2004

Visual Comparison of Datasets Using Mixture Decompositions

Alan Gous

Andreas Buja
University of Pennsylvania

Recommended Citation

Gous, A., & Buja, A. (2004). Visual Comparison of Datasets Using Mixture Decompositions. *Journal of Computational and Graphical Statistics, 13* (1), 1-19. http://dx.doi.org/10.1198/1061860043119

Visual Comparison of Datasets Using Mixture Decompositions

Abstract
This article describes how a mixture of two densities, f_0 and f_1, may be decomposed into a different mixture consisting of three densities. These new densities, f_+, f_-, and $f_=$, summarize differences between f_0 and f_1: f_+ is high in areas of excess of f_1 compared to f_0; f_- represents deficiency of f_1 compared to f_0 in the same way; $f_=$ represents commonality between f_1 and f_0. The supports of f_+ and f_- are disjoint. This decomposition of the mixture of f_0 and f_1 is similar to the set-theoretic decomposition of the union of two sets A and B into the disjoint sets $A\backslash B$, $B\backslash A$, and $A \cap B$. Sample points from f_0 and f_1 can be assigned to one of these three densities, allowing the differences between f_0 and f_1 to be visualized in a single plot, a visual hypothesis test of whether f_0 is equal to f_1. We describe two similar such decompositions and contrast their behavior under the null hypothesis $f_0 = f_1$, giving some insight into how such plots may be interpreted. We present two examples of uses of these methods: visualization of departures from independence, and of a two-class classification problem. Other potential applications are discussed.

Keywords
classification, data visualization, density estimation, exploratory data analysis, mixture decomposition

Disciplines
Applied Statistics | Other Statistics and Probability | Statistics and Probability

Visual Comparison of Datasets Using Mixture Decompositions

Alan GOUS and Andreas BUJA

This article describes how a mixture of two densities, f_0 and f_1, may be decomposed into a different mixture consisting of three densities. These new densities, f_+, f_-, and $f_=$, summarize differences between f_0 and f_1: f_+ is high in areas of excess of f_1 compared to f_0; f_- represents deficiency of f_1 compared to f_0 in the same way; $f_=$ represents commonality between f_1 and f_0. The supports of f_+ and f_- are disjoint. This decomposition of the mixture of f_0 and f_1 is similar to the set-theoretic decomposition of the union of two sets A and B into the disjoint sets $A\backslash B$, $B\backslash A$, and $A \cap B$. Sample points from f_0 and f_1 can be assigned to one of these three densities, allowing the differences between f_0 and f_1 to be visualized in a single plot, a visual hypothesis test of whether f_0 is equal to f_1. We describe two similar such decompositions and contrast their behavior under the null hypothesis $f_0 = f_1$, giving some insight into how such plots may be interpreted.

We present two examples of uses of these methods: visualization of departures from independence, and of a two-class classification problem. Other potential applications are discussed.

Key Words: Classification; Data visualization; Density estimation; Exploratory data analysis; Mixture decomposition.

1. INTRODUCTION

Figure 1 is a plot of $n = 329$ metropolitan areas in the United States. A score measuring housing cost in the area is plotted on the y-axis, and a score for the quality of the transportation infrastructure is plotted on the x-axis.

Are these two scores independent of one another? A standard test of, say, a zero correlation, confirms that they are not. This is also clear purely from visual evidence, if we compare this plot to Figure 2. The latter plot is of the same data, but with the x-values of all the points randomly permuted, while keeping the y-values fixed. This then is a sample of size n from the *permutation* distribution defined by the data, the product distribution

Alan Gous is Director of Research and Development, Cariden Technologies, Inc., 888 Villa Street, Suite 200, Mountain View, CA 94041. Andreas Buja is Professor, Statistics Department, The Wharton School, University of Pennsylvania, 471 Huntsman Hall, Philadelphia, PA 19104.

Journal of Computational and Graphical Statistics, Volume 13, Number 1, Pages 1–19
DOI: 10.1198/1061860043119

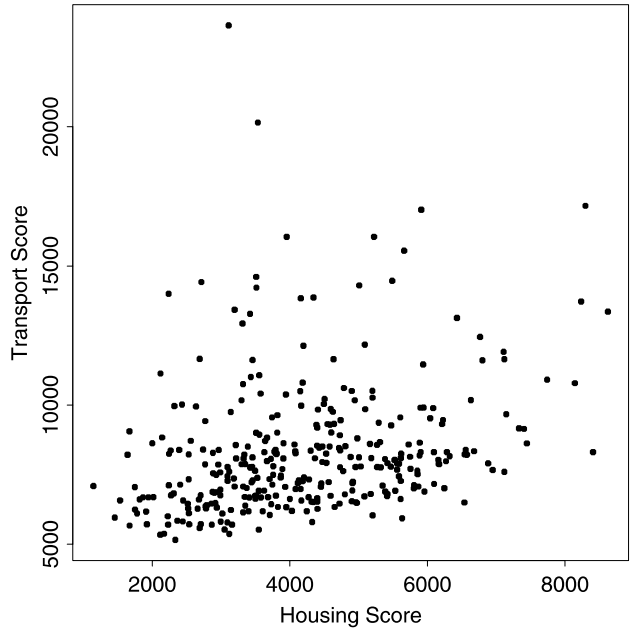

Figure 1. Housing and transportation ratings of 329 places to live in the U.S.

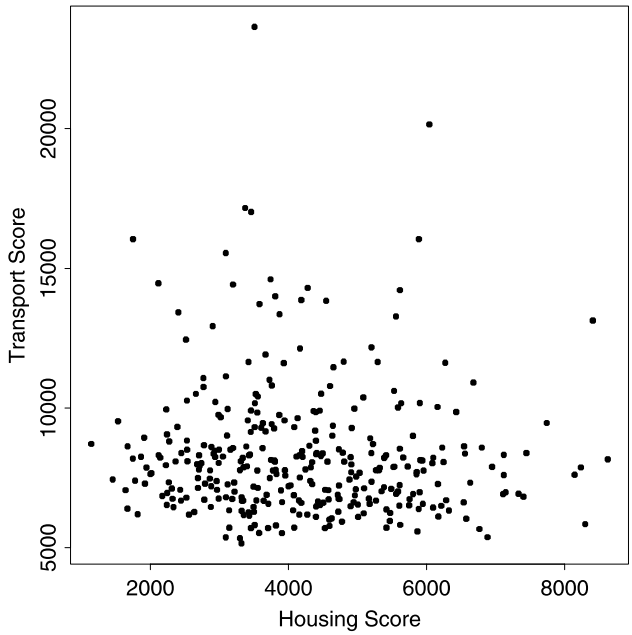

Figure 2. Places ratings data, with data on one axis randomly permuted.

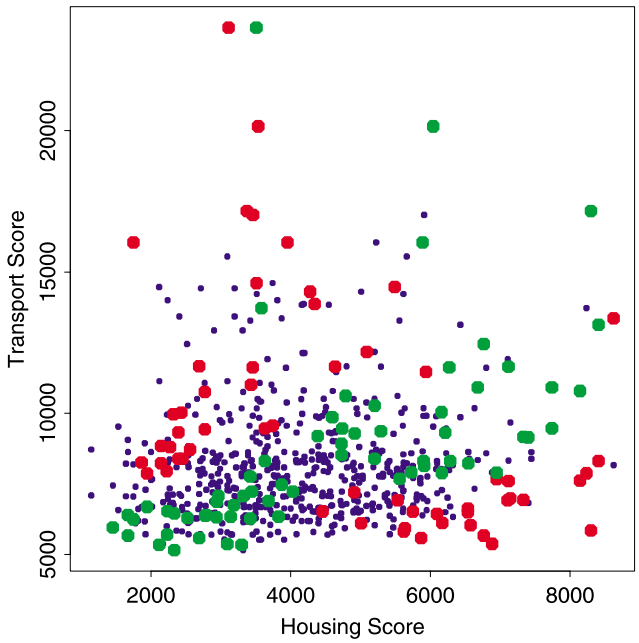

Figure 3. Combined data of Figures 1 and 2, with points colored according to allocation in the L_1 mixture model.

of the data margins, for which the x- and y-values are independent. There seems to be visual evidence that the two datasets creating the two plots are from different distributions, particularly when comparing the lower center and lower right-hand sides of the plots.

Figure 3 plots both sets of data on the same axes, and colors the points according to a scheme which provides immediate visual information about the differences between the two distributions. The combined data are regarded as having been drawn from a mixture of three distributions, and are colored accordingly. The distribution from which the green points are drawn has been defined so that it has high density where the original density is high compared to the permutation density. The red points, on the other hand, are drawn from a distribution defined so that it has high density where the original density is low compared to the (draw from the) permutation density. Areas of red or green points in the plot then provide evidence for differences between the two distributions. Points classified as being drawn from the third density in this mixture are colored blue. This density is defined to be high where the two distributions are similar, and so reflects a sort of consensus between the two.

With this interpretation, Figure 3 shows clearly the dependence between the x and y variables: The green points are funneled from the lower left to the upper right between two groups of red points in the the upper left and lower right. Keeping in mind that green/red points indicate preponderance/deficiency, of actual compared to permuted data, the plot lets us perceive the nature of the dependence in more detail than the plot of the raw data in Figure 1.

While interpreting the data, we must bear in mind the sampling variation that has been introduced into Figure 3 by the single draw from the permutation distribution used in its

creation. The patterns described above do, however, persist through multiple draws. Section 5 presents a second application of this visualization scheme in which such permutation sampling variation does not appear. The reader should keep in mind that the comparison of real and permuted data is just an illustration, albeit a useful one we think, of the proposed method for decomposing and visually comparing two distributions.

There exists a second source of sampling variation, given the original data, which is always present in these visualizations, and which will be described with the algorithm itself in Section 3.

A remark on plotting: Graphs such as Figure 3 are prone to overplotting. The order in which points are drawn is therefore important. It is recommended that points be drawn in reverse order of importance, such that more important points are drawn later to allow them to overplot the previously drawn less important points. In the present situation this means plotting green and red points after the blue ones because agreement between two distributions is less informative than their disagreement, represented here by the green and red points.

This article describes schemes such as that of Figure 3 for visualizing, in a single plot, the differences between two datasets. Besides the simple example above, which will be used for illustration throughout much of the article, there are a number of situations in which such plots may be useful:

1. Any testing situation where a unique null distribution can be simulated. The above example, in which the null is the permutation distribution, is a special case.
2. Two-class classification problems. An example is presented in Section 5.
3. Process data: visualizing differences between today's data and yesterday's data to monitor changes in the behavior of a process.
4. Model diagnostics: comparing actual data to data predicted by a given model, or to parametric bootstrap samples from a model fitted to the data.

Section 2 defines two schemes for using mixture decompositions to define differences between the two constituent densities. Section 3 describes how points may be drawn from the densities in the mixture. Section 4 describes some important theoretical differences between the schemes. An application of these methods to classification is presented in Section 5.

2. TWO CANONICAL MIXTURE DECOMPOSITIONS

Two univariate densities are depicted in Figure 4(a). One, which we will call f_1, is a mixture of two Gaussians of equal variance, with equal mixing probabilities. The other, f_0, is a single Gaussian which has been "fit" to f_1, matching its mean and variance. As a model of f_1, f_0 puts excess mass in the center of the density and in the tails. This excess is counterbalanced by deficiency of mass in the two regions between the center and the tails. Parts (b), (c), and (d) of Figure 4 are three different representations of these regions of excess and deficient mass.

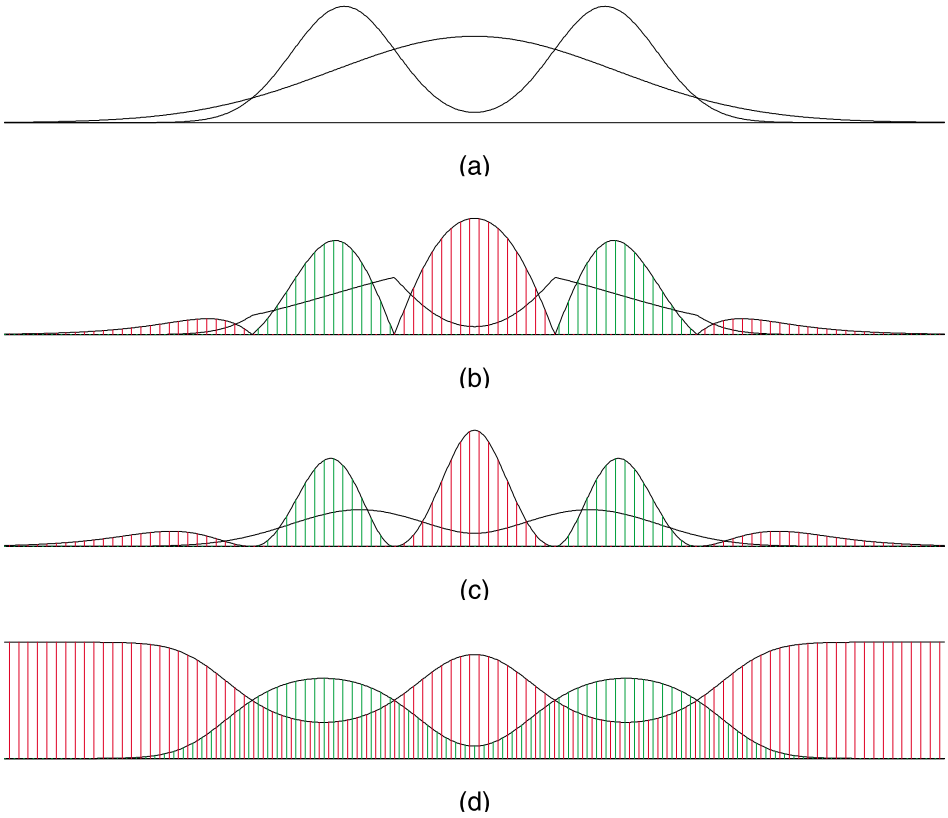

Figure 4. A simple mixture decomposition: (a) the original two densities. (b) The L_1 decomposition. The density shaded by green lines is f_+, that shaded by red lines is f_-. The unshaded density is $f_=$. (c) The L_2 decomposition with the same labeling as (b). (d) The functions F_+ and F_- are defined in (2.18).

2.1 THE CANONICAL L_1 MIXTURE DECOMPOSITION

Treating f_1 and f_0 simply as positive functions on $\mathbb{R}$, (or $\mathbb{R}^d$, $d \geq 1$ in general), we can write their sum as

$$f_0 + f_1 = (f_1 - f_0)_+ + (f_0 - f_1)_+ + 2\min\{f_0, f_1\}, \tag{2.1}$$

where min is the pointwise minimum, and for any function g, g_+ denotes $\max\{g, 0\}$. The three terms on the right-hand side of (2.1) are all positive functions, dominated by $f_0 + f_1$. Let

$$\omega = \int (f_1 - f_0)_+ \, d\mu = \int (f_0 - f_1)_+ \, d\mu, \tag{2.2}$$

where μ is Lebesgue measure on $\mathbb{R}^d$. If $f_1 \neq f_0$, so $\omega > 0$, then the following are densities on $\mathbb{R}^d$:

$$f_+ \quad = \quad \frac{1}{\omega}(f_1 - f_0)_+, \tag{2.3}$$

$$f_- = \frac{1}{\omega}(f_0 - f_1)_+, \tag{2.4}$$

$$f_= = \frac{1}{1-\omega}\min\{f_0, f_1\}. \tag{2.5}$$

The mixture density of f_1 and f_0, each with equal probability, may be written as

$$\frac{1}{2}(f_1 + f_0) = \frac{\omega}{2}f_+ + \frac{\omega}{2}f_- + (1-\omega)f_= , \tag{2.6}$$

that is, decomposed into a mixture of the three densities (2.3)–(2.5) with weights which are functions of ω.

These three densities, for the case of f_1 and f_0 in Figure 4(a), are shown in Figure 4(b). Note that f_+ is high where f_1 has excess mass compared to f_0, and f_- is high where f_1 is deficient in mass compared to f_0. The density $f_=$, proportional to the pointwise minimum of f_0 and f_1, is a measure of consensus between the two densities.

Because

$$2\omega = \|f_1 - f_0\|_1 , \tag{2.7}$$

the relative weights of the mixture densities on the right-hand side of (2.6) are determined by the L_1 distance between f_0 and f_1. We will call (2.6) the L_1 *mixture decomposition* of the densities f_1 and f_0.

The more these two densities differ from one another, the higher the combined weights of f_+ and f_- in the mixture will be. Note also that

$$1 - \omega = \int_{f_0 \geq f_1} f_1 \, d\mu + \int_{f_1 > f_0} f_0 \, d\mu , \tag{2.8}$$

which is the Bayes misclassification rate for the classification of a sample point drawn from f_1 or f_0 with equal prior probabilities.

Any sample value drawn from the mixture on the left-hand side of (2.6) can be interpreted as having been drawn from one of the three densities on the right-hand side of (2.6). Each point in Figure 3 has been colored according to such an allocation, as will be described in Section 3.

2.2 THE CANONICAL L_2 MIXTURE DECOMPOSITION

Another decomposition of f_1 and f_0, similar to (2.6), can be derived by replacing (2.1) with the decomposition

$$f_0 + f_1 = (\sqrt{f_1} - \sqrt{f_0})_+^2 + (\sqrt{f_0} - \sqrt{f_1})_+^2 + 2\sqrt{f_0 f_1} . \tag{2.9}$$

For functions g we write g_+^2 as shorthand for $(g_+)^2$. When expanding the squares on the right hand side, the two cross-product terms have disjoint nonzero regions, so the terms add up to $2\sqrt{f_0 f_1}$.

Again, if $f_1 \neq f_0$, the three terms on the right-hand side can be normalized into densities, this time by defining

$$\omega_+ = \int (\sqrt{f_1} - \sqrt{f_0})_+^2 \, d\mu, \quad \omega_- = \int (\sqrt{f_0} - \sqrt{f_1})_+^2 \, d\mu \,. \tag{2.10}$$

Note $\omega_+ \neq \omega_-$ in general. Define

$$f_+ = \frac{1}{\omega_+}(\sqrt{f_1} - \sqrt{f_0})_+^2, \tag{2.11}$$

$$f_- = \frac{1}{\omega_-}(\sqrt{f_0} - \sqrt{f_1})_+^2, \tag{2.12}$$

$$f_= = \frac{2}{2 - (\omega_+ + \omega_-)} \sqrt{f_0 f_1} \,. \tag{2.13}$$

Then we can write

$$\frac{1}{2}(f_1 + f_0) = \frac{\omega_+}{2} f_+ + \frac{\omega_-}{2} f_- + \left(1 - \frac{\omega_+ + \omega_-}{2}\right) f_= \,, \tag{2.14}$$

again decomposing the mixture of f_1 and f_0 into a mixture of three new densities representing excess and deficiency of f_1 with respect to f_0, and a measure of consensus between the two. The *consensus* density, $f_=$, is in this case proportional to the geometric mean of f_0 and f_1. For the f_1 and f_0 in Figure 4(a), these three densities are shown in Figure 4(c).

Note that

$$\omega_+ + \omega_- = \|\sqrt{f_1} - \sqrt{f_0}\|_2^2 \,, \tag{2.15}$$

the Hellinger distance between f_1 and f_0. For this reason we will call (2.14) the L_2 *mixture decomposition* of f_1 and f_0.

For simplicity we have used the same notation for the densities in this decomposition as for those defined in (2.3)–(2.5). The ambiguity is useful because the results in the next section may be applied to either case. It will also be useful, in the L_1 case, to define ω_+ and ω_- both to be equal to ω in (2.2).

Note: The L_1 and L_2 decompositions are invariant under changes in the underlying measure in the following sense. Let ν be a measure on $\mathbb{R}^d$ and let g_0, g_1 be densities with respect to ν so that

$$f_0 \, d\mu = g_0 \, d\nu, \quad f_1 \, d\mu = g_1 \, d\nu. \tag{2.16}$$

Let g_+, g_- and $g_=$ be defined with respect to g_0 and g_1 using either (2.3)–(2.5) or (2.11)–(2.13). Then clearly

$$f_+ \, d\mu = g_+ \, d\nu, \quad f_- \, d\mu = g_- \, d\nu, \quad f_= \, d\mu = g_= \, d\nu. \tag{2.17}$$

The probability distributions of these densities remain the same under this transformation, as, therefore, will samples drawn from the densities.

2.3 COMPARISON WITH THE TWO-CLASS CLASSIFICATION PROBLEM

There are other, more conventional functions of f_0 and f_1 besides f_+, f_- and $f_=$ which measure excess or deficiency of one over the other. For example, the two functions

$$F_+ = \frac{f_1}{f_0 + f_1}, \quad F_- = \frac{f_0}{f_0 + f_1}, \tag{2.18}$$

have been suggested in a very similar context by Friedman and Fisher (1999). Figure 4(d) illustrates these functions for the f_0 and f_1 of Figure 4(a). $F_+(x)$ and $F_-(x)$ are the probabilities that a sample X drawn from the mixture $\frac{1}{2}(f_0 + f_1)$ is from f_1 or f_0 respectively, given that $X = x$.

There are a number of advantages of the L_1 and L_2 decompositions over F_+ and F_-, however. First, F_+ and F_- are not densities with respect to Lebesgue measure μ, so one cannot visualize them by plotting sample points as we do for f_+, f_-, and $f_=$, using the method explained in the next section. Second, these functions provide a "multiplicative" comparison of f_0 and f_1, as opposed to the "additive" comparison provided by f_+ and f_-. As can be seen from Figure 4(d) this multiplicative comparison emphasizes differences in the very low density regions of f_0 and f_1, which are perhaps not as important as those in the high density regions, and are certainly more difficult to estimate. Last, the L_1 and L_2 decompositions define a third function, $f_=$, which describes the common variation in the heights of the densities f_0 and f_1, leaving f_+ and f_- to describe just the differences between the densities. The rôle of $f_=$ will be described further in Section 4.

3. SAMPLING FROM THE MIXTURE COMPONENTS

In the example of Section 1, we do not know either f_1, the joint distribution of the two variables, or f_0, the permutation distribution on these variables. Instead we have two samples from which they may be estimated. In general, let

$$X_1, \ldots, X_m \overset{\text{iid}}{\sim} f_1, \quad Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_0 \tag{3.1}$$

be defined for densities f_1 and f_0 on $\mathbb{R}^d$. In the example in Section 1 we have $d = 2$ and $m = n = 329$.

Simple estimates of f_1 and f_0 are the kernel density estimates

$$\hat{f}_1(x) = \frac{1}{mh^d} \sum_{i=1}^{m} K\left(\frac{X_i - x}{h}\right), \tag{3.2}$$

$$\hat{f}_0(y) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{Y_i - y}{h}\right). \tag{3.3}$$

Here K is some kernel (a Gaussian in $\mathbb{R}^d$, for example) and h is some bandwidth. In the examples in this article we chose h using "Scott's rule" (Scott 1992) on the combined sample size: we set

$$h = (n + m)^{-1/(d+4)}, \tag{3.4}$$

after first normalizing this combined sample to have unit variance in each coordinate. This rule is optimal in a certain sense if the underlying distribution is multivariate normal, and seemed to work adequately here.

Decompositions equivalent to (2.6) and (2.14) for f_1 and f_0 can be defined for $\hat{f}_1$ and $\hat{f}_0$, resulting in estimated densities $\hat{f}_+$, $\hat{f}_-$, and $\hat{f}_=$, and estimated constants $\hat{\omega}_+$ and $\hat{\omega}_-$. (The latter turn out to be irrelevant for our procedure; see below.) The estimated densities $\hat{f}_+$ and $\hat{f}_-$ may be used to characterize differences between the densities f_1 and f_0. These differences can be visualized using samples from $\hat{f}_+$ and $\hat{f}_-$. The following algorithm will sample from each of $\hat{f}_+$, $\hat{f}_-$, and $\hat{f}_=$, in the proportions of the mixture decomposition, with a total sample size N:

Repeat for i from 1 to N:

1. Draw $Z_i = z_i$ from $\frac{1}{2}(\hat{f}_1 + \hat{f}_0)$.
2. With probability

$$P_{+/-}(z_i) = \frac{\omega_+ \hat{f}_+(z_i) + \omega_- \hat{f}_-(z_i)}{\hat{f}_1(z_i) + \hat{f}_0(z_i)} \tag{3.5}$$

assign z_i to $\begin{cases} \text{the } \hat{f}_+ \text{ sample,} & \text{if} \quad \hat{f}_1(z_i) > \hat{f}_0(z_i) \\ \text{the } \hat{f}_- \text{ sample,} & \text{if} \quad \hat{f}_0(z_i) > \hat{f}_1(z_i) \end{cases}$

Otherwise, assign z_i to the $\hat{f}_=$ sample.

This algorithm can be applied using either the L_1 or L_2 mixture decompositions definitions. Note that we do not need to calculate ω_+ or ω_- since their values cancel with the definitions of $\hat{f}_+$ and $\hat{f}_-$ in the numerator of (3.5):

$$P_{+/-} \quad = \quad \frac{(\hat{f}_1 - \hat{f}_0)_+ \; + \; (\hat{f}_0 - \hat{f}_1)_+}{\hat{f}_1 \; + \; \hat{f}_0} \quad \text{and} \quad \frac{(\hat{f}_1^{1/2} - \hat{f}_0^{1/2})_+^2 \; + \; (\hat{f}_0^{1/2} - \hat{f}_1^{1/2})_+^2}{\hat{f}_1 \; + \; \hat{f}_0}$$

in the L_1 and L_2 case, respectively. Also, the numerator evaluates to zero if $\hat{f}_0 = \hat{f}_1$, in which case every z_i is allocated to $\hat{f}_=$.

The starting point of our procedure is of course data consisting of two samples, one each from f_1 and f_0. Although we state in Step 1 that samples are drawn from the estimate $\frac{1}{2}(\hat{f}_1 + \hat{f}_0)$, we prefer to avoid bias incurred in estimation by using the original data to form samples from $\frac{1}{2}(f_1 + f_0)$ and thus limiting the effects of estimation to the color scheme. If the sample sizes are equal, $m = n$, as in our example, or if they are approximately equal, then instead of sampling from the mixture in Step 1 we can simply set $N = n + m$ and let $Z_1, \ldots, Z_N$ be the combined sample of the X_i and Y_i. If m and n are widely different but the smaller (m, say) is large enough to be informative when displayed as a scatterplot, one could use subsamples of size m out of n from the larger sample.

If $m \approx n$ and the pooled data samples are used, the modified algorithm amounts to an assignment of each point in the original sample (3.1) to either $\hat{f}_+$, $\hat{f}_-$, or $\hat{f}_=$. Figure 3 is such an assignment. Points selected as samples from $\hat{f}_+$ have been colored green, those from $\hat{f}_-$ red, and those from $\hat{f}_=$ blue.

Note that since the algorithm samples randomly, different colorings are possible. It is certainly recommended that the algorithm be applied more than once to the data, and

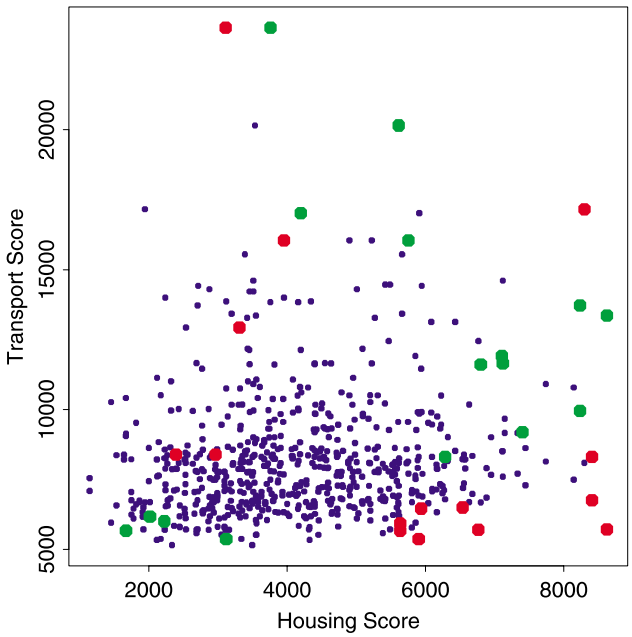

Figure 5. Combined data of Figures 1 and 2, with points colored according to the L_2 mixture model.

the resulting visualizations compared, to assess the impact of this sampling variation on interpretations of the data.

Figure 5 plots the same data as Figure 3, but colored according to a sampling from an L_2 mixture decomposition. Both plots characterize the positive dependence between the two variables. The L_2 method samples fewer red or green points, and so while the same pattern is there it is less distinct. The L_2 mixture decomposition has an important advantage over the L_1 decomposition, though, which will be explained in the next section.

Remarks:

1. One could raise the question of why we visualize samples when the results of our technique are really three density functions. One could argue that one should display these functions with function visualization, as opposed to samples with data visualization. This is a sensible proposal, but there are two arguments in favor of visualizing samples: (1) Comparing samples in more than two dimensions is more easily done than comparing functions of more than two variables. We use simple scatterplots of two variables only for didactic purposes; comparison of samples with more than two variables can be achieved with multivariate data visualizations such as scatterplot matrices and grand tours (Swayne, Cook, and Buja). (2) As already mentioned, when $m \approx n$ and Step 1 is replaced with using the pooled data samples, one prefers to see the original data and relegate the estimated densities to auxiliary devices.
2. We emphasize again that the general goal of this article is to develop the idea,

and illustrate the use of, decomposing two densities into a mixture of three components, each with a well-defined interpretation. The comparison of real and null data based on data permutations is just one of several possible applications. There exist many other techniques for detecting dependence, most in the form of tests. If visual testing of independence were our primary goal, we should refine our method by replacing the kernel density of the permuted data with the product of densities estimated from the two marginal distributions. That is, if the permuted data are the samples $Y_i = (Y_i^{(1)}, Y_i^{(2)})$, put $\hat{f}_0(y^{(1)}, y^{(2)}) = \hat{f}^{(1)}(y^{(1)}) \cdot \hat{f}^{(2)}(y^{(2)})$. Note that this estimate is independent of the specific draw $(Y_i)_{i=1\ldots n}$ from the permutation distribution because the marginal values are the same as those of the original data $(X_i)_{i=1\ldots n}$. For display, we would still use the same single draw $(Y_i)_{i=1\ldots n}$ from the permutation distribution, but the coloring would be derived from a density estimate $\hat{f}_0$ that mimics the independence assumption exactly. (We thank an anonymous referee for suggesting this point.)

4. BEHAVIOR UNDER THE NULL HYPOTHESIS

Plots such as Figures 3 and 5 provide visual information about differences in the two densities f_1 and f_0. To be able to interpret this information we need to know what we would expect to see in such plots if the two densities were the same.

Even if $f_1 = f_0$, we have $\hat{f}_1 \neq \hat{f}_0$ in general so points will be sampled from $\hat{f}_+$ and $\hat{f}_-$, not just $\hat{f}_=$, in the algorithm presented in the last section. Figures 6 and 7 illustrate such a situation. The figures plot the same data as Figure 3, except that in these cases the x-values of *both* samples have been randomly permuted, independently of one another, so that the two datasets in each of the two plots are drawn from the same underlying permutation distribution. In Figure 6 the colors are allocated according to the L_1 mixture decomposition, and in Figure 7 according to the L_2 mixture decomposition. There are obviously far fewer red and green points in the L_2 plot than in the L_1 plot. But there is another important, more subtle difference between the two.

The red and green points in both plots appear to be scattered approximately uniformly over the range of the data; far more uniformly, in fact, than the scattering of the original data. This is an interesting and useful property. If one were presented only with the red and green points, and they were distributed with the same variation in density as the original data, one might be misled into thinking that areas of high density in the original data were areas in which the two underlying densities differed the most. In fact, of course, these densities are identical everywhere.

The intuitive reason for the approximate uniformity, which we will formalize below, is the following. In regions of higher density in the original data $\hat{f}_0$ and $\hat{f}_1$ are estimated better, and therefore, since $f_0 = f_1$, are closer to one another. The densities $\hat{f}_+$ and $\hat{f}_-$ will then be low in these areas, with a corresponding lower proportion of red and green points compared to blue points sampled from these distributions. In regions of lower density in the original data on the other hand, $\hat{f}_0$ and $\hat{f}_1$ are estimated relatively badly, and so are farther

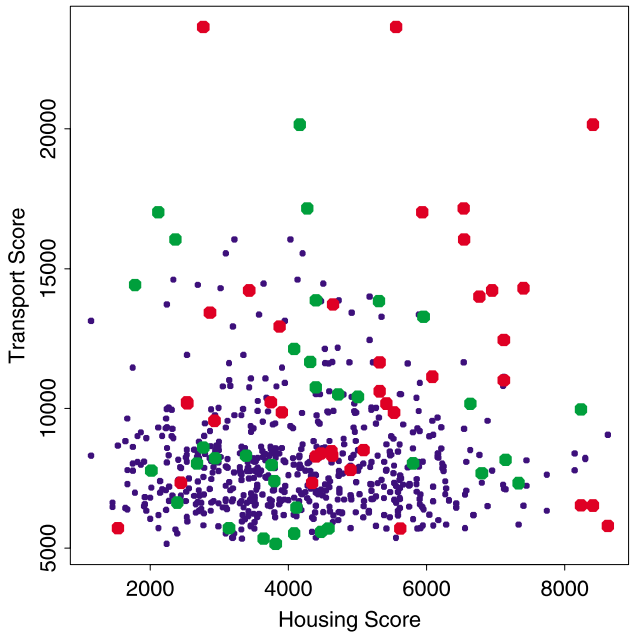

Figure 6. Combined data from two independent random permutations such as that of Figure 2, colored according to the L_1 mixture model.

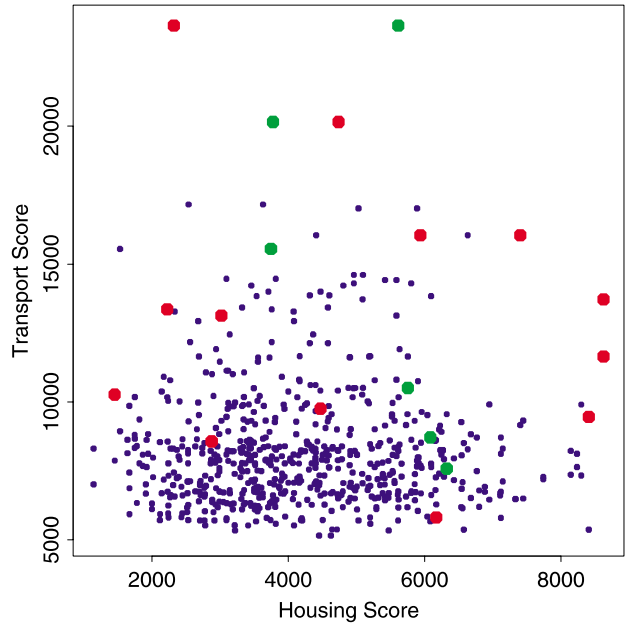

Figure 7. Combined data from two independent random permutations such as that of Figure 2, colored according to the L_2 mixture model.

apart from one another. So $\hat{f}_+$ and $\hat{f}_-$ are higher in these areas, resulting in higher proportions of red and green points sampled compared to the blue ones. Because fewer points overall are sampled in regions of lower density these differences in proportions balance out to a certain extent, resulting in the approximate uniformity.

The proportions of the reds and greens sampled in different areas of the densities depend on $P_{+/-}$, the sampling proportion defined in (3.5). The following theorem gives a precise asymptotic statement about the behavior of $P_{+/-}$. We will interpret the theorem below, and then in the next section investigate how these results compare to the small sample behavior of $P_{+/-}$.

Theorem 1. *Let f_0 and f_1 be two densities on $\mathbb{R}^d$, $d \geq 1$. Let $f_0(x) = f_1(x) > 0$ for $x \in \mathbb{R}^d$. Let K satisfy the usual conditions of a density estimation kernel (see e.g., Silverman 1986, chap. 4), so that using definitions (3.1)–(3.3),*

$$nh^d(\hat{f}_i(x) - \lambda) \to N(0, \lambda R), \quad i = 0, 1, \tag{4.1}$$

in distribution as $nh^d \to \infty$ and $h \to 0$. Here $\lambda = f_0(x)$ and $R = \int_{\mathbb{R}^d} K^2(x)dx$. Let $P_{+/-}$ be defined as in (3.5) with $\hat{f}_+$ and $\hat{f}_-$ defined from $\hat{f}_0$ and $\hat{f}_1$ using either the L_1 definitions (2.3), (2.4) or the L_2 definitions (2.11), (2.12).

Then as $nh^d \to \infty$ and $h \to 0$,

1. In the L_1 case:

$$(nh^d\lambda)^{\frac{1}{2}} P_{+/-}(x) \to \left(\frac{R}{2}\right)^{\frac{1}{2}} |Z|. \tag{4.2}$$

2. In the L_2 case:

$$(nh^d\lambda) P_{+/-}(x) \to \left(\frac{R}{4}\right) \chi_1^2. \tag{4.3}$$

Here Z is a standard Gaussian, and χ_1^2 is a chi-squared random variable with one degree of freedom.

A proof is given in the appendix. Because $E|Z| = \sqrt{2/\pi}$ and $E\chi_1^2 = 1$, the following corollary follows immediately:

Corollary 1. *Under the conditions of Theorem 1, as $nh^d \to \infty$ and $h \to 0$,*

1. In the L_1 case:

$$n\lambda E[P_{+/-}(x)] \asymp \left(\frac{Rn\lambda}{\pi h^d}\right)^{\frac{1}{2}}. \tag{4.4}$$

2. In the L_2 case:

$$n\lambda E[P_{+/-}(x)] \asymp \frac{R}{4h^d}. \tag{4.5}$$

We use the notation $a \asymp b$ here to mean $a/b \to 1$.

In a small neighborhood dx around a point x in $\mathbb{R}^d$ with $\lambda = f_0(x) = f_1(x)$, we expect $n\lambda dx$ points to be sampled in total, and

$$n\lambda E[P_{+/-}(x)]dx \tag{4.6}$$

points sampled from $\hat{f}_+$ or $\hat{f}_-$, and so colored green or red. The corollary gives asymptotic estimates of this quantity in both the L_1 and L_2 cases, showing how it depends on λ. For n large enough and h small enough, the first part of the corollary shows that (4.6) is proportional to $\sqrt{n\lambda}dx$: the number of points in this neighborhood colored red or green is approximately proportional to the square root of the total number of points in the neighborhood. One would expect then, as is observed in Figure 6, that the distribution of reds and greens over the range of the data is more uniform than the distribution of all the data together.

The second part of the corollary shows, however, that in the L_2 case the number of red and green points in such a neighborhood is asymptotically *constant* with respect to λ. Using a Gaussian kernel for example, we have

$$R = \left(\frac{1}{2\sqrt{\pi}}\right)^d, \tag{4.7}$$

and with $h = 1$ the right-hand side of (4.5) is 0.070, 0.020, and 0.0056 for $d = 1$, 2, and 3. These numbers are the approximate expected number of red and green points *per unit kernel volume* h^d for large n. The uniformity approximation appears then to be even better in the L_2 case than in the L_1 case.

To examine how large n has to be for this asymptotic approximation to be reasonable, we performed the following simulation, approximating the distribution $f_0 = f_1$ around a neighborhood of x by a uniform density of height λ. We used a standard Gaussian kernel, and set $h = 1$ throughout so that the results are scaled by kernel bandwidth:

1. Generate $X_1, \ldots, X_M$ and $Y_1, \ldots, Y_M$, iid from a uniform distribution on $[-r, r]^d$. Here $M = 2rn\lambda$, and r is chosen large enough so that $K(r) \approx 0$. The quantity $n\lambda$ is chosen so that M is an integer.
2. Calculate $\hat{f}_0(0)$ and $\hat{f}_1(0)$ from (3.2) and (3.3).
3. Calculate $P_{+/-}(0)$ from (3.5).
4. Repeat 1–3 a large number (1,000) of times, estimating $E[P_{+/-}(0)]$ from the sample mean.

The results of the simulation, performed for $d = 1$, 2 and 3, for both the L_1 and L_2 case, and over a range of $n\lambda$, are shown in Figure 8. The x-axis of the figure plots $n\lambda$ on a log scale, and the y-axis plots the simulation estimate of $n\lambda E[P_{+/-}(x)]$. For small values of $n\lambda$ all the graphs are approximately zero: no red or green points appear outside the range of the data. The graphs converge for high $n\lambda$ in the L_2 case, and rise at a rate $\sqrt{n\lambda}$ in the L_1 case, as predicted.

What is interesting, though, is how rapidly the L_2 graphs approach their asymptote. When $d = 2$, for example, convergence has occurred already at $n\lambda = 0.25$, a density of approximately only one data point per four kernel areas.

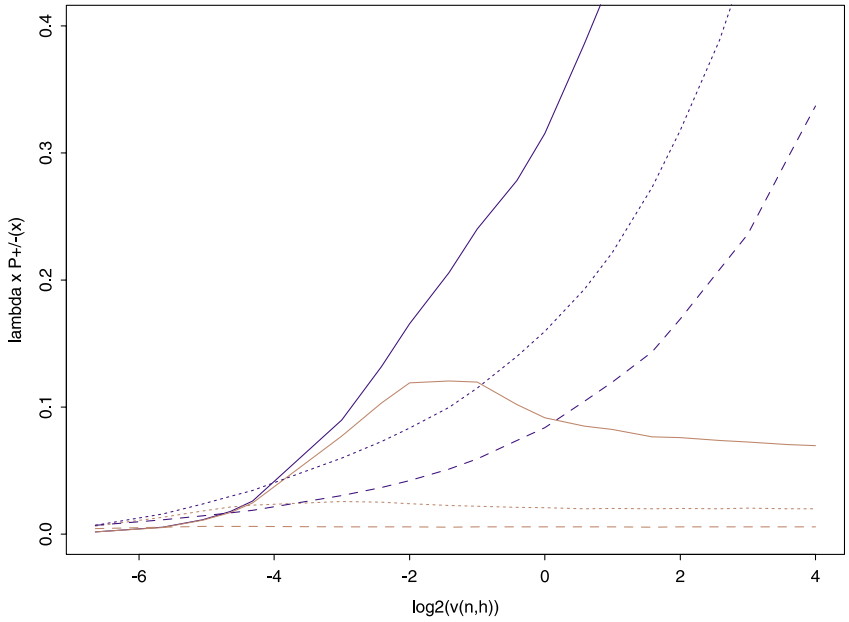

Figure 8. Results of the simulation to determine the numbers of red and green points compared to the total number of points, for a range of heights of densities λ and sample sizes n. Results for the L_1 decomposition are in blue, for the L_2 decomposition in brown. Dimensions $d = 1, 2, 3$ are represented by solid, dotted, and dashed lines, respectively.

The importance of the consensus density $f_=$ becomes apparent here. In both the L_1 and L_2 cases, and particularly in the L_2 case, the red and green points, sampled from f_+ and f_-, are distributed more uniformly than the original data. We have argued that this is a desirable property, but is only possible in a mixture decomposition of the original densities if a third density $f_=$ is included which has as much variation in height (slightly more, in fact) as the original densities.

We now turn to a different example of the use of these mixture decompositions, illustrating some further uses of the methods.

5. VISUALIZATION OF CLASSIFICATION

The example analysis we have used until now compares a real dataset to artificial data generated under a null hypothesis (in our case that of independence) on the original data. The mixture decomposition we have described can also be useful in visualizing differences between two classes of real data, for which there is no designated null dataset.

Figure 9 plots data from a study on diabetes in Pima Native American women, from Blake and Merz (1998). The plot is a two-dimensional projection of three-dimensional data obtained using the XGobi software (Swayne, Cook, and Buja). The three axes measure a body mass index, a plasma glucose concentration, and a diabetes pedigree function for each

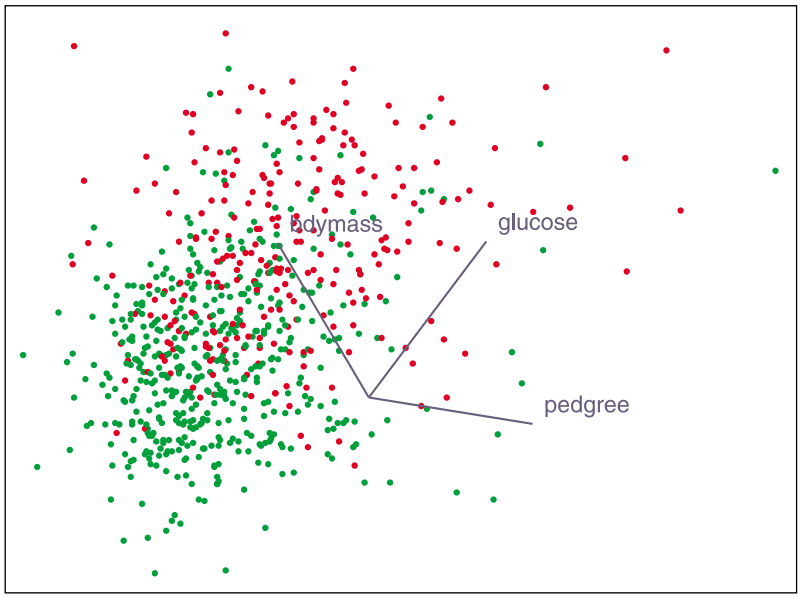

Figure 9. The Pima Native American diabetes data. Red points signify diabetics, green points nondiabetics. See text for details of axes.

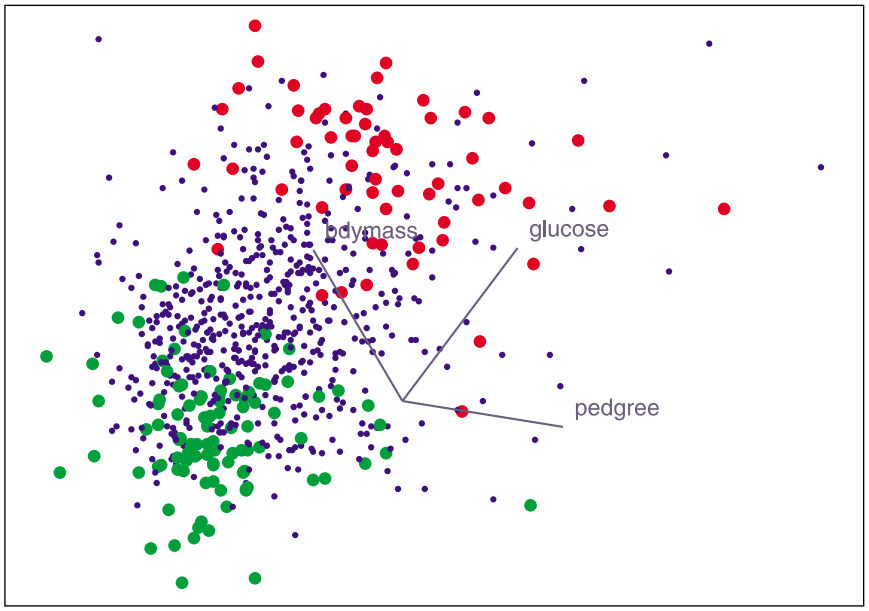

Figure 10. The data in Figure 9, with points colored using an L_2 mixture decomposition.

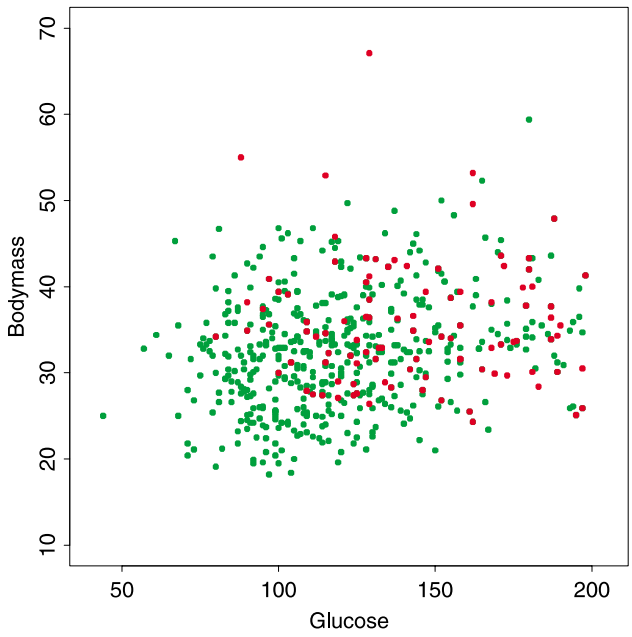

Figure 11. The data in Figure 9, with the pedigree variable removed.

of 488 Pima women. This is the subset of the original data for which all three measurements were available. The data are colored red in the case of a diabetic and green otherwise.

From the plot we could infer an approximate discrimination plane between the two groups, perpendicular to the glucose axis. There is a lot of overlap between the two groups, however, and the actual shape of the classification boundary is not clear. It is also not apparent, from this projection alone, whether the point clouds of the two classes merge completely in the center, or whether one is above or below the other and they just appear to merge because they are being projected on top of one another. In the latter case the plane described above would certainly not be an appropriate discrimination rule.

Figure 10 plots the same data, with the same projection, after computing an L_2 mixture decomposition. We have let f_1 be the density of the nondiabetics and f_0 be that of the diabetics, and colored the points using the same convention as before: green denotes samples from $\hat{f}_+$, red from $\hat{f}_-$, blue from $\hat{f}_=$. It is clear from this figure that the two point clouds must merge in the center. If one were being projected on top of the other, and there was in fact separation between the clouds when viewed from another angle, then the center of the plot would contain red and green points, indicating separate regions of high class purity. Instead there are only blue points. It is of course preferable, whichever coloring scheme is used, to rotate the data in 3-space and view many different projections. This is possible in XGobi, but not in an article.

The pedigree function is one of the less important predictor variables. Figure 11 plots the same data with this variable removed. Figure 12 is a decomposition of these points, this time using an L_1 scheme. As was noticed in the example in Section 1, the L_1 scheme colors proportionately more points red or green than the L_2 scheme. In Figure 12 enough points

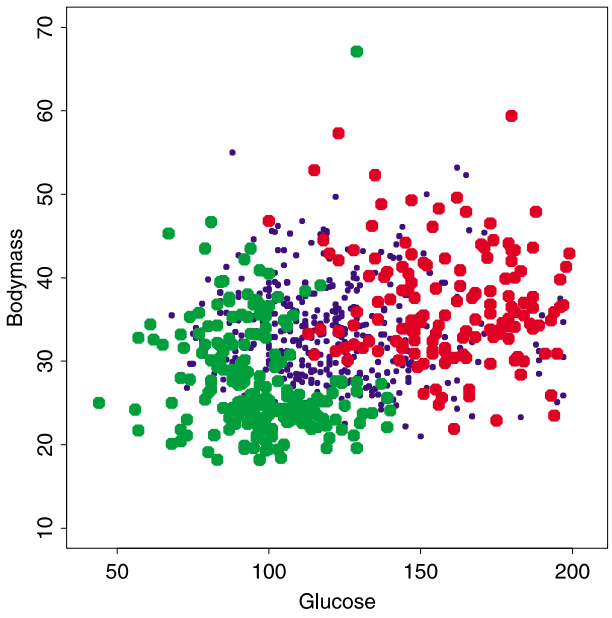

Figure 12. The data in Figure 11, with points colored using an L_1 mixture decomposition.

are colored red or green for an estimated classification boundary to become apparent. The boundary does not appear linear. There appears to be a central "peninsula" of high diabetic concentration projecting towards the nondiabetics. Whether such a pattern is real or appeared just by chance in the sampling can be checked by viewing plots from repeated samplings. This rather interesting boundary does seem to be reproduced in repeated sampling, and does therefore seem to be real.

A. APPENDIX: PROOF OF THEOREM 1

From (4.1),

$$nh^d\lambda(\hat{f}_i(x)/\lambda - 1) \to N(0, R), \quad i = 0, 1. \tag{A.1}$$

Let $v_{n,h} = nh^d\lambda$. Write (4.2) as $[v_{n,h}p_1(\hat{f}_0(x)/\lambda, \hat{f}_1(x)/\lambda)]^{\frac{1}{2}}$, where

$$p_1(a, b) = \frac{(a-b)^2}{(a+b)^2}, \quad a, b \in \mathbb{R}, \tag{A.2}$$

and write (4.3) as $[v_{n,h}p_2(\hat{f}_0(x)/\lambda, \hat{f}_1(x)/\lambda)]$, where

$$p_2(a, b) = \frac{(\sqrt{a}-\sqrt{b})^2}{(a+b)^2}, \quad a, b \in \mathbb{R}. \tag{A.3}$$

From a Taylor expansion of either p_i around $E[(\hat{f}_0(x)/\lambda, \hat{f}_1(x)/\lambda)] = (1,1)$, since $\nabla p_j\ (1,1) = (0,0)$, $j = 1, 2$,

$$v_{n,h} p_j(\hat{f}_0(x), \hat{f}_1(x)) \to \frac{1}{2} R Z_2^T \nabla^2 p_j(1,1) Z_2, \tag{A.4}$$

where $Z_2 \sim N_2(0, I)$. For $j = 1$ this is

$$\left(\frac{R}{2}\right) \chi_1^2, \tag{A.5}$$

and the result follows on taking square roots. For $j = 2$ this is

$$\left(\frac{R}{4}\right) \chi_1^2, \tag{A.6}$$

which is the required result.

ACKNOWLEDGMENTS

In their work on "Data Spelunking," Rick Becker and Alan Wilks first proposed methods for analyzing data by describing where the data are *not*, rather than where they are. We recast data spelunking as a comparison of data against a uniform distribution and extended the idea to a general scheme for comparing two arbitrary distributions, as described in this article. We would like to thank Rick Becker and Alan Wilks, as well as our manager at AT&T Labs, Daryl Pregibon, who initiated our exploration of data spelunking. We finally owe a debt of gratitude to two anonymous referees whose thoughtful suggestions helped clarify several issues in this article.

[Received December 1999. Revised October 2003.]

REFERENCES

Swayne, D., Cook, D., and Buja, A. "XGobi: Interactive Dynamic Data Visualization in the X Window System," www.research.att.com/areas/stat/xgobi/.

Blake, C., and Merz, C. (1998), UCI Repository of Machine Learning Databases www.ics.uci.edu/~mlearn/MLRepository.html.

Friedman, J., and Fisher, N. (1999), "Bump Hunting in High-Dimensional Data," *Statistics and Computing*, 9, 123–143.

Scott, D. W. (1992), *Multivariate Density Estimation. Theory, Practice, and Visualization*, New York: Wiley.

Silverman, B. W. (1986), *Density Estimation*, London: Chapman and Hall.