Final Report for

Model-Biased, Data-Driven Adaptive Failure prediction

NASA Cooperative Agreement NCC 2-1264

PI – Todd K. Leen
Period – March 1, 2001 – August 1, 2004

OGI School of Science and Engineering
Oregon Health & Science University
20000 N.W. Walker Road
Beaverton, OR 97006

503-748-1160

Summary of Research Activities and Findings

1. Helicopter Gearbox Anomaly Detection

The initial aim of this project was to provide machine learning support for failure prediction in helicopter gearboxes. We sought to develop anomaly, or outlier, detectors based on accelerometer measurements of gearbox vibration. Due to the large variability of vibration signatures with the aircraft's dynamical state (e.g. maneuvers), we recognized early that useful outlier detection would require knowledge of the state to allow conditioning.

Our initial studies were aimed at using features derived from vibration total RMS power and vibration spectra to identify -- via (unsupervised) clustering – aircraft maneuvers. The gearbox data for these studies consisted of the instantaneous signal from six accelerometers time-synchronously averaged at three different periods: the pinion, bevel, and rotor periods. The raw data is thus an 18-dimensional time series. These signals were available for 14 maneuvers, which we clustered into 9 classes based on symmetries.

RMS Power – RMS power was calculated from the entire time series (~34 sec) at each maneuver. To enhance clustering and aid visualization, we applied both PCA and discriminatory feature selection to reduce the signal dimension from 18 to 7.

Several clustering techniques, with cross-validation used to determine number of clusters, were applied. Gaussian mixture models severely underestimate the number of clusters (typically 3), yielding a poor discrimination between the maneuvers (37% classification rate). Entropy-constrained k-means (standard k-means with a regularizer consisting of

and entropy penalty to encourage small models) produces good classification (89% classification rate) but grossly overestimates the number of clusters (typically ~28). The k-means classifier accuracy is comparable to results obtained by NASA ARC scientists on the same data using a (supervised) neural network classifier. Entropy-constrained adaptive PCA typically gives 6 clusters and a 65% classification rate. The local dimension of the clusters range from 0 to 5.

Spectral Features – Spectra carry more detailed information than the total RMS power, and are expected to be important in anomaly detection. Auto-regressive spectral estimates failed to discriminate between maneuvers, so we turned to Welch-averaged estimates. We explored an extensive range of averaging windows reflecting a wide coverage of the bias-variance tradeoff in the spectral estimates. We further explored several techniques for concatenating and normalizing the spectra from the six accelerometers. Results indicated that with properly-chosen Welch-averaging and concatenation, unsupervised maneuver classification comparable to, but not better than, that resulting from RMS power features is obtained. Clustering based on the FFT of the time-synchronous average accelerometer traces did *not* perform as well as the best Welch-averaged spectra.

Nonstationarity – Marianne Mosher at NASA ARC determined that accelerometer signals are *not* stationary over the 34-second period used in the time-synchronous averages. She suggested timescales over which the signals *are* stationary. We found that the suggested short-time averaged spectra are *less-easily* clustered by maneuver. That is, maneuvers overlap more in this representation. Presumably, the shorter time averages contribute to noisy spectra, and the required Welch-averaging smoothes over discriminatory information.

The nonstationarity results suggested that maneuver are an insufficient specification of dynamical state. More refined indicators are required. Flight-bus data could provide the fine-scale dynamical state information required to understand the relationship between flight-state and vibration-signature during nonstationary flight. Based on this, and earlier results, in October 02, we requested pooled vibration and flight-bus data. These data were not available until late in March 03, by which time we had redirected our research thrust to remote earth observing data.

2. Application of Complexity-Penalized Clustering to Segmentation of EOS Data

In collaboration with Ashok Srivastava ar ARC, in early 2003 we began investigating the use of novel clustering techniques for exploration and segmentation of multi-channel imaging spectrometer data from NASA Earth Observing satellites. Dr. Srivastava had been using kernel-based clustering for segmentation of EOS images. His initial exploration on an image of Greenland turned up an unexpected identification of a possible ice-melt region.

The results of clustering algorithms are sensitive to initial conditions, and Dr. Srivastava voiced an interest in obtaining low-variability alternatives to the algorithms he has been using. The entropy-penalized clustering algorithms we had been exploring with helicopter data have a natural mechanism for suppressing variability, and like the kernel methods, have more flexible modeling capability than standard approaches. This led to our collaboration on these problems.

Our initial studies explored application of several entropy-constrained algorithms to portions of multi-channel spectrometer images of Sicily and of Greenland. We reproduced Dr. Srivastava's segmentation with slight differences in the boundary.

We found that an entropy-constrained k-means algorithm provides lower variability with respect to initial conditions than does unconstrained k-means, or our adaptive PCA algorithms. We have not yet compared our variability results with Dr. Srivastava's, though we find very robust replication of the segmentation feature he discovered, albeit with small variations of the boundary.

We explored the use of a genetic algorithm clustering to reduce variability. Our study showed that the computational complexity is unfavorable with respect to simple clustering with multiple restarts.

Finally, and most productively, we explored incorporating **hints** to help constrain clustering. These hints consist of human-induced biases that encourage, or discourage, co-clustering of a small number of pairs of datapoints. This is a form of prior knowledge that is weaker than class labels. Our resulting algorithm is a probabilistic clustering model (mixture model) that successfully generalizes the information in the hints to out-of-sample data.

This algorithmic development, and its application to the Greenland image data was published in NIPS 17 (see publications). We are also drafting a journal article on this material for submission in June 2005.

Educational Activity

This award supported a portion of the doctoral studies of Cynthia Archer. She received her Ph.D. degree in June, 2002. Dr. Archer is now employed at the Portland, OR office of Research Triangle Park.

This award supported a portion of the doctoral studies of Zhengdong Lu. Zhengdong is currently a Ph.D. student in the PIs lab.

The award also funded research activities of a postdoctoral research student, Dr. Alex Nelson, who worked with us on the helicopter gearbox data during the fall and early winter of 2002, and also working on preliminary aspects of the segmentation of EOS data. Dr. Nelson is now employed in biomedical signal processing at Inovise.

Publications

Zhengdong Lu and Todd K. Leen. Semi-supervised Learning with Penalized Probabilistic Clustering. In *Advances in Neural Information Processing Systems 17*, Saul, Weiss, and Bottou (eds), The MIT Press, 2005.

Zhengdong Lu and Todd K. Leen. Prior Knowledge for Probabilistic Clustering. In preparation for submission to *Neural Computation*. The target submission date is June 13, 2005.

Patent Activity – None

Ancillary Materials

Presentations from the IS PI workshops are appended below.

Building Better Clusters

Unsupervised Classification for Novelty Detection

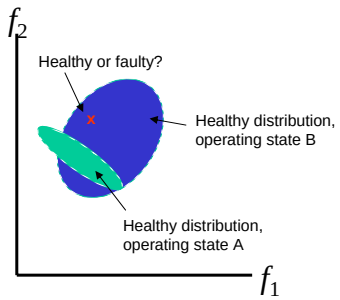
Towards Application to Failure Prediction

Sept 4, 2002

Cynthia Archer, Lu Zhengdong,
Todd Leen

Motivation and Algorithm Grounding

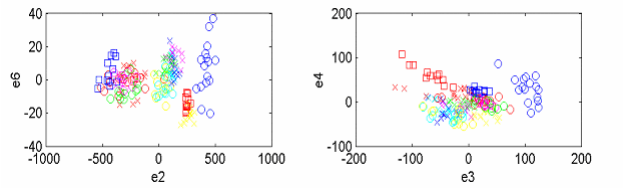
- **Outlier detection to identify anomalies**
- **Accurate models of healthy baseline**
  - **"healthy" must be conditioned on operating state – mixture or local models for nonstationarity**



Clustering Approaches

- Clustering ←→ Gaussian Mixture Density Models
  - How many clusters?
  - What shape (constraints of mixture components)?
  - Dimensionality for PCA-based clustering?

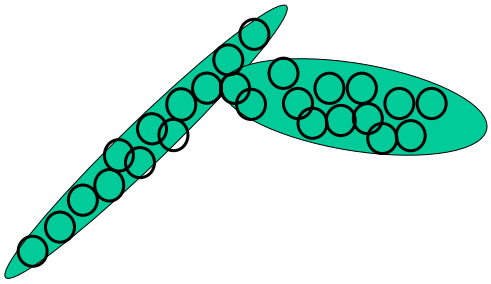
e.g. Helicopter gearbox RMS vibration signal from 6 accelerometers in 14 different maneuvers



Todd K. Leen
OGI – OHSU Sept. 4, 2002

Clustering Approaches

- For k-means (spherical 0-d clusters), how do model clusters correspond with true data clusters?



Visualization

Todd K. Leen
OGI – OHSU Sept. 4, 2002

New Algorithm

Entropy-Constrained Adaptive PCA

- **Clustering based on constrained Gaussian mixture model. Constraints related to PCA and factor analysis (Basilevsky, Tipping and Bishop)**
- **Structure includes model resolution parameter (or observation noise variance) σ^2.**

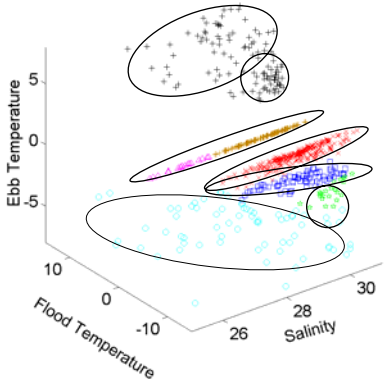
  Formalism leads to entropy-penalized (regularized) cost function directly from likelihood maximization.
- **Locally adjusts cluster dimensionality and shape to data.**
- **Includes unconstrained mixture models and entropy-penalized k-means as special cases.**
- **Number of clusters selected by cost minimization on holdout set.**
- **Makes inspired choices for number of clusters. Functions well for unsupervised classification.**

Todd K. Leen
OGI – OHSU Sept. 4, 2002

What's it do?

COlumbia RIver Estuary Modeling and Observation System (Antonio Baptista, ESE – OGI)

- Salinity and temperature measurements are correlated
- Conditions vary, changing correlation



Gaussian Mixture Model with full covariance matrix

Local PCA with 1 dimension

EC APCA, average dimension 1

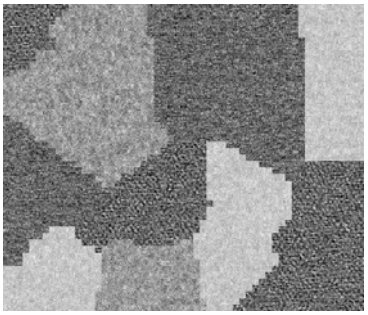
Todd K. Leen
OGI – OHSU Sept. 4, 2002

What Else Does it Do?

- **High-D example that can be visualized – unsupervised texture segmentation**

Training image blocked 9x9

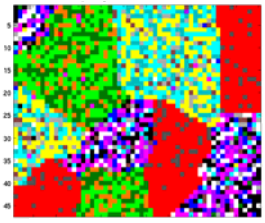
Four-texture test image →



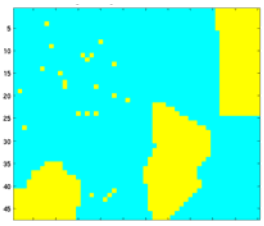
Texture Segmentation

Number of clusters chosen to minimize corresponding clustering cost – <u>not</u> to optimize texture segmentation performance.

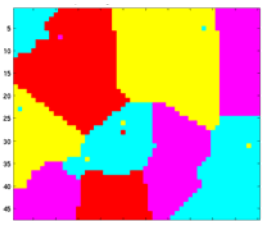
Entropy-constrained K-means



Standard Gaussian mixture



Entropy-constrained APCA



Gearbox Vibration

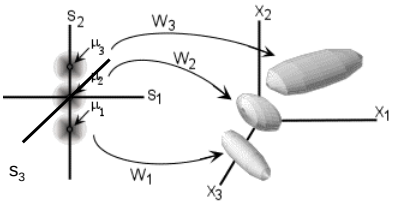
- 14 maneuvers, human-clustered into 9 classes
- Features – RMS power in each of 6 accelerometers from 3 different synchronous averaging periods, 18-dim space, pruned to 7 based on discriminative ability
- Clustering via entropy-constrained k-means, standard Gaussian mixtures, and entropy-constrained APCA. Evaluate clusters as classifier.
- Results
  - Unconstrained Gaussian mixtures severely underestimate number of clusters (3), poor discrimination between real classes (37% classification rate).
  - Entropy-constrained k-means produces good classification (89%) by grossly overestimating number of clusters (28)
  - Entropy-constrained APCA likes 6 clusters, gives 65% classification rate, cluster dimensions from 0 to 5.

Outstanding Issues

- Choosing model resolution σ^2 via cross-validation. Seems to consistently underestimate – estimation bias?
- How to do feature selection for clustering?
- Figure-of-merit for cluster-based unsupervised classifiers?
- How to do real-time operating state conditioning for helicopter data. Operating state – quantized or continuous?
- What about real texture segmentation?
- Applications to other environmental science datasets? Dynamical regime identification by clustering?

New Clustering Framework

- **Clustering based on constrained Gaussian mixture models**
  - **Latent variable generative model → constraint structure related to PCA / FA**
  - **Automatically tunes to local data dimensionality**
  - **Generates entropy-penalized (e.g. regularized) cost function directly from likelihood maximization**
  - **Automatic selection of number of clusters by likelihood maximization on holdout data.**
  - **Appears to work well for unsupervised classification.**



- Latent space *s*
- Maps W_i from *s* to data space *x* (fit)
- Additive noise – variance σ^2 (resolution control parameter, not fit)
- Rank(W_i) determined by data & σ^2 sets local cluster dimension

Todd K. Leen
OGI – OHSU Sept. 4, 2002

Entropy-Constrained Adaptive PCA

- **Density model with** $p(x) = \Sigma\, \pi_\alpha\, p(x|\alpha)$,

$$p(x|\alpha) = N(\mu_\alpha, \sigma_\alpha^2 \mathbf{I} + W_\alpha W_\alpha^T)$$

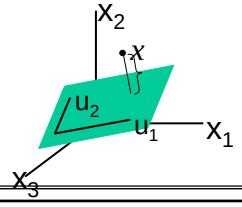
Defines position of local PCA subspaces

sets model resolution

Defines orientation and eigenvalues for local PCA subspaces

- **Hard-clustering limit of data likelihood is entropy-constrained cost**

$$C = \frac{1}{N} \sum_{\alpha=1}^{K} \sum_{x \in R_\alpha} (x - \mu_\alpha)^T (\mathbf{I} - U_\alpha U_\alpha^T)(x - \mu_\alpha) + \sigma^2 \sum_{\alpha=1}^{K} \pi_\alpha (-2 \log \pi_\alpha + h_\alpha)$$



Todd K. Leen
OGI – OHSU Sept. 4, 2002

Entropy-Constrained and Partially-Supervised Clustering

Unsupervised Classification for Novelty Detection and Segmentation

Feb. 4, 2004

Cynthia Archer, Alex Nelson,
Zhengdong Lu,
Todd Leen

Novel Clustering Applied to

- **Helicopter gear-box vibration segmentation/classification, towards anomaly detection**
- **Segmentation of satellite earth-observing data.**

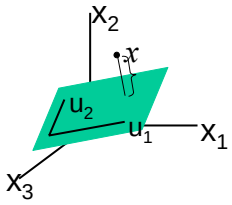
Mixture-PCA Density Model

- **Density model** $p(x) = \Sigma\, \pi_\alpha\, p(x|\alpha)$, **with**

$$p(x|\alpha) = N(\mu_\alpha,\ \sigma^2 \mathbf{I} + W_\alpha W_\alpha^T)$$

sets model resolution

Defines orientation and eigenvalues for local PCA subspaces



- **Soft-clustering through posterior** $p(\alpha|x)$
- **Hard-clustering limit of data likelihood leads to a cost function for entropy-constrained clustering -- entropy-constrained adaptive PCA (EC-APCA).**

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Entropy-Constrained Clustering

- **Automatically tunes to local data dimensionality**
- **Generates entropy-penalized (e.g. regularized) cost function directly from likelihood maximization**
- **Automatic selection of number of clusters by x-validation**
- **Includes unconstrained mixture models and entropy-penalized k-means as special cases.**
- **Number of clusters selected by x-validation.**
- **Selection of model resolution parameter σ^2 by x-validation (with variable results).**

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Application to Helicopter Gearbox Vibration

- Classify maneuver (flight "state") from vibration information. Surrogate task for fault detection.
- Features – RMS power in each of 6 accelerometers from 3 different synchronous averaging periods, 18-dim space, pruned to 7 based on discriminative ability
  - Clustering via entropy-constrained k-means, standard Gaussian mixtures, and entropy-constrained APCA. Evaluate clusters as classifier. (*Classification results ~ comparable to supervised learning*.)
- Features – Welch power spectra of time-synchronous averaged (TSA) time series.
  - Normalize spectra to unit power, concatenate spectra from several gear TSA. Marginally less accurate than clustering via RMS power.
- Long-term (~34 sec) TSA noted (Huff / Mosher, NASA Ames) to be non-stationary. But clustering over short-term TSA provides poor maneuver classification. Suggests need for more detailed state-description than maneuver only.

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Image Segmentation

- Success with texture segmentation (reported last year) suggested application to image segmentation of earth-observing data.
- Unlabeled image data – how to evaluate unsupervised segmentation?

  Compare with human clustering …

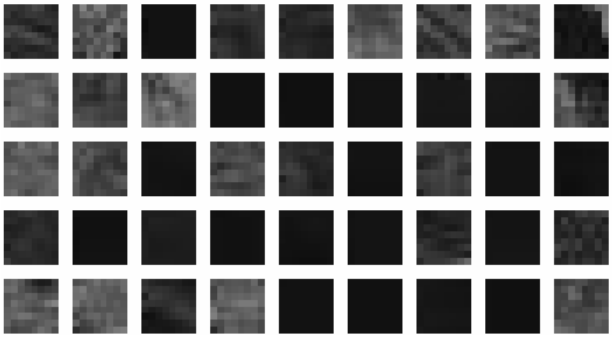
  ~ 68% agreement with 2 different human clusterings.

  Agreement between humans is about 70%

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Clustering Image blocks

- **Suppose we want to label following blocks by clustering**



. . .

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Clustering Image blocks

- **It is hard to tell which cluster a sample should go to, since we don't even know what those clusters look like**

Should go to cluster 2 or 5?

- **It is much easier to tell whether one pair of sample blocks should go into one cluster or not**

 and  should be in same cluster

 and  should be in different clusters

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Clustering Image Blocks

- **Led to partially-supervised mixture-based clustering**
  - **Gaussian mixture model for data density / clustering**
  - **Incorporate pairwise "opinions" into prior on assignment of image blocks to mixture components (clusters).**
  - **"Penalized Probabilistic Clustering" (PPC)**

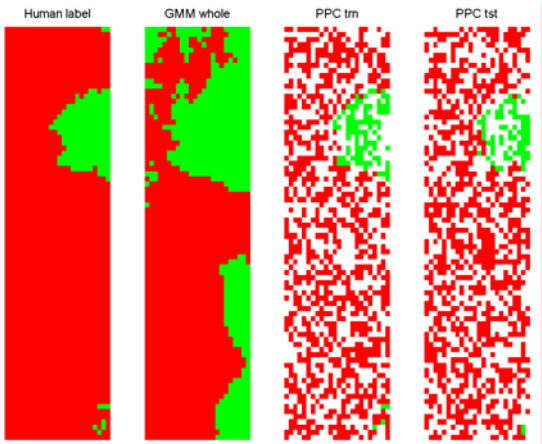
Todd K. Leen
OGI - OHSU Feb. 4, 2004

Satellite Image Data

- **Partially-labeled region**
  - **Labeled into 2** class-sets
    - Snow area: **wet snow, dry snow, melt ponds, bare ice**
    - Non-snow area: **water, clouds, bare land**

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Satellite Image - Generalization



- 50% data for training and 50% data for test
- Classification accuracy are averaged over 20 runs
- Effect of constraints in training properly generalizes to test set

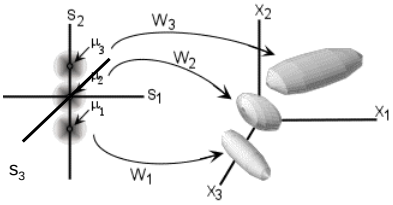
Classification accuracy	75.05%	99.23%	97.15%

Conclusion

- **Penalized Probabilistic Clustering**
  - **May be useful to bootstrap dataset labeling.**
  - **Partially-labeled datasets anyone?**

Clustering Framework

- **Clustering based on constrained Gaussian mixture models**
  - **Latent variable generative model → constraint structure related to PCA / FA**
  - **Automatically tunes to local data dimensionality**
  - **Generates entropy-penalized (e.g. regularized) cost function directly from likelihood maximization**
  - **Automatic selection of number of clusters by x-validation**



- Latent space *s*
- Maps W_i from *s* to data space *x* (fit)
- Additive noise – variance σ^2 (resolution control parameter, not fit)
- Rank(W_i) determined by data & σ^2 sets local cluster dimension

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Cost Functions

- **APCA (Adaptive Principle Components Analysis)**

$$Cost = \sum_{\alpha=1}^{M}\sum_{n=1}^{N} z(\alpha, x_n)\left(\overbrace{-2\log\pi_\alpha}^{\text{cluster selection entropy}} + \overbrace{\ln\left|\Lambda_\alpha / \sigma^2\right|}^{\text{coding entropy (in plane)}} + d\ln\sigma^2 + \overbrace{(x_n-\mu_\alpha)^T U_\alpha \Lambda_\alpha^{-1} U_\alpha^T (x_n-\mu_\alpha)}^{\text{distance in plane}} + \frac{1}{\sigma^2}\underbrace{(x_n-\mu_\alpha)^T(1-U_\alpha U_\alpha^T)(x_n-\mu_\alpha)}_{\text{distance from x to plane}} \right)$$

- **ECVQ (Entropy-Constrained Vector Quantization)**

$$C = \sum_{\alpha=1}^{M}\sum_{n=1}^{N} z(\alpha, x_n)\frac{1}{2}\left(\underbrace{-2\log\pi_\alpha}_{\text{cluster selection entropy}} + d\ln\sigma^2 + \frac{1}{\sigma^2}\underbrace{(x_n-\mu_\alpha)^T(x_n-\mu_\alpha)}_{\text{distance to mean}} \right)$$

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Incorporating Prior on Cluster Assignment

- **New complete data likelihood**

$$p(X, Z|\Theta)$$

$$p(X, Z \mid \Theta, W) = \frac{1}{K} \underbrace{p(X, Z \mid \Theta)} \underbrace{\prod_{i,j} \exp(-W(i,j) \sum_{\alpha} \left\| z(\alpha, x_i) - z(\alpha, x_j) \right\|^2)}$$

$$z(\alpha, x_i) = \begin{cases} 1, & \text{if } z_i = \alpha \\ 0, & \text{otherwise} \end{cases}$$

•W(i,j)>0, we prefer to assign x_i and x_j into same cluster – must-link

•W(i,j)<0, we prefer to assign x_i and x_j into different clusters – cannot-link

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Cluster Assignment
Posterior: standard mixture model

- **In standard GMM, W=0, the posterior that x_1 and x_2 are generated by the z_1^{th} and z_2^{th} components is**

$$p(z_1, z_2 \mid x_1, x_2, W, \Theta) = p(z_1, z_2 \mid x_1, x_2, \Theta)$$
$$= p(z_1 \mid x_1, \Theta) p(z_2 \mid x_2, \Theta)$$

- **The posterior of each sample x_i can be calculated separately as**

$$p(k \mid x_i, \Theta) = \frac{p(k, x_i \mid \Theta)}{p(x_i \mid \Theta)} = \frac{\pi_k p(x_i \mid \theta_k)}{\sum_{n=1}^{M} \pi_n p(x_i \mid \theta_n)}$$

where π_k and $p(x_i \mid \theta_k)$ are easy to calculate

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Cluster Assignment Posterior: PPC

- **The independence in assignment doesn't hold as for standard model**

$$p(z_1, z_2 \mid x_1, x_2, W, \Theta) \neq p(z_1 \mid x_1, W, \Theta) p(z_2 \mid x_2, W, \Theta)$$

- **Marginalization**

$$p(z_1 \mid x_1, W, \Theta) = \sum_{z_2} p(z_1, z_2 \mid x_1, x_2, W, \Theta)$$

- **Assume we have 20 samples set {x1,x2,…x20}, each 2 samples in that set are relevant to each other in assignment. To find the posterior of x1 to z1, we need to marginalize out x2,…x20**

$$p(z_1 \mid x_1, \Theta, W) = \sum_{z_2, \ldots z_{20}} p(\{z_1, z_2, \ldots, z_{20}\} \mid \{x_1, x_2, \ldots, x_{20}\}, \Theta, W)$$

  - **For model with M clusters, the time complexity is O(M^{20})**

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Posterior: PPC (cont'd)

- **More generally, we may have more samples relevant to each other**

- **Assume we have 20 samples set {x1,x2,…x20}, each 2 samples in that set are relevant to each other in assignment. To find the posterior of x1 to z1, we need to marginalize out x2,…x20**

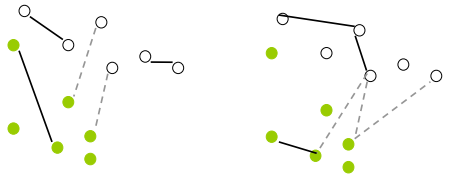
$$p(z_1 \mid x_1, \Theta, W) = \sum_{z_2, \ldots z_{20}} p(\{z_1, z_2, \ldots, z_{20}\} \mid \{x_1, x_2, \ldots, x_{20}\}, \Theta, W)$$

  - **For model with M clusters, the time complexity is O(M^{19})**

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Posterior: PPC (cont'd)

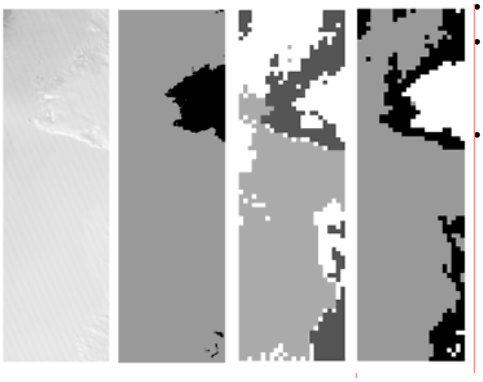
- **Here we only consider the situation where each sample can only be involved in at most one pairwise relation**



this NOT this

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Satellite Image

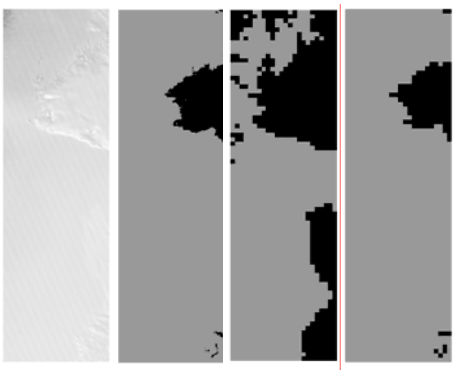


Channel 1 Label GMM PPC

- **This time we use 3-component model**
- **Here are typical runs of 3-component PPC and GMM**
- The clustering result of PPC is more consistent with label and human vision

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Satellite Image



Channel 1 Label GMM PPC

- **Model it with 2-component PPC with** only **Cannot-links**
- **Cannot-links are randomly chosen according the partial label**
- PPC works well on separating snow area from non-snow area

Todd K. Leen
OGI - OHSU Feb. 4, 2004

Features for Clustering

- **Welch-averaged power spectra of the TSA data,**
  - **select the appropriate FFT length using a qualitative bias/variance tradeoff.**
  - **combined the spectra of 6 accelerometers into a single feature vector.**
  - **Three different methods where investigated for this combination:**
    - **Concatenation without scaling.**
      - preserves frequency information, relative power between channels, total power
    - **Concatenation followed by Normalization to unit vector magnitude.**
      - preserves frequency information, relative power between channels, not total power
    - **Normalization followed by Concatenation.**
      - preserves frequency information only, no power information retained

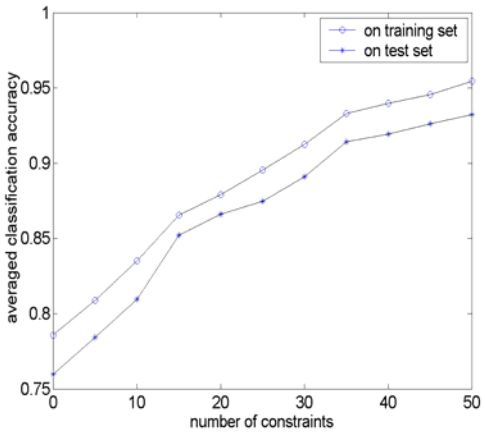
Todd K. Leen
OGI - OHSU Feb. 4, 2004

Combining Features

- **Normalize-Concatenate gave superior clustering accuracy for all three gear TSAs, and for both APCA and ECVQ. However,**
  - **normalization removes information about relative RMS power between accelerometers, as well as removing RMS differences between examples.**
  - **Cynthia reported 89% accuracy using RMS features from the 3 gear-TSAs and six accelerometers**
    - Handpicked 7 features using all gears and accelerometers.
- **So we do best by discarding RMS, even though Cynthia found it to be a useful feature!**

Fisher Iris data

Average classification accuracy vs. the # of constraints



- **150 samples, 3 classes**
- **each sample has 4 features**
- **90% data for training, 10% data for test**
- **Pairwise constraints are randomly chosen from training set**
- **Classification accuracy is used to measure the performance**
- **Result is averaged over 100 runs**
- **Effect of constraints in training properly generalizes to test set**