Chapter 1 Latent Variable Models

1.1 Statistical Modeling

We are given a set of observations $\mathcal{X} = \{x_i \mid i = 1 \ldots |\mathcal{X}|\}$. The data x_i may be multivariate and are not necessarily independent. We are interested in learning about the nature of the process that gave rise to these data. In particular, we believe that by investigating the relationships that exist between the various components of the x_i, or between the different x_i, we can arrive a succinct description of the data, and that this description will reveal the structure of the generating process. In this quest we shall follow a path that lies at the intersection of two fields: unsupervised learning and density estimation.

In the machine learning literature, the project that we have laid out is known as **unsupervised learning**. We shall focus on a subset of the machine learning techniques, defined by our belief that the underlying generative process is **stochastic**, where we seek to learn an explicit probabilistic model that describes the data. This will exclude from our purview some effective techniques, for example the Kohonen and ART networks; in general, however, there are probabilistic formulations that very closely resemble each of these, and so we expect the loss not to be too serious. In return, we gain access to a powerful collection of probabilistic analysis tools.

Thus, we seek a description of the probability distribution (or density, for continuous observations) function $\mathsf{P}(\mathcal{X})$[1]. As such, our objectives are similar to those of the field of **density estimation**. However, it is not the resultant distribution (or density) function that holds our interest, but rather the structure of the function and what that structure reveals about the process that generated the data. Thus, we will not pursue many useful, "non-parametric" techniques of density estimation on the basis that these will give us little insight into the underlying process.

It is important to note that the general task of density estimation – given data $\mathcal{X}$, estimate $\mathsf{P}(\mathcal{X})$ – is not well formed unless something is known *a priori* about the probability function. This prior knowledge may be as simple as a belief that the function must be smooth, but in the absence of any prior, any scheme for ranking two candidate distributions will fail at least as often as it will succeed. This point is made clearly by Wolpert (1996). In our case, the prior knowledge, dictated by scientific experience and intuition, will go towards the selection of one or more families of parameterized probability functions $\mathsf{P}_\theta(\mathcal{X})$. θ here denotes a set of parameters, each of which

[1] We shall use the notation $\mathsf{P}(\cdot)$ generically for probability distribution and density functions. The exact nature of the function should be clear from context and the arguments provided, when this is not so we shall identify particular functions with a subscript such as $\mathsf{P}_\theta(\cdot)$

may be discrete or continuous. There are two central problems to be addressed in the project of statistical modeling: the first, called **learning** or **fitting**, is to estimate a suitable set of parameters $\widehat{\theta}$, or, if one is of the Bayesian persuasion, a posterior distribution over the parameters $\mathsf{P}(\theta \mid \mathcal{X})$, that is appropriate for the observed data. The second, **model comparison**, is to choose from among a group of candidate models the one which is better supported by, or more probable given, the data. It is worth noting that in the strict Bayesian viewpoint there is no difference between these operations: we can simply introduce a **hyper-parameter** that identifies which model is to be used and then infer its posterior distribution. However, we are interested in the properties of the particular model that best describes the data, and so although we might accept a distribution over parameters, we insist on identifying a single best model.

1.2 Parameter Estimation

We are given a set of observations $\mathcal{X}$, along with a parameterized family of probability functions $\mathsf{P}_\theta(\mathcal{X})$. We would like to infer an "optimal" value of the parameters such that the corresponding function describes the data best. There are many competing definitions of "optimal" in this context.

It will be simplest to survey these definitions by starting from the Bayesian viewpoint. In the Bayesian framework, the parameters θ are treated as random variables, to be handled on a similar footing to the observations $\mathcal{X}$. In this case we can more aptly write our family of distributions as $\mathsf{P}_M(\mathcal{X} \mid \theta)$, where the subscript M identifies the particular model. Bayes' rule then allows us to find a **posterior** distribution of the θ,

$$\mathsf{P}_M(\theta \mid \mathcal{X}) = \frac{\mathsf{P}_M(\mathcal{X} \mid \theta)\,\mathsf{P}_M(\theta)}{\mathsf{P}_M(\mathcal{X})} \tag{1.1}$$

The function $\mathsf{P}_M(\theta)$ denotes the probability associated with particular value of the parameters under the model M *a priori* – that is, without reference to any observations. It is called the **prior** distribution. Similarly, $\mathsf{P}_M(\theta \mid \mathcal{X})$ gives the probability of the parameter values θ in the context of the observed data. This is the *a posteriori* or simply **posterior** distribution. The term $\mathsf{P}_M(\mathcal{X} \mid \theta)$ is the familiar function that describes the distributions within our model, however in the context of parameter estimation by (1.1) it is best viewed as a function of θ, rather than of $\mathcal{X}$. In this context it is given a different name; it is called the **likelihood** of the parameters in light of the data, and will be written $\mathcal{L}_\mathcal{X}(\theta)$ to emphasize the exchange of rôles between θ and $\mathcal{X}$. It is important to note that the the numerical value of the probability of data $\mathcal{X}$ under parameters θ, $\mathsf{P}_\theta(\mathcal{X})$ or $\mathsf{P}(\mathcal{X} \mid \theta)$, is identical to that of the likelihood of parameters θ given data $\mathcal{X}$, $\mathcal{L}_\mathcal{X}(\theta)$. The difference is only one of interpretation. The final term in (1.1) is the denominator $\mathsf{P}_M(\mathcal{X})$. This is also given a name, but one that will only really be relevant when we discuss model selection below. It is called the

evidence for the model M, or else the **marginal likelihood**, since it is obtained by integrating the likelihood with respect to θ. From the point of view of parameter estimation from observations it is usually of little importance, as it has a constant value with no dependence on the parameters.

In the strict Bayesian point of view the equation (1.1) represents all that there is to be said about parameter estimation. Once we know the posterior distribution of the parameters we have exactly expressed the complete extent of our knowledge about their value. In this view, any attempt to provide a single parameter estimate as a description of the situation must give up some useful information. This is most apparent if we ask how the parameter estimate is to be used. Typically, we are interested in predicting the value of some statistic that is dependent on the parameters; it might, for example, be the next data point to be drawn from the distribution. In this case we need to integrate over the posterior (this will also be true for model selection, treated below). Let us call the statistic that we wish to predict k. The probability distribution that describes our prediction will be

$$\mathsf{P}_M\left(k \mid \mathcal{X}\right) = \int d\theta\, \mathsf{P}_M\left(k \mid \theta\right) \mathsf{P}_M\left(\theta \mid \mathcal{X}\right) \tag{1.2}$$

Here we see the practical difficulty in the strict Bayesian point of view. For many models, this integral is impossible to compute exactly. One approach taken is to approximate the integral by a Monte-Carlo sampling technique such as the Gibbs or Metropolis samplers, or by various so-called "hybrid" Monte-Carlo methods (Gelfand and Smith 1990; Smith and Roberts 1993; Neal 1996). Such methods are asymptotically exact, although the number of samples needed to reach the asymptotic distribution can be probitively large.

In practice, we often use a single estimate of the values of the parameters. This approach may be understood from one of two points of view. In the first case, we will argue below that a suitable choice of estimate can, under certain circumstances, actually provide a reasonable approximation to the correct Bayesian predictor. In the second, it may be that the problem we are trying to solve requires a single estimate. If that is so, the problem will have introduced (perhaps implicitly) a **loss-function**, which we can then optimize to obtain the appropriate estimate.

In many cases the posterior distribution is very strongly peaked at its modal value, written θ^{MP} for *maximum a posteriori*. In this case we may assume that the only significant contribution to the integral comes from parameters very near the peak, and we may assume that the value of $\mathsf{P}_M\left(x \mid \theta\right)$ is approximately constant for these values of θ. Armed with these assumptions, along with the knowledge that $\int d\theta\, \mathsf{P}_M\left(\theta \mid \mathcal{X}\right) = 1$, we can make the approximation

$$\int d\theta\, \mathsf{P}_M\left(x \mid \theta\right) \mathsf{P}_M\left(\theta \mid \mathcal{X}\right) \approx \mathsf{P}_M\left(x \mid \theta^{\mathsf{MP}}\right) \tag{1.3}$$

That is, calculations made by simply plugging in the MAP estimator in the parameterized density approximate the Bayesian results. In general, this approximation improves with the number of

available data. The MAP value is also important in other, more accurate, approximations to the posterior which are based on the Laplace or saddle-point integral. In these techniques, the posterior is approximated by a Gaussian whose mean lies at the posterior mode and whose covariance is in the inverse of the Hessian of the posterior with respect to the parameters, evaluated at the mode (MacKay 1992). We will treat these in greater detail when we discuss model selection.

The MAP estimator maximizes the posterior (1.1). The denominator on the right hand side of Bayes' rule does not depend on θ, and so the maximization applies only to the numerator $\mathsf{P}_M(\mathcal{X} \mid \theta)\,\mathsf{P}_M(\theta)$. In many situations we may choose to neglect the prior and maximize only the first factor, the likelihood. This yields the maximum-likelihood or ML estimate, θ^{ML}. The ML estimate occupies an extremely prominent position in the classical (non-Bayesian) approach to statistics. In particular, the ML estimate can be shown to be asymptotically efficient, that is, as the sample size grows the expected square error of the ML estimate approaches the fundamental lower bound on such errors (known as the Cramér-Rao bound). In the presence of a "vague" prior (for example, a uniform prior in cases where this is well defined) the ML estimate enjoys all the properties of MAP estimator discussed above.

The MAP estimator can be seen to minimize the expected value of a probability-of-error loss function, which penalizes all errors equally. For continuous parameters we define the loss by the limit as $\epsilon \to 0$ of the function taking the value 0 in an ϵ-ball around the true parameter values and 1 elsewhere. An alternative loss function penalizes errors by the square of the departure from the true value. Minimizing the expected value of this loss leads to the minimum-square-error (MSE) estimator. The fact that the second moment of any distribution is smallest about its mean implies that the MSE estimator is the mean of the posterior. Finding this value may well involve integration of the posterior, with all its attendent pratical difficulties. The result about the asymptotic efficiency of the ML estimator quoted above implies that as the number of data grow larger the mode and mean of the posterior must converge.

We have argued that the MAP and ML parameter estimates are of considerable importance in statistical theory, either as legimate goals in their own part, or as inputs to approximations of Bayesian integrals. In much of this dissertation we shall focus on maximum-likelihood techniques, tacitly assuming a vague prior. In almost all cases, (in particular, in the EM algorithm that we shall encounter shortly and which will resurface throughout this dissertation) the techniques that we will discuss can easily be adapted in the presence of a strong prior to yield a MAP estimate.

1.3 Model Selection

We now consider the situation in which we do not have a single parameterized family of probability functions, but rather must choose between alternative families with different (and perhaps different

numbers of) parameters. These families might be very closely related. For example, we will discuss clustering models at some length in chapter 2, where the data are presumed to arise from some number of distinct distributions, one for each cluster. In this case we shall need to determine the appropriate number of clusters, given the data. This is a model selection problem.

Hyperparameters and stacked generalization

One approach to the model selection problem is to ignore it. We can combine the models into a single family, with a **hyperparameter** that selects between them. The parameter set is then the union of the parameters of the different models, along with the hyperparameter. In the case of **nested** models, where one family is a proper subset of the other, this is almost the same as selecting the most complex model with the addition of the new hyperparameter. If we proceed with the full Bayesian predictive procedure (1.2) this is, in fact, the correct approach. In the case of clustering, for example, we should use an unbounded number of clusters (Neal 1991). However, with such models, the posterior distribution will tend to be far more complex than with a single, continuously parameterized family. In particular, we expect modes corresponding to the MAP estimator for each model, along with the corresponding value of the hyperparameter. In the face of sufficient data one of these modes is likely to dominate, in which case we will have selected one model after all. With less data, we generally need to integrate this posterior, for example when making predictions, by Monte-Carlo means (Neal 1991).

A related approach, now termed **stacked generalization**, was proposed by Stone (1974) and has recently been explored by Wolpert (1992) and Breiman (1996). We can explicitly write the marginal of the predictive density over the model selection hyperparameter. If the models are labelled $\mathcal{M}_i$ this is

$$\mathsf{P}\left(k \mid \mathcal{X}\right) = \sum_i \mathsf{P}\left(\mathcal{M}_i \mid \mathcal{X}\right) \mathsf{P}_{\mathcal{M}_i}\left(k \mid \mathcal{X}\right) \tag{1.4}$$

where the rightmost factor is the predictive distribution derived from the ith model. Thus, the predictive distribution is the weighted sum of the predictions made by the different models. The weighting factor for the ith model is given by Bayes' rule,

$$\mathsf{P}\left(\mathcal{M}_i \mid \mathcal{X}\right) \propto \mathsf{P}\left(\mathcal{X} \mid \mathcal{M}_i\right) \mathsf{P}\left(\mathcal{M}_i\right) \tag{1.5}$$

that is, it is proportional to the product of the evidence or marginal likelihood $\mathsf{P}\left(\mathcal{X} \mid \mathcal{M}_i\right) = \mathsf{P}_{\mathcal{M}_i}\left(\mathcal{X}\right)$ and the prior probability of the model. The weights are normalized to add to one.

Choosing one model: the dangers of maximum likelihood

Such combined model approaches are attractive in situations where the goal is predictive, and the true family is unknown. In the case of statistical modeling as we have laid it out, however, we are often interested in identifying the particular model that is best supported by the data. In the example of clustering, one of our goals may well be to determine how many classes are present. If we are content with a probabilistic answer, then the marginal likelihood, or evidence, described above, indicates the relative probabilities of each model, as long as the prior weighting of each model is equal. If not, we may elect to choose the most probable model, thereby tacitly assuming a zero-one loss function as in the case of the MAP parameter estimate. In the following discussion we shall assume the latter point of view, arguing for the selection of a single, most probable model; however most of the approximations we will discuss can equally well be used to estimate the posterior probabilities of various models and thus used in techniques such as stacked generalization.

Note that choosing the model with the greatest *marginal* likelihood is different from choosing the model with the greatest *maximum* in the likelihood, which might have been the naïvely favoured policy. In general, more complex models will exploit the greater flexibility of their parameterizations to achieve higher likelihood maxima on the same data; however, such models will be able to explain all sorts of different data by adjusting their parameters appropriately, and can thus only assign a relatively low probability to any particular data set. In other words, the complexity is penalized in the integral, as the region of parameter space that assigns high likelihood to the data is likely to be proportionately smaller. Thus, the Bayes approach leads to the selection of the *simplest* model, within the group considered, that is adequate to explain the data; as a result this approach has been compared with the philosophical "razor" of William of Ockham.

We can express the difficulty with maximum-likelihood model choice in another way. The maximal likelihood for a given model, represents the suitability of one particular member of the model family to describe the data. The member chosen depends critically on the data provided. If the model is complex, and two equivalent, independent samples from the same probability distribution are available, the member functions chosen in the two cases may be very different. In either case, the function may well be far from the true density.

An example appears in figure 1.1. To produce this figure, one dimensional data, shown as filled half-circles on the lower axis, were generated from the Gaussian density shown by the solid line. These data were fit by two different models: one, a simple Gaussian density with mean and variance estimated from the data; the other a three-component mixture of Gaussians (basically the weighted sum of three Gaussian densities). Both models were fit by maximum likelihood estimation (the details of fitting the mixture model will be discussed in a subsequent chapter). The optimal estimates are shown: the simple Gaussian estimate is plotted with dashes; the more complex mixture estiamte with dashes and dots — the faint dotted lines show the shapes of the three mixture components.

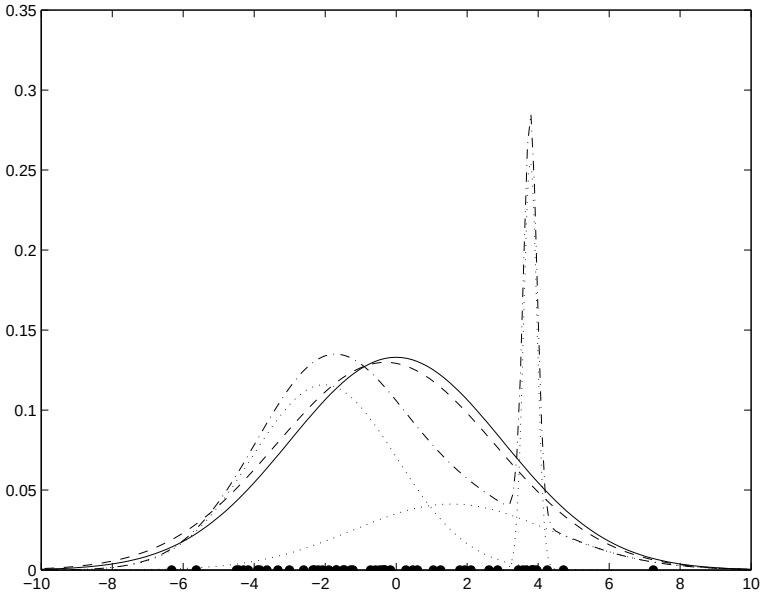


Figure 1.1: The dangers of over-fitting with a complex model.

The mixture model has a higher likelihood than the simpler one. In this case, the log likelihood per point for the simple model is -2.54, while that of the mixture model is -2.41. In part, this increase in likelihood has been achieved by adapting to the cluster of data that appears near the value 4, assigning high probability to this region. Different data, unlikely to cluster near 4, will probably yield a very different estimate.

It is obvious by inspection that the simple model has approximated the true density with greater accuracy. This tendency of complex models to fit the peculiarities of the given sample, rather than the underlying density funtion, is called **over-fitting**.

Bayesian analysis

We consider two candidate models, $\mathcal{M}_0$ and $\mathcal{M}_1$, to be used to describe the data $\mathcal{X}$. The two models have, respectively, p_0 and p_1 parameters, with $p_0 \leq p_1$. The parameter vectors will be written θ_0 and θ_1. In some cases we shall consider **nested** models, where the family of functions allowed under $\mathcal{M}_1$ is a proper superset of the functions available in $\mathcal{M}_0$. In this case we shall further assume that $\mathcal{M}_0$ can be obtained from $\mathcal{M}_1$ by fixing the values of $p_1 - p_0$ parameters, and that the remaining p_0 parameters of $\mathcal{M}_1$ are identical to the parameters of $\mathcal{M}_0$. Thus, $\mathcal{M}_1$ is to be thought of as the more complex model, and, in the nested case, may be a direct generalization of $\mathcal{M}_0$. The Bayesian model selection procedure (sometimes called **empirical Bayes**) dictates that we select model $\mathcal{M}_1$ if and only if the **posterior odds** in favour of $\mathcal{M}_1$, $\mathsf{P}\left(\mathcal{M}_1 \mid \mathcal{X}\right) / \mathsf{P}\left(\mathcal{M}_0 \mid \mathcal{X}\right)$ are greater than one.

Using Bayes' rule, we can write this as

$$\frac{\mathsf{P}\left(\mathcal{M}_1 \mid \mathcal{X}\right)}{\mathsf{P}\left(\mathcal{M}_0 \mid \mathcal{X}\right)} = \frac{\mathsf{P}_{\mathcal{M}_1}\left(\mathcal{X}\right)}{\mathsf{P}_{\mathcal{M}_0}\left(\mathcal{X}\right)} \times \frac{\mathsf{P}\left(\mathcal{M}_1\right)}{\mathsf{P}\left(\mathcal{M}_0\right)} \tag{1.6}$$

The second term on the right hand side of this expression is the **prior odds** of $\mathcal{M}_1$ being correct; the first term, which is the ratio of the marginal likelihoods, is called the **Bayes factor**. It is convenient to work with logarithms, and so the empirical Bayes criterion for selecting $\mathcal{M}_1$, in the face of equal prior probabilities for the two models (prior odds = 1), is

$$\log B_{10} = \log \mathsf{P}_{\mathcal{M}_1}\left(\mathcal{X}\right) - \log \mathsf{P}_{\mathcal{M}_0}\left(\mathcal{X}\right) > 0 \tag{1.7}$$

These are the same marginal likelihoods that appeared in the denominator of (1.1). While they do not play much of a rôle in parameter estimation, they can be seen to be vital to model selection.

The marginal likelihood is an integral over the parameter vector θ_i for the model $\mathcal{M}_i$,

$$\mathsf{P}_{\mathcal{M}_i}\left(\mathcal{X}\right) = \int d\theta_i \, \mathsf{P}_{\mathcal{M}_i}\left(\mathcal{X} \mid \theta_i\right) \mathsf{P}_{\mathcal{M}_i}\left(\theta_i\right) \tag{1.8}$$

As in the case of predictions using the posterior (1.2) this integral is often difficult to compute. Analytic solutions can be found for simple exponential family models, including multivariate normal linear regression models, with so-called **conjugate priors** on the parameters (these being priors chosen in part for the simplicity of the resulting integral). In the general case we need to estimate the integral via Monto-Carlo techniques (which we will not discuss here, but see Gelfand and Smith (1990), Smith and Roberts (1993) and Neal (1996)) or else employ analytic approximations which, while they may be asymptotically exact, yield biased estimates with realistic sample sizes.

Approximations to the Bayes factor

A simple and widely used approximation is called **Laplace's method** (Tierney and Kadane 1986; MacKay 1992). Let us write $\Phi(\theta)$ for the logarithm of the integrand in (1.8), the unnormalized posterior over the parameters. We have dropped the subscript i for simplicity. We can expand $\Phi(\theta)$ in a Taylor series about its maximum, which falls at θ^{MP}.

$$\Phi(\theta) = \Phi(\theta^{\mathsf{MP}}) + \nabla\Phi(\theta^{\mathsf{MP}}) \cdot (\theta - \theta^{\mathsf{MP}}) + \frac{1}{2}(\theta - \theta^{\mathsf{MP}})^T \nabla\nabla\Phi(\theta^{\mathsf{MP}})(\theta - \theta^{\mathsf{MP}}) + \ldots \tag{1.9}$$

where the notation $\nabla\nabla\Phi$ denotes the Hessian matrix of second derivatives $[\partial^2\Phi/\partial\theta_i\partial\theta_j]$ and should not be confused with the Laplacian, $\nabla^2\Phi = \mathrm{Tr}\left[\nabla\nabla\Phi\right]$. As θ^{MP} lies at a maximum of Φ, the gradient there is 0 and the linear term in the expansion vanishes. We ignore the terms of higher order than quadratic, a choice tantamount to approximating the posterior by a Gaussian, and write

$(K^{\mathsf{MP}})^{-1} = -(\nabla\nabla\Phi(\theta^{\mathsf{MP}}))^{-1}$ for the covariance of the approximation. The integral of (1.8) is then

$$\mathsf{P}_{\mathcal{M}_i}(\mathcal{X}) \approx \left|K_i^{\mathsf{MP}}/2\pi\right|^{-1/2} \exp\Phi_i(\theta_i^{\mathsf{MP}}) = \left|K_i^{\mathsf{MP}}/2\pi\right|^{-1/2} \mathsf{P}_{\mathcal{M}_i}\left(\mathcal{X} \mid \theta_i^{\mathsf{MP}}\right) \mathsf{P}_{\mathcal{M}_i}\left(\theta_i^{\mathsf{MP}}\right) \tag{1.10}$$

where we have reintroduced the model subscript. The log Bayes factor of (1.7) is thus approximated by

$$\log B_{10} \approx \Lambda_{10}^{\mathsf{MP}} + \Pi_{10}^{\mathsf{MP}} + \frac{1}{2}\log\frac{\left|K_0^{\mathsf{MP}}/2\pi\right|}{\left|K_1^{\mathsf{MP}}/2\pi\right|} \tag{1.11}$$

where $\Lambda_{10}^{\mathsf{MP}}$ is similar to the log likelihood ratio statistic for classical model comparison, although evaluated at the MAP estimates, and Π_{10}^{MP} is the difference in the log priors of the MAP estimators for the two models. Note that this is different to the log of the prior odds of $\mathcal{M}_1$, which we have assumed to be 0. The priors in this case are not the priors of the models, but rather the priors of the *parameters* of each model, evaluated at the maximum of the posterior. In general, the more complex model may be expected to spread its prior more thinly over a larger parameter space, and thus to provide a smaller prior density at any particular point. Thus, we expect the term Π_{10}^{MP} to be negative, *penalizing* the likelihood ratio. Similarly, the determinant of the Hessian of the more complex model is likely to be larger (if the parameters are all estimated with roughly equivalent error e and we rotate to a diagonal basis we see that it will scale as $(1/e)^{p_i}$) and so the ratio of $|K|$ will be less than one, also penalizing the likelihood. The Laplace approximation is asymptotically correct, with, under certain regularity conditions, relative error of order $O(N^{-1})$ where N is the number of observations (Kass *et al.* 1990).

In the discussion of parameter estimation, we argued that we would remain agnostic on the nature of the prior and choose the maximum-likelihood estimator, which is likely to be close to the MAP value for vague priors. Can we reduce (1.11) from the same standpoint? Assuming the prior is vague, and that θ^{ML} is close to θ^{MP}, we can approximate $\Lambda_{10}^{\mathsf{MP}}$ by the more conventional likelihood ratio, Λ_{10}, evaluated at the respective maxima of the likelihoods. Also, the prior will not have strong curvature, and so the Hessian of the log unnormalized posterior, evaluated now at θ^{ML} will be dominated by the likelihood term. Thus we can replace K_i^{MP} by the **observed information matrix** $K_i = -\nabla\nabla\ell_{\mathcal{X}}\left(\theta_i^{\mathsf{ML}}\right)$. This gives us

$$\log B_{10} \approx \Lambda_{10} + \Pi_{10}^{\mathsf{ML}} + \frac{1}{2}\log\frac{|K_0/2\pi|}{|K_1/2\pi|} \tag{1.12}$$

where Π_{10}^{ML} is the log ratio of priors evaluated at the maximum likelihood parameter values. This approximation exhibits relative error $O(N^{-1/2})$.

At first glance, it would seem that we cannot dispense with the term Π_{10}^{ML} as it reflects the difference in dimensionality of the two models and provides an important penalty. However, consideration of the asymptotic behaviour of (1.12) reveals that for large data sets it may be neglected. If we have

N data points, the likelihood ratio takes the form $\sum_{n=1}^{N} \log\left(\mathsf{P}_{\mathcal{M}_1}\left(x_n \mid \theta_1^{\mathsf{ML}}\right)/\mathsf{P}_{\mathcal{M}_0}\left(x_n \mid \theta_0^{\mathsf{ML}}\right)\right)$ and will therefore grow with N. A similar argument applies to the Hessian of the log-likelihood, so that the magnitude of the final term of (1.12) grows as $\log N$. Thus the term Π_{10}^{ML}, which is constant with changes in the number of data can be asymptotically neglected.

We can further simplify the ratio of Hessians that appears in the final term of (1.12). With N data points, we have

$$\begin{aligned}
\log|K_i/2\pi| &= \log\left|-\tfrac{1}{2\pi}\sum_{n=1}^{N}\nabla\nabla\mathsf{P}_{\mathcal{M}_i}\left(x_n \mid \theta_i\right)\right| \\
&\approx \log\left|N\hat{K}/2\pi\right| \\
&= \log\left(\left(N/2\pi\right)^{p_i}\left|\hat{K}\right|\right) \\
&= p_i(\log N - \log 2\pi) + \log\left|\hat{K}\right| \qquad (1.13)
\end{aligned}$$

where $\hat{K}$ is the expected value with respect to the distribution of x of the one-point Hessian $\nabla\nabla\mathsf{P}_{\mathcal{M}_i}\left(x \mid \theta_i\right)$ evaulated at θ_i^{ML}. Again we drop the terms that do not grow with N, and obtain

$$\log B_{10} \approx \Lambda_{10} - \frac{1}{2}(p_1 - p_0)\log N \qquad (1.14)$$

This approximation was introduced by Schwartz (1978) with a far more rigorous derivation in the case of multivariate linear regression with an exponential family noise distribution, and was extended by Haughton (1988) to regression on curves. The heuristic approach we have adopted here suggests that it should be useful for many model families, and indeed it is used quite widely. It is referred to in the literature as the Schwartz criterion, or as the Bayesian Information Criterion, BIC.

In general the BIC approximation to the Bayes factor introduces relative errors of order $O(1)$. Some authors attempt to reduce the BIC error in the context of particular models by approximating the constant (with respect to N) term that we have neglected. One approach, practical in this modern day of the computer, is to determine a suitable value of the constant empirically by simulating and fitting data from known distributions. Other authors pay close attention to the definition of the number N. In the above, we simply took it to be the total count of data; on other hand, from the derivation it is clear that it is really the growth rate of the Hessian. In some models, the parameters are local and are only affected by data that fall within a small region. The clustering models of chapter 2, for example, fall into this category. In this case it may be argued that N is not the total number of data, but rather the average number of data falling into each cluster. In practice, however, all of these corrections are of order $O(1)$ and, provided that the number of data are large, the BIC alone has been found to produce reasonable results. We shall see, however, that in the context of latent variable models care must be taken in the choice of the number of parameters (Geiger *et al.* 1998). We will postpone our discussion of this issue, along with treatment of another approximate

Bayes technique for latent variable models introduced by Cheeseman and Stutz (1996). Instead, we shall proceed to investigate another class of model selection methods based on direct estimates of the probability of over-fitting.

Validation

We have motivated much of our development of model selection criteria by the notion of predictive accuracy. One approach, then, is to try to measure the predictive performance of the various models directly by observing the probability they assign to data outside the observations used for training. This approach is called **validation**. In its simplest form the process of validation involves the division of the set of observations $\mathcal{X}$ into two parts, the **training data** for which we will continue to use the symbol $\mathcal{X}$, and the validation or **test data** for which we will write $\mathcal{V}$. The posterior over parameters for each model (or the parameter estimates) are obtained on the training data, and the models are ranked by the probability that they assign to the validation set

$$V_i = \int d\theta_i \, \mathsf{P}_{\mathcal{M}_i}\left(\mathcal{V} \mid \theta_i\right) \mathsf{P}_{\mathcal{M}_i}\left(\theta_i \mid \mathcal{X}\right) \approx \mathsf{P}_{\mathcal{M}_i}\left(\mathcal{V} \mid \theta_i^{\mathsf{MP}}\right) \tag{1.15}$$

The intuition behind this approach is appealing, but it is often a fairly noisy criterion. We usually have only a limited amount of data available, and the necessity to divide it in two means that both the estimate of the parameters, and the estimate of the expected off-training set error are likely to be noisy. Once we have chosen a model by validation, we can combine the training and validation data sets and then reestimate the parameters to improve our predictions. However, the noise due to small $\mathcal{X}$ and $\mathcal{V}$ may lead to the incorrect model being selected.

In the simplest validation procedure there is a single training set and a single validation set. However, we could equally well train on $\mathcal{V}$ and test on $\mathcal{X}$. This would yield two independent estimates of the off-training-set performance of a particular model. The average of the two will thus have smaller variance than any one of them. In general, we can split up the data set into N_V disjoint subsets. One by one, we take each of these subsets, call it validation data, train on its complement in the data set, and validate the resulting model. Thus we obtain N_V independent estimates of V_i, which we can average to reduce the error in the estimate by $O(1/\sqrt{N_V})$. This simple improvement on the basic validation scheme is called **cross-validation**. In the extreme case where $N_V = N$, the number of data, the term **leave-one-out** cross-validation is applied.

Non-Bayesian Penalties

The spirit of such validation techniques, along with approximations similar to those made during the Bayesian treatment above, can also be used to obtain alternative likelihood penalization schemes similar to the BIC. The goal here is to estimate by how much the observed training likelihood is

likely to differ from the likelihood of the validation set.

Let us suppose that the true distribution of the data is some distribution $\mathsf{P}_*(\cdot)$, which we are attempting to fit with a family $\mathsf{P}_\theta(\cdot)$. Let θ^* represent the parameters that come closest to the true distribution in the sense of the Kullback-Leibler divergence, that is

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathsf{KL}[\mathsf{P}_*\|\mathsf{P}_\theta] = \underset{\theta}{\operatorname{argmin}} \int dx\, \mathsf{P}_*(x) \log \frac{\mathsf{P}_*(x)}{\mathsf{P}_\theta(x)} \tag{1.16}$$

If the true distribution is actually a member of the parametric family then the minimum KL divergence will, of course, be 0. Asymptotically, the maximum likelihood estimator will approach θ^*. When discussing parameter estimation we made the well known observation that the maximum likelihood estimator is asymptotically efficient, which holds when the true distribution falls within the parameterized family. This result can be extended to the general case.

The ML estimator given data $\mathcal{X}$ has the property that $\nabla \ell_{\mathcal{X}}\left(\theta^{\mathsf{ML}}\right) = 0$. Assuming that θ^{ML} is close to θ^*, we can make a linear approximation to the gradient at θ^*

$$\nabla \ell_{\mathcal{X}}(\theta^*) \approx \nabla \ell_{\mathcal{X}}\left(\theta^{\mathsf{ML}}\right) + (\theta^* - \theta^{\mathsf{ML}}) \nabla\nabla \ell_{\mathcal{X}}\left(\theta^{\mathsf{ML}}\right) = (\theta^* - \theta^{\mathsf{ML}}) K \tag{1.17}$$

where K is the observed information matrix, as before. Thus the error $\theta^* - \theta^{\mathsf{ML}} \approx K^{-1} \nabla \ell_{\mathcal{X}}(\theta^*)$ Asymptotically, the expected value of the difference is 0. To calculate the variance we note that for iid data $\mathcal{E}[K] = N\hat{K}$ where N is the number of observations and $\hat{K}$ is the expected one-point Hessian. We write $\hat{J} = \mathcal{V}ar\left[\nabla \ell_{x_i}(\theta^*)\right]$ as the more conventional definition of the Fisher information, the variance of the one-point log likelihood gradient, so that $\mathcal{V}ar\left[\nabla \ell_{\mathcal{X}}(\theta^*)\right] = N\hat{J}$, and so

$$\mathcal{V}ar\left[\theta^* - \theta^{\mathsf{ML}}\right] \approx \frac{1}{N} \hat{K}^{-1} \hat{J} \hat{K}^{-1} \tag{1.18}$$

The expectations and variances are all with respect to the true density $\mathsf{P}_*(\cdot)$. If this is the same as $\mathsf{P}_{\theta^*}(\cdot)$ then the two definitions of the information are equivalent and $\hat{J} = \hat{K}$, so that the mean square error approaches the standard Crámer–Rao bound $1/N\hat{J}$.

Given the asymptotic behaviour of the ML estimate, we can ask what likelihood we will assign to a validation point, v generated from the true distribution $\mathsf{P}_*(v)$. We expand around the "correct" validation value at θ^*.

$$\begin{aligned} \ell_v\left(\theta^{\mathsf{ML}}\right) &\approx \ell_v(\theta^*) + (\theta^{\mathsf{ML}} - \theta^*)^T \nabla_\theta \ell_v(\theta^*) + \frac{1}{2}(\theta^{\mathsf{ML}} - \theta^*)^T \nabla\nabla_\theta \ell_v(\theta^*)(\theta^{\mathsf{ML}} - \theta^*) & (1.19) \\ &= \ell_v(\theta^*) + (\theta^{\mathsf{ML}} - \theta^*)^T \nabla_\theta \ell_v(\theta^*) + \frac{1}{2}\mathrm{Tr}\left[\nabla\nabla_\theta \ell_v(\theta^*)(\theta^{\mathsf{ML}} - \theta^*)(\theta^{\mathsf{ML}} - \theta^*)^T\right] & (1.20) \end{aligned}$$

If we now take the expectation with respect to the true distribution of the training data *and* of v, we can take the expected gradient at θ^* to be 0. Also, since v is independent of $\mathcal{X}$ and therefore of

θ^{ML}, we can factor the expectation within the trace.

$$\begin{aligned}
\mathcal{E}\left[\ell_v\left(\theta^{\mathsf{ML}}\right)\right] &= \mathcal{E}\left[\ell_v\left(\theta^*\right)\right] + \frac{1}{2}\mathrm{Tr}\left[\mathcal{E}\left[\nabla\nabla_\theta \ell_v\left(\theta^*\right)\right]\mathcal{E}\left[(\theta^{\mathsf{ML}} - \theta^*)(\theta^{\mathsf{ML}} - \theta^*)^T\right]\right] \\
&= \mathcal{E}\left[\ell_v\left(\theta^*\right)\right] - \frac{1}{2}\mathrm{Tr}\left[\hat{K}\mathcal{V}ar\left[(\theta^{\mathsf{ML}} - \theta^*)\right]\right] \\
&= \mathcal{E}\left[\ell_v\left(\theta^*\right)\right] - \frac{1}{2}\mathrm{Tr}\left[\hat{K}\frac{1}{N}\hat{K}^{-1}\hat{J}\hat{K}^{-1}\right] \\
&= \mathcal{E}\left[\ell_v\left(\theta^*\right)\right] - \frac{1}{2N}\mathrm{Tr}\left[\hat{J}\hat{K}^{-1}\right] \qquad (1.21)
\end{aligned}$$

This expression shows the approximate bias in the validation likelihood. On the training data we can expand $\ell_{\mathcal{X}}\left(\theta^*\right)$ around θ^{ML} (where the gradient is always 0) to obtain

$$\mathcal{E}\left[\ell_{\mathcal{X}}\left(\theta^*\right)\right] = \mathcal{E}\left[\ell_{\mathcal{X}}\left(\theta^{\mathsf{ML}}\right)\right] - \frac{1}{2}\mathrm{Tr}\left[\hat{J}\hat{K}^{-1}\right] \qquad (1.22)$$

Now, the expected values of the log-likelihoods at the fixed point θ^* are equal (up to a factor of the number of training data, N). Thus, we obtain

$$\mathcal{E}\left[\ell_v\left(\theta^{\mathsf{ML}}\right)\right] = \frac{1}{N}\left(\mathcal{E}\left[\ell_{\mathcal{X}}\left(\theta^{\mathsf{ML}}\right)\right] - \mathrm{Tr}\left[\hat{J}\hat{K}^{-1}\right]\right) \qquad (1.23)$$

This can be viewed as a prediction of the expected difference between the validation likelihood and the training likelihood. We might therefore rank models according to their training likelihoods penalized by the trace term.

This is the NIC (Network Information Criterion) of Murata *et al.* (1991, 1993, 1994). To use it we replace the expected values of the information measures $\hat{J}$ and $\hat{K}$ by their observed values,

$$\mathrm{NIC} = \ell_{\mathcal{X}}\left(\theta^{\mathsf{ML}}\right) - \mathrm{Tr}\left[JK^{-1}\right] \qquad (1.24)$$

with K the observed information and $J = \sum_i (\nabla \ell_{x_i}\left(\theta^{\mathsf{ML}}\right))^2/(N-p)$ where p is the number of parameters. If the true distribution lies within the parameterized family then $\hat{J} = \hat{K}$ and we can replace the trace penalty by the number of parameters p. This is the AIC of Akaike (1974). Akaike used AIC as an abbreviation for *An Information Criterion*, although it is usually taken to stand for the *Akaike Information Criterion*.

1.4 Graphical Representations

In most experiments we measure more than one variable simultaneously. Thus the observations x_i that we have described above are usually multivariate. It is often useful to partition the observations into a number of distinct random variables, each of which may still be multivariate. For example,

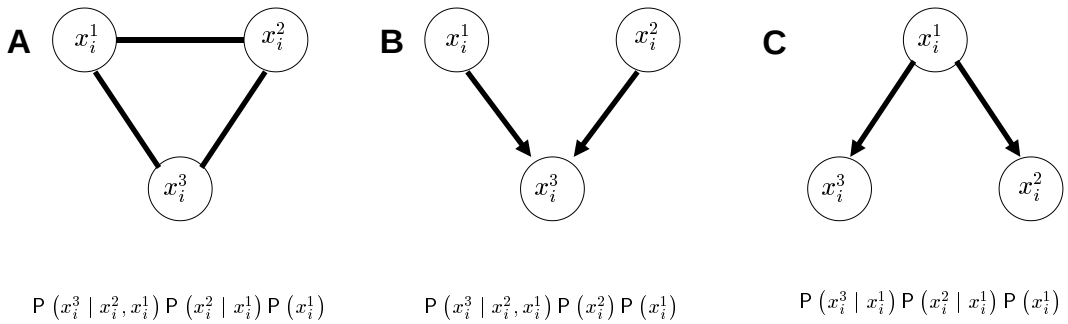


Figure 1.2: Graphical representation of conditional independence.

we may make measurements with different instruments and regard the output of each instrument, whether a single number or a vector, as a random variable. The advantage to such a partition is that it is often possible to write the parameterized model distribution $\mathsf{P}_\theta\left(x_i\right)$ more easily in terms of the partitioned variables. Why would this be so?

Let us consider a case where the observation x_i is partitioned into three random variables x_i^1, x_i^2, x_i^3. In general any probability function of the x_i may be written in conditional form:

$$\mathsf{P}\left(x_i\right) = \mathsf{P}\left(x_i^3 \mid x_i^2, x_i^1\right) \mathsf{P}\left(x_i^2 \mid x_i^1\right) \mathsf{P}\left(x_i^1\right) \tag{1.25}$$

However, it might be that x_i^2 is independent of x_i^1 and so we replace the second term on the right above with just $\mathsf{P}\left(x_i^2\right)$. Another possibility is that x_i^3 is conditionally independent of x_i^2 given x_i^1 so that we can write $\mathsf{P}\left(x_i^3 \mid x_i^1\right)$ in place of the first right hand term. This might seem like only a notational convenience, but, in fact, if we are to parameterize the probability distribution we have saved ourself some parameters. The factorized function is *simpler* (in the sense of model selection) than before.

The factorized structure of the distribution can be shown graphically as in figure 1.2. In panel A the case of no conditional or marginal independencies is shown as a fully connected undirected graph. Panel B represents the marginal independence of x_i^1 and x_i^2. Panel C represents the conditional independence of x_i^2 and x_i^3. Each of the latter two cases is represented by a **directed acyclic graph** or DAG.

It should be noted that the connection between probabilistic models and DAGs is far from cosmetic. An important and deep theory is available connecting reasoning about the probability distribution with algorithmic manipulations of the graph (Pearl 1988; Lauritzen 1996). However, we shall not exploit this theory at all; for us the graph will simply be a convenient tool for visualization.

When representing parameterized probability functions $\mathsf{P}_\theta\left(x_i\right)$ we will find it useful to introduce nodes in our graphical representation corresponding to the parameters. Since we have factorized our probability functions, we need to partition the parameters θ into the groups appropriate for each

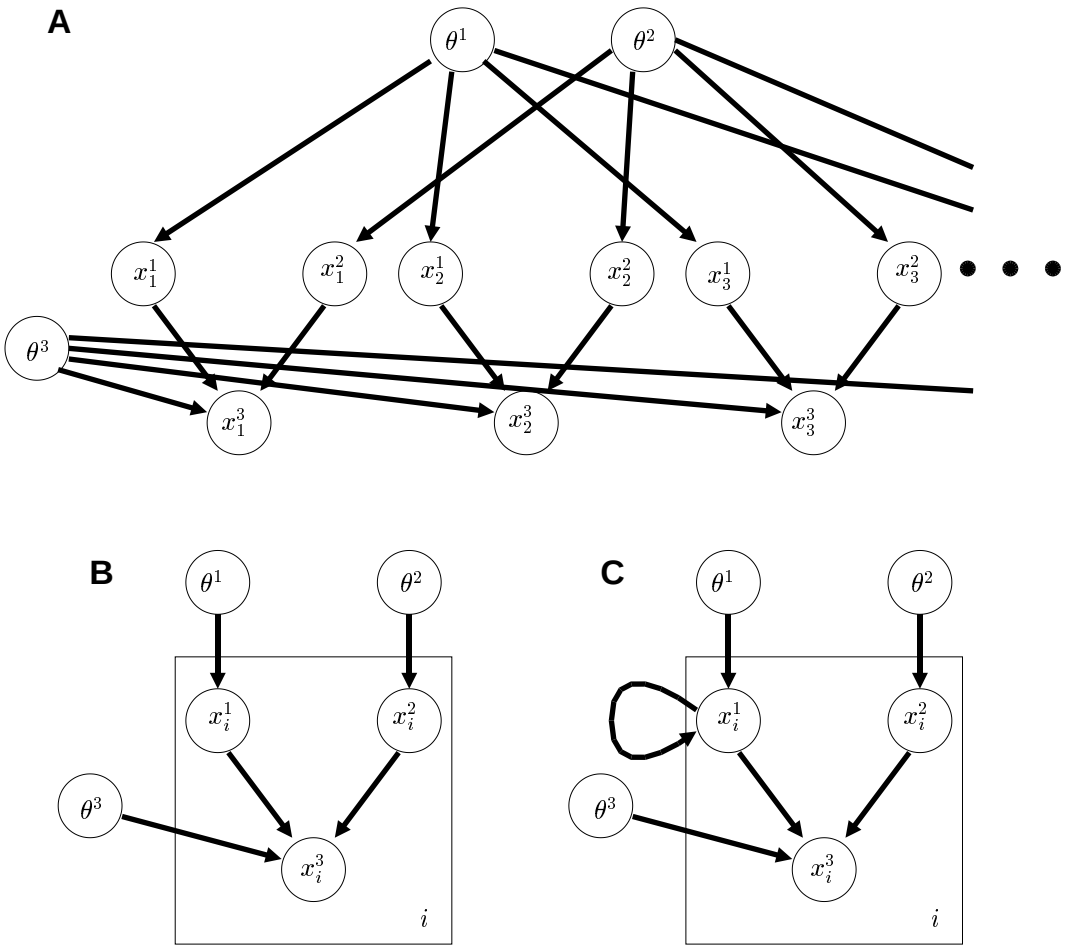


Figure 1.3: Graphical representations of repeated observation models.

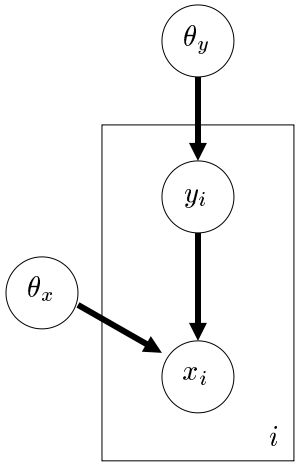
factor function. In general, we might write

$$\mathsf{P}_\theta\left(x_i\right) = \mathsf{P}_{\theta^3}\left(x_i^3 \mid x_i^2, x_i^1\right) \mathsf{P}_{\theta^2}\left(x_i^2 \mid x_i^1\right) \mathsf{P}_{\theta^1}\left(x_i^1\right) \tag{1.26}$$

where θ is the union of $\theta^r, r = 1 \ldots 3$. Figure 1.3A illustrates the representation. Whereas before it was sufficient to show the variables involved in a single observation i, with the implicit information that each observation is independent and identically distributed, we now need to make clear that the parameters are chosen exactly once and have the same value over all observations, whereas each observation has its own set of random variables x_i^r. This time the fact that the x_i^r are independent (conditioned on the parameters) is shown explicitly by the lack of edges between nodes at different values of i.

We can condense the representation as shown in Figure 1.3B[2]. The rectangle represents a single

[2] To the best of my knowledge, this representation was introduced in the computer program BUGS from the MRC biostatistics unit at Cambridge (Thomas 1994; Spiegelhalter *et al.* 1996).



$$\mathsf{P}_{\theta_x}(x_i \mid y_i)\,\mathsf{P}_{\theta_y}(y_i)$$

Figure 1.4: Graphical representation of a latent variable model.

observation with an index indicated its lower right hand corner; variables that appear within the rectangle are repeated across observations, while the parameters which are chosen only once for all observations appear outside it. As before, the lack of edges between nodes at differing i indicates that the observations are independent. Now, our decision to represent all the functions $\mathsf{P}_\theta(x_i)$ by a single subgraph indicates further that they are identical.

If the observations are not independent, say there are correlations between the variables x_i^1 at different i, we may represent this fact by an edge that crosses out of, and then back into, the rectangle, as in figure 1.3C. However, we cannot show the limits of this interaction. For example, if x_i^1 is generated by a Markov process, so that x_i^1 is conditionally independent of $x_1^1 \ldots x_{i-2}^1$ given x_{i-1}^1 we need the expanded time view of figure 1.3A, with additional edges for the Markovian dependence, to distinguish this from the other possible cross-observation dependency structures.

1.5 Latent Variables

We have seen that it can be useful to partition the observed variables so as to simplify the expression of the probability function by exploiting the conditional dependency structure of the problem. Another manipulation that can assist in this simplification is the introduction of **latent variables**. These are variables which are not observed. The parameters, of course, are also not observed; the latent variables are different in that they are presumed to be instantiated once for every observation, that is there is a latent y_i for each observation x_i. In graphical terms, the simplest latent variable model is sketched in figure 1.4. Note that the latent variable node appears within the rectangle.

In a latent variable model we can add a third operation to our pair of learning and model

selection, **inference**. This will refer to the estimation of value of the latent variables y_i given known parameters and the observations x_i. The difference from fitting, that is, estimating the parameters, is simply one of scale.

Again, it has been shown that certain algorithmic manipulations on the graph that defines the latent variable model can yield the correct form of inference (Pearl 1988). For most of the models we shall discuss, however, inference will be a simple matter of the application of Bayes' rule:

$$\mathsf{P}_\theta\left(y_i \mid x_i\right) = \frac{\mathsf{P}_\theta\left(x_i \mid y_i\right)\mathsf{P}_\theta\left(y_i\right)}{\mathsf{P}_\theta\left(x_i\right)} \tag{1.27}$$

1.6 The Expectation–Maximization Algorithm

How do we go about learning the parameter values of a latent variable model? It is possible to define a likelihood function for the parameters by integrating over the latent variables[3].

$$\ell_{\mathcal{X}}\left(\theta\right) = \log \int d\mathcal{Y}\ \mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right)\mathsf{P}_\theta\left(\mathcal{Y}\right) \tag{1.28}$$

where the integral is over all the y_i in the set $\mathcal{Y}$. However, in many cases this likelihood is quite difficult to optimize in closed form. Gradient- or Hessian-based numerical optimization schemes can be very effective for a number of problems. In the case of latent variable models, however, another algorithm exists that is frequently more straightforward and of comparable efficiency. This is the **Expectation–Maximization** (or **EM**) algorithm (Dempster *et al.* 1977). Quite complicated models may be fit efficiently by use of EM (Xu and Jordan 1996).

We shall first lay out the steps of the EM algorithm and only then offer two (informal) proofs of its validity. The second of these proofs will also provide the justification for various extensions.

If we had, in fact, observed the variables y_i we would be able to write the **joint data log likelihood**

$$\ell_{\mathcal{X},\mathcal{Y}}\left(\theta,\theta\right) = \log \mathsf{P}_\theta\left(\mathcal{X} \mid \mathcal{Y}\right) + \log \mathsf{P}_\theta\left(\mathcal{Y}\right) \tag{1.29}$$

This likelihood is often much easier to manipulate than the true likelihood of (1.28), since it avoids the awkward log-of-integral (or log-of-sum) expression. It will be the starting point for EM.

To begin the EM algorithm we provide seed guesses for the parameters. We will label successive outputs of the iterations by the iteration number in the superscript. Thus, the initial guesses will be called θ^0. At the nth iteration we estimate new values of the parameters by the following two steps.

E-step: Find the **expectation** of the joint data log-likelihood under the distribution of the y_i given

[3] In this general introduction we shall assume that the y_i are continuous, but discrete latent variables may be handled in the same fashion with the integral replaced by a sum.

the $n-1$th parameter estimates and the observations.

$$Q^n(\theta) = \mathcal{E}_{\mathcal{Y}|\mathcal{X};\theta^{n-1}}\left[\ell_{\mathcal{X},\mathcal{Y}}(\theta)\right] \tag{1.30}$$

M-step: Then **maximize** this expected joint data log-likelihood with respect to the parameters to obtain the new estimates.

$$\theta^n = \operatorname{argmax} Q^n(\theta) \tag{1.31}$$

Why does EM work? Let us consider the effect of the iterations on the true log-likelihood function given in (1.28). In each iteration we start with parameters θ^{n-1} and estimate new parameters θ^n. For notational simplicity we will write $\mathsf{P}_{n-1}(\cdot)$ for the various probability functions with parameters θ^{n-1} and similarly for $\mathsf{P}_n(\cdot)$. The resulting log-likelihood is

$$\ell_{\mathcal{X}}(\theta^n) = \log \int d\mathbf{y}\ \mathsf{P}_n(\mathcal{Y})\,\mathsf{P}_n(\mathcal{X} \mid \mathcal{Y}) \tag{1.32}$$

We introduce a factor of $\dfrac{\mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})}{\mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})}$ within the integral and rearrange to obtain

$$\ell_{\mathcal{X}}(\theta^n) = \log \int d\mathbf{y}\ \mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X}) \left(\frac{\mathsf{P}_n(\mathcal{Y})\,\mathsf{P}_n(\mathcal{X} \mid \mathcal{Y})}{\mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})}\right) \tag{1.33}$$

We can now use Jensen's inequality (see, for example, Cover and Thomas (1991)) applied to the convex function $\log(\cdot)$ to exchange the logarithm and integral. In this context, Jensen's inequality states that, for positive weights α_i that sum to 1,

$$\log(\sum_i \alpha_i x_i) \geq \sum_i \alpha_i \log(x_i) \tag{1.34}$$

We can generalize this for a positive continuous weight function with unit integral, in our case $\mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})$, to obtain

$$\begin{aligned} \ell_{\mathcal{X}}(\theta^n) &\geq \int d\mathbf{y}\ \mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})\ \log\left(\frac{\mathsf{P}_n(\mathcal{Y})\,\mathsf{P}_n(\mathcal{X} \mid \mathcal{Y})}{\mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})}\right) && (1.35) \\ &= \int d\mathbf{y}\ \mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})\ \log\left(\mathsf{P}_n(\mathcal{Y})\,\mathsf{P}_n(\mathcal{X} \mid \mathcal{Y})\right) - \\ &\quad \int d\mathbf{y}\ \mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})\ \log\left(\mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})\right) && (1.36) \end{aligned}$$

Thus the quantity on the right hand side of (1.36) is a lower bound on the likelihood at the nth iteration. The first term is readily identified as the quantity $Q^n(\theta)$ from our statement of the EM algorithm (1.30). The second term has no dependence on θ^n. Thus by maximizing $Q^n(\theta)$ as dictated by the m-step (1.31) we are maximizing a lower bound on the likelihood. Further, we know that the

maximum must be $\geq \ell_{\mathcal{X}}\left(\theta^{n-1}\right)$ since we can obtain that value by simply putting $\theta^n = \theta^{n-1}$. Thus we can be sure that as long as the EM algorithm does not converge, the likelihood of the model must increase.

We need also to show that when the EM algorithm does converge, we have reached a maximum of the true likelihood. This proof appears in (Dempster *et al.* 1977), and we will not reproduce it. Instead, we will follow Neal and Hinton (1998) and take a slightly different view of the algorithm; this approach will yield the necessary second component of the proof.

1.7 Free Energy and EM

Let us define a more general form of the function Q in (1.30) by taking the expectation with respect to an arbitrary probability function $p(\mathcal{Y})$, in place of the particular probability $\mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})$.

$$Q(p, \theta) = \mathcal{E}_p\left[\ell_{\mathcal{X},\mathcal{Y}}(\theta)\right] \tag{1.37}$$

We can then introduce a function that we will call the **free energy** by analogy with statistical mechanics,

$$F(p, \theta) = Q(p, \theta) + H(p) \tag{1.38}$$

where $H(p) = -\mathcal{E}_p[\log p]$ is the entropy of p. This function is familiar from above; it is the right hand side of (1.36) with the arbitrary function p replacing $\mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})$. Furthermore, in arriving at that expression our choice of weighting function to use in Jensen's inequality was arbitrary, so F also bounds the likelihood $\ell_{\mathcal{X}}(\theta)$ below. In drawing the physical analogy we should note that our F should, in fact, be regarded as the negative of the conventional free energy, which is consistent with the fact that we are interested in maximizing F, while physical systems evolve to minimize their free energy.

We observe (Neal and Hinton 1998) that, if θ is held constant, the free energy is, in fact, maximized by choosing $p(\mathcal{Y}) = \mathsf{P}_\theta(\mathcal{Y} \mid \mathcal{X})$. To see this, we maximize the quantity

$$L_\theta(p) = F(p, \theta) - \lambda \int d\mathcal{Y}\; p(\mathcal{Y}) \tag{1.39}$$

$$= \int d\mathcal{Y}\; p(\mathcal{Y})\left(\ell_{\mathcal{X},\mathcal{Y}}(\theta) - \log p(\mathcal{Y}) - \lambda\right) \tag{1.40}$$

where λ is a Lagrange multiplier enforcing the normalization constraint. From the theory of the calculus of variations (Mathews and Walker 1970) we find that at the maximum with respect to p the functional derivative of the integrand must be 0 (this is a trivial special case of the Euler-Lagrange

equations). Thus the maximum occurs when

$$\begin{aligned} 0 &= \frac{\partial}{\partial p}\left(p(\mathcal{Y})\left(\ell_{\mathcal{X},\mathcal{Y}}(\theta) - \log p(\mathcal{Y}) - \lambda\right)\right) \\ &= \left(\ell_{\mathcal{X},\mathcal{Y}}(\theta) - \log p(\mathcal{Y}) - \lambda\right) - \frac{p(\mathcal{Y})}{p(\mathcal{Y})} \end{aligned} \tag{1.41}$$

and so

$$p(\mathcal{Y}) = e^{-\lambda-1}\mathcal{L}_{\mathcal{X},\mathcal{Y}}(\theta) = e^{-\lambda-1}\mathsf{P}_\theta(\mathcal{X},\mathcal{Y}) \tag{1.42}$$

The requirement that p be normalized determines the multiplier λ and yields $p(\mathcal{Y}) = \mathsf{P}_\theta(\mathcal{Y} \mid \mathcal{X})$.

Thus we obtain a new interpretation of the EM algorithm.

E-step: Maximize F with respect to p holding θ constant.

M-step: Maximize F with respect to θ holding p constant.

We can now sketch the proof that if F achieves a local maximum at p^*, θ^* then $\ell_{\mathcal{X}}(\theta)$ achieves a local maximum at θ^* (Theorem 2 of Neal and Hinton (1998)). We first note that if $p(\mathcal{Y}) = \mathsf{P}_\theta(\mathcal{Y} \mid \mathcal{X})$ then

$$\begin{aligned} F(\mathsf{P}_\theta(\mathcal{Y} \mid \mathcal{X}), \theta) &= Q(\mathsf{P}_\theta(\mathcal{Y} \mid \mathcal{X}), \theta) + H(\mathsf{P}_\theta(\mathcal{Y} \mid \mathcal{X})) \\ &= \mathcal{E}_{\mathcal{Y}|\mathcal{X};\theta}\left[\ell_{\mathcal{X},\mathcal{Y}}(\theta)\right] - \mathcal{E}_{\mathcal{Y}|\mathcal{X};\theta}\left[\log \mathsf{P}_\theta(\mathcal{Y} \mid \mathcal{X})\right] \\ &= \mathcal{E}_{\mathcal{Y}|\mathcal{X};\theta}\left[\frac{\log \mathsf{P}_\theta(\mathcal{Y},\mathcal{X})}{\log \mathsf{P}_\theta(\mathcal{Y} \mid \mathcal{X})}\right] \\ &= \mathcal{E}_{\mathcal{Y}|\mathcal{X};\theta}\left[\log \mathsf{P}_\theta(\mathcal{X})\right] \\ &= \log \mathsf{P}_\theta(\mathcal{X}) \\ &= \ell_{\mathcal{X}}(\theta) \end{aligned} \tag{1.43}$$

Thus, writing $p^*(\mathcal{Y})$ for $\mathsf{P}_{\theta^*}(\mathcal{Y} \mid \mathcal{X})$, we have $\ell_{\mathcal{X}}(\theta^*) = F(p^*, \theta^*)$. Suppose there is some θ^{**} ϵ-close to θ^* at which the log-likelihood is larger, and that p^{**} is the corresponding $\mathsf{P}_{\theta^{**}}(\mathcal{Y} \mid \mathcal{X})$. Then it must be that $F(p^{**}, \theta^{**}) > F(p^*, \theta^*)$. But, assuming that $\mathsf{P}_{\theta^*}(\mathcal{Y} \mid \mathcal{X})$ varies continuously with θ^*, if θ^{**} is ϵ-close to θ^* then p^{**} is δ-close to p^*. This violates the assumption that F achieves a local maximum at p^*, θ^*, and so there can be no such θ^{**} close to θ^* with larger likelihood. Thus $\ell_{\mathcal{X}}(\theta^*)$ is a local maximum. A similar argument can be made for the global maximum (and we don't even need the continuity assumption).

1.8 Generalizations of EM

This formulation does not just provide straightforward access to the above proof; it also justifies a number of generalizations of the EM algorithm. The first actually follows from the argument

following (1.36) and appeared in (Dempster *et al.* 1977). This is the generalized M-step. As long as, at each iteration, the function Q is increased relative to its value at θ^{n-1}, all of the guarantees of increasing the likelihood are maintained. We do not need to maximize Q at each iteration, we can instead just take a step in the direction of its gradient (provided we are guaranteed that Q will indeed be maximized at convergence – see the comments below). This variant is called **gradient** or **generalized** EM (usually written GEM):

E-step: Find the **expectation** of the joint data log-likelihood under the distribution of the y_i given the $n-1$th parameter estimates and the observations. (This is unchanged.)

$$Q^n(\theta) = \mathcal{E}_{\mathcal{Y}|\mathcal{X};\theta^{n-1}}\left[\ell_{\mathcal{X},\mathcal{Y}}(\theta)\right] \tag{1.44}$$

GM-step: Change θ in the direction of the gradient of Q.

$$\theta^n = \theta^{n-1} + \eta \nabla_\theta Q^n(\theta^{n-1}) \tag{1.45}$$

where η is some learning rate parameter chosen in accordance with the usual principles of gradient ascent. Clearly, this is useful when Q cannot be maximized in closed form. In such situations it is usually computationally more efficient to use GEM rather than numerically optimizing Q in each M-step.

The free energy formulation suggests an alternative generalization. In principle, we could make a corresponding generalized E-step, and choose a function p different from $\mathsf{P}_{n-1}(\mathcal{Y} \mid \mathcal{X})$, provided it increases the free energy. We must be careful, however. We have shown that when the free energy reaches a local maximum, so does the likelihood. If we generate functions p by an algorithm that can converge even though F is not at a true local maximum, our guarantees of maximal likelihood evaporate. Such a situation arises when the functions p are restricted in functional form so that for most values of θ the function $\mathsf{P}_\theta(\mathcal{Y} \mid \mathcal{X})$ does not lie within the family of possibilities. In this case we can at best optimize F on the surface of constraint defined by the function family. An example is found in the Helmholtz machine (Dayan *et al.* 1995). The wake-sleep learning algorithm (Hinton *et al.* 1995) for the Helmholtz machine involves exactly such a constrained generalized E-step where the estimate p must be the output of a sigmoid belief network. As a result, it cannot guarantee convergence to the maximum likelihood parameters.

A similar caution, of course, can apply to generalized M-steps too. The usual choice of a gradient M-step, however, *is* guaranteed to converge to a local stationary point of F.

One example of an approximate E-step that maintains the convergence properties is provided by Neal and Hinton (1998). This is the incremental E-step, applicable when the x_i and y_i are independent. In this case, we can restrict the functions p to the family of functions with the form

$p(\mathcal{Y}) = \prod_i p_i(y_i)$ since the independence of the y_i guarantees that the optimal p will be in the family. We can now write

$$\begin{aligned} F(p,\theta) &= \sum_i F_i(p_i,\theta) \\ &= \sum_i \mathcal{E}_{p_i}\left[\ell_\theta\left(x_i,y_i\right)\right] + H(p_i) \end{aligned} \tag{1.46}$$

and maximize each component F in turn. The incremental EM algorithm now proceeds from initial guesses θ^0 and p_i^0 so:

IE-step: Choose some i. Maximize $F_i(p_i,\theta^{n-1})$ and leave the remaining $p_j, j \neq i$ unchanged.

$$\begin{aligned} p_i^n(y_i) &= \mathsf{P}_{n-1}\left(y_i \mid x_i\right) \\ p_j^n(y_j) &= p_j^{n-1}(y_j) \end{aligned} \tag{1.47}$$

M-step: Maximize F with respect to θ holding p constant.

In practice, for many distributions of interest, the M-step can be performed from sufficient statistics of the data, which are efficiently updated with respect to p_i (Neal and Hinton 1998). We shall, in fact, use a similar approach to track non-stationary mixture distributions efficiently.