From Kernels to Mixtures

David W. Scott
Department of Statistics, MS-138
Rice University
PO Box 1892
Houston, TX 77005-1892
scottdw@rice.edu

William F. Szewczyk
Mathematics Research Office
R&E Building, Suite 6515
National Security Agency
Fort George G. Meade, MD 20755-6515
wfszewc@afterlife.ncsc.mil

November 22, 2000

Abstract

Mixture models, which include kernel estimators, are widely used to model complex densities; however, one is faced with the challenge of determining an appropriate number of components. This task often involves identifying those components that are close enough to be combined. This paper introduces a new easily calculated measure of similarity between pairs of densities and illustrates its use in recursively collapsing components. This similarity measure naturally leads to a new algorithm (IPRA) for fitting mixture models sequentially. The algorithm is used to test for bumps in galaxy star velocity data and to examine MRI data.

Keywords: Kernel density estimation, Gaussian mixture models, BIC, Model simplification

1. Introduction

Mixture models have proven quite useful in modeling complex densities (Roeder 1990). Using a small number of normal components, one is able to model distributions that are far from normal (Marron and Wand 1992). Hence, mixtures are employed in everything from kernel density estimates (Scott 1992) to classification problems (MacLachlan 1992) to speech applications (Rabiner 1989). See the appendix for a brief review of kernel estimates.

Even though a number of books have been written on mixtures and methods for fitting them (Titterington, Smith, and Makov (1985), for example), most suffer from the drawback of only dealing with a small number of components. In real world applications, such as speech recognition, it is not unusual to have a multivariate model composed of thousands of components in tens of dimensions. Because of the computation involved in fitting one of these large models, the standard techniques of choosing an appropriate number of mixtures are inadequate. Instead, one attempts to choose a sufficiently large number of components and then fits the almost surely overparameterized model in a constrained manner to avoid singularities.

Figure 1 demonstrates an amazing fact about kernel and mixture models. The left-hand figure displays several mixtures with a varying number of components to the Lawrence Radiation Laboratory (LRL) data, consisting of 25752 observations, of Good and Gaskins (1980) with the right-hand figure zooming in to the region around the second mode. The four models used were a kernel density estimate using a normal kernel (i.e. a normal mixture with 25752 equally weighted and identical components), as well as a 50-, a 25-, and a 15-component normal mixture. (The smaller mixtures were derived from the kernel estimate using a method-of-moments procedure as described in Section 3.) There is no detectable visual difference between the kernel density estimate and the 50-component mixture and very little difference between those two and the 25-component model. It is only when one looks at the 15-component mixture (after reducing the number of components by 99.94%!) that one is able to easily see a difference from the original model, but its similarity with the original model (as defined in Section 2) is still high at 0.99998. While this example illustrates how well kernel estimates may be approximated by much smaller mixtures, it also raises the following question: How can one simplify and fit mixture models without losing significant information?

In this paper we address this question by the introduction and use of a new similarity measure between densities. A related difficult practical problem is obtaining good initial estimates for a mixture model. We will also show how it is possible to locally refit the parameters of a mixture model, thus making it possible to simplify large models by iteratively collapsing pairs of components. We will propose using a kernel estimate as an initial estimate of the *large* mixture model. The combination of these two methods will lead to new methods for fitting mixtures based on sequential simplification.

2. Similarity Measure for Densities

Let a general mixture density with m components be written as

$$f(x, \boldsymbol{\theta}) = \sum_{i=1}^{m} w_i \, f_i(x, \theta_i), \tag{1}$$

where $w_i > 0$, $\sum_{i=1}^{m} w_i = 1$, and f_i is a density with parameter vector θ_i. In theory the f_i's could be any parametric density, although in practice they are usually from the same parametric family, often normal.

The motivation for our similarity measure between two densities f_1 and f_2 is based on the expectation of one density, f_1, with respect to another, f_2. Intuitively, if f_1 and f_2 are similar then $\int f_1(x) f_2(x) \, dx$ should be larger than when f_1 and f_2 are not similar. (Since $E_{f_1} f_2 = E_{f_2} f_1$, the similarity definition is symmetric.) By Cauchy-Schwarz,

$$\int_{-\infty}^{\infty} f_1(x) f_2(x) \, dx \leq \left(\int_{-\infty}^{\infty} f_1^2(x) \, dx \int_{-\infty}^{\infty} f_2^2(x) \, dx \right)^{1/2}, \tag{2}$$

with equality *iff* $f_1(x) = f_2(x)$ almost surely. Dividing by the right-hand side of the inequality we define the similarity between two densities as the dimensionless quantity

$$\mathrm{sim}(f_1, f_2) = \frac{\int_{-\infty}^{\infty} f_1(x) f_2(x)\, dx}{\left(\int_{-\infty}^{\infty} f_1^2(x)\, dx \int_{-\infty}^{\infty} f_2^2(x)\, dx\right)^{1/2}} \tag{3}$$

Since f_1 and f_2 are nonnegative and using equation (2) we see that $0 \le \mathrm{sim}(f_1, f_2) \le 1$. Note that $\mathrm{sim}(f_1, f_2)$ can also be considered to be a "correlation" between the densities.

If f_1 and f_2 are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, then the similarity measure is given by the closed-form expression

$$\mathrm{sim}(f_1, f_2) = \frac{(2\sigma_1\sigma_2)^{1/2}}{(\sigma_1^2 + \sigma_2^2)^{1/2}} \exp\left[\frac{-(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right], \tag{4}$$

which allows for easy evaluation. Equation (4) follows from a useful fact: if $\phi(x|\mu, \sigma^2)$ denotes a normal density with mean μ and variance σ^2 evaluated at x, then

$$\int_{-\infty}^{\infty} \phi(x|\mu_1, \sigma_1^2)\, \phi(x|\mu_2, \sigma_2^2)\, dx = \phi(0|\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2). \tag{5}$$

We will use this well-known formula repeatedly in this paper. (A simple multivariate extension exists as well, see Wand and Jones (1995).)

It is instructive to examine equation (4) under the assumptions of equal variances or equal means. When $\sigma_1 = \sigma_2 = \sigma$ only the second factor contributes to the similarity, which reduces to a distance measure between the means, related to the coefficient of detection, $(\mu_1 - \mu_2)/\sigma$. On the other hand, when $\mu_1 = \mu_2$ equation (4) is proportional to the likelihood-ratio test statistic for equal versus non-equal variances (Rohatgi 1976).

In Figure 2 we display contours of the similarity measure between a $N(0, 1)$ and $N(\mu, \sigma^2)$ density over a range of means and standard deviations (note that the standard deviation is plotted on a log scale). The plot is symmetric about $\mu = 0$; however, for large values of σ, the similarity measure is fairly insensitive to μ.

To understand the empirical distribution of the similarity measure, we drew two samples of size 50 from a standard normal distribution and calculated the similarity measure using the sample moments. Figure 3 shows a histogram of the similarities for 1000 replications. The 5th percentile is 0.957; when $n = 500$, the 5th percentile increases to 0.995. However, we propose to use the measure more for ordering the closeness of pairs of densities rather than for inference.

3. Simplifying Mixtures

As an example of how to use this measure to simplify a mixture model, we generated 200 points from

$$f(x) = \frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2}, \frac{1}{9}). \tag{6}$$

Based on an ASH density estimate (Scott 1992), we chose $m = 3$ and used EM to fit a 3-component mixture to the data; see Figure 4. The estimated parameters of the mixture are

	$\hat{w}$	$\hat{\mu}$	$\hat{\sigma}$
Component 1	0.35	-0.48	1.15
Component 2	0.39	0.28	0.82
Component 3	0.26	1.55	0.32

To proceed, the definition of similarity must be extended to $\mathrm{sim}(w_1 f_1, w_2 f_2)$. It may be seen from directly plugging into equation (3) that the weights cancel from the numerator and denominator; hence we need only plug our estimates of the $\hat{\mu}$'s and $\hat{\sigma}$'s into equation (4). This results in the estimated similarity matrix:

	Component 1	Component 2	Component 3
Component 1	1.00		
Component 2	0.84	1.00	
Component 3	0.17	0.28	1.00

Based upon this table and Figure 4, components 1 and 2 appear to be good candidates for collapsing.

Once we have decided to combine the first two components, the question as to how naturally arises. For a small sample such as this we could just rerun the EM algorithm from scratch using just two components. However, this becomes very expensive when applied sequentially with a large dataset or a model with many components. Instead, we propose to refit the parameters locally.

3.1. Local Refitting Local refitting adjusts the parameters of only those components of interest, leaving the others fixed. There are a number of techniques which will accomplish this; we will examine three: (1) the method of moments, (2) L$_2$E (Scott 1998) without reference to the data, and (3) L$_2$E using the data.

3.1.1. Method of Moments In this case to refit locally, we replace the two components with one component whose parameters are calculated by the method of moments (MoM). If the parameters of the first component are w_1, μ_1, and σ_1^2 and those of the second are w_2, μ_2, and σ_2^2, then the updated component will have parameters $w_{new} = w_1 + w_2$, $\mu_{new} = w_1'\mu_1 + w_2'\mu_2$, and $\sigma_{new}^2 = w_1'\sigma_1^2 + w_2'\sigma_2^2 + w_1'w_2'(\mu_1 - \mu_2)^2$, where $w_1' = w_1/w_{new}, w_2' = 1 - w_1'$. Table 1 contains the parameters estimated by MoM as well as the other methods.

3.1.2. L$_2$E Scott (1998) introduced the L$_2$E method as a robust means of estimating the parameters of a density. The L$_2$E estimator $\hat{\boldsymbol{\theta}}$ is found by minimizing the L_2 error between

Technique	$\hat{w}$	$\hat{\mu}$	$\hat{\sigma}$
Method of Moments	0.74	-0.08	1.11
L$_2$E w/o Data	0.74	0.06	0.99
L$_2$E w/ Data	0.76	-0.01	1.06

Table 1: Estimated $\hat{w}, \hat{\mu}$ and $\hat{\sigma}$ for the three methods.

the estimated and target densities and is defined as,

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \int_{-\infty}^{\infty} \left(\hat{f}(x;\boldsymbol{\theta}) - f(x) \right)^2 dx. \tag{7}$$

There are two different ways that we can use L$_2$E to do a local refitting. The first, in analogy with MoM, estimates the parameters, apparently independent of the data (although it must be pointed out that the data were used in the original estimation of the parameters), while the second makes explicit use of the data in the estimation.

In the first approach we replace the two-component mixture by the single normal that minimizes the L_2 error. Substituting $\phi(x|\hat{\mu}, \hat{\sigma}^2)$ for $\hat{f}$ and $w_1' \, \phi(x|\mu_1, \sigma_1^2) + w_2' \, \phi(x|\mu_2, \sigma_2^2)$ for f in equation (7) we obtain

$$\underset{\hat{\mu},\hat{\sigma}}{\operatorname{argmin}} \int_{-\infty}^{\infty} \left\{ \phi(x|\hat{\mu}, \hat{\sigma}^2)^2 - \left[w_1'\phi(x|\mu_1, \sigma_1^2) + w_2'\phi(x|\mu_2, \sigma_2^2) \right] \right\}^2 dx.$$

Expanding the square and ignoring terms not involving $\hat{\mu}$ or $\hat{\sigma}$, the L$_2$E estimator is given by

$$\underset{\hat{\mu},\hat{\sigma}}{\operatorname{argmin}} \int_{-\infty}^{\infty} \left\{ \phi(x|\hat{\mu}, \hat{\sigma}^2)^2 - 2 \left[w_1'\phi(x|\hat{\mu}, \hat{\sigma}^2)\phi(x|\mu_1, \sigma_1^2) + w_2'\phi(x|\hat{\mu}, \hat{\sigma}^2)\phi(x|\mu_2, \sigma_2^2) \right] \right\} dx.$$

Using equation (5), this reduces to finding $\hat{\mu}$ and $\hat{\sigma}^2$ that minimize

$$\frac{1}{2\hat{\sigma}\sqrt{\pi}} - 2 \left[w_1 \, \phi(0|\hat{\mu} - \mu_1, \sigma_1^2 + \hat{\sigma}^2) + w_2 \, \phi(0|\hat{\mu} - \mu_2, \sigma_2^2 + \hat{\sigma}^2) \right]. \tag{8}$$

(Again, refer to Table 1 for the estimates for this example.)

In the second approach we wish to fit an m-component mixture to the data, keeping all but one component fixed. At first it would appear that we would not be able to use L$_2$E, since we do not know the target density, but that is not the case. Expanding the integral in equation (7) results in

$$\int_{-\infty}^{\infty} \hat{f}(x;\boldsymbol{\theta})^2 dx - 2\int_{-\infty}^{\infty} \hat{f}(x;\boldsymbol{\theta}) f(x)\, dx + \int_{-\infty}^{\infty} f(x)^2 dx.$$

The last term in the sum does not depend on $\boldsymbol{\theta}$ and thus does not enter into the optimization. The second integral is the expected height of the estimated density, $E\hat{f}(X;\boldsymbol{\theta})$, which we estimate by averaging over the sample. The L$_2$E formulation reduces to

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \int_{-\infty}^{\infty} \hat{f}(x;\boldsymbol{\theta})^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}(x_i;\boldsymbol{\theta}). \tag{9}$$

In the case of a mixture of normals, the integral in (9) can be easily evaluated in closed form using the useful fact given in equation (5); see Scott (1999) for further details.

4. Fitting Mixtures with IPRA

Fitting a mixture model with EM where the number of components m is deliberately inflated is extremely slow and highly dependent upon the initial values. Here we consider beginning with a large mixture model that is in fact the kernel estimate ($m = n$). By combining the similarity measure with local refitting, one can fit a sequence of reduced mixture models. We call such a procedure the Iterative Pairwise Replacement Algorithm (IPRA), which may be outlined as follows:

Phase 0: Construct a kernel density estimate.

Phase 1: Combine pairwise ties and near-ties using MoM.

Phase 2: Combine pairwise similar components using MoM or L$_2$E without data.

Phase 3: Combine pairwise similar components using L$_2$E with data until there is one component.

Phase 4: Using statistics collected in Phase 3, choose an appropriate number of components by an information criterion and perhaps re-estimate the parameters using L_2E with data.

We will explain each phase of the algorithm in detail before presenting examples.

Phase 0 We initialize the algorithm with an over-parameterized mixture, namely, a fixed-bandwidth kernel density estimate. There are many references on selecting an appropriate bandwidth (Scott (1992), Silverman (1986), and Wand and Jones (1995), among others), but our experience has shown that the results are similar for most reasonable selections of the bandwidth. In practice we use either the unbiased cross-validation (UCV) bandwidth (Rudemo 1982; Bowman 1984; Scott and Terrell 1987), for an undersmoothed start, or the normal reference rule (Scott 1992) for an oversmoothed start. (The appendix contains a short review of the UCV bandwidth.)

Phase 1 The first step in eliminating the redundancy in the mixture is to combine exact ties and near-ties. Recall that for equal variances, equation (4) reduces to a distance measure between means. We take advantage of that by declaring any two means separated by less than a certain threshold as a near-tie and combine them using the MoM (Section 3.1.1). The threshold is based on the sorted distances between adjacent means and on the sample size. For small sample sizes (less than, say 200) no near-ties need be declared. For moderate sizes (between 200 and 2000) the smallest 50 to 75% of the distances can be declared close enough to be near-ties without any noticeable change in the density estimate. For large sample sizes this fraction can be increased to 90 or 95% or more.

Phase 2 Once ties and near-ties have been eliminated we proceed by combining those pairs of components which change the density estimate the least. We calculate the similarity (equation (3)) between the current k-component mixture and the $(k-1)$-component mixture obtained by replacing each pair of adjacent components, in turn, by the MoM combination of the two. Choosing the pair that maximizes the similarity ensures that the density does not change too radically between iterations. (Phase 1 could, in reality, could be conducted as part of by Phase 2. However, by eliminating near-ties separately, we are able to speed up the implementation substantially.)

Phase 3 The boundary between Phase 2 and Phase 3, as well as the one between Phase 1 and Phase 2, is somewhat arbitrary. It has been chosen to balance the goals of computational efficiency and accuracy. In the current implementation we shift to Phase 3 when there are no more than 30 components in the mixture. Now that there are relatively few components at this point we can afford to use more computationally expensive methods. Since the MoM does not allow the variance of the components to shrink, it cannot correct any oversmoothing in the estimate. By using L_2E with data we can overcome that, by fitting a $(k-1)$-component model with all but one component fixed. By optimizing over the mean, variance, and weight of the new component the mixture can take on a character different from the original kernel estimate.

While comparing adjacent components often works, there are times when this is not the best strategy (see the third example in the next section for such a situation). To allow for "bridging" across components, in Phase 3 we also calculate the similarity between components two away instead of looking at only adjacent components. At this point it would also be feasible to do all k choose 2 comparisons, but we have found that comparing adjacent and two away components is sufficient.

Phase 4 The selection of an appropriate model is the most difficult phase of model building. We use a combination of the BIC (Kass and Raftery 1995), the value of L_2E criterion, and intuition. (The use of BIC is not rigorous, but our empirical work has found it a useful adjunct to L_2E.) Once a group of candidate models have been chosen, the parameters can then be refit globally using L_2E.

Returning to the two-component mixture from Section 3, Figure 5 on page 25 shows the density estimate after Phases 0–2, starting with the UCV bandwidth and a normal kernel. Phase 1 eliminates 99 near-ties using the median distance as the threshold. (There are no exact ties in these data.) Phase 2 further reduces the number of components to 30. The fact that there is no discernible difference between the Phase 0 estimate (a 200-component mixture) and the Phase 2 estimate (a 30-component mixture) illustrates how overparameterized the kernel estimator actually is.

At the end of Phase 3 there is no significant difference in the BIC for the 1-component model vs. the 2-component model (-335.687 and -335.454, respectively). The L_2E criterion, however, favors two components (-0.224 vs. -0.246). Figure 6 shows both candidate models plotted with the original kernel density estimate. Given the evidence of the BIC, the L_2E criterion, and the comparison with the kernel density, we decide on the two component model

$$\hat{f}(x) = 0.74\,\phi(x|\text{-}0.05, 1.18) + 0.26\,\phi(x|1.55, 0.12)$$

which compares favorably with the true density, equation (6) and Figure 7.

5. Simulated Examples

In this section we present a variety of examples to illustrate the performance of IPRA on both simulated and actual data.

5.1. Simulated Data Set One Figure 8 shows the result of applying IPRA to a four-component mixture starting from an oversmoothed kernel density estimate. The data are 1000 samples from

$$f(x) = \frac{4}{10}\,\phi(x|\text{-}2.5, 0.36) + \frac{3}{10}\,\phi(x|\text{-}0.5, 1) + \frac{2}{10}\,\phi(x|2, 0.09) + \frac{1}{10}\,\phi(x|3, 0.04).$$

We initialized the algorithm with a Gaussian kernel estimate using the normal reference rule bandwidth, $h = 1.06\hat{\sigma}n^{-1/5}$, where $\hat{\sigma}$ is an estimate of the overall standard deviation and

n is the length of the data (Scott 1992). As can be seen in the left-hand side of Figure 8 this produces an estimate with only 3 visible modes. After Phase 3, both the BIC and L_2E criteria favor a 4-component mixture. Figure 8 compares this solution with both the initial estimate (left) and the true density (right). Doing the local refit in Phase 3 using L_2E with data allows the estimated variances to shrink, overcoming the oversmoothed start.

5.2. Simulated Data Set Two Our second example illustrates a robustness property of IPRA and why we use both the BIC and the L_2E criteria for model selection. The data are as in the previous example with 4 additional points from a $N(6, (1/100)^2)$ distribution. The upper plots of Figure 9 show the BIC and L_2E criteria for 1- to 7-component models after Phase 3. If we were to use the BIC alone we would choose a model with 5 components (lower left). However, the L_2E criterion indicates that the 4-component model (lower right) is just as interesting as the 5-component one. Because of the robustness of the L_2E method the only change in going from 5 to 4 components is to increase the weight of the component with mean 3 by the weight of the mean 6 component. (The mean and scale estimates are not affected.) We leave the decision of which model is more appropriate to the reader.

5.3. Simulated Data Set Three Phase 2 of IPRA collapses adjacent components with maximum speed. Most of the time this causes no problems. Sometimes, however, two components which should be combined are separated by a third. Figure 10 on page 28 shows such a case. For this example we drew 180 samples from

$$f(x) = \frac{2}{3}\,\phi(x|0,1) + \frac{1}{3}\,\phi\,(x|0,0.01)\,.$$

The 3-component model is shown in the top plot of Figure 10. If we construct the 2-component model based on collapsing only adjacent components we obtain the model on the lower left, which is clearly inadequate. Allowing for bridging in Phase 3 produces the model on the lower right which captures the true density quite well.

6. Short Case Studies

To illustrate IPRA's performance on real data, we present three short case studies.

6.1. Galaxy Data The plausibility of many physical theories rests on testing the density function for the presence of modes or bumps at locations predicted by the theory. Good and Gaskins (1980) give an excellent account of this idea, including data collected from high-energy collider experiments (see Figure 1). At an NSF workshop in 1997 on clustering and bump-hunting (*http://www.stat.rice.edu/smoothing.html*), Jiayang Sun (Sun 1997) described competing theories of the formation of galaxies. The relevant data come from velocity measurements on some of the oldest stars in the Milky Way. If the process is rapid, then the velocity distribution should be smooth and unimodal, whereas if the process involves the destruction of smaller neighboring galaxies, then the velocity distribution should contain a narrow peak comprised of the stars originating in the smaller galaxy. In the data shown in Figure 11, a bump in the vicinity of 75 km/s is at issue. In her paper, Sun reports that most methods either fail to find the mode or underestimate its size; see also (Sun, Morrison, Harding, and Woodroofe 2000).

We re-analyzed these data with the IPRA method and also found the mode centered at $x = 74.92$ km/s (Figure 11). However, because of the long right tail in the data, we decided to repeat the analysis on a log-velocity scale. The top left frame of Figure 12 shows the starting kernel density estimate. The righthand figures are the resulting 6-component model on the log-velocity scale (top) and velocity scale (bottom). Both analyses found the same component around 75 km/s. This can be visually confirmed by overlaying the region of interest, in the velocity scale, from both analyses (Figure 12, bottom left).

The 7-component model (Figure 13) has a lower $\mathrm{L_2E}$ score than the 6-component one, and the BICs only differs by the penalty term, so it may be worthwhile to examine the 7-

component model as well. When we do, we are surprised to find another sharp peak around 87.96 km/s. To our knowledge, this peak has not been discussed in the literature. We do not have the astronomical background to comment on this component, but we highlight it and the peak at 22.16 km/s in the hope of sparking discussion.

6.2. British Income Data Our second example involves a large data set, the 1975 income of 7201 British households. The data were standardized by dividing by the mean income, and transformed by taking logs. Because of the size of the data set we chose to initialize IPRA with a density estimate based on the normal reference rule bandwidth (Figure 14 on page 32, top left) rather than the UCV to avoid the optimization. We also declared the smallest 85% of the distances as near-ties, reducing the number of components by 4966 (in addition to 1357 ties). There is no discernible difference between the density estimates at the beginning and the end of Phase 1. At the end of Phase 3, the two-component model (top right) is preferred by both the BIC and the L_2E criteria; however, expert opinion lends credence to a 3-component one, so we globally refit the 2-, 3-, and 4-component models starting from their Phase 3 values. The 2-component model did not change much in Phase 4, but the 3- and 4-component ones (Figure 14, bottom) are now either just as good, according to the BIC, or better, according to the L_2E criterion, as the 2-component one.

Examining the models in more detail, we see that the 2-component model makes an interesting compromise between the shoulder that appears in the kernel density estimate and the mode that IPRA discovers at around -0.5. The 3- and 4-component models are able to capture both the shoulder and the mode by placing a component around -0.5 whose sole purpose is to increase the density in that region (perhaps representing a group on a fixed-income). The densities for these last two models are essentially the same, however there is a preference for the 3-component model based on how well it fits the rightmost mode.

6.3. MRI Data Our last example is an application of IPRA to tissue typing. The image, Figure 15, left, is an MRI image from the S+WAVELETS™ toolkit (Bruce and Gao 1994) and is composed of 65536 grayscale values. Because the data are quantized values, nominally bounded below at 0, we first blurred the data and used the logs of the blurred data. The right side of Figure 15 shows the UCV-based density estimate (dotted line) and the preferred seven-component IPRA based model (solid line). In this case the UCV-based estimate produces an undersmoothed model, but IPRA is able to identify two shoulders at 3.9 and 4.6 missed by the kernel estimate.

Figure 16 shows the pixels associated with six of the seven clusters (the seventh cluster contained 13 isolated pixels with negative values and is not shown). The images have been masked to suppress the background. While groundtruthing of MRI data is not routinely done, the clusters appear to correspond to (from left to right, top to bottom) Air, Fluids, Fatty Tissue, Gray Matter/Muscle, White Matter, and Bone.

7. IPRA in Higher Dimensions

The similarity measure, equation (3), was originally developed to address a normal mixture problem in high dimensions. The model, based on a NIST dataset, had 2048 components in 39 dimensions (Reynolds 1995). The components were constrained to have arbitrary, but diagonal covariance matrices. Considerable computational resources were devoted to the determination of the model parameters by EM. The number of components was chosen to be more than sufficiently large enough to produce a good fit without any regard to parsimony or interpretability. Using a multivariate version of the similarity measure in equation (4), searching over local regions, and combining densities by MoM, the number of components was halved with no loss of discrimination power.

We believe that this number could be further reduced, but this raises the question of how to extend IPRA to higher dimensions. IPRA in one dimension was designed to take advantage of the natural ordering of the means of the components, reducing the number of comparisons from $O(n^2)$ to $O(n)$. Unfortunately, that advantage is lost in higher dimensions. One possible approach is to limit searches to adjacent nodes of a minimum spanning tree (MST), an idea also suggested by an anonymous referee, and updating the MST after each iteration. Another approach would be to probe the data using the stochastic mode tree (Scott and Szewczyk 2000) to identify putative modes. Starting with those modes one can then proceed as in Phase 3 of IPRA. We are currently pursuing both lines of research.

8. Discussion and Implementation Details

Mixture models can provide a parsimonious and highly interpretable fit for data sets. We believe IPRA has several distinct roles to play. First, IPRA can eliminate the need for guessing initial values for mixture models by sequentially simplifying a kernel estimate. Second, IPRA can provide a sieve estimate of successively simpler mixture models using the robust L_2E criterion. Finally in extremely challenging settings where $m \gg 1$, $d \gg 1$ and finding the optimal m is not feasible, an EM fit may be simplified into a more parsimonious version (without optimization) as in Figure 1.

In the preceding we glossed over some topics that need to addressed in detail. We will do so in this section.

8.1. Design Issues We designed our algorithm to capture in a parsimonious manner those features which are revealed by a good kernel density estimate. We chose an estimate based on the unbiased cross-validation bandwidth (UCV) for a number of theoretical, operational, and aesthetic reasons. Scott and Terrell (1987) show that the UCV bandwidth is a consistent estimator of the minimum asymptotic mean integrated squared error bandwidth, resulting in a good, if slightly undersmoothed, density estimate. It has also been our experience that the UCV bandwidth is the largest bandwidth that keeps unique data-points in separate components. That is, if we were to assign data to components based on maximum likelihood, a bandwidth larger than the UCV one would result in unique values being assigned to the same component. The fact that the UCV bandwidth is relatively noisy does not matter for this application.

Since we start with what we believe is a reasonable estimate, we designed IPRA to ensure that it doesn't stray too far from it. In order to do this, we constrain the optimization in Phase 3 by limiting the mean to a region bounded by other components and by setting a lower bound on the variance. A bound of 1/4 of the smallest variance of the components to be collapsed allows the variance to shrink, and perhaps reveal new modes, without causing severe undersmoothing or Dirac spikes. The only constraint we put on the weight is that it is in the interval $(0, 1)$, scaling the remaining ones to maintain $\sum w_i = 1$.

8.2. Partial Similarity The original motivation for the similarity measure was to identify redundant components in a mixture model; as defined in equation (3), similarity is independent of the weights assigned by the mixture to the two densities. To take proper account of these weights in this step would require a modification to the definition of "similarity."

While we have yet to solve this problem to our satisfaction, we have identified two potentially worthwhile approaches. The first approach defines the similarity between components as

$$\mathrm{sim}(w_1 f_1, w_2 f_2) \equiv \alpha(w_1, w_2)\mathrm{sim}(f_1, f_2).$$

One choice for $\alpha(w_1, w_2)$ is a function that assumes normalized weights ($w_1 + w_2 = 1$), attains its maximum when $w_1 = w_2 = 1/2$, and goes to 0 as either w_1 or w_2 goes to 0. One such function is related to the Epanechnikov kernel, $\alpha(w_1, w_2) = 4w_1 w_2 = 4w_1(1 - w_1)$. Normalized weights, however, may not be appropriate. Consider, for example, a situation where the weights of the two adjacent densities are 0.20 and 0.40, versus a situation where the weights are 0.10 and 0.20. In both cases, the normalized weights become 1/3 and 2/3, so the similarity measures are the same. However, it might make sense to have a measure that distinguishes these two cases, since we might prefer to merger the smaller pair (0.10, 0.20), before we would merge the larger pair (0.15, 0.30).

Consequently, we might consider a more intuitive approach which defines the similarity as

$$\mathrm{sim}(w_1 f_1, w_2 f_2) \equiv \mathrm{sim}(w_1 f_1 + w_2 f_2, \hat{w}\hat{f}_{12}),$$

where $\hat{w}\hat{f}_1 2$ is the component obtained by collapsing the other two. This approach requires an estimate of $\hat{w}\hat{f}_1 2$, the best method for which (e.g., ML, MoM, L2E) is presently unknown. Further research is needed to determine which of these two approaches is more appropriate.

8.3. Least Squares Mixture Decomposition Our approach is similar to Kim (1995), who focuses on adjusting the weights in the kernel estimate, $\sum_{i=1}^{n} w_i K_h(x - x_i)$, keeping h fixed. Using a least-squares criterion, he shows that his algorithm is a quadratic program and that many of the weights in the solution are zero. IPRA varies not only the weight of each mixture component, but also the location and scale.

8.4. Agglomerative EM As we were preparing our final version of this paper Mário Figueirdo called our attention to his recent work on agglomerative EM (AEM) (Figueiredo et al. 1999) as an alternative method of simplifying mixtures. AEM is initialized with, using their notation, a $k_{\max}$-component mixture. At each iteration of the algorithm, the current k-component mixture is used to initialize the EM algorithm. After convergence the two most similar components, based on the symmetric Kullback-Leibler divergence, are combined using the MoM. This new $(k-1)$-component mixture becomes the starting point for the next iteration.

We see two drawbacks to this method which IPRA does not have. The first is the need to choose $k_{\max}$. As the authors state "As long as $k_{\max}$ is large enough, AEM is quite insensitive to initialization." The question is, what is large enough? If $k_{\max}$ is chosen too large one may run into problems caused by the second drawback, the need to run EM at every iteration. IPRA avoids both of these by starting at the maximally over-parameterized mixture, the kernel density estimate, and only refitting locally.

Acknowledgments

This research was supported in part by the National Science Foundation grant #99-71797 and the National Security Agency. This work was completed while the first author was on sabbatical at the National Security Agency. The authors would like to thank the referees for their helpful comments.

Appendix: Kernel Density Estimators and the UCV Bandwidth

Let $x_i, i = 1, \ldots, n$, be a sample from a univariate distribution, and let $K(x)$ be a kernel function; that is $K(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} K(x)\,dx = 1$, and $\int_{-\infty}^{\infty} xK(x)\,dx = 0$ (Scott 1992). The fixed bandwidth kernel density estimate of the underlying density is

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i),$$

where $K_h(x) = K(x/h)/h$.

The UCV bandwidth, h_{ucv}, is then defined as

$$h_{ucv} = \operatorname*{argmin}_h \int_{-\infty}^{\infty} \left(\hat{f}(x) - f(x)\right)^2 dx.$$

In analogy to equation 7, the above can be expanded and simplified to

$$h_{ucv} = \operatorname*{argmin}_h \int_{-\infty}^{\infty} \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(i)}(x_i),$$

where $\hat{f}_{(i)}$ is the density estimate obtained by omitting the i-th data point. For the normal kernel this becomes (Rudemo 1982; Bowman 1984)

$$h_{ucv} = \underset{h}{\operatorname{argmin}} \frac{1}{2nh\sqrt{\pi}} + \frac{1}{n^2 h\sqrt{\pi}} \sum_{i<j} \left(e^{-((x_j - x_i)^2/4h^2)} - \sqrt{8}\, e^{-((x_j - x_i)^2/2h^2)} \right).$$

Multivariate versions of these formulae exist (Sain, Baggerly, and Scott 1994).

References

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika 71*, 353–360.

Bruce, A. and H.-Y. Gao (1994, December). *S+WAVELETS User's Manual*. Seattle, WA: DAPD, Mathsoft, Inc.

Figueiredo, M. A. T., J. M. N. Leitão, and A. K. Jain (1999). On fitting mixture models. In E. Hancock and M. Pellilo (Eds.), *Energy Minimization Methods in Computer Vision and Pattern Recogntion*, pp. 54–69. Springer-Verlag.

Good, I. J. and R. A. Gaskins (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by the scattering and meteorite data (with discussion). *Journal of the American Statistical Association 75*, 42–73.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kim, D. (1995). *Least-Squares Mixture Decomposition*. Ph. D. thesis, Virginia Polytechnic Institute.

MacLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley & Sons, Inc.

Marron, J. S. and M. P. Wand (1992). Exact mean integrated squared error. *Annals of Statistics 20*, 712—736.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257—286.

Reynolds, D. A. (1995). Automatic speaker recognition using gaussian mixture speaker models. *Lincoln Laboratory Journal 8*, 173–192.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxy. *Journal of the American Statistical Association 85*, 617–624.

Rohatgi, V. K. (1976). *An Introduction to Probability Theory and Mathematical Statistics.* New York: John Wiley & Sons.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics 9*, 65–78.

Sain, S. R., K. A. Baggerly, and D. W. Scott (1994). Cross-validation of multivariate densities. *Journal of the American Statistical Association 89*, 807–817.

Scott, D. W. (1992). *Multivariate Density Estimation.* New York: John Wiley & Sons, Inc.

Scott, D. W. (1998). Parametric modeling by minimum L_2 error. Technical Report 98-3, Department of Statistics, Rice University.

Scott, D. W. (1999). Remarks on fitting and interpreting mixture models. In K. Berk and M. Pourahmadi (Eds.), *Computing and Statistics*, Proceedings of the 31st Symposium on the Interface, pp. 104–109.

Scott, D. W. and W. F. Szewczyk (2000). The stochastic mode tree and clustering. in preparation.

Scott, D. W. and G. R. Terrell (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association 82*, 1131–1146.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

Sun, J. (1997). Bump hunting problems: A new test. Presented at NSF Workshop: Bumps, Jumps, Clustering and Discrimination, Houston Texas.

Sun, J., H. Morrison, P. Harding, and M. Woodroofe (2000). Mixture and bumps: Errors in measurement. Technical Report submitted, Dept. of Statistics, Case Western Reserve University.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley & Sons, Inc.

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. London: Chapman & Hall.

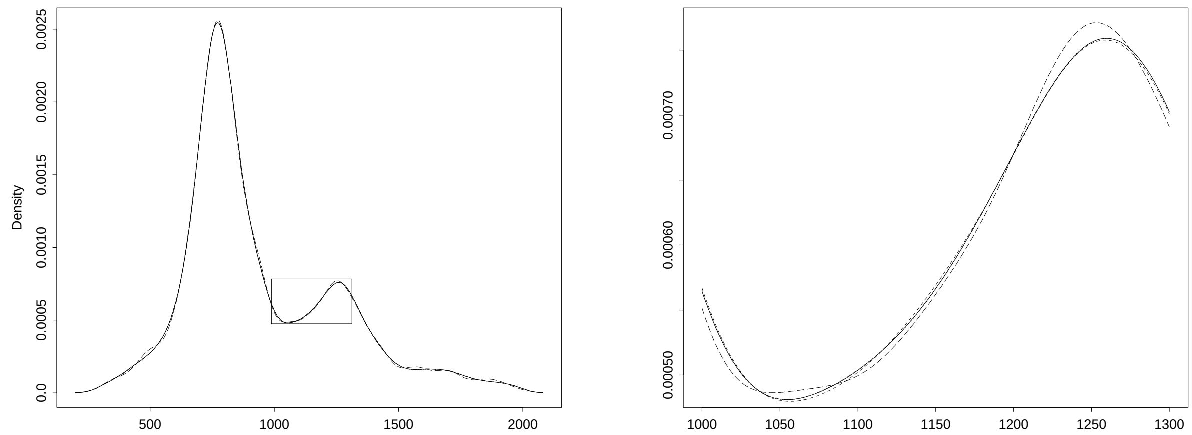

Figure 1: Left: Comparison of mixture models for LRL data with 25752, i.e., the sample size (solid), 50 (dotted), 25 (dash-dotted), and 15 (dashed) components. Right: Enlargement of boxed region.

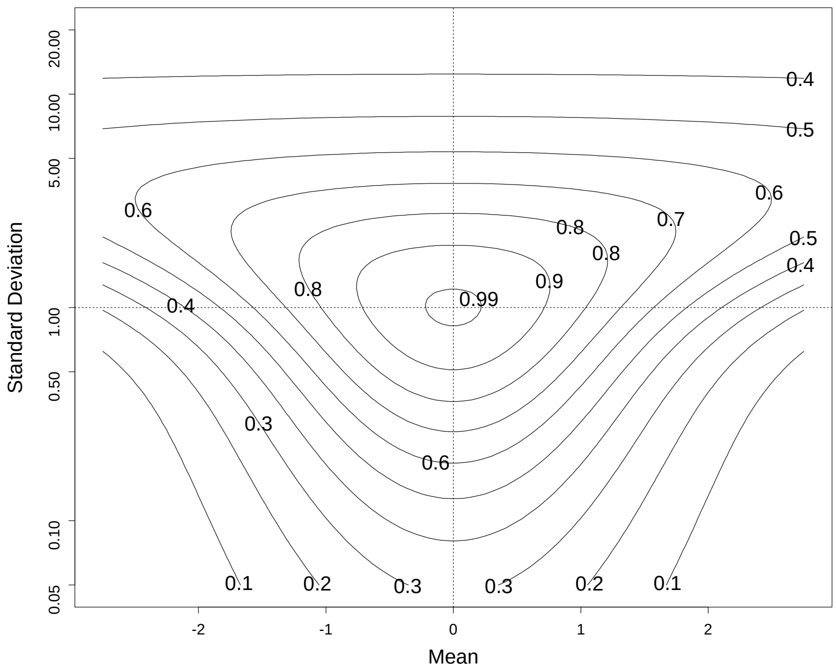

Figure 2: Similarity between $N(0,1)$ and $N(\mu, \sigma^2)$ densities.

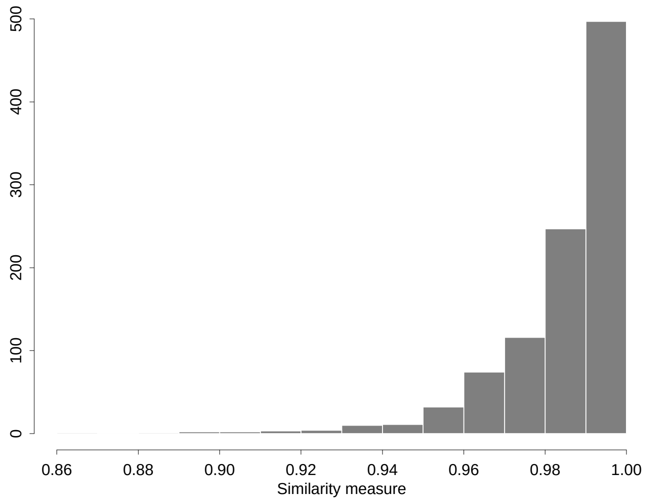

Figure 3: Sample similarity measure distribution when $n = 50$ and $f_1 = f_2 = N(0, 1)$.

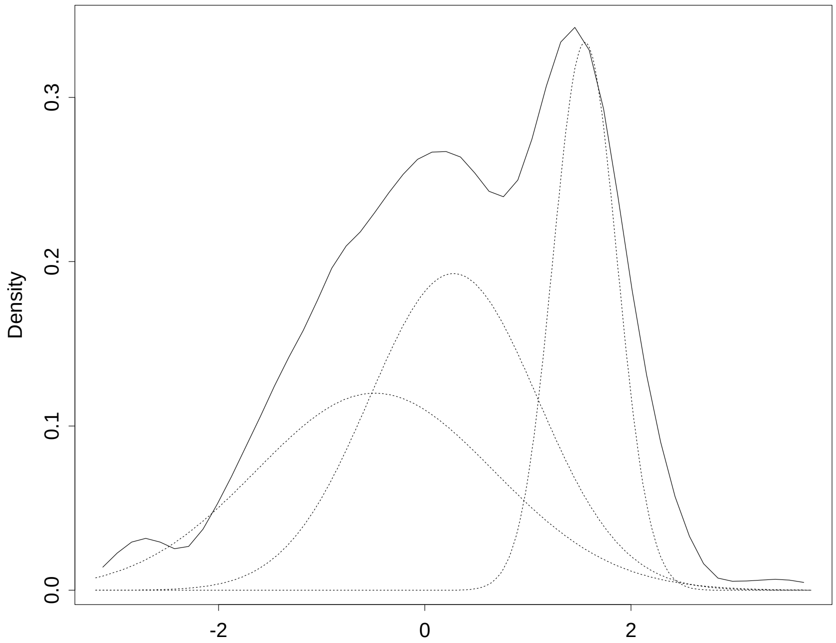

Figure 4: ASH density estimate and the three components obtained using EM.

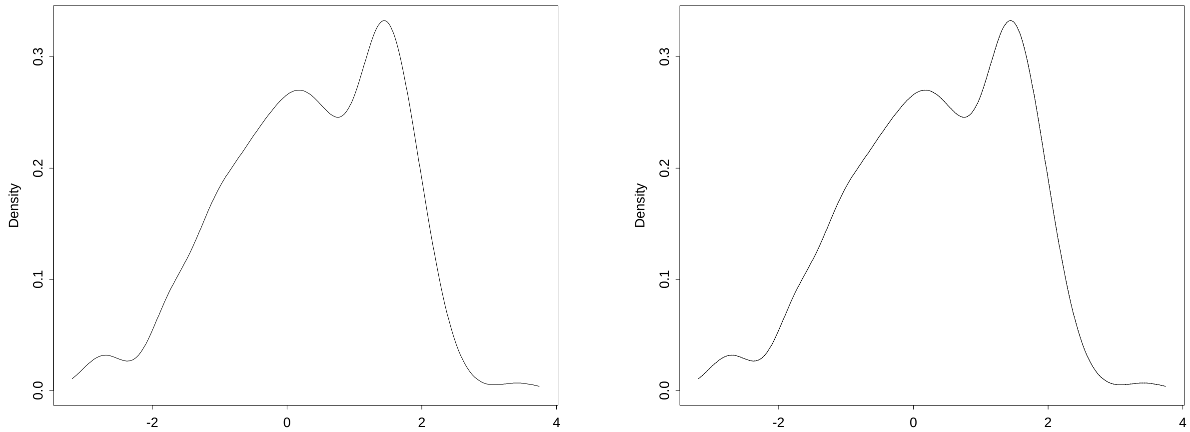

Figure 5: Initial Phase 0 estimate with 200 components (left). Density estimates after Phase 1 with 101 components (solid line) and Phase 2 with 30 components (dashed line, but not visible) (right).

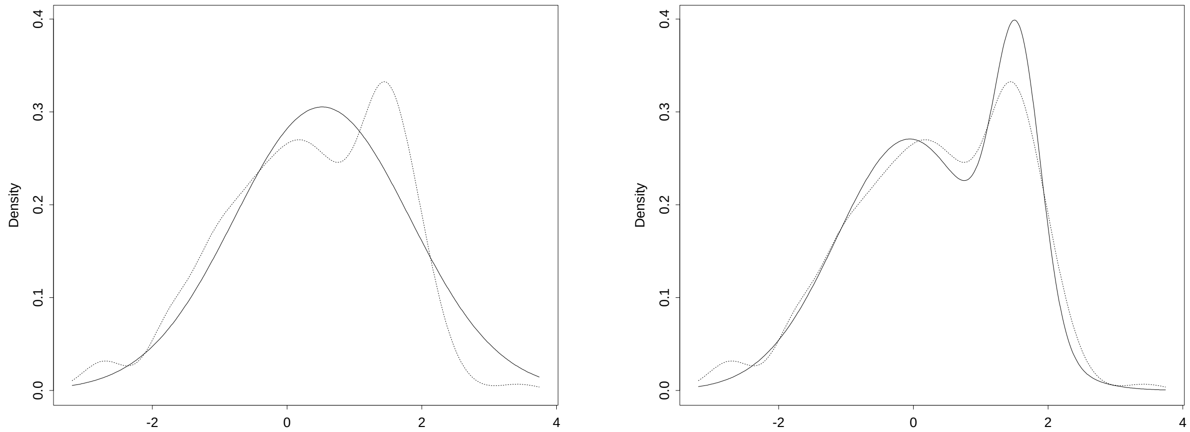

Figure 6: Preferred density estimates based on the BIC after Phase 3: one component (left) and two components (right). The kernel estimate with the UCV bandwidth is superimposed in both figures (dotted line).

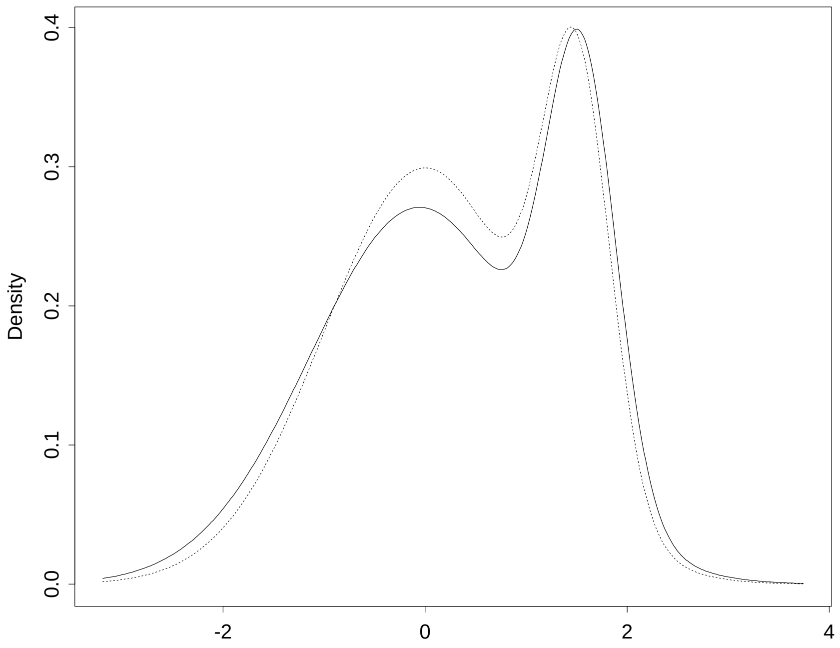

Figure 7: Two-component model chosen by IPRA and true density (dotted line).

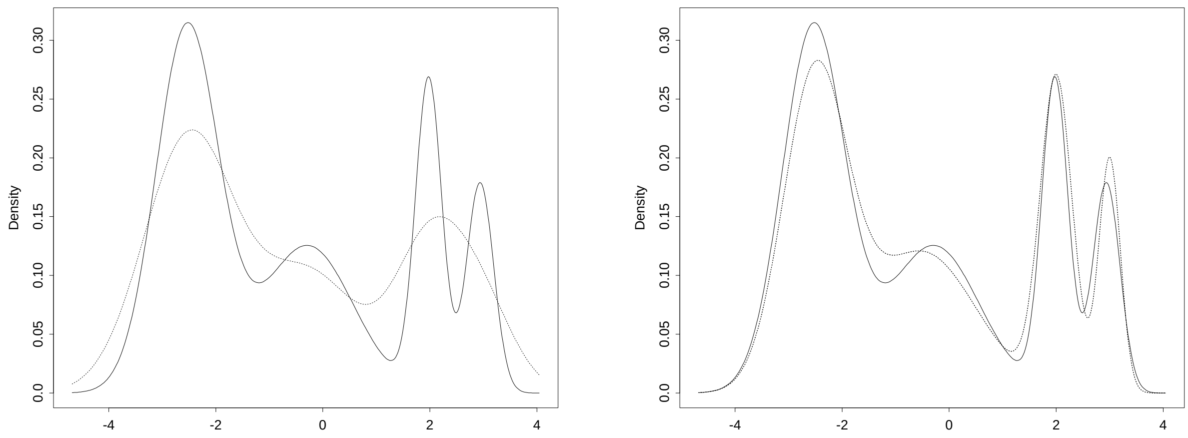

Figure 8: Left: Initial kernel density estimate (dotted line) and 4-component mixture selected by IPRA for the first example. Right: IPRA estimate and true density (dotted line).

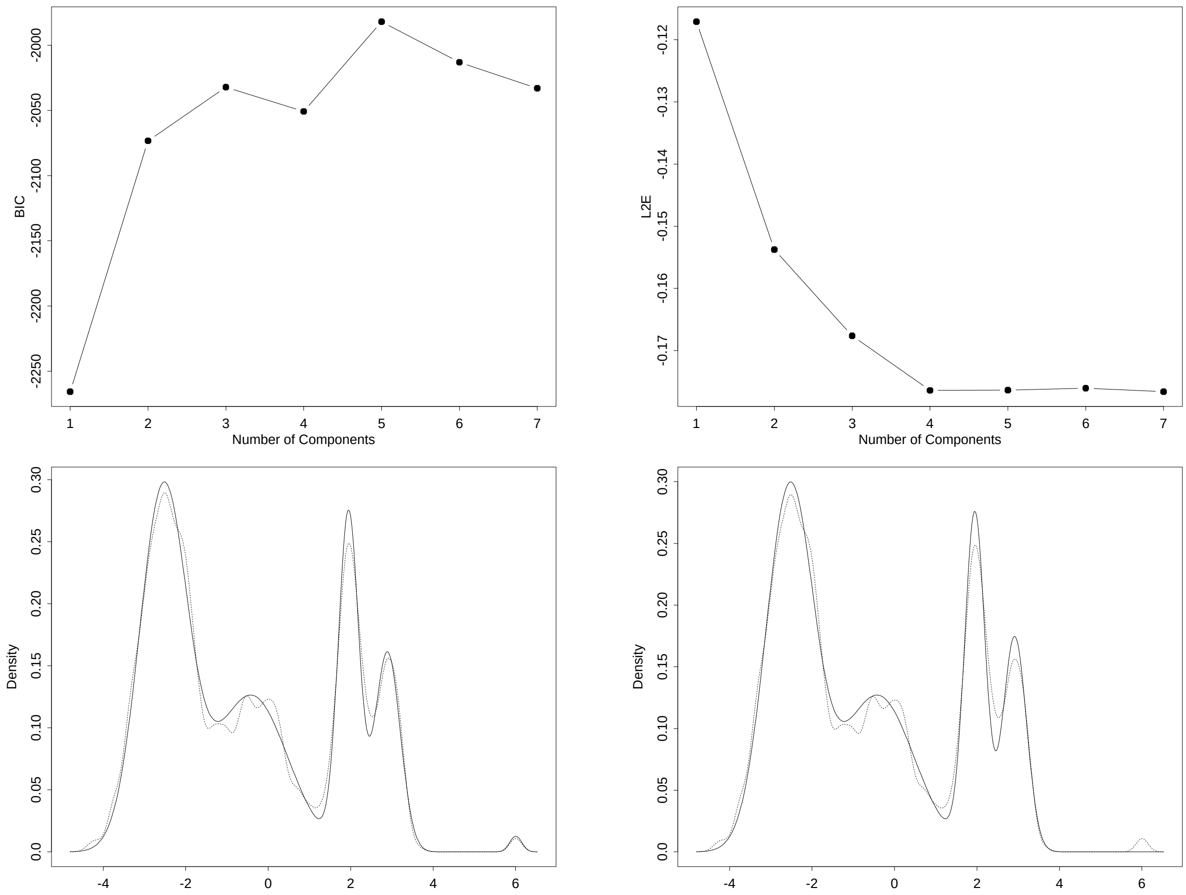

Figure 9: BIC and L_2E values for second example (top left and right). Five-component (bottom left) and four-component (bottom right) mixtures with kernel density estimate overlayed (dotted line).

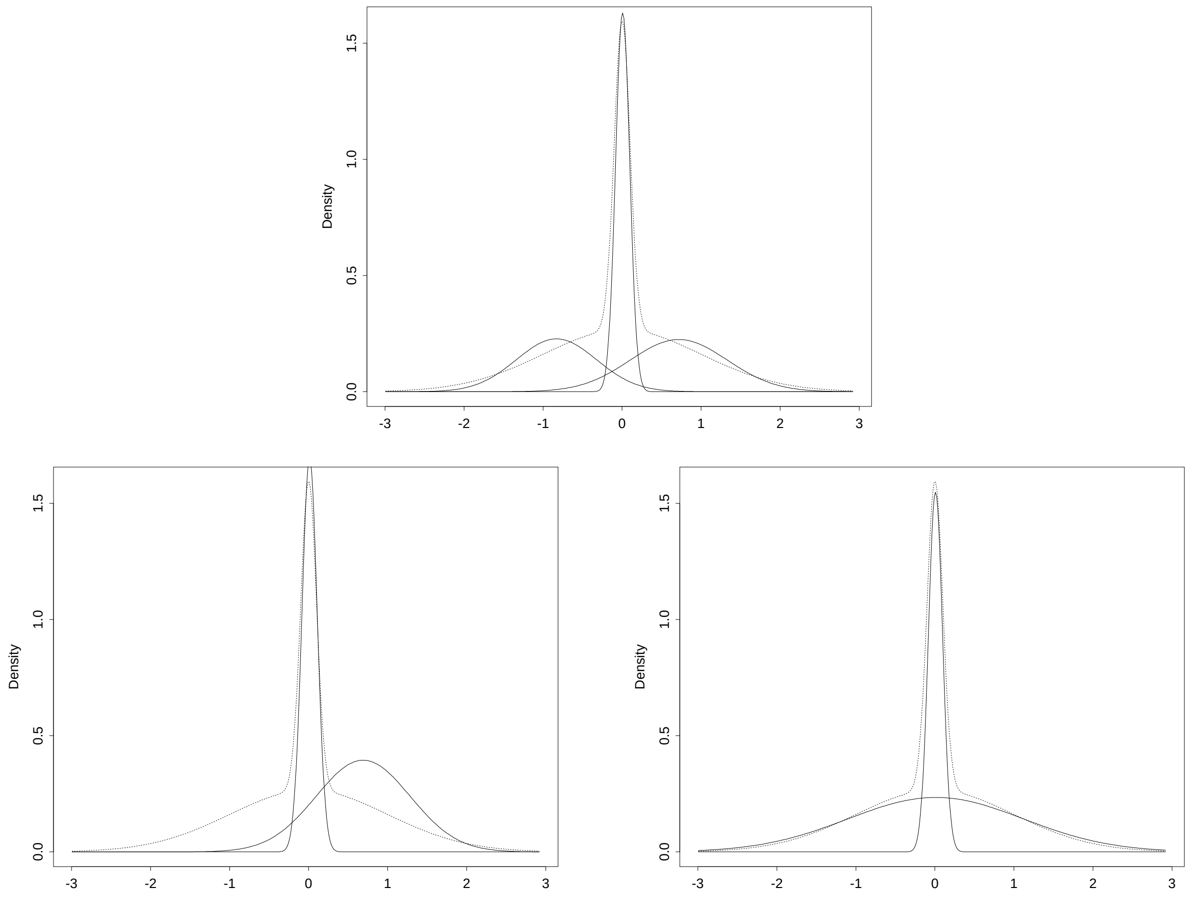

Figure 10: Three-component model for the third example (top). Two-component model without bridging (bottom left). Two-component model with bridging (bottom right). The true density is overlayed in all (dotted line).

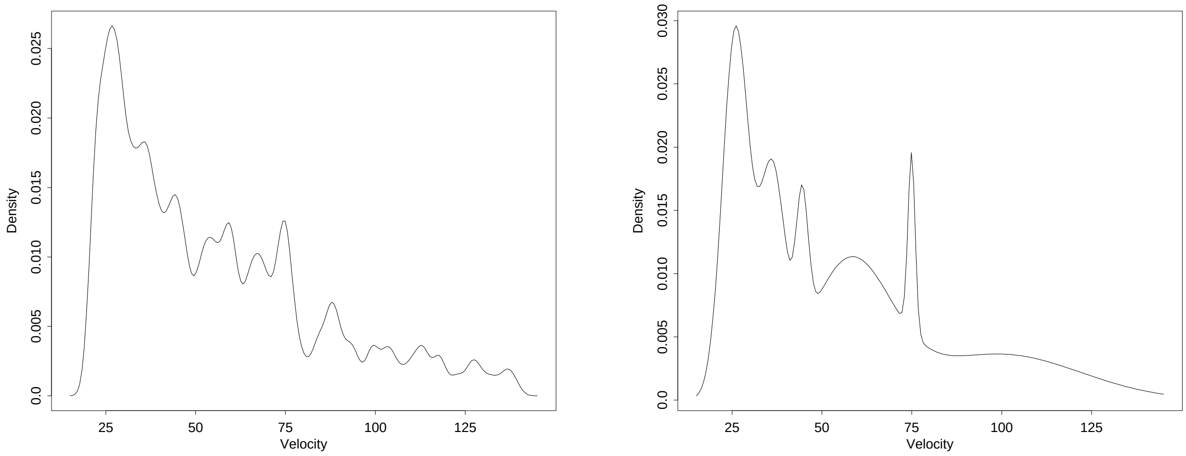

Figure 11: Kernel density estimate of the galaxy data, on the original scale, using the UCV bandwidth (left). IPRA based model with 6 components (right). The component of interest has $\mu = 74.92, \sigma^2 = 1.01^2$, and $w = 0.036$.

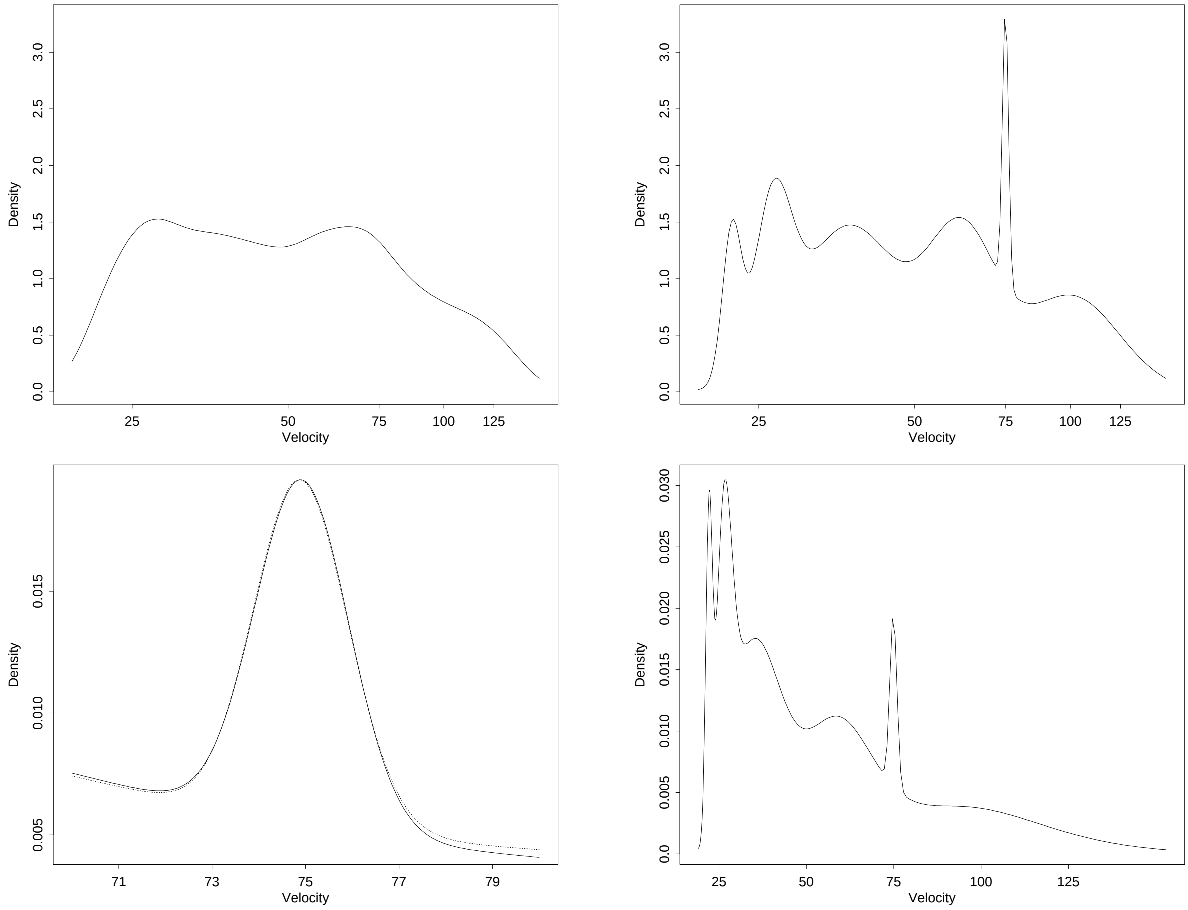

Figure 12: UCV-based density estimate of the galaxy data on a log-velocity scale (top left). Six component IPRA-based model on log-velocity scale (top right) and original scale (bottom right). The component of interest has $\mu = \log_{10}(74.91), \sigma^2 = 0.006^2$, and $w = 0.035$. The component of interest from the analysis on the original scale (solid line) and on a log-velocity scale (dotted line).

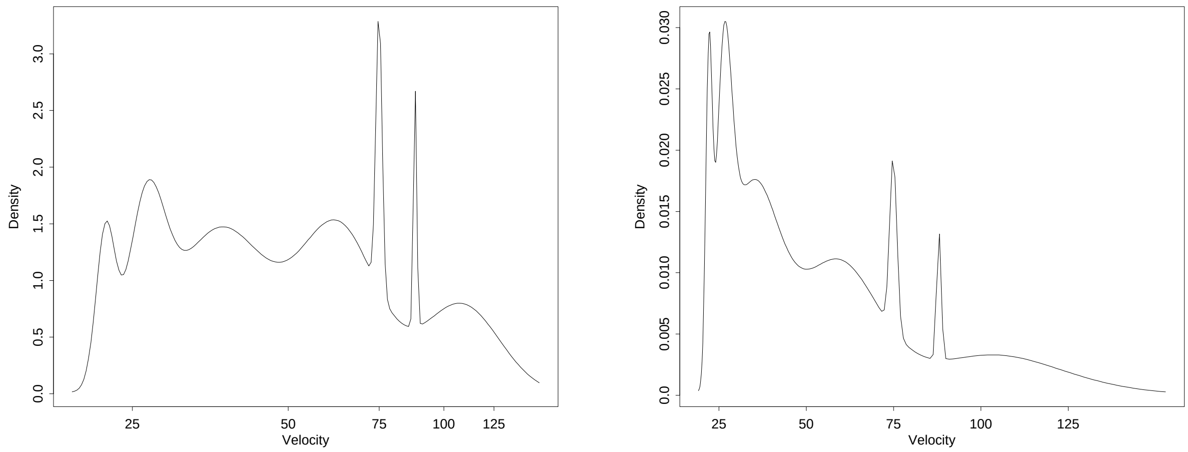

Figure 13: Seven component model for the galaxy data on the log-velocity scale (left) and velocity scale (right). The second sharp component has $\mu = \log_{10}(87.96), \sigma^2 = 0.003^2$, and $w = 0.02$

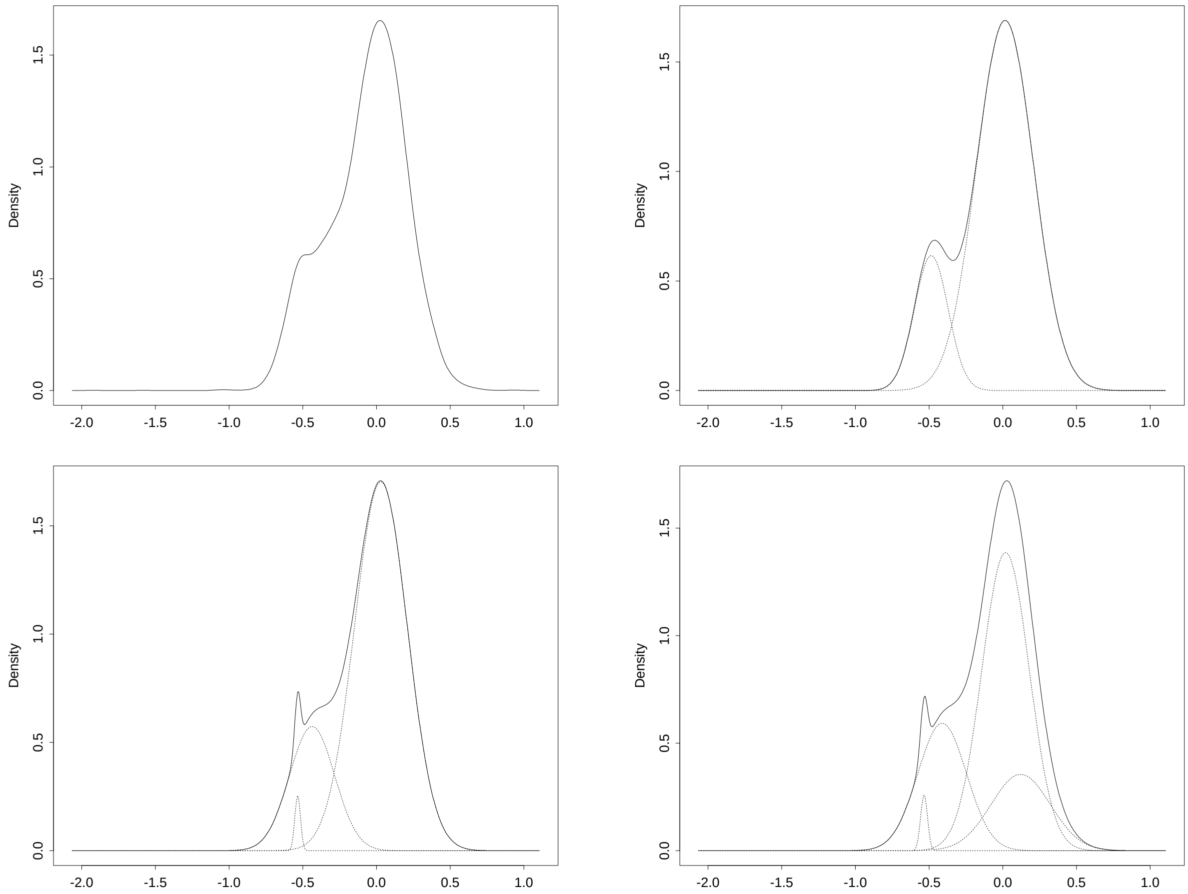

Figure 14: Kernel density estimate of income data using normal reference rule bandwidth (top left). Two-component model at the end of Phase 3 (top right). Three- and four-component models at end of Phase 3 (bottom left and right, respectively).

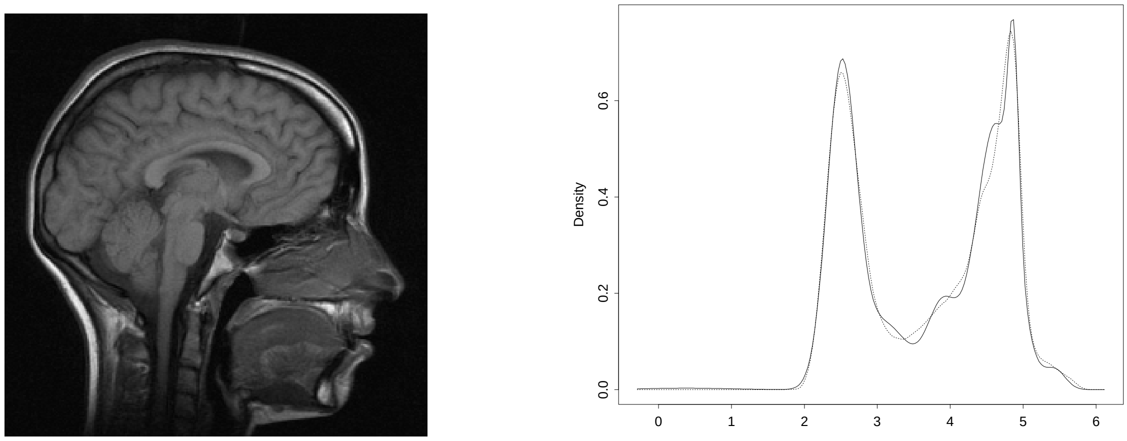

Figure 15: MRI brain scan (left). UCV-based density estimate of the log of the blurred MRI data (dotted line) and seven component IPRA-based model (solid line) (right)

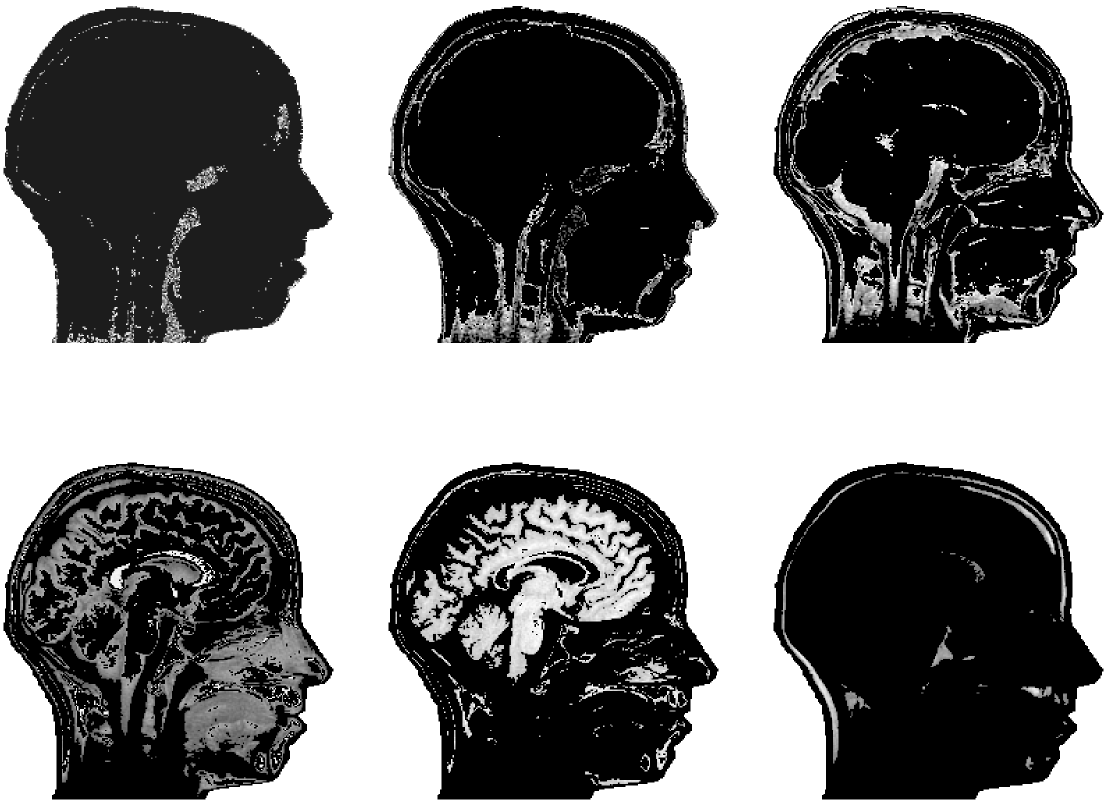

Figure 16: Tissue types from seven component model: Air, Fluid, Fatty Tissue (top row, left to right), Gray Matter/Muscle, White Matter, Bone (bottom row, left to right). (Note that the seventh component consisted of 13 isolated pixels with negative values and is not shown.)