PERGAMON

Pattern Recognition 35 (2002) 1821–1833

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Continuous mixture modeling via goodness-of-fit ridges

Stephen R. Aylward *

Department of Radiology, 129 Radiology Research Lab, MRI Building D, CB 7510, University of North Carolina, Chapel Hill, NC 27599-7510, USA

Received 14 August 2000; accepted 25 May 2001

Abstract

We present a novel method for representing "extruded" distributions. An extruded distribution is an M-dimensional manifold in the parameter space of the component distribution. Representations of that manifold are "continuous mixture models". We present a method for forming one-dimensional continuous Gaussian mixture models of sampled extruded Gaussian distributions via ridges of goodness-of-fit. Using Monte Carlo simulations and ROC analysis, we explore the utility of a variety of binning techniques and goodness-of-fit functions. We demonstrate that extruded Gaussian distributions are more accurately and consistently represented by continuous Gaussian mixture models than by finite Gaussian mixture models formed via maximum likelihood expectation maximization. © 2002 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Mixture modeling; Local maximum; Ridge traversal; Binning; Goodness-of-fit; Distribution representation; Maximum likelihood expectation maximization

1. Introduction

A critical aspect of data analysis and statistical pattern recognition is the accurate modeling of the distributions of data. When the shape, e.g., Normal, of a sampled distribution is matched by the mathematical function used to model it, e.g., Gaussian, the subsequent data analyses and classifications are accurate. When the shape of a distribution is unknown, the accurate and consistent modeling of the data is problematic.

The shapes of the distributions associated with medical images, speech, handwriting, and RGB color spaces are generally not known; the accurate and consistent modeling of these data has been difficult. We propose that these distributions are instances of a single class of shapes that we designate "extruded" Gaussian distributions [1,2].

Traditionally, finite Gaussian mixture models (FGMMs) defined via maximum likelihood expectation maximization (MLEM) have been used to represent extruded Gaussian distributions. We will show that extruded Gaussian distributions are more accurately and consistently represented by continua of means and variances; continuous Gaussian mixture models (CGMMs). The continua of means and variances of a CGMM form (possibly multidimensional) "traces" in the parameter space of a Gaussian. We will show that these traces can be extracted as multidimensional height ridges of Gaussian-goodness-of-fit functions.

CGMMs are just one type of continuous mixture model that can be formed by traces of goodness-of-fit functions. Continuous mixture models utilizing other component distributions, e.g., log–normal distributions, may be appropriate for other classification problems. Adaptation of the techniques presented in this paper to other continuous mixture models is accomplished by changing the shape of the goodness-of-fit function's expected distribution.

* Tel.: +1-919-966-9695; fax: +1-919-966-2859.
E-mail address: aylward@unc.edu (S.R. Aylward).

PII: S0031-3203(01)00133-9

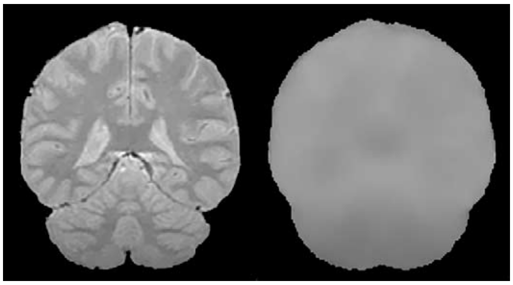


Fig. 1. (a) Proton density MR image. (b) Estimated intensity inhomogeneity.

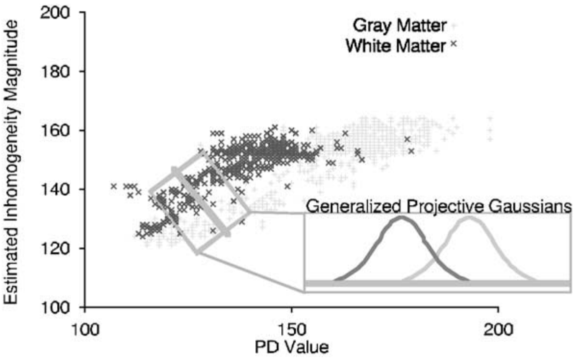


Fig. 2. Scatter plot of hand-labeled gray and white matter pixels (PD value versus estimated inhomogeneity value).

Section 2 discusses the existence of extruded Gaussian distribution in real-world data. Section 3 defines the methods by which we will assess a model's accuracy and consistency. Section 4 introduces finite and continuous Gaussian mixture modeling. Section 5 presents Gaussian-goodness-of-fit functions that respond maximally when their parameters (μ' and σ') match those of the distribution that generated the samples being tested. Section 5 also discusses how maximum curvature height ridges of Gaussian-goodness-of-fit functions define traces in the parameter space of a Gaussian, and, in turn, how those traces define CGMMs. Section 6 uses extruded Gaussian distributions, Monte Carlo simulations, and ROC analysis to compare FGMMs with CGMMs. Section 7 demonstrates the generation of an accurate representation of a tri-variate extruded Gaussian. Section 8 demonstrates the CGMM-based classification of tissues in an inhomogeneous MR image.

2. Extruded Gaussian distributions in real-world data

Extruded Gaussians occur frequently in "real-world" data. For example, they exist in the data associated with medical images, speech, and handwriting.

Within small, "localized" regions of an MR image, the intensities associated with a particular tissue type will be Gaussian distributed, yet MR intensity inhomogeneities cause the mean and variance of a tissue's localized intensities (i.e., the parameters of the localized Gaussian distributions) to change throughout the MR image. The proton density (PD) MR image in Fig. 1a was acquired and converted to byte pixel values as described in Ref. [3]. It contains an intensity inhomogeneity that is visible as a large scale dimming in the inferior cerebellum. The inhomogeneity can be quantified (Fig. 1b) by blurring the image at a scale of 15 pixels ($\sim$ 13 mm) using only those pixels having PD values between 100 and 200. These intensity limits correspond to the range of PD values associated with gray and white matter. More exact methods for measuring the inhomogeneity exist [4–6], but the stated approach is sufficient for this demonstration. A scatter plot reveals the non-linear correlation between PD value and inhomogeneity magnitude; Fig. 2 is formed from 984 hand-labeled white matter and 788 hand-labeled gray matter pixels from these images. In Fig. 2, localized collections of a tissue's intensities have Gaussian distributions, but a continuum of Gaussians is needed to represent each tissue's entire distribution; the tissue distributions are extruded Gaussians. Other research supports the existence of extruded Gaussians in X-ray CT images due to beam hardening and in SPECT images due to deficiencies in attenuation compensation.

In speech recognition, it is commonly accepted that hidden Markov models using Gaussian distributions can represent certain aspects of the speech of a single person in a controlled situation, e.g., given a fixed level of stress. Smooth warpings can be applied to the parameters of those Gaussians to transition them to new situations and speakers [7]. That is, to account for such variations in speaker and situation, multiple Gaussians are needed; the distributions resemble extruded Gaussians.

3. Assessing model accuracy and consistency

For this paper, the accuracy and consistency of the distribution models are quantified and compared based on the accuracy and consistency of the classifiers they define. That is, when a model Ψ of a class i is used to provide class conditional probability estimates $P(\underline{x}|\Psi^{(i)})$ to a classifier, the accuracy and consistency of the labelings produced by that classifier determine the accuracy and consistency of the model. Assuming equal class priors $P(\Psi^{(i)})$ and Bayesian classification, then

$$\mathrm{Label}(\underline{x}) = \underset{i=1\ldots\#\text{ of classes}}{\arg\max}\left[P(\Psi^{(i)}|\underline{x}) = \frac{P(\Psi^{(i)})P(\underline{x}|\Psi^{(i)})}{P(\underline{x})} = P(\underline{x}|\Psi^{(i)})\right]. \quad (1)$$

A classifier's labeling *accuracy* is quantified by its true positive and false positive rates. A classifier's labeling *consistency* is the standard error of those rates.

4. Mixture modeling

When the shape of a sampled distribution is not known, multiple "component" distributions can be combined to model the sampled distribution. Such modeling is called direct mixture modeling. For example, in a Gaussian mixture model the component distributions are multivariate (N-dimensional) normal densities that are parameterized by Φ.

$$F(\underline{x};\Phi)=\frac{1}{(2\pi)^{N/2}|\underline{\underline{\Sigma}}|^{(1/2)}}\mathrm{e}^{-(1/2)(\underline{x}-\underline{\mu})'\underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu})},$$

$$\text{where } \Phi=\{\underline{\mu},\underline{\underline{\Sigma}}\}. \tag{2}$$

4.1. Finite mixture modeling

If the number of components K is bounded, the mixture model is a finite mixture model Ψ. It provides a probability for a sample $\underline{x}$ via

$$P(\underline{x}|\Psi)=\sum_{j=1}^{K}\omega^{(j)}F(\underline{x},\Phi^{(j)}),\quad \text{where } 1=\sum_{j=1}^{K}\omega^{(j)}$$

$$\text{and } \Psi=\{\{\omega,\Phi\}^{(j)}\,|\,j=1\ldots K\}. \tag{3}$$

Most investigations involving mixture models use finite mixture models trained via maximum likelihood expectation maximization (MLEM). While no finite mixture model training algorithm is best in all situations, MLEM is easy to implement and provides several desirable convergence properties: monotonic convergence and nearly quadratic convergence rates [8–11].

MLEM, however, is an approximate gradient ascent algorithm, and it is subject to non-optimal local maxima. While MLEM is relatively robust to these non-optimal maxima [9,11,12], it will be shown that the component parameterizations produced via MLEM can vary greatly and can be far from optimal given different sets of samples from the same distribution; finite mixture models developed using MLEM do not perform consistently. This inconsistency is aggravated by the reliance on the user to specify the number of components. While much research has focused on algorithms for automatically determining an appropriate number of components for a given problem, a generally applicable approach has not been found [11,13]. A finite mixture model's expected accuracy does not vary monotonically as a function of the number of components. Additionally, the non-optimal maxima associated with likelihood maximization can lead to poorly utilized components; the effective number of components in a finite mixture model may be less than the user-specified number of components. Since extruded Gaussian distributions are comprised of an infinite number of components, determining an appropriate finite number of components to approximate them can be especially difficult.

4.2. Continuous mixture modeling

A continuous mixture model consists of an uncountably infinite number of components whose parameters Ψ span N_t traces $\boldsymbol{T}^{(j)}$ through the parameter space of its components, i.e., the domain of Φ. For this paper, we define that a continuous mixture model provides a likelihood estimate via

$$P(\underline{x}|\Psi)=\max_{\{\omega,\Phi\}\in\Psi}(\omega F(\underline{x};\Phi)),\quad \text{where}$$

$$\Psi=\left\{\{\omega,\Phi\}\,\middle|\,\begin{array}{l}\exists j\in 1\ldots N_t \text{ s.t. } \Phi\in\boldsymbol{T}^{(j)}\\ \text{and } \omega=P(\Phi)\end{array}\right\}. \tag{4}$$

This equation follows the simplifying assumptions made by Dempster et al. for MLEM [8]. That is, we assume that since the underlying distribution is assumed to be a mixture, each multivariate sample $\underline{\mathrm{x}}$ is, in fact, generated by just one of the mixture's components; the generating component is determined via maximum likelihood; and the generating component estimates the sample's likelihood, $P(\underline{x}|\Psi)$. In actuality, the probability at a point is an integration of the weighted probabilities provided by the continuum of component distributions. We have, however, found the above likelihood implementation to be expedient and sufficient for classification.

The function $F(\underline{x};\phi)$ can be interpreted as providing a trace-point conditional sample probability, and ω as providing a trace point a priori probability. Eq. (3) can therefore be rewritten as

$$P(\underline{x}|\Psi)=\max_{\{\Phi\}\in\boldsymbol{T}^{(j)}|j=1\ldots N_t}(P(\Phi)P(\underline{x}|\Phi)). \tag{5}$$

The focus of this paper is the definition of the traces $\boldsymbol{T}^{(j)}$ via height ridges of goodness-of-fit functions. CGMMs parameterized via such traces can accurately and consistently model the continua of means and variances that form an extruded Gaussian distribution. For this paper, analysis is limited to extruded Gaussians having one-dimensional traces; that is, CGMMs defined via one-dimensional height ridges in Gaussian-goodness-of-fit space.

5. Height ridges in goodness-of-fit space

A method has already been developed for representing objects in images as continua of centers and widths, i.e.,

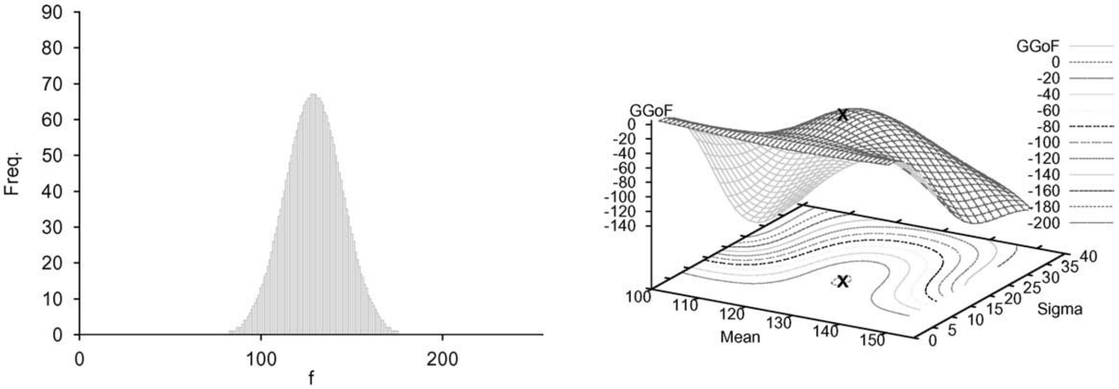


Fig. 3. (a) 2700 samples distributed so as to best match a Gaussian distribution. (b) The GGoF space of the samples in Fig. 3a. (Gaussian's actual parameter values are indicated by the X.)

medialness cores [14]. In medialness cores, the centers and widths of an object are approximated using multi-dimensional height ridges of a medialness function. Medialness cores have been proven to be invariant to rotation, translation, intensity, and scale and insensitive to a wide variety of image and boundary noise [14]. To apply medialness core methods to the representation of distributions, we substitute goodness-of-fit functions for medialness functions since the relevant property to be measured is sample density rather than boundariness. In certain situations, goodness-of-fit functions can be thought of as localizers of a distribution's means and variances instead of an object's centers and widths.

5.1. Univariate goodness-of-fit

One class of goodness-of-fit functions is the univariate chi-squared measure. This class includes Pearson's statistic χ^2_{P}, Read and Cressie's power divergent statistic $\chi^2_{\mathrm{R\&C}}$, and the log likelihood ratio χ^2_{LLR} [15]. These univariate statistics are binned, omnibus, and smooth statistical measures. That is, the expected distribution E and the observed samples O must be binned, the shape of the expected distributions is arbitrary, and the change in the measure's value will be smooth given small changes in the parameters of the expected distribution. We have modified these GGoF functions so as to be maximal when the parameters of the expected distribution best match the parameters of the population from which the samples originated. This is achieved by subtracting the standard goodness-of-fit functions from $\chi^2_{6-1}(\alpha = 0.99) = 15.09$ and then normalizing by that value. Since our goal in this paper is to develop mixture models using Gaussian components, the expected distribution is defined as a univariate Gaussian. The parameters of these functions are therefore μ' and σ', the mean and standard deviation to be tested. This paper uses six bins $B = 6$ centered at μ' and clipped so as to capture samples within $\pm 1.645\sigma'$ of μ'. We refer to these functions as Gaussian-goodness-of-fit (GGoF) functions.

$$\chi^2_{\mathrm{P}} = \left(15.09 - \sum_{i=1}^{B} \frac{(O^{(i)} - E^{(i)})^2}{E^{(i)}}\right) \Big/ 15.09, \tag{6}$$

$$\chi^2_{\mathrm{R\&C}} = \left(15.09 - \frac{9}{5}\sum_{i=1}^{B} O^{(i)} \left(\left(\frac{O^{(i)}}{E^{(i)}}\right)^{2/3} - 1\right)\right) \Big/ 15.09, \tag{7}$$

$$\chi^2_{\mathrm{LLR}} = \left(15.09 - 2\sum_{i=1}^{B} O^{(i)} \ln\left(\frac{O^{(i)}}{E^{(i)}}\right)\right) \Big/ 15.09. \tag{8}$$

The value of these goodness-of-fit functions will be greater than zero for 99% of the sets of samples that originate from a population parameterized by μ' and σ'.

The behavior of these GGoF functions for the data in Fig. 3a is illustrated in Fig. 3b. Specifically, Fig. 3a is a histogram of 2700 samples distributed so as to best match a Gaussian distribution. When the χ^2_{LLR} GGoF function is evaluated for a range of means μ' and standard deviations σ' given the data in Fig. 3a, then the GGoF function values shown in Fig. 3b result. The maximum of the GGoF function corresponds to the parameters of the Gaussian that the samples represent. This maximum is a zero-dimensional height ridge of GGoF; it accurately represents the zero-dimensional trace of the sampled (extruded) Gaussian.

Although these functions are smooth, we have found that the binning technique affects their realized smoothness and hence the consistency with which a data set's optimal local maximum can be found. We have studied four

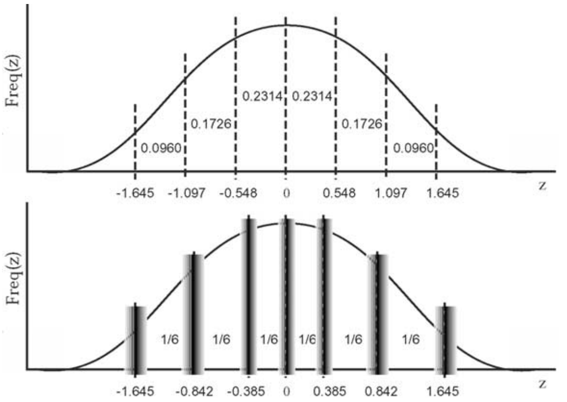


Fig. 4. (a) Equirange binning. (b) Overlapped-equiprobable binning; shades of gray indicate the weighting of samples near bin edges.

different binning techniques: equirange, equiprobable, overlapped-equirange, and overlapped-equiprobable [1]. Equirange binning refers to the use of non-overlapping bins which span equal sized ranges of values (Fig. 4a). Equirange binning is performed independent of the expected distribution. In equiprobable binning, each bin spans a value range such that the expected number of samples within each bin is equal (Fig. 4b). Equiprobable binning is therefore driven by the expected distribution. Overlapped binning techniques weight the allocation of a sample to a bin based on its distance from the bins' edges. Such weights, for example, can be based on a sigmoidal function as in Eq. (9). The variable y represents the normalized distance from a sample to the bin's edge. That normalization is with respect to half of the shorter range of the two bins that form the edge. The parameter Ω controls the amount of overlap. For all overlapped-binning work presented in this paper, $\Omega = 0.75$.

$$W(y, \Omega) = \frac{1}{1 + \mathrm{e}^{-(y/\ln(1+0.1*\Omega))}}. \tag{9}$$

The accuracy and consistency of the local maxima of the χ^2_{P}, $\chi^2_{\mathrm{R\&C}}$, and χ^2_{LLR} GGoF functions were evaluated using 96 Monte Carlo simulations. Each simulation consisted of 5000 runs. The simulations considered four different training set sizes (20, 40, 80, 160 samples) from two distributions (a Gaussian with $\underline{\mu} = 128$ and $\sigma = 16$ and a log–normal distribution using a $\overline{\text{log}}$ base of 1.6) and the four mentioned binning techniques. For each Monte Carlo run, the local maximum of the GGoF function was found via gradient ascent through the (μ', σ') parameter space. The starting points for gradient ascent were selected from a 2D Gaussian distribution centered at each population's ideal parameter values (μ, σ) having a standard deviation of 5% of those values. Figs. 5a and b illustrate the estimated parameter values (μ', σ') of 5000 local maxima in GGoF from two of the Monte Carlo simulations.

These scatterplots reveal improved accuracy when additional samples are used to calculate the GGoF values. Although the resolution of non-optimal local maxima occurs even when 160 samples are used, the GGoF values associated with the non-optimal points are generally negative and often are below -10. Such low values immediately suggest that a suboptimal local maxima has been found, and appropriate actions can then be taken to reject such maxima.

The accuracy of a local GGoF maximum is defined as the difference between the GGoF parameters (μ', σ') of that maximum and the population's actual parameters (μ, σ). The consistency of the maxima from a Monte Carlo simulation is the standard error associated with each parameter μ' and σ' of the maxima.

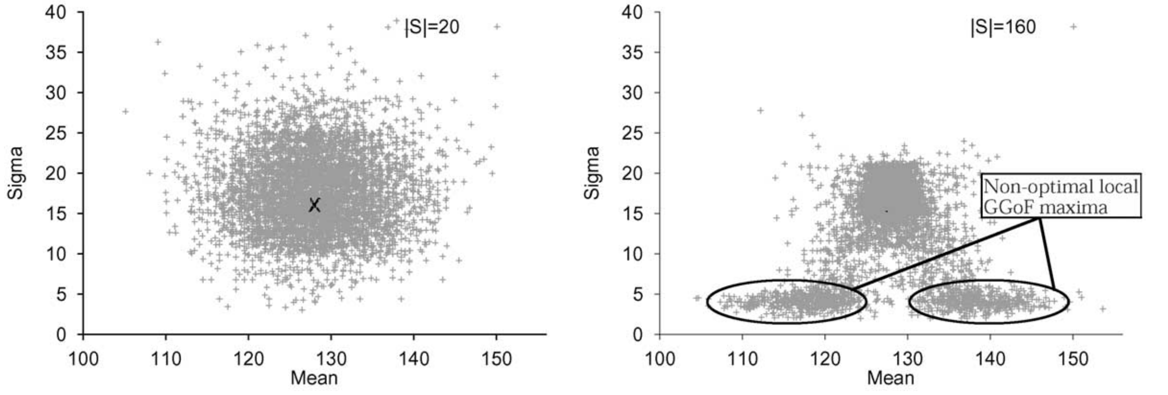


Fig. 5. (a) 5000 maxima from χ^2_{P} with equirange binning using 20 samples. (b) 5000 maxima from χ^2_{P} with equirange binning using 160 samples.

Conclusions drawn from the 96 simulations include that (1) the binning method has more influence on accuracy and consistency than the GGoF function; (2) the accuracy and consistency of the estimates of μ do not vary significantly as a function of the number of samples, the GGoF function, or the binning technique; and (3) χ^2_{LLR} with overlapped-equiprobable binning provides the most accurate and consistent estimates σ. As a result, χ^2_{LLR} with overlapped-equiprobable binning is used for all subsequent GGoF ridge calculations.

5.2. *Multivariate GGoF via GGoF ridge tangents and normals*

To calculate multivariate GGoF values, the multivariate data within a bounding box about a given $\underline{\mu}'$ are converted to multiple univariate distributions via projection onto a set of basis directions normal to the trace $\boldsymbol{T}$. The expected variance associated with each of those projections determines the size of the bounding box; the expected variance in each direction may differ. The multivariate GGoF value is the average χ^2_{LLR} value from each of those projections. We hypothesize that for an extruded Gaussian's trace, neighboring trace points capture a distribution's variance in the trace's tangent direction, so each of our GGoF ridge points needs only to capture variance normal to the ridge.

To estimate a ridge's normal (and tangent) directions as well as the magnitudes of expected variance in each of those directions, our algorithm extends the height ridge work of David Eberly and the concept of deriving geometric measures from local statistical moments developed by Yoo [16]. Specifically, we suggest that eigenvectors of the local data's covariance matrix $\underline{\underline{\Sigma}}^{(L)}$ approximate the normal (and tangent) directions of the GGoF height ridge, and the eigenvalues define expected variance ratios for each of the normal directions (Fig. 6).

$\underline{\underline{\Sigma}}^{(L)}$ is a function of two variables, a mean $\underline{\mu}'$ and a neighborhood size s$'$. $\underline{\underline{\Sigma}}^{(L)}$ is measured using a Gaussian weighting $G(*)$ of the samples S about $\underline{\mu}'$ so as to change smoothly given small changes in $\underline{\mu}'$ or s$'$.

$$\underline{\underline{\Sigma}}^{(L)}_{ij}(\underline{\mu}',s')=\frac{\sum_{\underline{z}\in S}\ G(\underline{z}|\underline{\mu}',3s')(\underline{z}_i-\underline{\mu}'_i)(\underline{z}_j-\underline{\mu}'_j)}{\sum_{\underline{y}\in S}\ G(\underline{y}|\underline{\mu}',3s')}. \tag{10}$$

Define λ_i for $i=1,\ldots,N$ as the descending ordered eigenvalues of $\underline{\underline{\Sigma}}^{(L)}$ and $\underline{\mathrm{v}}_i$ as their corresponding eigenvectors. In the absence of additional information, it can be assumed that the maximum eigenvalued eigenvector $\underline{\mathrm{v}}_1$ approximates the GGoF height ridge's tangent direction. The remaining eigenvectors approximate the normal directions. The magnitude of expected variance in each of the normal directions is specified via eigenvalue ratios; the magnitude of the expected variance in the eigendirection $\underline{\mathrm{v}}_i$ for $i=2,\ldots,N$, is

$$(\sigma_i')^2=(s')^2\lambda_i/\lambda_2. \tag{11}$$

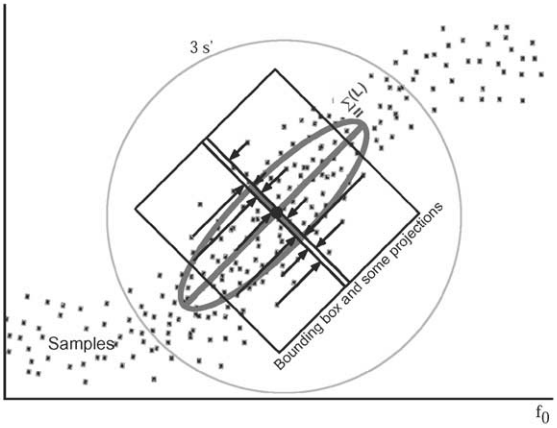

f1
f0
3 s'
$\underline{\underline{\Sigma}}$(L)
Samples
Bounding box and some projections

Fig. 6. The local data covariance matrix specifies the directions of projection and their expected variances for calculating a multivariate GGoF value.

By using expected variance ratios, extruded Gaussians composed of anisotropic, yet Gaussian shaped, cross-sections can be approximated by CGMMs.

To help understand the $N+1$ dimensional GGoF "space" $(\underline{\mu}',s')$ of an N dimensional distribution, slices through the three-dimensional GGoF space of a two-dimensional distribution in the feature space (f_0,f_1) can be calculated. Consider the scattergram shown in Fig. 7a. Those 900 samples were generated from a simulated extruded Gaussian designated Class A. Class A is defined by three approximating cubic B-splines and four isotropic control Gaussians. Each spline governs one of the three parameters of the Gaussians, i.e., $\mu_{f_0},\mu_{f_1},\sigma$. To generate a sample, a parametric value t is chosen from the uniform distribution $\boldsymbol{U}[0,1]$. The three splines are evaluated at that t value. An isotropic Gaussian distribution is thus defined, and from that distribution the sample is then generated. A one-dimensional GGoF trace exists along the extent of the distribution. Figs. 7b and c are volume renderings of GGoF space of the scattergram in Fig. 7a. The sliced view (Fig. 7c) clearly illustrates the track of maxima (height ridge/bright "core") forming a one-dimensional path through GGoF space.

5.3. *Goodness-of-fit ridge extraction*

Maximum-curvature height-ridge extraction techniques are applied to GGoF spaces to extract CGMM representations of extruded Gaussian distributions. The three steps involved in maximum-curvature height-ridge extraction are ridge stimulation, traversal, and traversal

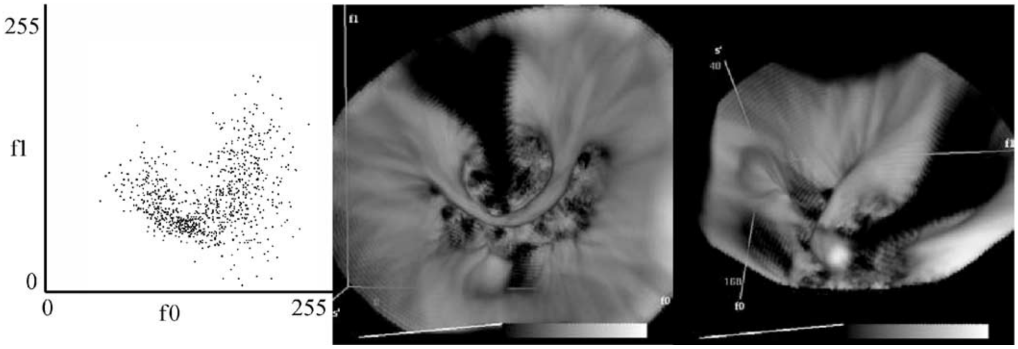


Fig. 7. Volume renderings of the GGoF space of the distribution in 8a. (a) "top–down" view of GGoF space; Viewed from large s' through to small s'. (b) Slicing GGoF space through s' and rotating the point of view, exposes the distribution's 1D GGoF ridge (bright "core").

termination. Other ridge definitions such as maxima of level curve curvature ridges or watershed ridges can be applied. The traces of an N-dimensional extruded Gaussian distribution can be of dimensionality 1 to $N-1$. While this paper concentrates on the extraction of one-dimensional GGoF ridges to approximate one-dimensional traces, the extraction technique can be generalized to higher dimensional traces.

Ridge stimulation. A ridge stimulation point has two components, $\underline{\mu}^0$ and s^0. A "stimulating" FGMM is used to specify $\underline{\mu}^0$. Specifically, the user must select the number of stimulating FGMM components to use (i.e., K), the data are then modeled using FGMM, and the component mean which is nearest (measured via Euclidean distance) to two other component means is chosen as $\underline{\mu}^0$. Consequently, $\underline{\mu}^0$ will generally be located within a dense region of a sampled extruded Gaussian. If multiple ridges are needed to represent a distribution, the remaining component means may be used. The number of stimulating FGMM components used appears to be non-critical; unless otherwise noted, all CGMMs developed in this paper were the stimulated using FGMMs with 7 components.

Specifying s^0 reduces to determining an initial neighborhood size for calculating $\underline{\underline{\Sigma}}^{(L)}$ at $\underline{\mu}^0$. By assuming that the trace tangent at $\underline{\mu}^0$ is well approximated by the maximum eigenvalued eigenvector of $\underline{\underline{\Sigma}}^{(L)}$, s^0 is the square root of the second largest eigenvalue. For this paper, the initial neighborhood size is set equal to the distance between $\underline{\mu}^0$ and its closest neighboring stimulating FGMM mean. For the data in Fig. 7a, $\underline{\mu}^0 = (163.66, 80.08)$ and $s^0 = 17.94$.

Ridge traversal. The ridge normals are approximated by the non-tangent eigenvectors of $\underline{\underline{\Sigma}}^{(L)}$ and a unit vector which points strictly in the s' direction. These directions define a hyperplane in GGoF space through which the local ridge segment passes. When this normal plane is slightly shifted in the local ridge tangent direction, a gradient ascent with respect to the GGoF values within that plane leads to a new ridge point (Fig. 8). For this paper, a step size of 0.1 feature space units is used to shift the normal plane, gradient ascent within that shifted plane is performed using Brent's line search method [17], and gradient ascent terminates when the gradient's projection onto the plane is less than 0.1% of its total magnitude. The point in the plane at which gradient ascent terminates is the new ridge point. The new point's tangent direction is approximated by the eigenvector of its local data's covariance matrix that has the maximum-magnitude dot product with the previous ridge point's tangent eigenvector. If the sign of the dot product is negative, the new tangent vector is negated to maintain the direction of traversal. This process is repeated until a traversal termination criterion is met.

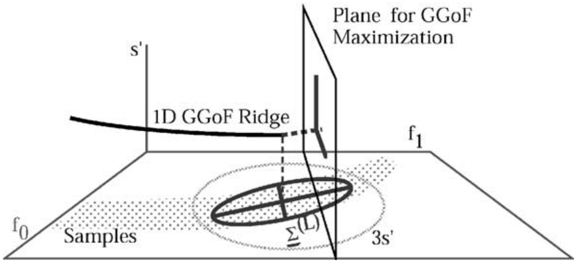


Fig. 8. The local data covariance matrix specifies the ridge's normal and tangent directions.

Ridge traversal termination and recovery. Trace traversal terminates when a "well fitting" Gaussian cannot be found. Empirical evidence suggests that encountering a GGoF value of -10 or less is a reasonable stopping criterion. This criterion was used to terminate the traversal of every trace presented in this paper.

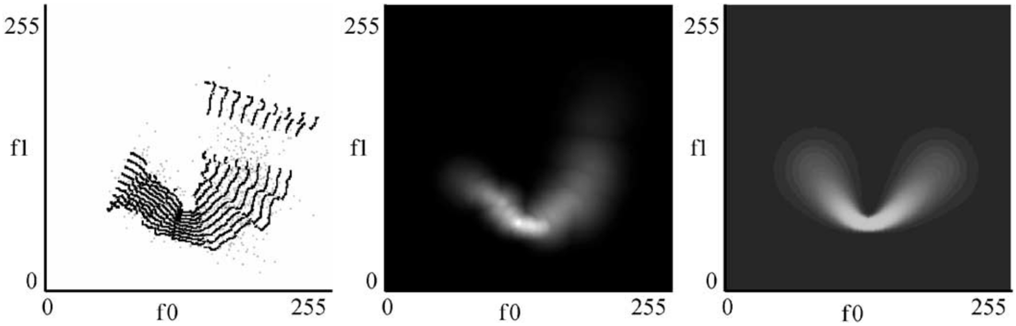


Fig. 9. (a) Isocontours of the variance estimates. (b) CGMM estimated probability density function of Class A. (c) Actual probability density function of Class A.

The rate of change of the trace is used to identify suspect trace points and prevent their inclusion into the trace without causing termination of the traversal process. Such points are "stepped over" using the tangents of the previous valid trace point.

The one-dimensional GGoF trace of the data in Fig. 7a is shown in Fig. 9a. The effect of stepping over is visible as a break in the trace. To visualize the variance estimates in the normal directions provided by the trace, the $0, \pm 0.5, \pm 1, \pm 1.5$, and ± 2 s' points along the normal at each trace point have been plotted.

Higher dimensional ridges. To generalize this technique to higher dimensions, multiple-tangent directions must be tracked and traversed so as to form a multidimensional mesh in the components' parameter space. Our work indicates that no other algorithmic modifications are required. Again the eigenvectors and eigenvalues of the local data covariance matrices are used. As the dimensionality of the ridge approaches the dimensionality of the data, other techniques for ridge traversal, discussed in [1], provide faster performance. The next section details the conversion of a GGoF trace to a CGMM.

5.4. CGMMs via GGoF traces

As defined in Eq. (4), two values, $P(\underline{x}|\phi)$ and $P(\phi)$, are required at each trace point ϕ to define a CGMM Ψ. To calculate $P(\underline{x}|\phi)$, a trace point covariance matrix $\underline{\underline{\Sigma}}(\phi)$ must be defined. The eigenvectors and eigenvalues of $\underline{\underline{\Sigma}}(\phi)$ are defined by (1) the $N-1$ approximate normal directions and expected variances which were used to calculate ϕ's GGoF value (Section 5.2) and (2) the approximate tangent direction, which is assigned a variance equal to the maximum expected variance in a normal direction.

A trace point's a priori probability $P(\phi)$ is defined as the portion of samples it is expected to represent. The number of samples that will be represented by a trace point can be extrapolated based on the number of observed samples within one standard deviation of that point.

The CGMM defined via the GGoF trace depicted in Fig. 9a produces the probability density function depicted Fig. 9b. Although the GGoF trace extended beyond the distribution, the low prior probabilities $P(\phi)$ associated with those points reduce their effect on the estimated density function. The estimated density function should be compared with the population's actual density function that is shown in Fig. 9c. There appears to be good correspondence. The next section focuses on quantifying that correspondence.

6. CGMM's accuracy and consistency

To determine the accuracy and consistency of a classifier and thus the accuracy and consistency of the distribution models it uses (Section 2), Monte Carlo simulations and ROC analyses are performed. This section begins by presenting an example classification result.

6.1. Example results

The classification problems used to evaluate the models make use of Class A, defined in Section 5.3. A competing class, Class B, is defined as an isotropic Gaussian with $\underline{\mu} = (128, 128)$ and $\sigma = 36$. Given 900 training samples from Class B, the stimulation point $\underline{\mu}^0 = (160.37, 123.30)$ and $s^0 = 17.94$ is automatically chosen. The resulting trace point conditional isoprobability curves overlaid onto the training data scattergram are shown in Fig. 10a. Fig. 10b is the density function estimated by that CGMM ($N_{\mathrm{T}} = 1$). Using the Class A and Class B CGMMs, every point in feature space can be assigned a label and an image can be developed which reflects those labelings by different shades of gray. Fig. 11a is such an image with the optimal decision bounds between the classes overlaid.

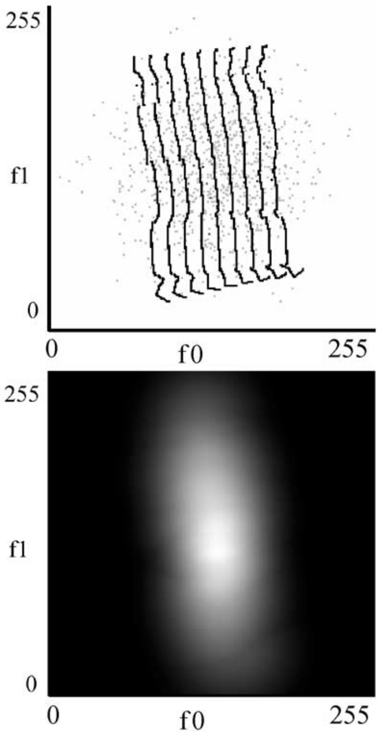


Fig. 10. (a) Isoprobability curves normal to the trace of Class B's CGMM. (b) Estimate of Class B's density function produced by CGMM.

The CGMMs of Classes A and B provide accurate labelings for most of feature space. To improve the CGMM's labelings, multiple traces can be used. While generally containing redundant information, additional traces do refine a CGMM. CGMMs using $N_T = 2$, 4, and 7 traces per class (called CGMM02, CGMM04, and CGMM07, respectively) produce the labelings shown in Figs. 11c, e, and g, respectively. FGMMs using $K = 1$, 2, 4, and 7 components per class (called FGMM01, FGMM02, FGMM04, and FGMM07, respectively) produce the labelings shown in Figs. 11b, d, f, and g respectively. Allocation to each trace/component is indicated by different shades of gray; light grays indicate allocation to Class A. The presence of non-optimal FGMM maxima is clear for FGMM07; one Class A component represents a sliver through feature space. That component is being poorly utilized, and its use does not correspond to the underlying distribution.

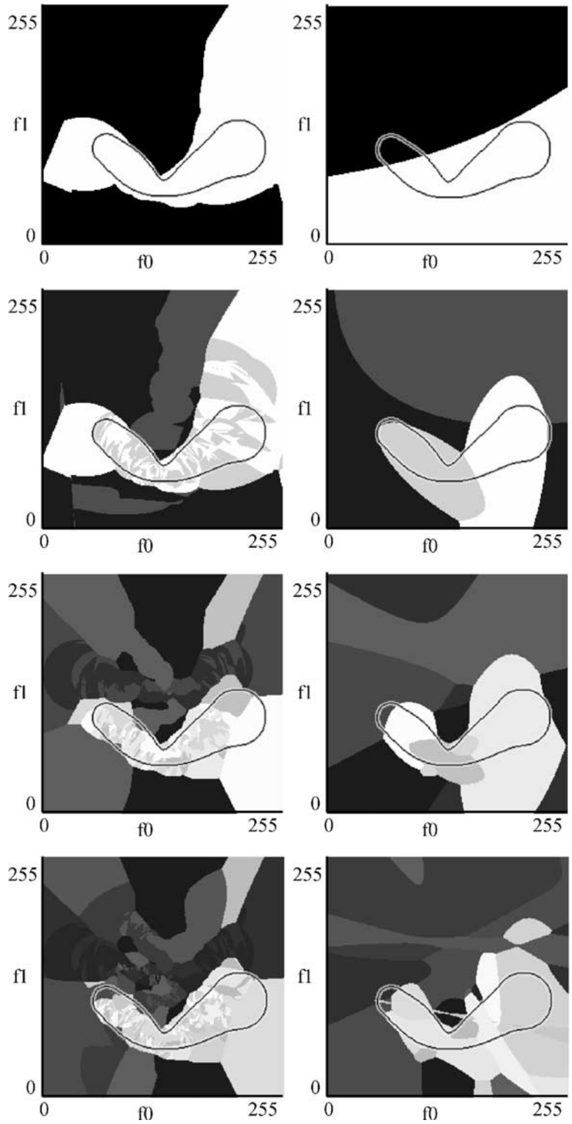


Fig. 11. Labeling of feature space produced using (a) CGMM ($N_T = 1$) (b) FGMM01s (c) CGMM02s (d) FGMM02s (e) CGMM04s (f) FGMM04s (g) CGMM07s (h) FGMM07s. Different shades of gray correspond to pixels' allocation to different traces/components; Light gray shades indicate assignment to Class A.

Given 2700 testing samples from each class, the Class A true positive rates (TPRs) and false positive rates (FPRs) in Table 1, Run 1 are produced. Compared to FGMM07s, CGMM07s provide a 718% decrease in the FPR with a less than 11% decrease in the TPR! To determine if these results were anomalous, new models were developed and tested using different samples from Classes A and B. Those results are summarized in Table 1, Run 2. CGMM07 again produced a low FPR, but the differences are less dramatic. Run 1's extreme performance resulted in an exceptionally good model in a central portion of feature space where the class samples were particularly dense.

While no conclusions should be drawn from these two runs, the results are quite encouraging. Not only does CGMM07 provide the lowest FPR values and

Table 1
Class B TPRs and FBRs from two different sets of training and testing data

	Run1		Run2	
	FPR	TPR	FPR	TPR
CGMM01	0.3233	0.8859	0.2281	0.6681
CGMM02	0.3215	0.8859	0.2178	0.7874
CGMM04	0.2604	0.8367	0.2200	0.8204
CGMM07	*0.0385*	*0.8237*	*0.2318*	*0.8485*
FGMM01	0.2933	0.8415	0.2878	0.8659
FGMM02	0.3259	0.9196	0.3185	0.9307
FGMM04	0.3315	0.9259	0.3218	0.9400
FGMM07	0.3152	0.9130	0.3067	0.9141

Table 2
Average TPR/FPR values and their standard error ranges

	Average		Standard error	
	FPR	TPR	FPR	TPR
CGMM01	0.2002	0.7181	0.0057576	0.0165489
CGMM02	0.2437	0.8192	0.0033732	0.0070245
CGMM04	0.2702	0.8658	0.0025880	0.0032410
CGMM07	0.2873	0.8862	0.0020565	0.0019929
FGMM01	0.2779	0.8364	0.0009231	0.0009339
FGMM02	0.2419	0.8660	0.0010374	0.0009371
FGMM04	0.2216	0.8495	0.0011087	0.0014111
FGMM07	0.1934	0.7990	0.0027022	0.0084882

competitive TPR values, but there is also an ordered progression in the TPR and FPR values for CGMM as the number of traces used is increased. For FGMM, the use of additional components does not always increase performance.

6.2. Monte Carlo results

To gain an understanding of the expected consistency with which CGMMs model extruded Gaussians, Monte Carlo simulations involving Classes A and B were performed. Initial simulations revealed that even after 5000 repetitions of the modeling and testing task of Section 6.1, classifiers using FGMMs demonstrated extremely poor consistency. So as to compare CGMMs with FGMMs on a problem for which FGMMs provide consistent performance, the Monte Carlo experiments were simplified by limiting their analysis to the FGMMs and CGMMs of the extruded Gaussian, Class A. Each classifier was provided with an exact model of Class B. Given 100 Monte Carlo runs involving 900 Class A training samples and 2700 Class A and 2700 Class B testing samples yielded the average TPRs, FPRs, and standard error ranges shown in Table 2.

These results are significant: (1) These results reveal that this is a problem on which FGMM performs well (i.e., consistently), therefore CGMM is being compared with FGMM on a problem for which FGMM performs well. (2) Both modeling techniques demonstrate an ordered progression in consistency based on their hyperparameter, i.e., number of components or number of traces. FGMM's consistency, however, monotonically declines as additional components are used. CGMM's consistency monotonically improves as additional traces are used. CGMM07 is shown to offer very competitive consistency. ROC analysis is needed to compare the accuracy of these classifiers.

6.3. ROC analysis

By changing the a priori probability (observer bias) associated with Class B while keeping each class model and the testing data fixed, a continuum of FPR and TPR values are defined. Qualitatively, these curves have very similar shapes. Using these curves, three measures can be made to quantitatively compare the classifiers' accuracy: the area under each curve; the maximum probability of generating a correct answer for each curve, i.e., $\max - P(C) = \max(\mathrm{TPR} + (1 - \mathrm{FPR}))$; and the TPR values of each curve at fixed FPR values [18]. Table 3 summarizes these measures for the ROC curves.

Table 3
Results of measures made on ROC curves

	Area of ROC	$\max - P(C)$	TPR @ FPR = 0.1	TPR @ FPR = 0.15	TPR @ FPR = 0.2
CGMM07	0.8752	1.5893	0.6160	0.7068	0.7741
FGMM01	0.8443	1.5530	0.5688	0.6704	0.7337
FGMM02	0.8665	1.6048	0.5889	0.6961	0.7844
FGMM04	0.8765	1.6126	0.6019	0.7166	0.7945
FGMM07	0.8793	1.6159	0.6047	0.7155	0.7935

Table 4
Probit's d' value for ROC curves based on Monte Carlo averages (Table 2)

	d'		d'
CGMM01	1.418	FGMM01	1.569
CGMM02	1.607	FGMM02	1.808
CGMM04	1.719	FGMM03	1.801
CGMM07	1.768	FGMM04	1.801
CGMM14	1.810	FGMM07	1.704

The area under the CGMM07 curve is comparable to that of FGMM04 and only slightly less than FGMM07. CGMM07 provides performance similar to FGMM02, but well below FGMM04 and FGMM07. As demonstrated in both experiments of Section 6.1, CGMM07 provides the best TPR value for the smallest FPR tested, i.e., FPR $= 0.1$. This ROC analysis, however, is based on a single instance of these modeling techniques and does not reveal expected accuracy.

To determine the expected accuracy of CGMMs and FGMMs on the Class A versus Class B problem, the Monte Carlo averaged TPR and FPR values reported in Section 6.2 are used. Specifically, the ROC curves passing through each classifier's Monte Carlo averaged TPR and FPR values can be explicitly calculated under the assumption that the class distributions are unit variance Gaussians. While that assumption is strictly incorrect for Class A, a Gaussian is a first order approximation to Class A's and B's actual distributions. The significant measure produced from this ROC analysis is the probit measure d', the spread of the means [18]. More accurate models have larger d' values. Table 4 lists the relevant d' values.

These values indicate that as additional cores are used, CGMMs can be expected to asymptotically outperform the best performing FGMM when representing Class A, an extruded Gaussian. That is, under first order assumptions for Classes A and B: (1) the area under CGMM14's ROC curve will be larger, (2) CGMM14's maximum probability of being correct will be higher, and (3) CGMM14 will provide a better TPR for every FPR value compared to the best performing FGMM, i.e., FGMM02.

In summary, every one of the experiments performed suggests that for low FPRs, CGMMs composed of a sufficient number of GGoF traces can be expected to provide better TPRs than any FGMM via MLEM. The next section demonstrates the application of CGMMs to a higher dimensional extruded Gaussian.

7. Trivariate extruded Gaussians and CGMMs

This section presents a trivariate distribution and shows its CGMM representation. This increase in the dimensionality of feature space allows anisotropic control Gaussians to be used to define a spline to generate extruded Gaussians having elliptical cross-sections (similar spline method was used in Section 5.2). This additional complexity illustrates the benefit of using expected variance ratios (Eq. (11)). Because of the higher dimensionality of feature space, 9000 samples are used to represent the population (Fig. 12a).

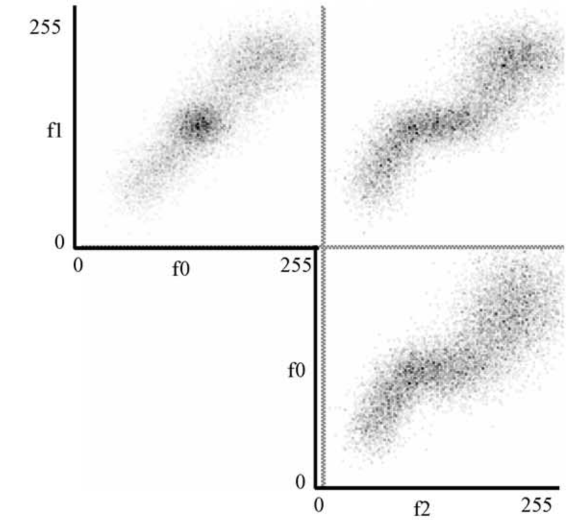
Using a stimulating FGMM with 7 components, a stimulation point at $\underline{\mu}^0 = (96.20, 98.94, 66.04)$ and $s^0 = 12.07$ is automatically generated. A one-dimensional GGoF ridge spanning approximately 192 feature space volume elements is automatically extracted. The central skeleton of the distribution is well tracked by the GGoF ridge. The GGoF ridge throughout the majority of the distribution's extent accurately estimates the local scale. A volume visualization of the estimated density function provided by the CGMM is illustrated in Fig. 12b. By slicing through this distribution at $f_2 = 64$, the fact that the model captures the anisotropic variance of the population is demonstrated (Fig. 12c). The values of a GGoF ridge point in that region match the expected variance ratios. The majority ($\sim$ 95%) of the ridge points evaluated demonstrated similarly accurate variance ratios.

In summary, the GGoF ridge approximates the trace of an anisotropic trivariate extruded Gaussian. A one-dimensional height ridge is tracked in the four-dimensional GGoF space. A CGMM is defined. A probability density function is estimated. No user interaction is required.

8. Inhomogeneous magnetic resonance images

This section demonstrates the efficacy of CGMMs using GGoF traces for medical image data. Using the hand-labeled samples shown in Fig. 2, four GGoF traces can be automatically extracted to represent each class. Using these CGMMs, all of the points in the image can be labeled as either gray or white matter. While there will be errors since other tissues are present, the results are very promising; the gray matter mask formed is given in Fig. 13a. The qualitative best FGMM was achieved using four components. FGMM04's gray matter mask is shown in Fig. 13b.

The differences between the CGMM and the FGMM masks are extremely small. The lack of a gold standard for this data prevents a quantitative comparison. These results are significant, however, in that they indicate that (1) CGMMs are a viable alternative for extruded Gaussians given "real-world" data and (2) CGMMs do not require the user to specify a hyperparameter value, i.e., the number of components.



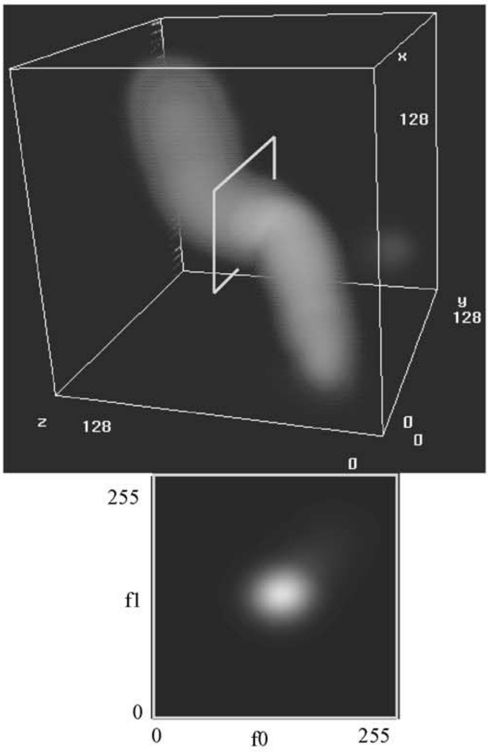


Fig. 12. (a) Estimated density function (b) Slice at $f_2 = 64$ through the estimated density function reveals its elliptical shape.

9. Conclusion

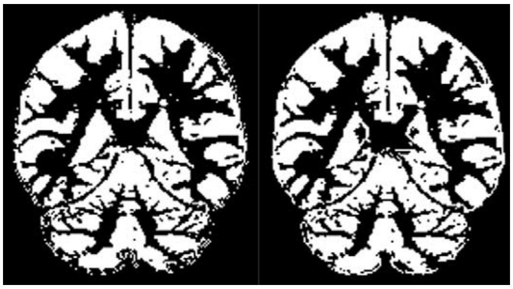


Fig. 13. (a) CGMM04's gray matter mask. (b) FGMM04's gray matter mask.

A CGMM of an extruded Gaussian can be defined using GGoF traces. Such models consistently yield accurate classifications. Initial experiments indicate that for small FPRs, this approach provides superior TPRs compared to FGMMs defined via MLEM. Given different collections of training data, the TPRs and FPRs associated with these labelings remain consistent relative to the consistency of the labelings produced by FGMMs. Furthermore,

- as additional GGoF traces are extracted, the accuracy and consistency of the CGMM improves asymptotically
- by defining CGMMs using GGoF traces, CGMM performance does not rely on the user to specify critical hyperparameters such as the number of components
- by using GGoF trace definitions, CGMM performances does not suffer from the problems associated with local maxima in an iterative parameter refinement processes, e.g., MLEM.

The application of CGMMs using GGoF ridges to medical image data and the existence of extruded Gaussians in medical images is demonstrated via the classification of tissues in an inhomogeneous MR image. Current work is focusing on the extraction of higher dimensional ($M > 1$) GGoF traces and the development of deformable distribution models using GGoF traces which adapt generic representations to form more optimal specific representations.

Acknowledgements

Special thanks goes to the members of Stephen Aylward's dissertation committee: James Coggins, Stephen Pizer, Dan Fritsch, Steve Marron, Jonathan Marshall, and Keith Muller. This work was supported in part by the Medical Image Display and Analysis Group with partial funding from NIH grant P01 CA47982.

References

[1] S.R. Aylward, Continuous mixture modeling via goodness-of-fit cores, Dissertation, Department of Computer Science, University of North Carolina, Chapel Hill, 1997.

[2] S.R. Aylward, S. Pizer, Continuous Gaussian mixture modeling, in: J. Duncan, G. Gindi (Eds.), Information Processing in Medical Imaging. Springer Lecture Notes in Computer Science 1230, Berlin, 1997, pp. 176–189.

[3] S.R. Aylward, J.M. Coggins, Spatially invariant classification of tissues in MR images. Visualization in Biomedical Computing, Rochester, MN, 1994.

[4] B.M. Dawant, A.P. Zijdenbos, R.A. Margolin, Correction of intensity variations in MR images for computer-aided tissue classification, IEEE Trans. Med. Imag. 12 (4) (1993) 770–781.

[5] C.R. Meyer, P.H. Bland, J. Pipe, Retrospective correction of intensity inhomogeneities in MRI, IEEE Trans. Med. Imag. 14 (1) (1995) 36–41.

[6] W.M. Wells III, W.E.L. Grimson, R. Kikinis, F.A. Jolesz, Adaptive segmentation of MRI data, IEEE Trans. Med. Imag. 15 (4) (1996) 429–442.

[7] J.R. Bellegarda, D. Nahamoo, Tied mixture continuous parameter modeling for speech recognition, IEEE Trans. Acoustics Speech Signal Process. 38 (12) (1990) 2033–2045.

[8] A. Depmster, N. Laird, D. Rubin, Maximum likelihood for incomplete data via the EM algorithm, Roy. Statist. Soc. 1 (1) (1977) 1–38.

[9] M.I. Jordan, L. Xu, Convergence results for the EM approach to mixtures of experts architectures, Technical Report, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, November 18, 1994.

[10] Z. Liang, R.J. Jaszezak, R.E. Coleman, Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing, IEEE Trans. Nucl. Sci. 39 (4) (1992) 1126–1133.

[11] X. Zhuang, Y. Huang, K. Palaniappan, Y. Zhao, Gaussian mixture density modeling, decomposition, and applications, IEEE Trans. Image Process. 5 (9) (1996) 1293–1302.

[12] D.M. Titterington, A.G.M. Smith, U.E. Markov, Statistical Analysis of Finite Mixture Distributions, Wiley, Chichester, 1985.

[13] G.J. McLachlan, K.E. Basford, Mixture Models, Marcel Dekker Inc., New York, vol. 84, 1988, p. 253.

[14] S.M. Pizer, D. Eberly, B.S. Morse, D.S. Fritsch, Zoom-invariant vision of figural shape: the mathematics of cores, Comput. Vision Image Understanding 69 (1998) 55–71.

[15] T.R.C. Read, N.A.C. Cressie, Goodness-of-fit Statistics for Discrete Multivariate Data, Springer, New York, 1988.

[16] T. Yoo, Image geometry through multiscale statistics, Dissertation, Department of Computer Science, University of North Carolina, Chapel Hill, 1996.

[17] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, Cambridge, 1990.

[18] J.P. Egan, Signal Detection Theory and ROC Analysis, Academic Press Inc., New York, 1975.

About the Author—STEPHEN R. AYLWARD is an Assistant Professor in the Department of Radiology at The University of North Carolina at Chapel Hill (UNC) and the leader of the Computer-Aided Diagnosis and Display Laboratory (http://caddlab.rad.unc.edu). He received his B.S. in computer science from Purdue University in 1988, his M.S. in computer science from Georgia Institute of Technology in 1989, and his Ph.D. in computer science from UNC in 1997. His research interests include forming 3D models of vasculature for interventional radiology planning and guidance, tumor differentiation using statistical pattern recognition methods, and data fusion to facilitate diagnosis.