SYMMETRIC NORMAL MIXTURES

Michael Turmon

Key words: symmetry constraint, algebraic group, EM algorithm
COMPSTAT 2004 section: clustering.

Abstract: We consider mixture density estimation under the symmetry constraint $x \stackrel{\mathcal{D}}{=} Ax$ for an orthogonal matrix A. This distributional constraint implies a corresponding constraint on the mixture parameters. Focusing on the gaussian case, we derive an expectation-maximization (EM) algorithm to enforce the constraint and show results for modeling of image feature vectors.

1 Introduction

We consider a simple constraint which captures underlying symmetry in density estimation problems. In particular, we are interested in cases where the target random variable $x \in R^d$ satisfies

$$x \stackrel{\mathcal{D}}{=} Ax \tag{1}$$

for a known linear transform A. It is immediate that A is nonsingular: otherwise Ax would concentrate in a proper subspace of R^d, and the law of x would fail to have a density with respect to Lebesgue measure on R^d. Indeed, $|A| = 1$ (writing $|A|$ for absolute value of the determinant) since

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

There are no general restrictions on A through its singular values. For example, consider for any orthonormal U the symmetry $A = U\left[\begin{smallmatrix} 0 & 2 \\ 1/2 & 0 \end{smallmatrix}\right]U^{\mathsf{T}}$. Choosing $x \sim N(0, \Sigma)$ where $\Sigma = U\left[\begin{smallmatrix} 2 & 0 \\ 0 & 1/2 \end{smallmatrix}\right]U^{\mathsf{T}}$ implies $A\Sigma A^{\mathsf{T}} = \Sigma$ so that $x \stackrel{\mathcal{D}}{=} Ax$.

Iterating (1), noting $|A| \neq 0$, shows $x \stackrel{\mathcal{D}}{=} A^p x$ for any integer p. For clarity we confine this paper to cyclic symmetries: $A^P = I$ for some period P. The set of symmetries $G = \{I, A, \ldots, A^{P-1}\}$ is then isomorphic to the cyclic group of order P. This can be relaxed in various ways. Some multiple symmetries are encoded by finite groups that are not cyclic. Also, continuous (e.g., scale) or aperiodic (e.g., translation) invariances are important in applications.

Mixture estimation problems for image data having symmetries motivated this work; see the figure on the last page. The upper left plot shows bivariate feature vectors taken from pairs of synchronized solar images from the MDI imager on the SoHO spacecraft. These densities have symmetry with respect to changing the sign of the magnetic flux, corresponding to $A = \left[\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. Similar data are gathered by other solar observatories. Taking A as a general rotation matrix can encode a variety of similar geometric constraints. Taking

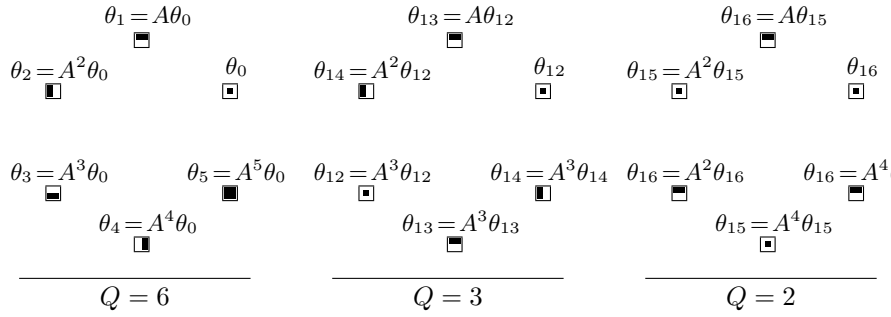


Figure 1: Schematic rendering of three cycles in a system with $P = 6$. All P versions of the component are shown; the P/Q aliases have the same markers.

A as a permutation enforces within-feature-vector distributional constraints. For complex x, $A = \sqrt{-1}I$ gives real-imaginary symmetry.

As density models for x we use finite normal mixtures [6]:

$$p(x) = \sum_{k=0}^{K-1} \gamma_k N(x;\, \mu_k, \Sigma_k) \tag{2}$$

where $\sum_{k=0}^{K-1} \gamma_j = 1$, the constituent mean vectors μ_k are arbitrary, and the covariance matrices Σ_k are symmetric positive-definite. We require that the (μ_k, Σ_k) be distinct to preserve identifiability. The free parameters

$$\Theta = \{(\gamma_k, \mu_k, \Sigma_k)\}_{k=0}^{K-1} \tag{3}$$

are chosen using training data $X = \{x_n\}_{n=1}^{N}$ and maximum likelihood:

$$\Theta_{\mathrm{ML}} = \arg\max_{\Theta \in \mathbf{\Theta}} \log p(X; \Theta) \quad . \tag{4}$$

To estimate these parameters, we use the well-known EM (Expectation-Maximization) algorithm, which leaves room to accommodate key physical constraints like (1). Constraints also ameliorate the problem of local maxima — which is especially troublesome in mixture estimation.

The sequel is organized as follows. In the next section we lay out the structure of the parameter constraints implied by the symmetry constraint in the context of normal mixtures, briefly examining related work. We then derive the EM algorithm for the general solution. Implementation issues and some representative results follow this derivation.

2 Constrained Mixture Parameters

Suppose x is governed by a normal mixture $\Theta = \{(\gamma_k, \mu_k, \Sigma_k)\}_{k=0}^{K-1}$. Then the constraint (1) is satisfied if and only if

$$(\gamma, \mu, \Sigma) \in \Theta \Rightarrow (\gamma, A\mu, A\Sigma A^{\mathsf{T}}) \in \Theta \quad . \tag{5}$$

Henceforth, for short: when $\theta = (\gamma, \mu, \Sigma) \in \Theta$, write $A\theta$ for $(\gamma, A\mu, A\Sigma A^{\mathsf{T}})$.

To see (5) suffices for (1), first note that (5) implies existence of a permutation π of $\{0, \ldots, K-1\}$ mapping mixture components according to A:

$$\pi(k) = \arg \min_{l:\theta_l = A\theta_k} (l - k) \bmod K \quad . \tag{6}$$

To see that π is a permutation, note that the set of l satisfying the condition is guaranteed to be nonempty by (5) so π is a well-defined function on $\{0, \ldots, K-1\}$. And, the inverse exists:

$$\pi^{-1}(l) = \arg \min_{k:\theta_l = A\theta_k} (k - l) \bmod K \tag{7}$$

which has the effect of counting down from l, looking for the first matching parameter tuple, while π counts up. Now π and $|A| = 1$ establish (1):

$$p(Ax) = \sum_{k=0}^{K-1} \gamma_k N(Ax;\, \mu_k, \Sigma_k) = \sum_{k=0}^{K-1} \gamma_{\pi(k)} N(x;\, \mu_{\pi(k)}, \Sigma_{\pi(k)}) = p(x) \quad .$$

The reverse implication, which is not so important for our purposes, follows from the linear independence of Gaussian functions [8].

The domain of π can be partitioned into cycles, each of the form $\mathcal{C} = (k_1, \ldots, k_Q)$ for some length Q. Cycles are the minimal subsets of the domain which are fixed by the permutation: $\pi(k_i) = k_{i+1}$ and $\pi(k_Q) = k_1$. Listing the cycles of π uniquely determines and succinctly describes its structure. This decomposition will prove key to compactly specifying the form of the mixture to be fit to X, e.g. section 4.

The cycles correspond to structural properties of the mixture. They partition the components, so write $[k]$ for the equivalence class of bump k under π. For instance, a component θ_k might itself satisfy $A\theta_k = \theta_k$, and $\pi(k) = k$: a cycle of length $Q = 1$. At the other end, a chain of $Q = P$ intermediate components, each having no symmetry properties, lead back to θ_k. Such a group is shown at left in figure 1, which takes $P = Q = 6$ and schematically represents application of A to some θ_k as rotation by 60°, and distinct components θ_l, $l \in [k]$ as different markers. (The figure shows them in sequence, although that is not true in general.) Cycles of length $Q > P$ cannot occur: otherwise, both θ and $\theta' = A^P\theta$ would exist as distinct members of Θ. Since $A^P = I$, this would violate identifiability.

More generally, cycles of $1 \leq Q \leq P$ components occur if and only if $Q \mid P$ (i.e., Q divides P). The middle panel of the figure shows the $Q = 3$ case where $\mathcal{C} = (12, 13, 14)$; there are only three distinct markers because $\theta_{12} = A^3\theta_{12}$. At right, $Q = 2$ and $\mathcal{C} = (15, 16)$. These diagrams illustrate why Q must divide P. Formally, this is just Lagrange's theorem applied to the cyclic group of order P: all its subgroups are cyclic and of an order dividing P.

Cycle	Mixture Indexes	Q	P'	Internal Constraint	Shared Parameter Constraint
1	0–5	6	1	none: $A^6 = I$	$\theta_5 = A\theta_4 = \cdots = A^5\theta_0$
2	6–11	6	1	none: $A^6 = I$	$\theta_{11} = A\theta_{10} = \cdots = A^5\theta_6$
3	12–14	3	2	$\theta_{12} = A^3\theta_{12}$	$\theta_{14} = A\theta_{13} = A^2\theta_{12}$
4	15–16	2	3	$\theta_{15} = A^2\theta_{15}$	$\theta_{16} = A\theta_{15}$
5	17–18	2	3	$\theta_{17} = A^2\theta_{17}$	$\theta_{18} = A\theta_{17}$
6	19	1	6	$\theta_{19} = A\theta_{19}$	—

Within this restriction, many component structures may coexist in Θ; we establish conventions for their ordering. A K-bump mixture corresponds to a tuple K_s, with entries summing to K, each giving the number of mixture components devoted to cycles of each possible length Q such that $Q \mid P$. For instance, if $P = 6$, a symmetry of $K_s = (12, 3, 4, 1)$ implies $K = 20$ and

$$\pi = ((0, 1, 2, 3, 4, 5)(6, 7, 8, 9, 10, 11)(12, 13, 14)(15, 16)(17, 18)(19)).$$

The table above itemizes the parameters, and figure 1 shows parameters corresponding to the first, third, and fourth cycles of π.

Suppose a given cycle contains Q components. In the conventional ordering, the components have a *shared parameter constraint*

$$\theta_{k+1} \equiv A\theta_k, \ldots, \theta_{k+Q-1} \equiv A^{Q-1}\theta_k \quad . \tag{8a}$$

Furthermore, each component also satisfies an *internal constraint*

$$(\forall l \in [k])\, \theta_l = A^Q\theta_l \quad . \tag{8b}$$

We will use Lagrange multipliers to enforce (8b). The Lagrangian term for $\mu = A\mu$ is $l_\mu = \lambda^\mathsf{T}(\mu - A\mu)$ for a vector λ to be determined. Enforcing $\Sigma = A\Sigma A^\mathsf{T}$ calls for a matrix Λ, one for each entry of $D = \Sigma - A\Sigma A^\mathsf{T}$:

$$l_\Sigma = \sum\nolimits_{i,j} \Lambda_{ij} D_{ij} = \operatorname{tr} D^\mathsf{T}\Lambda = \operatorname{tr}(\Sigma - A\Sigma A^\mathsf{T})\Lambda = \operatorname{tr}\Sigma(\Lambda - A\Lambda A^\mathsf{T}) \tag{9}$$

where we have used $\Sigma = \Sigma^\mathsf{T}$ and the trace identity, $\operatorname{tr} AB = \operatorname{tr} BA$. The constraint on Σ is equivalent to the same constraint on Σ^{-1}, so we use instead the more convenient $l_{\Sigma^{-1}} = \operatorname{tr}\Sigma^{-1}(\Lambda - A\Lambda A^\mathsf{T})$.

Earlier work on constrained mixtures imposes structure to compactly parameterize the covariance. Some structured covariances (e.g., $\Sigma_k = \sigma_k^2 I$) can be trivially handled in the EM algorithm. This idea has been extended using the eigendecomposition $\Sigma_k = \lambda_k H_k D_k H_k^\mathsf{T}$ where the H_k are orthogonal, $\lambda_k D_k$ is the diagonal eigenvalue matrix, and $|D_k| = 1$; a family of EM algorithms results [2, 3] from various parameter-sharing schemes. A "semi-tied" covariance model has been used in output modeling for hidden Markov models (HMMs) [4]. This parameterizes a subset $\mathcal{K} \subset \{0, \ldots, K-1\}$ of covariances by sharing H. Other subsets $\mathcal{K}'$ could have different structuring

matrices H. Mixtures of factor analyzers [5] are another twist: covariance models of the form $\Sigma_k = H_k H_k^{\mathsf{T}} + D_k$, with low-rank H_k and diagonal D_k. The constraints we consider give rather different structure to the covariance, and affect the means and weights as well. The structure imposed on the Gaussian distribution (i.e., $K = 1$) by symmetry expressed as an algebraic group has been deeply elucidated [1, App. A]. For concreteness, we have specialized in this paper to the cyclic group, while treating the more general class of K-component mixtures with a more computational viewpoint.

3 Normal Mixture Solution

Following the standard approach to fitting a mixture distribution via EM (e.g., [6, sec. 3.2]), define for each x_n a corresponding sequence of indicator variables $Z_n = (z_{n,0}, \ldots, z_{n,K-1})$. Exactly one of these indicators equals one, signaling which component of (2) generated x_n. We correspondingly denote $Z = \{Z_n\}_{n=1}^N$, and the pair (X, Z) becomes the complete-data of the EM algorithm. The log probability of the complete-data decouples as

$$\log p(X, Z) = \sum_{k=0}^{K-1} \sum_{n=1}^{N} z_{n,k} \log[\gamma_k N(x_n;\, \mu_k, \Sigma_k)]$$

and its expectation given the observation is

$$Q(\Theta) = E[\log p(X, Z) \,|\, X] = \sum_{k=0}^{K-1} \sum_{n=1}^{N} \tau_{n,k} \log[\gamma_k N(x_n;\, \mu_k, \Sigma_k)] \tag{10}$$

where the weights are regarded as known:

$$\tau_{n,k} := E[z_{n,k} \,|\, x_n] = \gamma_k N(x_n; \mu_k, \Sigma_k) \Big/ \sum_{l=0}^{K-1} \gamma_l N(x_n; \mu_l, \Sigma_l) \quad . \tag{11}$$

The quantity $\tau_{n,k}/N$ is a joint pmf. It is convenient to also define $\tau_k = \sum_{n=1}^N \tau_{n,k}$ and $\tau_{n|k} = \tau_{n,k}/\tau_k$. The latter is a correctly normalized conditional distribution. We maximize $Q(\Theta)$ at every EM iteration to update the parameters. We use the parameter ordering convention described above.

The update for the weights can be derived separately because the terms of Q involving γ_k separate out. Including the Lagrangian term for the unit mass constraint on the weights, the function to be maximized is

$$Q_C(\gamma_0, \ldots, \gamma_{K-1}) = \sum\nolimits_{k=0}^{K-1} \tau_k \log \gamma_k + \lambda\Big(1 - \sum\nolimits_{k=0}^{K-1} \gamma_k\Big) \quad .$$

To find γ_k, recall from (8a) that all of the weights γ_l, $l \in [k]$, are in fact the same parameter. Differentiating reveals the optimal weight is

$$\hat{\gamma}_k = \big(1/\#[k]\big) \sum\nolimits_{l \in [k]} \tau_l / N \tag{12}$$

where $\#[k]$ is the cardinality of the cycle. This is just the average class-membership in the cycle containing k, normalized to sum to unity.

Using the trace identity, the terms of (10) involving means and covariances are conventionally written via the weighted sufficient statistics:

$$Q(\mu_0, \ldots, \mu_{K-1}, \Sigma_0, \ldots, \Sigma_{K-1}) =$$
$$-\frac{1}{2} \sum_{k=0}^{K-1} \tau_k \left[\log|\Sigma_k| + (m_k - \mu_k)^{\mathsf{T}} \Sigma_k^{-1} (m_k - \mu_k) + \operatorname{tr} \Sigma_k^{-1} S_k(m_k) \right] \quad (13)$$
$$m_k := \sum_{n=1}^{N} \tau_{n|k} x_n \quad \text{and} \quad S_k(\eta) := \sum_{n=1}^{N} \tau_{n|k} (x_n - \eta)(x_n - \eta)^{\mathsf{T}} \quad . \quad (14)$$

The k subscript indicates weighting by the conditional probabilities $\tau_{n|k}$.

It is immediate from the sum in (13) that, in the usual unconstrained mixture problem, parameter updates for $(\hat{\mu}_k, \hat{\Sigma}_k)$ decouple across k. In the constrained case, differentiating with respect to μ_k or Σ_k will involve all components in $[k]$, but no others: components within a cycle are tied via (8a). In the remainder of this section, we suppose the cycle is indexed as $[k] = \{0, \ldots, Q-1\}$ to cut down on superfluous notation.

To enforce the shared parameter constraint (8a), let μ_0 be a free parameter and write $\mu_l = A^l \mu_0$, $0 < l < Q$, and similarly for the covariances. Use the Lagrangian mechanism to account for the internal constraint (8b), namely

$$\mu_l = A^Q \mu_l, \quad \Sigma_l = A^Q \Sigma_l A^{\mathsf{T}Q}, \quad 0 \le l < Q, \quad (15)$$

which is of course accomplished by constraining (μ_0, Σ_0) only. With this way of writing the parameters, the cycle-k terms of (13) are

$$Q(\mu_0, \Sigma_0) = -\frac{\tau_{[0]}}{2} \sum_{k=0}^{Q-1} \bar{\tau}_k \left[\log|\Sigma_0| + (A^{\mathsf{T}k} m_k - \mu_0)^{\mathsf{T}} \Sigma_0^{-1} (A^{\mathsf{T}k} m_k - \mu_0) + \right.$$
$$\left. \operatorname{tr} \Sigma_0^{-1} A^{\mathsf{T}k} S_k(m_k) A^k \right] \quad (16)$$

where $\tau_{[0]} := \sum_{k=0}^{Q-1} \tau_k$ and $\bar{\tau}_k = \tau_k / \tau_{[0]}$, a pmf on $\{0, \ldots, Q-1\}$.

Collapsing Q parameters to one makes, e.g., $m_0, \ldots, m_{Q-1}$ informative about μ_0. It aids understanding to write (16) with new sufficient statistics

$$\bar{m} := \sum_{k=0}^{Q-1} \bar{\tau}_k A^{\mathsf{T}k} m_k \quad \text{and} \quad \bar{S} := \sum_{k=0}^{Q-1} \bar{\tau}_k A^{\mathsf{T}k} S_k(A^k \bar{m}) A^k \quad . \quad (17)$$

Intuitively, the cycle's statistics are transformed back to the (μ_0, Σ_0) coordinates and averaged there. Formally, $\bar{m}$ arises by completing the square in the quadratic form involving μ_0 in (16). With this definition, and including Lagrangian terms, the objective function simplifies to

$$Q_C(\mu_0, \Sigma_0) = -\log|\Sigma_0| - (\bar{m} - \mu_0)^{\mathsf{T}} \Sigma_0^{-1} (\bar{m} - \mu_0) - \operatorname{tr} \Sigma_0^{-1} \bar{S} +$$
$$2\lambda^{\mathsf{T}} (\mu_0 - A^Q \mu_0) + \operatorname{tr} \Sigma_0^{-1} (\Lambda - A^Q \Lambda A^{\mathsf{T}Q}) \quad (18)$$

Differentiating with respect to μ_0 gives the necessary condition

$$\hat{\mu}_0 = \bar{m} + \Sigma_0 (I - A^Q)^\mathsf{T} \lambda$$

To satisfy the constraint, note that the average of $P' = P/Q$ transformed means $\hat{\mu}_0, A^{\mathsf{T}Q}\hat{\mu}_0, \ldots, A^{\mathsf{T}(P'-1)Q}\hat{\mu}_0$ telescopes to:

$$\hat{\mu}_0 = (1/P') \sum\nolimits_{r=0}^{P'-1} A^{\mathsf{T}Qr} \bar{m} \quad . \tag{19}$$

Substituting $\hat{\mu}_0$ into the Lagrangian (18) and differentiating with respect to the elements of Σ_0^{-1} reveals a necessary condition

$$\hat{\Sigma}_0 - \bar{S} - (\bar{m} - \hat{\mu}_0)(\bar{m} - \hat{\mu}_0)^\mathsf{T} + (\Lambda - A^Q \Lambda A^{\mathsf{T}Q}) = 0$$

Enforcing the constraint with the averaging method reveals

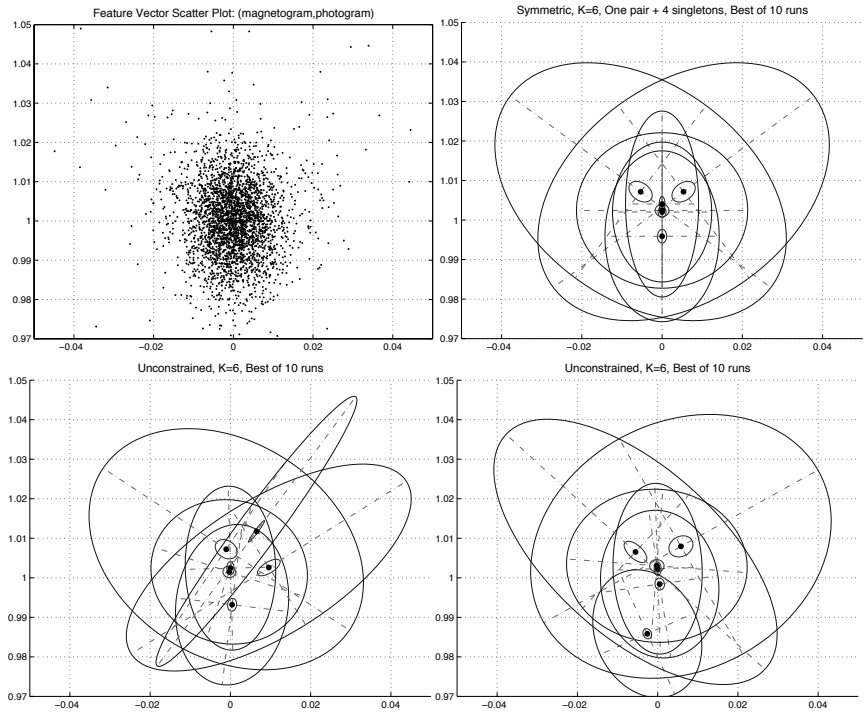
$$\hat{\Sigma}_0 = (1/P') \sum\nolimits_{r=0}^{P'-1} A^{\mathsf{T}Qr} \left[\bar{S} + (\bar{m} - \hat{\mu}_0)(\bar{m} - \hat{\mu}_0)^\mathsf{T} \right] A^{Qr} \quad ; \tag{20}$$

compare [1, Thm. A.2] for the $K = Q = 1$ case. To sum up, the parameters are updated with a nested average of transformed sufficient statistics. The inner averages (17) are across Q terms, one for each linked component in the cycle. The outer averages, in (19) and (20), sum over the symmetries in the order-P' cyclic subgroup of G to enforce invariance with respect to A^Q.

4 Implementation and Results

The new information needed is A and the symmetry vector K_s giving π: how many bumps to allocate to each symmetric configuration. (Unconstrained EM has $K_s = K$, $A = I$.) Standard EM finds $(m_k, \Sigma_k)_{k=0}^{K-1}$ as in (14). The new procedure follows these E and M steps with a constraint step which loops over each cycle of π, performing a $P = QP'$-fold averaging as in (17), (19), and (20). This 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 $(x, Ax, ..., A^{P-1}x)$ plus unconstrained EM, but requires P times less computation.

On the next page we compare unconstrained versus constrained methods with $K = 6$ on $N = 15032$ feature vectors from MDI images (top left). Each run selects the highest-likelihood model after ten, 1000-update EM sequences. The unconstrained models are unstable from run to run; the bottom panels show concentration ellipses and centers of two typical best-of-ten models. The constrained model (top right) uses $K_s = (2, 4)$. It does not have run-to-run instability, and its decomposition provides interpretable information: the symmetric pair is due to the chromospheric network, a small brightening distributed across the solar disk. One-bump cycles (i.e., $Q = 1$) are needed: models with $K_s = (K, 0)$ do not coalesce paired bumps and converge very slowly to inferior models. Three similar mixtures with K_s of $(4, 4)$, $(12, 2)$, and $(4, 2)$ are used operationally to identify three types of solar activity [7]. The constraint proved essential to estimate these more complex models.



References

[1] S. Andersson & J. Madsen. Symmetry & lattice conditional independence in a multivariate normal distribution. *Ann. Statist.*, 26(2):525–72, 1998.

[2] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–93, 1995.

[3] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *JASA*, 97:611–631, 2002.

[4] M. J. F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Trans. Speech and Audio Processing*, 7(3):272–281, 1999.

[5] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, U. of Toronto, 1997.

[6] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

[7] M. Turmon, J. Pap, & S. Mukhtar. Statistical pattern recognition for labeling solar active regions. *Astrophysical Journal*, 568(1):396–407, 2002.

[8] S. J. Yakowitz and J. D. Spregins. On the identifiability of finite mixtures. *Ann. Math. Statist.*, 39:209–214, 1968.

Acknowledgement: This research was carried out for the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA.
Address: M/S 126-347, Jet Propulsion Laboratory, Pasadena, CA 91109
E-mail: `turmon@jpl.nasa.gov`