Symmetry Constraints within the EM Algorithm for Gaussian Mixtures

Michael Turmon
Data Understanding Systems Group
Jet Propulsion Laboratory
Pasadena, CA 91109

May, 2002

Abstract

We consider mixture density estimation problems where physical knowledge indicates that a symmetry constraint of the form $x \overset{\mathcal{D}}{=} Ax$ holds for an orthogonal linear transform A. This symmetry constraint implies a corresponding constraint on the mixture density parameters. Focusing on the gaussian case, we propose an adaptation of the usual expectation-maximization (EM) algorithm which incorporates the constraint. Implementation is discussed and examples are given.

1 Introduction

Science data analysis involves problems about which much prior physical knowledge typically exists. Often this knowledge is to a significant degree uncertain; if the analysis is performed in a probabilistic framework there are various ways to incorporate such uncertain knowledge. Most often, a Bayesian framework models the uncertainty as satisfying the postulates of conventional mathematical probability, therefore finding expression in terms of a prior probability distribution. In this paper we consider quite certain physical knowledge which may be handled as a hard constraint on model parameters rather than via probabilistic preference as in the Bayesian setting.

We here 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

$$x \overset{\mathcal{D}}{=} Ax \qquad (1)$$

for some linear transform A. It is immediate that A is nonsingular: otherwise Ax would be concentrated in a proper subspace of R^d, and the law of x would fail to have a density function with respect to Lebesgue measure on R^d. Furthermore, we know $|A| = 1$ since

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

There are in general no further restrictions on A, e.g. through its singular values. For example, consider for some orthonormal U the symmetry matrix

$$A = U \begin{bmatrix} 0 & 2 \\ 1/2 & 0 \end{bmatrix} U^{\mathsf{T}} \quad .$$

Choosing $x \sim N(0, \Sigma)$ where

$$\Sigma = U \begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix} U^{\mathsf{T}} \implies A \Sigma A^{\mathsf{T}} = \Sigma \quad ,$$

so that $x \overset{\mathcal{D}}{=} Ax$. In algebraic terms, $A \in SL(d, R)$, the special linear group over R^d.

The data in figure 1 originally motivated us. We show bivariate feature vectors taken from a pair of synchronized solar images. The plots in the upper panels are from the Michelson Doppler Imager (MDI) on the SoHO spacecraft, and they show a symmetry of the density with respect to changing the sign of the magnetic flux, corresponding to

$$A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

The lower panels show similar data from Mt. Wilson Observatory (MWO), and exhibit the same symmetry. Especially in cases with limited training data, it is important to constrain the fitted model as much as possible.

To model the densities, we have employed the class of finite normal mixture distributions [MP00a] of the form

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

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. By letting K increase, the underlying distribution may be fit more exactly. The free parameters

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

are chosen using training data

$$X = \{x_n\}_{n=1}^{N} \tag{4}$$

and the maximum likelihood criterion:

$$\theta_{\text{ML}} = \arg\max_{\theta \in \Theta} \log p(X; \theta) \quad . \tag{5}$$

When $K > 1$, there is no longer a closed-form solution for these parameters so they must be estimated numerically. We use the well-known EM

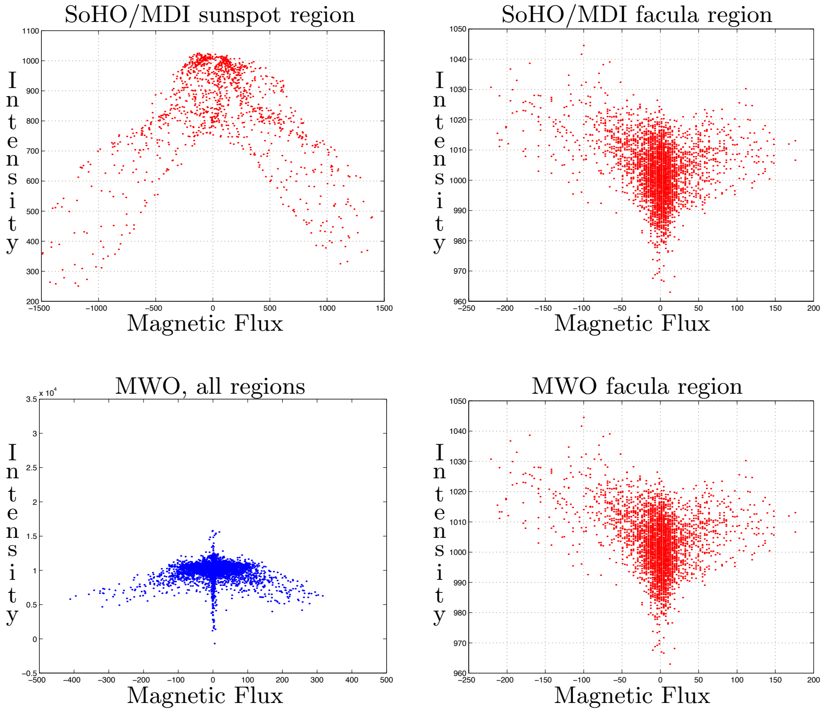

Figure 1: Synchronized magnetogram and photogram, with scatter plots of feature vectors in various regions.

(Expectation-Maximization) algorithm [MK97, sec. 2.7], which has the virtue of easily accommodating constraints like (1).

The sequel is organized as follows. In the next section we review the structure of the parameter constraints implied by the symmetry constraint in the context of normal mixtures. Then we derive the single-component ($K = 1$) solution because it is of independent interest, and it contains most of the elements of the general solution, which we discuss in the following section. Implementation issues and some representative results follow this derivation.

2 Constrained Mixture Parameters

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

$$(\gamma, \mu, \Sigma) \in \theta \Rightarrow (\gamma, A\mu, A\Sigma A^{\mathsf{T}}) \in \theta \quad . \tag{6}$$

To see this, first note that the condition (6) implies the existence of a permutation π of $\{0, \ldots, K-1\}$ which maps the components according to the transformation A:

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

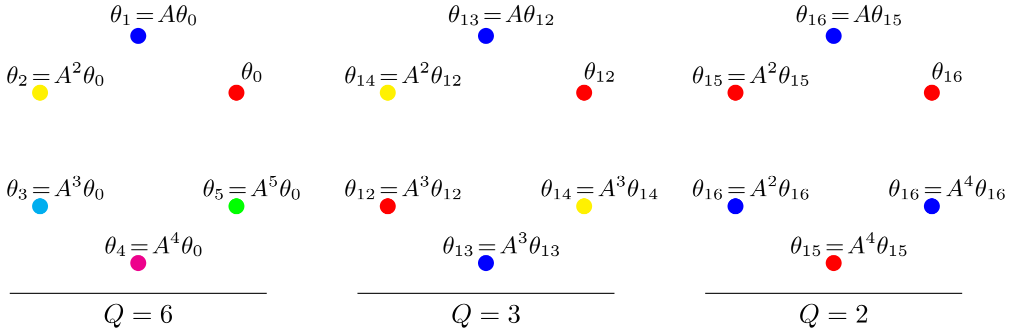
$\theta_1 = A\theta_0$
$\theta_2 = A^2\theta_0$
θ_0
$\theta_3 = A^3\theta_0$
$\theta_5 = A^5\theta_0$
$\theta_4 = A^4\theta_0$
$Q = 6$
$\theta_{13} = A\theta_{12}$
$\theta_{14} = A^2\theta_{12}$
θ_{12}
$\theta_{12} = A^3\theta_{12}$
$\theta_{14} = A^3\theta_{14}$
$\theta_{13} = A^3\theta_{13}$
$Q = 3$
$\theta_{16} = A\theta_{15}$
$\theta_{15} = A^2\theta_{15}$
θ_{16}
$\theta_{16} = A^2\theta_{16}$
$\theta_{16} = A^4\theta_{16}$
$\theta_{15} = A^4\theta_{15}$
$Q = 2$

Figure 2: A schematic parameter diagram shows three cycles in a symmetric system of period $P = 6$. All P versions of the component are shown on the diagram; the P/Q repeat versions are shown in the same color.

To see that π is a permutation, note that the set of l satisfying the condition is guaranteed to be nonempty by (6) so π is a well-defined function on $\{0, \ldots, K-1\}$. Furthermore,

$$\pi^{-1}(l) = \min_{k:\theta_l = A\theta_k} (k - l) \bmod K \tag{8}$$

which has the effect of counting down from l, looking for the first matching parameter tuple, while π counts up.

Returning to the proposition: if (6) holds,

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

where π is as above, showing that (6) is sufficient for (1).

Let us examine this structure more closely. The domain of any permutation 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 $\pi(k_Q) = k_1$. The permutation is uniquely determined, and succinctly described, by listing its cycles.

The cycles of π above correspond to structural properties of the mixture. Because the cycles partition the components, we use the standard notation $[k]$ for the equivalence class of component k under π. For instance, a component θ_k might itself satisfy the constraint, and $\pi(k) = k$: a cycle of length $Q = 1$. At the other extreme, a chain of $Q = P$ intermediate components, each of which in itself has no symmetry properties, might be needed to lead around to θ_k. Such a group is illustrated in the left panel of figure 2, 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-colored dots. (The figure shows them in sequence, although that is not true in general.) Note that cycles of length $Q > P$ cannot occur. If such a cycle existed, both $\theta_{k(1)}$ and $\theta_{k(P)} = A^P\theta_{k(1)}$ would exist in θ. But by periodicity of A, the latter equals $\theta_{k(1)}$, and the mixture is not identifiable, which is a possibility we exclude.

More generally, cycles of $1 \le Q \le P$ components are possible so long as Q divides P (written $Q \mid P$). The middle panel of the figure shows

Cycle	Mixture Indexes	Q	P'	Internal Constraint	External 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	$\theta_{19} = A\theta_{19}$	—

Table 1: An example of parameter groups and constraints. Here $P = 6$, $K = 20$, and $K_s = (12, 3, 4, 1)$.

the $Q = 3$ case where $\mathcal{C} = (12, 13, 14)$; here there are only three distinct colors because $\theta_{12} = A^3\theta_{12}$. The right panel shows $Q = 2$ and $\mathcal{C} = (15, 16)$. These diagrams illustrate why it is necessary that $Q \mid P$. Suppose $Q < P$ and note that $A^Q\theta_{k(1)} = \theta_{k(1)}$, which implies $A^{pQ}\theta_{k(1)} = \theta_{k(1)}$ for any integer p. Fix in particular the smallest p such that $pQ \geq P$; if $Q \not| P$ then $0 < pQ - P < Q$. Since $A^P\theta_{k(1)} = \theta_{k(1)}$, we have that $A^{pQ-P}\theta_{k(1)} = \theta_{k(1)}$. But this contradicts the minimality of Q as being the smallest integer l such that $A^l\theta_{k(1)} = \theta_{k(1)}$.

Within these restrictions, many component structures may co-exist in a given component-list θ. Since the ordering of the mixture components is immaterial to the distribution, it is convenient to establish conventions for the parameter organization. In the symmetric case, a K-component mixture corresponds to a vector K_s, with components 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 that $K = 20$ and

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

The corresponding parameter set consists of six cycles of parameters; see table 1. Figure 2 shows parameters corresponding to the first, third, and fourth groups of π. Of course, K_s can be recovered from θ, but the explicit notation is helpful. To sum up, suppose a given cycle contains Q constituents. In the standard ordering, all cycle parameters are related via

$$\theta_k, \theta_{k+1} \equiv A\theta_k, \ldots, \theta_{k+Q-1} \equiv A^{Q-1}\theta_k \quad ; \tag{9a}$$

each component also satisfies an internal constraint

$$(\forall l \in [k])\, \theta_l = A^Q\theta_l \quad . \tag{9b}$$

Earlier work has examined related constraints, often viewing the imposed structure as a way to achieve a compact parameterization of the covariance matrix — the most profligate consumer of degrees of freedom in high-dimensional mixtures. It is well-known that various kinds of sparse covariances (e.g., $\Sigma_k = \sigma_k^2 I$) can be accommodated in the framework of the EM algorithm. The idea of mixtures of factor analyzers [RT82, GH96, MP00b] is related, and amounts to covariance models of the form $\Sigma_k = H_k H_k^\mathsf{T} + D_k$, with H_k having relatively few columns and D_k diagonal. In this model, parameters are not shared among classes. Related multilevel models [LP98] do share some parameters. A "semi-tied" covariance model has been used in state-dependent output modeling for hidden Markov models (HMMs) [Gal99]. This parameterizes a subset $\mathcal{K} \subset \{0, \ldots, K-1\}$ of the component covariance matrices using common eigenvectors: for $k \in \mathcal{K}$, $\Sigma_k = H D_k H^\mathsf{T}$ where D_k is a component-specific diagonal matrix. Other component subsets $\mathcal{K}'$ could have different structuring matrices H, and all the parameters are estimated via an EM algorithm. Remarks on choosing an optimal basis H are in [Gop98].

3 Single Component Solution

The solution for a normal distribution is worth deriving on its own because it contains all the elements of the general case, presented in the next section, except the iterative re-estimation inherent in the EM algorithm. This model can be viewed as a mixture with $K = 1$, having one cycle with $Q = 1$ that obeys the constraint $\theta_0 = A\theta_0$, and by extension $\theta_0 = A^p\theta_0$ for any integer p. The likelihood is

$$\begin{aligned} l_1(\mu, \Sigma) = \log p(X) &= \sum_{n=1}^{N} \log N(x_n;\, \mu, \Sigma) \\ &= -(Nd/2)\log 2\pi - (N/2)\,|\Sigma| - \sum_{n=1}^{N} (x_n - \mu)^\mathsf{T} \Sigma^{-1} (x_n - \mu)/2 \end{aligned}$$

where $N(\cdot)$ is the gaussian density. Well-known manipulations using the "trace identity," $\operatorname{tr} AB = \operatorname{tr} BA$ (see the Appendix), yield

$$l_1(\mu, \Sigma) = -(N/2)[\kappa + \log|\Sigma| + (m-\mu)^\mathsf{T}\Sigma^{-1}(m-\mu) + \operatorname{tr}\Sigma^{-1}S(m)] \quad (10)$$

where $\kappa = d \log 2\pi$ and the sufficient statistics

$$m := N^{-1} \sum_{n=1}^{N} x_n \quad (11a)$$

$$S(\eta) := N^{-1} \sum_{n=1}^{N} (x_n - \eta)(x_n - \eta)^\mathsf{T} = S(m) + (m - \eta)(m - \eta)^\mathsf{T} \quad . \quad (11b)$$

The sample covariance is parameterized by an offset η; typically either $S(m)$ or $S(0)$ is computed. The expression (10) incidentally allows elegant derivations of the unconstrained maximum likelihood estimate (MLE). Indeed, $\Sigma > 0$ implies $\Sigma^{-1} > 0$, so the log-likelihood is quadratic in μ, and $\hat{\mu} = m$. Inserting into l_1 leaves two terms depending on Σ. Since Σ is in one-to-one correspondence with Σ^{-1}, l_1 may be differentiated with respect to the latter (see the Appendix) to obtain the condition $\hat{\Sigma} - S(m) = 0$, whence $\hat{\Sigma} = S(m)$.

To enforce the one-component structural constraints

$$\mu = A\mu, \quad \Sigma = A\Sigma A^\mathsf{T} \quad (12)$$

we use the standard method of lagrange multipliers. The lagrangian term corresponding to $\mu = A\mu$ is $l_\mu = \lambda^\mathsf{T}(\mu - A\mu)$ for a vector of lagrange multipliers λ to be determined. Enforcing $\Sigma = A\Sigma A^\mathsf{T}$ calls for a matrix Λ of lagrange multipliers, one for each entry of $D = \Sigma - A\Sigma A^\mathsf{T}$:

$$l_\Sigma = \sum_{i,j} \lambda_{ij} D_{ij} = \operatorname{tr} D^\mathsf{T}\Lambda = \operatorname{tr}(\Sigma - A\Sigma A^\mathsf{T})\Lambda = \operatorname{tr}\Sigma(\Lambda - A\Lambda A^\mathsf{T}) \tag{13}$$

where we have used $\Sigma = \Sigma^\mathsf{T}$ and the trace identity. The constraint (12) on Σ is equivalent to the same constraint on Σ^{-1}, allowing us to use the more convenient

$$l_{\Sigma^{-1}} = \operatorname{tr}\Sigma^{-1}(\Lambda - A\Lambda A^\mathsf{T})$$

instead of l_Σ.

The overall one-component lagrangian is the sum of these terms:

$$l_{1C}(\mu, \Sigma) = -(N/2)[\kappa + \log|\Sigma| + (m-\mu)^\mathsf{T}\Sigma^{-1}(m-\mu) + \operatorname{tr}\Sigma^{-1}S(m) + \\ 2\lambda^\mathsf{T}(\mu - A\mu) + \operatorname{tr}\Sigma^{-1}(\Lambda - A\Lambda A^\mathsf{T})] \quad . \tag{14}$$

Differentiating with respect to μ implies

$$\hat{\mu} = m + \Sigma(I - A)^\mathsf{T}\lambda \quad .$$

To choose λ to satisfy the constraint, note that

$$P\hat{\mu} = \sum_{r=0}^{P-1} A^r\hat{\mu} = P\tilde{m} + \sum_{r=0}^{P-1} A^r\Sigma(I-A)^\mathsf{T}\lambda = P\tilde{m} + \Sigma\Big[\sum_{r=0}^{P-1} A^r(I - A^\mathsf{T})\Big]\lambda \tag{15}$$

where $P\tilde{m} = \sum_{r=0}^{P-1} A^r m$, and we have used that Σ and A commute when (12) is in force. The quantity in brackets is in fact zero: it contains all powers of A in positive and negated versions. The constrained mean is

$$\hat{\mu} = (1/P)\sum_{r=0}^{P-1} A^r m \quad . \tag{16}$$

Substituting back into (14), rewriting the quadratic form as a matrix trace, and using a modified sample covariance matrix $S(\hat{\mu}) = S(m) + (m - \hat{\mu})(m - \hat{\mu})^\mathsf{T}$ leaves

$$l_{1C}(\hat{\mu}, \Sigma) = -(N/2)[\kappa + \log|\Sigma| + \operatorname{tr}\Sigma^{-1}(S(\hat{\mu}) + \Lambda - A\Lambda A^\mathsf{T})] \quad . \tag{17}$$

As in the unconstrained case, differentiating with respect to Σ^{-1} is more direct, and yields

$$\hat{\Sigma} = S(\hat{\mu}) + \Lambda - A^\mathsf{T}\Lambda A \quad .$$

To enforce the constraint, note that when $\hat{\Sigma}$ satisfies it,

$$P\hat{\Sigma} = \sum_{r=0}^{P-1} A^r\hat{\Sigma}A^{\mathsf{T}\,r} = \sum_{r=0}^{P-1} A^r S(\hat{\mu})A^{\mathsf{T}\,r} + \sum_{r=0}^{P-1} A^r(\Lambda - A^\mathsf{T}\Lambda A)A^{\mathsf{T}\,r}$$

The second term vanishes, and

$$\hat{\Sigma} = (1/P)\sum_{r=0}^{P-1} A^r S(\hat{\mu})A^{\mathsf{T}\,r} \quad . \tag{18}$$

The constrained MLE is summarized in equations 16 and 18. The solution is parallel to the unconstrained MLE, has the interpretation of being an average of appropriately transformed statistics of the data, and is readily computable from the sufficient statistics m and $S(m)$.

4 Normal Mixture Solution

Following the standard approach to fitting a mixture distribution via EM (e.g., [DLR77]), 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 probability distribution of the complete-data factors as

$$p(X,Z) = \prod_{n=1}^{N} \prod_{k=0}^{K-1} [\gamma_k N(x_n;\, \mu_k, \Sigma_k)]^{z_{n,k}} \tag{19}$$

implying that the loglikelihood neatly decouples

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

Its expectation given the observation is

$$l_K(\Theta) = E[\log p(X,Z) \,|\, X] = \sum_{k=0}^{K-1} \sum_{n=1}^{N} \alpha_{n,k} \log[\gamma_k N(x_n;\, \mu_k, \Sigma_k)] \tag{20}$$

where the weights

$$\alpha_{n,k} := E[z_{n,k} \,|\, x_n] = N(x_n; \mu_k, \Sigma_k) \Big/ \sum_{l=0}^{K-1} N(x_n; \mu_l, \Sigma_l) \quad . \tag{21}$$

It is readily seen that $\sum_{k=0}^{K-1} \alpha_{n,k} = 1$, a consequence of $\sum_{k=0}^{K-1} z_{n,k} = 1$. For the same reason, $0 \leq \alpha_{n,k} \leq 1$, so $\alpha_{n,k}/N$ is a joint probability distribution function. It is convenient to also define

$$\alpha_k = \sum\nolimits_{n=1}^{N} \alpha_{n,k}, \quad \alpha_{n|k} = \alpha_{n,k}/\alpha_k; \tag{22}$$

the latter is a correctly normalized conditional distribution.

The expectation $l_K(\Theta)$ is to be maximized at every EM iteration to update the parameters. These parameters can be ordered in any way, but here we assume without loss of generality that they have the structure laid out in equations 9 and table 1.

The update for the weights can be derived separately because the terms of l_K involving γ_k separate from those in which other parameters appear. Including the lagrangian term forcing the weights to be normalized, the function to be maximized is

$$l_{K,C}(\gamma_0, \ldots, \gamma_{K-1}) = \sum_{k=0}^{K-1} \alpha_k \log \gamma_k - \lambda\Big(\sum_{k=0}^{K-1} \gamma_k - 1\Big) \quad .$$

To find γ_k, recall from (9a) that all weights γ_l, $l \in [k]$, are in fact the same parameter. Differentiating reveals the necessary condition

$$\sum\nolimits_{l \in [k]} \alpha_l - \lambda \#[k] \hat{\gamma}_k = 0$$

where $\#[k]$ is the cardinality of the cycle. Summing all these conditions, one per cycle, and recalling that $\sum_{l=0}^{K-1} \alpha_l = N$ shows that

$$\lambda \sum\nolimits_{\text{cycles } [k]} \#[k] \hat{\gamma}_k = N$$

Since the sum must be unity, $\lambda = N$, and the optimal weight is

$$\hat{\gamma}_k = \bigl(1/\#[k]\bigr) \sum\nolimits_{l \in [k]} \alpha_l / N \quad . \tag{23}$$

This is just the average class-membership in the cycle containing k, normalized to sum to unity.

The remaining terms of $l_K(\Theta)$ involve the means and covariances, the weights having already been chosen. Similarly to section 3, we may rewrite the remaining terms of (20) as

$$l_K(\mu_0, \dots, \mu_{K-1}, \Sigma_0, \dots, \Sigma_{K-1}) = \\ -\frac{1}{2} \sum_{k=0}^{K-1} \alpha_k \Bigl[\log |\Sigma_k| + (m_k - \mu_k)^{\mathsf{T}} \Sigma_k^{-1} (m_k - \mu_k) + \operatorname{tr} \Sigma_k^{-1} S_k(m_k)\Bigr] \tag{24}$$

using the weighted sufficient statistics

$$m_k := \sum_{n=1}^{N} \alpha_{n|k} x_n \tag{25a}$$

$$S_k(\eta) := \sum_{n=1}^{N} \alpha_{n|k} (x_n - \eta)(x_n - \eta)^{\mathsf{T}} = S_k(m_k) + (m_k - \eta)(m_k - \eta)^{\mathsf{T}} \quad . \tag{25b}$$

For both averages, the k subscript indicates weighting by the conditional probabilities $\alpha_{n|k}$. It is easy to recover from (24) the standard unconstrained EM updates

$$\hat{\mu}_k = m_k, \text{ and } \hat{\Sigma}_k = S_k(m_k)$$

by inspection (for $\hat{\mu}_k$) and differentiation (for $\hat{\Sigma}_k$) just as below (10).

It is immediate from the sum in (24) that, in the unconstrained mixture problem, each bump's parameter updates $(\hat{\mu}_k, \hat{\Sigma}_k)$ 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 (9a). In the remainder of this section, we suppose the cycle is indexed as $[k] = \{0, \dots, Q-1\}$ to avoid superfluous notation.

To enforce the external constraints of (9a), we let μ_0 be a free parameter, and then write $\mu_l = A^l \mu_0$ for $0 < l < Q$, and similarly for the covariances. We again use the lagrangian mechanism to account for the internal constraints (9b), namely

$$\mu_l = A^Q \mu_l, \quad \Sigma_l = A^Q \Sigma_l A^{\mathsf{T}Q}, \quad 0 \le l < Q, \tag{26}$$

which can of course by accomplished by constraining (μ_0, Σ_0) only; compare (12). With this way of writing the parameters, the cycle-k terms of the objective function (24) are

$$\begin{aligned} l_K(\mu_0, \Sigma_0) &= -\frac{1}{2} \sum_{k=0}^{Q-1} \alpha_k \Big[\log|\Sigma_0| + (m_k - A^k\mu_0)^\mathsf{T} A^k \Sigma_0^{-1} A^{\mathsf{T}k}(m_k - A^k\mu_0) + \\ &\qquad \operatorname{tr} \Sigma_0^{-1} A^{\mathsf{T}k} S_k(m_k) A^k\Big] \\ &= -\frac{\alpha_{[0]}}{2} \sum_{k=0}^{Q-1} \bar{\alpha}_k \Big[\log|\Sigma_0| + (A^{\mathsf{T}k} m_k - \mu_0)^\mathsf{T} \Sigma_0^{-1} (A^{\mathsf{T}k} m_k - \mu_0) + \\ &\qquad \operatorname{tr} \Sigma_0^{-1} A^{\mathsf{T}k} S_k(m_k) A^k\Big] \end{aligned} \tag{27}$$

where $\alpha_{[0]} := \sum_{k=0}^{Q-1} \alpha_k$ and $\bar{\alpha}_k = \alpha_k / \alpha_{[0]}$, a probability mass function on $\{0, \ldots, Q-1\}$.

Collapsing the Q parameters to one has made, for example, $m_0, \ldots, m_{Q-1}$ informative about μ_0. It thus aids understanding to rewrite (27) via another set of sufficient statistics

$$\bar{m} := \sum_{k=0}^{Q-1} \bar{\alpha}_k A^{\mathsf{T}k} m_k \tag{28a}$$

$$\bar{S} := \sum_{k=0}^{Q-1} \bar{\alpha}_k A^{\mathsf{T}k} S_k(A^k \bar{m}) A^k \quad . \tag{28b}$$

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 (27):

$$\begin{aligned} &\sum_{k=0}^{Q-1} \bar{\alpha}_k (A^{\mathsf{T}k} m_k - \mu_0)^\mathsf{T} \Sigma_0^{-1} (A^{\mathsf{T}k} m_k - \mu_0) = \\ &\qquad (\bar{m} - \mu_0)^\mathsf{T} \Sigma_0^{-1} (\bar{m} - \mu_0) + \sum_{k=0}^{Q-1} \bar{\alpha}_k (A^{\mathsf{T}k} m_k - \bar{m})^\mathsf{T} \Sigma_0^{-1} (A^{\mathsf{T}k} m_k - \bar{m}) \end{aligned}$$

Substituting and combining summations into the trace converts (27) into

$$\begin{aligned} l_K(\mu_0, \Sigma_0) &= -\frac{\alpha_{[0]}}{2} \Big(\log|\Sigma_0| + (\bar{m} - \mu_0)^\mathsf{T} \Sigma_0^{-1} (\bar{m} - \mu_0) + \\ &\operatorname{tr} \Sigma_0^{-1} \sum_{k=0}^{Q-1} \bar{\alpha}_k \big[(A^{\mathsf{T}k} m_k - \bar{m})(A^{\mathsf{T}k} m_k - \bar{m})^\mathsf{T} + A^{\mathsf{T}k} S_k(m_k) A^k\big]\Big) \quad . \end{aligned} \tag{29}$$

Applying the covariance decomposition (25b) reveals the summation to be

$$\sum_{k=0}^{Q-1} \bar{\alpha}_k A^{\mathsf{T}k} \Big(S_k(m_k) + (m_k - A^k \bar{m})(m_k - A^k \bar{m})^{\mathsf{T}} \Big) A^k = \sum_{k=0}^{Q-1} \bar{\alpha}_k A^{\mathsf{T}k} S_k(A^k \bar{m}) A^k = \bar{S} \quad .$$

Ignoring the leading factor, which is irrelevant to the maximization, the objective, including the lagrangian terms, becomes

$$l_{K,C}(\mu_0, \Sigma_0) = -\log|\Sigma_0| - (\bar{m} - \mu_0)^{\mathsf{T}} \Sigma_0^{-1} (\bar{m} - \mu_0) - \operatorname{tr} \Sigma_0^{-1} \bar{S} + 2\lambda^{\mathsf{T}}(\mu_0 - A^Q \mu_0) + \operatorname{tr} \Sigma_0^{-1} (\Lambda - A^Q \Lambda A^{\mathsf{T}Q}) \quad . \tag{30}$$

Differentiating the lagrangian 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 on $\hat{\mu}_0$, reuse the averaging trick (15) with $P' = P/Q$ terms, noting that $(A^Q)^{P'} = I$ and that Σ_0 and A^Q commute in the presence of (26). The constrained mean is

$$\hat{\mu}_0 = (1/P') \sum_{r=0}^{P'-1} A^{\mathsf{T}Qr} \bar{m} \quad . \tag{31}$$

Substituting $\hat{\mu}_0$ into the lagrangian (30) and rewriting yields

$$l_{K,C}(\hat{\mu}_0, \Sigma_0) = -\log|\Sigma_0| - \operatorname{tr} \Sigma_0^{-1} \big(\bar{S} + (\bar{m} - \hat{\mu}_0)(\bar{m} - \hat{\mu}_0)^{\mathsf{T}} \big) + \operatorname{tr} \Sigma_0^{-1} (\Lambda - A^Q \Lambda A^{\mathsf{T}Q}) \tag{32}$$

As before, we differentiate with respect to Σ_0^{-1} to get a necessary condition

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

Using once again the observation that, when the constraint is satisfied, $P' \Sigma_0 = \sum_{r=0}^{P'-1} A^{\mathsf{T}Qr} \Sigma_0 A^{Qr}$, we find the constrained covariance

$$\hat{\Sigma}_0 = (1/P') \sum_{r=0}^{P'-1} A^{\mathsf{T}Qr} \big[\bar{S} + (\bar{m} - \hat{\mu}_0)(\bar{m} - \hat{\mu}_0)^{\mathsf{T}} \big] A^{Qr} \quad . \tag{33}$$

Equations 31 and 33, together with the sufficient statistics (28), encapsulate the constrained EM iteration. The parameters are updated with a weighted average of transformed sufficient statistics. The first averages, equations 28, are across Q terms, one for each linked component in the cycle. The second averages, in (31) and (33), are over the sufficient statistics influencing each individual bump.

This has allowed us to find $(\hat{\mu}_0, \hat{\Sigma}_0)$, the base parameters θ_k for cycle $[k]$, so it is immediate that $\theta_l = A^l \theta_k$ for $0 < l < Q$. The same procedure is used to find the parameters for the other cycles.

5 Conclusion

We have developed an EM algorithm for maximum-likelihood estimation of symmetry-constrained normal mixtures. A subsequent paper discusses implementation and results.

Acknowledgements

The research described in this paper was carried out for the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA.

6 Appendix

Some useful facts, which may need conditions on Σ. See also [BLW82, WS98].

$\nabla_\Sigma \log|\Sigma| = \Sigma^{-1}$.

$\nabla_\Sigma \operatorname{tr} \Sigma M = M$.

References

[BLW82] J. P. Burg, D. G. Luenberger, and D. L. Wenger. Estimation of structured covariance matrices. *Proceedings of the IEEE*, 70(9):963–74, September 1982.

[DLR77] A. D. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B-39:1–37, 1977.

[Gal99] M. J. F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Trans. Speech and Audio Processing*, 7(3), 1999.

[GH96] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, 1996.

[Gop98] R. A. Gopinath. Constrained maximum likelihood modeling with Gaussian distributions. In *Proc. of ARPA Workshop on Human Language Understanding*, 1998.

[LP98] S. Y. Lee and W. Y. Poon. Analysis of two-level structural equation models via EM type algorithms. *Statistica Sinica*, 8:749–766, 1998.

[MK97] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.

[MP00a] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

[MP00b] G. J. McLachlan and D. Peel. Mixtures of factor analyzers. In P. Langley, editor, *Proc. Seventeenth Intern. Conf. Machine Learning*, pages 599–606. Morgan Kaufmann, 2000.

[RT82] D. B. Rubin and D. T. Thayer. EM algorithms for factor analysis. *Psychometrika*, 47:69–76, 1982.

[WS98] M. Warmuth and Y. Singer. Batch and on-line parameter estimation of Gaussian mixtures based on the joint entropy. In *Advances Neural Inform. Processing Syst. 11*, pages 578–584. MIT, 1998.