Simplifying Mixture Models with Applications

William F. Szewczyk
National Security Agency
R&E Building, Suite 6515
Ft. George G. Meade, MD 20755-6515

David W. Scott
Department of Statistics, MS-138
Rice University
Houston, TX 77005

Abstract

Mixture models are widely used to model complex distributions; however, one is faced with the challenge of determining an appropriate number of components. This often involves identifying those components that are close enough to be considered the same. In this paper we introduce a new, easily calculated measure of similarity between distributions and illustrate its use in collapsing components. We also show how this measure naturally leads to a new clustering algorithm.

1 Introduction

Mixture models have proven quite useful in modeling complex distributions. Using a small number of normal components, one is able to model distributions that are far from normal (Marron and Wand, 1992). Because of this mixtures can be found in everything from kernel density estimates (Scott, 1992) to classification problems (MacLachlan, 1992) to speech applications (Rabiner, 1989).

Even though a number of books have been written on mixtures and methods for fitting them (Titterington et al. (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 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 usually chooses a sufficiently large number of components and fits, an almost surely, overparameterized model.

In this paper we address the problem of simplifying mixture models by the use of a new similarity measure for densities. We will also show how it is possible to locally refit the parameters of a mixture model, thus making it possible to update large models. These two methods will lead to a new clustering method based on mixtures.

2 Similarity of Densities

Let

$$f(x,\Theta) = \sum_{i=1}^{m} \alpha_i \, f_i(x,\theta_i), \tag{1}$$

with $\alpha_i > 0$ for $i = 1, \ldots, m$, $\sum_{i=1}^{m} \alpha_i = 1$, and each f_i a density with parameters θ_i. We say that equation 1 defines an m-component mixture model. In theory the f_i's could be any densities. In practice they are usually from the same parametric family. In this paper we will work with normal mixtures, although the similarity measure holds for all densities.

We base our similarity measure on the expectation of one density, f_1 with respect to another f_2. By Cauchy-Schwarz,

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

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

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

Since f_1 and f_2 are densities and using equation 2 we see that $0 \le \mathrm{sim}(f_1,f_2) \le 1$.

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 has the form

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

which allows for easy evaluation.

In figure 1 we display the similarity measure for a $N(0,1)$ density and other normals over a range of means and variances. (Because of the symmetry of μ_1 and μ_2 in equation 4 we only need to evaluate the measure for positive values of μ_2.) We notice that the similarity does attain its maximum value for $\mu_2 = 0$ and $\sigma_2^2 = 1$, and

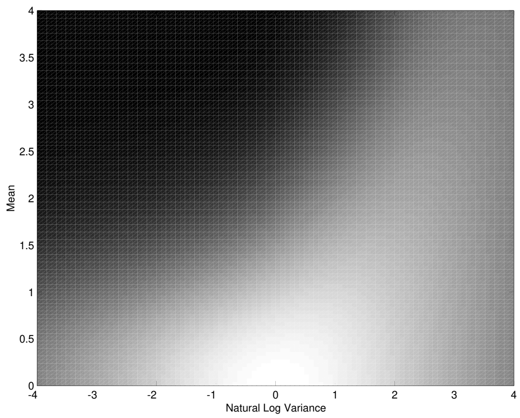

Figure 1: Similarity of $N(0,1)$ with other normal densities

then falls off, more rapidly for a change in mean than a change in variance.

To show the performance of the measure, we drew two samples of size 50 from a standard normal distribution and calculated the similarity measure using the sample moments. Figure 2 shows a histogram of the similarity for 1000 replications. While the measure performs well, since the left tail is dependent on the samplesize, we use the measure more for ordering densities that may be similar instead of for inference.

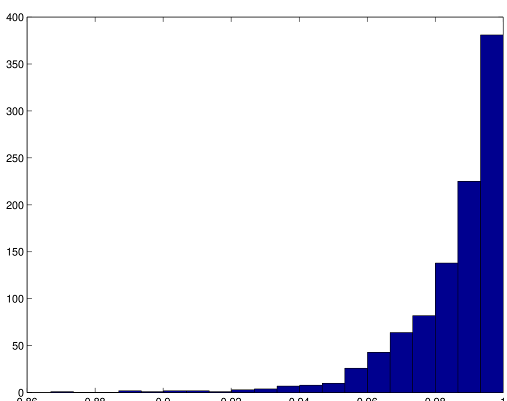

Figure 2: Results of simulation

3 Simplifying Mixtures

As an example of how to use this measure to simplify a mixture model, we fit a 3-component normal mixture model to the data in figure 3 below. Using EM we estimate the parameters of the mixture as

	$\hat{\alpha}$	$\hat{\mu}$	$\hat{\sigma}^2$
Component 1	0.32	-0.41	1.28
Component 2	0.44	0.25	1.10
Component 3	0.24	1.57	0.10

Using these estimates in equation 4 results in the similarities:

	Component 1	Component 2
Component 2	0.91	
Component 3	0.17	0.36

Based on this, components 1 and 2 are good candidates for collapsing.

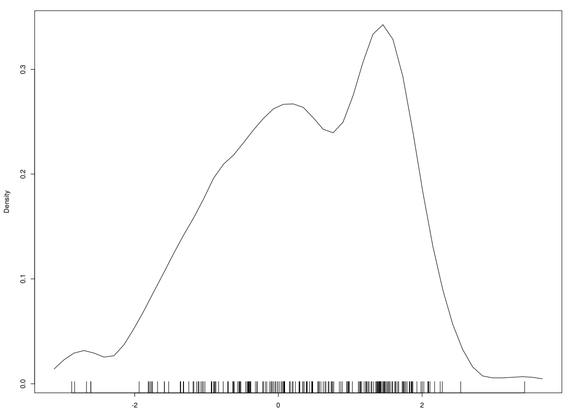

Figure 3: Data and density estimate for first example

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 with two components. However, this becomes very expensive with a large dataset or a model with many components. Instead, we choose to refit the components locally using the L_2E criterion.

3.1 L_2E and Local Refitting

L_2E (Scott, 1999) is a method to estimate parameters of a density by minimizing the L_2 error between the estimated and target densities. The L_2E estimator $\hat{\Theta}$ is defined as,

$$\hat{\Theta} = \underset{\Theta}{\text{argmin}} \int_{-\infty}^{\infty} \left(f(x) - \hat{f}(x;\Theta) \right)^2 dx.$$

For the problem at hand we are trying to minimize the distance between a mixture of two normals and a single normal by finding $\hat{\mu}$ and $\hat{\sigma}^2$ such that

$$\frac{1}{2\hat{\sigma}\sqrt{\pi}} - 2(\alpha_1\phi(\hat{\mu}-\mu_1|0, \sigma_1^2+\hat{\sigma}^2)+\alpha_2\phi(\hat{\mu}-\mu_2|0, \sigma_2^2+\hat{\sigma}^2)) \tag{5}$$

is a minimum, where $\phi(x|\mu, \sigma^2)$ is the $N(\mu, \sigma^2)$ density evaluated at x.

With the parameters for components 1 and 2 as estimated by EM equation 5 is minimized (using the Splus *nlmin* function) by $\hat{\mu} = -0.01$ and $\hat{\sigma}^2 = 0.99$. We take the the weight of this component as the sum of the weights of components 1 and 2, $\hat{\alpha} = 0.76$.

Our final estimate for the two component mixture is

$$\hat{f}(x) = 0.76N(-0.01, 0.99) + 0.24N(1.57, 0.10).$$

The data were actually sampled from

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

4 Fitting Mixtures with IPRA

Fitting a mixture model with EM where m is deliberately inflated is extremely slow and dependent upon the initial values. Here we consider beginning with a mixture model that is in fact the kernel estimate ($m = n$). By combining the similarity measure with local refitting one can fit mixture models using the Iterative Pairwise Replacement Algorithm (IPRA):

1. Start with a kernel density estimate.
2. Combine the two most similar components.
3. Repeat until there is a single component.
4. Choose an appropriate number of components using some criterion, such as the BIC.

To illustrate IPRA we generate 180 samples from $\frac{2}{3}N(0,1)+\frac{1}{3}N(0, \frac{1}{100})$ (figure 4). Starting with an undersmoothed kernel density estimate (Scott, 1992), at each step we replace the two most similar components with a single normal as determined by L_2E. Since at every step of the algorithm we have a likelihood and since, by construction we have no empty components, we are able to calculate the Bayesian Information Criterion (BIC) at every step.

Figure 5 shows the BIC for the last 30 iterations of IPRA. Clearly the two component mixture maximizes the BIC. However, notice that the difference between adjust values for models with from 2 to 30 components

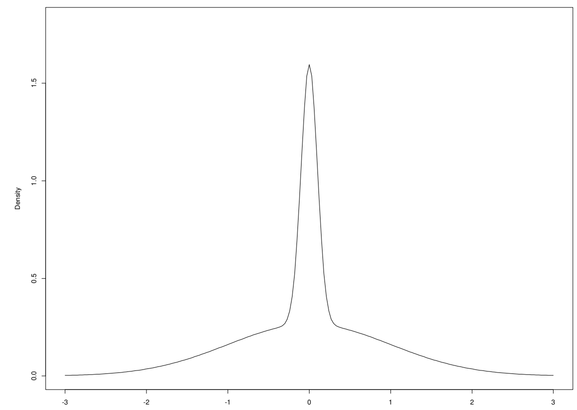

Figure 4: $\frac{2}{3}N(0,1) + \frac{1}{3}N(0, \frac{1}{100})$

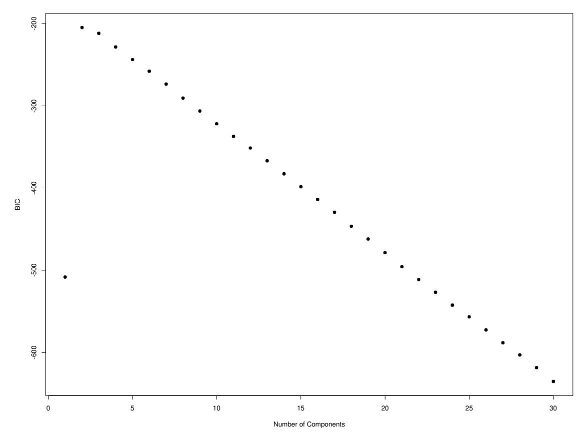

Figure 5: BIC values for the mixtures with 1—30 components

is due to the change in the penalty term in the BIC. In other words, the likelihood is essentially the same for all these models. This occurs in most applications of IPRA and is an indication of how fitting mixtures is such an ill-posed problem.

Choosing the model with 2 components we estimate the mixture as

$$\hat{f}(x) = 0.67N(0.00, 0.98) + 0.33N(-0.02, 0.01).$$

5 Clustering with IPRA

The procedure for fitting mixtures with IPRA is reminiscent of agglomerative clustering. Just a slight modification to IPRA produces a new clustering method:

1. Start with a kernel density estimate.

2. Combine the two most similar clusters, using the members of the cluster in the reestimation.

3. Repeat until there is a single cluster.

4. Choose an appropriate number of clusters using some criterion, such as the BIC.

In the previous section the data were used explicitly to generate the density estimate, and then implicitly subsequently. In order to do clustering we use the data explicitly in Step 2 by using the members of the combined clustering in the reestimation of the parameters. There are a number of ways to do this; we choose to again use the L_2E criterion.

For the estimation of the parameters of a normal density from a sample of n observations, L_2E implies that $\hat{\mu}$ and $\hat{\sigma}$ are those values that minimize

$$\frac{1}{2\sigma\sqrt{\pi}} - \frac{2}{n}\sum_{i=i}^{n} \phi(x_i;\mu,\sigma^2). \tag{6}$$

In our application, the summation ranges over only those observations in the particular cluster.

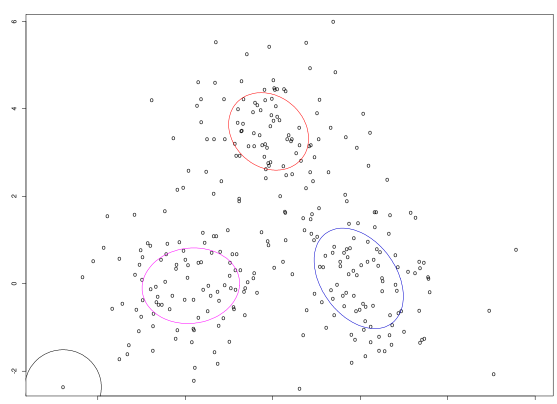

Figure 6: Result of using IPRA for clustering

Figure 6 shows the result of clustering IPRA on 300 samples drawn equally from bivariate normals with means $(0,0)$, $(4,0)$, and $(2, 2\sqrt{3})$ and identity covariance matrices. The BIC suggests 4 clusters, with the fourth one being the singleton in the lower left corner. The solution with 3 clusters combines the clusters centered $(0,0)$ and $(4,0)$.

This result can be explained by the robustness property of L_2E (Scott, 1999) which gives little weight to observations "far" from the mean. Because of this it may underestimate the covariance matrix, as in this case. The similarity measure therefore judges the singleton as being far enough removed from the cluster centered at the origin, due to the distance of the point from the mean and the size of the variance of the cluster, to prefer combining the clusters centered $(0,0)$ and $(4,0)$. In the next section we will indicate some possible solutions to this problem.

6 Discussion

In this paper we have introduced a new similarity measure between densities. This measure, along with local refitting, makes it possible to simplify a mixture model without having to refit the entire model. It also leads to a procedure IPRA which can be used for fitting mixtures and clustering.

The local nature of the algorithm and the closed form of the measure for some densities, such as the normal, makes IPRA easy to apply to large datasets. In one classification application, using this similarity measure, a mixture of 2048 components in 39 dimensions was reduced to 1024 components without any significant loss in discrimination ability. Of course, when working with such large datasets it would be at least as much work to calculate all pairwise similarities as it would be to fit a smaller model. In that case, it is sufficient to restrict one's attention to small neighborhoods around the component of interest.

Another improvement to the algorithm for dealing with large datasets is to use some simple clustering technique, such as k-means, to seed IPRA with a smaller number of components. Starting with, say 4 or 5 times the number of clusters one expects to find, should produce good results.

As stated in the previous section, a straightforward application of (data-driven) L_2E to fit the parameters of the components can lead to a covariance matrix that is too small. There are a few variations of the basic algorithm, such as replacing the normal density with a truncated normal, that could help.

The easiest one to implement would be to include the weight of the combined component, $\hat{\alpha}$ in the L_2E criterion. Equation 5 then becomes

$$\frac{\hat{\alpha}^2}{2\hat{\sigma}\sqrt{\pi}} - 2\hat{\alpha}(\alpha_1\phi(\hat{\mu}-\mu_1|0,\sigma_1^2+\hat{\sigma}^2)+\alpha_2\phi(\hat{\mu}-\mu_2|0,\sigma_2^2+\hat{\sigma}^2))$$

Of course, after $\hat{\alpha}$ is estimated all weights must be renormalized.

Acknowledgment

The second author's research was supported in part by the National Science Foundation grant DMS 96-26187.

References

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

Marron, J. S. and Wand, M. P. (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.

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

Scott, D. W. (1999). Parametric modeling by minimum L_2 error. Technical Report 98-3, Department of Statistics, Rice University.

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