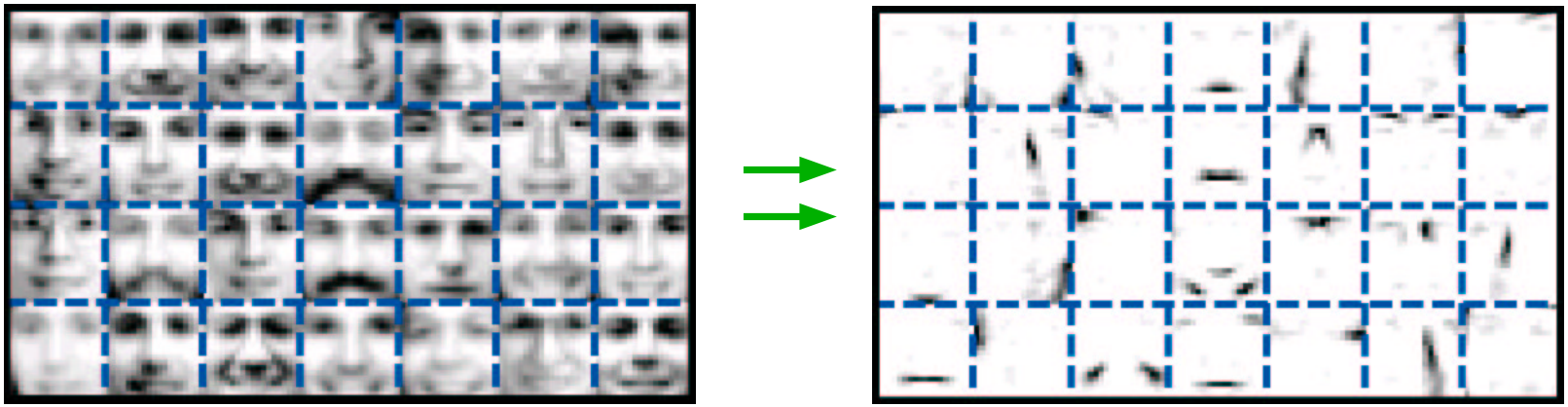
Motivation

- there are many unsupervised methods for learning parts-based representations of data
- e.g. non-negative matrix factorization (NNMF), given face images, learns the parts of which faces are composed

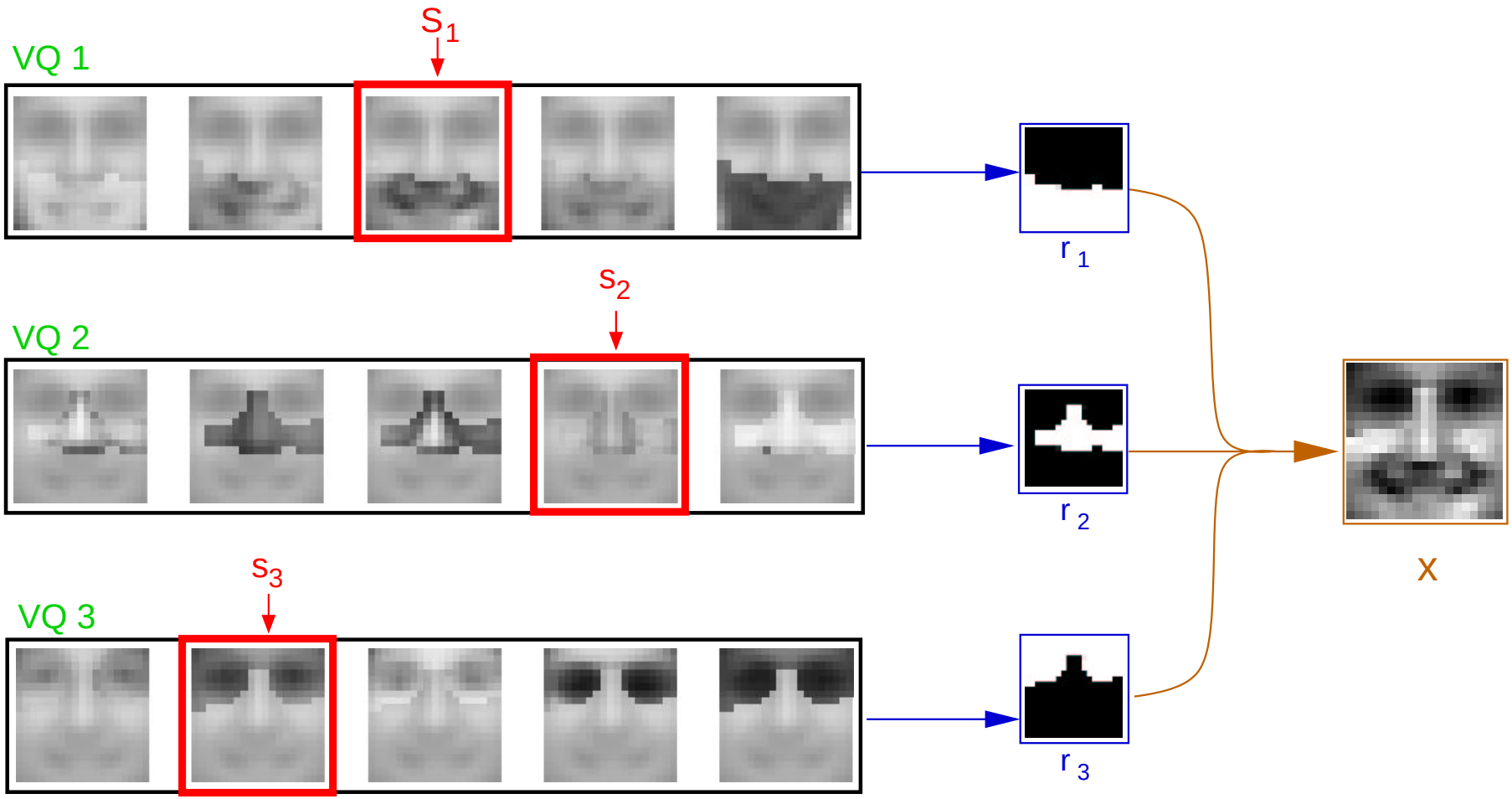


- NNMF does **not** (1) group related parts, or (2) learn how parts may be composed to create a valid whole
- we would like to learn the appearances of these parts, as well as their compositional structure

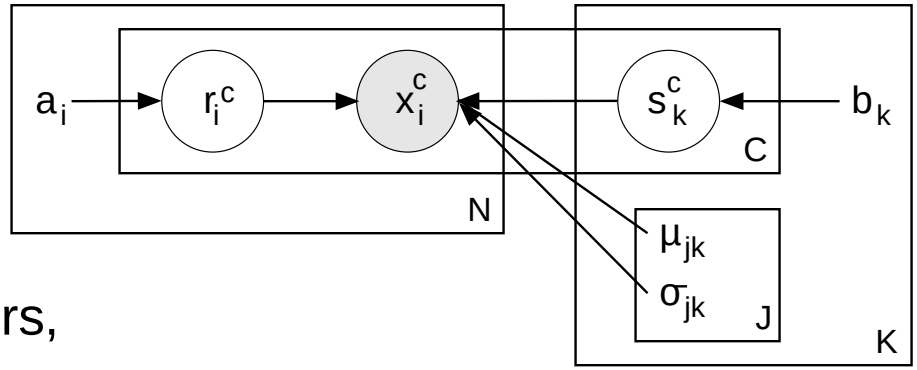
Overview

- Goal: learn a representation of vector data consisting of:
  - Parts: disjoint subsets of the data dimensions (multiple causes)
  - Appearances: a discrete characterization of the range of appearances for each part (vector quantization)
- Example: on face image data,
  parts could be *eyes*, *nose*, and *mouth*
  appearances could be different sizes and shapes of these parts
- Win: combinatorial power
  - VQ with N states represents N items
  - MCVQ with j states per N/j VQ's represents $j^{N/j}$ items
- formulate as a generative probabilistic graphical model, use variational EM to learn the maximum-likelihood parameters
- applications to image segmentation, text analysis, collaborative filtering

An Illustrative Example



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

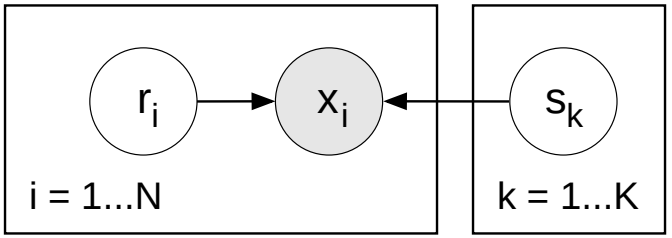
- Complete Likelihood

$$P(\mathbf{x}, R,S|\theta) = P(R|\theta)P(S|\theta)P(X|R,S,\theta)$$

$$= \left(\prod_{ik} a_{ik}^{r_{ik}}\right)\left(\prod_{jk} b_{jk}^{s_{jk}}\right)\prod_{ijk}\mathcal{N}(x_i;\mu_{ijk},\sigma_{ijk})^{r_{ik}s_{jk}}$$

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

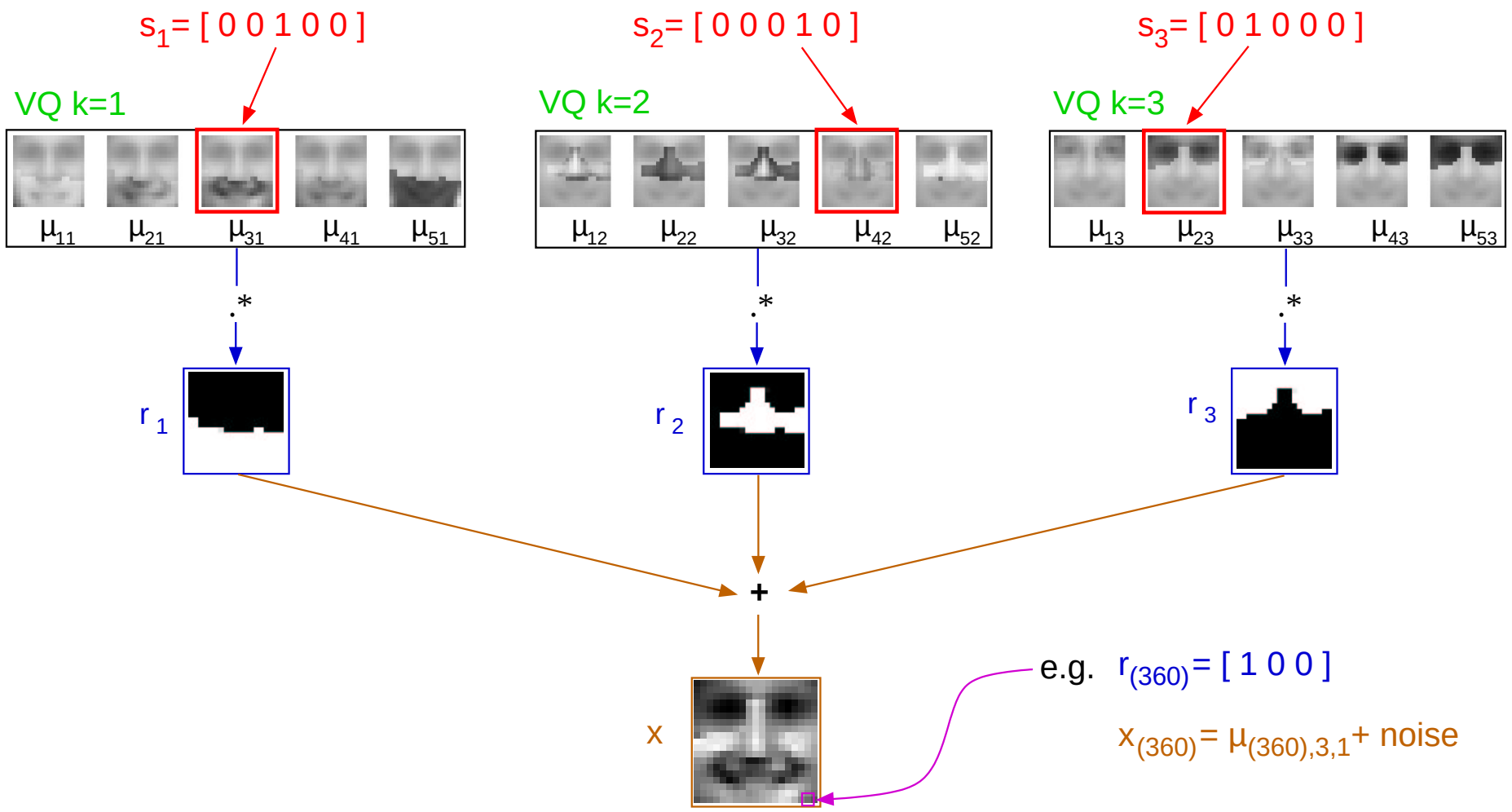


- $\mathbf{S} = \{\mathbf{s}_k\}$ J-dim. indicator vectors, select one state per VQ
- $\theta = \{\mu_{ijk}, \sigma_{ijk}\}$ parameters of dimension i, from j^{th} state of k^{th} VQ
- $\mathbf{a}_i$'s and $\mathbf{b}_k$'s prior distribution over $\mathbf{r}$'s and $\mathbf{s}$'s

- Complete Likelihood

$$P(\mathbf{x}, R,S|\theta) = P(R|\theta)P(S|\theta)P(X|R,S,\theta)$$

$$= \left(\prod_{ik} a_{ik}^{r_{ik}}\right)\left(\prod_{jk} b_{jk}^{s_{jk}}\right)\prod_{ijk}\mathcal{N}(x_i;\mu_{ijk},\sigma_{ijk})^{r_{ik}s_{jk}}$$


s_1= [0 0 1 0 0]
s_2= [0 0 0 1 0]
s_3= [0 1 0 0 0]
VQ k=1
VQ k=2
VQ k=3
μ_{11} μ_{21} μ_{31} μ_{41} μ_{51}
μ_{12} μ_{22} μ_{32} μ_{42} μ_{52}
μ_{13} μ_{23} μ_{33} μ_{43} μ_{53}
.*
.*
.*
r_1
r_2
r_3
+
x
e.g. $r_{(360)}$= [1 0 0]
$x_{(360)}$= $\mu_{(360),3,1}$+ noise

- E-Step: compute $P(R, S|\mathbf{x}, \theta)$
  computationally intractable since all r_{ik} and s_{jk} are mutually dependent (distribution cannot be factorized, and there are $J^K K^N$ possible combinations of (R, S))

- Variational E-Step: approximate posterior with

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

- Variational Free Energy:

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

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

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

M Step

$$g_{ik} = \exp\Big(-\frac{1}{C}\sum_{c,j} m_{jk}^c\, d_{ijk}^c\Big) \Big/ \sum_{\beta=1}^{K} \exp\Big(-\frac{1}{C}\sum_{c,j} m_{j\beta}^c\, d_{ij\beta}^c\Big)$$

$$\mu_{ijk} = \sum_c m_{jk}^c\, x_i^c \;/\; \sum_c m_{jk}^c \qquad \sigma_{ijk}^2 = \sum_c m_{jk}^c (x_i^c - \mu_{ijk})^2 \;/\; \sum_c m_{jk}^c$$

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

An Alternative Model

- restrict selections of VQ's, $\{r_{ik}\}$, to be the same for each training example
- update rule for g_{ik} becomes:

$$g_{ik} \propto \exp\Big(-\sum_{c,j} m^c_{jk}\ d^c_{ijk} \Big)$$

- in practice, we obtain good results by making $g_{ik} \propto \exp\big(-\frac{1}{T} \sum \ldots \big)$, and annealing the temperature, T, during learning

$\Rightarrow$ gradually moving from generative model in which r_{ik}'s can vary across examples, to one where r_{ik}'s are consistent across examples

Related Methods

Cooperative Vector Quantization (Zemel-Hinton; Ghahramani)

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

Non-Negative Matrix Factorization (Lee-Seung)

- $\mathbf{x} \sim$ 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 (* but uses Gaussian instead of Poisson noise)
- NNMF doesn't group related parts
- models differ in what novel examples they can generate

Credibility Networks (Hinton-Ghahramani-Teh)

Dynamic Trees (Williams-et al)

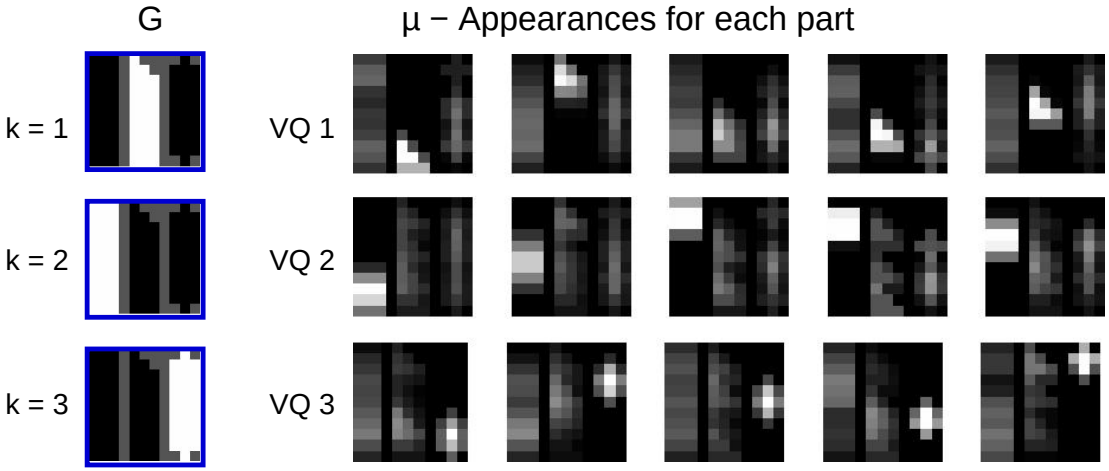
Flexible Sprites in Video Layers (Jojic-Frey)

- these methods focus mainly on the unknown pose (primarily position) of an object in an image
- they learn a single appearance for each object
  infer location & occlusion ordering
- MCVQ assumes fixed locations,
  learns locations & ranges of appearances of objects
  infers appropriate appearances
- our focus $\longrightarrow$ learning a parts-based decomposition

Experiments: Shapes

- data consists of 11x11 gray-scale images, each containing a box, triangle, and cross - vertical positions of shapes vary independently

- model trained with 3 VQ's, 5 appearances each

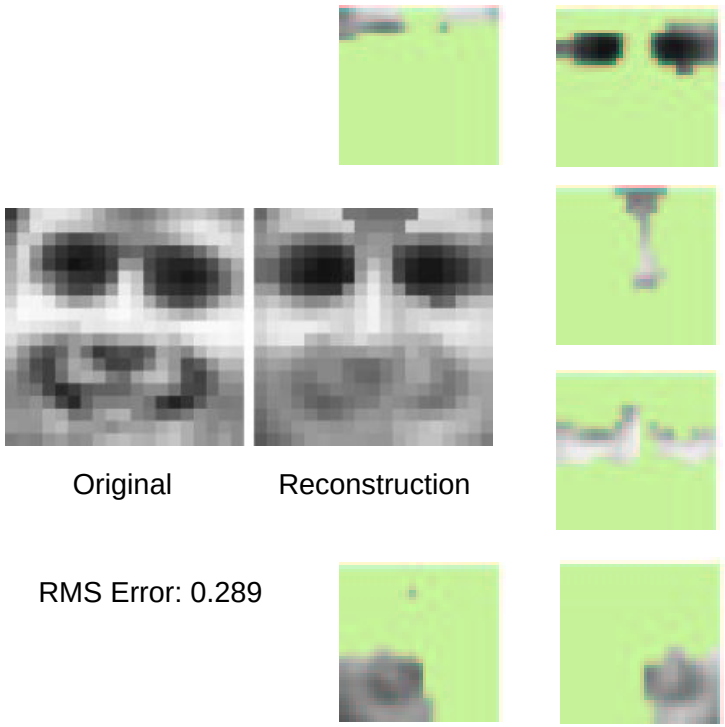


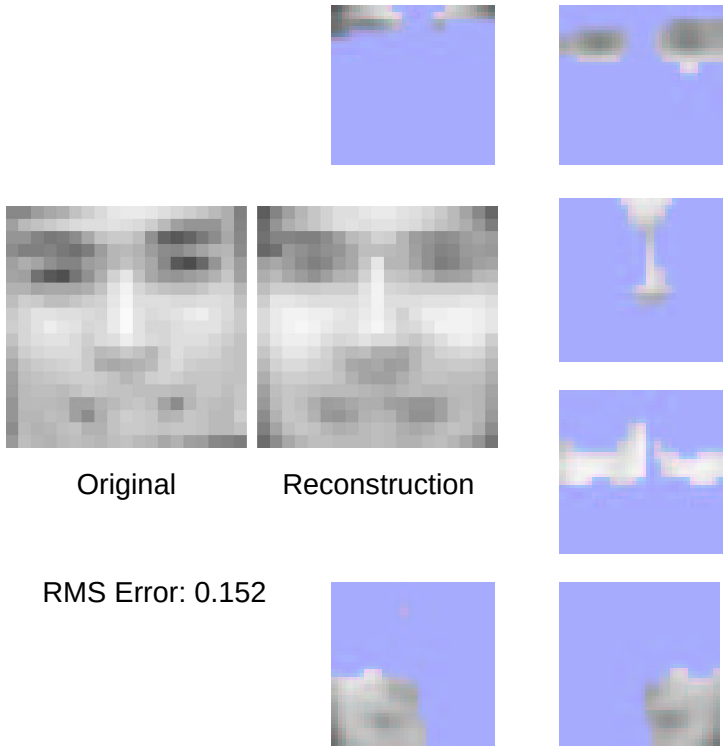
- comparison of RMS reconstruction error, versus other methods, on a novel shapes image:

RMS Error	0.502	0.7158	0.214

Experiments: Faces

- dataset: 19x19 gray-scale images of frontal faces
- model trained on 2000 images, using 6 VQ's, 12 appearances each
- reconstruction of two images from the test set - beside each are the specific appearances of each part (the most probable ones) used to generate it





Experiments: Text

- Bag of Words - represent document as a word count vector (one element per vocabulary word)
- each VQ state predicts a document word count
- learned parts 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 appearance is characterized by the words whose predicted count differs most from average
- experiments on NIPS Proceedings 0-12 Data (1740 documents, 14,265 word vocabulary) using a model of 8 VQ's, 8 appearances each

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

- each column is an appearance selected as most most likely for this document
- **bold** (plain) words have **highest** (lowest) predicted frequencies, relative to their averages

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
- the probability of each appearance can be inferred from the available observations for each part
- collaborative filtering - model can be learned on incomplete data, missing values for test vectors can be inferred
- strong ties between data dimensions and parts allow an active approach to inference $\rightarrow$ VQ responsibilities indicate relationships between data elements

Experiments: EachMovie

- EachMovie data consists of ratings on a scale from 1 to 6 on $\sim$1600 movies, by $\sim$74 000 users
- data very sparse - most users rated only a few movies
- we restricted the data to movies rated by > 125 users, and users rating ≥ 75 movies (still very sparse)
- trained an MCVQ model with 8 VQ's, 6 appearances each
- test set contained ratings vectors with some ratings "hidden", and the model was used to infer the hidden ratings
- preliminary results comparable with PLSA

User 1

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 (-)

User 2

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 (-)

- each column is an appearance selected as most most likely for this user
- **bold** (plain) movies have the **highest** lowest predicted ratings, relative to their average ratings

Current Directions

- Bayesian model selection to determine number of parts, and number of appearances per part
- hierarchical learning, e.g. condition appearance selection on an observed variable
  - Faces - condition on expression (happy, sad, angry, ...)
  - Text - condition on author

$\rightarrow$ ideally, treat these higher-level attributes as hidden variables and learn them in an unsupervised fashion