Transformation-based density estimation

Keith Knight
University of Toronto

ABSTRACT: A new non-parametric method of density estimation for univariate data is presented. The underlying idea is to estimate a transformation of the data so that the transformed data behave like a sample from some fixed distribution, for example, a standard Gaussian (or normal) distribution. The method described in this paper estimates the transformation by minimizing the integrated squared second derivative of the transformation subject to an upper bound on a weighted sum of squared deviations between observed and expected order statistics. This method has the attractive feature that an educated choice of the upper bound on the weighted sum of squares can be made *a priori*.

KEY WORDS: density estimation, goodness-of-fit, splines

1 Introduction

Let $X_1, \cdots, X_n$ be independent random variables from some (unknown) density f. The problem of nonparametrically estimating f has been extensively studied and a variety of methods exist, among them kernel estimation, nearest-neighbor estimation, penalized likelihood estimation and orthogonal series estimation; Silverman (1986) gives a complete and excellent account of these methods. The purpose of this paper is to outline a somewhat different method of density estimation based on estimating a transformation of the data so that the transformed data looks like a random sample from some known distribution. We concentrate on the case where the known distribution is standard Gaussian.

Suppose the data are transformed by some monotone function g so that the density of $g(X_i)$ is f_0; it then follows that the density of X_i is given by $f(x) = |g'(x)|f_0(g(x))$. (If f_0 is a uniform density on $(0,1)$ then g is the distribution function of the data and g' is the density.) This suggests that f can be estimated by estimating a smooth transformation g so that $g(X_1), \cdots, g(X_n)$ behave like a sample from f_0. A simple approach to estimating g can be obtained by looking at a quantile-quantile plot of the data for the distribution corresponding to f_0; if F_0^{-1} is the inverse of the distribution function corresponding to f_0 and $F_0^{-1}(i/(n+1))$ (say) is plotted against the i-th order statistic $X_{i:n}$ then an estimator of g is obtained by fitting a smooth (monotone) curve through these points. When viewed

this way, the estimation of g becomes similar to a scatterplot smoothing problem. To be somewhat more precise, take f_0 to be the standard Gaussian density and state the estimation problem as follows:

$$\text{minimize} \quad \int_{-\infty}^{\infty} (g''(x))^2 \, dx \quad \text{subject to} \quad \sum_{i=1}^{n} \omega^2(p_{i:n})(g(X_{i:n}) - \Phi^{-1}(p_{i:n}))^2 \leq S^2 \qquad (1)$$

where $p_{i:n} = (i - 0.375)/(n + 0.25)$ or $(i - 0.5)/n$ or $i/(n+1)$, for example; Φ^{-1} is the inverse distribution function of a standard Gaussian distribution and $\omega^2(t) = \phi^2(\Phi^{-1}(t))/(t(1-t))$ where ϕ is the standard Gaussian density function. If the sample is from a standard Gaussian distribution, then $\text{Var}(X_{i:n}) \approx \omega^{-2}(p_{i:n})/n$. Note also that for $i < j$,

$$\text{Cov}(X_{i:n}, X_{j:n}) \approx \frac{p_{i:n}(1 - p_{j:n})}{n\phi(\Phi^{-1}(p_{i:n}))\phi(\Phi^{-1}(p_{j:n}))};$$

it is not clear how much is lost by not considering the dependence between order statistics. The parameter S^2 obviously controls the smoothness of the estimated transformation g; a reasonable choice of S^2 will guarantee that for data from a Gaussian distribution, the estimated transformation $\widehat{g}$ will be nearly linear in the sense that $\int (\widehat{g}'')^2$ is small or zero.

To solve (1), we introduce a Lagrange multiplier λ, a slack variable z^2 and minimize

$$\int_{-\infty}^{\infty} (g''(x))^2 \, dx + \lambda \left[\sum_{i=1}^{n} \omega^2(p_{i:n})(g(X_{i:n}) - \Phi^{-1}(p_{i:n}))^2 + z^2 - S^2 \right]. \qquad (2)$$

This approach is, in many ways, an elaboration of an approach suggested by Parzen (1979). Good and Gaskins (1980) also use goodness-of-fit techniques (χ^2 and Kolmogorov-Smirnov tests) for selecting smoothness parameters for penalized likelihood density estimators.

A variation of (1) that is somewhat more familiar in the case of scatterplot smoothing is

$$\text{minimize} \quad \sum_{i=1}^{n} w_i(y_i - g(t_i))^2 + \lambda \int_{-\infty}^{\infty} (g''(t))^2 \, dt \qquad (3)$$

where λ is a fixed constant whose value is usually data dependent, the y_i's are responses and the t_i's covariates. This particular modification is necessary for scatterplot smoothing because a reasonable "target" value of the weighted residual sum of squares in (3) is seldom known *a priori*. (As noted by Reinsch (1967), if $y_i = g_0(t_i) + \varepsilon_i$ and $\text{Var}(\varepsilon_i) = \sigma^2/w_i$ with σ^2 known, then it would seem that we could put a reasonable upper bound on the weighted residual sum of squares in (3); for example, in the scatterplot smoothing setting, a naive value of S^2 would be $n\sigma^2$. However, as pointed out by several authors (Wold, 1974; Wahba, 1975a), such naive choices of S^2 typically undersmooth the data.) A density estimation

method similar in spirit to (1) is the penalized likelihood method of Silverman (1982) in which $\widehat{f}$ is chosen to maximize the penalized log-likelihood

$$\sum_{i=1}^{n} \ln f(X_i) - \alpha R(f)$$

subject to

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

for some $\alpha \geq 0$ and roughness penalty $R(f)$; however, as in the smoothing spline case, the "optimal" value of α is typically data dependent. In contrast to these latter two situations, we will show that the structure of (1) allows us to make a reasonably educated choice of S^2, which is independent of the data.

The solution $\widehat{g}$ of (1) is a cubic spline with knots at $X_1, \cdots, X_n$; if $\widehat{g}$ is monotone then the density estimator $\widehat{f}$ is

$$\widehat{f}(x) = C(\widehat{g}, n) \frac{\widehat{g}'(x)}{\sqrt{2\pi}} \exp(-\widehat{g}^2(x)/2) \tag{4}$$

where $C(\widehat{g}, n)$ is some normalizing constant. However, there is no guarantee that $\widehat{g}$ is monotone; this potential difficulty will be addressed later.

Throughout the remainder of this paper, we will assume that f_0 is the standard Gaussian density; however, the basic approach should be applicable to other densities. For example, one might be tempted to take f_0 to be a uniform density on $(0, 1)$ but this poses the additional constraints that $\widehat{g}(x)$ lie between 0 and 1 since $\widehat{g}(x)$ is an estimator of the underlying distribution function; this is similar to the approach taken by Wahba (1971, 1975b) where an interpolating polynomial is fitted to the empirical distribution function and then differentiated to obtain a density estimator. Another density estimation method using splines is the histospline approach of Boneva *et al* (1971), which uses splines to construct a density estimator from a histogram. Transforming to a distribution whose support is the real line is attractive for similar reasons to those used for analyzing probabilities on the logit or probit scale for binary data. The choice of goodness-of-fit criterion used here is also somewhat arbitrary but is attractive from a computational standpoint and from the fact that the decay of $\omega(t)$ to 0 near $t = 0$ or 1 provides some resistance against isolated extreme data points (although, obviously, we do not want resistance against clumps of extreme data points). Another approach that uses splines is due to Kooperberg and Stone (1991) who use maximum likelihood estimation to estimate $\ln(f)$ by a function from a space of cubic splines with predetermined knots; the number of knots is then determined according to some scheme.

2 Choice of smoothing parameter

In the case of smoothing splines, only rarely can the choice of S^2 in (1) or, equivalently, of λ in (3) be specified *a priori*, the reason being that typically we do not know the variance of the response at a given point. The standard approach, here, is to use the data to choose the parameter; some variation of cross-validation is commonly used. Similarly, most non-parametric density estimators depend on smoothness parameters that will typically be data dependent. However, the structure of (1) allows for an educated choice of the parameter S^2 (independent of the data) by considering the distribution of the minimized weighted sum of squares in (1) in a very special case.

Suppose that $X_1, \cdots, X_n$ are known to be a sample from a Gaussian distribution with unknown mean and variance. In this case, the natural density estimator is given by

$$\widehat{f}_n(x) = \frac{1}{\widehat{\sigma}\sqrt{2\pi}} \exp\left(-\frac{(x-\widehat{\mu})^2}{2\widehat{\sigma}^2}\right) = \frac{1}{\widehat{\sigma}}\phi\left(\frac{x-\widehat{\mu}}{\widehat{\sigma}}\right)$$

where $\widehat{\mu}$ and $\widehat{\sigma}^2$ are estimators of the mean and variance. A possible, albeit inefficient, method of estimating these parameters is to minimize

$$S_n^2(a,b) = \sum_{i=1}^{n} \omega^2(p_{i:n})(aX_{i:n} + b - \Phi^{-1}(p_{i:n}))^2 \tag{5}$$

over all a and b; if $\widehat{a}$ and $\widehat{b}$ are the minimizing values then $\widehat{\sigma} = 1/\widehat{a}$ and $\widehat{\mu} = -\widehat{b}/\widehat{a}$. The estimators of the mean and variance are not as important as is the distribution of $S_n(a,b)$ at its minimum. It seems reasonable to believe that knowledge of the distribution of $S_n^2(\widehat{a},\widehat{b})$ will give us some insight as to the choice of S^2 in (1); that is, we should try to estimate the transformation g so that

$$S^2 \sum_{i=1}^{n} \omega^2(p_{i:n})(\widehat{g}(X_{i:n}) - \Phi^{-1}(p_{i:n}))^2$$

lies somewhere in the centre of the distribution of $S_n^2(\widehat{a},\widehat{b})$, for example $E[S_n^2(\widehat{a},\widehat{b})]$ or some intermediate quantile of the distribution of $S_n^2(\widehat{a},\widehat{b})$. The rationale is that we should expect approximately the same "goodness of fit" for our transformed data as we would obtain for Gaussian data (with unknown location and scale). With a view to getting a good finite sample approximation to the distribution of $S_n^2(\widehat{a},\widehat{b})$, we will now consider the asymptotic behaviour of

$$A_n^2(u,v) = \sum_{i=1}^{n} \omega^2(p_{i:n})(un^{-1/2}X_{i:n} + vn^{-1/2} + X_{i:n} - \Phi^{-1}(p_{i:n}))^2 \tag{6}$$

where, without loss of generality, we assume that the X_i's are Gaussian with mean 0 and variance 1. Note that $A_n^2(u,v) = S_n^2(un^{-1/2}+1, vn^{-1/2})$. The following proposition says that the minimum of S_n^2 (or that of A_n^2) converges in distribution.

PROPOSITION. Define $S_n^2(a,b)$ as in (5) and let $S_n^2(\widehat{a},\widehat{b})$ be its minimum. If the X_i's are Gaussian then

$$S_n^2(\widehat{a},\widehat{b}) \to_d \text{some } A^2$$

with the exact form of A^2 defined below in (7).

Proof. To prove this result we start by defining $A_n^2(u,v)$ as in (6) and expanding; hence

$$\begin{aligned} A_n^2(u,v) &= \frac{1}{n}\sum_{i=1}^n \omega^2(p_{i:n})(uX_{i:n}+v)^2 \\ &\quad +\frac{2}{n}\sum_{i=1}^n \omega^2(p_{i:n})(uX_{i:n}+v)\sqrt{n}(X_{i:n}-\Phi^{-1}(p_{i:n})) \\ &\quad +\frac{1}{n}\sum_{i=1}^n \omega^2(p_{i:n})n(X_{i:n}-\Phi^{-1}(p_{i:n}))^2 \\ &\approx \int_0^1 \omega^2(t)(u\widehat{F}_n^{-1}(t)+v)^2\,dt \\ &\quad +2\int_0^1 \omega^2(t)(u\widehat{F}_n^{-1}(t)+v)\sqrt{n}(\widehat{F}_n^{-1}(t)-\Phi^{-1}(t))\,dt \\ &\quad +\int_0^1 \omega^2(t)n(\widehat{F}_n^{-1}(t)-\Phi^{-1}(t))^2\,dt \end{aligned}$$

where $\widehat{F}_n^{-1}$ is the empirical quantile function, that is, the inverse of the empirical distribution function. Now using some well-known results (Shorack and Wellner, 1986) concerning the convergence of the Gaussian quantile process $\sqrt{n}(\widehat{F}_n^{-1}-\Phi^{-1})$ and the fact that $\omega(t)(\widehat{F}_n^{-1}(t)-\Phi^{-1}(t))$ converges uniformly to 0 in probability, we get

$$A_n^2(u,v) \to_d \int_0^1 \omega^2(t)(u\Phi^{-1}(t)+v)^2\,dt + 2\int_0^1 \omega(t)(u\Phi^{-1}(t)+v)W(t)\,dt + \int_0^1 W^2(t)\,dt$$

where $W(t) = [t(1-t)]^{-1/2}B(t)$ with B a Brownian bridge process (that is, a Gaussian process with $E(B(s)) = 0$ for all s and $E(B(s)B(t)) = s(1-t)$ for $s \le t$). More importantly (noting that $\int_0^1 \omega^r(t)\Phi^{-1}(t)\,dt = 0$ for $r \ge 0$),

$$\begin{aligned} \min_{u,v} A_n^2(u,v) &\to_d A^2 \qquad (7) \\ &= \int_0^1 W^2(t)\,dt - \|\omega\Phi^{-1}\|^{-2}\left(\int_0^1 \omega(t)\Phi^{-1}(t)W(t)\,dt\right)^2 \\ &\quad -\|\omega\|^{-2}\left(\int_0^1 \omega(t)W(t)\,dt\right)^2 \end{aligned}$$

where $\|y\|^2 = \int_0^1 y^2(t)\,dt$. □

Note that the first term of A^2 is the weak limit of the so-called Anderson-Darling test statistic. To three decimal places, $\|\omega\Phi^{-1}\|^2 = 0.269$ and $\|\omega\|^2 = 0.479$. The mean of A^2 is 0.346 while the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th percentiles are 0.12, 0.16, 0.18, 0.23, 0.30, 0.41, 0.54, 0.64 and 0.93 respectively. (The mean of A^2 was evaluated by numerical integration while the quantiles are estimates based on 5000 simulations.) A gamma distribution fitted by matching first and second moments agrees fairly closely with the empirical simulation distribution; the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th percentiles of a gamma distribution with mean 0.346 and variance 0.0259 are 0.08, 0.13, 0.16, 0.23, 0.32, 0.44, 0.56, 0.65 and 0.82. The convergence of $\min A_n^2(u,v)$ to A^2 seems to be reasonably rapid although the quantiles of $\min A_n^2(u,v)$ seem to converge to those of A^2 from below.

The particular choice of S^2 depends on the amount of smoothness desired in the estimator $\widehat{f}$; choosing S^2 to be the 5th percentile of distribution of A^2 would result in a rougher and less Gaussian estimate than would result by choosing S^2 to be the mean or 95th percentile of the distribution of A^2. Obviously, decreasing S^2 will increase the chance of capturing fine details in the density of X but, at the same time, increases the chance of uncovering spurious details. In some sense, the choice of S^2 represents the data analyst's prior beliefs (or prejudices) regarding the smoothness of the underlying true density. Some practical experience has indicated that a "good" choice of S^2 is 0.16, the 5th percentile of A^2, although the density estimates themselves are not overly sensitive to the choice of S^2 provided that S^2 is not in the tails of the distribution of A^2. However, if the data come from a distribution whose transformation to a Gaussian distribution is highly "eccentric" then a much smaller value of S^2 may be needed to produce a good estimate of the underlying density. It should be remembered that, while the quantiles of two distributions might be close, their densities need not be close.

From the preceding discussion, it is clear that a similar approach could be taken with any density f_0 having infinite support simply by changing the weight function ω and noting that the distribution of A^2 will change depend on f_0; if F_0 is the corresponding distribution then one would replace Φ^{-1} by F_0^{-1} and $\omega^2(t)$ would become $f_0^2(F_0^{-1}(t))/(t(1-t))$. A further generalization would be to consider the limiting distribution of

$$\min_{\beta} \sum_{i=1}^{n} \omega^2(p_{i:n})(g_\beta(X_{i:n}) - F_0^{-1}(p_{i:n}))^2 \tag{8}$$

where $\{g_\beta\}$ is a set of monotone transformations (containing all linear transformations) indexed by a finite dimensional parameter β and the density of $g(X_i)$ is f_0 for some g in this set. The purpose of considering the distribution of (8) (when the density of the X_i's is f_0) rather than that of

$$\sum_{i=1}^{n} \omega^2(p_{i:n})(X_{i:n} - F_0^{-1}(p_{i:n}))^2$$

is to guard against oversmoothing the estimated transformation $\widehat{g}$ and the density estimate $\widehat{f}$.

3 Practical considerations

3.1 Normalization

The transformation $\widehat{g}$ cannot be unambiguously defined outside of the range of the data. As a result of this, the appropriate normalization for the density estimator $\widehat{f}$ in (4) is not obvious. Due to an absence of information, it may not make sense to define the estimator outside of $[X_{1:n}, X_{n:n}]$ as is done, for example, in kernel estimation; at the same time, it is somewhat presumptuous to set $\widehat{f}(x)$ to 0 outside the range of the data. A compromise is to leave $\widehat{f}(x)$ undefined outside of $[X_{1:n}, X_{n:n}]$ and normalize $\widehat{f}(x)$ so that the integral of $\widehat{f}(x)$ is slightly smaller than 1; more precisely, if F is the distribution function of the X_i's then $E[F(X_{n:n}) - F(X_{1:n})] = (n-1)/(n+1)$. This suggests that one might normalize the density estimator so that its integral from $X_{1:n}$ to $X_{n:n}$ is $(n-1)/(n+1)$; thus a natural choice for the normalizing constant $C(\widehat{g}, n)$ in (4) is

$$C(\widehat{g}, n) = \frac{n-1}{n+1}[\Phi(\widehat{g}(X_{n:n})) - \Phi(\widehat{g}(X_{1:n}))]^{-1}.$$

3.2 Tied observations

In theory, if the X_i's have a continuous distribution then the probability that $X_i = X_j$ for $i \neq j$ is 0; however, for a variety of reasons, data sets with tied observations occur frequently even when a continuous model is appropriate for the data. Also, in certain situations, an inherently discrete phenomenon is sometimes more conveniently approximated by a continuous distribution, for example, when the number of possible outcomes is large. These facts necessitate a change in the formulation of the basic optimization problem.

Suppose that

$$X_{i-1:n} < X_{i:n} = \cdots = X_{i+k:n} < X_{i+k+1:n}.$$

Then we replace $\omega^2(p_{i:n})$ in (1) by

$$\sum_{j=i}^{i+k} \omega^2(p_{j:n})$$

and $\Phi^{-1}(p_{i:n})$ by the weighted average

$$\left(\sum_{j=i}^{i+k} \omega^2(p_{j:n})\right)^{-1} \sum_{j=i}^{i+k} \omega^2(p_{j:n})\Phi^{-1}(p_{i:n}).$$

Finally, we change (1) so that we sum only over the unique X_i's. An alternative approach that might be used sometimes is to "jitter" the data by adding random noise to the original observations.

3.3 Non-monotone transformations

So far, we have assumed that the transformation $\widehat{g}$ is monotone. There is, however, no guarantee that the solution to (1) is monotone; in fact, if the data contain extreme outliers then the solution to (1) will not always be monotone. (The "effective" kernel of the spline smoother is not positive at all values (Silverman, 1984), hence the possibility of non-monotonicity.) There are several possible remedies to this, for example, adding the constraint that $\widehat{g}' \geq 0$ or increasing S^2 until a monotone solution to (1) exists; in fact, it is possible to reformulate (1) slightly so that the solution is a monotone spline although no general computing algorithm seems to exist. Two recent papers consider different approaches to monotone splines: Ramsay (1988) uses integrated B-splines while Kelly and Rice (1990) impose a monotonicity constraint on the B-spline coefficients. It should be noted that, in this situation, most other scatterplot smoothers are guaranteed to produce a monotone $\widehat{g}$ due to the monotonicity of the points $\{(X_{i:n}, \Phi^{-1}(p_{i:n})); i = 1, \cdots, n\}$. However, it is still possible to find the density of X even if g is non-monotone and the density of $g(X)$ is f_0. For each y in the range of g, we define a number $N(y)$ such that

$$N(y) = \#\{x : g(x) = y\}.$$

Then the density of X is simply

$$f(x) = \frac{|g'(x)|}{N(g(x))} f_0(g(x))$$

(Bickel and Doksum, 1977, p.45).

It is worth noting here that the non-monotonicity of $\widehat{g}$ is seldom a problem from a practical point of view; that is, almost all data sets encountered in practice will yield monotone $\widehat{g}$'s. Non-monotone solutions to (1) occur when a few points are isolated from the rest of the data set; in this case, no non-parametric procedure can hope to deal with this behaviour except by pre-transforming the data so that the spacings between adjacent points become more homogeneous, estimating the density of the transformed data and then transforming back to the original scale. In a certain sense, a non-monotone solution can serve as a warning that the data contain outliers and that some remedial action should be taken. However, pre-transformation is useful even when monotone $\widehat{g}$'s exist on the untransformed scale. The pre-transformation approach could be formalized by considering a parametric class of transformations (for example, Box-Cox transformations), estimating the (parametric) transformation from the data and then estimating the density of the transformed data; however, the precise choice of transformation is not as important as is pulling extreme data closer to the body of the data.

Pre-transforming the data seems also to improves the performance of this method when the data are restricted to a bounded interval or a half line and a non-negligible fraction of the data occurs at or near the boundary of the interval or half line; uniform- or exponential-like data are two such examples. The Gaussian transformation estimators, like many other nonparametric density estimators, tend to underestimate the density near the boundaries in such cases if the data are not pre-transformed. Fortunately, these situations are often very easy to recognize in practice and can be dealt with accordingly. Wand, Marron and Ruppert (1991) use transformation in connection with kernel estimation to try to minimize the mean integrated squared error. It should be noted, as Kooperberg and Stone (1991) point out, that minimizing integrated squared error does not necessarily lead to a qualitatively good estimate of the density.

3.4 Other smoothers

It is certainly possible to estimate the transformation using other smoothing methods; for example, one could used a kernel smoother (with bandwidth h) to estimate g so that

$$\widehat{g}_h(x) = \frac{\sum_{i=1}^n w_h(x - X_{i:n})\Phi^{-1}(p_{i:n})}{\sum_{i=1}^n w_h(x - X_{i:n})}$$

or

$$\widehat{g}_h(x) = \frac{\sum_{i=1}^n \omega^2(p_{i:n}) w_h(x - X_{i:n}) \Phi^{-1}(p_{i:n})}{\sum_{i=1}^n \omega^2(p_{i:n}) w_h(x - X_{i:n})}$$

and then choose h so that

$$\sum_{i=1}^n \omega^2(p_{i:n})(\Phi^{-1}(p_{i:n}) - \widehat{g}_h(X_{i:n}))^2 = S^2.$$

The problem with the estimator given above is the fact that, while the estimator of the transformation itself is quite good, the derivative of the estimator is not so well behaved. Thus the resulting density estimate can be very wiggly although this can be rectified to a certain extent by increasing S^2, the target value for the weighted sum of squares. This highlights one very positive feature of smoothing splines: the ability to give good estimates of the derivatives of a function as well as the function itself.

4 Examples

In this section, we will apply the method described above to five data sets that have been analyzed using a variety of parametric and non-parametric methods. These examples are given merely for illustration of the method and as such should not in any sense be construed as analyses of these data sets. It should be noted, however, that this method produces estimates that are qualitatively similar to those produced by more "intensive" estimation techniques. In addition, we will apply the method to artificial data generated from two distributions. For the first five examples, the raw data (possibly jittered) are plotted on the x-axis.

4.1 Old Faithful geyser data

Silverman (1986) considers the lengths of 107 eruptions of Old Faithful geyser in Yellowstone National Park. The density estimates shown by Silverman indicate that the data are multimodal. Figure 1 shows the Gaussian transformation density estimates taking S^2 to be 0.16, 0.346 and 0.64; all estimates have at least two modes with the smallest value of S^2 producing the most pronounced modes.

4.2 Long Beach SO_2 data

Leadbetter *et al* (1980) fit a type I extreme value distribution to 228 monthly maxima (of hourly averages) of sulfur dioxide concentrations at Long Beach, California from 1956 to

1974. Figure 2 gives both the Gaussian transformation density estimate (using $S^2 = 0.16$) and the fitted type I extreme value density

$$\widehat{f}(x) = 0.115 \exp(-0.115(x - 14.5)) \exp(-\exp(-0.115(x - 14.5))).$$

The parametric estimate is indicated by a dashed line, the Gaussian transformation estimate by a solid line.

4.3 Buffalo snowfall data

Data sets consisting of seasonal snowfall amounts in Buffalo, New York have been considered by Parzen (1979) and Tukey (1977). Data from 99 years (1884/85 to 1982/83) (courtesy of Ned Glick) will be considered here. Two estimates are shown in Figure 3; the first estimate uses the raw data with $S^2 = 0.16$ while the second estimate uses the logarithms of the data with $S^2 = 0.16$ and then transforms back to the original scale. The pre-transformed estimate is indicated with a dotted line; the two estimates are virtually identical. This is not surprising since, over the range of the data, $\ln(x)$ is very nearly linear.

4.4 Hidalgo stamp thickness data

Izenman and Sommer (1988) consider estimating the number of modes in the distribution of paper thicknesses in the 1872 Hidalgo stamp issue of Mexico. Using "bump-hunting" techniques in connection with kernel estimation, they conclude that 7 modes exist while fitting a mixture of Gaussian densities indicates either 3 or 5 modes. The Gaussian transformation estimate (based on 485 observations and using $S^2 = 0.16$) in Figure 4 shows 5 definite modes and one very small mode.

4.5 Chondrite meteorite data

Data for the distribution of silica in 22 chondrite meteorites are considered by Good and Gaskins (1980). Their best fitting density estimate was trimodal; they say that this result "is not surprising because there are several types of chondrite." The Gaussian transformation estimate (using $S^2 = 0.16$) as shown in Figure 5 is also trimodal.

4.6 Bimodal logistic data

Five samples of size 200 were drawn from a density which is a mixture of two logistic densities:

$$f(x) = \frac{3}{4}\frac{\exp(x+2.5)}{[1+\exp(x+2.5)]^2} + \frac{1}{4}\frac{\exp(x-2.5)}{[1+\exp(x-2.5)]^2}$$

Figure 6 shows the 5 estimated densities using $S^2 = 0.16$; the true density is indicated by the solid line.

4.7 Log-normal data

Five samples of size 200 were drawn from a log-normal distribution; the logarithms have a Gaussian distribution with mean 0 and variance 0.25. The density estimates were computed on the raw data and Figure 7 shows the 5 estimated densities using $S^2 = 0.16$. As before, the true log-normal density is indicated with a solid line.

5 Comments

5.1 Convergence results

At this point, very little can be said about the convergence properties of the density estimators described in this paper; however, it seems reasonable to believe that the rate of convergence will be similar to that of kernel or penalized likelihood estimators. The rate of convergence of $\widehat{f}$ to f depends most critically on the rate of convergence of the derivative of $\widehat{g}$ to its population analogue. In fact, even the most naive estimator of the transformation g (namely $\Phi^{-1}(\widehat{F}_n(\cdot))$) is a $\sqrt{n}$-consistent estimator of g (since $\widehat{F}_n$ is a consistent estimator of F). It is easy to show that consistency also hold for the spline estimator $\widehat{g}$.

Approximating the constraint in (1) as an integral suggests that

$$\begin{aligned}\frac{S^2}{n} &\geq \int_0^1 \omega^2(t)(\widehat{g}(F^{-1}(t)) - \Phi^{-1}(t))^2\,dt + o_p(1) \\ &= \int_{-\infty}^{\infty} \omega^2(F(x))f(x)(\widehat{g}(x) - \Phi^{-1}(F(x)))^2\,dx + o_p(1).\end{aligned}$$

Since $S^2/n \to 0$, this suggests that for any x,

$$\widehat{g}(x) \to_p \Phi^{-1}(F(x))$$

and that the convergence would be uniform over compact sets contained strictly within the support of the density. Moreover, letting $g_0(x) = \Phi^{-1}(F(x))$ be the true transformation to a Gaussian distribution, we get

$$\begin{aligned} S^2 \;\geq\; & \int_0^1 \omega^2(t) n(\widehat{g}(\widehat{F}_n^{-1}(t)) - g_0(\widehat{F}_n^{-1}(t)))^2 \, dt \\ & +2\int_0^1 \omega^2(t)\sqrt{n}(\widehat{g}(\widehat{F}_n^{-1}(t)) - g_0(\widehat{F}_n^{-1}(t)))\sqrt{n}(g_0(\widehat{F}_n^{-1}(t)) - \Phi^{-1}(t)) \, dt \\ & + \int_0^1 \omega^2(t) n(g_0(\widehat{F}_n^{-1}(t)) - \Phi^{-1}(t))^2 \, dt + o_p(1) \end{aligned}$$

From the fact that

$$\sup_{\epsilon \leq t \leq 1-\epsilon} \sqrt{n}(g_0(\widehat{F}_n^{-1}(t)) - \Phi^{-1}(t)) = O_p(1)$$

it follows that

$$\sup_{\epsilon \leq t \leq 1-\epsilon} \sqrt{n}(\widehat{g}(\widehat{F}_n^{-1}(t)) - g_0(\widehat{F}_n^{-1}(t))) = O_p(1),$$

that is, $\widehat{g}$ is a uniformly $\sqrt{n}$-consistent estimator of g_0 on compact sets contained strictly within the support of the distribution. Of course, as mentioned above, it is not the behaviour of $\widehat{g}$ which determines the behaviour of $\widehat{f}$ as much as the behaviour of $\widehat{g}'$; $\widehat{g}$ simply allows us to estimate the (cumulative) distribution function via $\Phi(\widehat{g}(\cdot))$. For example, a piecewise linear function through the points $\{(X_{i:n}, \Phi^{-1}(p_{i:n})); i = 1, \cdots, n\}$ will also be a $\sqrt{n}$-consistent estimator of g but will not give a consistent estimator of g'.

6 Concluding Remarks

The purpose of this paper has been to introduce a new method of density estimation based upon transforming data to a target distribution. This method, while not demonstrably superior to any existing method, has two attractive features. First, it is intuitively appealing; all we are doing is estimating a transformation from a quantile-quantile plot and then plugging in the estimated transformation and its derivative to obtain a density estimate. Second, it can be viewed as a nearly automatic method in the sense that we can make a reasonable choice of the parameter independently of the data; for example, there is no cross validation needed to choose the parameters. However, this automatic aspect is, of course, only a positive feature if used with an appropriate amount of caution. However, this method produces estimates which compare favourably with more computationally intense methods.

A possible drawback of this method is that it does not generalize "nicely" to multivariate data; for example, estimating a vector of transformations seems a non-trivial task. However,

one could use this method with the projection pursuit approach of Friedman *et al* (1984) which uses univariate density estimates along certain directions to construct a multivariate density estimate.

Acknowledgment: This research was supported, in part, by the Natural Sciences and Engineering Research Council of Canada.

References

Bickel, P.J. and Doksum, K.A. (1977) *Mathematical Statistics: Basic Ideas and Selected Topics.* San Francisco: Holden-Day.

Boneva, L.I., Kendall, D. and Stefanov, I. (1971) Spline transformations: three new diagnostic aids for the statistical data-analyst. (with discussion) *Journal of the Royal Statistical Society (series B).* **33**, 1-70.

Friedman, J.H., Stuetzle, W. and Schroeder, A. (1984) Projection pursuit density estimation. *Journal of the American Statistical Association.* **79**, 599-608.

Good, I.J. and Gaskins, R.A. (1980) Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. (with discussion) *Journal of the American Statistical Association.* **75**, 45-73.

Hall, P., DiCiccio, T.J. and Romano, J.P. (1989) On smoothing and the bootstrap. *Annals of Statistics.* **17**, 692-704.

Izenman, A.J. and Sommer, C.J. (1988) Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association.* **83**, 941-953.

Kelly, C. and Rice, J. (1990) Monotone smoothing with application to dose response curves and the assessment of synergism. *Biometrics.* **46**, 1071-1085.

Kooperberg, C. and Stone, C.J. (1991) A study of logspline density estimation. *Computational Statistics and Data Analysis.* **12**, 327-347.

Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1980) *Extremes and Related Properties of Random Sequences and Processes.* New York: Springer.

Parzen, E. (1979) Nonparametric statistical data modeling. (with discussion) *Journal of the American Statistical Association.* **74**, 105-131.

Ramsay, J.O. (1988) Monotone regression splines in action. (with discussion) *Statistical Science.* **3**, 425-461.

Reinsch, C.H. (1967) Smoothing by spline functions. *Numerische Mathematik.* **10**, 177-183.

Shorack, G.R. and Wellner, J.A. (1986) *Empirical Processes with Applications to Statistics.* New York: Wiley.

Silverman, B.W. (1982) On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics.* **10**, 795-810.

Silverman, B.W. (1984) Spline smoothing: the equivalent variable kernel method. *Annals of Statistics.* **12**, 896-916.

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

Tukey, J.W. (1977) *Exploratory Data Analysis.* Reading, Mass: Addison-Wesley.

Wahba, G. (1971) A polynomial algorithm for density estimation. *Annals of Mathematical Statistics.* **42**, 1870-1886.

Wahba, G. (1975a) Smoothing noisy data by spline functions. *Numerische Mathematik.* **24**, 383-393.

Wahba, G. (1975b) Interpolating spline methods for density estimation I: Equi-spaced knots. *Annals of Statistics.* **3**, 30-48.

Wand, M.P., Marron, J.S. and Ruppert, D. (1991) Transformations in density estimation (with discussion). *Journal of the American Statistical Association.* **86**, 343-361.

Wold, S. (1974) Spline functions in data analysis. *Technometrics.* **16**, 1-11.

Figure 1: Old Faithful data with S=0.16, 0.346 and 0.64

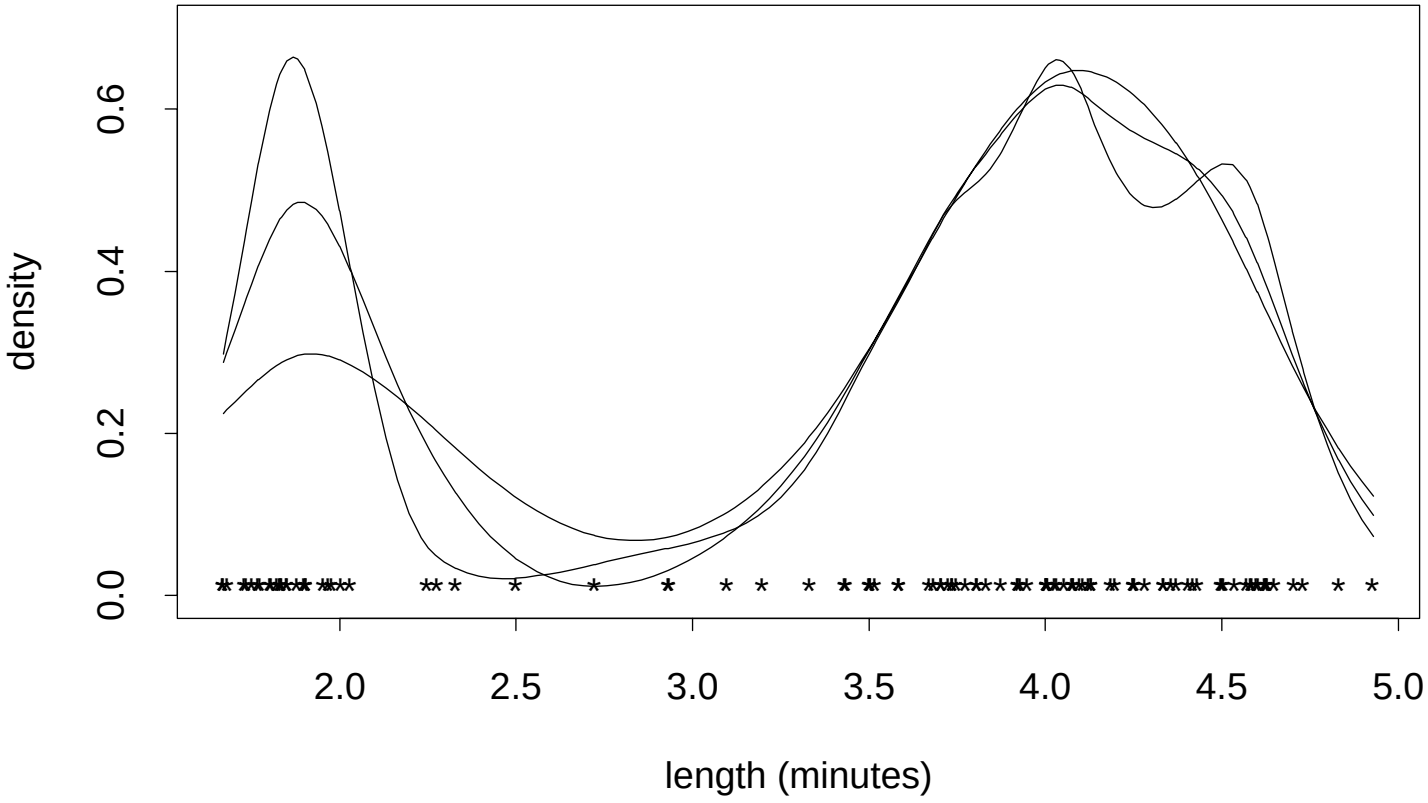

Figure 2: Sulfur dioxide data

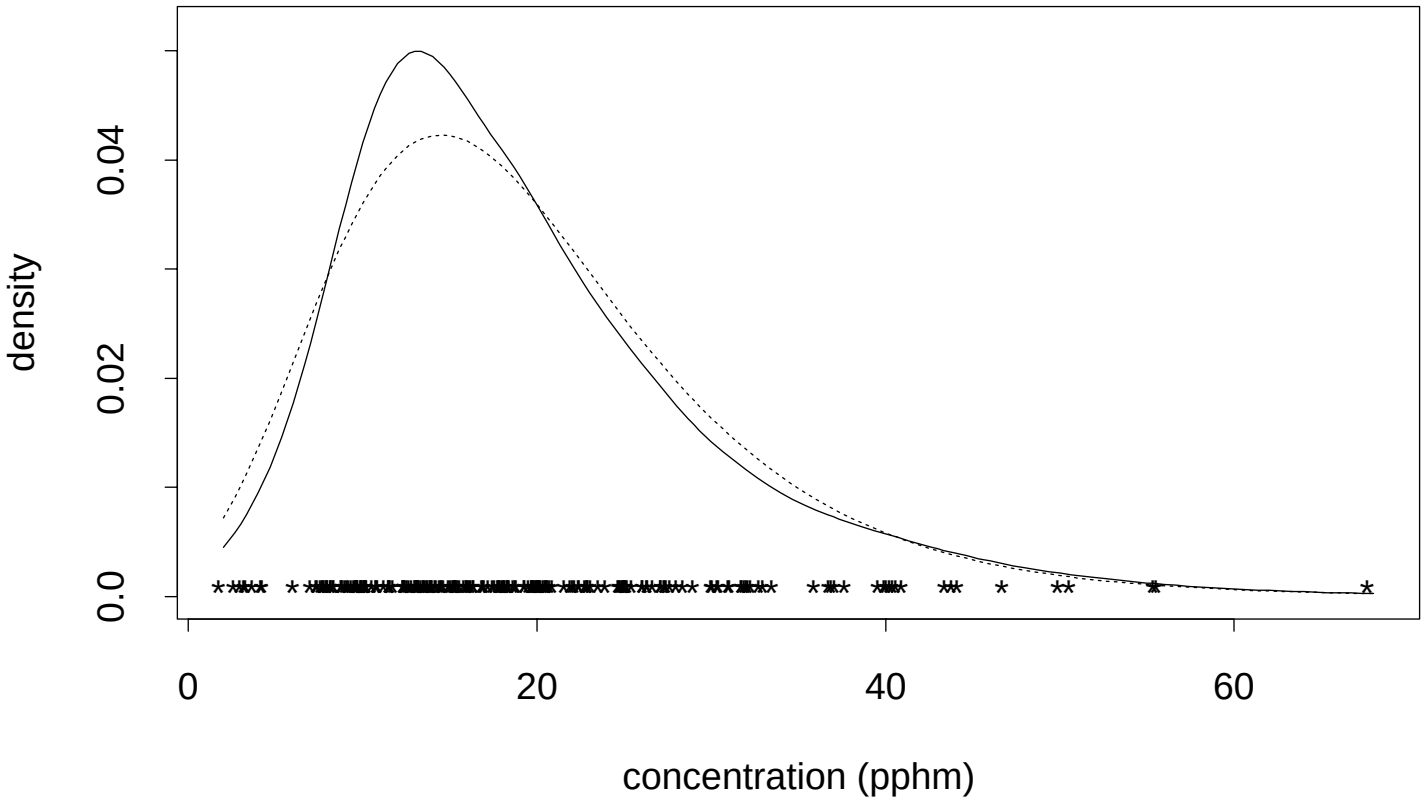

Figure 3: Buffalo snowfall data

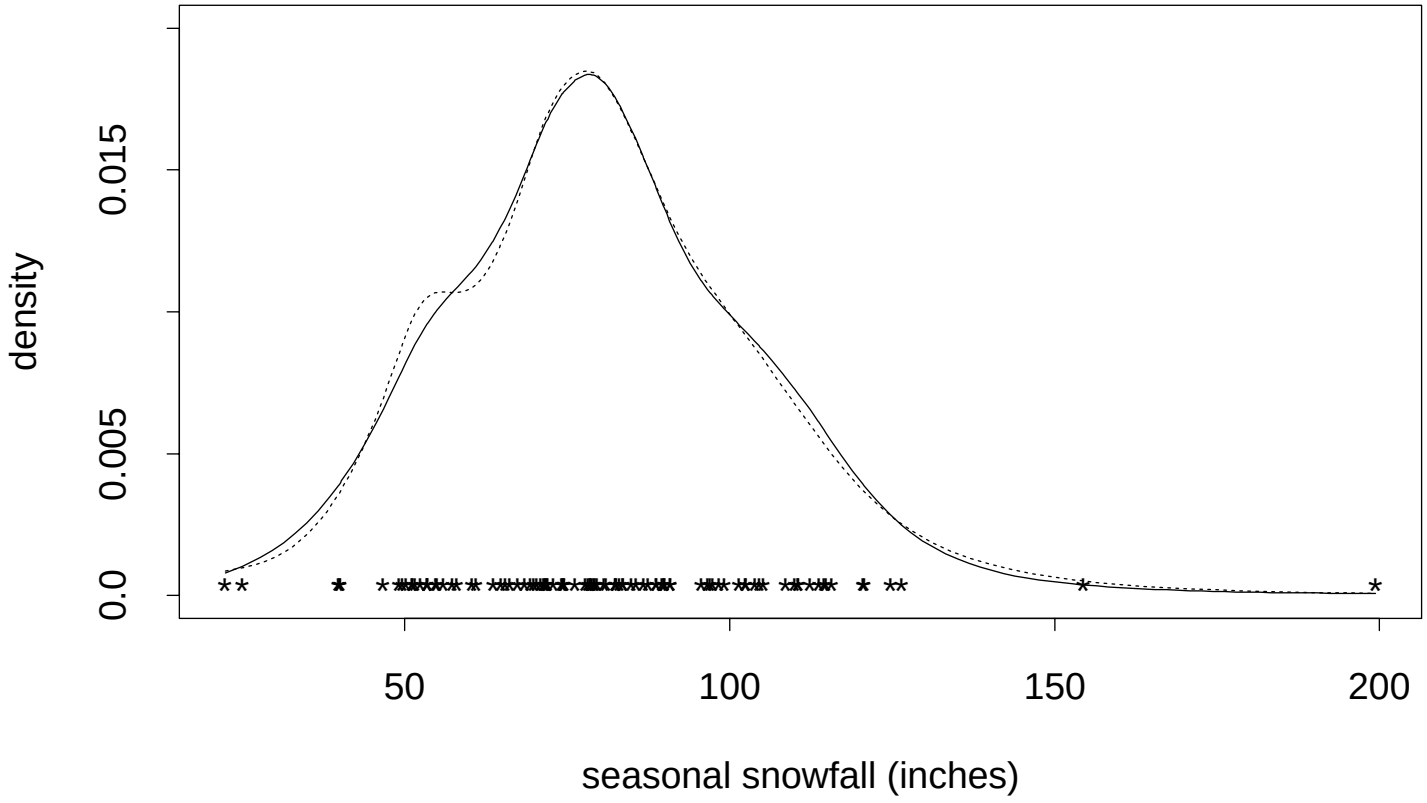

Figure 4: Chondrite meteorite data

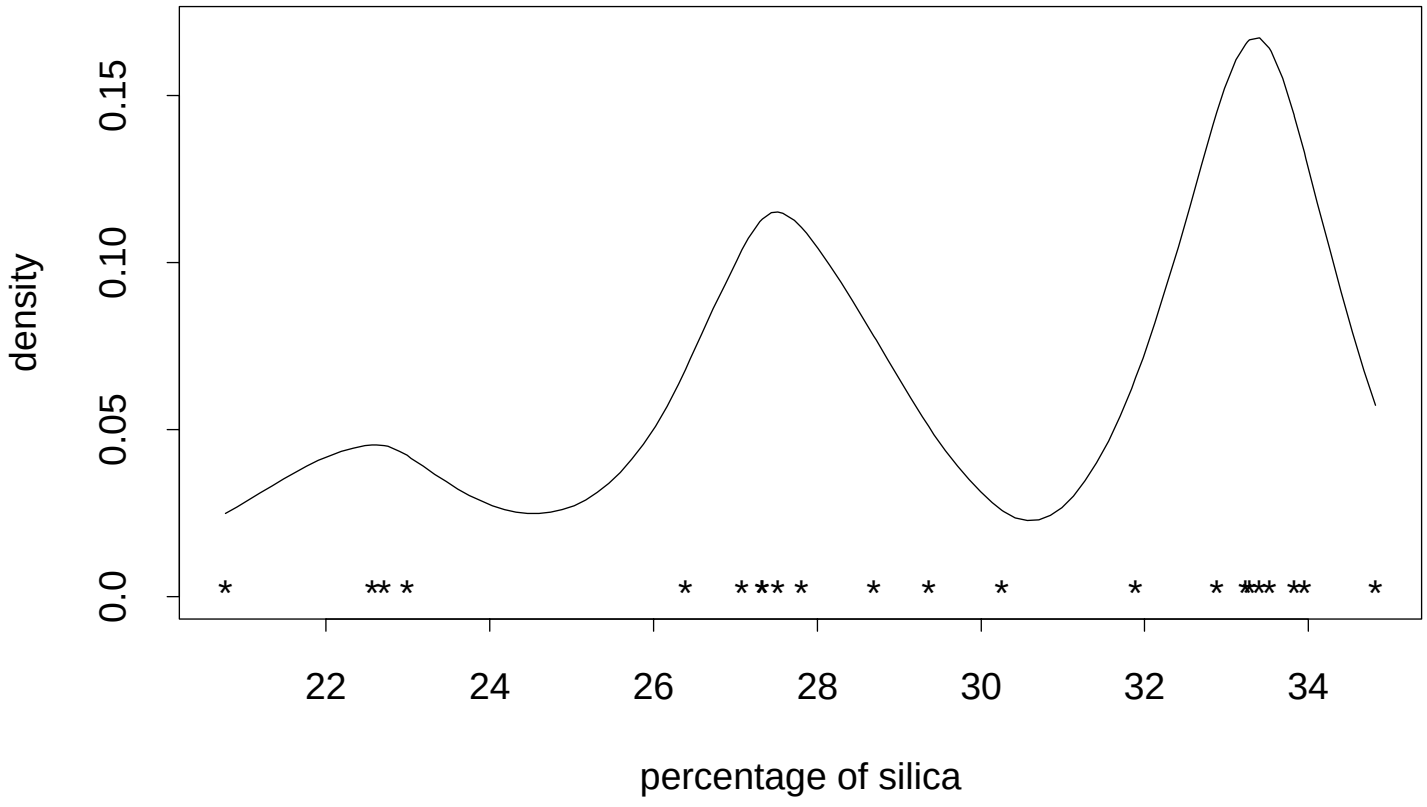

Figure 5: Hidalgo stamp thickness data

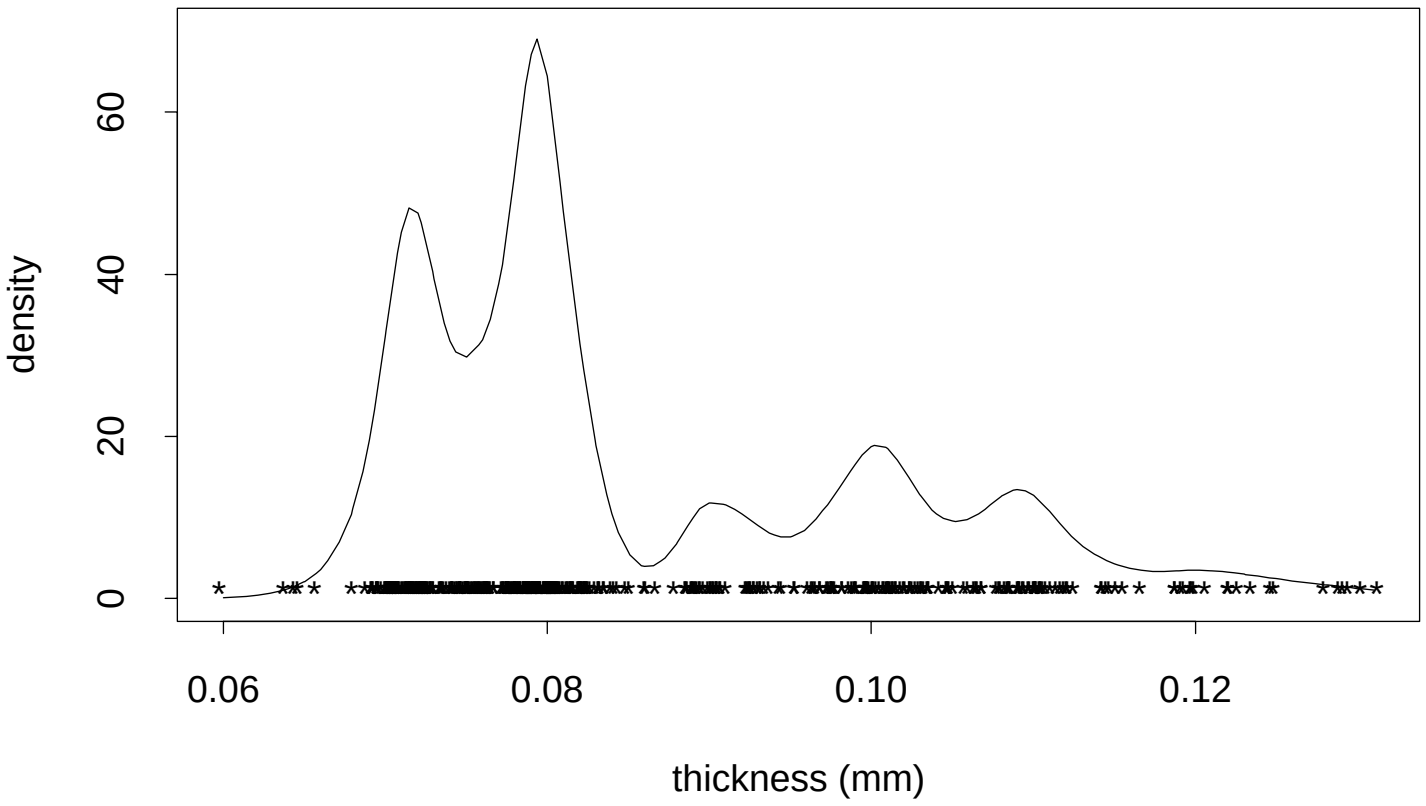

Figure 6: Bimodal logistic data

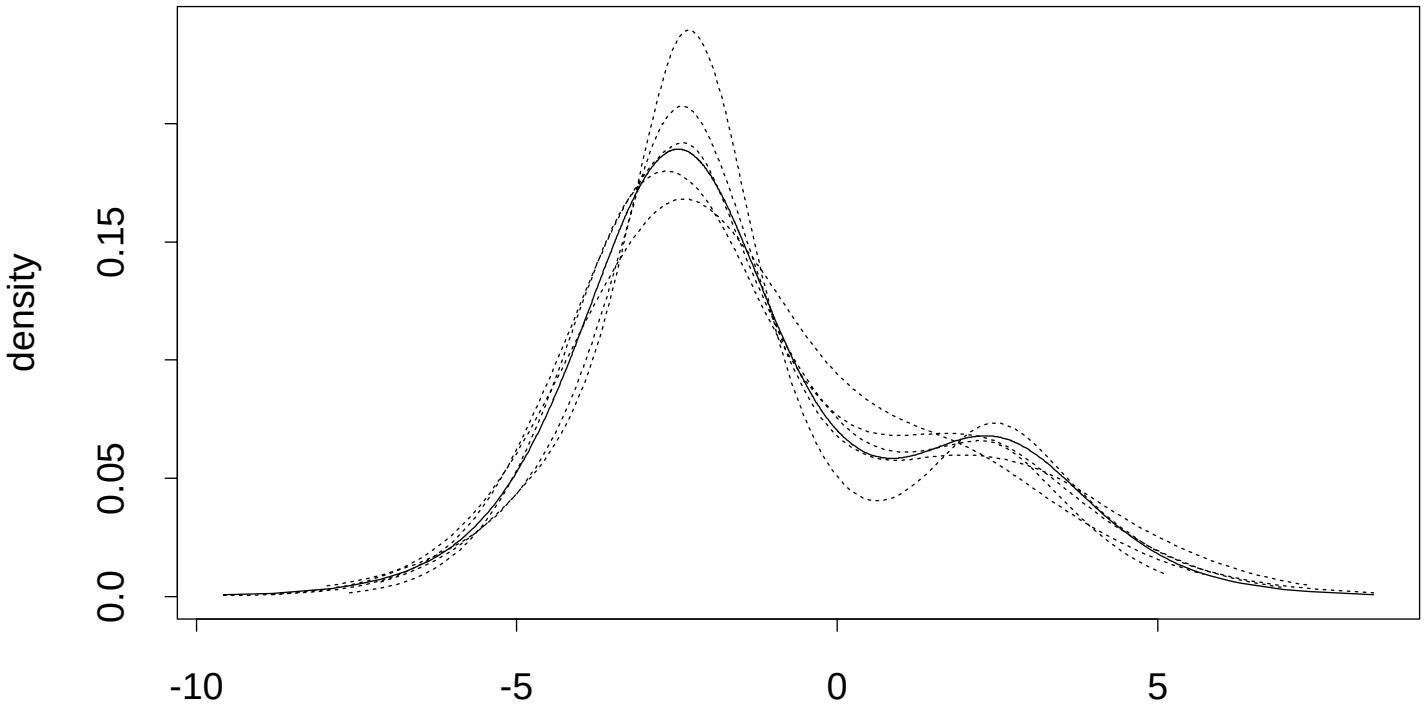

Figure 7: Log-normal data