TITLE: ROBUST, AUTOMATIC SPIKE SORTING USING MIXTURES OF MULTIVARIATE *t*-DISTRIBUTIONS

Shy Shoham, Ph.D.

Dept. of Molecular Biology, Princeton University

Matthew R. Fellows

Dept. of Neuroscience, Brown University

Richard A. Normann, Ph.D.

Dept. of Bioengineering, University of Utah

Keywords: spike sorting; multi-unit recording; electrode array; unsupervised classification; mixture models; expectation-maximization; multivariate t-distribution.

Correspondence to: Shy Shoham, Princeton University, Department of Molecular Biology, Washington Road, Princeton, NJ, 08544; E-mail: sshoham@princeton.edu; Phone: (609) 258-0374; Fax: (609) 258-1035

Abstract

A number of recent methods developed for automatic classification of multiunit neural activity rely on a gaussian model of the variability of individual waveforms and the statistical methods of gaussian mixture decomposition. Recent evidence has shown that the gaussian model does not accurately capture the multivariate statistics of the waveform samples' distribution. We present further data demonstrating non-gaussian statistics, and show that the multivariate *t*-distribution, a wide-tailed family of distributions, provides a significantly better fit to the true statistics. We then adapt a new Expectation-Maximization (EM) based competitive mixture decomposition algorithm and show that it can efficiently and reliably performs mixture decomposition of *t*-distributions. Our algorithm determines the number of units in multiunit neural recordings, even in the presence of significant noise contamination resulting from random threshold crossings and overlapping spikes.

Introduction

Extracellular recordings of neural activity provide a noisy measurement of action potentials produced by a number of neurons adjacent to the recording electrode. Automatic and semiautomatic approaches to the reconstruction of the underlying neural activity, or 'spike-sorting' were the subject of extensive development over the past 4 decades and reviews of early and recent efforts can be found in the literature (Schmidt 1984; Lewicki 1998). It is generally assumed that each neuron produces a distinct, reproducible shape, which is then contaminated by noise that is primarily additive. Identified sources for noise include: Johnson noise in the electrode and electronics, background activity of distant neurons (Fee et al. 1996), waveform misalignment (Lewicki 1994), electrode micromovement (Snider and Bonds 1998) and the variation of the action potential shape as a function of recent firing history (Fee et al. 1996; Quirk and Wilson 1999). Given this signal+noise structure, the problem of automatically classifying the different shapes is a clustering problem and can be addressed either in the context of the full time-sampled spike-shape or of a reduced feature set, such as the principal components or a wavelet basis (Hulata et al. 2002).

While the application of general clustering methods such as k-means (Salganicoff et al. 1988), fuzzy c-means (Zouridakis and Tam 2000), a variety of neural-network based unsupervised classification schemes (Ohberg et al. 1996; Garcia et al. 1998; Kim and Kim 2000) and ad-hoc procedures (Fee et al. 1996; Snider and Bonds 1998) have been pursued by some authors, a number of other studies (Lewicki 1994; Sahani et al. 1997; Lewicki 1998; Sahani 1999), attempting to provide statistically plausible, complete and efficient solutions to the waveform clustering problem have focused their attention on clustering based on a gaussian mixture model. The assumption underlying the latter approach is that after accounting for non-additive noise sources (e.g., misalignment, changes during neural bursts), the additive noise component is gaussian-distributed. As a

result, the waveforms resulting from each neuron are samples from a multidimensional gaussian distribution with a certain mean and covariance matrix. Given this statistical structure, it is possible to construct an appropriate statistical model of the data and apply the powerful method of gaussian mixture decomposition to solve the clustering problem (Jain et al. 2000; McLachlan and Peel 2000). This allows estimation of model parameters such as the shape of the individual waveforms and the noise characteristics; further, the number of clusters (or 'components') itself can be the subject of statistical selection, allowing objective determination of the unknown number of neurons being observed. The estimated model parameters are used to classify each 'spike' to one of several source neurons.

Although the statistical framework resulting from the multivariate gaussian model is extremely powerful and well studied, recent evidence suggests that it may provide an inaccurate description of the spike statistics (Harris et al. 2000). Examination of the distribution of Mahalanobis distances of spikes produced by a single unit reveals a discrepancy between the expected χ^2 distribution and the empirical distribution, which exhibits wider tails. Algorithms based on the gaussian assumption may therefore be ill suited for the task of automatic spike sorting, in particular as it is well known that they are not robust against a significant proportion of outliers. Moreover, the estimation of the number of clusters present is likely to be inaccurate.

In this study, we provide additional evidence for the non-gaussian nature of spike-shape statistics and demonstrate that an alternative model, one using *multivariate t-distributions* instead of multivariate gaussians is better suited to model the observed statistics. Multivariate *t*-distributions have attracted some recent attention in the applied statistics literature, and we show here that, combined with a new approximation, mixture decomposition algorithms for *t*-distributions based on the Expectation-Maximization (EM) algorithm have comparable computational complexity to the popular gaussian algorithm. Finally, we combine the new model with a recent powerful modification of

the EM algorithm that allows automatic determination of the number of sources through a process involving competitive elimination of components.

Experimental methods

The extracellular signals analyzed were recorded with a 100-microelectrode array (Jones et al. 1992) (Bionic Technologies, LLC, Salt Lake City, Utah). The array consists of a rectangular grid of silicon electrodes with platinized tips (200-500 $k\Omega$ impedances measured with a 1kHz, 100 nA sine wave). The array was chronically implanted in the arm region of a macaque monkey's (*M. mulatta*) primary motor cortex using surgical implantation procedures described elsewhere (Maynard et al. 2000), with the electrode tips approximately located in layers IV and V. A chronic connector system was used, allowing simultaneous access to signals from 48 electrodes. Recordings were obtained while the monkey was awake and performing a manual tracking task (Paninski et al. 2001 submitted). Signals were band-pass filtered (250-7500 Hz, 5^{th} order Butterworth), amplified (5000x), digitized (30 kHz sampling), and acquired to a Pentium-based PC using a 100-channel data acquisition system (Guillory and Normann 1999) (Bionic Technologies, LLC, Salt Lake City, Utah). Thresholds were manually set, at relatively low values, and threshold-crossing events were saved to disk. The events consisted of 48 time samples (1.6 ms), 10 of which preceded the threshold crossing. Of the 48 available electrodes, 14 provided single or multiunit activity. All of the subsequent data analysis procedures were performed using Matlab (Mathworks, Natick, MA.).

Figure 1 shows data collected from a well-isolated unit with signal-to-noise ratio of 16.9 (peak to peak/noise RMS), which was selected for much of the analysis below. Of the nearly 200,000 threshold-crossing events recorded in one behavioral session, 10,000 were selected. Random threshold-crossing events, which constituted nearly one half of the events, were very identifiable and manually removed using amplitude windows. This left approximately 5,300 events to be considered as unit waveforms. The absence of

detectable waveform overlaps in the raw events further suggests that this is a single unit. The unit displayed cosine modulation (Georgopoulos et al. 1982) with the instantaneous direction of arm motion (data not shown).

The waveform peak locations were estimated with subsample resolution by up-sampling the waveform at a 10 times finer resolution, and finding the new peak (Sahani 1999). All peaks were then aligned, and the waveforms interpolated at the original sampling resolution. Five points on the waveform edges were discarded to eliminate the need for extrapolation, leaving 43-sample waveforms. Our simulations indicate that this technique achieves an alignment accuracy of roughly 0.1 samples (standard deviation).

Statistics of spike-shape variability

In mixture modeling we assume that each sample $\mathbf{x}_i$ (in general, a vector) originates from one of g components. In spike sorting, $\mathbf{x}_i$ represents a sampled spike waveform or a vector of features, and the different components correspond to g different units. Assuming that each unit accounts for a proportion π_j of the total spikes, and that the distribution of spikes from unit j has parameters θ_j, the likelihood of the data (the probability of obtaining the given data set from this model) is (Lewicki 1998; McLachlan and Peel 2000):

$$p(\mathbf{x}_1 ... \mathbf{x}_N) = \prod_{i=1}^{N} p(\mathbf{x}_i) = \prod_{i=1}^{N} \sum_{j=1}^{g} \pi_j \, p(\mathbf{x}_i \mid \theta_j) \qquad (1)$$

The best-fitting model parameters $\{\pi_{1...g}, \theta_{1...g}\}$ are determined by maximizing the model likelihood, or its logarithm (the 'log-likelihood', *L*).

What is $p(\mathbf{x}_i \mid \theta_j)$, the distribution of spikes from unit j? The *p*-dimensional multivariate gaussian with mean $\boldsymbol{\mu}_j$ and covariance $\boldsymbol{\Sigma}_j$:

$$p(\mathbf{x}_i \mid \theta_j) = p(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{1}{(2\pi)^{p/2} \left| \boldsymbol{\Sigma}_j \right|^{1/2}} \exp(-\delta(\mathbf{x}_i, \boldsymbol{\mu}_j; \boldsymbol{\Sigma}_j)/2) \qquad (2)$$

has been used by a number of authors (Lewicki 1998; Sahani 1999) as a model, where $\delta(\mathbf{x}_i, \boldsymbol{\mu}_j; \boldsymbol{\Sigma}_j) = (\mathbf{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j)$ is the squared Mahalanobis distance between $\mathbf{x}_i$ and the template $\boldsymbol{\mu}_j$. The distribution of Mahalanobis distances of the different samples from the multivariate gaussian is expected to approximately follow the chi-square distribution with *p* degrees of freedom (only approximately, since we are dealing with sample mean and covariance). The left panel in Figure 2 illustrates that the empirical and χ^2 distributions have significant discrepancies, over the entire data range. These discrepancies are further illustrated in Figure 3a where the quantiles of the cumulative χ^2 distribution and the cumulative distribution of squared distances are compared. The two figures present complementary views of the overall disagreement and the large discrepancy of a few outlier data points. The solid line in Figure 3a presents the expected cumulative distribution of χ^2 with 43 degrees of freedom ($\chi^2(43)$), while the dashed line is the best fitting line plotted by Matlab on Quantile-Quantile (Q-Q) distribution plots of this type. The discrepancy between the best-fit line and the data are limited to the last few percent of data, while the disagreement with the expected $\chi^2(43)$ model is essentially everywhere.

Multivariate *t*-distributions (Lange et al. 1989; Peel and McLachlan 2000) represent an alternative to multivariate gaussians. Similar to gaussians, multivariate *t*-distributions are parameterized by a unique mean $\boldsymbol{\mu}_j$, and covariance matrix $\boldsymbol{\Sigma}_j$. In addition, they have a 'degrees of freedom' (DOF) parameter ν. Effectively, ν parameterizes the distribution's 'robustness', that is, how wide the tails are, or how many outliers are expected. The case $\nu \to \infty$ corresponds to a gaussian distribution and when $\nu=1$ we obtain the wide tailed multivariate Cauchy distribution (the expected covariance is infinite for $\nu \leq 2$). The *p*-dimensional *t*-distribution probability density function is:

$$p(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{p/2}\left|\boldsymbol{\Sigma}_j\right|^{1/2}} \times \frac{1}{\left[1+\frac{\delta(\mathbf{x}_i, \boldsymbol{\mu}_j; \boldsymbol{\Sigma}_j)}{\nu}\right]^{(\nu+p)/2}} \quad (3)$$

where Γ is the Gamma function. The distribution of squared Mahalanobis distances in the case of t-distributions can be evaluated analytically, and is equal to:

$$p(\delta(\mathbf{x}_i, \boldsymbol{\mu}_j; \boldsymbol{\Sigma}_j) \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \nu) = beta\left(\frac{1}{1+\delta(\mathbf{x}_i, \boldsymbol{\mu}_j; \boldsymbol{\Sigma}_j)/\nu}; 2+\nu/2, p/2\right) \quad (4)$$

where $beta(x;\alpha,\beta)$ is the beta probability density function with parameters α and β at point x.

Figures 2 (right panel) and 3b demonstrate the superior performance of the *t*-distributions as models of neural waveform variability. The model parameters were fit using a procedure that will be explained in the next section. In Figure 3b the expected distribution (solid line) and the best fit exactly overly each other. The *t*-distributions are, however, not a perfect fit. They clearly fail to explain a small proportion of points (0.1%-0.2%) with extremely large Mahalanobis distances. In a typical sample often used for spike sorting (2000-3000 waveforms) this proportion amounts to two to six spikes. In fact, removing the 6 outliers in our example had only a small effect on the optimal distribution parameters ($\nu = 51.9$ vs. $\nu = 46.7$).

The optimal DOF parameter for *t*-distributions becomes smaller (more non-gaussian) as we try to fit a projection onto a smaller subset of the leading principal components. Principal components analysis finds high-variance dimensions in the data, which appear to be less gaussian. This point is illustrated in Figures 2 (lower panel) and 4. As the data suggests that gaussians are a poor model of the waveform statistics while *t*-distributions may provide an adequate description, we continue by discussing and developing algorithms for fitting mixtures of *t*-distributions.

Clustering with mixtures of multivariate *t*-distributions

The most widely used method for estimating the parameters of mixture models is through an iterative procedure called the Expectation-Maximization (EM) algorithm (Dempster et al. 1977; Jain et al. 2000; McLachlan and Peel 2000). The EM algorithm for mixtures of gaussian distributions has been widely used for over three decades. Recently, an EM algorithm for estimating the parameters of mixtures of multivariate *t*-distributions was presented (Peel and McLachlan 2000). Without repeating the derivation here, we summarize the E and M steps of the solution below. The EM algorithm for this problem requires the introduction of two sets of auxiliary variables, twice their number in gaussian mixture-decomposition (in the gaussian case only the memberships are used):

z_{ij} - Membership of spike *i* to unit *j* ($0 \le z_{ij} \le 1$, 1 indicates unit j produced spike i).

u_{ij} - Weights indicating 'typicality' of spike *i* w.r.t. unit *j* ($u_{ij} << 1$ for outliers)

These variables are recalculated in the E step, and subsequently used to generate new estimates of the model parameters in the M step. The required calculations at step k of the algorithm are:

E step

Update the memberships and weights using:

$$\begin{cases} \hat{z}_{ij} = \dfrac{\pi_j P_{ij}}{\sum_{j=1}^{g} \pi_j P_{ij}} \\ \hat{u}_{ij} \equiv \dfrac{p+\nu}{\delta(\mathbf{x}_i, \boldsymbol{\mu}_j^{(k-1)}; \boldsymbol{\Sigma}_j^{(k-1)}) + \nu} \end{cases} \tag{5}$$

with $P_{ij} \equiv p(\mathbf{x}_i \mid \boldsymbol{\mu}_j^{(k-1)}, \boldsymbol{\Sigma}_j^{(k-1)}, \nu)$ as defined in (3)

M step

Update the proportions $\pi_{1\ldots g}$, component means, and common covariance using:

$$\begin{cases} \pi_j^{(k)} = \sum_{i=1}^{N} \hat{z}_{ij} \Big/ N \\ \boldsymbol{\mu}_j^{(k)} = \dfrac{\sum_{i=1}^{N} \hat{z}_{ij}\hat{u}_{ij}\mathbf{x}_i}{\sum_{i=1}^{N} \hat{z}_{ij}\hat{u}_{ij}} \\ \boldsymbol{\Sigma}_j^{(k)} = \dfrac{\sum_{i=1}^{N} \left(\hat{z}_{ij}\hat{u}_{ij}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_j^{(k)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_j^{(k)}\right)^T}{\sum_{i=1}^{N} \hat{z}_{ij}\hat{u}_{ij}} \end{cases} \tag{6}$$

Estimation of the DOF parameter ν, which allows tuning the shape of the distribution to the empirical spike distribution, involves solving the following implicit nonlinear equation (Peel and McLachlan 2000):

$$\sum_{i=1}^{N}\sum_{j=1}^{g} \hat{z}_{ij}\left[\psi\left(\frac{p+\nu}{2}\right) + \log\left(\frac{2}{\delta(\mathbf{x}_i, \boldsymbol{\mu}_j; \boldsymbol{\Sigma}_j) + \nu}\right) - \hat{u}_{ij} + \log\frac{\nu}{2} + 1 - \psi(\frac{\nu}{2})\right] = 0 \tag{7}$$

Where ψ is the digamma function. Solving this implicit equation typically involves a one-dimensional search, which adds significant computational overhead to the EM algorithm. Instead, we have found the following empirical approximation to provide a very accurate and fast approximation to the solution of (7) ($\left|\nu - \nu^*\right| < 0.03$ tested on simulated data with $5 < \nu < 50$) :

$$\nu^* = \frac{2}{y + \log y - 1} + 0.0416\left(1 + erf\left(0.6594 * \log\left(\frac{2.1971}{y + \log y - 1}\right)\right)\right) \tag{8}$$

Where y is an auxiliary variable defined by:

$$y \equiv -\sum_{i=1}^{N}\sum_{j=1}^{g} \hat{z}_{ij}\left[\psi\left(\frac{p+\nu}{2}\right) + \log\left(\frac{2}{\delta(\mathbf{x}_i, \boldsymbol{\mu}_j; \boldsymbol{\Sigma}_j) + \nu}\right) - \hat{u}_{ij}\right] \Big/ N \tag{9}$$

and *erf* is the error function.

Practical problems often encountered with EM-based mixture decomposition algorithms involve the initialization procedure, avoiding convergence to local likelihood maxima and parameter singularities and the determination of the number of components. In the next section, we present a new strategy that successfully avoids these problems.

Determining the number of units

Determination of the number of components in a mixture model has been the subject of extensive research (reviewed in (Sahani 1999; McLachlan and Peel 2000; Figueiredo and Jain 2002)). The methods most widely used for this task were based on selecting the best mixture models from a set of candidates with different numbers of components. After fitting the parameters of the different models (using the EM algorithm) the different models are compared using a penalized likelihood function, which penalizes the likelihood for 'complexity' (i.e., a larger number of components) and an "optimal" model is found. This class of methods has the disadvantage of requiring estimation of the parameters of multiple mixture models. Other approaches include the use of stochastic model estimation using model-switching Markov-Chain Monte-Carlo methods (Richardson and Green 1997), and deterministic annealing based approaches (Sahani 1999), which we have recently adapted to the case of the multivariate *t*-mixture model (Shoham 2002). These approaches suffer from significant computational complexity.

A recently introduced algorithm (Figueiredo and Jain 2002), appears currently to offer the best overall profile in terms of computational simplicity, efficiency and selection accuracy, and was adapted by us for the case of multivariate t-distributions. The algorithm attempts to maximize a penalized log-likelihood where the penalty is based on the Minimum Message Length criterion (Wallace and Freeman 1987):

$$L = \sum_{i=1}^{n} \log \sum_{j=1}^{g} \pi_j P_{ij} - \left[\frac{N}{2} \sum_{j} \log \frac{n\pi_j}{12} + \frac{g}{2} \log \frac{n}{12} + \frac{g(N+1)}{2} \right] \qquad (10)$$

Where N is the number of parameters per mixture component. Based on this cost function, an EM algorithm is derived, where different components compete for data points and are eliminated when they become singular. The algorithm is initialized with a large number of components, and subsequently eliminates components until convergence. The 'competition' is achieved using a new update of the components mixing proportions that replaces (6):

$$\pi_j^{(k)} = \frac{\max\left\{0, \sum_{i=1}^{N} \hat{z}_{ij} - N/2\right\}}{\sum_{j=1}^{g} \max\left\{0, \sum_{i=1}^{N} \hat{z}_{ij} - N/2\right\}} \tag{11}$$

The new algorithm differs in two additional ways from the usual EM implementation. First, the solution achieved by simply allowing the algorithm to converge may not be optimal, so after convergence to a solution the smallest component is eliminated, and a new cycle of iterations is started. Second, (11) has a very simple failure mode: if initialized with many small components, all, or many of them will be set to zero immediately, which will lead to an erroneous solution. In order to circumvent this problem (Figueiredo and Jain 2002) turns to a component-wise EM procedure (Celeux et al. 1999) that serves to re-normalize the component proportions after each step and therefore avoids full annihilation. We have found that this approach offers significant disadvantages when used with our model. In particular, fitting common parameters like ν becomes problematic. A simple alternative does present itself. Maximizing (10) with respect to π_j under the constraint $\sum_j \pi_j = 1$ leads after simple rearrangement to the following update rule:

$$\pi_j = \left[\sum_{i=1}^{n} \frac{\pi_j P_{ij}}{\sum_l \pi_l P_{il}} - \frac{N}{2}\right] \Big/ \left(n - \frac{gN}{2}\right) \tag{12}$$

Reapplying this expression in an iterative manner until convergence leads to a (possibly local) maximization of the penalized likelihood (10) with respect to the π_j. The term $\sum \pi_l P_{il}$ in the denominator assures that the proportions are re-normalized when a component is eliminated, and the expression is otherwise closely related to (11). This expression differs from (11) in that it maximizes the penalized likelihood (10) directly rather than maximize the 'complete data' likelihood as is typically done in the EM algorithm. Such maximization steps are a part of the ECME framework, an extension of the classical EM algorithm (Liu and Rubin 1994).

Simulation study

In order to avoid the inherent uncertainty in assessing the true number of units in extracellular recordings, we first tested the new algorithm on simulated random mixtures. We compared the clustering results of a randomly initialized EM algorithm (with the correct number of components) and of the new algorithm using 100 mixtures consisting of five components with different covariance matrices and proportions (1000 five-dimensional vectors in each mixture). The mixture components were *t*-distributed, and simulations were performed with three levels of "contamination", $\nu = \{3,5,20\}$. When comparing our results to the penalized-loglikelihood (10) of the underlying "true" distribution of points, we found that in all cases the new algorithm markedly outperformed the unmodified EM algorithm, which obtained incorrect and significantly less likely solutions in 40-50% of the trials (see Figure 5). The new algorithm correctly determined the number of components (5) in 90%-98% of the mixtures, and in over half the cases where it found an incorrect number (always either 4 or 6) the "wrong" answer corresponded to a higher penalized-loglikelihood than that of the underlying model used to generate the data. In all cases where the correct number of components was found, it either corresponded to the underlying model or had better penalized-loglikelihood. In fact in 5%-30% of the trials it obtained solutions with a much-higher penalized-loglikelihood

than that of the underlying model. The algorithm's performance therefore appears to be limited by the uncertainty inherent to the maximum-likelihood approach.

While performing this simulation study we found that the theoretical value of N (the number of parameters per component - $N = \frac{p(p+1)}{2} + p$ for an unconstrained mean and covariance) led to over-clustering, and we replaced it with an empirically obtained value (i.e., we consider it to be a user-assigned parameter). We continued this practice when applying the algorithm to real data.

Spike classification results

Results of applying our algorithm to real multi-unit motor data appear in Figures 6 and 7. The algorithm in both cases was initialized with 10 components, and rapidly converged to a result that appears to have the correct number of clusters as illustrated in Figure 6 (a). In both figures there are 'noise collection' clusters that are not a neural unit. These results were obtained using the full sampled waveforms, however the algorithm works well with a reduced feature set, such as the leading principal components. The results also illustrate that the performance is successful in spite of large noise contamination. The automatic tuning of the DOF parameter helps achieve this performance. The range of DOF in the solutions to these examples was 10-15, while isolated spike distributions have DOF parameters in the range 30-50. When using the projection on the first five principal components, DOF solutions obtained were in the range 3-8.

We summarize the new unsupervised spike classification scheme in Table 1.

Discussion

One of the most promising recent advances in basic and applied neuroscience research is the fabrication of arrays of electrodes that allow multiple site recording and stimulation in various neural systems (Jones et al. 1992; Hoogerwerf and Wise 1994; Rousche et al. 2001). Recording neural activity with such arrays can be used to address a

multitude of basic neuroscience questions, and has also been suggested as a brain-computer interface for use by paralyzed individuals (Maynard et al. 1997; Wessberg et al. 2000; Schwartz et al. 2001; Serruya et al. 2002). However, the traditional practice of optimizing SNR by micro-manipulating the electrode placement is no longer possible or practical when using these arrays. In practical terms this means that significant effort must be expended to signal detection and classification under "low" SNR scenarios (Kim and Kim 2000). This need motivated the present study, in particular as studies suggest that automatic methods potentially possess a significant accuracy advantage over manual spike sorting (Lewicki 1994; Harris et al. 2000), and are clearly more suitable for high electrode count arrays.

As model-based mixture decomposition clustering algorithms appear to currently offer the best prospects for the classification subunit in a fully automatic spike sorting routine (Lewicki 1998; Sahani 1999), we started out by testing the popular gaussian model, and replacing it with an improved, *t*-distribution model, at the cost of adding a single global parameter ν. An EM algorithm for mixture decomposition of *t*-distributions has recently been developed, and is well suited for this application, providing a robust alternative to the use of gaussian mixture algorithms by automatically down-weighing the effect of outlier waveforms (Peel and McLachlan 2000). Previous studies of *t*-distribution model fitting have focused on the difficult problem of fitting the degrees of freedom parameter, which we circumvent by adding a simple and accurate numerical approximation. We finally adapt a new EM based competitive agglomeration algorithm to this problem, which solves some of the difficulties associated with the use of the EM algorithm, and simplifies the determination of the number of units in the data. Unfortunately, at present this algorithm relies on an empirically determined penalty parameter, which weakens the advantage of using the superior statistical model. Our current implementation of the algorithm clusters 2000 five-dimensional waveforms in 10-

20 seconds on a Pentium-800 computer, and can therefore probably be implemented in a fully automatic multi-channel data acquisition system.

Our results regarding the statistics of waveform variability support those of a recent study (Harris et al. 2000) (figure 3A) where intracellular recordings were used to reliably identify the action potentials fired by individual neurons. Our results are, in fact, stronger in rejecting the gaussian model, possibly because Harris et al. (Harris et al. 2000) presented the best fitting line in their χ^2 distribution plot, rather than the distribution with the correct degrees of freedom (see Figure 3). Interestingly, previous studies concluded that the statistics were normal. The discrepancy can be explained by noting that the distribution plots presented in (Lewicki 1994) (Figure 2b) and (Fee et al. 1996) (Figure 1e) collapse together residuals at different time-delays. By eliminating the dependence structure between the different time-delays this operation reduces a multivariate distribution to a univariate one. In contrast, the distribution of Mahalanobis distances is a measure that is well suited for looking at the distribution of multivariate elliptical distributions. Moreover, close examination of the distribution plots appearing in these studies reveals larger-than-normal tails, and that the plotted gaussians were matched to the central region of the bell curve, rather than the standard variation. An additional study looked at the multivariate statistics of the background noise (Sahani 1999) (Figure 5.4), examining the marginal distributions along different principal directions, and demonstrated that the distribution exhibited extra kurtosis along the first few (i.e. most significant) principal directions.

The reason for the superior fit provided by the multivariate *t*-distributions is clearly the flexibility provided by the degrees of freedom parameter, and its wider tails. However, it may also be viewed as related to underlying characteristics of the background noise process. A previous study (Fee et al. 1996) provided compelling evidence that the neural background noise is highly nonstationary, and therefore the spike waveform distribution results from the mixed contributions of noise samples with

different characteristics. This "double randomness" is a characteristic of compound probability models of which the *t*-distribution is a member (Johnson et al. 1994). *t*-distributed variables can be generated as normally distributed with covariance matrix Σ / u where u is a random variable itself with a gamma distribution (Peel and McLachlan 2000). The nonstationarity of the background noise thus provides a potential reason why the noise statistics do not follow the normal distribution, in spite of the central limit theorem.

Alternatives and possible extensions

One solution to the problem of outlier waveforms (Banfield and Raftery 1993; Sahani 1999) is adding an additional large component whose influence encompasses the entire data set and serves as a 'garbage collector'. We found that adding the resulting component is highly sensitive to the definition of the 'data range'. Instead, in our implementation, following the clustering procedure we use heuristics to select those components thought to contain random threshold crossings and overlapping waveforms. Additional robust mixture-based clustering algorithms found in the literature are based on Huber's M-estimators (Huber 1982) like the hybrid of a gaussian distribution with laplacian tails called the ψ-function (Tadjudin and Landgrebe 2000) or Least Trimmed Squares estimators (Medasani and Krishnapuram 1998). It is quite possible that mixtures with nonelliptical mixture components (in contrast to multivariate gaussian or *t*-distributions) will improve the fit to the real statistics.

The problem of nonadditive waveform changes (resulting from rapid bursts or slow drift) and the detection problem (of single or overlapping spikes) remain as the two most significant hurdles to the general applicability of any unsupervised spike classification method. One approach to overcoming the problems resulting from changing templates is to abandon the mixture decomposition approach and develop ad-hoc clustering algorithms (Fee et al. 1996; Snider and Bonds 1998). Another approach is to

extend the basic mixture model to one in which the waveform shape depends on the recent spiking history, e.g., by the use of a sophisticated Hidden Markov Model (Sahani 1999). To address the problem of slow nonstationarity, one can replace the EM algorithm with an adaptive version that updates the model parameters recursively as the spikes are acquired (Neal and Hinton 1998; Sahani 1999). The quality of spike detection in the raw electrophysiological signal, significantly affects the amount and nature of contamination, and thereby the clustering results. One solution to this problem relies on using filters that are optimally matched to the distribution of templates and noise (Fee et al. 1996). Generic preemphasis techniques based on Teager's Energy operator (Kim and Kim 2000) and wavelet packets (Hulata et al. 2002) have been suggested to improve the detection quality. It remains to be evaluated how such algorithms perform in the nongaussian noise environment we have described.

Acknowledgments

We wish to thank Professors Sri Nagarajan and Mario Figueiredo for valuable input in preparing this manuscript. The work was supported by a State of Utah Center of Excellence contract #95-3365.

References

Banfield, J. D. and A. E. Raftery (1993). "Model-based Gaussian and non-Gaussian clustering." Biometrics **49**: 803-821.

Celeux, G., S. Chertien, et al. (1999). A component-wise EM algorithm for mixtures, INRIA, France.

Dempster, A. P., N. M. Laird, et al. (1977). "Maximum Likelihood from incomplete data using the EM algorithm (with discussion)." Journal of the Royal Statistical Society B **39**: 1-39.

Fee, M. S., P. P. Mitra, et al. (1996). "Automatic sorting of multiple unit neuronal signals in the presence of anisotropic and non-Gaussian variability." J Neurosci Methods **69**(2): 175-88.

Fee, M. S., P. P. Mitra, et al. (1996). "Variability of extracellular spike waveforms of cortical neurons." J Neurophysiol **76**(6): 3823-33.

Figueiredo, M. and A. Jain (2002). "Unsupervised learning of finite mixture models." IEEE Trans on PAMI **24**(3): 381-396.

Garcia, P., C. P. Suarez, et al. (1998). "Unsupervised classification of neural spikes with a hybrid multilayer artificial neural network." J Neurosci Methods **82**(1): 59-73.

Georgopoulos, A. P., J. F. Kalaska, et al. (1982). "On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex." J Neurosci **2**(11): 1527-37.

Guillory, K. S. and R. A. Normann (1999). "A 100-channel system for real time detection and storage of extracellular spike waveforms." J Neurosci Methods **91**(1-2): 21-9.

Harris, K. D., D. A. Henze, et al. (2000). "Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements." J Neurophysiol **84**(1): 401-14.

Hoogerwerf, A. C. and K. D. Wise (1994). "A three-dimensional microelectrode array for chronic neural recording." IEEE Trans Biomed Eng **41**(12): 1136-46.

Huber, P. J. (1982). Robust Statistics. New York, John Wiley and Sons.

Hulata, E., R. Segev, et al. (2002). "A method for spike sorting and detection based on wavelet packets and Shannon's mutual information." J Neurosci Methods **117**: 1-12.

Jain, A. K., R. P. W. Duin, et al. (2000). "Statistical Pattern Recognition: A Review." IEEE Trans. Pattern Anal. Mach. Intell. **22**(1): 4-37.

Johnson, N. L., S. Kotz, et al. (1994). Continuous Univariate Distributions. New York, John Wiley & Sons.

Jones, K. E., P. K. Campbell, et al. (1992). "A glass/silicon composite intracortical electrode array." Ann Biomed Eng **20**(4): 423-37.

Kim, K. H. and S. J. Kim (2000). "Neural spike sorting under nearly 0-dB signal-to-noise ratio using nonlinear energy operator and artificial neural-network classifier." IEEE Trans Biomed Eng **47**(10): 1406-11.

Lange, K. L., R. J. A. Little, et al. (1989). "Robust Statistical Modeling Using the t Distribution." Journal of the American Statistical Association **84**(408): 881-896.

Lewicki, M. S. (1994). "Bayesian modeling and classification of neural signals." Neural Computation **6**(5): 1005-1030.

Lewicki, M. S. (1998). "A review of methods for spike sorting: the detection and classification of neural action potentials." Network **9**(4): R53-78.

Liu, C. and D. B. Rubin (1994). "The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence." Biometrika **81**(4): 633-648.

Maynard, E. M., E. Fernandez, et al. (2000). "A technique to prevent dural adhesions to chronically implanted microelectrode arrays." J Neurosci Methods **97**(2): 93-101.

Maynard, E. M., C. T. Nordhausen, et al. (1997). "The Utah intracortical electrode array: a recording structure for potential brain-computer interfaces." Electroencephalogr Clin Neurophysiol **102**(3): 228-39.

McLachlan, G. J. and D. Peel (2000). Finite Mixture Models. New York, Wiley.

Medasani, S. and R. Krishnapuram (1998). Categorization of Image Databases for Efficient Retrieval Using Robust Mixture Decomposition. IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, IEEE.

Neal, R. M. and G. E. Hinton (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. Learning in Graphical Models. M. I. Jordan. Dordrecht, Kluwer Academic Publishers**:** 355-368.

Ohberg, F., H. Johansson, et al. (1996). "A neural network approach to real-time spike discrimination during simultaneous recording from several multi-unit nerve filaments." J Neurosci Methods **64**(2): 181-7.

Paninski, L., M. R. Fellows, et al. (2001 submitted). "Temporal tuning properties for hand position and velocity in motor cortical neurons." J Neurophysiol.

Peel, D. and G. J. McLachlan (2000). "Robust mixture modelling using the t distribution." Statistics and Computing **10**: 339-348.

Quirk, M. C. and M. A. Wilson (1999). "Interaction between spike waveform classification and temporal sequence detection." J Neurosci Methods **94**(1): 41-52.

Richardson, S. and P. Green (1997). "On Bayesian analysis of mixtures with unknown number of components." Jour. of the Royal Statistical Society B **59**: 731-792.

Rousche, P. J., D. S. Pellinen, et al. (2001). "Flexible polyimide-based intracortical electrode arrays with bioactive capability." IEEE Trans Biomed Eng **48**(3): 361-71.

Sahani, M. (1999). Latent Variable Models for Neural Data Analysis. Computation and Neural Systems. Pasadena, California, California Institute of Technology.

Sahani, M., J. S. Pezaris, et al. (1997). On the Separation of Signals from Neighboring Cells in Tetrode Recordings. Advances in Neural Information Processing Systems 11, Denver, CO.

Salganicoff, M., M. Sarna, et al. (1988). "Unsupervised waveform classification for multi-neuron recordings: a real-time, software-based system I Algorithms and implementation." J Neurosci Methods **25**(3): 181-7.

Schmidt, E. M. (1984). "Computer separation of multi-unit neuroelectric data: a review." J Neurosci Methods **12**(2): 95-111.

Schwartz, A. B., D. M. Taylor, et al. (2001). "Extraction algorithms for cortical control of arm prosthetics." Curr Opin Neurobiol **11**(6): 701-7.

Serruya, M. D., N. G. Hatsopoulos, et al. (2002). "Brain-machine interface: Instant neural control of a movement signal." Nature **416**(6877): 141-2.

Shoham, S. (2002). "Robust clustering by Deterministic Agglomeration EM of mixtures of multivariate t-distributions." Pattern Recognition **35**.

Snider, R. K. and A. B. Bonds (1998). "Classification of non-stationary neural signals." J Neurosci Methods **84**(1-2): 155-66.

Tadjudin, S. and D. A. Landgrebe (2000). "Robust parameter estimation for a mixture model." IEEE Trans. Geoscience, Remote Sens. **38**(1): 439-445.

Wallace, C. and P. Freeman (1987). "Estimation and inference via compact coding." Jour. of the Royal Statistical Society B **49**(3): 241-252.

Wessberg, J., C. R. Stambaugh, et al. (2000). "Real-time prediction of hand trajectory by ensembles of cortical neurons in primates." Nature **408**: 361-365.

Zouridakis, G. and D. C. Tam (2000). "Identification of reliable spike templates in multi-unit extracellular recordings using fuzzy clustering." Comput Methods Programs Biomed **61**(2): 91-8.

algorithm

Initialization: use simple clustering method (e.g. k-means or FCM) to determine centers $\boldsymbol{\mu}_{1...g\,\max}$ of $g_{max}>>g_{true}$ components. Set

$\pi_{1...g} = \frac{1}{g_{\max}}; \boldsymbol{\Sigma}_{1...g} = \mathbf{I}; \nu = 50; L_{\max} = -\infty;$ pre – determine N

While $g \geq g_{min}$

Repeat

E Step

Update memberships z_{ij} and weights u_{ij} (5)

M step

While $\left| \sum_{j=1}^{g} \pi_j - 1 \right| > 10^{-4}$

For j=1:g

Update π_j (12)

End For

g ← # of $\pi_j > 0$

End While

Purge components where $\pi_j = 0$

Update $\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$ (6)

Update ν (9),(8)

Update P_{ij} (3)

Update L (10)

Until Convergence ($\Delta L < 0.1$ & $\Delta \nu < 10^{-2}$)

If $L > L_{\max}$

$L_{\max}=L$; store parameters $\{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \nu\}$ as 'optimal';

Set smallest component to zero; g=g-1;

Else

Break

End if

End While

Table 1

Figure Legends

Figure 1. Representative statistics for a well-isolated motor cortical unit. (a) collection of ~5300 aligned waveforms. (b) Projection of waveforms from (a) onto their first two principal-components. (c) ISI histogram. Cell fired at an average rate of roughly 30Hz. (d) Histogram of collapsed residuals from (a) after the removal of the mean waveform. Inset shows right 'tail'.

Figure 2. Comparison of predicted and actual distributions of Mahalanobis squared distances. Plots show results for the same unit as in Figure 1, using both gaussian (left panels) and *t*-distribution (right panels) models. Both upper panels are the distributions using the full sampled waveforms (43 dimensions), and the lower panels are calculated using the first two principal components. The *t*-distributions used had $\nu = 46.7$ (upper) and $\nu = 7.4$ (lower). The predicted distributions (solid lines) are chi-square (gaussian), and a beta distribution (*t*).

Figure 3. Analysis of cluster shape using cumulative distribution plots (a) χ^2 cumulative distribution plot (43 degrees of freedom). $\chi^2(43)$ is the expected distribution of distances for normally distributed residuals. (b) Cumulative distance distribution using a beta distribution model. A beta distribution of the distances is expected for *t*-distributed residuals. Inserts in both plots are blow-up views of the central region.

Figure 4. Q-Q plots for the Mahalanobis squared distances for the gaussian (black) and t-distribution (light gray) models. The data used were from same unit as in Figure 1. A good model fit is indicated when the plot falls along the slope 1 line. Both axes of all plots are in squared distance units.

Figure 5. Failures of EM and new algorithm with simulated data. Plots display projections on first two principal components. Left panels: clustering results with regular EM (a) and new algorithm (b-c). Right panels: Underlying mixture. Ellipses indicate 2σ lines. Note that the while the EM failures are gross, erroneous solutions obtained by the new algorithm are nearly equivalent.

Figure 6. Clustering of multi-unit motor data I. (a) Snapshots of the algorithm progress illustrated in the space of the first two principal components (ellipses mark the 2σ lines). Top: initialization (10 components). Middle: Intermediate stage (four components). Bottom: final (three components representing two units + noise waveforms). (b) Top: Aligned raw data (3000 events). Middle: Classified waveforms. Bottom: ISI histogram for the two units and the noise cluster (shown in black). Insert shows the waveform templates.

Figure 7. Clustering of multiunit motor data II. (a) Raw data (3000 events), and its projection on the first two principal components. (b) Results of automatic clustering algorithm. Insert in right panel shows the learned templates. Ellipses on right mark the 2σ lines. Gray cluster consists of random threshold crossings and local field potential waveforms. Black cluster includes overlapping waveforms and noise waveforms.

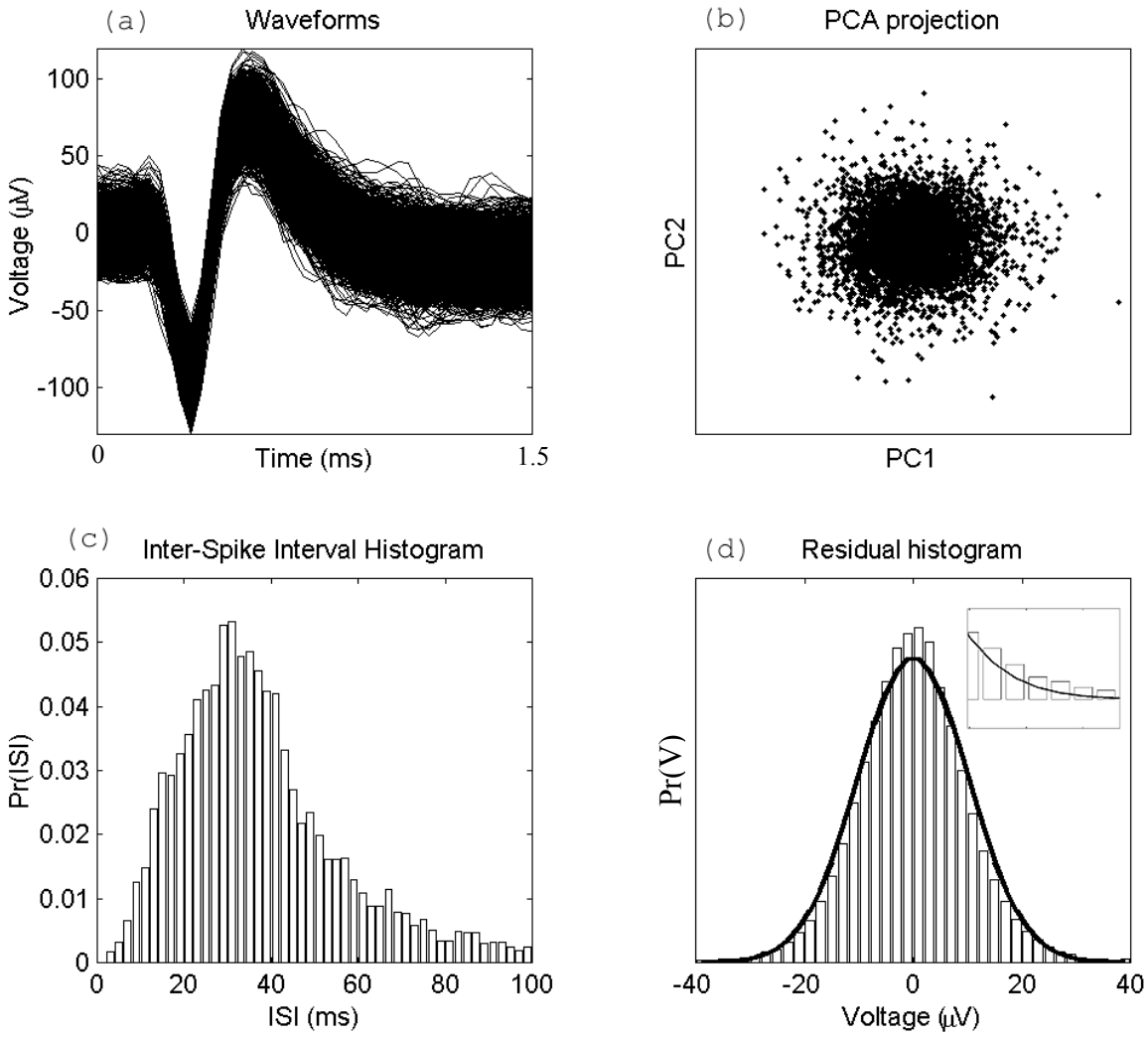
(a)
Waveforms
100
50
0
-50
-100
Voltage (μV)
0
Time (ms)
1.5
(b)
PCA projection
PC2
PC1
(c)
Inter-Spike Interval Histogram
0.06
0.05
0.04
0.03
0.02
0.01
0
Pr(ISI)
0
20
40
60
80
100
ISI (ms)
(d)
Residual histogram
Pr(V)
-40
-20
0
20
40
Voltage (μV)

Figure 1

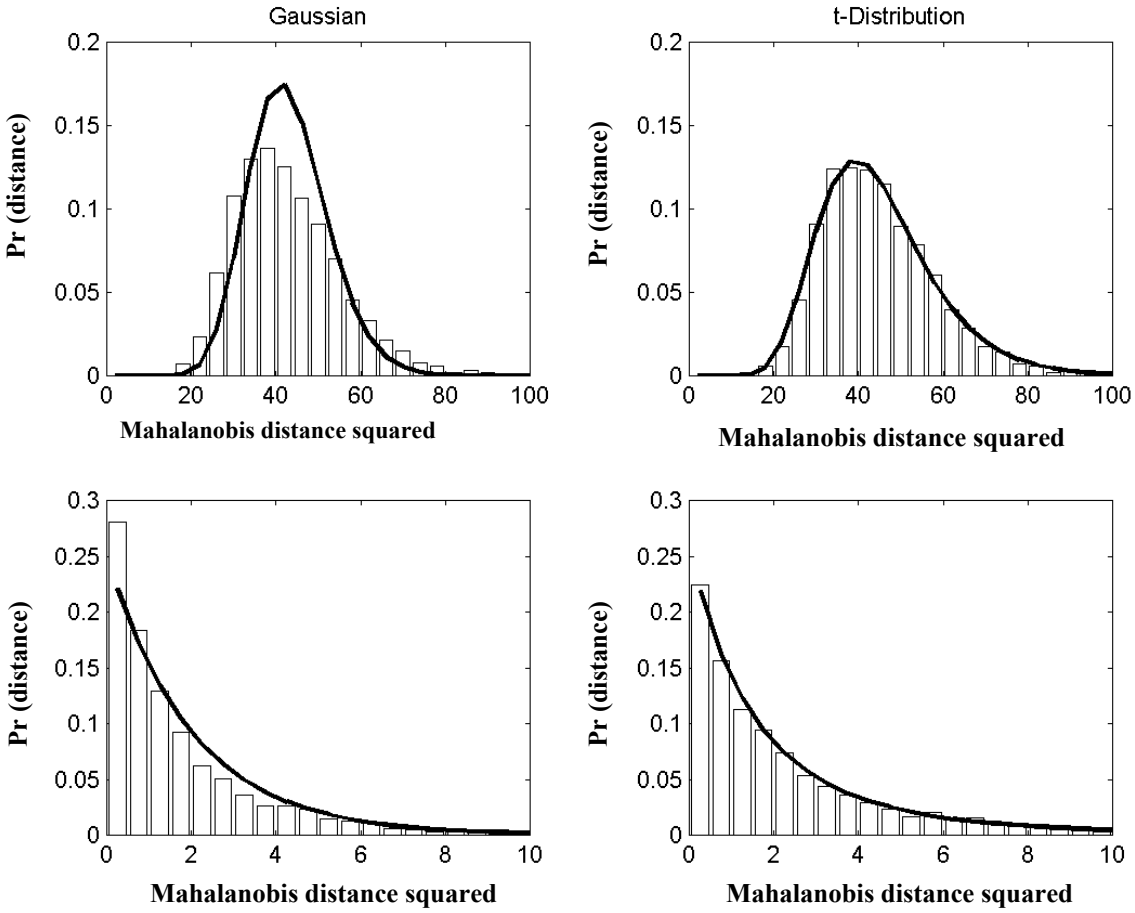
Gaussian
t-Distribution
Pr (distance)
Mahalanobis distance squared
0.2
0.15
0.1
0.05
0
20
40
60
80
100
0.3
0.25
2
4
6
8
10

Figure 2

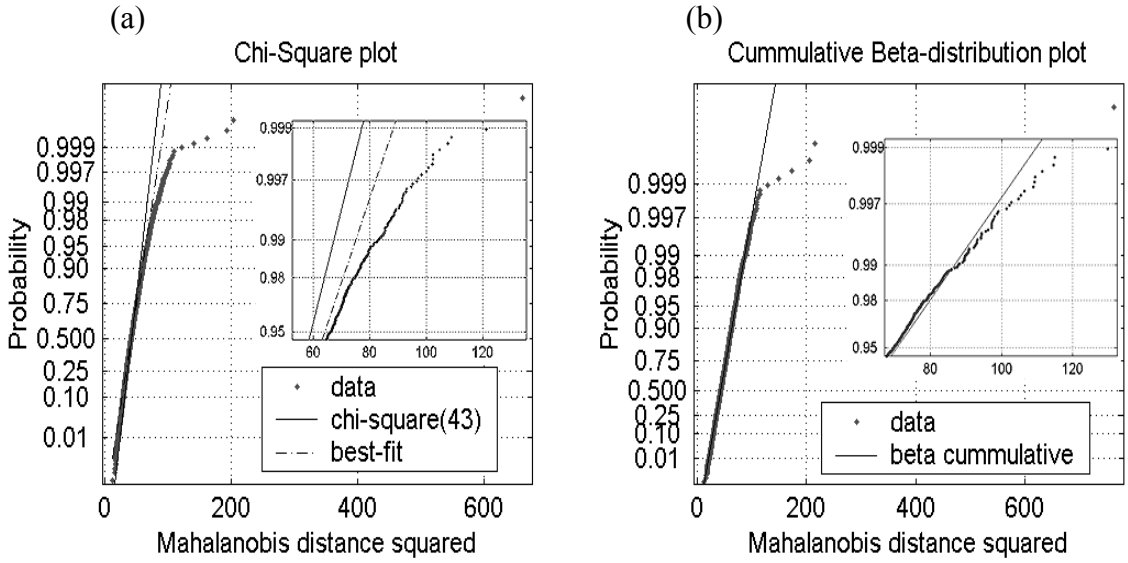
(a)
Chi-Square plot
Probability
0.999
0.997
0.99
0.98
0.95
0.90
0.75
0.500
0.25
0.10
0.01
0
200
400
600
Mahalanobis distance squared
data
chi-square(43)
best-fit
(b)
Cummulative Beta-distribution plot
Probability
0.999
0.997
0.99
0.98
0.95
0.90
0.75
0.500
0.25
0.10
0.01
0
200
400
600
Mahalanobis distance squared
data
beta cummulative

Figure 3.

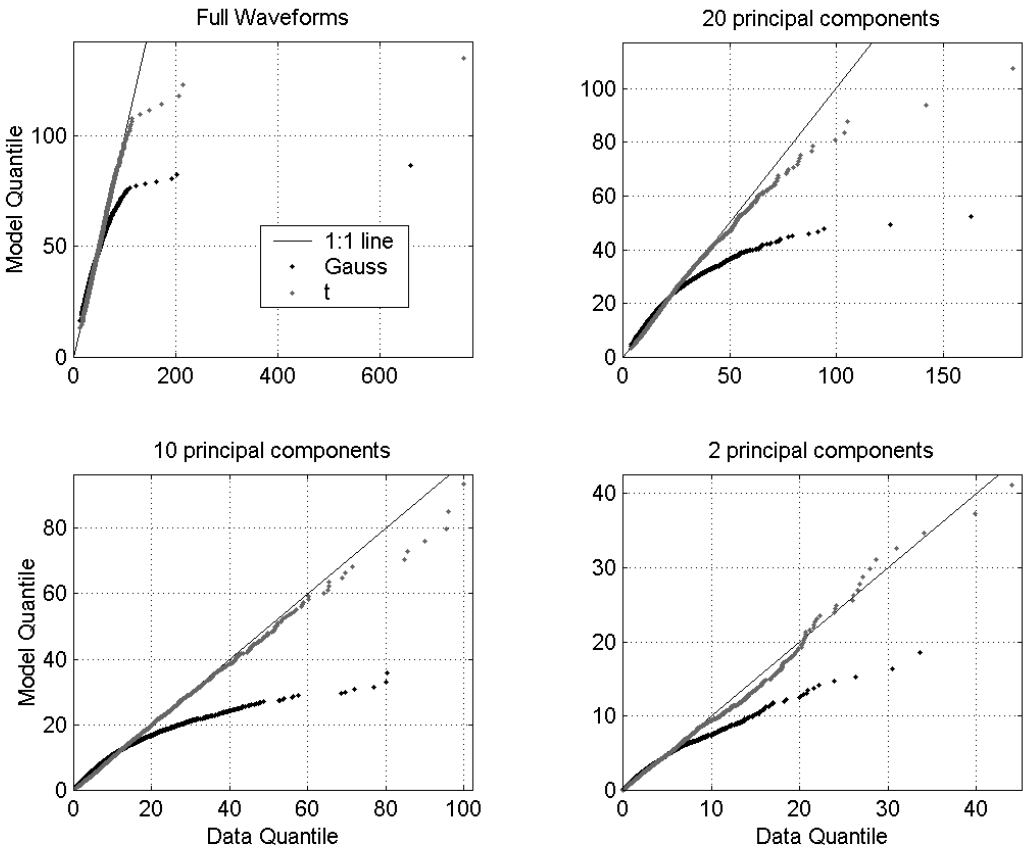
Full Waveforms
20 principal components
10 principal components
2 principal components
Model Quantile
Data Quantile
1:1 line
Gauss
t

Figure 4.

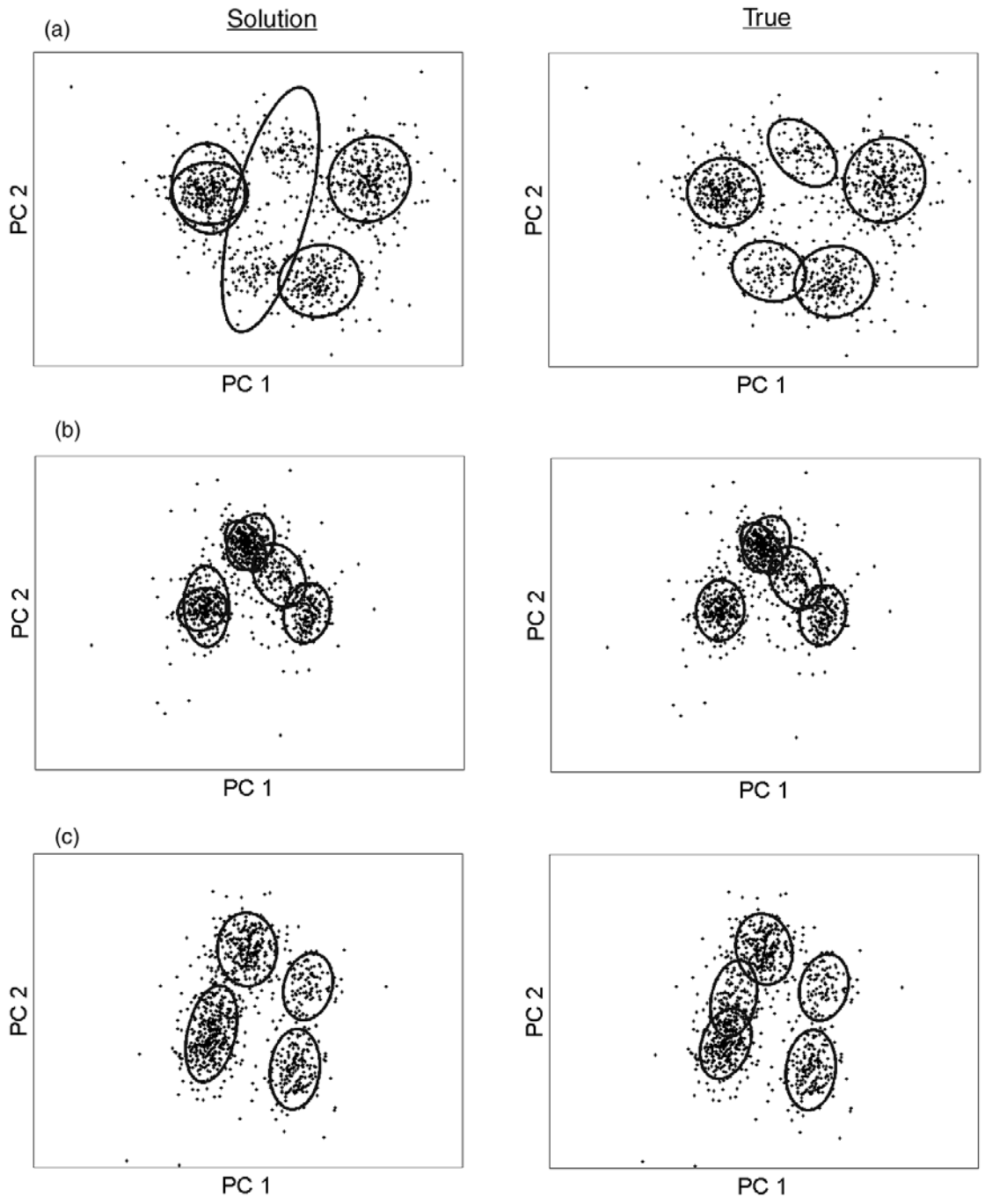

Figure 5

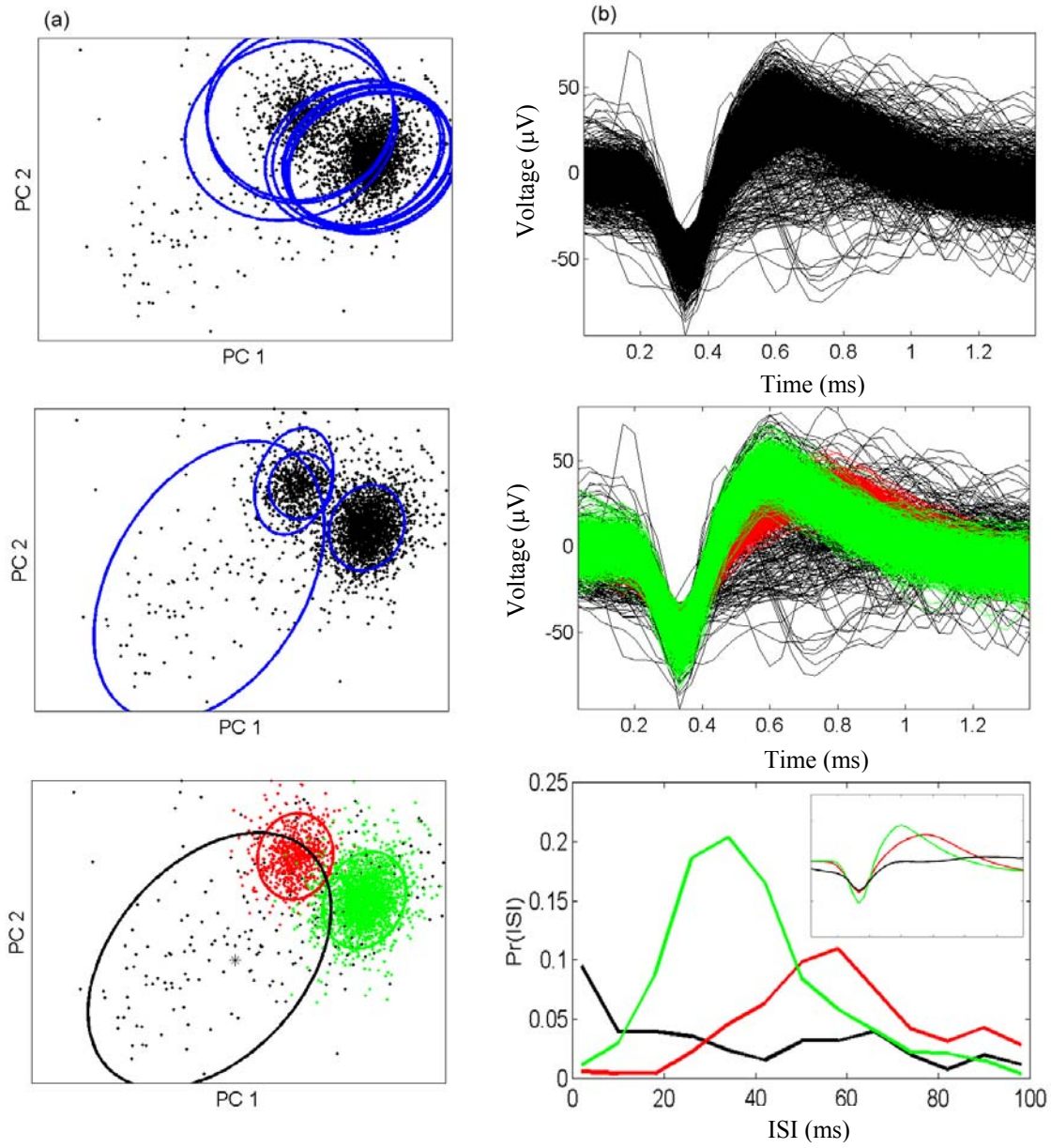
(a)
PC 2
PC 1
(b)
Voltage (μV)
50
0
-50
0.2
0.4
0.6
0.8
1
1.2
Time (ms)
PC 2
PC 1
Voltage (μV)
50
0
-50
0.2
0.4
0.6
0.8
1
1.2
Time (ms)
PC 2
PC 1
Pr(ISI)
0.25
0.2
0.15
0.1
0.05
0
0
20
40
60
80
100
ISI (ms)

Figure 6.

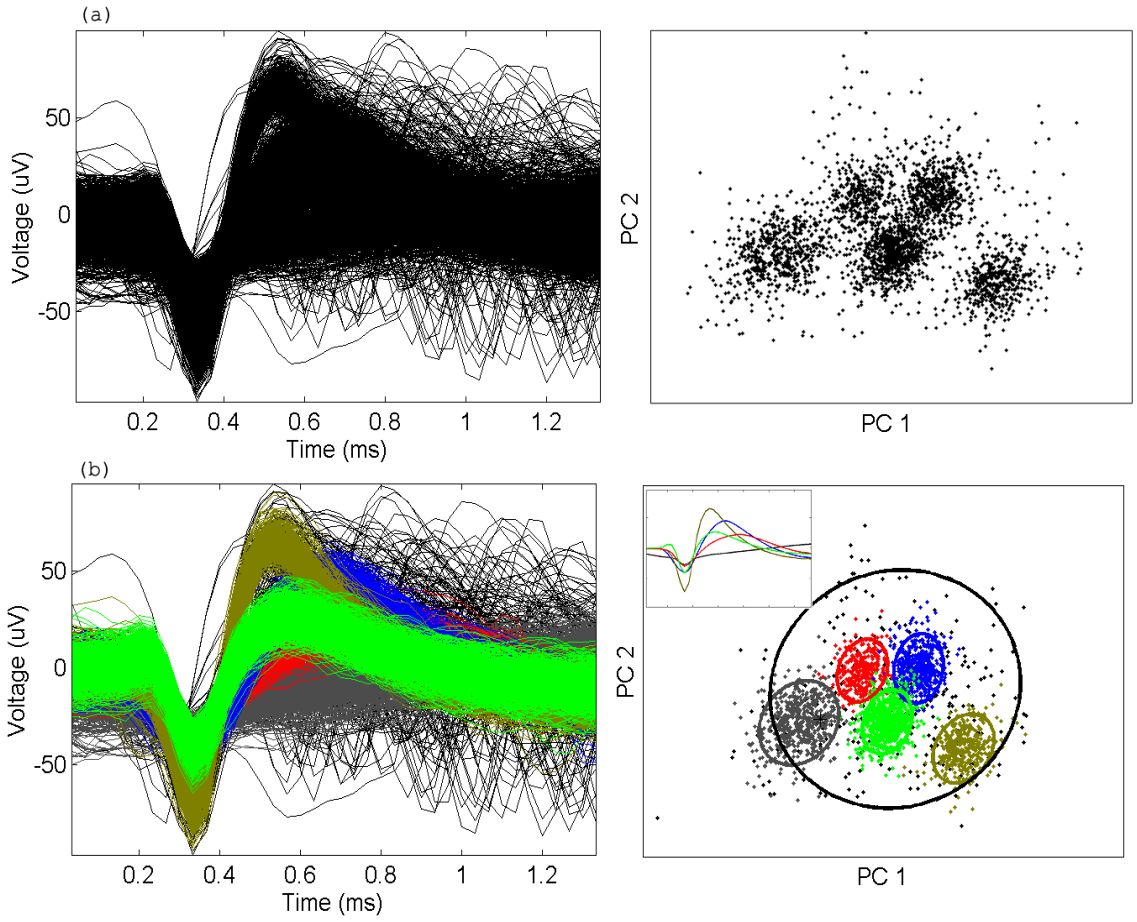
(a)
50
0
-50
Voltage (uV)
0.2
0.4
0.6
0.8
1
1.2
Time (ms)
PC 2
PC 1
(b)
50
0
-50
Voltage (uV)
0.2
0.4
0.6
0.8
1
1.2
Time (ms)
PC 2
PC 1

Figure 7.