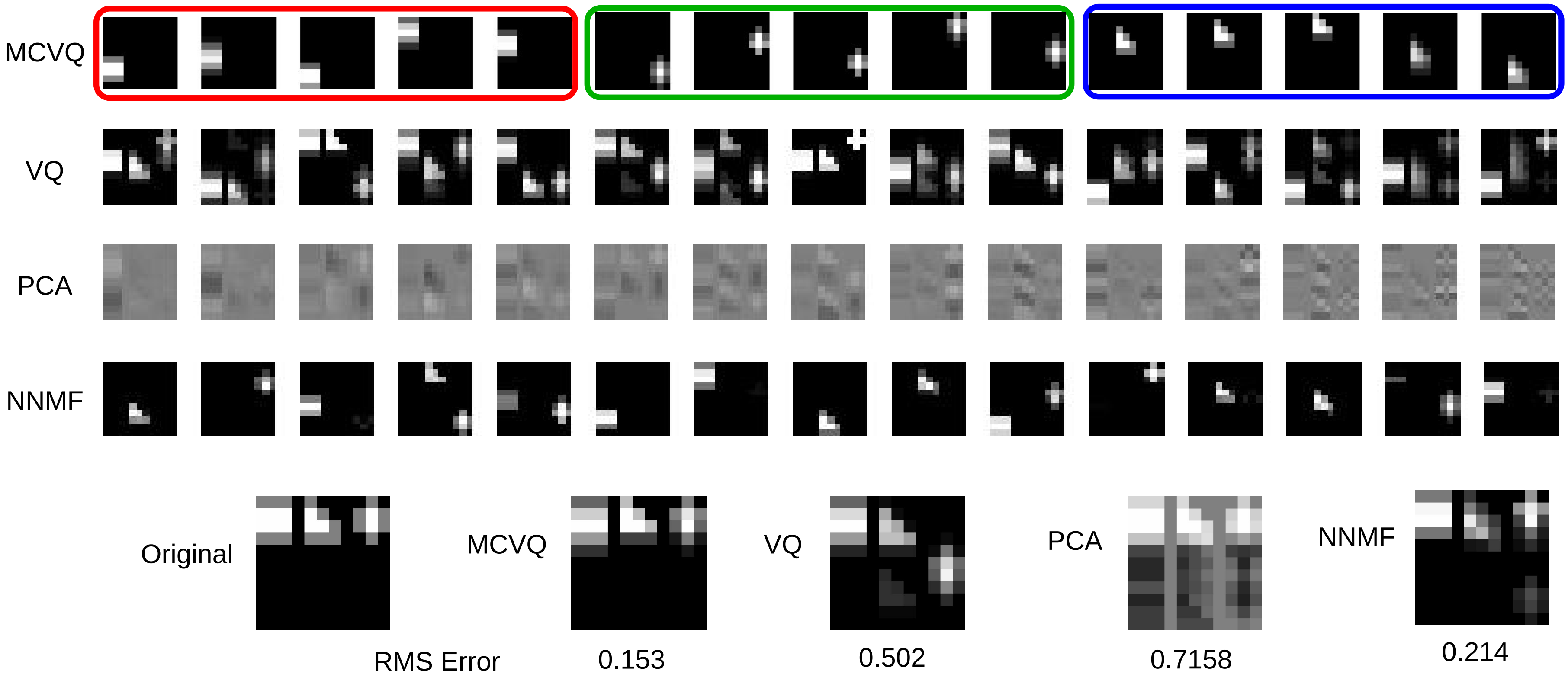
Experiments 1. Shapes

Data Examples:





Experiments 1. Shapes: Comparing Methods



Related Models

Cooperative Vector Quantization

- x_i is generated by the VQ's cooperatively (linear combination), rather than competitively (stochastic selection)

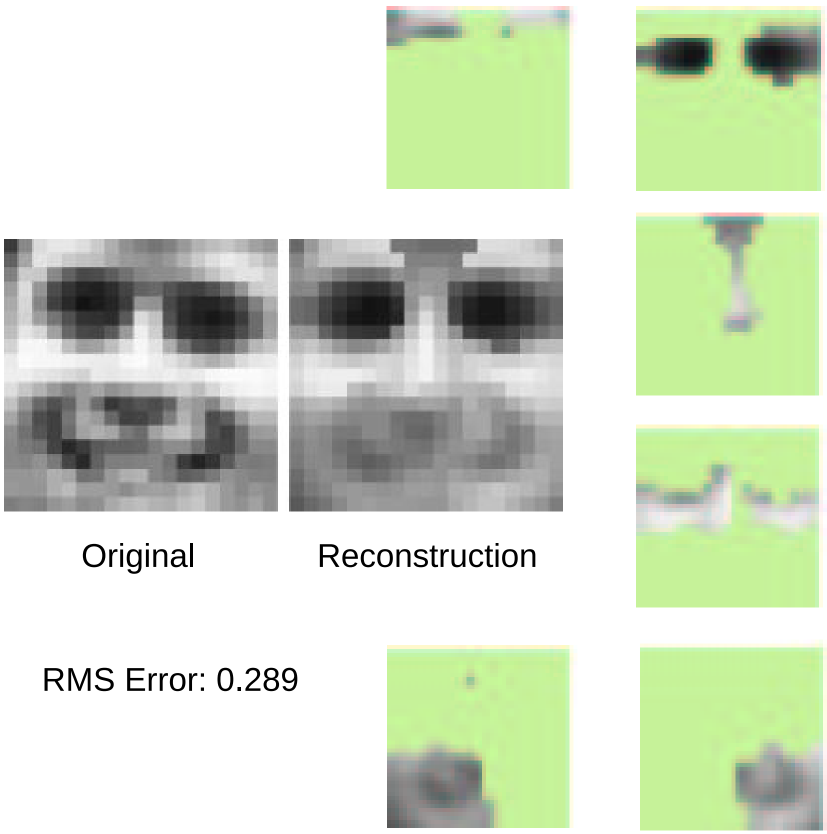
Non-Negative Matrix Factorization

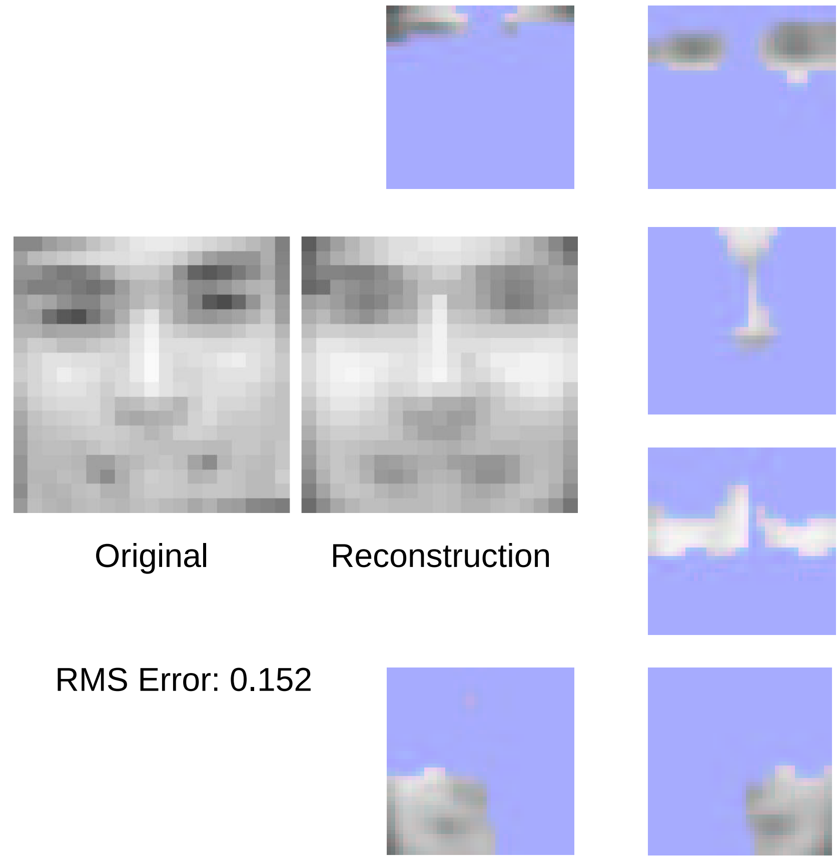
- $\mathbf{x} \sim$ Poisson with mean $\mathbf{W}\mathbf{v}$, where $\mathbf{W}, \mathbf{v} \geq 0$
- non-negativity constraints result in sparse, parts-based, basis vectors $\mathbf{w}_j$
- MCVQ is similar*, with $\mathbf{W} = [\boldsymbol{\mu}_{jk} * \mathbf{g}_k]$, and $\mathbf{v} =$ concatenation of $\mathbf{s}_k$'s
- NMF doesn't group related parts
- models differ in what novel examples they can generate

Flexible Sprites in Video Layers

- learns a single appearance for each object
  infers location & occlusion ordering
- MCVQ assumes fixed locations,
  learns locations & ranges of appearances of objects
  infers appropriate appearances

Experiments 2. Faces





Experiments 3. Text

- Bag of Words - represent document as a word count vector (one element per vocabulary word)
- each VQ state predicts a document word count
- values of g_{ik} provide a segmentation of the vocabulary into subsets of words with correlated frequencies
- within a particular subset, words can be
  - related - tend to appear in the same documents
  - contrasting - seldom appear together
- a particular VQ state is characterized by the words whose predicted count differs most from average

Predictive Sequence Learning in Recurrent Neocortical Circuits

R. P. N. Rao & T. J. Sejnowski

afferent	**ekf**	**latent**	**ltp**
lgn	**niranjan**	**som**	**gerstner**
interneurons	**freitas**	**detection**	**zador**
excitatory	**kalman**	**search**	**soma**
membrane	**wp**	**data**	**depression**
query	critic	mdp	spline
documents	stack	pomdps	tresp
chess	suffix	prioritized	saddle
portfolio	nuclei	singh	hyperplanes
players	knudsen	elevator	tensor

The Relevance Vector Machine
Michael E. Tipping

svms	**hme**	**similarity**	**extraction**
svm	**svr**	**classify**	**net**
margin	**svs**	**classes**	**weights**
kernel	**hyperparameters**	**classification**	**functions**
risk	**kopf**	**class**	**units**
jutten	chip	barn	mdp
pes	ocular	correlogram	pomdps
cpg	retinal	interaural	littman
axon	surround	epsp	prioritized
behavioural	cmos	bregman	pomdp

Missing Data

- model naturally handles case of unobserved data
- all data dimensions are leaves in the graphical model, so unobserved values play no role in learning or inference
- collaborative filtering application - EachMovie dataset
- active approach to learning - VQ responsibilities indicate relationships between data elements

Experiments 4. EachMovie

The Fugitive 5.8 (6)
Terminator 2 5.7 (5)
Robocop 5.4 (5)

Kazaam 1.9 (-)
Rent-a-Kid 1.9 (-)
Amazing Panda Adventure 1.7 (-)

Pulp Fiction 5.5 (4)
The Godfather: Part II 5.3 (5)
The Silence of the Lambs 5.2 (4)

The Brady Bunch Movie 1.4 (1)
Ready to Wear 1.3 (-)
A Goofy Movie 0.8 (1)

Cinema Paradiso 5.6 (6)
Touch of Evil 5.4 (-)
Rear Window 5.2 (6)

Jean de Florette 2.1 (3)
Lawrence of Arabia 2.0 (3)
Sense & Sensibility 1.6 (-)

Best of Wallace & Gromit 5.6 (-)
The Wrong Trousers 5.4 (6)
A Close Shave 5.3 (5)

Robocop 2.6 (2)
Dangerous Ground 2.5 (2)
Street Fighter 2.0 (-)

Tank Girl 5.5 (6)
Showgirls 5.3 (4)
Heidi Fleiss: Hollywood Madam 5.2 (5)

Talking About Sex 2.4 (5)
Barbarella 2.0 (4)
The Big Green 1.8 (2)

Mediterraneo 5.3 (6)
Three Colors: Blue 4.9 (5)
Jean de Florette 4.9 (6)

Jaws 3-D 2.2 (-)
Richie Rich 1.9 (-)
Getting Even With Dad 1.5 (-)

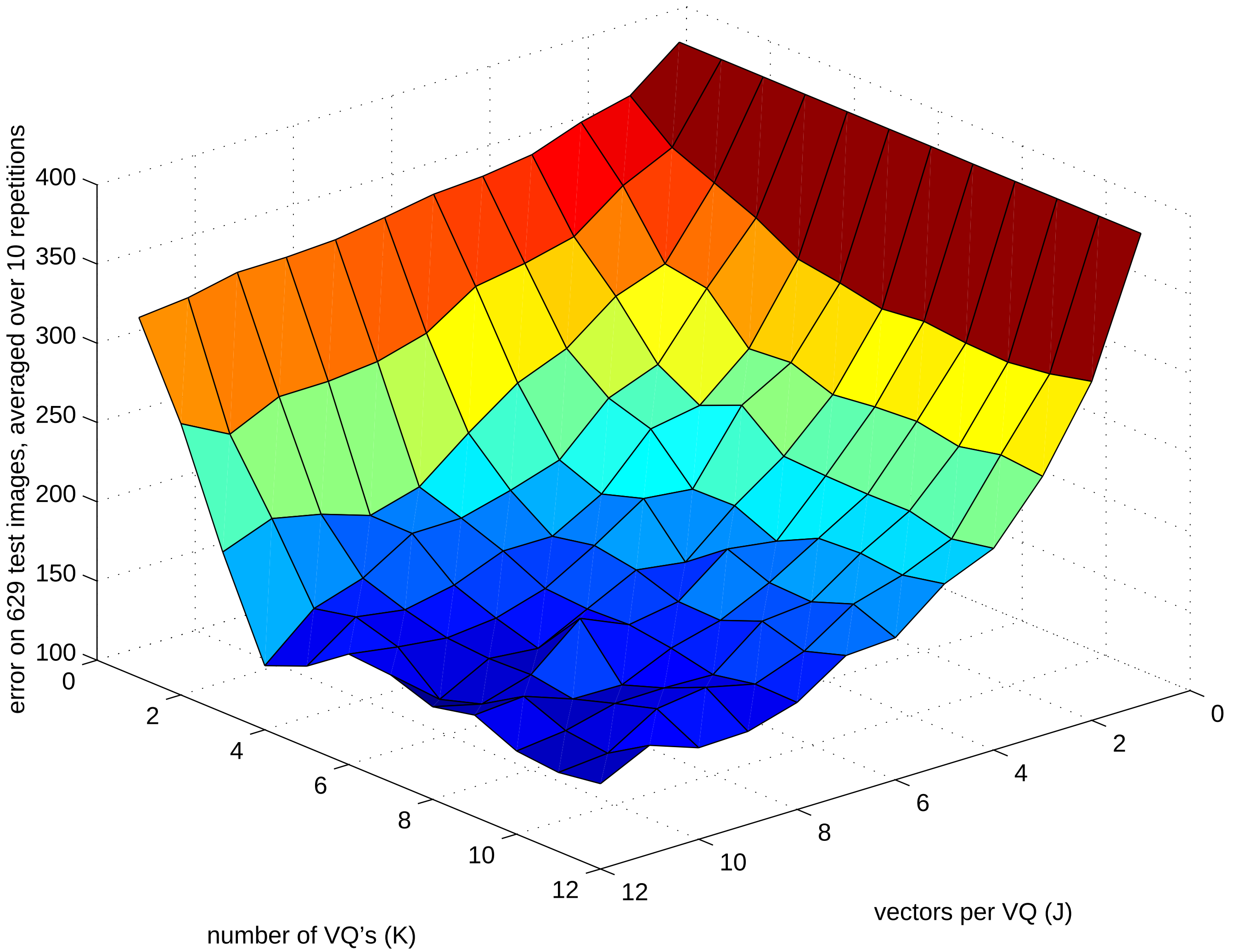
Current Directions

1. model selection

2. relaxing ownership restriction

3. sequential/incremental learning

Cross-Validation on Shapes Data



Model Selection

- quality of learned representation depends strongly on selecting correct # of factors, K
- Goal: want to determine best K (and J)
- compare likelihood estimates for various K's
  - ML doesn't penalize for model complexity
- cross-validation
  - computationally expensive
  - explicitly trains & tests all possible models under consideration

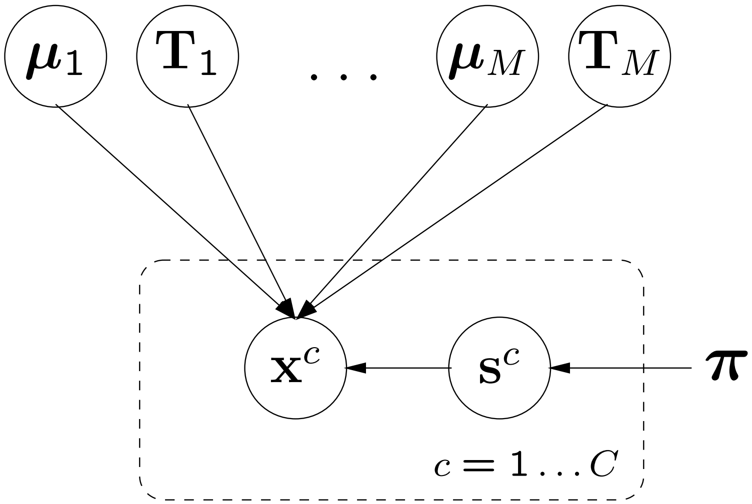
Variational Bayesian Learning

- select model, $\mathcal{M}$, with highest evidence, integrating over choice of parameters, θ :

$$P(X|\mathcal{M}) = \int P(X|\theta)P(\theta|\mathcal{M})d\theta$$

- penalizes models with more degrees of freedom
- avoids overfitting, since parameters are not fit to the data
- requires computing a difficult integral
- use a variation approximation, $Q(\theta)$ to $P(\theta|X, \mathcal{M})$
  $\rightarrow$ optimize a lower bound, $\mathcal{L}(Q)$, on the log-evidence
- Variational EM: maximize $\mathcal{L}(Q)$ wrt Q (E-Step), then $\mathcal{M}$(M-Step)

VB Mixture of Gaussians (Corduneanu & Bishop)



$\boldsymbol{\mu} \sim \mathcal{N}(0, aI)$

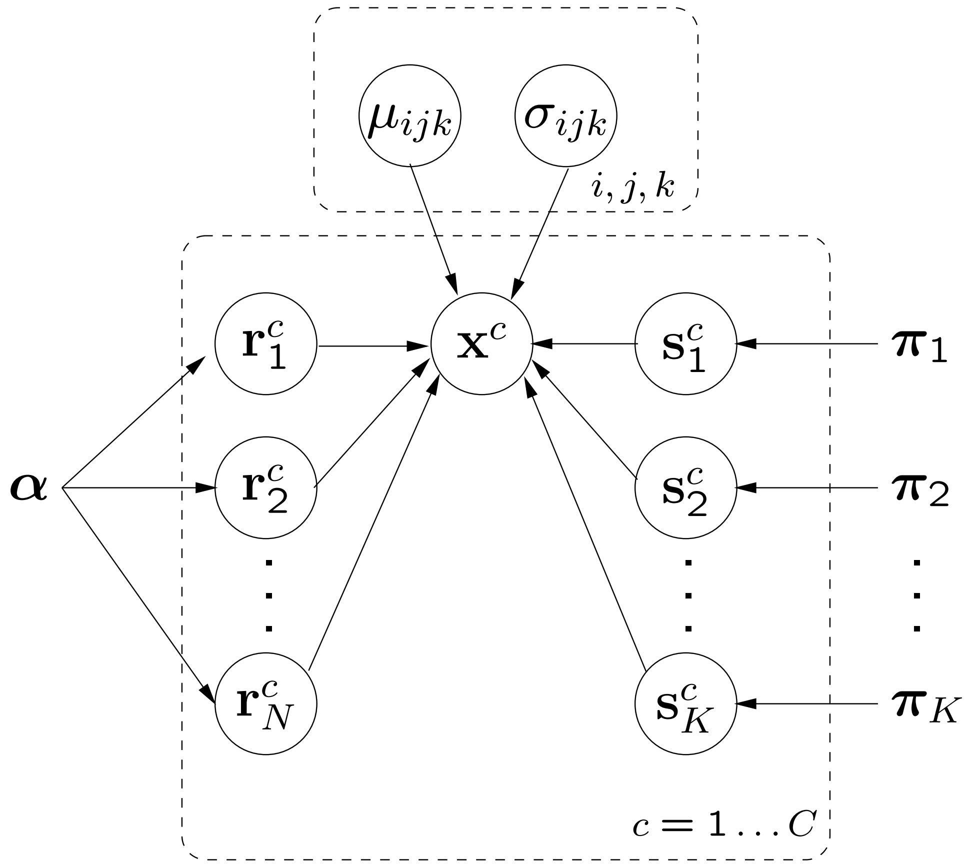
$\mathbf{T} \sim$ Wishart

$\mathbf{s} \sim$ Discrete($\boldsymbol{\pi}$)

$$\mathcal{L}(Q) = \int Q(\mu)Q(T)Q(s) \ \ln \ \frac{P(D, \theta|\pi)}{Q(\mu)Q(T)Q(s)} \, d\theta$$

- start with a fixed number of potential components (the maximum # considered)
- optimize using variational EM
  $\longrightarrow$ causes priors of unwanted components (π's) to go to zero

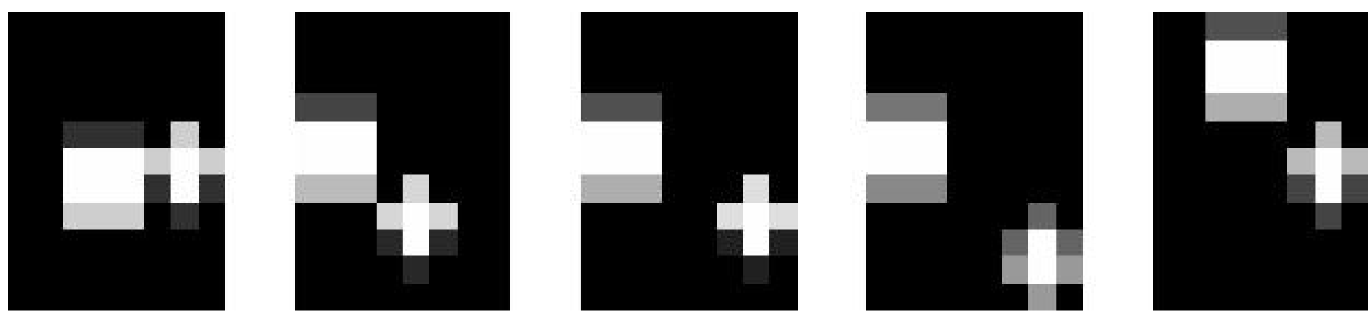
VB MCVQ



- remove VQ k when $\alpha_k \approx 0$
- remove state j of VQ k, when $\pi_{jk} \approx 0$

Overlapping Causes

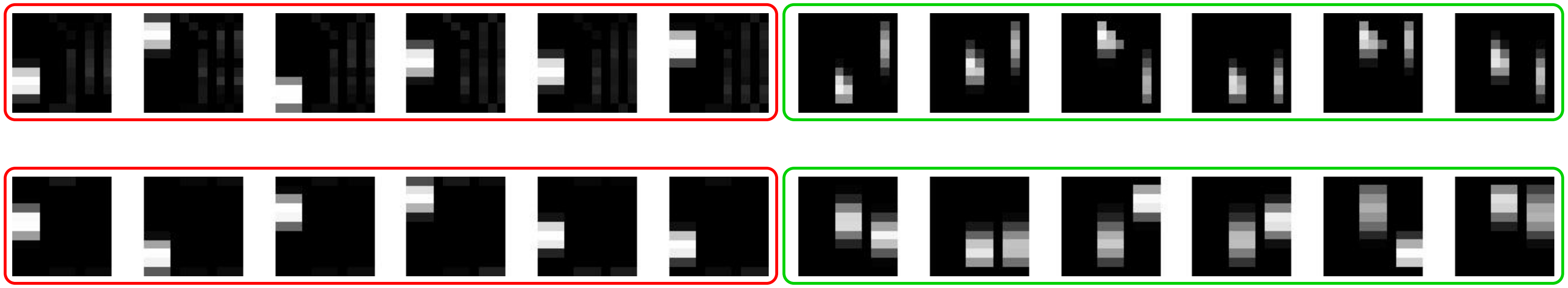
- with current implementation, g_{ik}'s always binary
- would like non-binary g's in some cases,
  e.g. at object borders in natural images
- Sample Data:



- Results: still binary!

Incremental MCVQ

- learn causes one at a time, as per Williams & Titsias
- train model with one (or more) ordinary VQ's, and one VQ with fixed, high variance
- hopefully ordinary VQ's will learn one cause each, high variance VQ will learn the remainder
- Results:



- Issue: choosing variances?
- Next: try this on text data
- Alternatively: a single low variance VQ, collects static data dimensions