ELSEVIER

Journal of Neuroscience Methods 127 (2003) 111–122

JOURNAL OF NEUROSCIENCE METHODS

www.elsevier.com/locate/jneumeth

Robust, automatic spike sorting using mixtures of multivariate *t*-distributions

Shy Shoham [a,*], Matthew R. Fellows [b], Richard A. Normann [a]

[a] *Department of Bioengineering, University of Utah, Salt Lake City, UT 84112, USA*
[b] *Department of Neuroscience, Brown University, Providence, RI 02912, USA*

Received 11 October 2002; received in revised form 8 April 2003; accepted 14 April 2003

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 introduce an adaptation of a new Expectation-Maximization based competitive mixture decomposition algorithm and show that it 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.

Keywords: Spike sorting; Multi-unit recording; Electrode array; Unsupervised classification; Mixture models; Expectation-Maximization; Multivariate *t*-distribution

1. 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' have been 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., 1996b), 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., 1996b; 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., 1996a; Snider and Bonds, 1998) have been pursued by some authors, a number of other studies (Lewicki, 1994, 1998; Sahani et al., 1997; 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

* Corresponding author. Department of Molecular Biology, Princeton University, Washington Road, Princeton, NJ 08544, USA. Tel.: +1-609-258-0374; fax: +1-609-258-1035.
E-mail address: sshoham@princeton.edu (S. Shoham).

0165-0270/03/$ - see front matter
doi:10.1016/S0165-0270(03)00120-1

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 multi-dimensional 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. The estimated model parameters are used to classify each 'spike' to one of several mixture components that correspond to different neural units (or possibly noise).

Although the statistical framework resulting from the multivariate Gaussian model is 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 squared 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. 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 (Lange et al., 1989), and a mixture decomposition algorithm for multivariate t-distributions was developed (Peel and McLachlan, 2000), based on the expectation-maximization (EM) algorithm. This algorithm requires computation of twice as many hidden variables as in Gaussian mixture decomposition algorithms, and involves an additional computational step for adapting the 'degrees of freedom' (DOF) parameter.

In addition to the choice of a statistical model for the mixture components, practical EM-based mixture decomposition algorithms need to address a number of issues including the determination of the number of components, the choice of an initialization procedure and avoiding convergence to local likelihood maxima or parameter singularities. Determination of the number of components in a mixture model has been the subject of extensive research (reviewed by Sahani (1999), McLachlan and Peel (2000) and Figueiredo and Jain (2002)). The methods most widely used for this task are 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, and, in addition, annealing approaches are quite sensitive to the specific choice of an annealing schedule. A recently introduced algorithm (Figueiredo and Jain, 2002), provides a new strategy where a process involving competitive elimination of mixture components drives a modified EM algorithm towards the optimal model size, simultaneously with the model parameter estimation. This approach appears currently to offer the best overall profile in terms of computational simplicity, efficiency and selection accuracy, and tends to avoid the usual difficulties of initialization sensitivity and convergence to singularities associated with the EM algorithm. We provide an adaptation of this algorithm for the case of multivariate t-distributed components. Our final algorithm is statistically plausible, simple and well-behaved and can effectively deal with many real data sets.

2. Theory: statistics of spike-shape variability

In mixture modeling we assume that each sample $\mathbf{x}_i$ (in general, a p-dimensional 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 n 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):

$$p(\mathbf{x}_1 \ldots \mathbf{x}_n) = \prod_{i=1}^{n} p(\mathbf{x}_i) = \prod_{i=1}^{n} \sum_{j=1}^{g} \pi_j p(\mathbf{x}_i|\theta_j) \quad (1)$$

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

What is $p(\mathbf{x}_i|\theta_j)$, the distribution of spikes from unit j? The p-dimensional multivariate Gaussian with parameters $\theta_j = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$:

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

has been used by a number of authors (Lewicki, 1998; Sahani, 1999) as a model. Here $\boldsymbol{\mu}_j$ is the mean, $\boldsymbol{\Sigma}_j$ is the covariance and $\delta(\mathbf{x}_i,\ \boldsymbol{\mu}_j;\ \boldsymbol{\Sigma}_j) = (\mathbf{x}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_j)$ is the Mahalanobis squared distance between $\mathbf{x}_i$ and the template $\boldsymbol{\mu}_j$. The distribution of Mahalanobis squared distances of the different samples from the multivariate Gaussian is expected to approximately follow the χ^2-distribution with p degrees of freedom (only approximately, since we are dealing with sample mean and covariance).

Multivariate t-distributions (Lange et al., 1989; Peel and McLachlan, 2000) represent a heavy-tailed elliptically symmetric 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 'DOF' parameter v, which is a positive scalar. Effectively, v parameterizes the distribution's 'robustness', i.e. how wide the tails are or how many outliers are expected relative to a Gaussian distribution with the same mean and covariance. The case $v \to \infty$ corresponds to a Gaussian distribution and when $v = 1$ we obtain the wide tailed multivariate Cauchy distribution (the expected covariance is infinite for $v \leq 2$). The p-dimensional t-distribution probability density function with parameters $\theta_j = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, v\}$ is:

$$p(\mathbf{x}_i|\theta_j) = \frac{\Gamma\left(\frac{v+p}{2}\right)}{\Gamma\left(\frac{v}{2}\right)(\pi v)^{p/2}|\boldsymbol{\Sigma}_j|^{1/2}} \frac{1}{\left(1 + \frac{\delta(\mathbf{x}_i,\ \boldsymbol{\mu}_j;\ \boldsymbol{\Sigma}_j)}{v}\right)^{(v+p)/2}} \qquad (3)$$

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

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

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

3. Algorithms: clustering with mixtures of multivariate t-distributions

The most widely used method for estimating the parameters of mixture models is through an iterative loglikelihood maximization procedure called the 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). As noted in the introduction, rather than apply the EM algorithm directly, we would like to apply it in conjunction with an efficient model selection scheme developed recently (Figueiredo and Jain, 2002). This approach maximizes a penalized log-likelihood with a penalty based on the minimum message length criterion (Wallace and Freeman, 1987):

$$L_p = \sum_{i=1}^{n}\log\sum_{j=1}^{g}\pi_j P_{ij} - \left[\frac{N}{2}\sum_{j=1}^{g}\log\frac{n\pi_j}{12} + \frac{g}{2}\log\frac{n}{12} + \frac{g(N+1)}{2}\right] \qquad (5)$$

Where N is the number of parameters per mixture component. This penalized-loglikelihood function leads to a different update of the mixing proportions in the M-step, which causes mixture components to compete for data points and be eliminated when they become singular. The algorithm is initialized with a large number of components, and subsequently eliminates components until convergence. This basic algorithm has a problematic failure mode: when it is initialized with many very small components they are all immediately eliminated. To circumvent this problem (Figueiredo and Jain, 2002) use the component-wise EM procedure (Celeux et al., 1999) to re-normalize the component proportions after each sub-step. We have found that this particular implementation offers significant disadvantages when used with the t-distribution model; in particular, fitting common parameters like the DOF parameter becomes problematic. Instead, we found that maximizing Eq. (5) directly with respect to π_j also provides the desired effect without the associated difficulty (see Appendix A).

The full algorithm (Table 1) consists of the EM algorithm for fitting mixtures of t-distributions (Peel and McLachlan, 2000), repeated here without derivation, together with a modified M-step for maximizing Eq. (5), derived in Appendix A. The algorithm uses two sets of auxiliary variables (in the Gaussian case only the memberships are used):

z_{ij}—membership of spike i to unit j ($0 \leq z_{ij} \leq 1$, 1 indicates unit j produced spike i).
u_{ij}—weights indicating 'typicality' of spike i with respect to unit j ($u_{ij} \ll 1$ 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:

3.1. E-step

Update the memberships and weights using:

$$\begin{cases} \hat{z}_{ij} = \dfrac{\pi_j P_{ij}}{\sum_{l=1}^{g} \pi_l P_{il}} \\ \hat{u}_{ij} \equiv \dfrac{p + v^{(k-1)}}{\delta(\mathbf{x}_i,\ \boldsymbol{\mu}_j^{(k-1)};\ \boldsymbol{\Sigma}_j^{(k-1)}) + v^{(k-1)}} \end{cases} \tag{6}$$

with $P_{ij} \equiv p(\mathbf{x}_i|\boldsymbol{\mu}_j^{(k-1)}, \boldsymbol{\Sigma}_j^{(k-1)}, v^{(k-1)})$ as defined in Eq. (3). Since the expectation of the Mahalanobis squared distances $\delta(\mathbf{x}_i, \boldsymbol{\mu}_j^{(k-1)}; \boldsymbol{\Sigma}_j^{(k-1)})$ is p, $\hat{u}_{ij} \approx 1$, except for outliers.

3.2. M-step

(1) Update the proportions $\pi_{1\ldots g}$ by iterating until convergence:

$$\pi_j^{(k)} = \frac{\max\left(\sum_{i=1}^{n} \dfrac{\pi_j P_{ij}}{\sum_{l=1}^{g} \pi_l P_{il}} - \dfrac{N}{2},\ 0\right)}{n - \dfrac{gN}{2}} \tag{7}$$

(2) Update the component means and covariances using:

$$\begin{cases} \mu_j^{(k)} = \dfrac{\sum_{i=1}^{n} \hat{z}_{ij}\hat{u}_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} \hat{z}_{ij}\hat{u}_{ij}} \\ \boldsymbol{\Sigma}_j^{(k)} = \dfrac{\sum_{i=1}^{n} (\hat{z}_{ij}\hat{u}_{ij})(\mathbf{x}_i - \boldsymbol{\mu}_j^{(k)})(\mathbf{x}_i - \boldsymbol{\mu}_j^{(k)})^T}{\sum_{i=1}^{n} \hat{z}_{ij}\hat{u}_{ij}} \end{cases} \tag{8}$$

(3) Estimate the DOF parameter v (tunes the tails of the distribution) by solving the following nonlinear equation (Peel and McLachlan 2000):

$$\begin{aligned} &\sum_{i=1}^{n} \\ &\quad \times \sum_{j=1}^{g} \hat{z}_{ij} \left[\psi\left(\frac{p + v^{(k-1)}}{2}\right) \right. \\ &\quad + \log\left(\frac{2}{\delta(\mathbf{x}_i,\ \boldsymbol{\mu}_j^{(k-1)};\ \boldsymbol{\Sigma}_j^{(k-1)}) + v^{(k-1)}}\right) - \hat{u}_{ij} \\ &\quad \left. + \log\frac{v^{(k)}}{2} + 1 - \psi\left(\frac{v^{(k)}}{2}\right)\right] \\ &= 0 \end{aligned} \tag{9}$$

Where ψ is the digamma function. Solving this equation typically involves a one-dimensional search, which adds significant computational overhead to the EM algorithm. Instead, we found empirically an approximation that provides a very accurate and fast approximate solution to Eq. (9) ($|v - v^*| < 0.03$ tested on simulated data with $5 < v < 50$):

$$\begin{aligned} v^{(k)} = &\frac{2}{y + \log y - 1} \\ &+ 0.0416\left(1 \right. \\ &\left. + erf\left(0.6594^* \log\left(\frac{2.1971}{y + \log y - 1}\right)\right)\right) \end{aligned} \tag{10}$$

Where y is an auxiliary variable defined by:

$$\begin{aligned} y \equiv &-\sum_{i=1}^{n} \\ &\times \sum_{j=1}^{g} \hat{z}_{ij} \left[\psi\left(\frac{p + v^{(k-1)}}{2}\right) \right. \\ &\left. + \log\left(\frac{2}{\delta(\mathbf{x}_i,\ \boldsymbol{\mu}_j^{(k-1)};\ \boldsymbol{\Sigma}_j^{(k-1)}) + v^{(k-1)}}\right) - \hat{u}_{ij}\right] \Big/ n \end{aligned} \tag{11}$$

and *erf* is the error function.

3.3. Clustering simulations

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 ($\pi = \{0.3, 0.3, 0.2, 0.1, 0.1\}$). There were 1000 five-dimensional vectors in each mixture, and the individual components had random means that were uniformly distributed in the range $\{-5, 5\}$ in each dimension, and diagonal covariance matrices with random elements uniformly distributed between 0.5 and 2. The data vectors were *t*-distributed, and simulations were performed with three levels of 'contamination', $v = \{3, 5, 20\}$. When comparing the penalized-loglikelihood (Eq. (5)) of the clustering results to those 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 Fig. 1). The new algorithm correctly determined the number of components (Eq. (5)) in 90–98% of the mixtures, and in over

Table 1
Algorithm

Initialization: use simple clustering method (e.g. k-means or FCM) to determine centers $\boldsymbol{\mu}_{1\ldots g_{\max}}$ of $g_{\max} \gg g_{\text{true}}$ components. Set $\pi_{1\ldots g} = 1/g_{\max}$; $\boldsymbol{\Sigma}_{1\ldots g} = \mathbf{I}$; $\nu = 50$; $L_{\max} = -\infty$; pre-determine N
While $g \geq g_{\min}$
  Repeat
    E Step
      Update memberships z_{ij} and weights u_{ij} (Eq. (6))
    M-step
      While $|\Sigma_{j=1}^{g}\pi_j - 1| > 10^{-4}$
        For $j = 1{:}g$
          Update π_j (Eq. (7))
        End For
        $g \leftarrow$ #of $\pi_j > 0$
      End While
      Purge components where $\pi_j = 0$
      Update $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$ (Eq. (8))
      Update ν (Eq. (11), Eq. (10))
    Update P_{ij} (Eq. (3))
    Update L (Eq. (5))
  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

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, either it corresponded to the underlying model or it 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-$N = p(p+1)/2+p$ 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.

4. 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 kΩ impedances measured with a 1 kHz, 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., submitted for publication). Signals were band-pass filtered (250–7500 Hz, fifth order Butterworth), amplified (5000×), digitized (30 kHz sampling), and acquired to a Pentium-based PC using a 100-channel data acquisition system (Guillory and Normann, 1999) (Bionic Technologies). 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).

5. Results

5.1. Spike waveform statistics

Fig. 2 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 easily identifiable and manually removed using amplitude windows. This left approximately 5300 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 point waveforms. Simulation tests indicate that this technique achieves an alignment accuracy of roughly 0.1 samples (standard deviation).

The left panel in Fig. 3 illustrates that the empirical and χ^2-distributions have significant discrepancies, over the entire data range. These discrepancies are further illustrated in Fig. 4a where the quantiles of the

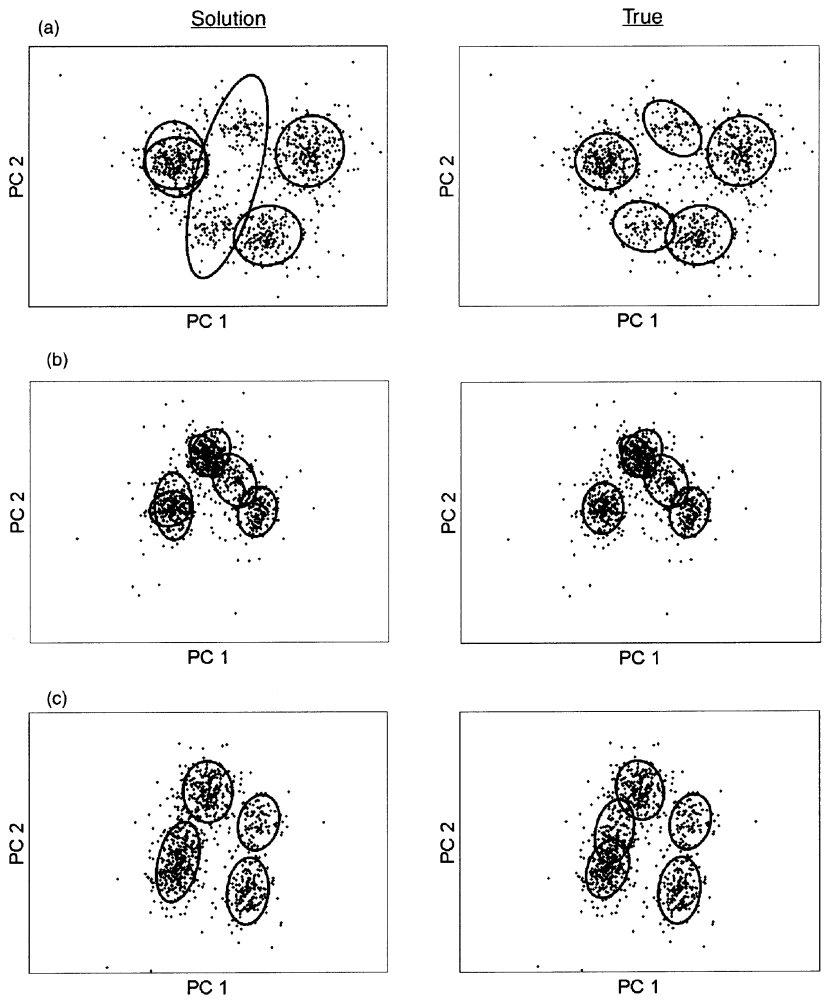

Fig. 1. Clustering simulations—failure modes of EM and new algorithm. Plots display projections on first two principal components. Left panels: clustering results with regular EM (a) and new algorithm (b–c). Right panels: Underlying mixture. Ellipses indicate 2σ lines. Note that while the EM failures are gross, erroneous solutions obtained by the new algorithm are nearly equivalent.

cumulative χ^2-distribution and the cumulative distribution of squared distances are compared. The two figures present complementary views of the overall disagreement. A few outlier data points with particularly large deviation are not shown in this figure. The solid line in Fig. 4a presents the expected cumulative distribution of χ^2 with 43 DOF (χ^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 (χ^2 (43)) model is essentially everywhere.

Fig. 3 (right panel) and Fig. 4b demonstrate the superior performance of the t-distributions as models of neural waveform variability. In Fig. 4b the expected distribution (solid line) and the best fit exactly overlie 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 squared distances. In a typical sample often used for spike sorting (2000–3000 waveforms) this proportion amounts to two to six spikes.

To obtain a quantitative measure of the goodness-of-fit of the two distributions, we calculated the Kolmogorov–Smirnov statistics using the Mahalanobis squared distances of the observed data, and simulated data generated from distributions with the best-fitting parameters (5000 waveforms generated in each case). The KS statistic was 0.11 ($p < 10^{-25}$, highly significant difference) for the multivariate Gaussian distribution and 0.013 ($p = 0.78$, insignificant difference) for the multivariate t-distribution. These numbers demonstrate the superior fit provided by the multivariate t-distribution.

The overall shape of the distribution, not merely the presence of a few outliers, is the source of the discrepancy with the Gaussian distribution. Removing the six outliers in our example had only a small effect on the optimal distribution parameters ($\nu = 51.9$ vs. $\nu =$

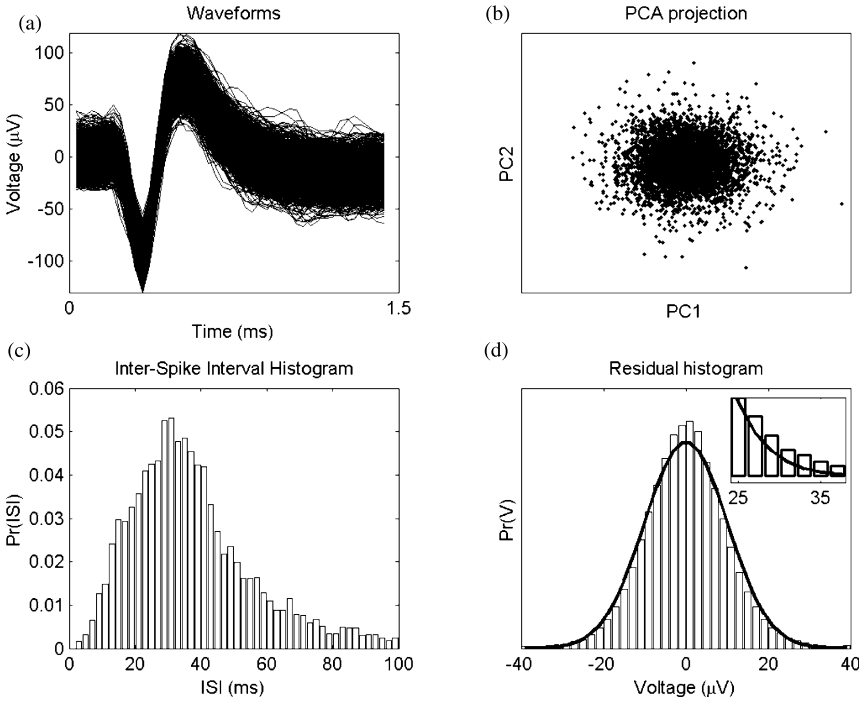

Fig. 2. Representative statistics for a well-isolated motor cortical unit. (a) Collection of ~5300 aligned waveforms. (b) Projection of waveforms from (a) onto their first two principal-components. (c) ISI histogram. Cell fired at an average rate of roughly 30 Hz. (d) Histogram of collapsed residuals from (a) after the removal of the mean waveform. Inset shows right 'tail'.

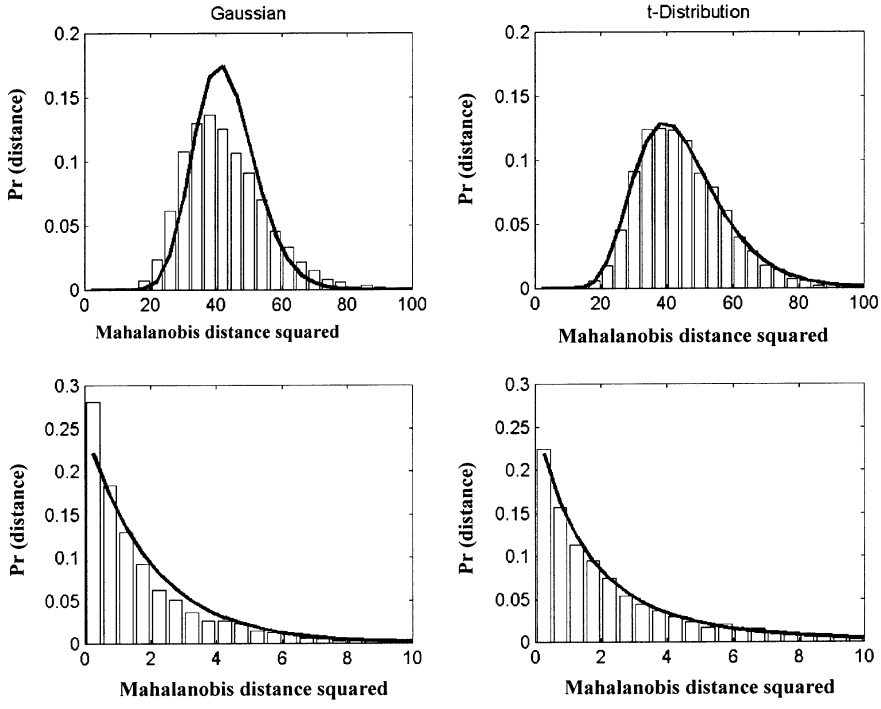

Fig. 3. Predicted and actual distributions of Mahalanobis squared distances. Plots show results for the same unit as in Fig. 2, using both Gaussian (left panels) and t-distribution (right panels) models. Both upper panels are the distributions using the full sampled waveforms (43 dimensions), and the 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 χ^2 (Gaussian), and a beta distribution (t).

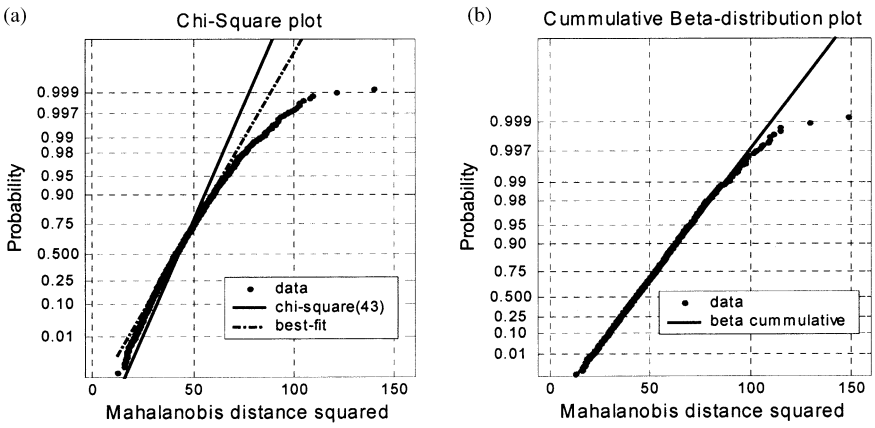

Fig. 4. Analysis of cluster shape using quantile–quantile plots (a) χ^2 cumulative distribution plot (43 degrees of freedom). χ^2 (43) is the expected distribution of distances for normally distributed residuals. Both the expected and the best fitting line are shown. (b) Cumulative distance distribution using a beta distribution model. A beta distribution of the distances is expected for t-distributed residuals. The four most distant points (Mahalanobis Squared distance >150) were excluded for display purposes.

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 (Fig. 3 (lower panel) and Fig. 5). The best fitting model for the waveform projections on the first 10 PCs has $v = 11.9$ and on the first two PCs has $v = 7.8$. The first 10 PCs capture $\sim 92\%$ of the entire ensemble variance. A consistent picture emerges when fitting the projections individually dimension by dimension (Fig. 5d). The most significant dimensions are best fit with a t-distribution with DOF 7–15.

5.2. Clustering

Results of applying our algorithm to real multi-unit motor data appear in Figs. 6 and 7. Waveforms were realigned (as above), but not subjected to any additional preprocessing. 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 Fig. 6a. In both figures there are 'noise collection' clusters that are not a neural unit, but rather capture outlier waveforms produced by noise or overlapping waveforms. 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.

6. 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). Neural activity recorded 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 (Shoham, 2001; Donoghue, 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 in signal detection and classification under 'low' SNR scenarios (Kim and Kim, 2000). This need motivated the present study, in particular because 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 mixture model-based 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 v. Using a t-distribution provides a robust alternative to the use of Gaussian mixture models, automatically down-weighing the effect of outlier waveforms. Our parameter estimation relies on a new algorithm that combines a recent EM algorithm for mixture decomposition of t-distributions (Peel and McLachlan, 2000), a new EM-based competitive agglomeration algorithm (Figueiredo and Jain, 2002), and

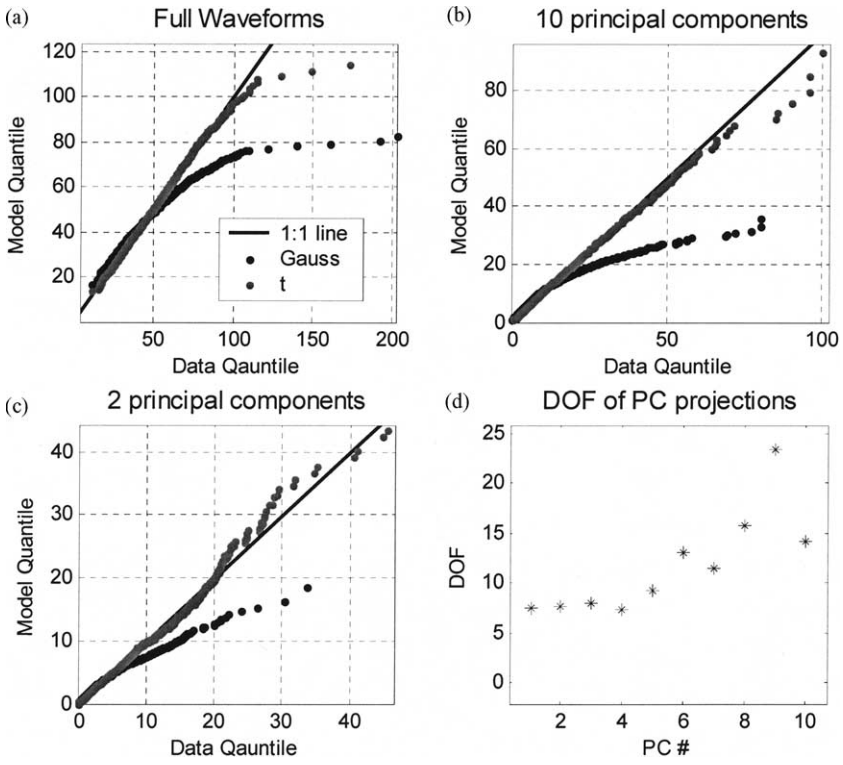

Fig. 5. Q–Q plots for the Mahalanobis squared distances for the Gaussian (black) and t-distribution (light gray) models (a)–(c). The data used were from same unit as in Fig. 1. A good model fit is indicated when the plot falls along the slope 1 line. Both axes of all plots are in squared distance units. (d) DOF obtained from fitting the data projection onto individual PCs.

a simple approximation for determining ν. Unfortunately, at present this algorithm relies on an empirically determined penalty parameter, which weakens the advantage of using the superior statistical model. A MATLAB implementation of the presented algorithm (available online: http://www.bionictech.com/support.html) is currently used for off-line sorting of electrode-array data by a number of laboratories, mainly in conjunction with Bionic Technologies electrode arrays and data acquisition systems (CyberKinetics Inc., Providence, RI). The current algorithm typically clusters a sample of 2000 five-dimensional waveforms in 5–6 s on a Pentium 2.4 GHz computer, and can therefore potentially be implemented as part of 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) (Fig. 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. (2000) presented the best fitting line in their χ^2-distribution plot, rather than the distribution with the correct DOF (see Fig. 4). Two earlier studies of waveform variability (Lewicki (1994) (Fig. 2b) and Fee et al. (1996b) (Fig. 1e)) used a different data analysis approach, collapsing together the residuals from different time-delays thereby reducing a multivariate distribution to a univariate one (in contrast, the distribution of Mahalanobis squared distances is a measure that is well suited for looking at the distribution of multivariate elliptical distributions). Close examination of the distribution plots appearing in these studies reveals larger-than-normal tails (in fact, 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 DOF 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., 1996b) 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

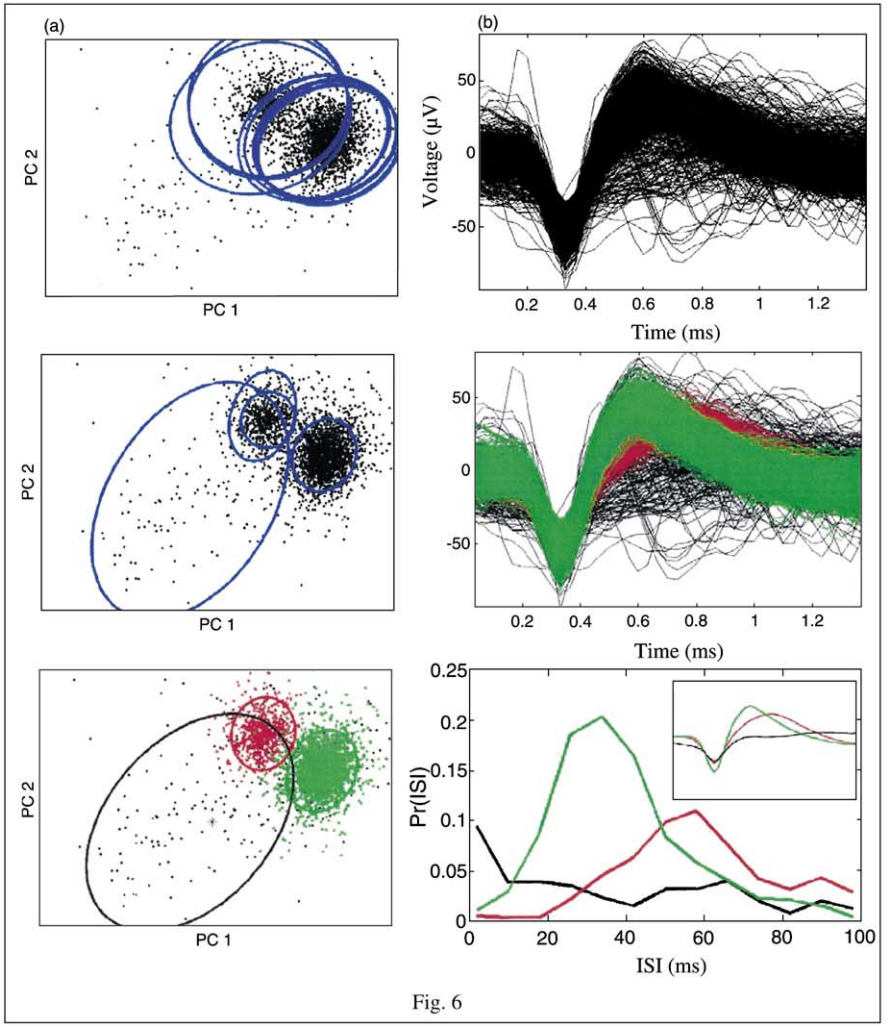

Fig. 6

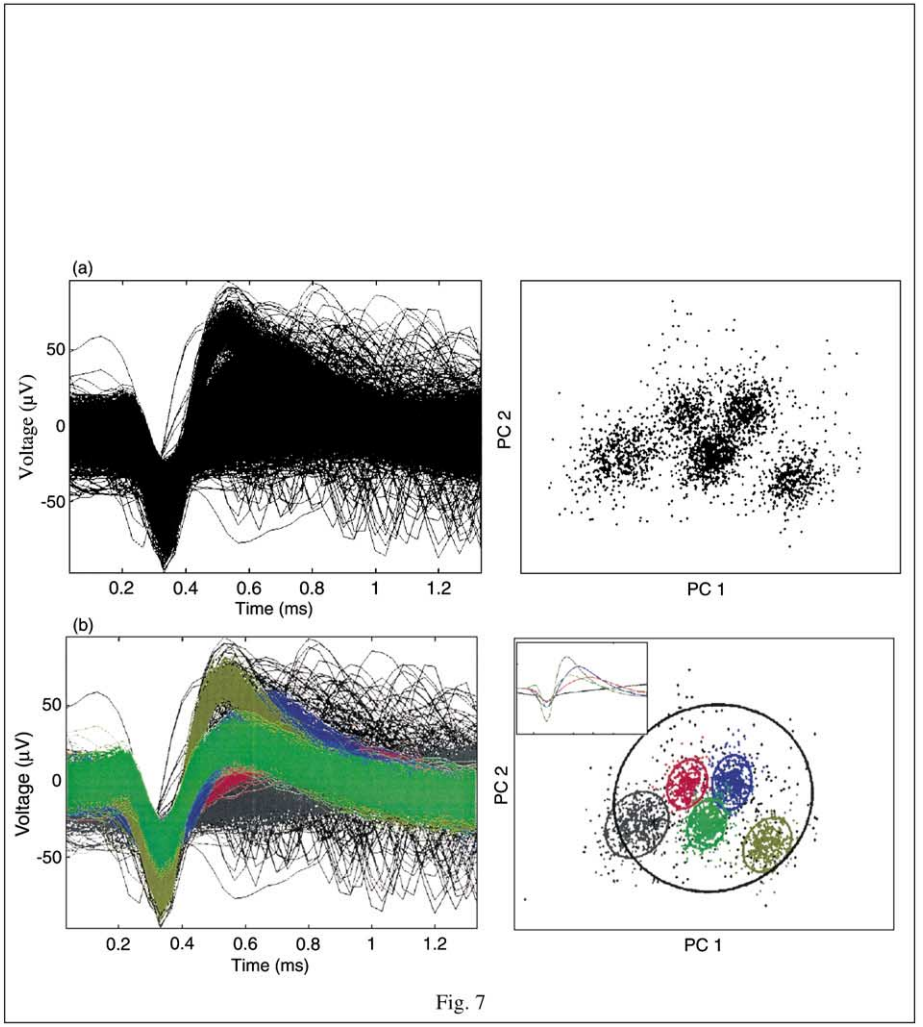

Fig. 7

Fig. 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). Inset: waveform templates.

Fig. 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. Inset 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.

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.

6.1. Alternatives and possible extensions

Another solution to the problem of non-Gaussian waveform distributions (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 (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. Next generations of this algorithm can also incorporate additional information regarding the behavior of spike trains into the process of spike sorting, including the existence of refractory periods and waveform changes during bursts. Examples of how to extend the probabilistic modeling approach we have used to include this domain-specific information are provided in a recent study (Sahani, 1999).

Acknowledgements

We wish to thank Professors Sri Nagarajan, Mario Figueiredo, and John Donoghue for valuable input and support during the preparation of this manuscript. We thank the two anonymous reviewers for their insightful comments. The work was supported by a State of Utah Center of Excellence contract #95-3365 to R.A.N., and NIH grant # R01NS25074 to Professor Donoghue.

Appendix A

Following the ideas of the ECME algorithm (Liu and Rubin, 1994), we are interested in maximizing the penalized-loglikelihood (Eq. (5)) directly with respect to π_j:

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

The maximization is subject to the constraint: $\Sigma_{j=1}^{g} \pi_j = 1$. To solve this constrained optimization problem we use a Lagrange multiplier; we now have to maximize:

$$L_{p'} = \sum_{i=1}^{n} \log \sum_{j=1}^{g} \pi_j P_{ij} - \left[\frac{N}{2} \sum_{j=1}^{g} \log \frac{n\pi_j}{12} + \frac{g}{2} \log \frac{n}{12} + \frac{g(N+1)}{2} \right] + \lambda \left[\sum_{j=1}^{g} \pi_j - 1 \right] \tag{13}$$

Differentiating with respect to π_j, we obtain:

$$\sum_{i=1}^{n} \frac{P_{ij}}{\sum_{l=1}^{g} \pi_l P_{il}} - \frac{N}{2\pi_j} + \lambda = 0 \tag{14}$$

Multiplying by π_j/g and summing with respect to j:

$$\frac{1}{g} \sum_{i=1}^{n} \frac{\sum_{j=1}^{g} \pi_j P_{ij}}{\sum_{l=1}^{g} \pi_l P_{il}} - \sum_{i=1}^{g} \frac{N}{2g} + \frac{\lambda}{g} \sum_{j=1}^{g} \pi_j = \frac{n}{g} - \frac{N}{2} + \lambda = 0 \tag{15}$$

Substituting λ from Eq. (15) back into Eq. (14) and rearranging, we get the formula:

$$\pi_j = \sum_{i=1}^{n} \frac{\pi_j P_{ij}}{\sum_{j=1}^{g} \pi_j P_{ij}} - \frac{N}{2} \Bigg/ n - \frac{gN}{2} \tag{16}$$

Which can be solved iteratively. Following Figueiredo and Jain (2002) we also enforce the additional constraint $\pi_j \geqslant 0$ during the iterations, which leads to Eq. (7).

References

Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. Biometrics 1993;49:803–21.

Celeux G, Chertien S, Forbes F, Mkhadri A. A component-wise EM algorithm for mixtures. France: INRIA, 1999.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data using the EM algorithm (with discussion). J R Stat Soc B 1977;39:1–39.

Donoghue JP. Connecting cortex to machines: recent advances in brain interfaces. Nat Neurosci 2002;5(Suppl):1085–8.

Fee MS, Mitra PP, Kleinfeld D. Automatic sorting of multiple unit neuronal signals in the presence of anisotropic and non-Gaussian variability. J Neurosci Methods 1996a;69(2):175–88.

Fee MS, Mitra PP, Kleinfeld D. Variability of extracellular spike waveforms of cortical neurons. J Neurophysiol 1996b;76(6):3823–33.

Figueiredo M, Jain A. Unsupervised learning of finite mixture models. IEEE Trans PAMI 2002;24(3):381–96.

Garcia P, Suarez CP, Rodriguez J, Rodriguez M. Unsupervised classification of neural spikes with a hybrid multilayer artificial neural network. J Neurosci Methods 1998;82(1):59–73.

Georgopoulos AP, Kalaska JF, Caminiti R, Massey JT. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. J Neurosci 1982;2(11):1527–37.

Guillory KS, Normann RA. A 100-channel system for real time detection and storage of extracellular spike waveforms. J Neurosci Methods 1999;91(1–2):21–9.

Harris KD, Henze DA, Csicsvari J, Hirase H, Buzsaki G. Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. J Neurophysiol 2000;84(1):401–14.

Hoogerwerf AC, Wise KD. A three-dimensional microelectrode array for chronic neural recording. IEEE Trans Biomed Eng 1994;41(12):1136–46.

Huber PJ. Robust statistics. Wiley: New York, 1982.

Hulata E, Segev R, Ben-Jacob E. A method for spike sorting and detection based on wavelet packets and Shannon's mutual information. J Neurosci Methods 2002;117:1–12.

Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 2000;22(1):4–37.

Johnson NL, Kotz S, Balakrishnan N. Continuous Univariate Distributions. New York: Wiley; 1994.

Jones KE, Campbell PK, Normann RA. A glass/silicon composite intracortical electrode array. Ann Biomed Eng 1992;20(4):423–37.

Kim KH, Kim SJ. Neural spike sorting under nearly 0-dB signal-to-noise ratio using nonlinear energy operator and artificial neural-network classifier. IEEE Trans Biomed Eng 2000;47(10):1406–11.

Lange KL, Little RJA, Taylor JMG. Robust statistical modeling using the t distribution. J Am Stat Assoc 1989;84(408):881–96.

Lewicki MS. Bayesian modeling and classification of neural signals. Neural Comput 1994;6(5):1005–30.

Lewicki MS. A review of methods for spike sorting: the detection and classification of neural action potentials. Network 1998;9(4):R53–78.

Liu C, Rubin DB. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika 1994;81(4):633–48.

Maynard EM, Fernandez E, Normann RA. A technique to prevent dural adhesions to chronically implanted microelectrode arrays. J Neurosci Methods 2000;97(2):93–101.

McLachlan GJ, Peel D. Finite mixture models. New York: Wiley, 2000.

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

Ohberg F, Johansson H, Bergenheim M, Pedersen J, Djupsjobacka M. A neural network approach to real-time spike discrimination during simultaneous recording from several multi-unit nerve filaments. J Neurosci Methods 1996;64(2):181–7.

Paninski L, Fellows MR, Hatsopoulos NG, Donoghue JP. Temporal tuning properties for hand position and velocity in motor cortical neurons. J Neurophysiol, in review.

Peel D, McLachlan GJ. Robust mixture modelling using the t distribution. Stat Comput 2000;10:339–48.

Quirk MC, Wilson MA. Interaction between spike waveform classification and temporal sequence detection. J Neurosci Methods 1999;94(1):41–52.

Richardson S, Green P. On Bayesian analysis of mixtures with unknown number of components. J R Stat Soc B 1997;59:731–92.

Rousche PJ, Pellinen DS, Pivin DP, Jr., Williams JC, Vetter RJ, Kipke DR. Flexible polyimide-based intracortical electrode arrays with bioactive capability. IEEE Trans Biomed Eng 2001;48(3):361–71.

Sahani M, Pezaris JS, Andersen RA. On the Separation of Signals from Neighboring Cells in Tetrode Recordings. Advances in Neural Information Processing Systems 11, Denver, CO: 1997.

Sahani M. Latent Variable Models for Neural Data Analysis. Ph.D. Dissertation. Computation and Neural Systems. California Institute of Technology, 1999

Salganicoff M, Sarna M, Sax L, Gerstein GL. Unsupervised waveform classification for multi-neuron recordings: a real-time, software-based system. I. Algorithms and implementation. J Neurosci Methods 1988;25(3):181–7.

Schmidt EM. Computer separation of multi-unit neuroelectric data: a review. J Neurosci Methods 1984;12(2):95–111.

Shoham S. Advances towards an implantable motor cortical interface. Ph.D. dissertation. Dept. of Bioengineering. University of Utah, 2001.

Shoham S. Robust clustering by Deterministic Agglomeration EM of mixtures of multivariate t-distributions. Pattern Recognition 35: 2002.

Snider RK, Bonds AB. Classification of non-stationary neural signals. J Neurosci Methods 1998;84(1–2):155–66.

Tadjudin S, Landgrebe DA. Robust parameter estimation for a mixture model. IEEE Trans Geosci Remote Sens 2000;38(1):439–45.

Wallace C, Freeman P. Estimation and inference via compact coding. J R Stat Soc B 1987;49(3):241–52.

Zouridakis G, Tam DC. Identification of reliable spike templates in multi-unit extracellular recordings using fuzzy clustering. Comput Methods Prog Biomed 2000;61(2):91–8.