Chapter 3 Relaxation Expectation–Maximization

In chapter 2 we noted a number of practical difficulties that arise in the use of the Expectation–Maximization (EM) algorithm to find maximum likelihood fits of mixture models. Two among these were the sensitivity to initial conditions and the computational overhead involved in carrying out model selection. In this chapter we shall introduce a modified EM algorithm which addresses both of these issues in a natural fashion. Our modifications will rely on the statistical mechanics notion of **relaxation**.

3.1 Annealing and Relaxation

3.1.1 Simulated annealing

Relaxation methods are well known in data analysis, primarily due to the popularity of the **simulated annealing** technique for the solution of non-convex optimization problems (Kirkpatrick *et al.* 1983). This being the most common example, we will review it briefly so as to provide a point of departure for our discussion.

The objective is to find the global minimum of a function $E(x)$. The approach taken is to simulate the motion (in x space) of a thermally excited particle under the influence of a potential energy landscape given by $E(x)$. In principle, at zero temperature the particle will be found at the global minimum. Of course, in practice, if it starts at a position far from the lowest energy point it will most likely travel to a local minimum and come to rest there. At higher temperatures, the particle will travel rapidly all over the landscape, spending relatively more time in regions where the function $E(x)$ is minimal. The annealing procedure lowers the simulated temperature gradually. As the temperature falls, the bias towards regions of lower energy increases, while the particle is still able to cross barrier regions of higher energy. If the rate of cooling is sufficiently gradual, these two tendencies — the attraction to regions of low energy and the thermal activation to cross energy barriers — combine in such a way as to inevitably leave the particle at the global minimum once the temperature reaches 0. Cooling schedules which guarantee this result can be shown to exist in principle (Geman and Geman 1984); however, they invariably take impractically long. Fortunately, less than perfect cooling schedules usually yield good results.

This physical picture of the optimization process is appealing, but it is difficult to build intuition for why the trade-off between activation energy and attraction to potential wells should work out so conveniently. Also, while it will be valuable to contrast this view with the "deterministic annealing" or relaxation procedure we will discuss later, it is not the most convenient starting point for the development of the new approach. Therefore we reexamine the algorithm from a more statistical viewpoint.

3.1.2 Annealed sampling

The fundamental logic behind annealing schemes is best illustrated by the simulated annealing of Markov chain Monte-Carlo (MCMC) samplers (Neal 1993; Bertsimas and Tsitsiklis 1993). The objective here is to sample from some complicated target probability function $\mathsf{P}(x)$. For convenience, we will introduce an energy function given, up to an arbitrary additive constant, by $E(x) = -\log \mathsf{P}(x)$. The density is thus given by the Boltzmann equation $\mathsf{P}(x) = \frac{1}{Z}\exp(-E(x))$, for some normalizing constant Z. We are able to evaluate $E(x)$ for any point x, but the energy does not have a simple functional form that makes direct sampling by analytic means tractable. The MCMC sampling approach constructs an ergodic Markov-chain[1] over the target space such that the stationary distribution of the chain is $\mathsf{P}(x)$. In other words, we obtain a scheme for making probabilistic transitions from one point in the space to another in a memory-less (Markov) fashion, and such that, in the long run, the probability of visiting some point x is exactly $\mathsf{P}(x)$. A number of schemes to construct a suitable Markov chain exist, the most prominent being the Gibbs sampling and the Metropolis algorithms. The details of the process are unimportant for our purposes; we seek only to gain an intuitive picture of the value of annealing; the reader interested in more detail is referred to the excellent review by Neal (1993).

When using an MCMC sampler, we need to begin the chain at some point in the domain, say x_0. Since we cannot sample directly from the target density, this point must be chosen from an arbitrary density, probably quite different to the target one. Let us say this initial density is uniform on the domain of interest, although the argument is not crucially dependent on this choice. The density of the next point, call it x_1, is then the product of this uniform distribution and the transition density of the Markov chain, marginalized over x_0, $\mathsf{P}_1(x_1) = \int dx_0 \; \mathsf{P}_0(x_0)\, \mathsf{P}(x_1 \mid x_0)$. (For discrete domains we can picture multiplying a vector representing the uniform distribution by a transition matrix.) The resultant density will also be far from the target, as will the densities of many subsequent samples. Thus, our necessarily poor choice of $\mathsf{P}_0(x_0)$ results in a "burn-in" period of incorrectly distributed samples. The typical length of this period is related to the mismatch between the initial distribution and the target (or stationary) distribution, and to the magnitude of the non-unit eigenvalues of the transition operator, which set the decay rate of the non-stationary modes in $\mathsf{P}_0(\cdot)$. In general, the mixing time cannot easily be calculated, but in experiments with practical examples it is often impractically long.

The difficulty is that in many problems $\mathsf{P}_0(x_0)$ is likely to ascribe a relatively large mass to

[1] The basic theory of Markov chains will be reviewed in section 4.1.1.

regions where the target function is vanishingly small, and furthermore, has small log-gradients. For domains of high dimensionality, the probability of falling in such regions can approach 1. The structure of the usual MCMC samplers (in particular, a feature called **detailed balance** which is needed to guarantee ergodicity) results in the sampler executing an almost unbiased random walk within that region until it finally emerges into a region of higher probability.

How can annealing help reduce this burn-in period? We create a sequence of probability functions $\mathsf{P}_0(x)$, $\mathsf{P}_1(x)$, ..., $\mathsf{P}(x)$ which starts with the uniform distribution and ends in the target. In the case of the Boltzmann distribution this sequence is easily constructed using a "inverse-temperature" parameter, β. We choose a sequence of β_i, starting with 0 and ending in 1, and write $\mathsf{P}_i(x) = \frac{1}{Z(\beta_i)} \exp(-\beta_i E(x))$, where $Z(\beta_i)$ is the partition function. By analogy with statistical physics, these densities correspond to the canonical distributions of a system with energy E cooled through a sequence of temperatures $T = 1/\beta$. We now choose an initial point from $\mathsf{P}_0(x)$ as before, but then use the MCMC sampler corresponding to the density $\mathsf{P}_1(x)$, with $0 < \beta_1 \ll 1$, rather than the target sampler. The mismatch between these two distributions is small by construction, and so this Markov chain will soon achieve the stationary distribution for $\mathsf{P}_1(x)$. Once enough time has elapsed to make convergence likely, we switch to sampling from $\mathsf{P}_2(x)$, where the same argument about quick convergence holds. Eventually, we reach the target distribution (at $\beta = 1$). In many situations, the total burn-in time for all of the annealing steps is much smaller than the burn-in encountered stepping directly to the target.

What does all this have to do with the physical picture of optimization by simulated annealing that we saw before? The Metropolis sampling algorithm used in some MCMC simulations has its origins in the physical simulation of particle motion, and, indeed, is precisely the simulation algorithm used by Kirkpatrick *et al.* (1983). If we extend to temperatures close to 0 ($\beta \gg 1$) the sequence of distributions discussed above, virtually all of the probability mass becomes concentrated near the global energy minimum. Provided the MCMC sampler is maintained in equilibrium, then, samples drawn in this limit will be arbitrarily close to the optimum. This is precisely the simulated annealing optimization algorithm.

3.1.3 Relaxation

We have examined the simulated annealing algorithm from two different points of view. In the first, the underlying energy landscape was fixed by the function to be optimized, while the motion of a thermally active particle in the landscape was simulated at steadily decreasing temperatures. In the second, the energy landscape was transformed from a flat initial condition to the target function *and beyond*, while samples were drawn from the corresponding Boltzmann distribution. This gradual transformation of the energy surface is called **relaxation**; for this reason, simulated annealing is also known as **stochastic relaxation**.

Optimization within a relaxation framework need not be stochastic. Let us focus on the energy functions themselves rather than on the implied Boltzmann densities. We can construct a sequence of functions, $E_0(x) \ldots E(x)$ such that the first function $E_0(x)$ is easily optimized — it might, for example, have a single extremum — while the final function is the target. Our goal in constructing this sequence is for the global optimum of the ith function $E_i(x)$ to lie within the domain of convergence of the global optimum of the next function $E_{i+1}(x)$. We then pass along the sequence of functions, optimizing each one by a hill-climbing (or, for minima, descending) algorithm, which is seeded with the location of the previous optimum. Thus, we hope to track the global optimum from $E_0(x)$, where it was easily found, to $E(x)$. Unfortunately, unlike the case of stochastic relaxation, there is no simple strategy that is guaranteed to provide a suitable sequence of functions in the case of such deterministic relaxation, even with exponentially long relaxation schedules, and indeed schemes devised for particular classes of energy (say mixture likelihoods) may not work even in all examples of that class. Nevertheless, in practice, this approach often does yield good results.

3.2 Deterministic Annealing

One example of a non-stochastic relaxation process has been called **deterministic annealing**. This algorithm was introduced by Rose *et al.* (1990) as a maximum entropy approach to clustering and vector quantization, following earlier work on **elastic net** algorithms for the traveling salesman problem (Durbin and Willshaw 1987; Durbin *et al.* 1989; Simic 1990; Yuille 1990). In this form, the algorithm is strongly motivated by physical analogy. Below, we will see that it can be generalized beyond its statistical physics origins, to yield a powerful procedure that can be applied to any problem in which the EM algorithm is used for learning. We shall refer to the generalization as Relaxation Expectation–Maximization, reserving the term "deterministic annealing" for the original formulation.

Rose *et al.* view clustering as a squared-distance distortion minimization operation. They introduce a cost function, $E_m(x_i)$, describing the distortion due to association of the the ith data point with the mth cluster. We shall take this cost to be the squared Euclidean distance $E_m(x_i) = \|\mu_m - x_i\|^2$, although other distortions may be considered. The cost of adopting a particular set of cluster parameters $\theta = \{\mu_m\}$ *and* a particular assignment of points to clusters, represented by indicator variables $\mathcal{Z} = \{z_{m,i}\}$, is given by

$$E(\theta, \mathcal{Z}) = \sum_i \sum_m z_{m,i} E_m(x_i) \tag{3.1}$$

We have chosen notation different from that of Rose *et al.* (1990) in order to highlight the similarity to the mixture model development in chapter 2. This cost, $E(\theta, \mathcal{Z})$, may be viewed as the energy of a microstate, identified by the pair $(\theta, \mathcal{Z})$, of a physical system and we may proceed by analogy to statistical physics (as we will see below, this analogy is not vital; the results follow directly from the maximum-likelihood framework and the EM algorithm). We expect the system to display a distribution over microstates $\mathsf{P}(\theta, \mathcal{Z})$. For a fixed average energy, E, this distribution will maximize the entropy under the constraint $\mathcal{E}[E(\theta, \mathcal{Z})] = E$ (see, for example, Kittel and Kroemer (1980)). We can find this maximizer by the method of Lagrange multipliers, optimizing the entropy $H = -\int d\theta \ \sum_{\mathcal{Z}} \mathsf{P}(\theta, \mathcal{Z}) \log \mathsf{P}(\theta, \mathcal{Z})$ while enforcing the constraint $E - \int d\theta \ \sum_{\mathcal{Z}} \mathsf{P}(\theta, \mathcal{Z}) E(\theta, \mathcal{Z}) = 0$ with the multiplier β. Doing so, we obtain the well-known Boltzmann distribution

$$\mathsf{P}_{\beta}(\theta, \mathcal{Z}) \propto e^{-\beta E(\theta, \mathcal{Z})} \tag{3.2}$$

The value of the multiplier β can be obtained by solving for the constraint energy. Rose *et al.* argue, as we have, that the distribution of interest in the case of modeling or prediction problems is not the joint, but rather the marginal

$$\mathsf{P}_{\beta}(\theta) = \sum_{\mathcal{Z}} \mathsf{P}(\theta, \mathcal{Z}) \propto \prod_{i} \sum_{m} e^{-\beta E_m(x_i)} \tag{3.3}$$

For the case of the squared distance cost, this is seen to be the same as the likelihood of a mixture of Gaussians with mixing probabilities $\pi_m = \frac{1}{M}$ and covariances $\Sigma_m = \frac{1}{2\beta} I$.

Given this "likelihood", they proceed to derive heuristically re-estimation equations similar to those of the EM algorithm (written here for the squared error distortion metric):

$$\begin{array}{rcl} r_{i,m} & \leftarrow & e^{-\beta E_m(x_i)} / \sum_l e^{-\beta E_l(x_i)} \\ \mu_m & \leftarrow & \sum_i r_{i,m} x_i / \sum_i r_{i,m} \end{array} \tag{3.4}$$

We have again chosen notation to emphasize the connection to our previous development. The deterministic annealing algorithm then involves varying the value of the parameter β from 0 to a final value chosen either through some knowledge of the expected final distortion (due, say, to a known noise-floor), or else by a validation-based stopping criterion (or else by operator fiat). At each step the re-estimations (3.4) are iterated to convergence.

The intuitions that underlie this algorithm can be used to obtain similar solutions to a number of other problems (Rose *et al.* 1993; Buhmann and Kuhnel 1993; Miller *et al.* 1996; Kloppenburg and Tavan 1997; Rao *et al.* 1997; Rao *et al.* 1999). Many of these are reviewed by Rose (1998). In general, however, each such problem presents the need for a fresh derivation. Furthermore, it is not always clear how best to generalize the approach to some problems. For example, Kloppenburg and Tavan (1997) provide an extension to a mixture of multivariate Gaussians with arbitrary covariances; but they are forced to introduce multiple annealing parameters, leaving serious questions about the choice of relative annealing schedules.

In the next section we will encounter a generalized relaxation method which subsumes the various deterministic annealing algorithms, and allows extremely straightforward generalization.

3.3 REM-1

In this section, we will develop a novel relaxation scheme within the framework of the EM algorithm, to obtain an algorithm that we call the first **Relaxation Expectation–Maximization** algorithm[2] (REM-1).

In section 1.7 we introduced a free-energy F, a function of the model parameters, θ, and a probability distribution on the latent variables, p,

$$F(p, \theta) = Q(p, \theta) + H(p) = \mathcal{E}_p \left[\ell_{\mathcal{X},\mathcal{Y}}(\theta)\right] - \mathcal{E}_p \left[\log p(\mathcal{Y})\right] \tag{3.5}$$

We showed that if this function achieved a maximum at (θ^*, p^*) the true model likelihood (marginalized over the latent variables) achieved a maximum at θ^*. This allowed us to interpret the EM algorithm as an alternation of optimization steps, maximizing F first with respect to p, and then with respect to θ. This view of EM forms the basis for our relaxation scheme.

Let us introduce an annealing parameter β so as to construct a family of free-energy functions,

$$F_\beta(p, \theta) = \beta Q(p, \theta) + H(p) \tag{3.6}$$

The analogy to statistical mechanics inherent in the term "free-energy" is maintained by this choice (modulo an overall minus sign). We may view β as the inverse of a (dimensionless) temperature, in which case it enters into the free-energy definition in the physically appropriate position. When β takes the value 1 (that is, $T = 1$) we recover the original free-energy, which is the target function whose maximum we seek. On the other hand, when β is 0 ($T \to \infty$) F is equal to the entropy $H(p)$. In general, there is a single, easy to find, global maximum of this entropy. For discrete latent variables, for example, it is achieved by the uniform distribution. For the case of the mixture model, in which the latent variables indicate with which cluster each point is associated, and we see that F_0 is maximized by associating all of the points uniformly with all of the clusters. The $\beta = 0$ case does not constrain the parameters θ at all, however it is convenient to choose θ as before, maximizing Q with p fixed at its maximum-entropy value.

Thus, the sequence of functions $F_{\beta_i}(p, \theta)$, $0 = \beta_0 < \beta_1 < \cdots < \beta_R = 1$ satisfies at least two of the conditions we desired for a relaxation progression: it starts with an easily maximized function

[2] The same formulation has been independently proposed under the name "Deterministic Annealing Expectation Maximization" by Ueda and Nakano (1998). A slightly different development, which we call REM-2, will appear below.

and ends with the target. To be sure of finding the global maximum of the target function we need another condition to be satisfied: the global maximum of each function in the sequence must lie within the basin of attraction of the global maximum of the next function. Provided that the location of global maximum changes continuously with β, this can be assured by choosing sufficiently small annealing steps.[3] Unfortunately, we will see below that even for the particularly simple example of the mixture model, the maximum does not move smoothly. In general it is not guaranteed that REM will find the global maximum of the target. However, in many common examples it does find a good maximum.

Any hill-climbing technique may be used to find the optimum of each succeeding free-energy in the relaxation sequence; however, we choose to employ the same approach as in the EM algorithm, alternately optimizing with respect to p and θ, in each case holding the other variable fixed. Note first that, for fixed p, the relaxation factor β has no effect on the optimal value of θ. Thus, the M-step of the algorithm is exactly as for the normal EM algorithm. The E-step, however, does differ.

We showed previously (1.42) that the target free-energy is maximized with respect to p (for fixed θ) by choosing $p(\mathcal{Y}) = \mathsf{P}_\theta\left(\mathcal{Y} \mid \mathcal{X}\right)$. In the case of the relaxation free-energies we can proceed in the same fashion as we did at that point. We introduce a Lagrange multiplier λ enforcing the constraint $\int d\mathcal{Y}\; p(\mathcal{Y}) = 1$ and obtain

$$\begin{aligned} 0 &= \frac{\partial}{\partial p}\left(F_\beta(p,\theta) - \lambda \int d\mathcal{Y}\; p(\mathcal{Y})\right) \\ &= \frac{\partial}{\partial p}\left(\int d\mathcal{Y}\; p(\mathcal{Y})(\beta \ell_{\mathcal{X},\mathcal{Y}}(\theta) - \log p(\mathcal{Y}) - \lambda)\right) \end{aligned} \tag{3.7}$$

from which, by the calculus of variations,

$$\begin{aligned} 0 &= \frac{\partial}{\partial p}\left(p(\mathcal{Y})(\beta \ell_{\mathcal{X},\mathcal{Y}}(\theta) - \log p(\mathcal{Y}) - \lambda)\right) \\ &= (\beta \ell_{\mathcal{X},\mathcal{Y}}(\theta) - \log p^*(\mathcal{Y}) - \lambda) - \frac{p^*(\mathcal{Y})}{p^*(\mathcal{Y})} \end{aligned} \tag{3.8}$$

and so

$$p^*(\mathcal{Y}) = e^{-\lambda-1}(\mathcal{L}_{\mathcal{X},\mathcal{Y}}(\theta))^\beta = e^{-\lambda-1}(\mathsf{P}_\theta\left(\mathcal{X},\mathcal{Y}\right))^\beta \tag{3.9}$$

But $\mathsf{P}_\theta\left(\mathcal{X},\mathcal{Y}\right) = \mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right)\mathsf{P}_\theta\left(\mathcal{Y}\right)$ and so

$$p^*(\mathcal{Y}) = \frac{1}{Z(\beta)}(\mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right)\mathsf{P}_\theta\left(\mathcal{Y}\right))^\beta \tag{3.10}$$

[3] This assertion can be proved by noting that a global maximum must have at least an ϵ-sized basin of attraction and that continuity guarantees that there exists some δ so that for a δ-sized step in β the change in global maximum is smaller than this ϵ.

with $Z(\beta)$ and appropriate normalizing constant.

Thus we obtain the steps of the REM-1 algorithm, repeated until $\beta = 1$.

R-step: Increment β according to the relaxation schedule.

Repeat the following EM steps until convergence:

E-step: Maximize F_β with respect to p holding θ fixed.

$$p(\mathcal{Y}) \leftarrow \frac{1}{Z(\beta)}(\mathsf{P}_\theta\,(\mathcal{X} \mid \mathcal{Y})\,\mathsf{P}_\theta\,(\mathcal{Y}))^\beta \tag{3.11}$$

M-step: Maximize F_β with respect to θ holding p fixed.

$$\theta \leftarrow \mathrm{argmax}\,\mathcal{E}_p\left[\ell_{\mathcal{X},\mathcal{Y}}\,(\theta)\right] \tag{3.12}$$

Relationship to deterministic annealing

The deterministic annealing algorithm for vector quantization described in section 3.2 is easily seen to arise from REM-1 applied to a simple mixture model. Consider an M-component model in which each component is a Gaussian with identity covariance matrix and mean μ_m. We will refer to this as a mixture of unit Gaussians. Any model in which the all of the components are known to share the covariance matrix Σ can be transformed to this canonical form by multiplying each data vector by the whitening matrix $\Sigma^{-1/2}$. The relaxation free-energy for such a model is

$$F_\beta(p,\theta) = \beta \sum_i \sum_m r_{m,i}(\log \pi_m - \frac{1}{2}\|x_i - \mu_m\|^2) - \sum_i \sum_m r_{m,i} \log r_{m,i} \tag{3.13}$$

where the distribution p is expressed in terms of the responsibilities $r_{m,i}$. For notational simplicity we have left out the normalization factor from the Gaussian. For a model with fixed, equal, covariances this factor does not change and careful inspection reveals that it does not survive in any of our eventual results.

The REM-1 iterations for such a model are easily seen to be given by

$$\begin{array}{rcl} r_{i,m} & \leftarrow & \frac{1}{Z_i}\pi_m^\beta e^{-\frac{1}{2}\beta\|x_i-\mu_m\|^2} \\ \pi_m & \leftarrow & \sum_i r_{i,m}/|\mathcal{X}| \\ \mu_m & \leftarrow & \sum_i r_{i,m}x_i/\sum_i r_{i,m} \end{array} \tag{3.14}$$

If we further constrain the mixing probabilities to remain equal, that is, $\pi_m = 1/M$, we obtain exactly the iterations of (3.4).

Note that in the case of the fixed mixing probabilities, the relaxation likelihoods correspond to true likelihoods for other models, in this case, a mixture of Gaussians with covariance $\beta^{-1}I$. This allows us to interpret the relaxation procedure as the successive optimization of a sequence of models with shrinking covariances. This is actually a special case and for the majority of models no such equivalence holds. Given even the simple step of allowing unconstrained mixing probabilities, the iterations (3.14) do not correspond to EM for any model.

It is instructive to note that the maximization of the free-energy with respect to p, which is motivated in REM entirely by the maximum likelihood considerations of chapter 1, may indeed be interpreted as a maximization of the entropy of p under a "constraint" set by the expected joint log-likelihood and enforced by a Lagrange multiplier. This is in accordance with the physical analogy of Rose *et al.* (1990), although it is obtained directly without resort to the physics.

Yuille *et al.* (1994) remarked on a connection between the heuristic optimization steps usually employed within deterministic annealing solutions and the EM algorithm. However, they seem to regard EM simply as an optimization technique embedded within the physically motivated deterministic annealing framework. Notably, they appear to have failed to observe the deep connection between the free-energy formulation of EM and the relaxation procedures of deterministic annealing; in particular, they make no mention of the availability of a simple generalization of any EM algorithm to yield a relaxation (or "annealing") procedure.

3.4 Phase Transitions in REM

An important feature of deterministic annealing and relaxation EM is best illustrated in a simple example. We will use the mixture of unit Gaussians described in the preceding section. We will write $(r^*_{m,i}, \pi^*_m, \mu^*_m)$ for the optimum of the relaxation free-energy. Clearly, these values satisfy the recurrence relations

$$\begin{aligned} r^*_{m,i} &= \frac{\pi_m^{*\beta} e^{-\frac{1}{2}\beta\|x_i-\mu^*_m\|^2}}{\sum_l \pi_l^{*\beta} e^{-\frac{1}{2}\beta\|x_i-\mu^*_l\|^2}} && (3.15) \\ \pi^*_m &= \frac{\sum_i r^*_{m_i}}{|\mathcal{X}|} && (3.16) \\ \mu^*_m &= \frac{\sum_i r^*_{m,i} x_i}{\sum_i r^*_{m,i}} && (3.17) \end{aligned}$$

When $\beta = 0$ the relaxation E-step finds the maximum entropy distribution over the latent variables. For a mixture distribution, where the latent variables are discrete, this is the uniform distribution and

$$r^*_{m,i} = \mathsf{P}\left(z_{m,i} = 1 \mid x_i\right) = \frac{1}{M} \qquad (3.18)$$

In this limit the relaxation free-energy is independent of θ and so the M-step is unconstrained. However, we can choose it to maximize $Q(\theta, p^*)$ where p^* is the maximum entropy distribution described above, thereby preserving consistency with the $\beta > 0$ case. As the responsibilities for each

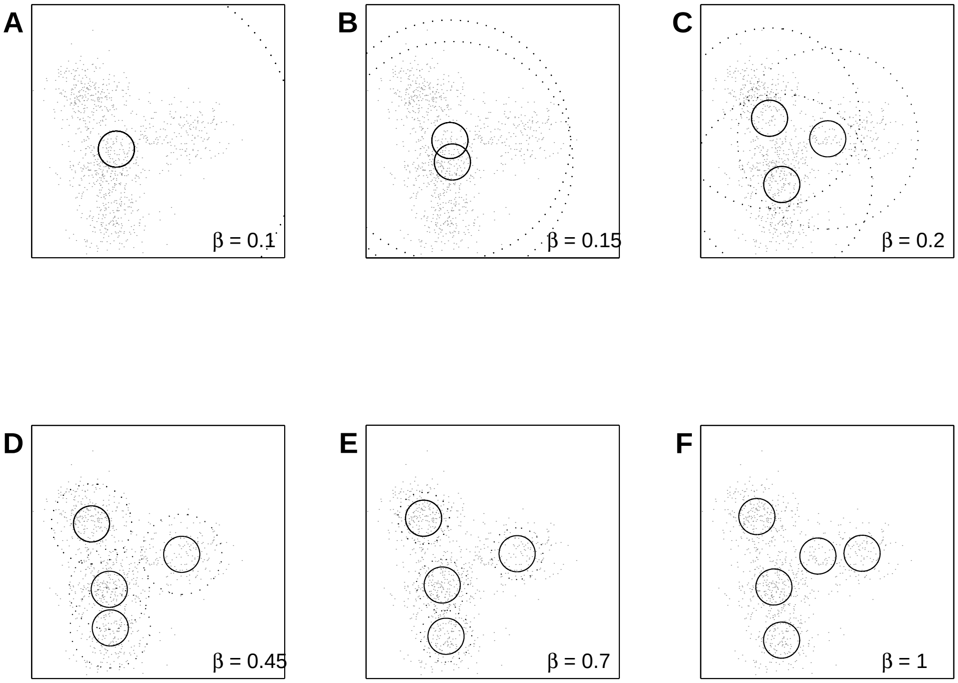

Figure 3.1: Phase transitions in REM-1 for fixed-variance Gaussians

data point are shared equally between all of the components, the maximizing μ_m are all identical. The solution in the $\beta = 0$ case, then, has all the components located at the overall mean of the data.

A remarkable fact is that even as the temperature decreases (that is, β increases) this solution remains the global maximum of the likelihood for some range of temperatures. Once the relaxation process reaches a critical temperature, the solution undergoes a **phase transition** and the former stationary point (where all the components are identical) ceases to be a maximum. A new maximum appears, usually dividing the components into two groups, so that all of the components assume one of only two distinct parameter values. As the system cools further, the optimal solution again continues with only two distinct component values, although the values of those components may change. Eventually, though, it undergoes another phase transition and more distinct components are observed.

Figure 3.1 shows an example of the optimal mixtures at various stages of relaxation. We fit two dimensional data, shown by the scattered points, by a mixture of five unit Gaussians. Each panel of the figure shows the mixture at a different temperature. The inner, solid, circle shows the 1σ boundary of the Gaussian; the outer, dashed, circle shows the effective variance ($\beta^{-1}I$) boundary. In the first few diagrams, fewer than five components are visible due to the exact coincidence of the means.

3.4.1 Critical temperatures

In the case of this simple model it is possible to calculate the critical temperatures at which the mixture will undergo a phase transition.

Suppose we were to start the EM algorithm with parameters θ^0 in which two (or more) of the components were identical. Without loss of generality we shall take these two be the first two components, setting $\mu_1^0 = \mu_2^0$ and $\pi_1^0 = \pi_2^0$. At each E-step the responsibilities of these two components for each of the data points will be the same. Thus, at the M-step they will both be updated in exactly the same way, and will remain identical. The EM algorithm will thus preserve the duplication, and will converge to a stationary point with $\mu_1^* = \mu_2^*$ and $\pi_1^* = \pi_2^*$.

Is this stationary point a maximum, or merely a saddle point? The stability of the solution θ^* can be evaluated by examining the value of the Hessian of the free-energy at that point. In fact, we know that for any parameter value, F_β is maximized with respect to the $r_{m,i}$ by the relaxation E-step. Thus, we need only evaluate the Hessian within the surface of constraint set by the equation (3.11). With the responsibilities chosen optimally, we can reduce the free-energy thus,

$$\begin{aligned}
\ell_\beta(\theta) &= F_\beta\left(\frac{\pi_m^\beta e^{-\frac{1}{2}\beta\|x_i-\mu_m\|^2}}{\sum_l \pi_l^\beta e^{-\frac{1}{2}\beta\|x_i-\mu_l\|^2}}, \theta\right) \\
&= \beta \sum_i \sum_m r_{m,i} \log\left(\pi_m e^{-\frac{1}{2}\|x_i-\mu_m\|^2}\right) - \sum_i \sum_m r_{m,i} \log r_{m,i} \\
&= \sum_i \sum_m r_{m,i} \log\left(\frac{\pi_m^\beta e^{-\frac{1}{2}\beta\|x_i-\mu_m\|^2}}{r_{m,i}}\right) \\
&= \sum_i \sum_m r_{m,i} \log \sum_l \pi_l^\beta e^{-\frac{1}{2}\beta\|x_i-\mu_l\|^2} \\
&= \sum_i \log \sum_l \pi_l^\beta e^{-\frac{1}{2}\beta\|x_i-\mu_l\|^2} \qquad (3.19)
\end{aligned}$$

where, in the last step we have used the fact that $\sum_m r_{m,i} = 1$. This form is quite similar to the log-likelihood of the underlying model. We refer to it as the **relaxation log-likelihood**. Precisely the same relationship exists between the relaxation free-energy and the relaxation log-likelihood as does between the true free-energy and log-likelihood.

Evaluation of the Hessian of $\ell_\beta(\theta)$ proves to be notationally challenging. Rose (1998) suggests an alternative which is more tractable and which we shall adopt. We consider a perturbation $\epsilon\delta_m$ applied to each of the means μ_m^* respectively, with $\delta_m = 0$ for all but the identical components. We then evaluate the derivative $\frac{d^2}{d\epsilon^2}\ell_\beta(\{\pi_m^*\}, \{\mu_m^* + \epsilon\delta_m\})$ at the point in question. This is equivalent to finding the projection of the Hessian on the direction defined by the perturbation δ_m.

We begin with the first derivative.

$$\frac{d}{d\epsilon} \sum_i \log \sum_l \pi_l^{*\beta} e^{-\frac{1}{2}\beta\|x_i-\mu_l^*-\epsilon\delta_l\|^2} = \sum_i \sum_l \frac{\pi_l^{*\beta} e^{-\frac{1}{2}\beta\|x_i-\mu_l^*-\epsilon\delta_l\|^2}}{\sum_k \pi_k^{*\beta} e^{-\frac{1}{2}\beta\|x_i-\mu_k^*-\epsilon\delta_k\|^2}} \beta\delta_l^T(x_i - \mu_l^* - \epsilon\delta_l)$$

$$= \sum_i \sum_l \beta r_{l,i} \delta_l^T (x_i - \mu_l^* - \epsilon \delta_l) \tag{3.20}$$

with the responsibilities evaluated at the perturbed θ. We note that when $\epsilon = 0$ we can write this derivative as $\beta \sum_l \delta_l^T \left(\sum_i r_{l,i}^* x_i - \mu_l^* \sum_i r_{l,i}^* \right)$ which is always zero by (3.17). This simply verifies that parameters which satisfy the recurrence relations (3.15)–(3.17) are indeed stationary points of the relaxation log-likelihood.

The second derivative is

$$\frac{d}{d\epsilon} \sum_i \sum_l \beta r_{l,i} \delta_l^T (x_i - \mu_l^* - \epsilon \delta_l) = \sum_i \sum_l \left(\beta \frac{dr_{l,i}}{d\epsilon} \delta_l^T (x_i - \mu_l^* - \epsilon \delta_l) - \beta r_{l,i} \|\delta_l\|^2 \right) \tag{3.21}$$

with the derivative of the responsibility given by

$$\begin{aligned}
\frac{dr_{l,i}}{d\epsilon} &= \frac{d}{d\epsilon} \left(\frac{\pi_l^{*\beta} e^{-\frac{1}{2}\beta \|x_i - \mu_l^* - \epsilon \delta_l\|^2}}{\sum_k \pi_k^{*\beta} e^{-\frac{1}{2}\beta \|x_i - \mu_k^* - \epsilon \delta_l\|^2}} \right) \\
&= \frac{\pi_l^{*\beta} e^{-\frac{1}{2}\beta \|x_i - \mu_l^* - \epsilon \delta_l\|^2} \beta \delta_l^T (x_i - \mu_l^* - \epsilon \delta_l)}{\sum_k \pi_k^{*\beta} e^{-\frac{1}{2}\beta \|x_i - \mu_k^* - \epsilon \delta_l\|^2}} - \\
&\quad \frac{\pi_l^{*\beta} e^{-\frac{1}{2}\beta \|x_i - \mu_l^* - \epsilon \delta_l\|^2} \sum_j \pi_j^{*\beta} e^{-\frac{1}{2}\beta \|x_i - \mu_j^* - \epsilon \delta_l\|^2} \beta \delta_j^T (x_i - \mu_j^* - \epsilon \delta_j)}{\left(\sum_k \pi_k^{*\beta} e^{-\frac{1}{2}\beta \|x_i - \mu_k^* - \epsilon \delta_l\|^2} \right)^2} \\
&= \beta r_{l,i} \left(\delta_l^T (x_i - \mu_l^* - \epsilon \delta_l) - \sum_j \beta r_{i,j} \delta_j^T (x_i - \mu_j^* - \epsilon \delta_j) \right)
\end{aligned} \tag{3.22}$$

Combining these equations we arrive at

$$\begin{aligned}
\frac{d^2}{d\epsilon^2} \ell_\beta (\theta^\epsilon) &= \beta^2 \sum_i \sum_l r_{l,i} \left(\delta_l^T (x_i - \mu_l^* - \epsilon \delta_l) \right)^2 - \beta^2 \sum_i \left(\sum_l r_{l,i} \delta_l^T (x_i - \mu_l^* - \epsilon \delta_l) \right)^2 \\
&\quad - \beta \sum_i \sum_l r_{l,i} \|\delta_l\|^2
\end{aligned}$$

and so, evaluating at $\epsilon = 0$ and exploiting the facts that $\delta_l = 0$ for $l > 2$ and that the means and responsibilities of components 1 and 2 are identical by construction.

$$\begin{aligned}
\frac{d^2}{d\epsilon^2} \ell_\beta (\theta^*) &= \beta \sum_l \delta_l^T \left(\beta \sum_i r_{l,i}^* (x_i - \mu_l^*)(x_i - \mu_l^*)^T - \sum_i r_{l,i}^* \right) \delta_l \\
&\quad - \beta^2 \sum_i \left(r_{1,i}^* (x_i - \mu_1^*)^T \sum_l \delta_l \right)^2
\end{aligned} \tag{3.23}$$

The second term in this expression, a sum of squares, is always non-negative. We can force it to 0 by choosing the perturbations so that $\sum_l \delta_l = 0$. The first part will be negative for all choices of δ as long as the matrix $\beta \sum_i r_{l,i}^* (x_i - \mu_l^*)(x_i - \mu_l^*)^T - \sum_i r_{l,i}^*$ is negative definite. Let $\sigma_{l,s}$ be the sth eigenvalue of the matrix $\sum_i r^*_{l,i}(x_i - \mu^*_l)(x_i - \mu^*_l)^T / \sum_i r^*_{l,i}$. The condition for negative definiteness is thus

$$\beta < \frac{1}{\max(\sigma_{l,s})}; l = \{1, 2\} \tag{3.24}$$

This condition is both necessary and sufficient for the solution θ^* with components 1 and 2 identical to be a stable maximum. We have shown that if it holds then the derivative of (3.23) is negative for any choice of δ_m. If it fails we can choose δ_1 and δ_2 pointing in opposite directions along the eigenvector corresponding to the largest $\sigma_{l,s}$ so as to obtain a positive Hessian.

Thus, a critical temperature is reached whenever the temperature β^{-1} becomes smaller than the leading eigenvalue of the covariance of the data assigned to any of the mixture's components. If we interpret the parameter β^{-1} as the effective scale of the covariance matrix of each Gaussian, this result is intuitively appealing. When the observed covariance of the data assigned to a component becomes larger than the component can "handle", a transition to more distinct component centers occurs.

3.4.2 Model-size

It is tempting to interpret the phase transition structure of relaxation models as indicating a progressive change in the underlying model-size (for example, the number of components in a mixture). Take the mixture model shown in figure 3.1, for example. Initially, only one distinct set of component parameters exists, and we might think of the mixture as containing only that one component. As the relaxation progresses, each phase transition introduces more distinct component values. We would like to view these as new components being added to the mixture, thus growing the underlying model-size.

Unfortunately, under the REM-1 algorithm (as well as the basic deterministic annealing algorithm), such an interpretation does not hold up. In the ground-state ($\beta = 1$) mixture likelihood, if two components, say the first two, have identical parameters, so that $\mathsf{P}_1(x_i) = \mathsf{P}_2(x_i)$, they may be replaced by a single component with the same parameters and mixing proportion $\pi_1 + \pi_2$ without any change in the likelihood. This is made clear by inspection of the likelihood

$$\ell_{\mathcal{X}}(\theta) = \sum_i \log \sum_m \pi_m \mathsf{P}_m(x_i) \tag{3.25}$$

In particular, if the larger model is at a maximum in the likelihood, then the smaller one will be too.

This convenient behaviour does not carry through to higher temperatures. Recall the form of the relaxation log likelihood

$$\ell_{\mathcal{X},\beta}(\theta) = \sum_i \log \sum_m \pi_m^\beta \mathsf{P}_m(x_i)^\beta \tag{3.26}$$

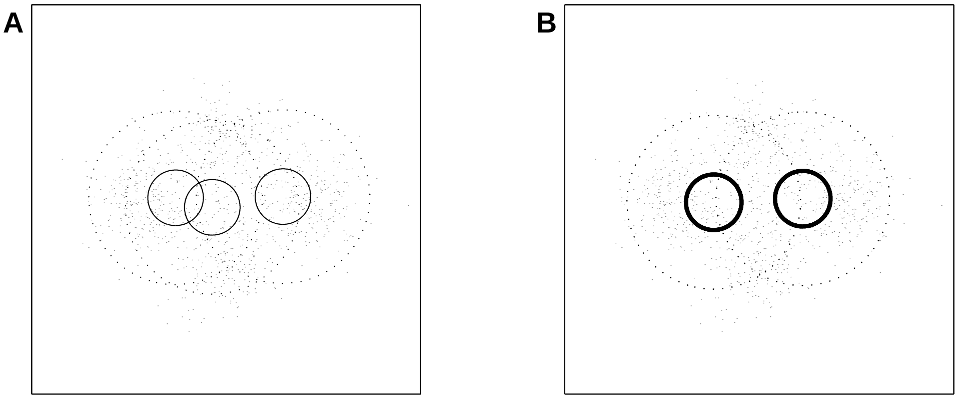

Figure 3.2: Inequivalence of different size models

Clearly, with $\beta < 1$ we cannot replace the identical components as before, since $\pi_1^\beta + \pi_2^\beta \neq (\pi_1 + \pi_2)^\beta$. Nor can we simply set the mixing proportion of the new component to $(\pi_1^\beta + \pi_2^\beta)^{1/\beta}$, since this violates the normalization of $\mathsf{P}_\theta\left(y_i\right)$. In general, then, the relaxation likelihood changes between the two models. Furthermore, a maximum in the more complex model may not correspond to a maximum in the simpler one, indeed the number of distinct component values in the two models may not be the same.

Figure 3.2 illustrates the point. Panel A shows a maximum in the relaxation likelihood of a three-component mixture of unit Gaussians at the stage $\beta = 0.3$. Panel B shows the optimal configuration, at the same temperature, of a four-component mixture, which was constructed by replacing the rightmost component of the mixture of panel A with two identical Gaussians. Both visible contours in B represent two identical components (indicated by the dark lines — other than this the representation of the components is as in figure 3.1). Thus, the duplication of one component has, in effect, driven the relaxation of the mixture in reverse, to a smaller phase.

Thus, the view of the model changing in size during the relaxation process cannot be maintained consistently under REM-1.

A further issue emerges from this analysis. Consider the mixture of figure 3.2B, where a four component mixture is being fit, but where only two distinct component values are visible. How do we know how to distribute these duplicated components? Clearly, each choice will yield a different intermediate solution; but the final result may also be affected since subsequent phase transitions will be constrained by the availability of components. We would like to be able to introduce the additional component wherever it is needed, but we cannot "move" the component around without changing the likelihood landscape. The result is that the choice of how to group the various components, a choice that must be made at each phase transition, will affect the outcome of the relaxation process.

Both of these issues can be rectified by the introduction of a variant of the basic relaxation algorithm, which we call REM-2.

3.5 REM-2

It is instructive to examine the structure of the relaxation free-energy of REM-1 for clues to the origin of the inequivalence of different model-sizes described above. Recall that the term $Q(p, \theta)$ is the expected value of the joint data log-likelihood under the distribution p. Using the fact that $\ell_{\mathcal{X},\mathcal{Y}}(\theta) = \log\left(\mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right)\mathsf{P}_\theta\left(\mathcal{Y}\right)\right)$ we can write the free-energy of (3.6) as

$$F_\beta(p, \theta) = \beta\mathcal{E}_p\left[\log\mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right)\right] + \beta\mathcal{E}_p\left[\log\mathsf{P}_\theta\left(\mathcal{Y}\right)\right] - \mathcal{E}_p\left[\log p\right] \qquad (3.27)$$

If we introduce a new hidden state, we increase the entropy of the latent variables. However, provided the new state is identical to some old one, the cross-entropy $-\mathcal{E}_p\left[\log\mathsf{P}_\theta\left(\mathcal{Y}\right)\right]$ decreases by the same amount. When $\beta = 1$, then, such an addition has no net effect on the free-energy. However, at higher temperatures the free-energy increases with the introduction of the new state. The size of this increase depends on both p and θ and so the location of the maxima of the free-energy may also change, as we saw above.

This formulation suggests a resolution of the difficulty. We introduce a slightly different relaxation free-energy which will form the basis of our second Relaxation Expectation–Maximization algorithm (REM-2).

$$\begin{aligned} F'_\beta(p, \theta) &= \beta\mathcal{E}_p\left[\log\mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right)\right] + \mathcal{E}_p\left[\log\mathsf{P}_\theta\left(\mathcal{Y}\right)\right] - \mathcal{E}_p\left[\log p\right] \\ &= \beta Q'(p, \theta) - \mathsf{KL}[p(\mathcal{Y})\|\mathsf{P}_\theta\left(\mathcal{Y}\right)] \qquad (3.28) \end{aligned}$$

Here $\mathsf{KL}[f\|g]$ stands for the Kullback-Leibler divergence between the distributions f and g. This form no longer enjoys the analogy with the familiar free-energy of statistical physics. Nonetheless, from the point of view of optimization it provides just as valid a relaxation progression as does the more traditional form.

Again, we optimize each free-energy in the relaxation sequence using the EM approach of alternate optimizations with respect to p and with respect to θ. The E-step is derived in the same manner as before. We introduce a Lagrange multiplier λ enforcing the constraint $\int d\mathcal{Y}\; p(\mathcal{Y}) = 1$ to obtain

$$\begin{aligned} 0 &= \frac{\partial}{\partial p}\left(F'_\beta(p, \theta) - \lambda\int d\mathcal{Y}\; p(\mathcal{Y})\right) \\ &= \frac{\partial}{\partial p}\left(\int d\mathcal{Y}\; p(\mathcal{Y})(\beta\log\mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right) + \log\mathsf{P}_\theta\left(\mathcal{Y}\right) - \log p(\mathcal{Y}) - \lambda)\right) \qquad (3.29) \end{aligned}$$

from which, by the calculus of variations,

$$\begin{aligned} 0 &= \frac{\partial}{\partial p}\left(p(\mathcal{Y})(\beta \log \mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right) + \log \mathsf{P}_\theta\left(\mathcal{Y}\right) - \log p(\mathcal{Y}) - \lambda)\right) \\ &= \left(\beta \log \mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right) + \log \mathsf{P}_\theta\left(\mathcal{Y}\right) - \log p^*(\mathcal{Y}) - \lambda\right) - \frac{p^*(\mathcal{Y})}{p^*(\mathcal{Y})} \end{aligned} \tag{3.30}$$

and so

$$p^*(\mathcal{Y}) \propto \mathsf{P}_\theta\left(\mathcal{Y}\right)\left(\mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right)\right)^\beta \tag{3.31}$$

The multiplier λ ensures that p is correctly normalized.

At first glance it might seem that the M-step, involving the maximization of $\beta\mathcal{E}_p\left[\log \mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right)\right]+$ $\mathcal{E}_p\left[\log \mathsf{P}_\theta\left(\mathcal{Y}\right)\right]$ will be different from standard EM and REM-1. In most models, however, the parameters θ can be partitioned into two disjoint and independent sets, one responsible for the distribution of the latent variables and the other for the conditional of the observables given the latent variables. If this is the case, F'_β can be optimized with respect to each of these sets separately, and clearly the resulting update rules will be exactly as in standard EM.

Now, when $\beta = 0$, this free-energy is optimized by any choice of p and θ for which $p(\mathcal{Y}) = \mathsf{P}_\theta\left(\mathcal{Y}\right)$. Although p need not be the maximum entropy distribution, the resulting parameter values are very similar to the initial conditions for REM-1. In particular, the distribution p must be independent of the observations $\mathcal{X}$. For the mixture model, for example, we have $r_{m,i} = \pi_m$, which implies that each component is fit with equal weight given to all of the data (although that weight may be different for the different components) and so all the component parameters are identical. For consistency with REM-1, and in the spirit of maximum entropy statistical methods where unknown distributions are assumed to be maximally uncertain, we will adopt the convention that the initial choice of parameters governing $\mathsf{P}_\theta\left(\mathcal{Y}\right)$ does indeed maximize the entropy of the latent variables under the constraints of the model. This is merely a convention, though. Any initial choice of $\mathsf{P}_\theta\left(\mathcal{Y}\right)$, provided every possible outcome has non-zero probability, will produce the same results.

In figure 3.3 the REM-2 algorithm is used to fit a 5-component mixture to the same data as was used in figure 3.1. This figure illustrates the fact that REM-2 exhibits the same type of phase transition structure as we saw previously in REM-1. Indeed, we can follow through the analysis of section 3.4.1 and find that exactly the same condition for stability holds, except that now the responsibilities that appear in (3.23) are those of the new algorithm

$$r_{m,i} = \frac{\pi_m e^{-\frac{1}{2}\beta\|x_i - \mu_m\|^2}}{\sum_l \pi_l e^{-\frac{1}{2}\beta\|x_i - \mu_l\|^2}} \tag{3.32}$$

(note that the mixing probabilities π_m are not raised to the power β). This results is a small change in the actual values of the critical temperatures between the two algorithms on the same data set;

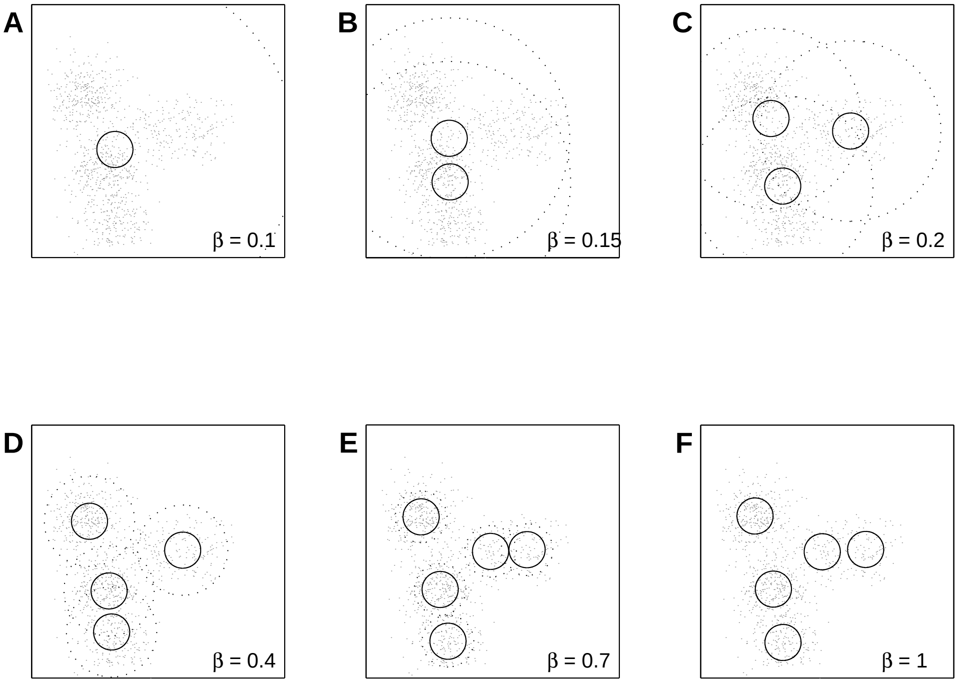

Figure 3.3: Phase transitions in REM-2 for fixed-variance Gaussians

an example of this is evident in a comparison of figures 3.3 and 3.1.

We can verify that the issues raised in section 3.4.2 are resolved by REM-2 by consideration of the implied relaxation likelihood for a mixture model.

$$\begin{aligned}\ell_\beta(\theta) &= F'_\beta(\{r_{i,m}\},\theta)\\ &= \beta\sum_i\sum_m r_{m,i}\log\mathsf{P}_m(x_i)+\sum_i\sum_m r_{m,i}\log\pi_m-\sum_i\sum_m r_{m,i}\log r_{m,i}\\ &= \sum_i\sum_m r_{m,i}\log\frac{\pi_m\mathsf{P}_m(x_i)^\beta}{r_{m,i}}\\ &= \sum_i\sum_m r_{m,i}\log\sum_l\pi_l\mathsf{P}_m(x_i)^\beta\\ &= \sum_i\log\sum_l\pi_l\mathsf{P}_m(x_i)^\beta \qquad (3.33)\end{aligned}$$

Clearly, the two identical components can be replaced by one (with mixing probability given by the sum of the weights of the duplicate components) without disturbing the likelihood. Thus, we can legitimately regard the model-size as increasing during the relaxation process. Furthermore, we need not make any choice about how to group components: any grouping will yield the same sequence of likelihoods and extra components can be assigned as needed when a critical temperature is reached.

3.6 Cascading Model Selection

In our development to this point, we have tacitly assumed that the size of the eventual model is known. If we use REM-1, the model size is set at the outset and maintained throughout. If we use REM-2, the model-size grows during the relaxation, but is capped at the correct value. In practice, however, this knowledge is often not available *a priori*. In using a mixture model for clustering, for example, we may not know in advance the appropriate number of clusters. Instead, the model-size needs to be learnt along with the parameters of the appropriate model.

This is an example of the more general problem of model selection. We have already visited this problem twice in the course of this dissertation. Section 1.3 discussed the general theory and described a number of likelihood-penalty techniques that are used in practice, as well as related approaches such as cross-validation. Section 2.7.3 added a further technique, called the Cheeseman-Stutz criterion, which is suitable for latent variable models such as mixtures. In this section we will investigate the relationship between these techniques and REM.

3.6.1 A natural answer?

It is tempting to think that in certain situations, the phase transition structure of REM provides a natural answer to such problems, and, indeed, a number of authors have assumed this (see, for example, Rose (1998) or Weiss (1998)). Take the mixture of unit Gaussians that has been our running example in this chapter. Suppose we were to fit by relaxation a mixture with a very large number of components. Once the relaxation had run its course, we would find that only a small number of distinct component values existed in the final mixture. Furthermore, whether we had used REM-1 or REM-2 to find that mixture, it would always be the case that at unit temperature the equivalence between a mixture with duplicate components and a smaller one with all duplications removed would hold. Thus, we can safely assert that the relaxation procedure has found a solution with limited model-size. Is this the correct model-size?

Unfortunately, despite the suggestions to that effect that appear in the literature, it is not. This should be clear from the fact that ultimately, the technique by which the final mixture was found is not important. That mixture is simply a maximum — with luck, the global maximum — of the model likelihood. Choosing a number of components in the manner suggested is thus the same as choosing between different models solely on the basis of their unpenalized likelihoods. Such a choice is prone to over-fit for all of the reasons that were discussed in section 1.3. The estimate of the model-size will be biased upwards.

We can drive the point home by means of a simple example. Suppose that the data to be modeled have actually arisen from a single Gaussian distribution with zero mean and unit covariance matrix. We attempt to model this data with a mixture of Gaussians, each with unit covariance, fitting the mixture by REM. As we have seen, at low values of the relaxation parameter, β, all of the mixture components coincide. However, once β reaches the inverse of the leading eigenvalue of the observed covariance matrix, more than one distinct mean will be observed. The eigenvalues of the observed covariance are asymptotically symmetrically distributed about 1 (the exact density is given by Anderson 1963). Thus, with a probability of approximately $1-2^{-p}$, where p is the dimensionality of the Gaussian, the leading eigenvalue will be greater than 1. In this case, the phase transition will occur with $\beta < 1$. If relaxation were to proceed to completion at $\beta = 1$, we would arrive at a solution with more than one component.

The situation is even more dire for other latent variable models. For example, if the covariances of the Gaussians are unknown (and perhaps unequal) the maximum likelihood solution given a sufficiently large number of components has each component concentrated around exactly one data point, giving rise to as many distinct components as data. Clearly, this is not a reasonable solution.

Another suggestion is as follows. The relaxation procedure is carried out using a large number of components, just as before. Now, however, a section of the data — a validation set — is held out and the (relaxation) likelihood of the optimal model at each temperature is evaluated on these data. After relaxation is complete, we select the model at which the validation likelihood was greatest.

This scheme is only meaningful in situations where the relaxation likelihood corresponds to an actual model. Even in such situations, though, it will tend to return the wrong answer; in this case the bias appears in the parameter estimates. Take the simple example of data from a single Gaussian. It is plausible that this scheme would correctly identify the optimal model-size as containing only one component. However, selecting this component will require choosing a solution at a non-unit temperature. Thus, the Gaussian will have a larger variance than appropriate.

The resolution would appear to be to use a model selection scheme (validation in this example) to choose the model-size, but then continue to relax the model of this size to unit temperature. We shall discuss a local version of this scheme in the next section.

3.6.2 Cascading model selection

Careful consideration of the nature of the relaxation likelihood has indicated that, despite the appealing natural limits that appear in the fixed-variance models commonly used in conjunction with deterministic annealing, to avoid bias the model-size must be chosen by a more traditional model selection technique. Nonetheless, the hierarchical “division” due to the phase transition structure that we saw in the case of the mixture model does still form an attractive basis for model selection. We shall see that it is indeed possible to exploit this structure. Through a progressive development we will arrive at an efficient method for choosing the correct model size, within the relaxation framework, that we call **cascading model selection**.

In what follows we shall consider the mixture model, with the selection of model-size being

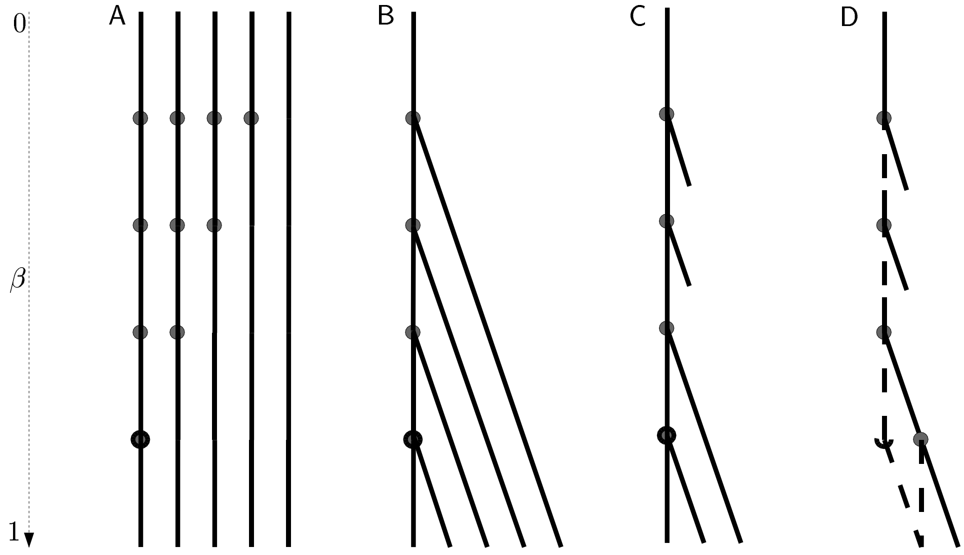

Figure 3.4: Schematic of model selection using REM

equivalent to choosing the correct number of components. The method is, however, quite general and can be applied with ease to any latent variable model for which an EM algorithm can be written.

The standard approach to model selection is as follows. Using some algorithm, which might just as well be REM, we obtain maximum likelihood fits for a variety of models with differing numbers of components. These models are then compared using one the methods discussed in sections 1.3 or 2.7.3. Many of these methods involve a comparison of the maximal log-likelihood values of the different models, reduced by a term that reflects the number of free parameters in the model. It is such penalized-likelihood methods that we shall consider first.

The various model selection schemes that we will discuss are shown schematically in figure 3.4. Panel A represents the basic procedure. The solid lines each represent the relaxation of a model, while the circles indicate the occurrence of phase transitions. The five models being fit are of different sizes, which is why they undergo different numbers of phase transitions. Roughly speaking, the total length of the lines in each panel represents the computational cost associated with each model selection strategy. The remaining panels will be described below.

If the optimization is carried out using REM-2 then the process of fitting the different size models can be made considerably more efficient. The relaxation process for models with M and with $M+1$ components is identical until the final phase transition of the larger model. Thus, there is no need to repeat the fitting process up to that point. As a result, we fit all of the models in a linear tree structure, shown in figure 3.4B, with a new branch emerging at each phase transition. (The schematic adopts the convention that the line emerging on the right of the circle has not undergone the phase transition, while the one that continues below has.) We note that this process is not possible with either the conventional deterministic annealing algorithm or REM-1.

We can improve further on this scheme by allowing early pruning of some branches. This is facilitated by the following important result, which holds for models being fit by REM-2. Suppose we have an M component model in which one component is unstable in the sense of section 3.4.1, that is, if additional components are available it would undergo a phase transition. We compare the likelihoods of two models: $\mathcal{M}_1$ has only M components and therefore exhibits no phase transition, while $\mathcal{M}_2$ has a model-size of $M+1$ and thus has allowed the unstable component to "split". If the relaxation log-likelihood at some $\beta < 1$ of $\mathcal{M}_2$ exceeds that of $\mathcal{M}_1$ by Δ, then the final log-likelihood of $\mathcal{M}_2$ will exceed that of the smaller model by an amount larger than Δ. We offer an informal proof of this point.

Recall first that $\mathcal{M}_1$ is identical in likelihood to an $(M+1)$-component model $\mathcal{M}_{1*}$ in which the unstable component is duplicated, but both copies retain the same parameters. By assumption the relaxation log-likelihood of $\mathcal{M}_2$ exceeds that of $\mathcal{M}_{1*}$. Recall that this log-likelihood is obtained from the free-energy

$$F'_\beta(p,\theta) = \beta Q'(p,\theta) - \mathsf{KL}[p(\mathcal{Y}) \| \mathsf{P}_\theta(\mathcal{Y})] \tag{3.28}$$

by setting $p(\mathcal{Y}) = \mathsf{P}_\theta(\mathcal{Y} \mid \mathcal{X})$. Now it must be the case that the Kullback-Leibler term for $\mathcal{M}_2$ is greater than that for $\mathcal{M}_{1*}$. If that were not true, the more complex model would be preferred even at $\beta = 0$, which we know not to be the case. Thus, it must also be true that the Q' term in the likelihood of $\mathcal{M}_2$ exceeds that of $\mathcal{M}_{1*}$ (and thus of $\mathcal{M}_1$).

How will the log-likelihoods of the two models change as relaxation progresses? Let $\ell_\beta(\theta^*)$ be the optimal relaxation log-likelihood, that is, the value of $F'_\beta(p,\theta)$ with $\theta = \theta^*$, the optimal parameters, and $p(\mathcal{Y}) = \mathsf{P}_{\theta^*}(\mathcal{Y} \mid \mathcal{X})$. The maximizing value of the model parameter vector, θ^*, is, of course, a function of the relaxation parameter β. Thus, we may differentiate the maximal log-likelihood with respect to β using the chain rule

$$\frac{d}{d\beta}\ell_\beta(\theta^*) = \frac{\partial}{\partial\beta}\ell_\beta(\theta^*) + \frac{\partial}{\partial\theta}\ell_\beta(\theta^*)\frac{d\theta^*}{d\beta} \tag{3.34}$$

But, since θ^* maximizes the log-likelihood, the gradient of $\ell_\beta(\theta)$ at θ^* for fixed β is 0. The partial with respect to β is obtained trivially from (3.28), and thus we find that

$$\frac{d}{d\beta}\ell_\beta(\theta^*) = Q'(\mathsf{P}_{\theta^*}(\mathcal{Y} \mid \mathcal{X}), \theta^*) \tag{3.35}$$

We have argued that the Q' term for $\mathcal{M}_2$ is greater than that for $\mathcal{M}_1$. Thus, we find that the optimal log-likelihood of the larger model is growing more rapidly than that of the smaller one (if both gradients are negative, then it is shrinking less rapidly). As a result, any difference in likelihoods at $\beta < 1$ can only grow as β increases.

Thus, it is possible to further streamline the model selection process. If, at any stage in the relaxation, the penalized relaxation log-likelihood of some model is exceeded by that of a larger model (that is, the difference in log-likelihoods is greater than the difference in penalties) we can immediately neglect the smaller model, effectively pruning that branch of the tree. This is indicated in figure 3.4C, where the first two models are pruned.

Finally, we arrive at the approach that we call **cascading model selection**. We assume that the penalized likelihood rises monotonically with model-size until the optimal value is reached. While this is not guaranteed to be the case, it is an intuitively appealing assumption and the experiments below suggest that, at least for simple mixture models, it is typically valid. Under these conditions, we need not even consider a model of size $M+2$ until the model with M components has been rejected in favour of one with $M+1$.

In our implementation of cascading model selection we think of a particular model size as being "current" at all times. This is indicated by the solid line in figure 3.4D. When a critical temperature is reached, the current model retains its size. However, we begin to track the optimum of a "shadow" model of larger size (and thus, which undergoes the phase transition). If the penalized likelihood of this shadow model exceeds that of the current one, we abandon the current model and make the shadow current. Sometimes, it will be the case that the shadow model reaches a critical temperature without having replaced the current model. If this happens, we simple maintain the shadow model's size and continue to relax; we do not introduce the larger model.

It might also be the case that the current model will encounter another critical temperature, even though it remains more likely than the shadow. In this case we need to introduce another shadow model, usually of the same model-size as the previous one, but resulting from a different phase transition. In the case of the mixture model, it is useful to think of a different component having "split". If, as relaxation progresses, we reach a point where either of these shadow models becomes more likely than the current one, we make that model current and abandon all the others.

The cascading model selection procedure is capable of find optima that the basic REM algorithm is not. To see why, consider the case described above where a second shadow model may be introduced. This shadow model is different from any that might be obtained by REM; to achieve it we have "disallowed" one phase transition but allowed another. If this model proves to have greater likelihood than the first shadow, and also to be preferred to the current model according to the penalized likelihoods, then we will arrive at a model with greater likelihood than that obtained by REM with the same number of components. Intuitively, the cascading model selection prevented us from "wasting" a component due to the phase transition at the higher temperature, instead reserving it for the more advantageous split. This point will be illustrated below.

Finally, we note that the core result of cascading model selection has been obtained only for a penalized likelihood style model selection procedure. However, to the extent that such methods approximate techniques such as Bayesian model selection or cross-validation, we might believe that such techniques can be used in the same way. In particular, for mixture models the Cheeseman-Stutz criterion of section 2.7.3 often provides good results.

3.7 Experiments

As we first encountered the REM algorithm in section 3.3, we noted that, because the maximum of the free-energy does not, in fact, vary continuously with the relaxation parameter, the algorithm process cannot be guaranteed to find the global optimum of the likelihood. Instead, we appealed to an intuitively founded expectation that it would tend to find a good optimum. In this section we examine the results of numerical experiments to see if this is actually the case.

The experiments described here all involve the simple mixture of two-dimensional unit Gaussians model, which we have seen throughout this chapter. In all cases the relaxation is performed using the REM-2 algorithm. The basic outline of the experiments is as follows: we select a random mixture of unit Gaussians, generate data from it, and fit mixture models to these data using both the REM-2 and standard EM algorithms. We then compare the performance of the algorithms by computing the likelihoods of the resultant models. Any solution in which the likelihood of the fit model is greater than the likelihood of the true (that is, data-generating) model will be called "good."

The parameters of the generating mixture are all chosen randomly within pre-specified intervals. The number of components, M, is chosen from the discrete uniform distribution on the values 3, 4, 5 and 6. The mixing proportions are chosen by randomly partitioning the interval $(0, 1)$ as follows: $M - 1$ numbers in the interval $(0, 1)$ are chosen from a uniform distribution on the interval and then ordered, thereby inducing a partition into M subintervals; the lengths of these subintervals are taken to be the mixing probabilities. The means are generated from the two-dimensional uniform distribution on the rectangular region bounded by ± 5 in both dimensions. The covariances are all set to the identity matrix.

500 data points are generated randomly from this mixture distribution. Mixtures of the correct number of Gaussians are then fit both by REM-2 and by standard EM. For each data set, the standard EM algorithm is started 10 times, from 10 randomly selected initial conditions (see section 2.7.2). Both algorithms are iterated to the same convergence criterion, which is that the relative change in likelihood after a complete EM step should fall below 10^{-7}. The likelihoods of all of the models, including the generating one, are then evaluated. We call a fit model "poor" if its likelihood is less than that of the generating model on the given data.

This entire procedure is repeated for 200 different generating mixtures.

Figure 3.5 shows the number of "poor" optima achieved under the different algorithms. The 10 bars on the left show how the rate of success of the standard EM algorithm increases as a

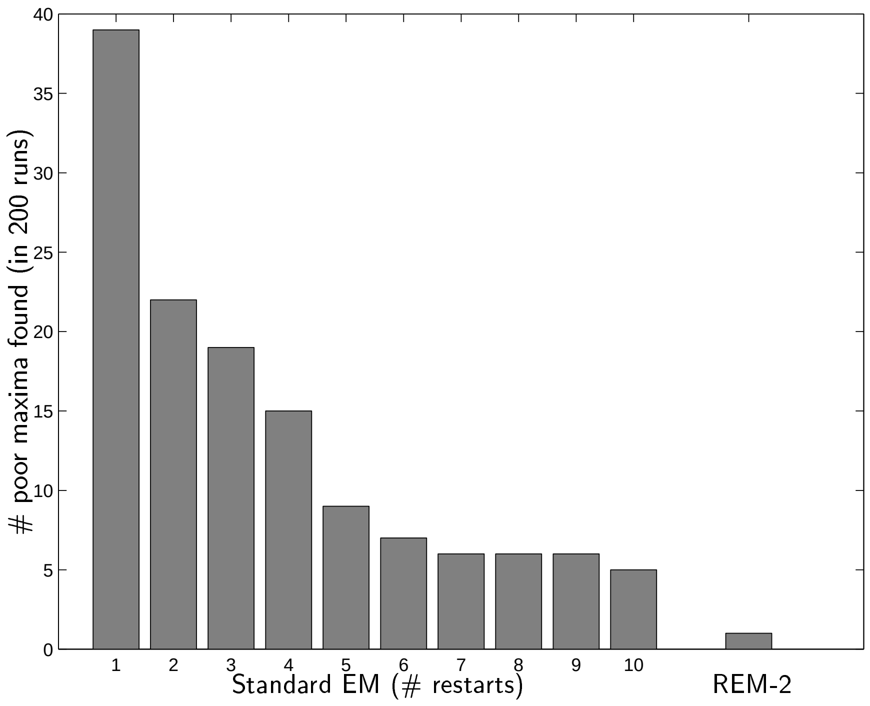

Figure 3.5: Frequency of poor maxima

progressively larger number of restarts are used. The likelihood used in the calculation of the bar labeled n is the largest of the likelihoods obtained from the first n restarts. The single bar on the right indicates that, for REM-2, only a single run achieved a poor optimum.

It is instructive to examine the single example in which REM-2 converged to a poor maximum. This is shown in figure 3.6. Panel A shows the model from which the data were generated. Panel B shows the optimum found by the REM-2 algorithm. Evidently, a phase transition that split the component in the middle-right was encountered before the phase transition that would correctly split the bottom-left component. In panel C we show the results of running REM-2 in conjunction with cascading model selection (using the BIC likelihood-penalty with no corrective constant). Whereas the standard REM-2 algorithm ran on a model with the correct number of components provided *a priori*, with cascading model selection this number could be determined from the data. Furthermore, it is evident that by incorporating on-line model selection, the early phase transition was rejected on the basis of the penalized likelihood , whereas the later, correct, one was subsequently accepted. It should be clear that without the cascading property this maximum could not have been found: had the different model sizes been compared after optimization (as is usual) then the model of size 5 would have been that of panel B. Thus, we observe that — as was suggested at the end of section 3.6 — besides the obvious benefits of automatic model size determination, the cascading model selection process can sometimes improve the optima found by REM.

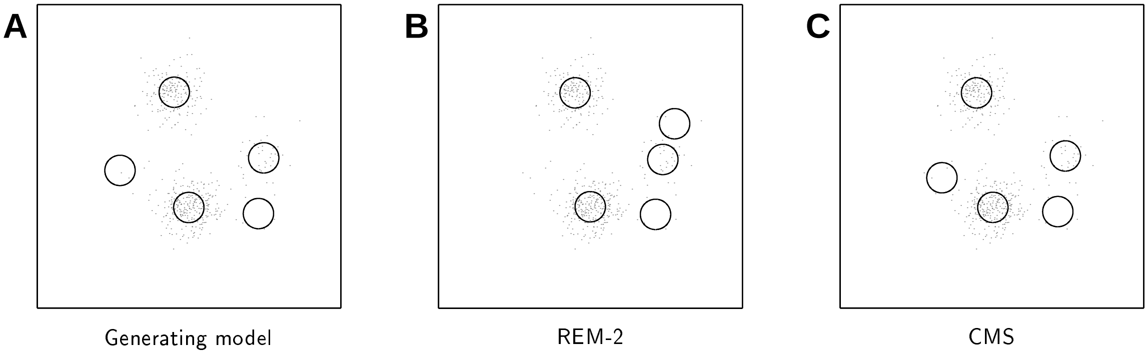

Figure 3.6: Cascading model selection can improve optima

A word of caution needs to appear here. The data shown in figure 3.5 suggest that, at least in this simple case, REM — perhaps in combination with cascading model selection — might well converge reliably to the global maximum of the likelihood. This is not actually the case. Closer inspection reveals that for 11 of the random mixtures at least one of the standard EM runs found a model with a likelihood more than 10^{-4} log-units larger than that found by REM-2. Furthermore, it is possible that even for the remaining mixtures the relaxation solution is not globally optimal, but that none of the standard EM iterations found the maximum either. Thus, REM does not always find the global optimum; indeed we cannot expect any algorithm of polynomial complexity to reliably do so. Nonetheless, figure 3.5 does suggest that it tends to find an optimum at least as good as the model that actually generated the given data with remarkable regularity.