MODELLING WITH MIXTURES

Mike West
ISDS, Duke University

DP# 91-A05
Modelling with mixtures

By
Mike West
Institute of Statistics and Decision Sciences
Duke University, Durham, NC 27706, USA.

SUMMARY
Discrete mixtures of distributions of standard parametric forms commonly arise in statistical modelling and with methods of analysis that exploit mixture structure. This paper discusses general issues of modelling with mixtures that arise in fitting mixtures to data distributions, using mixtures to approximate functional forms, such as posterior distributions in parametric models, and development of mixture pruning methods useful for reducing the number of components of large mixtures. These issues arise in problems of density estimation using mixtures of Dirichlet processes, adaptive importance sampling function design in Monte Carlo integration, and Bayesian discrimination and cluster analysis.

Invited paper for the Fourth Valencia International Meeting on Bayesian Statistics, Peñíscola, Spain, April 15th–20th 1991. Research reported here was partially financed by the National Science Foundation under grant DMS-8903842.
1. INTRODUCTION

Several areas of application of discrete mixture distributions are reviewed here, involving recent methodological and computational developments in fitting mixtures to data and in approximating posterior distributions via mixtures of standard parametric forms.

The first area concerns the approximation of prior and posterior distributions by discrete mixtures of normal or, more generally, multivariate $T$ distributions. This work stems from use of such mixtures as importance sampling functions in Monte Carlo integration, extending previous approaches to adaptive importance sampling. Mixtures have the flexibility to represent possibly very irregular forms of posterior densities, though there has to date been little application of mixtures, either as importance sampling distributions or as more direct approximations to posteriors, the main reason being the lack of analytic techniques for constructing suitable mixtures. A constructive approach, based on adaptive resampling and refinement of approximating mixtures (West, 1990a), is reviewed and discussed, with illustration, in Section 2.

In data analysis using discrete mixtures, several themes arise according to application area. Direct fitting of mixtures of standard exponential family distributions, such as mixtures of exponentials or mixtures of normals, is technically possible now using iterative resampling schemes, such as Gibbs sampling. Some work in this area, with a focus on mixture deconvolution, discrimination and classification (Lavine and West, 1991), is discussed in Section 3.

Section 4 concerns discrete mixtures arising in nonparametric Bayesian density estimation using mixtures of Dirichlet processes. Some review of the basic structure of such models is given, and issues of posterior and predictive calculations in such models described. Particular methods of approximation of predictive distributions in these models lead to Bayesian analogues of traditional kernel density techniques, and simpler forms based on mixtures of, typically, small numbers of components. Various practical issues in traditional kernel estimation, including smoothing parameter selection and modelling varying degrees of smoothness across the sample space, for instance, reduce to standard problems of modelling and inference in a model based framework. Other issues discussed include mixture deconvolution, inference about numbers of mixture components, and the modality characteristics of data distributions. More recent work on the computational side, involving the development of Gibbs sampling techniques for posterior analysis in mixtures of Dirichlet processes (Escobar, 1990; Escobar and West, 1991), is also discussed and illustrated.

2. MIXTURE MODELS FOR PRIORS AND POSTERIORS

2.1 Sampling posteriors and kernel reconstruction

In simulation based analysis of posterior distributions, marginal posterior densities are typically approximated by some form of discrete mixture. With the availability of complete conditional distributions for the application of Gibbs sampling techniques, mixtures of various conditionals
Modelling with mixtures

Valencia, April 1991

arise naturally. Elsewhere, kernel estimation techniques have been used by several authors to construct smooth approximations to posteriors when other, more efficient forms of approximation are unavailable (Gelfand and Smith, 1990; Wolpert, 1991). Suppose a simulation analysis provides an approximate sample from a posterior \( p(\theta) \), with \( \theta \) in \( p \) dimensions. Denote the sampled points \( \Theta = \{ \theta_j, j = 1, \ldots, n \} \). A kernel estimate of \( p(\theta) \) has the form \( n^{-1} \sum_{j=1}^n d(\theta_j, V h^2) \) where \( d(\theta|m, M) \) is a \( p \)-variate, elliptically symmetric density function with mode \( m \) and scale matrix \( M \), \( V \) is an estimate of the variance matrix of \( p(\theta) \), and \( h \) is a smoothing parameter. Important kernels are the multivariate T family, when the estimates are mixtures of equally weighted T distributions. More generally, in Monte Carlo with importance sampling, a posterior approximation \( g_0(\theta) \) provides the random sample \( \Theta \), and importance weights are defined by \( w_j \propto p(\theta_j)/g_0(\theta_j) \), subject to unit sum. Then posterior expectations are estimated by weighted sums such as \( E[a(\theta)] \approx \sum_{j=1}^n w_j a(\theta_j) \) for given functions \( a(\theta) \). Now the kernel approach extends to give a posterior density estimate of the form

\[
g(\theta) = \sum_{j=1}^n w_j d(\theta_j, V h^2),
\]

appropriately inserting the Monte Carlo weights (Wolpert, 1991; West, 1990a).

Posterior reconstruction and graphical display using (1) require specification of \( h \) and \( V \). Choice of smoothing parameter \( h \) may be guided by traditional kernel estimation theory, and also by the derivation of kernel densities as approximations to predictive distributions in Bayesian density estimation using mixtures of Dirichlet processes (Ferguson, 1983; West, 1990c) where the smoothing parameter \( h \) has interpretation as a function of model parameters and sample size. Routine recommendations (eg: Silverman, 1976) are broadly applicable though tend to over-smooth more erratic, multimodal densities (West, 1990a,b). Choosing \( h \) as a slowly decreasing function of simulation sample size \( n \), as traditionally recommended (Silverman, 1986), ensures large sample consistency of \( g(\theta) \) as an estimate of the true posterior \( p(\theta) \). For the scale matrix \( V \), a natural choice is the Monte Carlo estimate of posterior variance \( V = \sum_{j=1}^n w_j (\theta_j - \bar{\theta}) (\theta_j - \bar{\theta})' \) where \( \bar{\theta} = \sum_{j=1}^n w_j \theta_j \) is the Monte Carlo mean. However, as a direct approximation to \( p(\theta) \) the mixture \( g(\theta) \) will then be overdispersed — if the kernels are T on \( a > 0 \) degrees of freedom then the variance under \( g(\theta) \) is \( V(1 + h^2 a/(a - 2)) \), not \( V \). Guided again by theory of Bayesian density estimation (West, 1990c), shrinkage of the kernel locations may be applied to reduce this overdispersion: with shrinkage factor \( x \) in \( (0, 1) \), replacing the sampled points \( \theta_j \) by \( m_j = x \theta_j + (1 - x) \bar{\theta} \) as the modes of the kernels, the resulting mixture has variance \( V(x^2 + h^2 a/(a - 2)) \) which reduces to \( V \) at \( x^2 = 1 - h^2 a/(a - 2) \). In practice, \( h \) will be specified as a slowly decreasing function of sample size \( n \) (West, 1990a) so that the shrinkage will be negligible with large sample sizes.

Incidentally, note that simulation and kernel reconstruction provides a technique by which any specified prior (that may easily be sampled) may be approximated by a mixture of distributions of standard forms – the kernels may be any standard form, not necessarily symmetric. At the second Valencia meeting, Diaconis and Ylvisaker (1985) discuss theoretical issues in approximating
Modelling with mixtures

Valencia, April 1991

priors by mixtures of conjugate forms; see also Dalal and Hall (1983). More practically, Alspach and Sorenson (1972) develop approximating mixtures of normals in signal processing problems, although there are really no general, direct and automatic techniques for constructing such approximations. Sampling the prior, or an importance sampling function approximating the prior, and using a kernel techniques with normal kernels is one possibility, which also allows neatly for approximate prior to posterior updating, as follows. Suppose the kernels in (1) are multivariate normal with modes $m_j$, and that $n$ is large, say several thousands. Then each of the mixands will be closely concentrated about its mode $m_j$, since $h$ will be small for large $n$. Multiplying by a likelihood function $l(\theta)$ we have posterior density $g(\theta)l(\theta) \propto \sum_{j=1}^n w_j d(\theta|m_j, Vh^2)l(\theta)$. With $n$ large enough, hence $h$ small enough, the likelihood function will be diffuse compared to each of the the mixands, though not necessarily relative to the overall mixture. Hence a local log-quadratic approximation to the products $d(\theta|m_j, Vh^2)l(\theta)$ leads to approximate normal posteriors $d(\theta|m_j^*, V_j^*)$ multiplied by (data dependent) constants $v_j$, say; thus the posterior is approximately proportional to $\sum_{j=1}^n w_j^* d(\theta|m_j^*, V^*)$, with updated weights $w_j^* \propto w_j v_j$, and summing to unity. In some problems, and with informed priors, this can provide excellent approximations to posteriors of direct use; one area in which this has been employed is in providing initial guesses at posteriors in adaptive development and refinement of mixture approximations using importance sampling, as described below. This is of particular interest in application in sequential analyses, as in Section 2.4 below, where the priors to be approximated are quite typically fairly concentrated relative to likelihood functions, and so the mixands of kernel reconstructions are really very precise, and simple analytic updating as just described can be extremely effective.

2.2 Adaptive approximation of posteriors

In West (1990a) mixtures of T distributions as just described are used as importance sampling distributions in a scheme to adaptively refine posterior approximations. The basic idea is to develop an initial guess at the posterior, say $g_0(\theta)$, to be used as a starting importance sampling function. Based on a sample of size $n_0$ from this distribution, a weighted kernel estimate is constructed as in (1), and used as a second guess; call this $g_1(\theta)$. Repeating this procedure, with a possibly different sample size $n_1$ from $g_1(\theta)$, we can revise the mixture importance sampling density to $g_2(\theta)$, and further if required. At each stage, the smoothing parameter may be chosen to increase or decrease the spread of the next importance function if that currently is thought to poorly represent the posterior, such as when the current distribution of the importance sampling weights varies dramatically away from uniformity; a more diffuse distribution will spread out the sampled points at the next stage, with a view to increasing support in regions of the parameter space previously undersupported. In cases when $g_0(\theta)$ is a rather crude initial guess, adaptive refinement is most appropriate, since then direct use of $g_0(\theta)$ will be highly inefficient, computationally and statistically. At the final stage, a relatively small sample may be sufficient for the desired accuracy, and often just one refinement may be sufficient to adjust a very crude approximation, say a single
Modelling with mixtures

Valencia, April 1991

multivariate T density, to a mixture $g_1(\theta)$ of, say, several hundred T densities, much more closely representing the true $p(\theta)$. In approximating moments and probabilities, a Monte Carlo sample of a few thousand draws from $g_1(\theta)$ may do as well as, or better than, a sample of several times that from the original $g_0(\theta)$. At a final stage, the sampled points and weights provide for estimation of posterior expectations, as usual, and kernel reconstructions provide marginal posterior density and distribution function estimates.

An example, with $p = 5$ parameters, concerns a special case of a model introduced in Migon and Harrison (1985), and discussed in West and Harrison (1989, Section 14.4). The data comprise the first 56 weeks of observations on two series, $Z_t$, a binomial response variate, and $X_t$, an independent variable used to explain and predict $Z_t$. $Z_t$ is a count, out of a total of 66 randomly sampled, of individuals who, when questioned, demonstrate their awareness of a current television advertisement; $X_t$ is a summary measure of advertising expenditure in week $t$, referred to as the weekly TVR (standing for television rating). The model is a binomial regression of $Z_t$ on current and past values of $X_t$, defined by $(Z_t | p_t) \sim \text{Bin}[66, p_t]$ with

$$\begin{align*}
p_t &= \alpha + E_t, \\
E_t &= \gamma - (\gamma - \rho E_{t-1}) \exp(-\kappa X_t),
\end{align*}$$

for $t = 1, \ldots, 56$. The success probability $p_t$ measures awareness, $E_t$ represents the 'effect' of current and past TVR levels on current awareness, $\alpha$ is a lower threshold on the effect, the sum $\alpha + \gamma$ an upper threshold, $\rho$ is a measure of retention, or 'memory', of past advertising effects, and $\kappa$ is a measure of the immediate, penetrative effect of current advertising. The second equation in (2) may be used to recursively evaluate each $E_t$, $(t = 1, \ldots, 56)$, for any specified values of the five uncertain quantities $(\alpha, \gamma, \rho, \kappa, E_0)$. For analysis using importance functions based on mixtures of multivariate T-9 distributions, transformation is made to real-valued parameters. Parameter restrictions are $1 > \rho > 0, \kappa > 0, 0.2 > \alpha > 0, 0.8 > \gamma > 0$ and $E_0 < \gamma$, although the final restriction is ignored since $E_0$, and all subsequent values $E_t$, typically lie well below threshold. Parameter transforms used here are then $\theta_1 = \log(\alpha/(0.2 - \alpha)), \theta_2 = \log(\gamma/(0.8 - \gamma)), \theta_3 = \logit(\rho), \theta_4 = \log(\kappa)$, and $\theta_5 = \logit(E_0)$. The initial prior for $\theta = (\alpha, \gamma, \rho, \kappa, E_0)^t$ is a 5-variate T with 9 degrees of freedom whose mode and scale matrix are specified in line with the prior used in dynamic model analysis in Section 14.4 of West and Harrison (1989). The prior mode vector is $(0.0, 2.0, 2.0, -3.5, -0.75)^t$ and the scale matrix is diagonal, the square root of diagonal elements given by $(0.15, 1.0, 0.2, 0.25, 0.25)$.

Illustrated analysis uses three stages of refinement, with sample sizes $n_0 = 1500$, $n_1 = 2000$ and $n_2 = 6500$, and with $g_0(\theta)$ simply the prior. Repeat analyses with varying, and larger, sample sizes verify the results of this particular analysis. Final approximations to the univariate posterior marginals for $\rho$, $\kappa$, $\alpha$ and $\gamma$ appear in Figure 1 (as full lines), and contours of some of the bivariate marginals in Figure 2. The contours are graphed at heights of 0.001, 0.01, 0.1, 0.25, 0.5, 0.75 and 0.9 times the maximum. Superimposed on the posteriors in Figure 1 are, as dashed lines, the prior
Modelling with mixtures

Valencia, April 1991

densities. Over the 56 weeks of data, variation in the regression variable \( X_t \) is such that there are periods when \( X_t = 0 \) for a while, leading to potential for learning about the ‘memory decay’ parameter \( \rho \); the realisation of this potential is reflected in the figure for \( \rho \). Similarly, variation in \( X_t \) when non-zero is necessary to inform about the ‘penetration’ parameter \( \kappa \), and the graph reflects a reasonable degree of information for \( \kappa \) from this dataset. The posteriors for \( \alpha \) and \( \gamma \), however, differ only slightly from the priors; the response series \( Z_t \) never reaches very low or very high levels, so that there is little opportunity to gain information about these two ‘threshold’ parameters.

— Figure 1 goes about here, with caption:

Figure 1. Advertising model: Marginal priors (broken lines) and margins of final posterior (full lines).

— Figure 2 goes about here, with caption:

Figure 2. Advertising model: Some bivariate posterior densities.

2.3 Reduction of mixtures

Sampling and, especially, evaluating the density functions of mixtures of multivariate T distributions is computationally expensive if the number of components is very large. Also, mixtures of very many components can rather often be very well matched by reduced mixtures of far fewer, as experiences in density estimation show (West, 1990c). West (1990a) describes the reduction of weighted kernel estimates to mixtures of much smaller numbers of components, often lower than 10\% of the original sample size, by simply dropping components of negligible weights and clustering others that are close. Hierarchical nearest neighbour clustering, ‘averaging’ the component of smallest weight at each stage with its nearest neighbour, has been found to be effective. This involves sequentially identifying the nearest neighbour of that \( \theta_j \) with smallest weight at each stage, and so is computationally intensive in problems in several dimensions. It is both statistically and computationally effective, however, in the context of adaptive mixture refinement when the simulation sample size is large, say several thousands, and the mixture is later to be sampled and its density evaluated.

In the advertising model example, this mixture reduction was actually made between stages, reducing the mixtures of 1500, 2000 and 6500 components, respectively, to 500. The posteriors in the figures are based on the reduced 500 component mixtures. For comparison, Figure 3 redisplay the final univariate margins; the 500 component mixtures appear as full lines, and the 6500 component mixtures from which they were constructed as broken lines. The adequacy of the reduced approximation is clear here, and similar agreements are found between contour plots of bivariate margins.

— Figure 3 goes about here, with caption:
2.4 Sequential analyses and dynamic models

An original motivation for the development of the preceding sections was sequential analysis of non-linear dynamic models (West and Harrison, 1989). Simulation based analyses involve the propagation over time of discrete approximations to prior and posterior distributions based on approximate samples from those distributions, or from approximations to them. In West (1990b), a general approach is worked out in detail, with illustrations.

A very simple example illustrating the novel features of the computational problems arising in dynamic models is given by a non-normal, first-order polynomial model (Pole and West, 1988, 1990). Here the data, $y_t$, ($t = 1, 2, \ldots$), are modelled as $y_t = \theta_t + \nu_t$ and $\theta_t = \theta_{t-1} + \omega_t$ where $t$ indexes time, $\theta_t$ is the level of the series at time $t$, $\nu_t$ and $\omega_t$ are independent, zero-mean noise terms with some specified, non-normal densities $f_\nu(.)$ and $f_\omega(.)$ respectively. Heavy-tailed noise distributions, such as Student T, permit the automatic modelling of outliers and jumps in level in the series, which may be a component of a larger system model. Let $D_t = \{y_t, y_{t-1}, \ldots, y_1\}$ for all $t$. Analysis involves sequential updating of distributions at each time point; assume at time $t-1$ that the current posterior for the level, namely $p(\theta_{t-1}|D_{t-1})$, is available and summarised. The computations at time $t$ involve the steps

$$p(\theta_{t-1}|D_{t-1}) \rightarrow p(\theta_t|D_{t-1}) \rightarrow p(\theta_t|D_t) \downarrow$$

$$p(y_t|D_{t-1})$$

(3)

The first step is the evolution, involving computing the implicitly defined convolution $p(\theta_t|D_{t-1}) = \int f_\omega(\theta_t - \theta_{t-1})p(\theta_{t-1}|D_{t-1})d\theta_{t-1}$. Given an approximate sample from $p(\theta_{t-1}|D_{t-1})$, or from an appropriate importance function, this convolution may be evaluated directly by Monte Carlo. Similarly, the evolution may be directly simulated to generate points from $p(\theta_t|D_{t-1})$ given sampled values from $p(\theta_{t-1}|D_{t-1})$, or from a smooth reconstruction of the latter such as a weighted kernel estimate. So a Monte Carlo approximation may be transformed through the first step in (3). The forecasting step to simulate and evaluate $p(y_t|D_{t-1})$ may be similarly performed, since this involves just another convolution of distributions, $p(y_t|D_{t-1}) = \int f_\nu(y_t - \theta_t)p(\theta_t|D_{t-1})d\theta_t$. The final step involves updating the Monte Carlo approximation for $p(\theta_t|D_{t-1})$ to one for $p(\theta_t|D_t)$, processing the current observation $y_t$, and here the previous development of mixture approximations in Section 2.2 may be directly applied.

Take $f_\nu(.)$ to be standard Student T-5, and $f_\omega(.)$ to be T-5 with a scale factor of 0.1 (compare example 8.1 in Pole and West, 1988). In such models, prior and posterior densities can easily become multimodal and exhibit marked skewness and shoulders. To illustrate, suppose observations 1–6 are
Modelling with mixtures

given by 6,6,0,0,0,6, and that the initial prior \( p(\theta_1 | D_0) \) is standard T-9. Figure 4 gives estimated
priors \( p(\theta_t | D_{t-1}) \) (broken lines) and posteriors \( p(\theta_t | D_t) \) (full lines), for \( t = 1, \ldots, 6 \). Initially, the
‘conflict’ between the prior for \( \theta_1 \) and the likelihood function based on observing \( y_4 = 6 \) results
in the posterior strongly favouring the prior (which dominates due to larger degrees of freedom)
though a fair amount of posterior mass is spread over the interval 0–6. Additional uncertainty is
introduced in evolving to \( \theta_2 \), and then the second observation \( y_2 = 6 \) confirms the first, resulting in
\( p(\theta_2 | D_2) \) concentrated near 6, though with a longer left-hand tail. Spreading this tail through the
evolution to time 3 leads to \( y_3 = 0 \) shifting mass back towards zero, this is reinforced by \( y_4 = 0 \) and
\( y_5 = 0 \), and then sufficient credibility has been invested in \( \theta_5 \) values near zero that the conflicting
observation \( y_6 = 6 \) is largely rejected as an outlier.

— Figure 4 goes about here, with caption:

Figure 4. Prior and posterior densities for level of non-normal process.

The computations here use mixtures of T-9 distributions as approximating forms. Updating
from \( p(\theta_t | D_{t-1}) \) to \( p(\theta_t | D_t) \) at each \( t \) involves three successive refinements with sample sizes \( n_0 = 250, n_1 = 350 \) and, finally, \( n_2 = 1500 \). Following each refinement, these are clustered into mixtures
of just 100 components before sampling for the next iteration. Thus the densities in the figure are
mixtures of just 100 T-9 densities.

3. MIXTURES MODELS FOR DATA 1: DISCRIMINATION AND DECONVOLUTION

A common application of mixtures is in problems of discrimination, classification and clustering
(Bernardo and Girón, 1985; Binder, 1978; Hartigan, 1975, Chapter 5; McClacllan and Basford,
(Gelfand and Smith, 1990), now provide for many relevant calculations to be performed in practical
models, particularly models based on mixtures of standard exponential family distributions, such
as mixtures of normals or exponentials. This is demonstrated and illustrated in Lavine and West
(1991), in the context of mixtures of multivariate normals, and is currently being explored in
exponential, and other, mixtures for application in survival and reliability analysis.

Suppose \( p \)-dimensional data \( y_j, (j = 1, \ldots, n) \), are a random sample from a mixture of \( k \)
normals, \( y_j \sim \mathcal{N}(\mu_i; \Sigma_i) \) with probability \( \theta_i \), for \( i = 1, \ldots, k \); \( k \) is known and we have mean vectors
\( \mu = \{ \mu_i; i = 1, \ldots, k \} \), variance matrices \( \Sigma = \{ \Sigma_i; i = 1, \ldots, k \} \), and component probabilities \( \theta = (\theta_1, \ldots, \theta_k) \). Introduce latent classification variables \( z_j \), where \( z_j = i \) implies that \( y_j \) is drawn from
component \( i \) of the mixture, or classified into group \( i \); then the \( z_j \) are conditionally independent with
\( P(z_j = i | \theta) = \theta_i \). Inference is typically required for \( \mu, \Sigma, \theta \) and the classifiers \( z = \{ z_j; j = 1, \ldots, n \} \).
Prediction of future observations, and classification of future cases according to possible mixture
component of origin, is also often at issue.

Write \( y = \{ y_j; j = 1, \ldots, n \} \). Then, with an appropriate prior structure, the joint posterior of
all quantities \( (z, \theta, \mu, \Sigma | y) \) factorises as follows.
Modelling with mixtures

Valencia, April 1991

(a) Fixing $z$ means the data are classified as $k$ independent normal samples, and a standard, conjugate analysis may be performed for each group if the component means and variances are assigned independent normal/inverse-Wishart priors (which may be based on previous analysis of completely classified training data). This leads to conditionally independent posteriors of the same form, with easily computed parameters.

(b) Assigning an initial Dirichlet prior to $\theta$ implies that $p(\theta|y, z, \mu, \Sigma) \equiv p(\theta|x)$ is also Dirichlet with simply updated parameters.

(c) For $z$ we have, for all $i$ and $j$, $P(x_j = i|y, \theta, \mu, \Sigma) \propto \theta_i p(y_j|\mu_i, \Sigma_i, x_j = i)$ and summing to unity over $i = 1, \ldots, k$. Here $p(y_j|\mu_i, \Sigma_i, x_j = i)$ is just the normal density function for group $i$, with mean vector $\mu_i$ and variance matrix $\Sigma_i$, evaluated at the point $y_j$.

This structure provides the basis for an iterative resampling scheme to draw approximate samples from the joint posterior, and such samples are also the basis of predictive inference for future cases. In particular, a further case $y_f$ drawn from the mixture has predictive density $p(y_f|y)$ evaluated by Monte Carlo as a mixture over (many) conditional predictive $T$ distributions. For such a case, classification is based on predictive classification probabilities $P(x_f = i|y)$, similarly approximated. Full details appear in Lavine and West (1991).

An example from that reference involves a $k = 3$ component mixture of bivariate normals. With priors based on a training sample of just 15 classified cases, a further 300 unclassified observations are processed using the Gibbs sampling analysis to produce an approximate posterior sample of size 500, and used to evaluate the predictive density for a future case $y_f$; the contours of this multimodal density appear in Figure 5(a). Figure 5(b) provides contours of the predictive classification probabilities $P(x_f = 3|y)$ that may be used for discrimination – similar contours for groups 1 and 2 appear in Lavine and West (1991). Note that the usual linear or quadratic discrimination rules are clearly inappropriate – the contours of constant classification probability are neither linear nor quadratic. One particular feature of some interest is the fact that, while the chance of a future case coming from the third component naturally decreases away from the mass of the predictive density of that component, it increases again far from the mass, noticeably in the top right corner of the graph. This is a general phenomenon, and a feature that is completely missed by traditional discrimination techniques; on reflection, it is clear that, moving outwards in any direction from the mass of the predictive density, the classification probabilities will eventually favour that mixture components having largest variance resolved in that direction.

--- Figure 5 goes about here, with caption:---

Figure 5. Predictive contours and classification contours for normal mixture.
4. MIXTURE MODELS FOR DATA 2: DENSITY ESTIMATION

4.1 Mixtures of Dirichlet processes

Perhaps the most common use of mixture models is in non-parametric kernel density estimation (Silverman, 1976). Mixtures of Dirichlet processes (Ferguson, 1983) provide a natural model based foundation for Bayesian approaches to non-parametric density estimation, interpretation of kernel approaches as approximations to predictive densities, and insight into practical problems of smoothing parameter selection, local versus global smoothing, and multivariate and non-normal extensions (West, 1990c; Escobar and West, 1991). This class of models also provides a suitable setting for some problems of cluster analysis, although this remains to be explored in detail.

In perhaps the most useful normal based model, scalar data \( y_j, (j = 1, \ldots, n) \), are assumed conditionally independent with \( y_j \sim \mathcal{N}[\mu_j; V] \), the means \( \mu_j \) drawn independently from an uncertain prior \( G(.) \), and \( G(.) \) is modelled as a Dirichlet process with known base measure \( \alpha G_0(.) \). Given \( V \), and assuming \( G_0(.) \) to be \( \mathcal{N}[m_0; \tau V] \) for some scale factor \( \tau \), we can derive the following.

\( a \) \( k \)-configurations.

The \( n \) means \( \{\mu_j\} \) are not necessarily distinct — there will be some number \( k \leq n \) such that the \( \mu_j \) are drawn from some \( k \) distinct quantities \( \theta_1, \ldots, \theta_k \). The configuration of the means is defined in West (1990c) by an integer \( n \)-vector \( c = (c_1, \ldots, c_n) \) whose elements take values between 1 and \( k \), each such value appearing at least once, so that \( c_j = i \) implies \( \mu_j = \theta_i \). We label this arrangement of the means \( C_k(c) \), and refer to it as a \( k \)-configuration. Further, let \( n_j \) be the number of the \( \{\mu_j\} \) equal to \( \theta_j \), given by \( n_j = \#\{c_i = j; i = 1, \ldots, n\} \), for \( j = 1, \ldots, k \). Given \( C_k(c) \), the setup is a one-way classification; the \( y_i \) are classified into \( k \) groups with group means \( \theta_j \). Further, the \( \theta_j \) are independently drawn from the marginal prior \( G_0(.) \) so that, given \( C_k(c) \), posterior and predictive calculations can be performed using standard theory.

\( b \) Predictions under \( C_k(c) \).

Given the observed data \( y = \{y_1, \ldots, y_n\} \), the following conditional predictive distributions for a future case \( y_f \) are simply derivable (West, 1990c). Firstly, given the \( \mu_j \), or just the distinct values \( \theta_i \), the conditional predictive distribution is easily obtained. Writing \( \mu = \{\mu_1, \ldots, \mu_n\} \), we have

\[
(y_f|\mu, V, y) \sim \alpha a_n \mathcal{N}[m_0; (1 + \tau)V] + a_n \sum_{j=1}^{n} \mathcal{N}[^{\mu_j}; V] \\
\sim \alpha a_n \mathcal{N}[m_0; (1 + \tau)V] + a_n \sum_{i=1}^{k} n_i \mathcal{N}[\theta_i; V]
\]

(4)

where \( a_n = 1/(\alpha + n) \). This is a mixture of \( k + 1 \) normals, one for each of the existing \( k \) groups and one corresponding to the possibility of a new group. Similarly, given the observed data \( y \) and
Modelling with mixtures

Valencia, April 1991

knowledge of $C_k(c)$ (though not of the actual values of the $\theta_i$), it follows that

\[(y_f|V, C_k(c), y) \sim \alpha a_n \mathcal{N}(m_0; (1 + \tau)V) + a_n \sum_{j=1}^{k} n_j \mathcal{N}(m_j; (1 + w_j/n_j)V)\]  \hspace{1cm} (5)

where, for each $j = 1, \ldots, k$, $m_j = w_j \bar{y}_j + (1 - w_j)m_0$ and $w_j = \tau n_j/(1 + \tau n_j)$, with $\bar{y}_j$ the sample mean of the $n_j$ observations in group $j$.

Simple, but practically very important extensions provide for learning about $V$ (West, 1990c). Suppose $V^{-1}$ is assigned a gamma prior with specified shape $s_0/2 > 0$ and scale $S_0/2$, so that $\nu_0 = S_0/s_0$ is a prior estimate of $V$ and $s_0$ the associated prior degrees of freedom. For each group $j = 1, \ldots, k$, define $R_j = (\bar{y}_j - m_0)^2 n_j/(1 + n_j \tau)$ and $S_j = \sum(x - \bar{y}_j)^2$, the sum being over all data points $x$ in group $j$. Further, set $s = s_0 + n$ and $S = S_0 + \sum_{j=1}^{k}(S_j + R_j)$, and let $\nu_n = S/s$ be the corresponding posterior estimate of $V$. Then (5) is simply modified to a mixture of $(k + 1)$ Student T distributions,

\[(y_f|C_k(c), y) \sim \alpha a_n T_s[m_0; (1 + \tau)\nu_n] + a_n \sum_{j=1}^{k} n_j T_s[m_j; (1 + w_j/n_j)\nu_n].\]  \hspace{1cm} (6)

Further generalisation to allow for different scale factors $V = V_j$ for group $j$ is important in some applications – this induces different degrees of data-dependent smoothing across the sample space. a point of difficulty for non-Bayesian kernel approaches but trivially incorporated in these models. Escobar and West (1991) describe this extension, and show how (6) is modified; see also Ferguson (1983).

(c) Prior for $k$.

The number of groups $k$, $(1 \leq k \leq n)$, has a prior determined by $\alpha$ and $n$, given in Antoniak (1974), which favours larger values of $k$ when $\alpha$ is large, and, for fixed $\alpha$, decays rapidly as $k$ increases. For $\alpha = 1$, corresponding to a reasonably imprecise prior, $E(k) \propto \ln(n)$ for large $n$ so that small numbers of mixture components are anticipated relative to $n$. For finite $n$, $P(k)$ is rather diffuse relative to a Poisson prior, but can be shown to be approximately Poisson for very large $n$.

(d) Multimodality and smoothing.

A major feature of mixture modelling is the capacity to model multimodal data distributions. For fixed $\alpha$ and $n$, (6) may be multimodal depending on the observed data configuration and the value of the prior variance $\tau$; larger $\tau$ is consistent with more spread amongst the group means and hence greater chance of multimodality. West (1990c), and Escobar and West (1991), relate choice of $\tau$ to traditional smoothing parameter selection, and show how $\tau$ may be estimated (together with the observational variance $V$, possibly varying across groups). It is clear, however, that the information about $\tau$ available from the data is often scant, and informed priors are needed in the same way that subjective choice of smoothing parameters is typically advocated by kernel estimators (Silverman, 1976). Some applications of density estimation focus on ‘bump-hunting’ to identify the number
of modes, often using this as a proxy for inference about the number of components of a mixture (Silverman, 1981; Roeder, 1990). A key reason is that methods exist for inference about the modality of mixtures, but traditional approaches do not easily lead to inferences about numbers of components. The framework here, by contrast, provides priors, and hence posteriors, for the number of components \( k \) directly.

The number of modes \( h \), say, will always be less than \( k \), and the prior \( P(h) \) will depend only on \( \alpha \), \( n \) and \( \tau \). By simulating from the model for given values of these parameters, the predictive densities (5) or (6) can be evaluated and the number of modes identified — this provides a Monte Carlo estimate of \( P(h) \). Figure 6 provides summary information on \( P(h) \) for the model with \( V = 1 \), \( n = 50 \) observation, the two cases \( \alpha = 1 \) and \( \alpha = 5 \), and as a function of the smoothing parameter \( \tau \) over the range \( 0 < \tau \leq 350 \). Here the model was simulated 10,000 times, the number of modes in (5) being counted by evaluation over a very fine grid each time. The figure indicates prior probabilities \( P(h) \), at each value of \( \tau \), by circles with radii proportional to probability; also, the full line in each frame connects prior modes (to the nearest integer). Though larger \( h \) values are favoured for larger \( \tau \), the insensitivity of the prior for larger values, say \( 50 < \tau \), is apparent, and this is consistent with practical experiences that indicate robustness of posterior predictions across such ranges (West, 1990c).

— Figure 6 goes about here, with caption:

Figure 6. Priors for number of modes \( h \) in predictions from mixture of Dirichlet processes.

4.2 Simple mixture approximations to predictive distributions

In practice, of course, the number of groups and configuration of the data into groups is unknown. Predictions should be based on the marginal predictive density defined by marginalisation over configurations, \( p(y_j|y) = \sum p(y_j|C_k(c), y)P(C_k(c)|y) \), the sum being over all configurations and all \( k \). The computations implied here are, for realistic sample sizes, impossible to perform, as is well known (Antoniak, 1974; Berry and Christenson, 1979; Ferguson, 1983; West, 1990c). In developing approximations, several authors identify the extreme configuration \( C_n(1) \), where 1 is the unit \( n \)-vector, as leading to a Bayesian analogue of more traditional kernel procedures. Under \( C_n(1) \), the \( \mu_j \) are distinct and (6) becomes a mixture of \( n + 1 \) terms; with \( n \) at all large relative to the initial precision \( \alpha \), this is close to a kernel form with \( n \) components located near the observations, since the component modes are \( m_j \approx y_j \) for large \( n/\alpha \). This configuration, however, typically has very low prior probability \( P(k = n) \), and in practice will usually have extremely low posterior probability too. West (1990c), develops an hierarchical search procedure to explore the space of configurations and to identify values of \( k \) and configurations \( C_k(c) \) of high posterior probability. This search is trivially implemented, as described in that reference, and leads to simply computed relative posterior probabilities over a single chosen configuration for each value of \( k \). These probabilities are typically maximum over smaller values of \( k \), and decay rapidly otherwise. The favoured
Modelling with mixtures

configurations lead to predictive densities that are mixtures of smaller numbers of components, and these may be used either directly as conditional predictions or approximately marginalised to estimate $p(y_f | y)$. The results in West (1990c) may be summarised as follows: (i) though the extreme $n-$configuration $C_n(1)$ typically will have little support under the posterior over configurations, the distribution function of the resulting ‘Bayesian kernel density estimate’ (equation (6) at $C_k(c) = C_n(1)$) will typically be close to that of other configurations with much greater support; (ii) the search algorithm identifies approximate posterior modal configurations that are typically mixtures of small numbers of components, and the resulting predictive densities (6) are useful, and simple, approximations to marginal predictions; (iii) using the iterative sampling based analysis discussed below in Section 4.3, these simple mixture approximations can be compared with Monte Carlo approximations to the full marginal $p(y_f | D)$, and often turn out to be quite similar. The hierarchical search algorithm used to investigate configurations provides one means of identifying reduced mixtures, in point (ii), the computational benefits of reducing from $n$ to $k << n$ being of interest, for example, in applications where $n$ is very large and the predictive distributions is to be heavily evaluated, simulated, and so forth.

4.3 Sampling analyses

Escobar (1990) introduced a Gibbs sampling technique that applies to produce approximate samples from the posterior distribution of the means $\mu = \{ \mu_1, \ldots, \mu_n \}$ in (4). This is easily extended to analysis with $V$ uncertain, as described and illustrated in West (1990c), and further to models with $V$ varying across groups, and also to cover the case of uncertain $\tau$; details appear in Escobar and West (1991). Restricting attention for theoretical discussion here to (4), suppose we have available a sample $\{ \mu(r), V(r) ; r = 1, \ldots, N \}$ from the joint posterior $p(\mu, V | y)$. Thus, for each $r$, $\mu(r) = \{ \mu_1(r), \ldots, \mu_n(r) \}$ is a single sample of the mean values, and $V(r)$ the corresponding sampled value of $V$. There will be some $k(r)$ distinct values in the set $\mu(r)$, a sample from the posterior distribution over the number of groups $P(k | y)$. The latter permits inference about $k$ directly, and the former may be used to compute a Monte Carlo estimate of the predictive density $p(y_f | y)$ via $p(y_f | y) \approx N^{-1} \sum_{r=1}^{N} p(y_f | \mu(r), V(r), y)$, with summands defined in (4).

Complete details of the computational issues appear in the above references, where an example from Roeder (1990) is re-examined. The data there are measured velocities of a number of galaxies identified in six distinct conic sections of space. Clustering of the velocities of neighbouring galaxies is consistent with the Big Bang theory of the origin of the universe, so that interest lies in the number of components of a mixture of velocity distributions. While Roeder addresses the effects of uncertainty about density estimates on the assessment of multimodality, particularly on the hypothesis of unimodality, the scientifically interesting issue is that of clustering. The analysis reported in Escobar and West (1991) supports Roeder’s broad conclusion that there are most likely four modes in the distribution, though the uncertainty about modality is considerable, and goes further to the more relevant issue of inference about $k$. In a model with $\alpha = 1$ for this sample size
of \( n = 82 \), the prior \( P(k) \) supports \( k \) heavily supports values between 3 and 7 but is quite diffuse over 1–11. Posterior, across a range of plausible prior parameters, support rather large values of \( k \) around 7 or 8. The conclusions are rather typical of inferences with data consistent with heavily overlapping mixtures, but, while there is high uncertainty about \( k \), the computations here at least provide a formal assessment of such uncertainty, unlike traditional density estimation approaches.

As an example, consider the raw histogram of \( n = 159 \) observations in Figure 7(a). This data comes from neurological experimentation concerned with the assessment of neuronal responses to simulated current inputs. The study of neuronal responses typically focusses on the identification of patterns of neural response in terms of the distributions of numbers of neurons ‘firing’ and the distributions of magnitudes of response. This data comes from a collaborative study with Dr. D.A. Turner, of Duke Medical Center, and is part of a much bigger study just beginning. The mixture models of this section are perfectly suited to the data structure in this area, since the underlying neurological science hypothesises a mixture of normals for the sort of data displayed in the histogram. The uncertain number of components of the mixture and the distribution of the distinct means of the components relate to the numbers of neurons responding to external stimuli and the durations and magnitudes of responses. For illustration here, one analysis along the lines detailed in Escobar and West (1991) is summarised in Figure 7. The background to the data collection procedure provides a substantial amount of information on the observational variance \( V \) of the data. In this particular experiment, a similar dataset is available from a previous experiment designed to estimate \( V \) alone – based on similar neuronal measurements but without simulated stimuli. This routine calibration experiment provides \( n = 159 \) observations with roughly zero-mean and a sample standard deviation of 0.047, which suggests a prior for \( V \) of the inverse gamma form with \( s_0 = 158 \) and \( S_0 = 158 \times 0.047 = 7.05 \). Other, more diffuse priors may be used, of course, but this is taken here – existing techniques directly guess a value for \( V \) based on such calibration data. Notice that this prior strongly supports values of \( V \) that are very small compared to the observed spread of the histogram in Figure 7(a), indicating that there is substantial variation in the stimulus response data above that contributed by background noise. The analysis summarised actually uses the model of Escobar and West (1991) that allows for different values of \( V \) across the distinct normal components, though the concentration of the prior (with 158 degrees of freedom) heavily restricts the values to be close, and a reanalysis with common \( V \) would give substantially similar results. The analysis also includes the prior smoothing parameter \( \tau \), producing approximate posterior samples for the group means, hence the number of distinct components \( k \), the observational variances, and \( \tau \). The initial priors are specified with \( m_0 = 10.0 \), \( s_0 \) and \( S_0 \) as above, and an initial inverse gamma prior for \( \tau \) with shape 1.0 and scale 5.0; this latter is a fairly diffuse prior for \( \tau \), the density appearing as the dashed line in Figure 7(e). Iterative posterior sampling provides an approximate sample of size \( N = 500 \) on which the summary figures are based. The smooth line in Figure 7(a) is the Monte Carlo estimate of the predictive density \( p(y_f | y) \) for a future case \( y_f \) – a Bayesian density estimate. This is the Monte Carlo average of 500 sampled predictive densities, some 100 of which
are plotted in Figure 7(c) to give a rough idea of the posterior uncertainty about the estimate in Figure 7(a). Figures 7(b) and 7(d) give similar plots of the cumulative distribution function; as is typical, apparently large uncertainties on the density scale appear much less relevant on the probability scale (West, 1990c). The estimated posterior $p(\tau|y)$ appears as the full line in Figure 7(e), along with the prior. Finally, the Monte Carlo estimate of $P(k|y)$ appears as the full bars in Figure 7(f), the prior as dashed bars. With this data, the very precise prior on small values of $V$ is reconciled with the much greater spread in the observed data through the posterior favouring rather larger numbers of mixture components than the prior.

Further analysis under different priors for $V$ and $\tau$ are easily performed to explore sensitivity of posterior conclusions, and extensions to include uncertainty about the parameter $\alpha$ in the analysis are possible, though not explored here.

---

Figure 7 goes about here, with caption:

Figure 7. Inferences for neurological response data

REFERENCES


Figure 1

rho

kappa

alpha

gamma
Figure 2
Figure 3

rho

kappa

alpha

gamma
Figure 5

Predictive density

P(z=3|y)
Figure 6

(a) $n=50$, $\alpha=1$

(b) $n=50$, $\alpha=5$
Figure 7

(a) Predictive pdf

(b) Predictive cdf

(c) Sampled pdfs

(d) Sampled cdfs

(e) Pdfs $p(\tau)$ and $p(\tau | y)$

(f) Probabilities $p(k)$ and $p(k | y)$