Method for Unsupervised Classification of Multiunit Neural Signal Recording Under Low Signal-to-Noise Ratio

Kyung Hwan Kim*, *Member, IEEE* and Sung June Kim, *Member, IEEE*

***Abstract*—Neural spike sorting is an indispensable step in the analysis of multiunit extracellular neural signal recording. The applicability of spike sorting systems has been limited, mainly to the recording of sufficiently high signal-to-noise ratios, or to the cases where supervised classification can be utilized. We present a novel unsupervised method that shows satisfactory performance even under high background noise. The system consists of an efficient spike detector, a feature extractor that utilizes projection pursuit based on negentropy maximization (Huber, 1985 and Hyvarinen *et al.*, 1999), and an unsupervised classifier based on probability density modeling using mixture of Gaussians (Jain *et al.*, 2000). Our classifier is based on the mixture model with a roughly approximated number of Gaussians and subsequent mode-seeking. It does not require accurate estimation of the number of units present in the recording and, thus, is better suited for use in fully automated systems. The feature extraction stage leads to better performance than those utilizing principal component analysis and two nonlinear mappings for the recordings from the somatosensory cortex of rat and the abdominal ganglion of *Aplysia*. The classification method yielded correct classification ratio as high as 95%, for data where it was only 66% when a k-means-type algorithm was used for the classification stage.**

***Index Terms*—Mixture of Gaussians, neural spike sorting, projection pursuit, unsupervised classification.**

I. Introduction

THE EXTRACELLULAR recording of neural signals is of prime importance in investigating information transmission within nervous system, since it enables the simultaneous monitoring of the activities of multiple neurons. Recorded waveform usually consists of action potentials from several neurons that are in close proximity to the electrode site under investigation, thus, they must be classified into spike trains from each cell for further analysis in which each spike train is considered to be a point process [1]. This procedure for classifying multiunit neural signals into multiple spike trains is referred to as neural spike sorting, and a considerable amount of study has occurred during the past several decades as summarized in [2] and [3]. Currently it is possible to obtain a satisfactory result even under very high background noise as shown in [4] and [5], when a supervised classifier is used. For the online or first offline analysis of experimental data where no template waveforms for each unit are available and the number of units in the recording is unknown, the application of a supervised classifier is not possible. Although several studies on unsupervised neural spike sorting such as those described by Zouridakis and Tam [25], Fee *et al.* [6], and Sahani [7] have been reported recently, their performance under low signal-to-noise ratio (SNR) conditions has not been demonstrated.

The goal of this study was to develop an unsupervised neural spike sorting system that shows high performance under low SNR. Our approach was to combine action potential detectors that have been partially described in [5] and [8], a dimensionality reduction scheme that provides effective discriminative features, and a proper unsupervised classification method. For feature extraction we use a linear projection that shows higher separability than conventional methods. Our classification method permits the robust estimation of the distribution shape of each cluster and is insensitive to the parameters that are to be predetermined. The importance of the feature extraction stage arises from "the curse of dimensionality" [9]. The training of employed unsupervised classifier was not successful under the low SNR targeted in this paper, if raw data of approximately 25–40 dimensions (time samples of action potential) were used without dimensionality reduction. In the case where the SNR is sufficiently high, conventional methods such as principal component analysis (PCA) or simple *ad hoc* features such as peak-to-peak amplitude and/or spike duration, can be useful [10]. However, under high background noise, they become inadequate as discriminative features. Here, we employ a projection pursuit method based on negentropy maximization (PP/NEM) [11], [12] and show that it is more capable of extracting discriminative features for spike sorting. These features usually form hyper-ellipsoidal clusters and, thus, the failure rate of classification is prohibitively high if a clustering method based solely on Euclidean distance is used. Since the aim was to develop a method by which the shape of distribution can be considered, a clustering algorithm based on the modeling of probability density function (pdf) by mixture of Gaussians (MoG) was used. This type of algorithm is frequently used for this purpose [13]. Our focus here is on solving the problem of determining the number of Gaussians in the mixture model. We show that it is possible to obtain robust unsupervised spike sorting, by a rough estimation of the number of Gaussians and subsequent mode-seeking. We demonstrate that it is possible

Manuscript received March 4, 2002; revised November 22, 2002. This work was supported by the Korea Science and Engineering Foundation (KOSEF) through Nano Bioelectronics and Systems Research Center. *Asterisk indicates corresponding author.*

*K. H. Kim is with the Human-Computer Interaction Laboratory, Samsung Advanced Institute of Technology, Yongin 499-712, Korea (khkim@ieee.org).

S. J. Kim is with the School of Electrical Engineering and Computer Science, Nano Bioelectronics and Systems Research Center, Seoul National University, Seoul 151-742, Korea.

Digital Object Identifier 10.1109/TBME.2003.809503

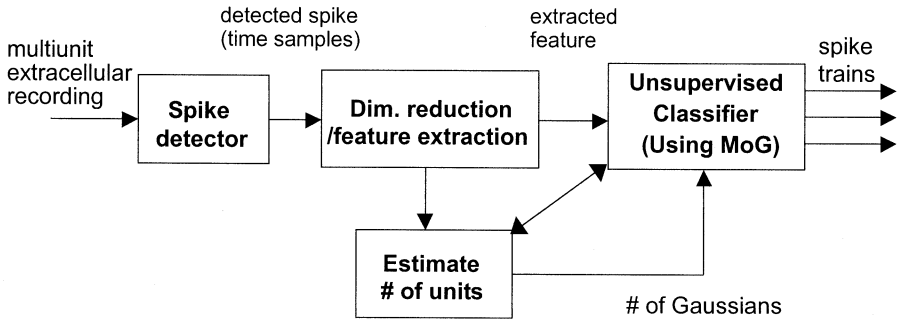


Fig. 1. Block diagram of the overall unsupervised sorting system.

to obtain a high success rate by combining this classification scheme with PP/NEM.

Fig. 1 shows the block diagram of the proposed unsupervised action potential classification system. First, the raw waveform from the electrode and recording electronics is passed to the action potential detector. We proposed two effective action potential detectors, based on the Teager energy operator [5] and on the prudent combination of discrete wavelet transform coefficient at several scales [8]. Thorough descriptions of the detection methods can be found in a companion paper [8] and, thus, the details are omitted in this paper. The detected waveforms are given to the block that performs the dimensionality reduction, after which the extracted feature vectors are given as input to the unsupervised classifier.

II. Experiments and Test Data Generation

A performance test must be done for the data where exact information such as the number of units in the recording and the firing times of each unit are known preliminarily. This is possible only for synthetic test data. In order to generate test data that faithfully represents the actual characteristics of experimental recordings, we extracted template waveforms of distinct units, and autoregressive (AR) models of background noise from the neural signal recordings of the abdominal ganglion of *Aplysia* and the somatosensory cortex of the Sprague–Dawley rat. The details of the recording experiments have been described thoroughly in [5] and [14]. Both were recorded using thin-film semiconductor microelectrodes, the impedance ranges of which were 2–3 MΩ at 1 kHz. Bandpass filtering was employed for both recordings (100 Hz–5 kHz for the *Aplysia* recording, 300 Hz–3 kHz for rat recording, respectively). The sampling rate was 10 ksamples/s and 20 ksamples/s, for the *Aplysia* and rat recording, respectively. The extraction of template waveforms was done by a human supervisor assisted by efficient detectors described in [5] and [8], and fuzzy c-means (FCM) clustering of waveforms at a reasonably high SNR. The action potential segments consist of 25 samples and 40 samples for the *Aplysia* and for the rat recording, respectively. The template waveforms of the three units extracted from the rat recording are shown in Fig. 2, and those of the *Aplysia* recording can be found in [5]. The AR model coefficients were identified from background noise segments of approximately 300 samples by Burg's algorithm [15] along with order determination using Akaike information criteria [15]. Accurate determination of the order was not critical for effective modeling. From the template waveforms of several units and the AR model of the background noise, it becomes possible to generate the waveforms of arbitrary SNR that have the characteristics of real experimental recordings. The SNR is defined as shown in the equation at the bottom of the page.

III. Algorithm Description

A. Feature Extraction Stage: Projection Pursuit Based on Negentropy Maximization

We use a linear transform for feature extraction because it preserves the underlying shape of distribution in high dimension and, thus, the overall pdf modeling of the extracted feature vectors by the Gaussian mixture is feasible. A linear transform can be expressed as $\mathbf{y} = \mathbf{W}^T\mathbf{x}$ where $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$ is m-dimensional observed vector, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is n-dimensional random vector $(n \leq m)$, and $\mathbf{W}$ is an $m \times n$ matrix. The actual dimension of $\mathbf{y}$ is different from case to case, and determined so that the required classification accuracy is met by the employed classifier. The projection matrix $\mathbf{W}$ must be found so that components of $\mathbf{y}$ become discriminative features, i.e., separability among clusters is maximized. This type of problem is referred to as projection pursuit [11]. Appropriate objective function must be defined to find $\mathbf{W}$ that maximizes the separability. The measure of non-Gaussianity is appropriate for this purpose considering the well-known fact that the multimodality of given high dimensional data might be represented most lucidly in the direction where non-Gaussianity is maximized [16], and the multimodality in the resulting distribution is desirable for clustering [11], [16]. It is also well-known that entropy is minimized for the most non-Gaussian distribution; entropy has a small value for the distribution that is concentrated on certain

$$\text{SNR} = \left(\frac{\text{peak-to-peak value of action potential with minimum amplitude}}{\text{root-mean-square value of pure noise segment}} \right)^2$$

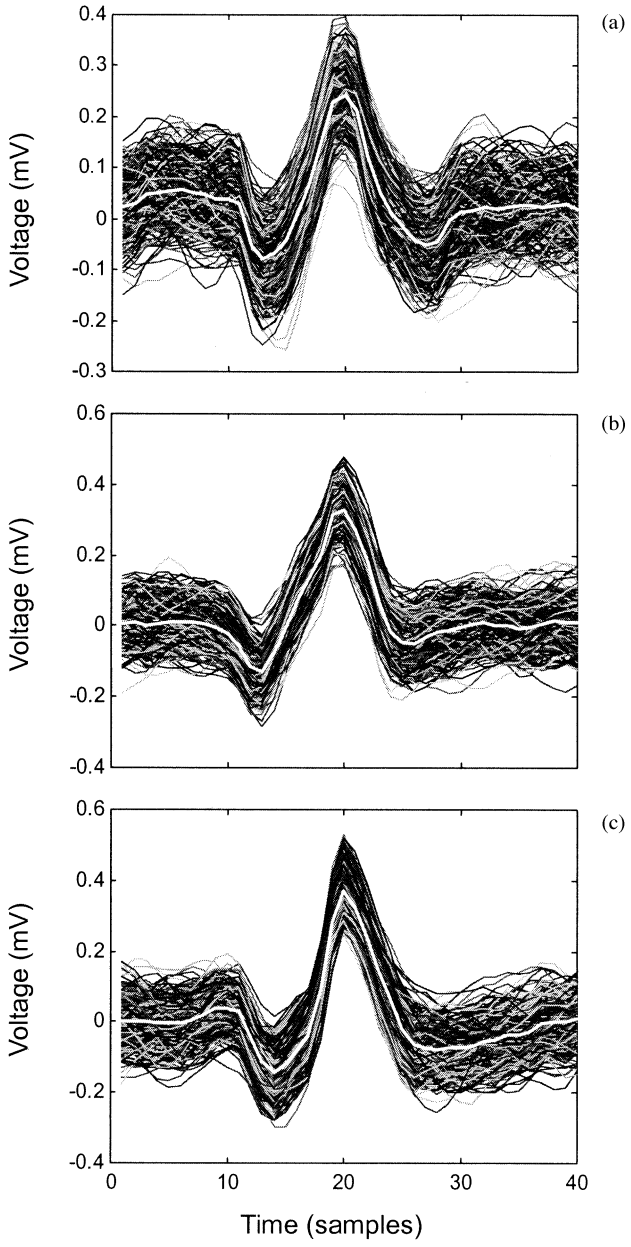


Fig. 2. Overlapped action potential waveforms at SNR $\approx$ 2.4 along with template waveforms (thick white lines) for three units of recording from rat somatosensory cortex. Each unit consists of 40 samples (2 ms).

values, i.e., when the variable is clearly clustered. This relation of non-Gaussianity and entropy can be used to seek the projection that maximizes separability.

Negentropy $J(\mathbf{y})$ is defined as $J(\mathbf{y}) = H(\mathbf{y}_G) - H(\mathbf{y})$ where $H(\mathbf{y})$ denotes the entropy of $\mathbf{y}$ [11], so that it directly signifies the measure of non-Gaussianity, as it is set to zero for a Gaussian random vector. Here, $\mathbf{y}_G$ denotes a Gaussian random vector with the same mean and variance as $\mathbf{y}$. Our goal can be achieved by maximizing $J(\mathbf{y})$, and as a result, we refer to our feature extraction technique as PP/NEM. To find the value of $\mathbf{W}$ that maximizes $J(\mathbf{y})$, we utilize a batch-type learning algorithm based on the study of Hyvarinen [16]. His algorithm is mainly intended for independent component analysis (ICA) for source separation. In fact, independence among each component is sought indirectly by maximizing the summation of the negentropies of components since these two criteria have been shown to be identical [16], [17]. Hence, the algorithm is naturally fitted for our purpose. An online (adaptive-filter-type) algorithm based on the gradient ascent method was not employed because of its slow convergence and dependence on the appropriateness of the choice of learning rate sequence [17].

Hyvarinen's technique is described briefly below. The input data ($\mathbf{x}$s) are first whitened and centered so that their mean becomes zero. Based on a simple expression of the negentropy $J(y)$, a basis of projection, $\mathbf{w}$ (a column vector of $\mathbf{W}$) can be obtained by maximizing $J(y) = J(\mathbf{w}^T\mathbf{x})$. The maximization can be transformed into an equation-solving problem by means of the Kuhn–Tucker theorem [18]. Subsequently, a solution, i.e., a basis of projection, can be found using Newton–Raphson iteration, which does not require a predetermined learning rate sequence. One step in this procedure can be expressed as follows [16]:

$$\mathbf{w}^+ = E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w},$$

and

$$\mathbf{w}^{new} = \mathbf{w}^+ / \left\|\mathbf{w}^+\right\| \tag{1}$$

where $g(\cdot)$ denotes the derivative of some nonquadratic function that approximates $J(y)$ [16]. It is suggested that $g_1(y) = \tanh(a_1 y)(1 \le a_1 \le 2)$, $g_2(y) = y\exp(-y_2/2)$, or $g_3(y) = y^3$ are suitable choices.

Iteration by (1) finds just *one* direction of projection. There are several values of $\mathbf{w}$ where the objective function has a local maxima, and we wish to find more than one direction, i.e., more than one discriminative feature. Suppose that we have already found $p-1$ basis $\mathbf{w}_j$s ($j = 1, 2, \ldots, p$). In order to prevent a newly found basis $\mathbf{w}_p$ from converging to the already found local maxima, Gram-Schmidt orthogonalization is performed for each iteration as [16]

$$\mathbf{w}_p^* = \mathbf{w}_p - \sum_{j=1}^{p-1}(\mathbf{w}_p^T\mathbf{w}_j)\mathbf{w}_j \quad \text{and} \quad \mathbf{w}_p^{new} = \mathbf{w}_p^* / \left\|\mathbf{w}_p^*\right\| \tag{2}$$

This requirement is justified from the fact that the correlation matrix of the data must be the identity matrix due to the whitening, and this requires, in turn, the orthonormality of basis vectors [26]. This procedure is not computationally intensive, because the actual dimension of the feature vector is limited to a low value. In this paper we extract two-dimensional (2-D) features in order to facilitate visualization and MoG model estimation. For this purpose, we first find three $\mathbf{w}$s (thus, $p = 3$) and subsequently choose two of them according to resulting $J(y)$.

B. Unsupervised Classification Stage: Mixture of Gaussian Model

After dimensionality reduction, the projected data points are given to the unsupervised classifiers. Most algorithms for the unsupervised classification are based on either a mixture model of pdf, or a k-means-related algorithm [13]. We opted to use the

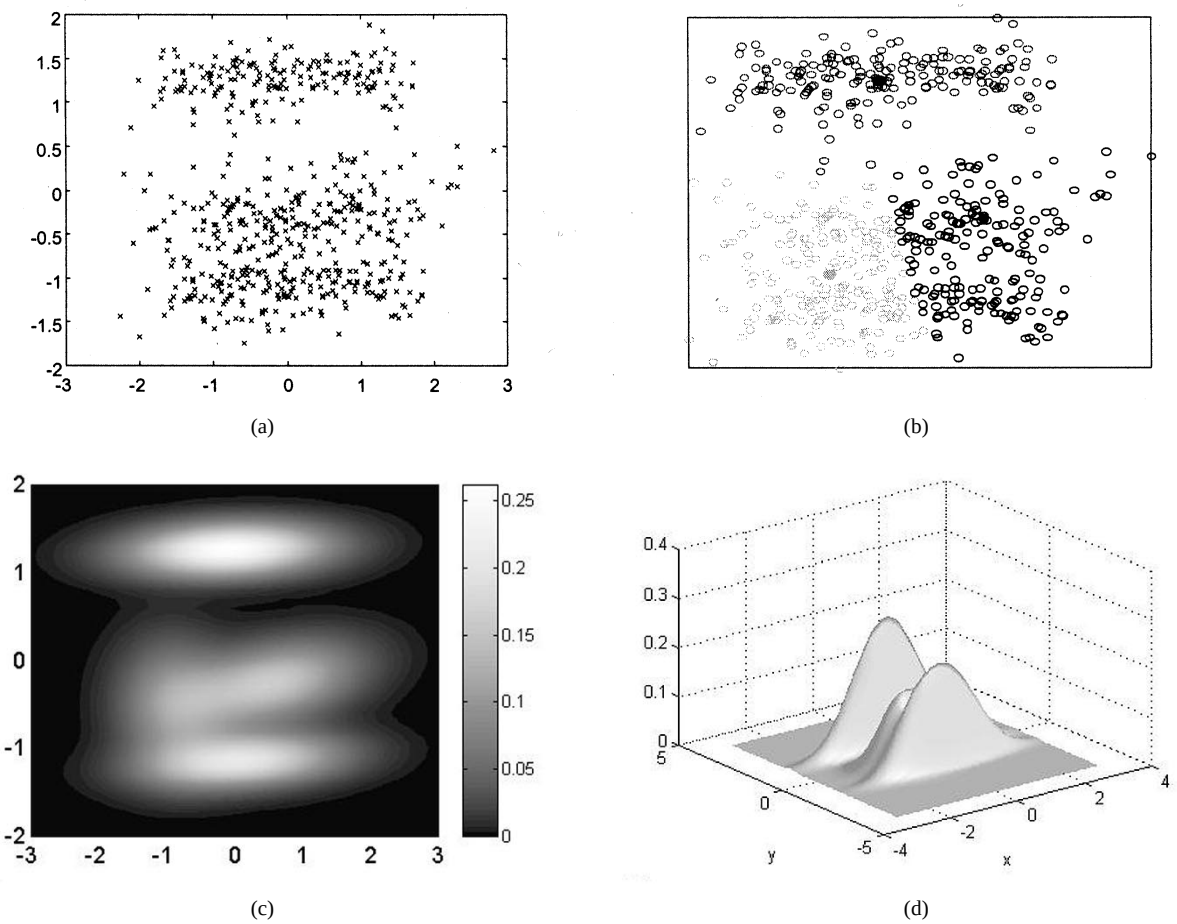


Fig. 3. Application of the proposed projection and pdf model identification for an *Aplysia* recording with SNR ≈ 1.25. (a) Projection result using PP/NEM. (b) Clustering result using the FCM algorithm (different clusters appear in varying shades of gray scales). (c) Two-dimensional and (d) three-dimensional view of the identified MoG pdf model.

former because the k-means-related algorithms determine the membership of each data point solely by Euclidean distance, and cluster shape cannot be considered. As shown in Fig. 3(a), in many cases where the feature extraction is performed by PP/NEM, the extracted feature vectors form a distribution with elongated shape because the degree of scatter is different for each component. Therefore, it is obvious that we must utilize an unsupervised classification method by which the cluster shape can be considered. We exploited the modeling of the overall pdf by MoG. This method has been utilized for many unsupervised pattern classification problems, and was also used for the neural spike sorting by Lewicki [19]. Here, we concentrate on the problem of determining the number of Gaussians in the mixture, and the number of units in the recording. We show that this problem can be settled by using a roughly estimated number of Gaussians and then subsequently seeking the modes of the obtained pdf model.

The MoG model is defined as follows:

$$p(\mathbf{x}) = \sum_{m=1}^{K} p(m)p(\mathbf{x}|m) = \sum_{m=1}^{K} \pi_m p(\mathbf{x}|m). \tag{3}$$

Here, π_m is the prior probability of the mth Gaussian, and $\sum_{m=1}^{M} \pi_m = 1$, $\pi_m \in (0, 1)\ \forall m = 1, \ldots, K$. K denotes the number of Gaussians used to represent the pdf of the given data. Each $p(\mathbf{x}|m)$ is a Gaussian distribution function, i.e.,

$$\begin{aligned} p(\mathbf{x}|m) &= g(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = g(\mathbf{x}|\boldsymbol{\theta}_m) \\ &= |2\pi\boldsymbol{\Sigma}_m|^{-(1/2)} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^T\boldsymbol{\Sigma}_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)\right). \end{aligned} \tag{4}$$

Here, $g(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ is the mth Gaussian whose mean vector and covariance matrix are $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$, respectively. It is well known that the parameter vector of the MoG $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K) = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ can be iteratively estimated by the application of expectation-maximization (EM) algorithm [13].

Fig. 3(c) and (d) shows an example of the pdf modeling of the data with three ellipsoidal clusters, shown in Fig. 3(a). Note that it is not possible to obtain a satisfactory clustering result using the k-means-related algorithm (FCM) as shown in Fig. 3(b), primarily because information on the shape of each cluster cannot be considered. As shown in Fig. 3(c) and (d), we were able to find a reasonable estimate of the pdf of the data in Fig. 3(a) with the MoG model estimated by the EM algorithm.

The number of Gaussians for the mixture model, K, must be determined prior to the use of the EM algorithm. A similar problem, the determination of the number of clusters, arises

for the case of the k-means-related algorithms. Although many studies on cluster validity indexes [20], which give best estimate of the number of clusters, have been proposed in order to use the k-means-related algorithms without knowledge of the exact number of clusters, their performance does not seem to be satisfactory for many practical problems, and a manual determination is necessary. At first sight it also appears that a considerable amount of variability in the resulting MoG pdf model might occur when different K values are used, and similar problems might occur in our method. However, although the resultant parameter vector $\boldsymbol{\theta}$ can vary considerably, the pdf models obtained are quite similar for different Ks in the case of neural spike sorting, when the K values are slightly larger that the actual number of clusters. This can be explained as shown in Fig. 4(a). Here, K is assumed to be three, when the actual number of clusters is 2. In several attempts to estimate the MoG parameter, we consistently obtained the result shown by the solid line in Fig. 4(a), where the means of two specific Gaussians are so close to each order that a single peak is formed by merging them, and the overall estimated pdf (dashed line) has a bimodal shape.

The problem of determining K for the MoG parameter estimation can be considered to be a decision problem where the maximum-(log)likelihood estimation can be applied. However, the likelihood is a monotonically increasing function of the number of parameters; this is universal in other estimation problems such as the AR modeling of time-series. A number of criteria, such as Akaike's information criteria and minimum description length have been proposed to determine the model order [15]. These criteria consist of a log-likelihood term and a penalty term the purpose of which is to compensate for the monotonic increase of the log-likelihood as a function of the model order (the number of parameters). A similar approach can be considered for the determination of K. However, it is not easy to define the penalty term that is generally applicable. Instead, we use a method based on the behavior of the log-likelihood versus K curve. As shown in Fig. 4(b), this curve typically shows rapid initial increment behavior, followed by a slow increment [21]. The actual number of clusters is located slightly above the "*knee*" of this curve. Because a satisfactory estimation of the MoG parameters is possible when the value of K is set to a slightly larger value than the actual number of clusters as explained above, and does not need to be accurate, it is possible to determine the value of K to be used as follows.

1) Calculate the log-likelihood, Q, for several values of the number of Gaussians, k.
2) Find the number of Gaussians to be used in MoG estimation, K, as follows:

$$K = \arg\max_{k} \left(\frac{\Delta Q}{\Delta k} \right) + 2. \tag{5}$$

The value of K determined by (5) is denoted by an arrow in Fig. 4(b).

An actual procedure for applying the identified MoG model to unsupervised neural spike sorting is described below. Because the number of Gaussians can be different from the true number of units, in order to use the learned MoG model for classification, it is necessary to identify the number and position of the

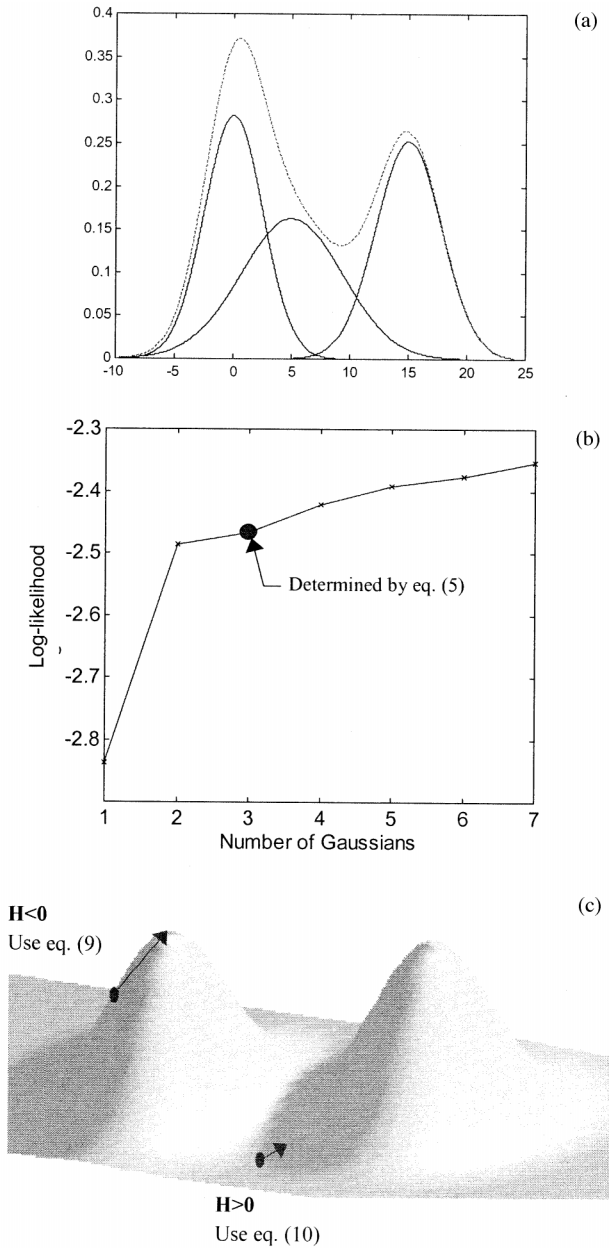


Fig. 4. Robustness of the pdf identification algorithm on the number of clusters, which should be predetermined. (a) Two-mode pdf estimation as a mixture of three Gaussians. Solid line: Each Gaussian component. Dashed line: Overall estimated pdf. (b) Log-likelihood as a function of the number of Gaussians used for the MoG estimation. (c) Graphical illustration of the mode-finding algorithm by using quadratic maximization or gradient ascent method.

"modes" (local maxima of MoG) in the mixture model of pdf, and to assign each Gaussian in the mixture to a specific mode. After identifying the local maxima, each mean of Gaussians $\boldsymbol{\mu}_m$ is assigned to the closest local maximum so that the pdf of a single cluster (i.e., a single unit) can be represented by the sum of the Gaussians as follows:

$$p_l(\mathbf{x}) = \sum_{m=1}^{L} \pi_m g(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{6}$$

where $p_l(\mathbf{x})$ denotes the pdf of the lth cluster, and L is the total number of Gaussians the center of which is closest to the lth

mode. Hence, $l(\mathbf{x})$, the class label of a particular data point $\mathbf{x}$ is determined as follows:

$$l(\mathbf{x}) = \arg\max_{l} p_l(\mathbf{x}) \tag{7}$$

The method of finding the local maxima of a given MoG is shown below. A more elaborate algorithm on this subject has recently been reported [22], but following this simple method was sufficient for our purpose. We use the quadratic maximization or gradient ascent method [18] and, thus, the expression of the Hessian and gradient is necessary. Suppose we are starting the quadratic search at a data point $\mathbf{x}_0$. By Taylor series expansion, $p(\mathbf{x})$ is expressed as

$$p(\mathbf{x}) \approx p(\mathbf{x}_0) + (\mathbf{x}-\mathbf{x}_0)^T \mathbf{g}(\mathbf{x}_0) + \tfrac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T \mathbf{H}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0). \tag{8}$$

Here, $\mathbf{g}(\mathbf{x}_0) = \nabla p(\mathbf{x}_0)$ and $\mathbf{H}(\mathbf{x}_0) = (\nabla\nabla^T)p(\mathbf{x}_0)$. The data point where the gradient becomes zero is given by

$$\begin{aligned} \nabla p(\mathbf{x}) &= \mathbf{g}(\mathbf{x}_0) + \mathbf{H}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) = \mathbf{0} \\ &\Rightarrow \mathbf{x} = \mathbf{x}_0 - \mathbf{H}^{-1}(\mathbf{x}_0)\mathbf{g}(\mathbf{x}_0). \end{aligned} \tag{9}$$

According to (9), we proceed from $\mathbf{x}_0$ to the maximum in a single step, if the Hessian is negative definite ($\mathbf{H} < 0$). If the Hessian is positive definite, indicating that we have not yet reached a hill cap that is defined as the region around a mode where $\mathbf{H} < 0$, the gradient ascent method is then used as follows:

$$\mathbf{x} = \mathbf{x}_0 + \eta \mathbf{g}(\mathbf{x}_0) \tag{10}$$

where η is the step size. Once the point where $\mathbf{g} = \mathbf{0}$ is reached, we can check whether it is at a maximum by verifying that $\mathbf{H} < 0$. The means of all the Gaussians in the MoG model are used as starting points for the search. This hill-climbing procedure is graphically illustrated in Fig. 4(c). While doing the quadratic or gradient ascent search of (9) or (10), we utilize the closed-form expressions of the gradient and Hessian, instead of numerically calculated values. They can be derived as follows:

$$\mathbf{g} = \nabla p(\mathbf{x}) = \sum_{m=1}^{M} \pi_m p(\mathbf{x}|m)\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\mu}_m - \mathbf{x}) \tag{11}$$

$$\begin{aligned} \mathbf{H} &= (\nabla\nabla^T)p(\mathbf{x}) \\ &= \sum_{m=1}^{M} \pi_m p(\mathbf{x}|m)\boldsymbol{\Sigma}_m^{-1}((\boldsymbol{\mu}_m - \mathbf{x})(\boldsymbol{\mu}_m - \mathbf{x})^T - \boldsymbol{\Sigma}_m)\boldsymbol{\Sigma}_m^{-1}. \end{aligned} \tag{12}$$

The overall procedure of the proposed spike sorting system is illustrated in a pseudocode form in Fig. 5(a). Fig. 5(b) and (c) shows the feature extraction procedure by the PP/NEM and mode-seeking procedure, respectively.

IV. Performance of the Overall System

First we show the result of a projection using the PP/NEM onto 2-D feature space, applied to a three-unit recording from the somatosensory cortex of rat. Fig. 6 shows the scatter plot of the projected data points using the PP/NEM [Fig. 6(a)] along with the scatter plots of the feature vectors extracted by the PCA [Fig. 6(b)], and other two nonlinear projection methods [Sammon's mapping in Fig. 6(c) and generative topographic mapping (GTM) in Fig. 6(d)]. Details of the GTM and Sammon's mapping are given in [23] and [24], respectively. There are 700 data points from each cluster. For the linear projections [Fig. 6(a) and (b)], two basis vectors of the projection are also shown in the inset. For the PCA, the first two principal components were extracted as in [10]. The SNR of the recording used to generate Fig. 6 was 2.4. When the SNR is lowered to this level, overlap among the clusters in the PCA scatter plot becomes so severe as to prevent even manual sorting, in which human supervisors determine the decision boundary. In contrast, the feature vectors extracted by the PP/NEM form a clearly clustered structure so that the decision boundary can be easily determined and automated classification is possible using unsupervised classification algorithms. Although it is not certain solely from Fig. 6 whether the two nonlinear methods are superior to the PCA or not, they are evidently inferior to the PP/NEM for this case. We obtained a similar trend for the three-unit recording from the abdominal ganglion of *Aplysia*. For the purpose of quantitative comparison of the above methods for the clustering, we used a performance index based on a scatter matrix. This index is used for the Fisher's linear discriminant analysis, and is defined as follows [9]:

$$J_1(\mathbf{W}) = \frac{\det(\tilde{\mathbf{S}_B})}{\det(\tilde{\mathbf{S}_W})} = \frac{\det(\mathbf{W}^T\mathbf{S}_B\mathbf{W})}{\det(\mathbf{W}^T\mathbf{S}_w\mathbf{W})} \tag{13}$$

$$\mathbf{S}_w = \sum_{i=1}^{c} \mathbf{S}_i, \qquad \mathbf{S}_i = \sum (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T$$

$$\mathbf{S}_B = \sum_{i=1}^{c} (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \tag{14}$$

Because the determinant of the scatter matrix corresponds to the square of the volume of hyper-ellipsoidal scattering, $J_1(\mathbf{W})$ is the ratio between the separation of each cluster and the scattering within a single cluster. Thus, the larger $J_1(\mathbf{W})$ signifies a better projection in that the underlying cluster structure is revealed more clearly. Table I shows the separability index $J_1(\mathbf{W})$ computed for the PCA, GTM, Sammon's mapping, and PP/NEM, applied to the data in Fig. 6, and to the *Aplysia* data with SNR ≈ 1.6. The PP/NEM consistently shows superior performance, thus justifying our choice of the PP/NEM for the unsupervised spike sorting for low SNR data.

We applied a supervised linear classifier [9], [13] to the data shown in Fig. 6(a) (PP/NEM), and Fig. 6(b) (PCA) to test their efficacy for classification. The test result is summarized in contingency table form in Table II.

We demonstrate the performance of the entire system by showing an estimation of the overall pdf using the MoG model, for the data where successful clustering was impossible by FCM as in the case of Fig. 3(b), and then the classification result. The scatter plot of the extracted features from *Aplysia* data (SNR ≈ 1.6) is shown in Fig. 7(a). Fig. 7(b) shows the

Overall procedure (a)

Inputs: $\mathbf{X}=[\mathbf{x}_1\ \mathbf{x}_2\ \dots\ \mathbf{x}_N]$, where $\mathbf{x}_i$ is i'th action potential detected from raw recording

Outputs: $\mathbf{L}=[l(\mathbf{x}_1)\ l(\mathbf{x}_2)\ \dots\ l(\mathbf{x}_N)]$ and N_{mode}, where $l(\mathbf{x}_i)=\{1,2,..,C\}$, $l(\mathbf{x}_i)$ is class label of $\mathbf{x}_i$, and N_{mode} is number of units in the recording

1. Get $\mathbf{W}$ and extracted feature vector $\mathbf{y}$ by PP/NEM (*subroutine 1*)
2. *for* k=2, … M (M: predetermined number of steps)

   get MoG model of pdf of $\mathbf{y}$, $p(\mathbf{y})$ and log-likelihood $Q(k)$ (by EM algorithm)

   end for k
3. Determine K by equation (5)
4. Find the N_{mode} modes (local maxima) of $p(\mathbf{y})$ at $k=K$ (*subroutine 2*)
5. Assign each Gaussians in MoG to one of the modes found in step 4
6. Form the pdf of l'th cluster , $p_l(\mathbf{x})$, by equation (6)
7. Classify each data points into C clusters, i.e., determine $\mathbf{L}$ by equation (7)

Subroutine 1: PP/NEM (b)

Inputs: $\mathbf{X}=[\mathbf{x}_1\ \mathbf{x}_2\ \dots\ \mathbf{x}_N]$

Outputs: $\mathbf{W}=[\mathbf{w}_1\ \dots\ \mathbf{w}_p]$

1. Center the data: $\mathbf{x} \leftarrow \mathbf{x}-E\{\mathbf{x}\}$
2. Whiten the data so that $E\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\}=\mathbf{I}$
3. *for* all m=1,…p

   1. Randomly initialize $\mathbf{w}_m$ so that $\|\mathbf{w}_m\|=1$
   2. *do*

      iteration $\mathbf{w}_m^+ = E\{\mathbf{x}g(\mathbf{w}_m^T\mathbf{x})\} - E\{g'(\mathbf{w}_m^T\mathbf{x})\}\mathbf{w}_m$, and $\mathbf{w}_m = \mathbf{w}_m^+ / \left\|\mathbf{w}_m^+\right\|$

      orthogonalization $\mathbf{w}_m^* = \mathbf{w}_m - \sum_{j=1}^{m-1}(\mathbf{w}_m^T\mathbf{w}_j)\mathbf{w}_j$ and $\mathbf{w}_m = \mathbf{w}_m^* / \left\|\mathbf{w}_m^*\right\|$

      until $\mathbf{w}_m$ converges

   end for m
4. Form projection matrix $\mathbf{W}=[\mathbf{w}_1\ \dots\ \mathbf{w}_p]$
5. *for* i=1,2, … N

   $\mathbf{y}_i=\mathbf{W}^{\mathrm{T}}\mathbf{x}_i$

   end for i

Subroutine 2: Find the modes (local maxima) of p($\mathbf{y}$) (c)

Inputs: MoG model $p(\mathbf{x}) = \sum_{m=1}^{K} \pi_m \left|2\pi\boldsymbol{\Sigma}_m\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^T\boldsymbol{\Sigma}_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)\right)$

Outputs: Location of modes $\{\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_{Nmode}\}$ and number of modes N_{mode}

1. *for* i=1, …, K (K: Number of Gaussians in the mixture model $p(\mathbf{x})$)

   1. Set initial location of mode $\mathbf{x}_0 \leftarrow \boldsymbol{\mu}_i$
   2. *if* $\mathbf{H}(\mathbf{x}_0)<0$ $\mathbf{x}_0 \leftarrow \mathbf{x}_0 - \mathbf{H}^{-1}(\mathbf{x}_0)\mathbf{g}(\mathbf{x}_0)$ and go to step 1.3

      else

      $\mathbf{x}_0 \leftarrow \mathbf{x}_0 + \eta\mathbf{g}(\mathbf{x}_0)$

      go to step 1.2

      end if
   3. $\mathbf{m}_i=\mathbf{x}_0$
   4. Initialize $\mathbf{x}_0$ again

   end for i
2. Choose $\{\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_{Nmode}\}$ from $\{\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_K\}$ so that all the elements are different

Fig. 5. Pseudocode of (a) the overall procedure of the proposed spike sorting system, (b) the PP/NEM procedure, and (c) the mode-seeking procedure.

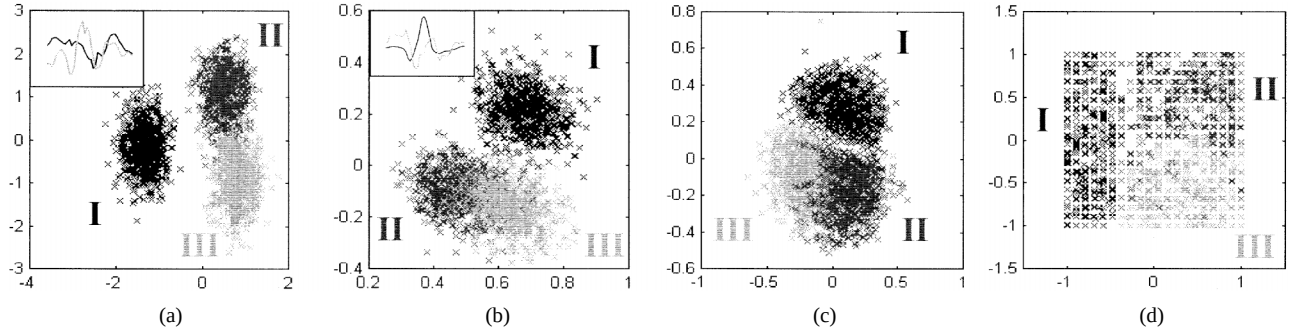


Fig. 6. Scatter plot of the projection obtained from a rat cortex recording with SNR ≈ 2.4.by (a) by PP/NEM, (b) PCA, (c) Sammon's mapping, and (d) GTM.

TABLE I
QUANTITATIVE COMPARISON OF THE SEPARABILITY OF PROJECTION METHODS BY $J_1(\mathbf{W})$

	Rat cortex recording, SNR≈2.46	*Aplysia* recording, SNR≈1.6
PP/NEM	29.1666	3.8272
PCA	3.7739	0.0758
GTM	2.3170	0.0162
Sammon's mapping	2.3360	0.0584

TABLE II
LINEAR CLASSIFICATION RESULT (LEFT: PP/NEM, RIGHT: PCA)

true \ estimated	Cluster I	Cluster II	Cluster III
Cluster I	700	0	0
Cluster II	0	667	33
Cluster III	0	42	658

true \ estimated	Cluster I	Cluster II	Cluster III
Cluster I	674	11	15
Cluster II	8	527	165
Cluster III	10	174	516

estimated pdf using four Gaussians. It is clear that we can obtain a quite reasonable result having three peaks. Next, Fig. 7(c)–(e) shows the results of the pdf estimation using 5–7 Gaussians, respectively. All of these also yielded reasonable results and, thus, the pdf estimation using the MoG model was not greatly dependent on the number of Gaussians, as we previously claimed. Hence, the classification result should be much less affected by a parameter that must be predetermined than the case of the k-means-related algorithms, so that the fully automated system based on the MoG would be expected to be much more reliable.

We next present the results of the estimation of the local maxima (modes) of the MoGs, which correspond to the average (or template) waveforms of each unit. The modes identified by the algorithms described in Section III-B are shown as triangles, and the true templates are denoted by squares in Fig. 7. When seven Gaussians were used, slight mismatches between the true templates and the estimated modes were found, but it still gave reasonable estimates of the pdfs and the location of the local maxima. Fig. 8 illustrates the step-by-step procedure for applying our algorithm to the *Aplysia* data when four Gaussians were used for the pdf modeling. "×" in the MoG model plot denotes the mean of each Gaussians. Two of these were assigned to one mode, and formed a single peak pdf corresponding to one cluster.

In Table III, the classification success rates of the overall system using various numbers of Gaussians are presented. As expected, the results using four, five, and six Gaussians were virtually the same, and the result obtained using seven Gaussians was also quite satisfactory. Note that when we used PP/NEM-FCM with the proper number of clusters (three clusters), the best success rate obtained was 66%.

V. DISCUSSION AND CONCLUSION

Our goal was to realize a fully automated, unsupervised neural spike sorting system that does not require interactive input from a human supervisor, and shows high performance under low SNR conditions. In order for the performance of this system to be limited only by the performance of the action potential detector, we introduced an efficient linear projection method, PP/NEM, for feature extraction. We were able to achieve separability higher than that of PCA, and it was superior to nonlinear methods such as GTM and Sammon's mapping. Since the PP/NEM is a linear projection, it also has the advantage that the characteristics of the original data in a high dimension are maintained. The computational burden of the PP/NEM in training is somewhat higher than that of PCA, but, was much lower than the nonlinear methods and in an acceptable range. 0.12 and 1.876 s (average of ten trials)

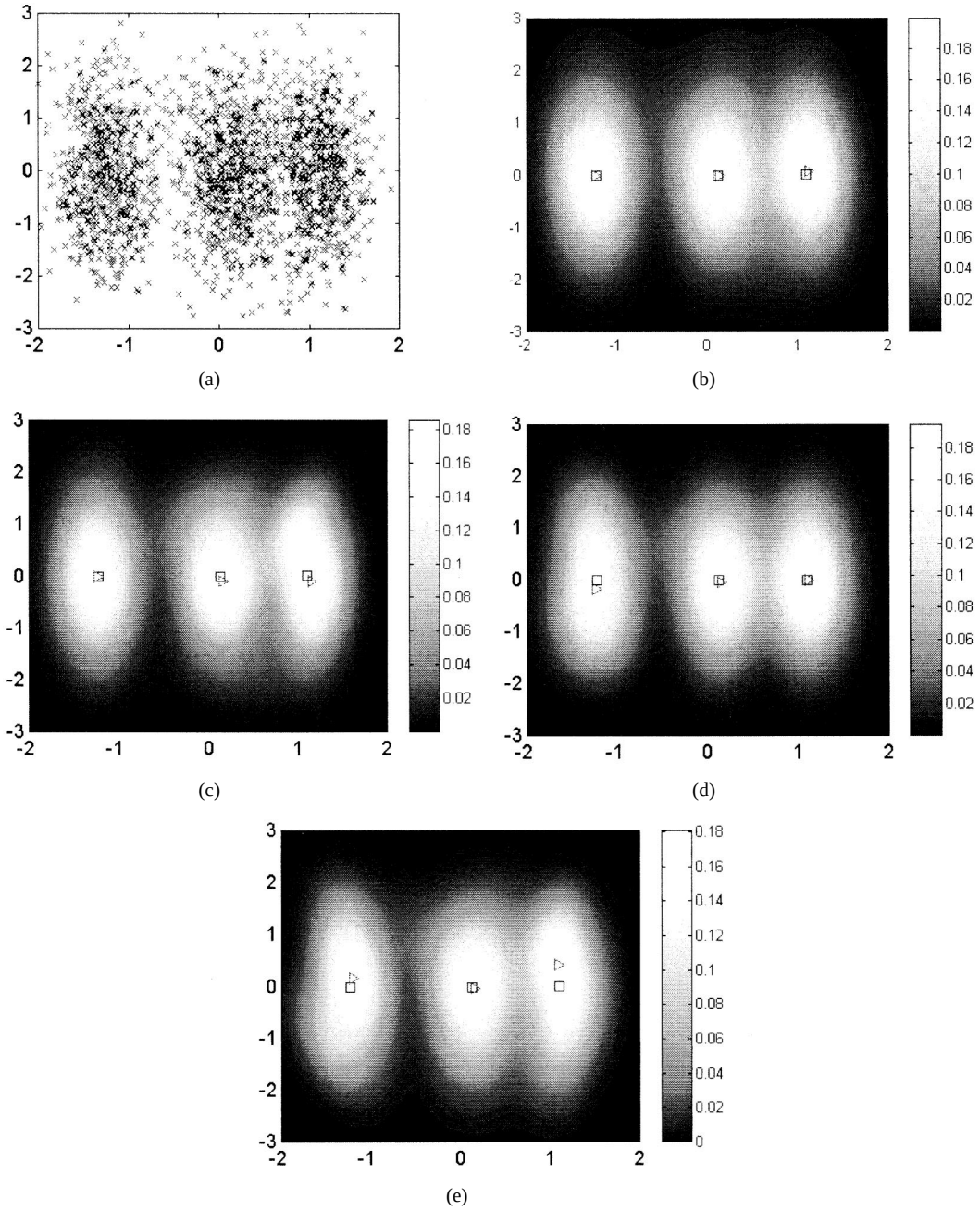


Fig. 7. (a) Another scatter plot of a projection using PP/NEM, *Aplysia* data, SNR ≈ 1.6. Identified MoG model, using (b) four Gaussians, (c) five Gaussians, (d) six Gaussians, and (e) seven Gaussians. The local maxima discovered by the algorithm of Section IV is denoted by "∇," and the true means of each unit are denoted by "$\square$."

were required to perform a projection on 2100 action potentials from three units by PCA and PP/NEM, respectively (MATLAB implementation on Pentium-III PC with 1000-MHz clock). Once the projection matrix is found, the amount of computation required for PP/NEM is exactly same as that for PCA.

In many previous studies on unsupervised spike sorting, the employed classifiers were the k-means-related methods (FCM or ISODATA) [13], [25]. We attempted to utilize the FCM, which is generally considered to be superior in that partial membership to multiple classes can be considered and cluster validity indexes are available. Nevertheless, it was impossible to obtain a reasonable classification result using FCM for the data where a manual determination of the decision boundary by a human supervisor seemed to be easy when the features were extracted by the PP/NEM. Thus, we focused on employing an unsupervised classification method that is capable of modeling shape of the distribution, and exploited the MoG-based technique. When a combination of PP/NEM and MoG was used, it was possible to obtain a high success rate of approximately 95%, for the data where the success rate using the PP/NEM-FCM was much lower (66%). Yet another advantage gained from the use of the MoG is the fact that classification result is much less sensitive to the parameter that must be predetermined. In contrast, in the case of FCM, the clustering result has a large dependence on the choice of the predetermined number of clusters. Zouridakis and Tam [25]

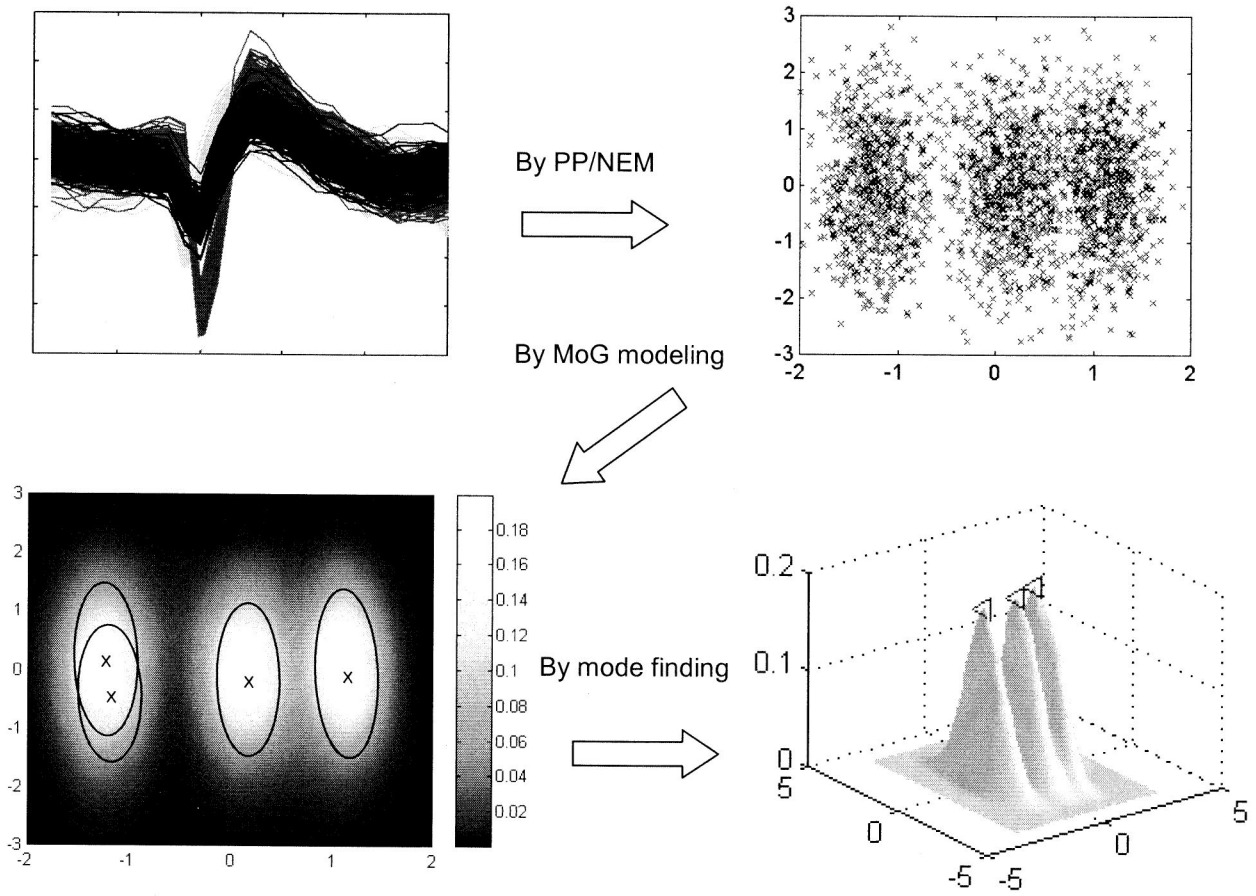


Fig. 8. Step-by-step procedure for applying our algorithm when four Gaussians are used for the pdf modeling

TABLE III
SUCCESS RATE OF PP/NEM-MoG, APPLIED TO THE DATA SHOWN IN FIG. 7(a)

K=4	K=5	K=6	K=7
95.62%	95.67 %	95.7%	94.81%

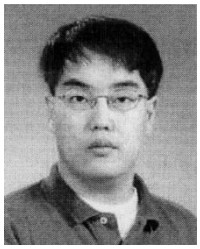
recently reported on an unsupervised spike sorting system that utilizes FCM along with a cluster validity index [20] for determining the number of clusters. They claimed that it is possible to obtain fully automated spike sorting. However, it appears that their result was obtained using very high SNR data. Actually when we used the FCM for the data obtained from low SNR recordings, neither successful clustering nor the estimation of the number of clusters using the validity index were possible.

In conclusion, a novel unsupervised neural spike sorting system that shows high performance under low SNR, and is advantageous for applications that require a fully automated system is described. Our strategy was to combine a linear projection with high separability and an unsupervised classification method that can take into account the anisotropic distribution of feature vectors and that does not require a precise parameter that must be predetermined by the user. Experimental neurophysiological studies that apply the proposed system to various neural signal recordings will be helpful to tune and improve performance of the system.

REFERENCES

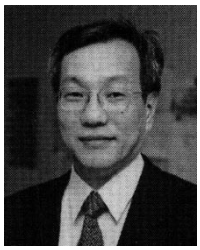
[1] F. Rieke, D. Warland, R. De Ruyter Van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code*. Cambridge, MA: The MIT Press, 1996.
[2] M. S. Lewicki, "A review of methods for spike sorting: The detection and classification of neural action potentials," *Network: Computation Neural Syst.*, vol. 9, pp. R53–R78, 1998.
[3] E. M. Schmidt, "Computer separation of multi-unit neuroelectric data: A review," *J. Neurosci. Meth.*, vol. 12, pp. 95–111, 1984.
[4] R. Chandra and L. M. Optican, "Detection, classification, and superposition resolution of action potentials in multiunit single channel recordings by an on-line real-time neural network," *IEEE Trans. Biomed. Eng.*, vol. 44, pp. 403–412, May 1997.
[5] K. H. Kim and S. J. Kim, "Neural spike sorting under nearly 0 dB signal-to-noise ratio using nonlinear energy operator and artificial neural network classifier," *IEEE Trans. Biomed. Eng.*, vol. 47, pp. 1406–1411, Oct. 2000.
[6] M. S. Fee, P. P. Mitra, and D. Kleinfeld, "Automatic sorting of multiple unit neuronal signals in the presence of anisotropic and non-Gaussian variability," *J. Neurosci. Meth.*, vol. 69, pp. 175–188, 1996.
[7] M. M. Sahani, "Variable models for neural data analysis," Ph.D. dissertation, California Inst. Technol., Pasadena, CA, 1999.
[8] K. H. Kim and S. J. Kim, "Method for action potential detection from extracellular neural signal recording with low signal-to-noise ratio," IEEE Trans. Biomed. Eng., 2003, to be published.
[9] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[10] B. C. Wheeler and W. J. Heetderks, "A comparison of techniques for classification of multiple neural signals," *IEEE Trans. Biomed. Eng.*, vol. BME-29, pp. 752–759, 1982.
[11] P. J. Huber, "Projection pursuit," *Ann. Statist.*, vol. 13, pp. 435–475, 1985.

[12] M. Girolami, A. Cichocki, and S. I. Amari, "A common neural-network model for unsupervised exploratory data analysis and independent component analysis," *IEEE Trans. Neural Network*, vol. 9, pp. 1495–1501, 1998.
[13] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 4–37, Jan. 2000.
[14] T. H. Yoon, E. J. Hwang, D. Y. Shin, S. I. Park, S. J. Oh, S. C. Jung, H. C. Shin, and S. J. Kim, "A micromachined silicon depth probe for multi-channel neural recording," *IEEE Trans. Biomed. Eng.*, vol. 47, pp. 1082–1087, Aug. 2000.
[15] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. New York: Wiley, 1996.
[16] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Network*, vol. 10, pp. 626–634, May 1999.
[17] T. W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources," *Neural Computat.*, vol. 11, pp. 417–441, 1999.
[18] D. G. Ruenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.
[19] M. S. Lewicki, "Bayesian modeling and classification of neural signals," *Neural Computat.*, vol. 6, pp. 1005–1030, 1994.
[20] X. L. Xie and G. A. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 841–847, Aug. 1991.
[21] D. A. Langan, J. W. Modestino, and J. Zhang, "Cluster validation for unsupervised stochastic model-based image segmentation," *IEEE Trans. Imag. Processing*, vol. 7, pp. 180–195, Feb. 1998.
[22] M. A. Carreira-Perpinan, "Mode-finding for mixtures of Gaussian distribution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 1318–1323, Nov. 2000.
[23] C. M. Bishop, M. Svensen, and C. K. I. Williams, "GTM: The generative topographic mapping," *Neural Computat.*, vol. 10, pp. 215–234, 1998.
[24] T. Kohonen, *Self-Organizing Maps*, 2nd ed. Berlin, Germany: Springer-Verlag, 1997.
[25] G. Zouridakis and D. C. Tam, "Identification of reliable spike templates in multi-unit extracellular recordings using fuzzy clustering," *Comp. Meth. Prog. Biomed.*, vol. 61, pp. 91–98, 2000.
[26] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.



Kyung Hwan Kim (M'02) was born in Seoul, Korea, in 1973. He received the B.S. degree from Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejon, in 1995. He received the M.S. and Ph.D. degrees from School of Electrical and Computer Engineering, Seoul National University, Seoul, in 1997, and 2001, respectively.

Since March 2001, he has been working as a member of technical staff in the Human-Computer Interaction Laboratory, Samsung Advanced Institute of Technology, Yongin, Korea. His research interests include biomedical signal processing, pattern recognition, and instrumentation with emphasis on their application to neural signals, for neuroscience research and neural prosthesis. He is also working on physiological signal processing for the implementation of affective human-computer interface.



Sung June Kim (S'79–M'84) received the B.S. degree in electronics engineering from Seoul National University, Seoul, Korea, in 1978. He received M.S. and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, NY, in 1981, and 1983, respectively. His Ph.D. degree thesis was on the photo-induced fabrication of trans-substrate microelectrode arrays based on silicon substrate and their neurophysiological applications.

From 1983 to 1989, he worked as an MTS in Bell Laboratories, Allentown, PA, and in Murray Hill, NJ. At Bell Labs, he studied design and processing of silicon VLSI, and the process and device development of optoelectronic integrated circuits (OEICs) for long-wavelength optical communications. In 1989, he returned to Korea to join the Department of Electronics Engineering and the Inter-university Semiconductor Research Center (ISRC), Seoul National University, where he is now a Full Professor in the School of Electrical Engineering. Since 2000, he has been the Director of Nano-Bioelectronics and System Research Center (NBS-ERC) which is funded by Korea Science and Engineering Foundation (KOSEF). His research interests are in areas of bioelectronics, bioinstrumentation, neural prosthesis, and optoelectronic semiconductor devices. He has published about 60 papers, and has managed many research projects in the areas mentioned. His web site is at http://helios.snu.ac.kr/.