

Symmetric Normal Mixtures

Michael Turmon

Machine Learning Group
Caltech / Jet Propulsion Laboratory
Pasadena, CA, USA

COMPSTAT

23 August 2004

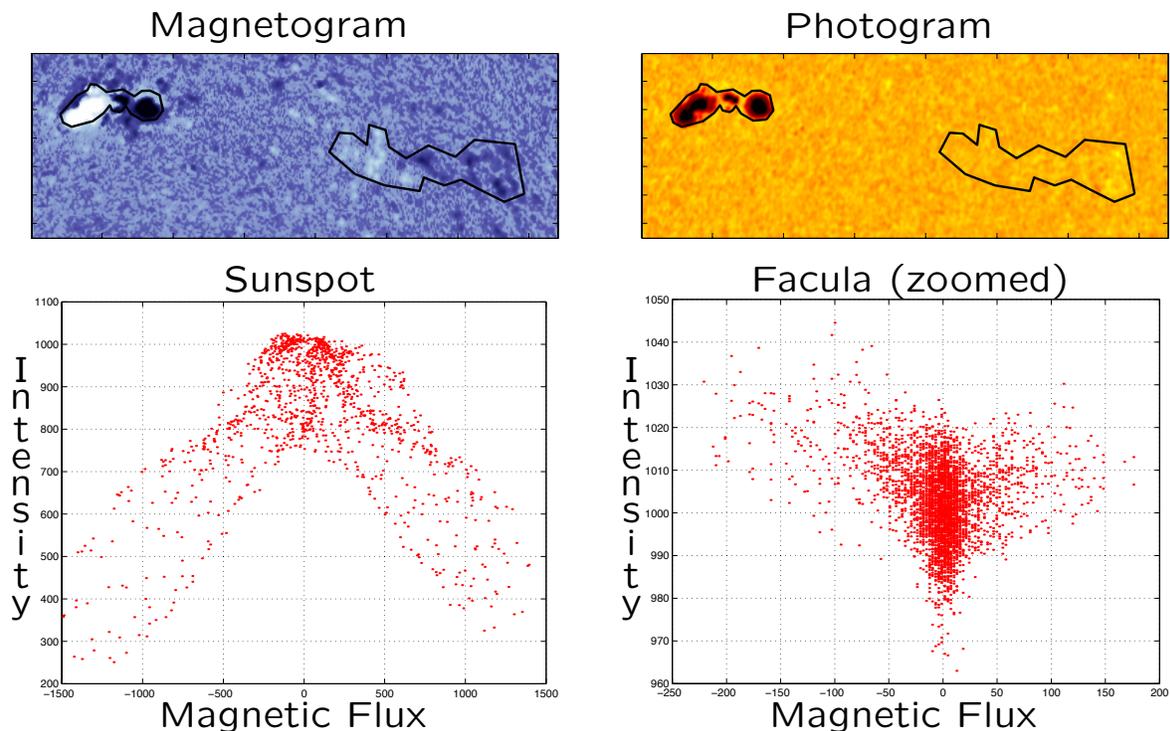
- Introduction and Motivation
- Algebraic Structure of Symmetric Mixtures
- Solution via EM Algorithm
- Examples

`turmon@aig.jpl.nasa.gov`

`http://www-aig.jpl.nasa.gov/home/turmon/`

Symmetric Data

Model and thus classify pixels (2-d feature vectors) from these images



Use **probability under a normal mixture** which has been trained on identified chips (supervised) or on pooled pixels (unsupervised)

Model should honor symmetry with respect to flipping the magnetic field

As models become more complex (more features, more pixels) the fit must be as well-constrained as possible: **find better models faster**

A Simple “Wrapper” Fails

- **A clean fix?**

1. Start the EM iteration at a symmetric model
2. Add symmetric ghost data points

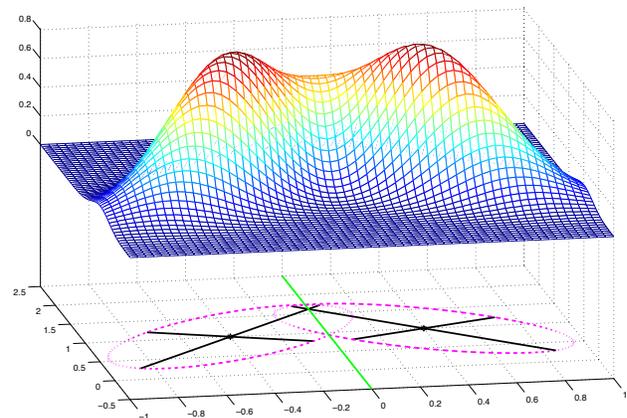
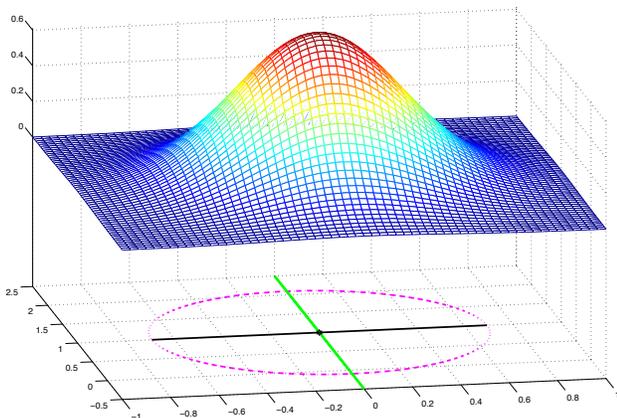
Alas, must also alter the EM algorithm internals:

3. Re-apply symmetry constraint at every iteration to patch floating-point inconsistencies

- **Structural problem arises**

EM cannot condense paired bumps into self-symmetric bumps: EM updates stagnate

So, must introduce two types of bumps:



Finally, doubling runtime motivated a closer look at symmetry in mixture density estimation

Basic Constraint: $x \stackrel{\mathcal{D}}{=} Ax$

A must be nonsingular

...else Ax would fail to have a density

...iterating shows $x \stackrel{\mathcal{D}}{=} A^p x$ for integer p

$$1 = \int p(x) dx = \int p(Ax) dx = |A|^{-1} \int p(y) dy = |A|^{-1}$$

Cyclic restriction: $A^P = I$ for some period P .

$G = \{I, A, \dots, A^{P-1}\}$ is isomorphic to the cyclic group of order P .

Some multiple symmetries are encoded by finite groups that are not cyclic

Continuous (scale) or aperiodic (translation) invariances: integrate wrt Haar measure

- **Examples**

Original example: $A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$, $P = 2$

A is a general rotation matrix: encode a variety of geometric constraints

A is a permutation: enforce within-feature-vector distributional constraints

$A = \sqrt{-1} I$: provide real-imaginary symmetry for complex x ($P = 4$)

Specialize to Normal Mixtures

- **Usual normal mixture setup**

$$p(x) = \sum_{k=0}^{K-1} \gamma_k N(x; \mu_k, \Sigma_k)$$

with $\sum_{k=0}^{K-1} \gamma_j = 1$, μ_k arbitrary, $\Sigma_k > 0$

(μ_k, Σ_k) are distinct to preserve identifiability

Choose the free parameters

$$\Theta = \{(\gamma_k, \mu_k, \Sigma_k)\}_{k=0}^{K-1}$$

using maximum likelihood

$$\Theta_{\text{ML}} = \arg \max_{\Theta \in \Theta} \log p(X; \Theta)$$

with training data $X = \{x_n\}_{n=1}^N$ and EM

- **Account for the symmetry constraint**

Symmetry of p is enforced by a modified EM iteration which we will derive

But first, we examine the effect of the constraint on the structure of the mixture model

Reinterpret the Constraint

The distributional constraint is **equivalent** to

$$(\gamma, \mu, \Sigma) \in \Theta \Rightarrow (\gamma, A\mu, A\Sigma A^T) \in \Theta \quad (*)$$

Note: When $\theta = (\gamma, \mu, \Sigma) \in \Theta$, write $A\theta$ for $(\gamma, A\mu, A\Sigma A^T)$.

A permutation π of $\{0, \dots, K-1\}$ *groups mixture components that jointly have symmetry*

π counts up, looking for the first match:

$$\pi(k) = \arg \min_{l: \theta_l = A\theta_k} (l - k) \bmod K$$

...the inverse counts down from l :

$$\pi^{-1}(l) = \arg \min_{k: \theta_l = A\theta_k} (k - l) \bmod K$$

Can show $p(x) = p(Ax)$ using π ; the reverse implication follows from L.I. of Gaussians

- **Cycles of π are key**

Recall: Describe permutations by their **cycles**, partitioning mixture components $\{0, \dots, K-1\}$

Each cycle $\mathcal{C} = (k_1, \dots, k_Q)$ is a group of mutually constrained components $(\gamma_k, \mu_k, \Sigma_k)$, $k \in \mathcal{C}$

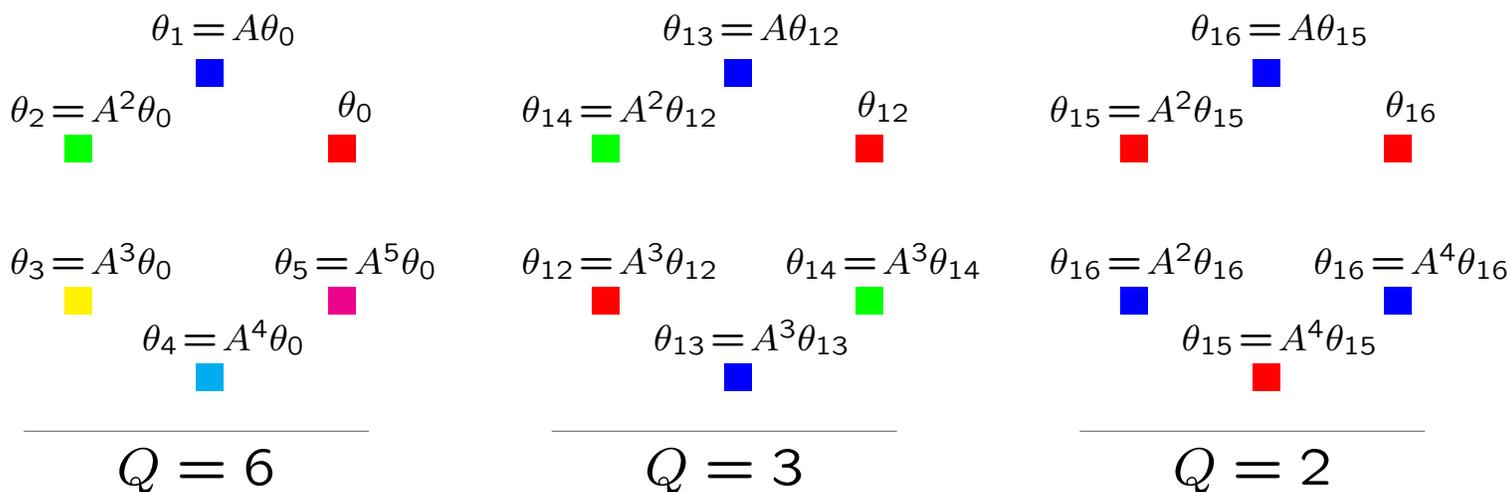
The Structure of Cycles

A single component $\theta = (\gamma, \mu, \Sigma)$ can satisfy the constraint itself: $\theta = A\theta$, $Q = 1$

A component wraps around via $Q = P$ distinct components: $\theta, A\theta, \dots, A^{P-1}\theta$ (below left)

Each intermediate component is *unconstrained*

General case: θ wraps around via $Q < P$ distinct intermediaries, each satisfying $\theta = A^Q\theta$



Above schematic shows component relationships $P = 6$ ("rotation by 60° "), three cycles shown

All P versions of θ are drawn: $\theta, A\theta, \dots, A^{P-1}\theta$

The P/Q aliases are drawn with the same color

Enforcing the Constraint

- **Specify the cycles**

Given a symmetry A with period P

Give the number of cycles for each integer Q dividing P (cycle lengths must evenly divide P)
The chosen sets \mathcal{C} partition $\{0, \dots, K-1\}$

Convention: The Q components of a cycle are numbered sequentially, say $k, k+1, \dots, k+Q-1$.

Internal constraint: $\theta_k = A^Q \theta_k$ for each $k \in \mathcal{C}$

Sharing constraint: $\theta_{k+1} = A \theta_k$ for $k, k+1 \in \mathcal{C}$

- **Enforce with Lagrange multipliers**

Enforcing $\mu - A\mu = 0$ implies a Lagrangian term
 $l_\mu = \lambda^\top (\mu - A\mu)$

Enforcing $\Sigma - A\Sigma A^\top = 0$ calls for a matrix Λ , one for each entry of $D = \Sigma - A\Sigma A^\top$:

$$\begin{aligned} l_\Sigma &= \sum_{i,j} \Lambda_{ij} D_{ij} = \text{tr } D^\top \Lambda \\ &= \text{tr}(\Sigma - A\Sigma A^\top) \Lambda = \text{tr } \Sigma (\Lambda - A\Lambda A^\top) \end{aligned}$$

Equivalently, use $l_{\Sigma^{-1}} = \text{tr } \Sigma^{-1} (\Lambda - A\Lambda A^\top)$

Related Work

- **Parameter-sharing**

Reduce the dimensionality of Θ as that of x grows

Covariances like $\sigma^2 I$ are trivial in EM

Generally: Zero-ing blocks of Σ^{-1} is simple

- **Eigendecomposition**

$\Sigma_k = \lambda_k H_k D_k H_k^T$ for orthogonal H_k

$\lambda_k D_k$ is the diagonal eigenvalue matrix; $|D_k| = 1$
(Fraley, Raftery; Celeux, Govaert)

Related speech models (Gales)

- **Mixtures of factor analyzers**

$\Sigma_k = H_k H_k^T + D_k,$

low-rank H_k and diagonal D_k

e.g., Ghahramani and Hinton

- **Symmetric Normals**

The Gaussian ($K = 1$) case with symmetry expressed as an algebraic group has been deeply elucidated by Andersson, Madsen, et al. (sufficiency, estimation, consistency, ...)

Sufficient Statistics

Rewrite the log-likelihood of component k

$$\begin{aligned} & \log |\Sigma_k| + \sum_{n=1}^N \tau_{n|k} (x_n - \mu_k)^\top \Sigma_k^{-1} (x_n - \mu_k) \\ &= \log |\Sigma_k| + (m_k - \mu_k)^\top \Sigma_k^{-1} (m_k - \mu_k) + \text{tr} \Sigma_k^{-1} S_k(m_k) \end{aligned}$$

using $\text{tr} AB = \text{tr} BA$ idiom & K sufficient stats:

$$m_k = \sum_{n=1}^N \tau_{n|k} x_n \quad S_k(m) = \sum_{n=1}^N \tau_{n|k} (x_n - m)(x_n - m)^\top$$

Account for **sharing constraint** by writing terms of one cycle $k = 0, \dots, Q - 1$ with new statistics

$$\begin{aligned} \bar{m} &= \sum_{k=0}^{Q-1} \bar{\tau}_k A^{-k} m_k \\ \bar{S} &= \sum_{k=0}^{Q-1} \bar{\tau}_k A^{-k} S_k(A^k \bar{m}) A^k \end{aligned}$$

Transform back to the θ_0 coordinates and average there since $\theta_0 = A\theta_1 = \dots = A^{Q-1}\theta_{Q-1}$

This cycle's log-likelihood, with **two terms for the internal constraint** ($\theta_0 = A^Q \theta_0$) becomes

$$\begin{aligned} & -\log |\Sigma_0| - (\bar{m} - \mu_0)^\top \Sigma_0^{-1} (\bar{m} - \mu_0) - \text{tr} \Sigma_0^{-1} \bar{S} + \\ & \quad 2\lambda^\top (\mu_0 - A^Q \mu_0) + \text{tr} \Sigma_0^{-1} (\Lambda - A^Q \Lambda A^\top) \end{aligned}$$

EM Parameter Updates

We update one cycle of Q linked components which we assume are indexed $0, \dots, Q - 1$

Differentiating to solve for (μ_0, Σ_0) gives

$$\hat{\gamma}_0 = \frac{1}{NQ} \sum_{k=0}^{Q-1} \sum_{n=1}^N \tau_{n,k}$$

$$\hat{\mu}_0 = \frac{1}{P'} \sum_{r=0}^{P'-1} A^{-Qr} \bar{m} \quad \text{where } P' = P/Q$$

$$\hat{\Sigma}_0 = \frac{1}{P'} \sum_{r=0}^{P'-1} A^{-Qr} \left[\bar{S} + (\bar{m} - \hat{\mu}_0)(\bar{m} - \hat{\mu}_0)^T \right] A^{Qr}$$

Updates $(\hat{\mu}_0, \hat{\Sigma}_0)$ are transformed repeatedly by A and used for $(\mu_1, \Sigma_1), \dots, (\mu_{Q-1}, \Sigma_{Q-1})$

(μ_0, Σ_0) are updated with a nested average of transformed sufficient statistics (m_k, S_k)

The **inner averages** (prior page) are across Q terms, one for each component in the cycle

The **outer averages**, above, sum over the symmetries in the order- P' cyclic subgroup of G to enforce invariance with respect to A^Q

Implementation

New information required:

symmetry matrix A (from which P is known)
generalized K : the number of bumps per
feasible Q , where $Q \mid P$

- **Procedure**

Standard EM finds $(m_k, \Sigma_k)_{k=0}^{K-1}$

Constrained EM follows these E and M steps with
a constraint-enforcement step

This operation loops over each cycle of
components, performing a $P = QP'$ -fold
averaging in two phases as above

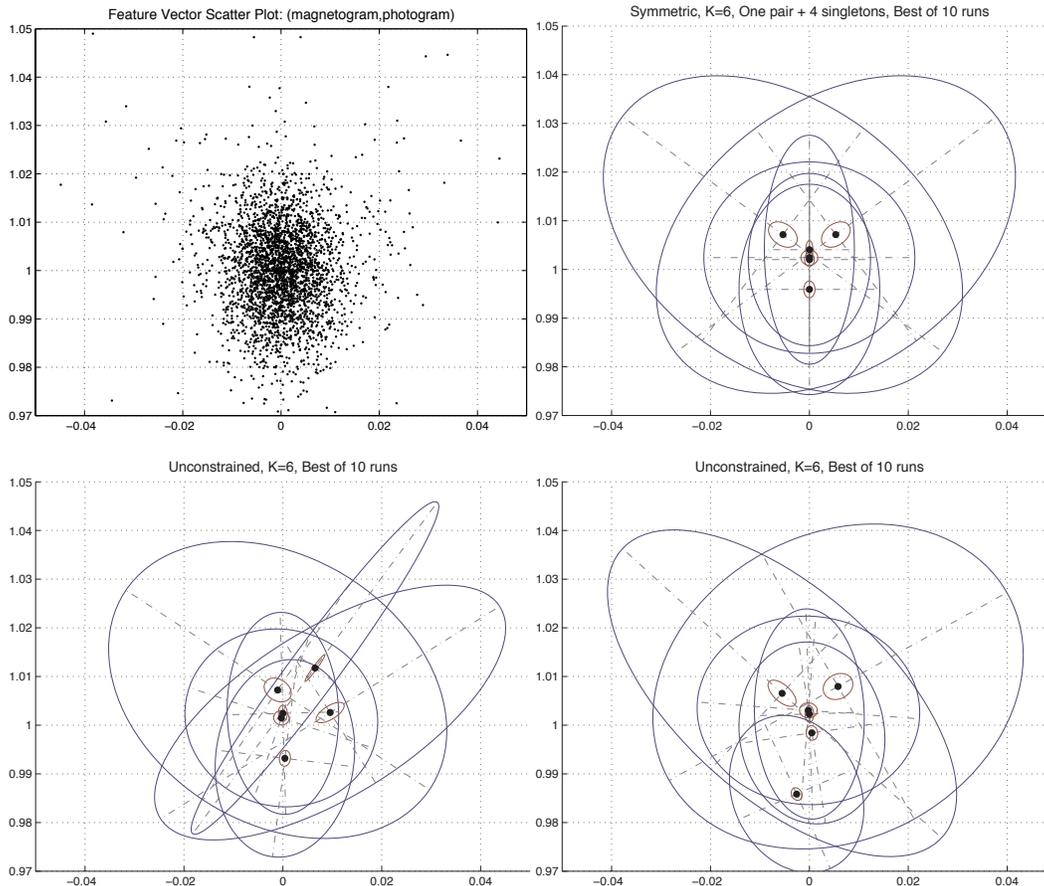
- **Computation time**

Constraining takes $O(Kd^3)$ operations, dwarfed by
the $O(NKd^3)$ in each ordinary EM step

If all cycles have $Q = P$, the constrained algorithm
is equivalent to copying each $x \in X$, P times
($x, Ax, \dots, A^{P-1}x$) plus unconstrained EM, but
requires P times less computation.

Constraint in Action

Shown: $A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$, $P = 2$, $K = 6$, $N = XXX$,
best of 10 runs



Fits are more stable

repeated runs are consistent

Fits are higher-quality

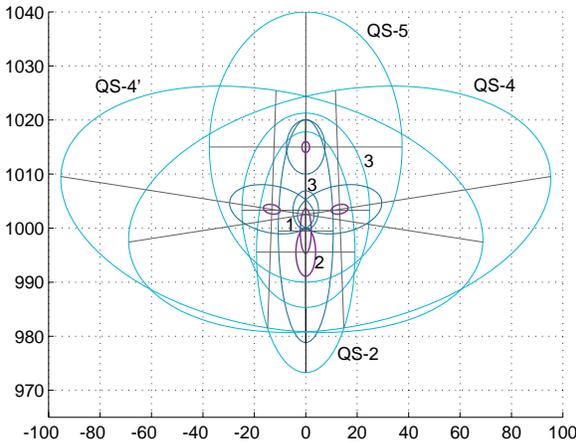
obey constraint (of course)

fewer elongated components

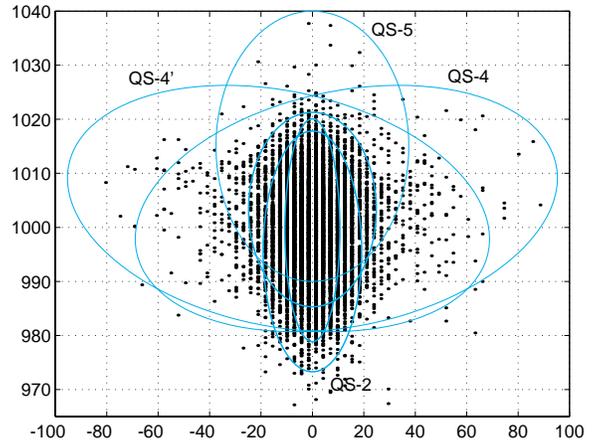
robust identification of supergranulation

Learned Clusters: SoHO/MDI

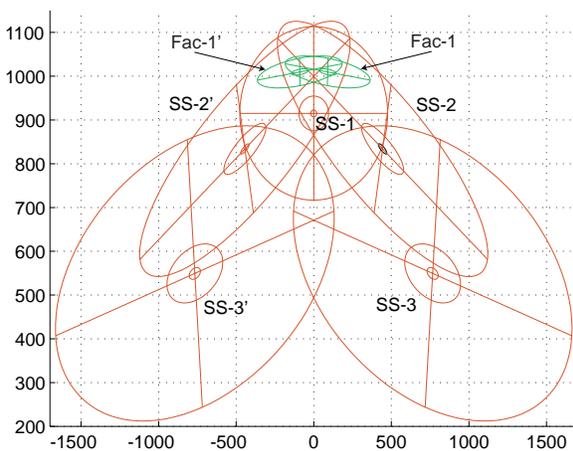
Class 1 (“Quiet”) Model



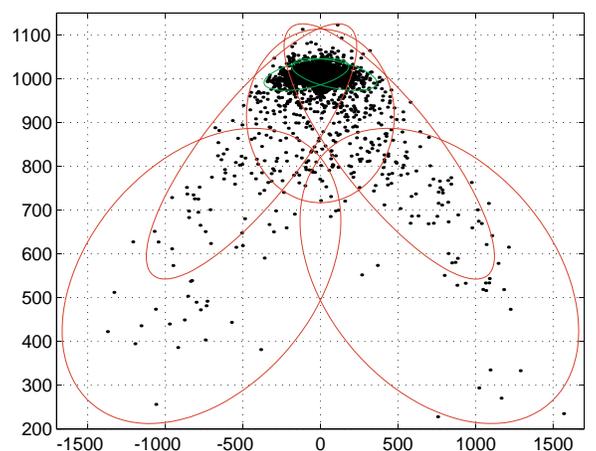
Class 1 Model Fit



Class 2 & 3 (“Active”) Models



Class 2 & 3 Model Fits



Shown:

Class 1: $K = 6$, one pair, four singletons

Classes 2, 3: $K = 7$, three pairs, one singleton

Currently:

Class 1: $K = 8$, two pairs

Classes 2, 3: $K = 20$, 8 pairs