BAYESIAN ANALYSIS OF MIXTURES
APPLIED TO POSTSYNAPTIC POTENTIAL FLUCTUATIONS

Dennis A. Turner
Neurosurgery and Neurobiology
Duke University Medical Center

Mike West
Institute of Statistics & Decision Sciences, Duke University

DP #92-A01
Bayesian Analysis of Mixtures
Applied to Postsynaptic Potential Fluctuations

By

Dennis A. Turner & Mike West
Neurosurgery and Neurobiology Institute of Statistics & Decision Sciences
Duke University Medical Center Duke University
Durham NC 27710 USA Durham NC 27708-0251 USA

Abstract

Bayesian inference techniques have been applied to the analysis of fluctuation of postsynaptic potentials in the hippocampus. The underlying statistical model assumes that the varying synaptic signals are characterized by mixtures of (unknown) numbers of individual gaussian, or normal, component distributions. Each solution consists of a group of individual components with unique mean values and relative probabilities of occurrence and a predictive probability density. The advantages of Bayesian inference techniques over the alternative method of maximum likelihood estimation of the parameters of an unknown mixture distribution include the following: 1) prior information may be incorporated in the estimation of model parameters; 2) conditional probability estimates of the number of individual components in the mixture are calculated; 3) flexibility exists in the extent to which the estimated noise standard deviation indicates the width of each component; 4) posterior distributions for component means are calculated, including measures of uncertainty about the means; and 5) probability density functions of the component distributions and the overall mixture distribution are estimated in relation to the raw grouped data, together with measures of uncertainty about these estimates. This expository report describes this novel approach to the unconstrained identification of components within a mixture, and provides demonstration of the usefulness of the technique in the context of both simulations and the analysis of distributions of synaptic potential signals.

Keywords: Bayesian inference; Hippocampus; Mixture model; Postsynaptic potentials; Quantal analysis; Stochastic release probability
Introduction

Trial to trial fluctuation of synaptic potentials occurs in many different physiological preparations, including the neuromuscular junction and most CNS neurons, such as motoneurons and hippocampal neurons (Del Castillo and Katz, 1954; Martin, 1966; McLachlan, 1978; Redman, 1990). These fluctuations have been analyzed statistically using a variety of different models, depending upon both the underlying hypothesis of synaptic potential generation and experimental conditions (particularly the concentration of divalent cations). For example, amplitudes and release probabilities are assumed to differ minimally across synaptic release sites at the neuromuscular junction. Thus, the statistics of release at the neuromuscular junction show a high degree of site-to-site uniformity for multiple-site release kinetics, which are spatially uniform in terms of location; these conditions favor a simple binomial summation of relatively uniform events with similar probability of occurrence (Del Castillo and Katz, 1954; Korn et al, 1981; Redman, 1990).

However, most excitatory synaptic terminations in the central nervous system (CNS) possess only individual release sites and are likely to be spatially segregated from each other due to their predominantly dendritic location, creating an underlying physiological model quite different from that observed at the neuromuscular junction. These individual release sites may not be uniform in terms of either postsynaptic effectiveness at the recording site or release probability. Additionally, the recording conditions at CNS sites tend to be complicated by a significantly higher noise level, for either sharp intracellular recording techniques (Foster and McNaughton, 1991; Foster et al, 1992; Friedlander et al, 1990; Jack et al, 1981; Larkman et al, 1991; Sayer et al, 1990; Turner, 1987, 1988; Walmsley et al, 1987, 1988) or using whole cell patch clamp techniques (Bekkers et al, 1990; Bekkers and Stevens, 1990; Malinow, 1991; Stern et al, 1992). The higher background noise levels and uncertainty about the origination site of the synaptic potential considerably complicate the evaluation of underlying events in a composite raw response histogram since the location and probability of either ‘miniature’ or summed events cannot be inferred easily.

In spite of these significant difficulties with transferring the neuromuscular junction statistical model to CNS synapses, simple Poisson and binomial release models have been predominantly applied to CNS synaptic potential data (Bekkers et al, 1990; Bekkers and Stevens, 1990; Foster and McNaughton, 1991; Foster et al, 1992; Malinow, 1991). However, other authors have tried to discern differences between synaptic sites on CNS neurons, either in terms of the site-to-site variability of release (using a compound binomial model) or site-to-site variability in amplitude (Friedlander et al, 1990; Jack et al, 1981; Redman, 1990; Sayer et al, 1990; Walmsley et al, 1987, 1988). Significant concerns also exist as to the stationarity of long runs of synaptic potential data, the statistical distribution of the underlying ambient noise and the relationship of calculated component parameters to the noise data (Brown et al, 1976; Clamann et al, 1991; Clements, 1990; Kullman, 1989; Larkman et al, 1991; Redman, 1990). Techniques for estimation of underlying parameters in CNS neurons include least-squares optimization methods (Wong and Redman, 1980) and maximum likelihood techniques (Kullman, 1989; Kullman and Nicoll, 1992; Ling and Tolhurst, 1983; Redman, 1990). However, these techniques require considerable constraints to achieve a solution, frequently ensembles cannot be analyzed at all and different solutions cannot be easily ranked according to goodness of fit. Thus, techniques for unconstrained analysis of synaptic potential ensembles may be very helpful for identifying components, and presumably individual synaptic sites, within a raw data histogram, particularly if these techniques allow coherent estimation of uncertainties regarding component parameters and assessment of goodness-of-fit.

The unconstrained mixture model assumes that a multi-site synaptic potential ensemble can be considered as a finite mixture of components, each possessing (initially) a normal, or gaussian, distribution. Our development is novel in several respects, particularly in the application of
Bayesian inference techniques. Statistical issues and models pertinent to mixture modeling generally are covered in Lavine and West (1992), Titterington et al (1985), and West (1992), with further details specific to the current approach in Ferguson (1983), and Escobar and West (1992). The Bayesian approach illustrated here overcomes many of the practical and technical difficulties associated with maximum likelihood and older techniques of mixture estimation. This stems primarily from the foundation in Bayesian statistics, deriving, notably, from the formal and direct use of probability to represent uncertainties about parameter values. A proposed model form combines with the observed data values to provide inferences about all model parameters summarized via a posterior probability distribution that represents the uncertainties about the parameters. Readers completely unfamiliar with Bayesian statistics may consult Lee (1989) for a brief but useful introduction, and Lindley (1965) for more thorough coverage. Some brief coverage of essential concepts in the mixture modeling paradigm appears as a technical appendix to this paper.

This report centers on solutions to the following questions regarding an ensemble of observed synaptic potential values: 1) how many components are there in the mixture and what is the likelihood for different numbers of such components? 2) what are likely values for the mean and variance of each component? Current techniques to solve these estimation problems are based on the EM-algorithm for computing maximum likelihood estimates (Kullman, 1989; Kullman and Nicoll, 1992; Ling and Tolhurst, 1983; Redman, 1990; Wong and Redman, 1980). One benefit of likelihood-based approaches is the clear definition and interpretation of derived estimates of parameters. However, as in all multiparameter problems, likelihood estimation does not adequately deliver appropriate and computable assessments of uncertainties about parameter estimates, particularly inferential statements about functions of several parameters, such as the problem of estimating the number of components. Thus, the techniques presented here may give additional information about the components underlying mixtures and the uncertainty regarding potential solutions of a finite mixture of gaussian components.

Models and Methods

**Synaptic data model and assumptions**

A generalized model of synaptic terminations onto cortical neurons (such as pyramidal cells in the hippocampus) points to the high likelihood that the individual synaptic locations (for even a small input stimulus) will be spread over a considerable dendritic surface. A major question arises as to whether the synaptic current is similar between synaptic sites at different dendritic locations, particularly with the possibility of enhanced receptor density at more distal sites (Redman, 1990). This variability between sites may be enhanced by the cable filtering occurring prior to the signal reaching the recording electrode (distal sites would be attenuated more). The end effect will be that synaptic potentials of often different shape indices (such as rise time and half-width) and highly variable amplitude will be summed together, resulting in an irregular location of components on histograms, since the individual sites may not be "equal" in any sense and thus may not sum to demonstrate "quantal" peaks. Rather, the amplitude histogram of an ensemble may exhibit "primary" components, representing the contribution from one or a few synaptic sites, and "secondary components", which would represent the necessary sums of the irregularly spaced components. For this reason, unconstrained mixtures, with components that are not necessarily quantally spaced, are appropriate to identify the (unknown) locations of components.

The data ensembles undergoing Bayesian analysis are assumed to possess a number of characteristics, including: 1) the evoked synaptic signal is adjusted to be sufficiently small so that only a few components may be present (biased into the range of less than 8–10 primary and secondary
peaks); 2) a set of interleaved noise samples are available, sampled as close as possible to the region in which the signal is evoked; 3) the distribution of the interleaved noise samples appears approximately gaussian (in this initial analysis); 4) the data set is approximately stationary, particularly in the sense of not exhibiting trends over time in overall level or significant changes in the fluctuation pattern; and 5) the time course of the standard deviation (Time SD) around the mean synaptic signal is similar to that of the mean (indicating the fluctuations are occurring primarily at the peak of the signal). Clearly, the smaller the number of components which comprise the signal data the more likely each component and its representative sums will be resolvable. In addition, the smaller the noise SD the more likely peaks are resolvable (Redman, 1990). Thus, the physiological data is assumed to be acquired under optimal conditions to maximize the possibility of resolving individual components, by minimizing the number of components involved and decreasing as much as possible the noise fluctuations in the baseline. The physiological data is summarized as a list of peak amplitudes together with the sample variance and degrees of freedom of the accompanying interleaved noise samples.

**Statistical concepts and Bayesian approach**

We assume that a single experiment generates data whose values are modeled as independent draws from a mixture of gaussian distributions - each draw comes from a single gaussian with mean and variance to be estimated, but it is not known from which of this finite collection of gaussians the draw is selected. To introduce notation, $x_1, x_2, \ldots$ are the data samples modeled as independent draws from the mixture distribution, and the probability density function of the mixture is:

$$ f(x) = \sum_{j=1}^{k} p_j N(x|t_j, v_j). \quad (1) $$

Here $N(x|t, v)$ denotes the normal density function $N(x|t, v) = (2\pi v)^{1/2} \exp(-(x-t)^2/2v)$, $k$ is the number of distinct normal components in the mixture, and the $j^{th}$ component normal density has a weight determined by the probability $p_j$. Clearly, these probabilities must sum to unity. The parameters of the model are the collection of component means, variances and weights, together with the number $k$ of components. The class of mixture models considered here lead to mixtures of the form (1) arising from a more fundamental description of uncertainty about the data distribution. Fuller conceptual and technical details of the models, referred to as Dirichlet mixture models, appears in the Appendix, and a basic description follows.

A Dirichlet mixture model actually implies a population distribution that is a mixture of the form (1) but with an infinite number of components. However, the Bayesian analysis of these models naturally implies use of only finite numbers of components. Assume a data sample size $n$, so we are to observe data values $x_1, x_2, \ldots, x_n$. The questions of inference focus on two key points:

- How many distinct normal components are these $n$ observations likely generated from?
- What are likely values for the moments of such components?

In maximum likelihood, and other, approaches, the number of components is typically assumed known or fixed prior to analysis, often by guessing a suitable value in the light of graphical displays of the recorded data. With clearly separated components, whose component variances are small in comparison with the distances between means, this can be done with some confidence that uncertainties about the numbers of components is negligible. Otherwise, it is very difficult to precisely estimate $k$. In some problems, there may be underlying theory or structure that suggests plausible values, or ranges of values, for the number of components (just as there may be relevant
prior information regarding values of the component means and variances). In a Bayesian approach, this relevant information can, and should, be reflected in specified prior probability distributions over values of \( k \). The Dirichlet mixture framework implicitly defines a class of priors for \( k \). These prior distributions are determined by the choice of a single scalar parameter. How this enters into determination of \( k \) and of the mixture weights \( p_j \) is described below.

Bayesian inference naturally encompasses a predictive perspective on modeling. From such a perspective, the estimation of a population density or distribution function, whatever the model, translates to that of identifying a predictive distribution for a further observation assumed drawn from that population. In the Dirichlet mixture model, suppose that we identify the observed values \( x_1, \ldots, x_n \), as arising from a collection of some \( k \) mixture components (in practice this is the theoretical structure, though \( k \) is unknown). Let these be the components indexed \( j = 1, \ldots, k \) in equation (1). In predicting a further observation \( x = x_{n+1} \) in the Dirichlet mixture framework, the relevant predictive distribution is a minor modification of the mixture (1); the number of components is extended to \( k + 1 \) giving a mixture \( \sum_{j=0}^{k} p_j N(x|t_j, v_j) \). The interpretation is that the further observation \( x \) may come from one of the initial \( k \) components already represented in the observed data and in (1), but that it may also be generated from a further component not identified for the first \( n \). This view of sequential generation of data is useful in describing the key features of this class of Dirichlet mixture models, and is developed further in the following section. In addition, these models naturally and simply provide mechanisms for generating appropriate values of, and describing uncertainties about, the means and variances of the normal components.

**Mixture model parameters**

The characteristics of data derived from this statistical model depend on the specified values of model parameters, to be described below. This statistical approach also incorporates a probabilistic description of parameter uncertainty that is fundamental to the Bayesian approach to scientific inference.

Consider sampling the data set \( x_1, \ldots, x_n \) sequentially, beginning with \( x_1 \), then \( x_2 \), and so on. To start, \( x_1 \) comes from some normal distribution - one of the components of the mixture, but which component is unknown. Suppose, for notational convenience, \( x_1 \) is drawn from \( N(x|t_1, v_1) \) for some moments. Next, \( x_2 \) may come from that component, or a distinct component \( N(x|t_2, v_2) \), say. A deduction from the Dirichlet mixture framework is that \( x_2 \) comes from the same component with probability \( 1/(1+a) \) where \( a > 0 \) is a positive parameter of the Dirichlet mixture model, called the precision parameter. The value of \( a \) must either be specified initially or may be estimated from the data in the analysis (together with \( k \) and the component moments) as we shall see later. Here assume for illustration \( a \) has been specified. According to the physiological assumptions defined above, the number of distinct components represented in a sample of size \( n \) is expected to be small relative to \( n \) (for example, 8-10 components at most for a sample size of \( n=500 \)). In this particular analysis \( a \) will take small values. The \( a \) value directly, and exclusively, determines the stochastic process by which the distinct normal distributions are generated as the data is sequentially sampled.

For example, proceeding to sample \( x_3 \) we identify \( x_3 \) as coming from a normal distribution distinct from those of \( x_1 \) and \( x_2 \) (whatever they are) with probability \( a/(2+a) \); continuing this process, we have probability \( a/(j+a) \) that the observation \( x_{j+1} \) is drawn from a normal component distinct from those already represented in sampling \( x_1, \ldots, x_j \). In this process, some component normal distributions will, with some probability, be represented more than once, resulting in some total number \( k \leq n \) of distinct components. In addition, we will have some \( k \) sample sizes \( n_1, \ldots, n_k \) allocated to the distinct components; that is, exactly \( n_j \) of the data values are drawn from compo-
nent \( j \) of the mixture, with \( n_1 + \cdots + n_k = n \). Then, in drawing a further observation \( x = x_{n+1} \), we select a further distinct component with probability \( a/(n+a) \) and reselect one of the existing components with probability \( n_j/(n+a) \) of choosing component \( j \). So inference about future observations is based on a density of the form (1) with \( p_j = n_j/(n+a) \) (setting \( n_0 = 1 \)).

An initial distribution is generated, over \( k \) values, which depends only on \( a \) and the sample size \( n \) according to the sequential selection probabilities for distinct components. Call this prior distribution \( P(k) \). Note that these probabilities depend on both \( n \) and \( a \) although, at this point, we do not make this explicit in the notation for clarity. The Dirichlet mixture model theory tells us that \( P(k) > 0 \) for all values of \( k \leq n \), so that no value of \( k \) is ruled out initially. However, unless \( a \) is very large, the function \( P(k) \) concentrates on smaller numbers of components, having a unimodal form similar to a Poisson distribution. A basic guideline is that, for reasonably large sample sizes \( n \) and so long as \( k \) is unlikely to be very large, we expect \( k \) to be of the order of \( a \ln(1+n/a) \). For example, with \( n \) ranging between, say, 300 and 1000, and \( a = 1 \), \( k \) values will likely lie between 2 and 10. Larger values of \( a \) are consistent with higher chances of larger numbers of components. Table 1 gives example prior probabilities for the distribution of \( k \) for sample sizes \( n = 300, 500 \) and 1000. The table gives chances of \( k \) components for values \( a = 0.001, 0.01, 0.1 \) and 1.0. Note especially the insensitivity of table entries to the sample size \( n \), most marked for the smaller value of \( a \). With \( a \) towards the lower end of the unit interval the prior probabilities of \( k = 1, 2, 3 \) and 4 are essentially unaffected by sample sizes \( n \) in the 300–1000 range, a range of interest here.

In many practical problems, population distributions are adequately modeled using mixtures of small numbers of components, consistent with small values of \( a \) in our framework. If data distributions show marked evidence of multimodality, larger numbers of components may be necessary to adequately reproduce such features, consistent with larger values of \( a \). A facility for estimating \( a \) might be desirable in some applications; otherwise, \( a \) may be viewed as a control parameter to be chosen externally, or to be subject to sensitivity analysis involving exploration of inferences based on various values of \( a \). These points are illustrated below for simulated and example data sets.

Now consider the means and variances \( t_j \) and \( v_j \) of the components. The values of these moments determine the characteristics of data in the model, once \( k \) is generated and given a component index \( j \). In the Dirichlet mixture model, as in any Bayesian approach, the moments are assumed to have an initial, or prior, distribution. Following previous authors (Ferguson, 1983; Escobar and West, 1992) it is natural to adopt conjugate prior forms, and the use of such priors is illustrated below. The structure assumes that initial uncertainty about the moments is represented as follows:

1. Firstly, the pairs \((t_j, v_j)\) are independent and identically distributed. This is a neutral initial standpoint imposing no constraints on possible relationships among the component moments in advance of the data analysis. Note that this assumption is a direct implication of the Dirichlet mixture model.

2. Secondly, the mean \( t_j \) has a conditional normal prior distribution, \( N(t_j|m, wv_j) \), conditional on any fixed value of \( v_j \). The values \( m \) and \( w \) are prior parameters that will be estimated in fitting the mixture model to a particular data set. The positive number \( w \) appears as a common variance multiplier and is responsible for determining the relative spread of the distinct means \( t_j \). If \( w \) is small, then the \( t_j \) are likely to be grouped closely together relative to the random data variation measured by the variances \( v_j \). If \( w \) is large, the means \( t_j \) will be more dispersed, so that the resulting mixture of normals can become quite asymmetric, heavily skewed and multimodal. So two parameters \( a \) and \( w \) together control and determine the extent of non-normality in the model; \( a \) partially determines the number of components in the mixture and \( w \) partly determines how separated and relatively diffuse the components
are. In the next section, we will explore the dependence of the analysis on these parameters and describe how they may be estimated.

Table 1 about here

(3) Thirdly, and finally, we must define an initial prior distribution for the component variances \( \nu_j \). There are two distinct approaches possible here. The first assumes \( \nu_j = \nu \), constant over \( j \), with a specified prior distribution for the common and constant variance \( \nu \). Taking \( \nu \) as the noise variance corresponds to the assumption that the signal data are obtained exactly as component means plus noise, with no other source of variance (such as variability in actual component size). In many other applications of mixture modeling, the idea that component variances may not be constant is critical. Data distributions often and typically evidence greater dispersion in the tails, where uncertainty about distributional form is greatest, than in the center of the range. This calls for component variances differing if the means differ appreciably, and allowing the data to assist in estimating the differences. If a constant variance model is used and such features are indeed apparent in the data, then resulting inferences will be subject to various possible biases. Over-estimation of the global variance \( \nu \) will result if \( k \) is fixed or constrained to smaller values. On the other hand, over-estimation of \( k \) will result if \( \nu \) is fixed or constrained to small values. Estimation procedures that effectively allow for differing \( \nu_j \) across components can avoid these biases. Thus the second, and most appropriate, prior structure assumes the \( \nu_j \) differ but that, initially, the prior describes these as independently drawn from some prior distribution; the conjugate prior is an inverse gamma distribution (a generalization of a scaled \( \chi^2 \) distribution, as detailed further in the Appendix). This prior structure allows for variation of the component variances rather than a strict reliance on the estimated noise variance.

Mixture model estimation and inference

Bayesian analysis involves learning from the data \( x_1, \ldots, x_n \) about the \( k \) and the pairs \( t_j, \nu_j \), and incidentally about the additional nuisance parameters \( a, m \) and \( w \). Write \( D \) for the observed data \( D = \{ x_1, \ldots, x_n \} \). Inference involves computing and summarizing the joint posterior distribution for the parameters of interest. For example, the mean of each \( t_j \) under this posterior distribution is one possible point estimate of the actual values of \( t_j \); the mode is another and ranges can also be calculated. The marginal posterior distribution for \( k \), denoted \( P(k|D) \), provides probabilities that indicate numbers of components that are judged likely and unlikely in the light of the data \( D \). This represents a revised or updated version of the prior \( P(k) \) discussed in the previous section.

In practical application, the critical parameters \( w \) and \( a \) will usually be estimated in the Bayesian analysis. Resulting posterior probabilities for \( k \) are then data-weighted averages over all possible values of \( w \) and \( a \), the weighting based on the posterior distribution \( p(w,a|D) \) delivered in the analysis. Explicitly, \( P(k|D) = \int P(k|a,w,D)p(a,w|D)\,da\,dw \) where \( p(a,w|D) \) denotes the posterior density for the two parameters based on the analysis. We note that inference about, for example, \( k \) using \( P(k|D) \) formally and completely accounts for the estimation and uncertainty about all the other parameters - the means and variances of components and the additional parameters. In order to be able to do this, the Bayesian analysis requires initial specification of prior distributions for the additional parameters \( a, m \) and \( w \) (the priors for the \( t_j \) and \( \nu_j \) already being specified as discussed above). The following assumptions for the prior distributions are noted:
(1) The routine default or reference prior distribution is assumed for the pair \( m, w \); this has the form \( p(m, w) \propto w^{-1} \) (Lindley, 1965, volume II, p40) and is taken as representing initial indifference about the values of \( m \) and \( w \).

(2) The noise variance estimate provides a baseline guide to use in specifying the inverse gamma prior for each of the \( \nu_j \). The degrees of freedom of the noise variance may be reduced to allow for some deviation of the \( \nu_j \) away from the previously specified noise value if the signal data are in substantial conflict with the single restricted value. Bias will be towards larger values of \( k \) if the noise standard deviation actually underestimates some of the component variances. However, the results may in many instances not be sensitive to this reduction in the degrees of freedom of the noise variance.

(3) The precision parameter \( a \) is a critical smoothing parameter for the model. Learning about \( a \) from the data may be addressed by assuming an initial gamma prior distribution; Escobar and West (1992) give full details based on the development in West (1992). It is natural to assume priors for \( a \) that induce priors for \( k \) (cf. Table 1) supporting numbers of components that are initially viewed as likely; with sample sizes in the 300–1000 range and expected numbers of components likely to be in single digits, this implies priors for \( a \) that place most mass on values around or less than unity, but are diffuse enough so that the data may substantially alter inferences about \( a \). The gamma prior with density \( p(a) \propto a^{a_0-1}e^{-ra_0/a_0} \), for \( a > 0 \), has a mean of \( E(a) = a_0 \) and shape parameter \( r_0 \). For small \( r_0 \) this is very diffuse over a wide range of \( a \) values; large values of \( r_0 \) imply greater concentration around the initial point estimate \( a_0 \). Just how inferences about mixture structure depend on \( r_0 \) and \( a_0 \) are explored below. As noted in Table 1, larger values of \( a \), hence larger values of the prior estimate \( a_0 \), predispose the analysis towards larger values of \( k \) and so may result in overestimating the number of components underlying the mixture. Significantly lower values of \( a \) or \( a_0 \) (such as in the 0.001–0.01 range) are required to counter this tendency. Even with prior parameters supporting such small values, however, strong indications of several components in the data will still result in the posterior for \( k \) supporting appropriately large numbers.

The analysis program using this statistical model currently calculates the following summary of posterior distributions:

(1) The mean and standard deviation for the locations of each component, the posterior marginal probability for the presence of additional components, the relative probability of the component and the predictive distribution of each component. These values define the significant components (relative probability > 0.01), the estimated marginal probability for additional components, and the estimated moments for each component.

(2) A mean predictive probability density function as a function of the control parameters \( r_0 \) and \( a_0 \), with a range determined by the individual sampled distributions (see Appendix for details, with 5000 sampled distributions calculated). A cumulative probability function can also be calculated. These represent the Bayesian estimates of the underlying mixture density and distribution functions. Assessments of uncertainties about these estimated functions are available in terms of point-wise intervals inducing uncertainty "bands" around the estimates.

(3) Posterior probability density functions for the control parameters \( a \) and \( \omega \), as determined by the data set. Marginal likelihood functions for these parameters can also be computed.

(4) Estimates of goodness of fit of the Bayesian predictive probability distribution to the raw histogram for the data. These includes traditional \( \chi^2 \) estimates and Kolmogorov-Smirnov assessments, plus more scientifically based measures of relative fit based on posterior distributions for the control parameters, especially \( a \). The \( \chi^2 \) values are given as numbers, since
assessment of appropriate degrees of freedom is obscure. However, the \( \chi^2 \) values are very low, well below traditional significance levels for any degrees of freedom. Hence the \( \chi^2 \) measure is effectively meaningless in any absolute sense, though may allow a relative comparison between different parameter values.

Results

Simulated data reconstruction

Simulations of mixtures were first analyzed, to illustrate the approach and the relative separation of components, compared to published EM simulation data (Kullman, 1989; Ling and Tolhurst, 1983). The main distribution modeled was a mixture of 3 components with locations at either (0.0, 1.5, 3.0) or (0.0, 2.0, 4.0), and with relative probabilities of 0.25/0.50/0.25. The standard deviation was set at 1.0 and studies have been performed across sample sizes \( n = 300, 500 \) and 1000. Figure 1 illustrates a typical example of the degree of separation and the estimates of variability for these simulated data; the results were similar with the two critical separations of either 1.5 or 2.0 noise SD values and thus only the latter is shown (Redman, 1990).

Figure 1A shows the histogram, predictive probability density and ranges from a single simulation analysis \( (n = 500) \) in which the underlying discrete distribution has location \( (0.0, 2.0, 4.0) \). Figure 1B indicates the theoretical distribution by the filled circles - these are located at the component means and have heights equal to the component probabilities. The resulting predictive pdf [calculated with \( (a_0, r_0) = (0.01, 10) \)] is plotted as a continuous line in Figure 1A, with upper and lower ranges or bands (from the 5000 iterations of the Bayesian analysis) shown as dashed lines. The corresponding predictive pdfs of each of the individual components appear in Figure 1B, scaled by the estimates of their relative probabilities. The "lollipops" in this figure are topped by boxes centered horizontally at the posterior means of component location, and centered vertically at the posterior means of the relative probabilities; the horizontal and vertical widths of the boxes are set at \( \pm 1.0 \) posterior standard deviation for location (the vertical stem of the box) and probability respectively. In addition to an excellent fit to the data, apparent in Figure 1A \( (\chi^2 = 0.38) \), the components are well identified as is clear from Figure 1B. The posterior marginal probabilities \( P(k|a_0 = 0.01, r_0 = 10, D) \) indicated a strong preference for three underlying components: \( P(k = 3|D) = 0.95 \) and \( P(k > 3|D) = 0.05 \). Thus, for this one example, the number, location and probabilities of the components were accurately predicted for the simulated data set.

Figure 1C shows similar results for 15 repeat simulations \( (a_0 = 0.01, r_0 = 10, n = 500) \) with component locations also at \( (0.0, 2.0, 4.0) \). Because the analysis is completely unconstrained (neither the number of components nor the spacing between components is specified or limited) there exists considerable diffusion of the points around the underlying distribution (shown as filled circles). As with Figure 1A, the fidelity of the predictive distribution function to each simulated data set is excellent (and certainly dramatically so by the \( \chi^2 \) measure), but the dispersion of predicted component means and relative probabilities indicates the high degree of variability across sampled data sets that is always to be expected. The variability regions around the components clearly overlap the underlying distribution but also point out the uncertainty associated with a single, sampled data set. This point is also shown in the histogram of the sampled data set in Figure 1A, which demonstrates neither peaks nor inflections, yet is accurately reconstructed by the Bayesian analysis.

The simulated data sets were analyzed with a variety of \( a_0 \) and \( r_0 \) values, ranging from \( a_0 = 0.0001 \) to \( a_0 = 1.0 \) and \( r_0 = 100 \) to \( r_0 = 2 \). Values for \( a_0 \) smaller than 0.01 led to the analysis indicating insufficient components and values larger than 0.1 showed excess components. Variations
in the shape value \( r_0 \) led to changes in the degree of uncertainty around the posterior component locations, with \( r_0 \) in the range of 2–10 showing appropriate overlap with the underlying distribution (as in Figure 1C).

*Figure 1 about here*

**Physiological data set 1: C3P1**

The data set shown in Figures 2 and 3 (C3P1) is an ensemble obtained during intracellular recording from a CA1 pyramidal neuron in a hippocampal slice preparation (Turner, 1988; Turner and Deupree, 1991; Turner and Wheal, 1991). The excitatory postsynaptic potentials (EPSPs) were evoked by focal microstimulation of the stratum radiatum, which includes the afferents leading to the postsynaptic cell. The stimulation intensity was set at the threshold for the occurrence of a detectable postsynaptic potential, attempting to bias the data to contain only a few (5–6) components. The data was obtained at 2 Hz stimulation. Figure 2A shows the mean of the individual responses (\( n = 551 \)) and Figure 2B shows a compressed time series plot of these individual signal responses and also the interleaved noise samples. Figure 2A shows the region of the peak of the mean EPSP, from which values for the individual responses were analyzed.

The interleaved noise values averaged 0.009 ± 0.26 mV, which is representative for sharp electrode data in this preparation and the interleaved noise traces averaged very close to zero. The appearance of the plots in Figure 2B conveys a subjective impression of stationarity of the noise samples, without evidence for a trend in the mean value. Trends and stationarity can be more formally assessed using statistical time series methods, and Bayesian approaches using dynamic linear models (West and Harrison, 1989, first-order polynomial models of chapter 2) have been applied here to confirm the stability over time in the level of the noise series and the lack of significant autocorrelations. In addition, the noise values were tested for normality using both a \( \chi^2 \) test (\( \chi^2 = 10.8, df = 14, P = 0.63 \)) and a Kolmogorov-Smirnov analysis (K-S = 0.03, P = 0.56). Thus, the noise appeared to be stationary and reasonably well-characterized as a single gaussian function.

The signal values overall averaged 1.72±0.80 mV, exhibiting considerably increased fluctuation compared to that of the noise alone (F value = 9.51, P < 0.001). The waveform parameters of the trace (shown in Figure 2A) are rise time of 3.95 msec and half-width of 26.2 msec; the companion Time SD values are similar (not shown). The issues of stationarity may again be addressed via time series analysis, though now with caution. If the signal data arise from non-gaussian mixtures, then appropriate time series methods must be designed to either directly estimate the mixture structure while assessing trends (which raises a research agenda well beyond our current scope), or be insensitive to departures from normality of the underlying distribution. Simple first-order polynomial dynamic linear models (West and Harrison, 1989, chapter 2) again prove suitable if they are extended to automatically cater for apparent outliers in the data (outliers with respect to an assumed normal model) and other observed features of non-normality. Suitable techniques for model monitoring and adaptation appear in West and Harrison (1989, chapter 11), and are implemented in the software package BATS (West, Harrison and Pole, 1987; Pole, West and Harrison, 1992). As a result of analyses using these methods, the signal data in Figure 3B may be accepted as exhibiting insignificant changes in overall level and negligible autocorrelations.

The noise and signal data sets were analyzed using a prior \( a_0 \) value similar to that for the simulations (\( a_0 = 0.01 \)) but with the shape parameter \( r_0 \) at either 2 or 100 to represent very flat and very concentrated priors, respectively. However, the estimates of the first six means of the components are relatively stable across the two parameter values, with one component split into
two with $r_0 = 2$ as opposed to $r_0 = 100$ and the variability estimates slightly larger for $r_0 = 2$. For $r_0 = 100$ and $r_0 = 2$ the components are, respectively (mean±SD mV; 1) 0.58±0.01 versus the sum of 0.36 ± 0.07 and 0.90 ± 0.06; 2) 1.20 ± 0.03 versus 1.38 ± 0.08; 3) 1.82 ± 0.06 versus 1.92 ± 0.14; 4) 2.26 ± 0.06 versus 2.37 ± 0.23; 5) 2.71 ± 0.07 versus 2.98 ± 0.44; 6) 3.71 ± 0.18 versus 3.62 ± 0.32. Thus, the values for the component locations (and relative probabilities) are very close. The marginal probabilities are as follows: $P(k = 6|D) = 0.97$ with $r_0 = 100$; $P(k = 5|D) = 0.31$, $P(k = 6|D) = 0.60$ and $P(k = 7|D) = 0.09$ with $r_0 = 2$. The goodness of fit value are $\chi^2 = 0.63$ for $r_0 = 100$ and $\chi^2 = 0.62$ for $r_0 = 2$, indicating a similar good fit to the histogram for both values.

Figure 3A plots histograms of the noise and signal samples, with the predictive probability densities and ranges (shown as dashed lines), analyzed with $(a_0, r_0) = (0.01, 100)$. Note the close approximations of the histogram and ranges as well as the predictive pdf. Figure 3B shows the corresponding predictive pdfs for the noise and individual mixture components. The vertical bars are again located at estimated component means, with heights set at estimated component probabilities. The dashed boxes again indicate uncertainties, in terms of estimates ±1.0 estimated standard deviation, for locations and probabilities. The prior of $a_0 = 0.01$ suggests a concentration on 1 or 2 components (Table 1), and the large value $r_0 = 100$ indicates strong prior concentration on small values, but the posterior strongly favors at least 6 components for this data set. Thus, the data are overwhelming the prior, even though the prior is quite informative and precise.

Figures 2 and 3 about here

Further analyses were also performed with significantly lower degrees of freedom in the prior for the component noise variances. Thus, instead of df = 550 corresponding to the noise data sample number, reanalyses were explored using df = 50 thereby increasing initial uncertainty about the variances of the signal components. This permits a check on the issue of whether or not the use of the component variances are overly constrained by the use of the full noise degrees of freedom $(n - 1)$ in the prior. For this data set, the decrease in the degrees of freedom resulted generally in slightly larger estimates of the component variances and, as a consequence, small decreases in posterior probabilities on larger numbers of components. However, this change in the degrees of freedom for the noise variance did not lead to significant changes in inference about components.

Data set 2: paired-pulse analysis - changes in a parameter

Figure 4 shows a paired-pulse stimulation ensemble, in which an identical stimulus was repeated at an interval of 75 msec after the first. This second stimulus resulted in an increase in the second pulse on the average compared to the first, though there was marked variation in the difference between first and second pulses from trial to trial. The peak amplitude of the first EPSP averaged 1.48±0.51 mV and was significantly smaller than the second EPSP (1.79±0.48 mV, t value = 7.70, P < 0.001 for rejection of null hypothesis that the means are equal). The time course of the two responses was similar: rise time of 3.7 msec and half-width of 31 msec, with the Time SD exhibiting a similar time course of variation. The significant enhancement of the second pulse over the first indicates the presence of one form of short-term synaptic plasticity exhibited in neuronal responses.

Figure 4B shows a compressed time series plot for the first EPSP (and also representative for the second EPSP), indicating the apparent absence of trends and evident stability of the responses and noise. The interleaved noise samples showed means of −0.002±0.26 for the first response and −0.004±0.31 for the second. The EPSP signals exhibited less overall variation compared to the noise than the previous example, but remained significantly greater in terms of fluctuation (F values
Analysis of Postsynaptic Signal Mixtures

August 26, 1992

= 3.85 and 2.40, both P < 0.001). The \( \chi^2 \) value for these noise samples showed likely conformity to a single gaussian, with \( \chi^2 \) values of 7.1 (df = 13, P = 0.9) for the first response and 28.4 (df = 14, P = 0.12) for the second response. Thus the assumptions of approximate normality of the noise data and stationarity of signal and noise samples appeared reasonable by these criteria.

Figures 5A and 5B display the signal histograms and the predictive density functions with the ranges shown as dashed lines, with a prior value of \((a_0, r_0) = (0.01, 2)\). The \( \chi^2 \) values for these fits give \( \chi^2 = 1.32 \) for the first response and \( \chi^2 = 1.41 \) for the second response. Figures 5C and 5D display the predictive estimates of the mixture components, with essentially four, maybe five significant components predicted in the first EPSP \((P(k = 4|D) = 0.84 \text{ and } P(k = 5|D) = 0.16)\), and similar results in the second EPSP \((P(k = 4|D) = 0.86 \text{ and } P(k = 5|D) = 0.14)\). The locations of the components are stable for the first and second EPSP, respectively (mean±SD, mV): 1) 0.69±0.12, only in the first EPSP; 2) 1.16±0.14 versus 0.97±0.09; 3) 1.64±0.13 versus 1.61±0.10; 4) 2.20±0.20 versus 2.20±0.16; 5) 2.28±0.18 versus 2.25±0.10. The critical differences between these two responses appears to be a shift to higher amplitude components and decreased probability of the lower amplitude components, without a significant change in the component location. These components are also irregular, though the spacings correspond to roughly 0.5 mV. Thus, the Bayesian analysis appears able to identify the physiological changes induced by such a short-term plastic response (Manabe et al, 1992).

Figures 4 and 5 about here

The issue of extraneous components and the reduction of the analysis to the few critical components is of paramount importance in assessing mixture structure. What is happening is clear: the model (as with any mixture model allowing \( k \) to vary) prefers larger numbers of components since marginal improvements in fidelity to the data can be achieved via additional components with small probabilities interleaved between the more significant components. This is an over-fitting phenomenon that results from allowing the number of parameters in the model to increase, and the fact that component separations are unconstrained. Some form of control adjustments are required to partially (and stochastically) constrain the number of parameters and act against this over-fitting tendency, which is otherwise theoretically guaranteed (again, in any suitable model). The phenomenon can be illustrated by examining the integrated or marginal likelihood functions for \( a \) in any analysis; since the posterior for \( a \) is proportional to the prior multiplied by the likelihood function, then, in any analysis the likelihood function can be evaluated (approximately) as \( l(D|a) \propto p(a|D)/p(a) \) where \( p(a) \) is the prior (gamma) density and \( p(a|D) \) is the computed posterior density. This function provides the formal and absolute measure of support in the data for different values of \( a \), with the bias inherent in the prior effectively removed.

Figure 6 and Table 2 summarize several features from repeat analyses of the paired-pulse responses (Figures 4 and 5), with parameters \( r_0 = 2 \) and \( a_0 \) ranging between 0.000001 and 1.0. First, we show summary estimates of the number of components \( k \) (open circles and squares) plotted against the prior estimate \( a_0 \) used for each analysis, in Figure 6. In each analysis, the estimate chosen is just the posterior expectation \( E(k|D) \) under the posterior distribution \( P(k|D) \) from the analysis (the full distribution for the posterior \( P(k|D) \) is given in Table 2). Notice that the estimates are stable across the range \( a_0 < 0.1 \), but increase substantially for larger initial values of \( a \) as expected. Also plotted are the \( \chi^2 \) values for the fit of predictive distributions to the data in each case. Note the stability across \( a_0 \) values (apart from the very high values for the first pulse at \( a_0 < 0.001 \)), again pin-pointing the inability to distinguish fits based on this single summary measure.

Finally, the likelihood functions \( l(D|a) \) are given, one for each pulse data set (for \( a_0 = \)
0.01, from the analysis shown in Figure 5, plotted over the $a_0$ axis. Note that these do not (theoretically) vary with $a_0$, though minor variations exist across analyses due to the numerical approximations inherent in the Bayesian simulation. The over-fitting tendency is very clearly highlighted in the likelihood functions, which would predict approximately 16–17 components, though this large number of components leads to only marginally better $\chi^2$ values (0.66 for the first EPSP and 1.23 for the second EPSP). We stress again that increasing the number of components does not necessarily translate into improved goodness of fit by any method that measures “distance” between the fitted predictive density function and the raw data histogram; the $\chi^2$ measure is an example. In addition, low probability components are usually superfluous, particularly those added into the upper tail of the histogram, and often overlap with significant components (as noted in Figures 1B, 5C and 5D).

Table 2 and Figure 6 about here

In traditional methods of fitting mixtures, especially maximum likelihood, it has been routine practice for investigators to increase the number of components sequentially, fitting 1 then 2 components and so on, until the estimated means of additional components begin to coincide (modulo estimated uncertainty about the means) with existing components. Using small estimated probabilities of additional components as a stopping guide is also common. This clearly addresses the over-fitting issue, albeit in an ad-hoc fashion. Our analyses indicate the need for some such correction - note that, as further components are added, uncertainty about their location increases, implying increasing overlap in the posteriors. From our more formal Bayesian perspective, it is more natural to use the prior distributions for model parameters to effect control over the tendency to over-fit, particularly the prior for $a$. Taking a rather low prior estimate (say $a_0 \approx 0.01$) with a reasonable shape parameter (such as $r_0 = 10$) constrains the posterior for $a$ to much smaller values than a prior supported by the over-fitting, data-fidelity conscious likelihood function (such as $a_0 = 1$, as noted in both Table 1 and Figure 6).

Theoretical studies show that, for sample sizes in the 300–1000 range, priors $P(k|a, n)$ (as in Table 1) put decreasing probabilities over increasing values of $k$ (starting at $k = 1$) if $a \leq 0.2$; so taking $a = 0.01$ or thereabouts induces priors for $k$ that favor smaller values than are perhaps anticipated in the data ensemble but which are consistent with the need to effectively counter the over-fitting phenomenon. Additional theoretical and empirical investigations are needed to explore parameter combinations, our current benchmark or reference being to take $(a_0, r_0) = (0.01, 10)$. With this prior reference the information in the data appears to counteract the effect of the prior to induce small numbers for $k$, while also preventing the over-fitting of the data with large numbers of components.

Discussion

Model advantages

Our Bayesian analyses share the underlying mixture models with other analysis techniques, particularly the alternative method of maximum likelihood estimation (MLE) of the parameters of an unknown mixture distribution (if an unconstrained model is applied - Walmsley et al, 1988; Kullman, 1989). Unlike the MLE method the Bayesian analysis provides uncertainty estimates in a number of ways, including the conditional probabilities over numbers of components and estimates of variability around both the component means and the relative probability of each component. Additionally, and similar to the MLE technique, prior or estimated information can be incorporated
into the analysis, including approximate knowledge of the expected number of components and the noise variance. However, in the Bayesian analysis this prior information can be more or less important in terms of the posterior distribution, depending particularly on the parameter values. The posterior distributions reflect "shifts" from the prior expectations to more relevant parameter values in light of the data set under consideration. These shifts may be considered as the information within the data set leading to partial constraints on one or another parameter. One clear example is that the posterior distribution for the smoothing parameter $w$ (not illustrated in our examples, but see Escobar and West, 1992; West and Cao, 1992) shows a clear preference for values which are not necessarily indicated as plausible under the unbiased, reference prior for this parameter; this preference is determined by the data set rather than by the prior value.

Another considerable advantage of this approach is the calculation of the predictive probability density and cumulative density functions, for both the overall mixture and for each individual component. The iterative Bayesian simulation technique also permits estimation of the variability of these predictive functions, particularly through indications via range bands as illustrated. These predictive density functions likewise demonstrate excellent goodness of fits as assessed by conventional criteria (such as the $\chi^2$ test and K-S test).

The unconstrained model may be the most generalized in terms of attempting to pin down the activity of a variety of synaptic sites which may not be equal in terms of either amplitude or probability. Bayesian analysis using this model provides the location and uncertainty estimates of the probable location of the signal and allows for considerable variability between components. Most of the data sets analyzed (including many not shown) have demonstrated irregular component spacing, possibly pointing to the diffuse dendritic terminations of the synaptic afferents which have been stimulated. A logical sequence to include in the analysis would be an identification of the primary peaks from individual sites and also the necessary irregular sums which would be expected. Thus, many of the components noted to occur in the upper tail with the analysis may represent sums of smaller amplitude responses. On the other hand, the use of higher parameter values (such as $a_0 > 0.2$) leads to a large number of components with much smaller spacings, creating considerable ambiguity as to the "proper" number and spacing of components. However, the simulations point out a more appropriate range of initial parameter values ($a_0 \approx 0.01$) which effectively allow the data to determine directly the number of components and their locations. Clearly, though, the analysis remains constrained by the level of the ambient noise, and peak separation likely continues to be a function of the noise SD, with 1.5–2.0 noise SD as the limiting separation (Clamann et al, 1991).

**Future issues and applicability to data sets**

Many issues raised in the foregoing sections need further empirical and theoretical investigation. One or two other issues not explicitly considered above are as follows. To begin, there is the question of choice of sample size. Is it practical to perform the Bayesian analysis on as few as 100 points or are 300–500 the necessary minimum? For resolution of closely spaced peaks, sample sizes in the range of 750–1,000 points have been previously recommended. However, the length of the sample size is severely constrained by the period of stationarity in the cells being recorded at the time of the physiological experiment. Subgroup analysis (such as every other sample being considered, rather than the present linear sequence) may help to reveal trends as well as significant deviations from similarity of the components. However, stationarity is a severely constraining factor for large data sets, in which the sampling may stretch out over at least 15–20 minutes (Larkman et al, 1991; Malinow, 1991). Additional techniques to define degrees of stationarity may also be helpful, such as time series analysis, to analyze appropriate data sets. Likewise, a time series
approach in a general sense, using small subgroups and analyzing signal changes with respect to experience and manipulation, may be the most helpful direction, rather than attempting to isolate large regions which are approximately stationary. Such a time series approach would model ongoing synaptic function and plasticity in a phasic sense rather than attempt to freeze such a highly unstable process.

A mechanism to improve the number of data sets which can be analyzed would be to include the actual noise values (from a Bayesian analysis of the noise) rather than just the gaussian equivalent (Kullman, 1989). Various classes of noise and their application to the maximum likelihood estimation technique appeared to increase significantly the number of analyzable data sets (Kullman, 1989). This modification would include a full Bayesian description of the noise (including means, relative probability and conditional probability values) in addition to the parameter values used currently. This addition may aid considerably in allowing analysis of a larger number of data sets which are reasonable except in terms of non-gaussian noise elements present (which is the rule rather than the exception). Thus, there are a number of extensions to the present level of analysis which may improve considerably the confidence in parameter settings and adequacy of the predictive density functions, in light of the information contained within the entire data set. Physiological and anatomical confirmation of the location and effectiveness of synaptic connections, together with this novel form of analysis, may lead to significantly improved knowledge regarding synaptic function and neuronal processing of signals.

Statistical Appendix

Begin by taking the special case of constant variances so that \( f(x) = \sum_{j=0}^{k} p_j N(x|t_j, \nu) \). This is a special case of the general convolution formula

\[
 f(x) = \int N(x|z, \nu) dG(z) \tag{A1}
\]

which has the following interpretation: a random quantity \( x \) drawn from the distribution is conditionally distributed as \( N(x|z, \nu) \), and the mean \( z \) is initially uncertain and drawn from some mixing distribution \( G(z) \). The special case of (1) (with constant variances) simply corresponds to \( G(z) \) being a discrete distribution with probabilities \( p_0, \ldots, p_k \) on the values \( t_0, \ldots, t_k \). In real problems, \( G(z) \) is uncertain. Our class of models describes uncertainty about the entire distribution function \( G(z) \) using a Dirichlet process model. The key ingredients of a Dirichlet process are

(a) an initial or prior estimate \( G_0(z) \) of \( G(z) \), and

(b) a precision parameter \( a > 0 \),

which enter into the notation \( G \sim D(aG_0) \).

\( G_0(z) \) is a prior 'guess' at the form of the unknown distribution, and \( a \) determines how close this initial estimate is likely to be. Very large values of \( a \) imply that \( G(z) \) is expected to lie very close to \( G_0(z) \), vesting a great deal of faith in the initial estimate. To clearly illustrate the roles played by \( G_0(z) \) and \( a \) in determining mixture structure, suppose we observe a random sample of values \( z_1, \ldots, z_n \) from the distribution \( G(z) \), and ask what these values tell us about \( G(z) \). Part of the answer is that, at any chosen point \( z \), we should revise or update the initial estimate \( G_0(z) \) to the posterior estimate
\[ G_n(z) = a_n a G_0(z) + a_n \sum_{i=1}^{n} \delta(z - z_i), \quad (A2) \]

where \( a_n = 1/(a + n) \) and \( \delta(z) \) is the usual delta or indicator function \( \delta(z) = 1 \) if \( z = 0 \), \( \delta(z) = 0 \) otherwise. Formally, \( G_n(z) \) is the conditional expectation of the number \( G(z) \) under the (posterior) distribution of the uncertain function \( G \) given the values \( z_1, \ldots, z_n \). The precision parameter \( a \) determines how much these 'data' are 'weighted' relative to the initial estimate \( G_0(z) \). Note that

- If \( a \) is small relative to \( n \), \( G_n(z) \) is essentially just the empirical cumulative distribution function of the data, the natural non-parametric estimate of \( G(z) \). This is the form for very large \( n \).

- For larger values of \( a \), the weight \( a_n a \) is appreciable and \( G_n(z) \) modifies the empirical estimate by 'mixing' with the initial estimate \( G_0(z) \) in a common Bayesian manner.

- If further draws are taken from \( G(z) \), then \( G_n(z) \) is the distribution used to predict their values. In formal terms, based on the data values \( z_1, \ldots, z_n \), the estimate \( G_n(z) \) is the predictive distribution for further draws from the unknown distribution \( G(z) \). If \( G_0(z) \) is a continuous distribution, then \( G_n(z) \) is a mixture of discrete and continuous, resulting in new values drawn with chance \( a_n a \) and a repeat value drawn with chance \( 1 - a_n a \). The chance of a repeat value increases with \( n \).

- Consequently, in any set of \( n \) sampled values \( z_1, \ldots, z_n \) there is positive probability that there will be only a smaller number, say \( k \leq n \), of distinct values; if we knew the \( z_i \) we would know \( k \) and the actual distinct values, and also the precise arrangement, or configuration, of the \( z_i \) amongst these distinct values.

Given \( k \) and the configuration, label the distinct values \( t_1, \ldots, t_k \) and let \( n_j \) be the number of the \( z_i \) taking the value \( t_j \). Obviously \( n = n_1 + \ldots + n_k \) for any given configuration. As a result, \( (A2) \) becomes

\[ G_n(z) = a_n a G_0(z) + a_n \sum_{j=1}^{k} n_j \delta(z - t_j). \quad (A3) \]

Return now to the model \( (A1) \) for the actual data values \( z_1, z_2, \ldots \). Each \( z_i \) comes from a gaussian distribution with variance \( v \) some mean \( z_i \) which is itself drawn from \( G(z) \). Suppose for the moment that \( v \) is known and fixed. In real life, the means \( z_i \) are not themselves observed and so we need to be concerned with their estimation. Putting this aside for the moment to further explore the structure and implications of the model, combining equations \( (A1) \) and \( (A3) \) implies

\[ f(x) = a_n \int N(x|m, v) dG_0(z) + a_n \sum_{j=1}^{k} n_j N(x|t_j, v), \quad (A4). \]

If \( G_0(z) \) is taken as the gaussian distribution \( N(z|m, w v) \), for some numbers \( m \) and \( w \), then the convolution in the first term of \( (A4) \) is also gaussian for \( z_i \); writing \( p_0 = a_n \) and, for \( j = 1, \ldots, k \), \( p_j = a_n n_j \), we then see that \( (A4) \) is just a gaussian mixture of the form \( (1) \) (specialized to constant variances \( v_j = v \) for \( j = 1, \ldots, k \)) with \( t_0 = m \) and \( v_0 = v(1 + w) \). Practically important extensions to include differing variances across the mixture components, and some other
frills, alter details in (A4) but not the basic mixture structure. The analysis in the current paper assumes a gaussian $G_0(x)$. Note, however, that other choices are possible and might be considered in alternative analyses; an obvious generalization is to assume $G_0(x)$ itself is a discrete mixture of gaussians.

The generalization to provide for the possibility that the $v_j$ vary across $j$ is straightforward as described in Escobar and West (1992). Equation (A4) becomes

$$f(x) = a_n N(x|t_0, v_0)) + a_n \sum_{j=1}^{k} n_j N(x|t_j, v_j), \quad (A5)$$

where:

- the different variances $\nu_0, \nu_1, \ldots, \nu_k$ are conditionally independent and drawn from the inverse gamma prior distribution that is implied by assuming $s/v_j$ to be $\chi^2$ with $h$ degrees of freedom.
- conditional on the values of the $\nu_j$, the distinct locations $t_0, t_1, \ldots, t_k$ are conditionally independent with $t_j$ normally distributed about a mean $\mu$ and having variance $\nu_j$.  

We turn now to the Bayesian analysis of these models. Analysis is simulation based. For notation, write $u = \{k, \mu, \nu, a, t_0, \ldots, t_k, \nu_0, \ldots, \nu_k\}$ - the complete collection of uncertain parameters, including $\mu, \nu$ and $a$, and explicitly recognize the dependence of the data density (A4) on these parameters by rewriting $f(x)$ in (A4) as $f(x|u)$. Then the formal definition of the density for the future $x$ is

$$f(x|D) = \int f(x|u)f(u|D)dy$$

where $f(u|D)$ is the density of the posterior distribution for all the parameters $u$ based on the observed sample $D$. Computation of this posterior distribution lies at the heart of the statistical analysis; once it is computed and summarized, we use the summaries to make inferential statements about those parameters ... providing summary answers to the key questions of how many components, and what are the moments of the components, and so forth. Before the problem of posterior computation can be addressed, the model specification must be completed. In particular, we require prior distributions for the remaining parameters $\mu, \nu$ and $a$. As described in Escobar and West, priors for $\mu$ that are normal with variance proportional to $\nu$ are natural choices, as are inverse gamma priors for $\nu$. We may also assess priors for these parameters using plausible location and ranges of data values, anticipated spread amongst the locations of components, and numbers of components. Alternatively, and for $\mu$ and $\nu$ alone, a routine default prior distribution may be specified - the reference prior for the pair $\mu, \nu$ has the form $f(\mu, \nu) \propto \nu^{-1}$ (Lindley, 1965, volume II, p40). This improper prior is supposed to represent initial vagueness or indifference about the values of $\mu$ and $\nu$, hence the term reference. It is the standard reference prior for normal mean and variance parameters, and is used in the earlier analyses.

Central to this analysis is the precision parameter $a$ of the Dirichlet mixture model - a critical smoothing parameter for the model. Learning about $a$ from the data may be addressed by assuming an initial gamma prior distribution; Escobar and West (1992) give full details based on the development in West (1992). For sample sizes in the hundreds to a few thousand when it is anticipated (as is the norm) that there will be a relatively small number of mixture components, then suitable priors for $a$ place most mass on values around or less than unity, but will be diffuse enough so that the data may substantially impact inferences about $a$. This suggest gamma distributions
with shape parameters in the low integers and scale parameters around two or three times the shape. The problems of over-fitting associated with the fact that the number of parameters increases as \( k \) increases (Escobar and West, 1992) mitigate against this, however, and suggest priors that more heavily constrain \( a \) to smaller values - for sample sizes in the hundreds, to values near \( a = 0.01 \). Gamma priors with locations near 0.01 and reasonably large shape parameters, are thus implied.

Finally, the parameters \( h \) and \( s \) of the inverse gamma, or scaled \( \chi^2 \), prior for the variances \( \nu_j \) must be assigned. In the context of evoked potential responses, the noise ensemble provides an initial data set for preliminary estimation of the noise variance. Assuming the noise variance as a baseline estimate of each of the \( \nu_j \), we recommend setting the prior estimate \( s \) to the sample variance of the noise data, and choosing the prior degrees of freedom parameter to allow for some deviation of the \( \nu_j \) away from the noise value if the signal data are in substantial conflict with the single restricted value. Taking \( h \) as low as, say, 10–15 degrees of freedom will typically permit some such variation, though in the present instance the data analysis did not appear sensitive to such changes. Using very much larger values, such as the degrees of freedom associated with the noise data, may overly restrict the \( \nu_j \) to lie very close to \( s \) and therefore may run the risk of biasing the analysis. Bias will be in favor of larger numbers of components if, as may often be the case, the noise standard deviation actually underestimates some of the component variances. The preceding analyses assumed \( h = 1 \) and set \( s \) to the observed noise variance.

Proceeding to analysis, Escobar and West (1992) describe the technical and computational difficulties involved in the computation of the required posterior \( f(u|D) \), and present a feasible computational approach for its solution. The approach uses Monte Carlo simulation methods designed to produce simulated sample values of \( u \) drawn from the required posterior distribution \( f(u|D) \). This development represents application of what is known as Gibbs sampling - a class of simulation techniques that finds very wide use in modern statistical analysis (Gelfand and Smith, 1990) and which has promoted routine implementation of Bayesian methods in complex statistical models. Full details for the mixture model analysis appear in Escobar and West (1992). The basic idea is one of approximately simulating a collection of values of the parameter vector \( u \) from the posterior \( f(u|D) \) - with a specified Monte Carlo sample size \( r \), a collection of simulated parameters \( u_1, \ldots, u_r \) provides a collection of 'estimates' of the parameters on which inference about \( u \) may be based. For example, the first element of \( u \) is the unknown number of components \( k \), and the simulated values \( k_1, \ldots, k_r \) represent an approximate sample from the posterior distribution \( P(k|D) \) relevant for inference about \( k \). In each example in the Results section, the quoted posterior probabilities are just the relative frequencies of \( k \) values in this sample, with \( r = 5000 \) in each case. Similarly, the means, medians, quantiles and standard deviations of component locations \( \ell_i \) are just the sample versions based on the corresponding sampled values of the parameter \( \ell_i \). The graphs of approximate posterior distribution and density functions for the parameters, and predictive densities, are rather more efficient, smoothed estimates of the actual marginal posteriors. Refer to Escobar and West (1992) for full details of this, and of other technical aspects of the simulation analysis.

Acknowledgements

The research of Dennis A. Turner was supported by a Merit Review award from the Durham VAMC, a Burroughs-Wellcome Travel Award, and NINDS Grant RO1 NS29482-01. The research of Mike West was partially supported by the National Science Foundation under grant DMS-9024793. We thank Guoliang Cao for considerable programming effort and for help with the simulations.
References


Foster, T.C. and McNaughton, B.L. (1991) Long-term enhancement of CA1 synaptic transmission is due to increased quantal size, not quantal content, Hippocampus, 1:79-91.


Pole, A., West, M. and Harrison, P.J. (1992), Bayesian Analysis of Time Series, manuscript and computer software, ISDS, Duke University.


TABLE 1

PRIOR PROBABILITIES $P(k)$ AS A FUNCTION OF PARAMETER $a$ AND SAMPLE SIZE $n$

<table>
<thead>
<tr>
<th>Case $a = 0.001$:</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$ values</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 300$</td>
<td>0.99</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 500$</td>
<td>0.99</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.99</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case $a = 0.01$:</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$ values</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 300$</td>
<td>0.94</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 500$</td>
<td>0.94</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.93</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case $a = 0.1$:</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$ values</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$n = 300$</td>
<td>0.54</td>
<td>0.34</td>
<td>0.10</td>
<td>0.02</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>0.51</td>
<td>0.35</td>
<td>0.11</td>
<td>0.02</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.48</td>
<td>0.36</td>
<td>0.13</td>
<td>0.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case $a = 1.0$:</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$ values</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$n = 300$</td>
<td>0.02</td>
<td>0.06</td>
<td>0.12</td>
<td>0.17</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>0.01</td>
<td>0.05</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.01</td>
<td>0.03</td>
<td>0.06</td>
<td>0.11</td>
</tr>
</tbody>
</table>

(all probabilities quoted to two decimal places)
TABLE 2

POSTERIOR PROBABILITIES FOR $k$ IN PAIRED-PULSE DATA ANALYSIS ($n = 299$)

First pulse conditional probabilities $P(k|a_0, r_0 = 2, D)$

<table>
<thead>
<tr>
<th>$k$ values</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>≥ 13</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0 = 0.001$</td>
<td>0.88</td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_0 = 0.01$</td>
<td></td>
<td>0.85</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_0 = 0.1$</td>
<td></td>
<td></td>
<td>0.37</td>
<td>0.39</td>
<td>0.18</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_0 = 1.0$</td>
<td></td>
<td></td>
<td></td>
<td>0.02</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.09</td>
<td>0.10</td>
<td>0.11</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Second pulse conditional probabilities $P(k|a_0, r_0 = 2, D)$

<table>
<thead>
<tr>
<th>$k$ values</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>≥ 13</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0 = 0.001$</td>
<td></td>
<td></td>
<td></td>
<td>0.96</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_0 = 0.01$</td>
<td></td>
<td></td>
<td></td>
<td>0.86</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_0 = 0.1$</td>
<td></td>
<td></td>
<td></td>
<td>0.52</td>
<td>0.23</td>
<td>0.15</td>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_0 = 1.0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.03</td>
<td>0.05</td>
<td>0.08</td>
<td>0.11</td>
<td>0.12</td>
<td>0.12</td>
</tr>
</tbody>
</table>

(all probabilities quoted to two decimal places)
Figure 1. Figures 1A and 1B come from a single simulation analysis of \( n = 500 \) observations drawn from the mixture with locations \((0.0, 2.0, 4.0)\) and weights \((0.25, 0.50, 0.25)\). Analysis is based on \( a_0 = 0.01 \) and \( r_0 = 10 \). Figure 1A shows the data histogram and fitted/predictive pdf, with uncertainty bands, and 1B shows the predictive pdfs of components (scaled by their estimated probabilities). Figure 1B also indicates the distribution underlying the simulation by the filled circles - these are located at the component means with heights set at the component probabilities. Inference is summarized by dashed boxes located at the posterior estimates of component locations and probabilities; the horizontal and vertical widths of these boxes are \( \pm 1.0 \) estimated standard deviations for the locations (around the vertical bar) and probabilities, respectively. Figure 1C shows similar results from 15 simulations with component locations at \((0.0, 2.0, 4.0)\). Here the filled circles again represent the underlying component means and weights of the sampled distribution. Note the clustering around the expected distribution and overlapping of the dashed boxes, which are estimates of variability obtained from the Bayesian analysis.

Figure 2. This figure shows the mean EPSP response for data set C3PR1 \((n = 551)\) in 2A, and the signal and noise time series plot, against data sample number, in 2B. The small box around the peak shows the time window within which each of the peak EPSP values was calculated, for each response. The highly variable response is clearly shown in 2B, though there is no evident trend in either the noise or signal time series plot.

Figure 3. Figure 3A displays the raw histograms for the noise and signal data, together with the predictive pdfs and, for the signal analysis, uncertainty bands about the predictive pdf (dashed lines). This analysis is based on \((a_0, r_0) = (0.01, 100)\), though there was little difference with \( r_0 = 10 \) or 2. Figure 3B shows the predictive pdfs of individual components as estimated in this analysis. This frame also displays dashed boxes located at the posterior estimates of component locations and probabilities; the horizontal and vertical widths of these boxes are \( \pm 1.0 \) estimated standard deviations for the locations and probabilities, respectively.

Figure 4. Another example of a synaptic potential, with a paired-pulse shown (mean of \( n = 299 \) responses). The second EPSP at a 75 msec delay is significantly larger than the first, which is a common form of evoked short-term potentiation, called paired-pulse facilitation. The small boxes at the peak of the first and second EPSP again indicate the region from which the peak values were analyzed for each response. The noise and signal time series presented in 4B indicate apparent stationarity of the data series over the time period of the sample; these were similar for the first and second EPSP (second not shown).

Figure 5. Figures 5A and 5B show the data histograms and predictive pdfs for the first and second pulses, with uncertainty bands shown as dashed lines (calculated with \( a = 0.01 \) and \( r_0 = 2 \)). The predictive densities of the individual components for the first and second pulse are shown in 5C and 5D, respectively, with inferences about the components shown in terms of vertical bars and dashed boxes. Comparing the two pulse signal distributions, the peaks appear similar but in slightly different locations, indicating that the probabilities have likely changed towards larger values though the absolute amplitude (indicative of the location and number of sites) did not appear to change. Thus, the enhancement of the second pulse appeared due to an emphasis on larger components.

Figure 6. Figure 6 shows a comparison of \( \chi^2 \) values and posterior means for predicted numbers of components, together with the estimated likelihood functions over \( a_0 \), for the analysis of the paired pulse data \((r_0 = 2)\). The lowest values for \( a_0 \) show a decreased goodness of fit, according to the level of the calculated \( \chi^2 \) value. The likelihood function predicts a large number of components \((15-17)\). Values for the prior \( a_0 \) parameter in the range of 0.001 and 0.01 appear to provide the appropriate numbers of components for this data set, since the \( \chi^2 \) fit appears to point to a relatively constant value in this range.
Figure 3

A.

B.

EPSP Amplitude (mV)
Figure 4

A.

Voltage (mV)

Mean (n = 299)

Time (msec)

B.

EPSP Amplitude (mV)

Signal

Noise

Data Sample Number