Ann. Inst. Statist. Math.
Vol. 49, No. 4, 601 614 (1997)

FITTING A NORMAL DISTRIBUTION WHEN THE MODEL IS WRONG

J. B. COPAS AND C. B. STRIDE

Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K.

(Received May 22, 1996; revised March 10, 1997)

Abstract. The paper discusses a likelihood based method of estimation which allows for a small amount of misspecification in the assumption of normality. Asymptotic results suggest that the new method can give an estimated model which is closer to the true model. An application to hearing threshold data is discussed.

Key words and phrases: Local likelihood, semi-parametric inference, robust estimation, model misspecification.

1. Introduction

Most elementary statistical methods used in practice assume normality. Essentially, although this is usually implicit rather than explicit, a normal distribution is fitted to data and then the estimated parameters of the model are used to make the desired inference. For example a tolerance interval or control limit may be taken as the sample mean plus or minus a fixed multiple of the standard deviation, this multiple being chosen on the assumption of normality. But of course normality never holds exactly, and much recent research concerns diagnostics for model fit, detection of outliers and robust methods. In practice some suitable plot of the data is usually examined and a subjective choice made either to keep the model as it is, to use a transformation or another model, or to abandon parametric inference altogether and use a non-parametric approach. An alternative is to keep within the parametric framework of normality but to allow maximum likelihood estimation to adapt to local departures from the assumed model. We aim to show that allowing for a small amount of model misspecification in this way gives asymptotic inferences which are, in a number of senses to be discussed, closer to those for the true distribution than are the usual inferences derived from an incorrect assumption of normality. Our concern is with departures from the whole model and not just the downweighting of a small proportion of outliers, which is the usual focus of robust methods.

Copas (1995) uses ideas of stochastic censoring in an approach to maximum likelihood which can adapt to local departures from an assumed parametric model.

To motivate the idea here, suppose we have a random sample of size n from a probability density function $f(x,\theta)$, but that the data are interval censored to $(t-h, t+h)$. That is, we only observe the exact values of x_i for those sample points which happen to fall in this interval. Then the likelihood is

$$L_t(\theta) = \sum l_t(x_i,\theta) \tag{1.1}$$

where

$$l_t(x,\theta) = K\left(\frac{x-t}{h}\right)\log f(x,\theta) + \left(1-K\left(\frac{x-t}{h}\right)\right)\log\left(1-\int K\left(\frac{x-t}{h}\right)f(x,\theta)dx\right) \tag{1.2}$$

and $K(u)$ is 1 if $|u|<1$ and 0 otherwise. The semi-parametric likelihood of Copas (1995) is exactly equations (1.1) and (1.2) but with $K(u)$ replaced by a smooth kernel function scaled to take the maximum value 1 at $u=0$.

If $\hat{\theta}_t$ is the value of θ which maximizes (1.1), then this "local maximum likelihood estimate" should give a better local fit to the data in the neighbourhood of t, and so adapt to model departures in that region of the sample space. We use the term "adaption" to describe the way in which estimation adapts to such model departures. The amount of adaption is controlled by h. Clearly (1.1) tends to the ordinary log likelihood $\sum \log f(x_i,\theta)$ as h tends to ∞, and so for large h, $\hat{\theta}_t$ will be close to the ordinary maximum likelihood estimate $\hat{\theta}$. Copas (1995) shows that for small h, $f(t,\hat{\theta}_t)$ is close to the non-parametric kernel estimate of the density of X at t, the kernel estimate using the same $K(u)$ and the same value of h. Here we are interested in a small amount of adaption, and hence in large h, for which the only relevant property of K is

$$K\left(\frac{x-t}{h}\right) = 1 - \frac{(x-t)^2}{2h^2} + O(h^{-4}). \tag{1.3}$$

Without loss of generality we have assumed that the argument of K is scaled so that $K''(0)=-1$. In the numerical example below $K(u)$ is taken as $\exp(-u^2/2)$, but normality of the kernel is not needed for the theory, simply that $K(u)$ is smooth and locally quadratic around $u=0$.

The paper is concerned with asymptotic properties of $\hat{\theta}_t$ and related quantities when both n and h are large. We assume that $f(x,\theta)$ is the normal distribution, but that this model may be misspecified. Section 2 sets out some basic theory. Section 3 compares the three distributions $f(t,\hat{\theta}_t)$, $f(t,\hat{\theta})$ and $g(t)$, the true distribution from which the data are sampled. Using suitable distance measures we show that $f(t,\hat{\theta}_t)$ is on average closer to $g(t)$ than is $f(t,\hat{\theta})$. Section 4 considers the estimation of the true expectation of some given function $s(t)$, showing that the asymptotic mean squared error of $\int s(t)f(t,\hat{\theta}_t)dt$ is always less than that of $\int s(t)f(t,\hat{\theta})dt$ when $s(t)$ is a low order polynomial, and often so for other functions. Section 5 applies the method to the estimation of the upper percentiles of the distribution of hearing threshold. Here the estimates with local adaption fit the tail

much better than the percentiles based on normality alone. A generalization is mentioned briefly in Section 6.

The estimated density $f(t, \hat{\theta}_t)$ for *small* h, which as mentioned is closely related to kernel density estimation, is discussed in Copas (1996). Approximating the mean squared error for large n and small h leads to formulae for the optimum choice of h, rather similar to the corresponding formulae in the theory of kernel density estimation (Silverman (1986)). The small-h case is also related to Hjort and Jones (1996), who use a slightly different approach to semi-parametric density estimation. Here we are concerned with large h, appropriate for small differences between g and f. Interest in the large-h case derives from the fact that, in practice, a parametric model will be chosen because it is sensible in the light of the data and context, and so in applications it is usually only small departures from the model which are important. Typically, large departures would lead to a different model being chosen. By keeping h large the sampling variances of estimates remain similar to those of ordinary maximum likelihood.

Further and more general discussion of the censoring approach to local inference, and of links to dynamic regression modelling, is given in Copas (1995).

2. Some basic theory

Taking $f(x, \theta)$ to be the probability density function of $N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma)^T$, and using equations (1.1) to (1.3) we find for large h

$$l_t(x, \theta) = \log f(x, \theta) - \frac{1}{2h^2}(x - t)^2(\log f(x, \theta) - \log(\sigma^2 + (\mu - t)^2)) + O(h^{-4}). \tag{2.1}$$

Hence, omitting from now on terms of order h^{-4} and smaller,

$$\frac{\partial}{\partial \theta} L_t(\theta) = \sum u(x_i, \theta) - \frac{nT}{2h^2} \tag{2.2}$$

where

$$u(x, \theta) = \frac{\partial}{\partial \theta} \log f(x, \theta),$$

$$T = n^{-1} \sum (x_i - t)^2 (u(x_i, \theta) - v(\theta)), \tag{2.3}$$

and

$$v(\theta) = \frac{\partial}{\partial \theta} \log(\sigma^2 + (\mu - t)^2).$$

It turns out that T in (2.3) is the crucial quantity which determines the difference between $\hat{\theta}_t$ and $\hat{\theta}$. Explicitly, the two components of the vector T are

$$\frac{1}{n} \sum (x_i - t)^2 \left(\frac{x_i - \mu}{\sigma^2} - \frac{2(\mu - t)}{\sigma^2 + (\mu - t)^2} \right)$$

and

$$\frac{1}{n}\sum(x_i - t)^2\left(\frac{(x_i-\mu)^2-\sigma^2}{\sigma^3} - \frac{2\sigma}{\sigma^2+(\mu-t)^2}\right),$$

showing that up to the order of $O(h^{-2})$ the local likelihood inference make use of the skewness and kurtosis of the sample.

Now the ordinary maximum likelihood estimate $\hat{\theta}$ is the solution of the equation

$$\frac{1}{n}\sum u(x_i,\theta) = 0 \tag{2.4}$$

and so, as $n \to \infty$, $\hat{\theta}$ will converge strongly to θ, the solution of the equation

$$\int u(x,\theta)g(x)dx = 0. \tag{2.5}$$

But

$$u(x,\theta) = \left(\frac{x-\mu}{\sigma^2}, \frac{(x-\mu)^2-\sigma^2}{\sigma^3}\right)^T \tag{2.6}$$

and so equation (2.5) implies

$$\mu = E_f(X) = E_g(X) \tag{2.7}$$

and

$$\sigma^2 = E_f(X-\mu)^2 = E_g(X-\mu)^2 \tag{2.8}$$

where the suffix on the expectation operator indicates the distribution over which the expectation is taken. This is the obvious property of the normal distribution that it matches the mean and variance of any distribution to which it is fitted. It follows that

$$E_f(X-t)^2 = E_g(X-t)^2 \tag{2.9}$$

and so

$$E_g(T) = E_g((X-t)^2u(X,\theta)) - (\partial/\partial\theta)(\sigma^2+(\mu-t)^2) \tag{2.10}$$

$$= \int (x-t)^2u(x,\theta)(g(x)-f(x,\theta))dx. \tag{2.11}$$

Note that if $g = f$ then (2.11) is zero and so $E_f(T) = 0$.

Another consequence of (2.7) and (2.8) is that the Hessian $(\partial^2/\partial\theta\partial\theta^T)\log f(x,\theta)$, which involves at most quadratic terms in x, has the same expectation under both f and g, namely

$$H = -\sigma^{-2}\begin{pmatrix}1 & 0\\ 0 & 2\end{pmatrix}.$$

Differentiating (2.2) with respect to θ, and noting that $\hat{\theta}_t$ is the zero of (2.2) and $\hat{\theta}$ is the zero of (2.4), we obtain the first order approximation

$$\hat{\theta}_t - \hat{\theta} = \frac{1}{2h^2} H^{-1} T.$$

Substituting (2.6) into (2.3) and taking expectations thus gives

$$E_g(\hat{\theta}_t - \hat{\theta}) = -\frac{\sigma^2}{4h^2} \begin{pmatrix} 2\sigma b_1 \\ \sigma b_2 - 2(t-\mu) b_1 \end{pmatrix}, \tag{2.12}$$

where b_1 and b_2 are the skewness and kurtosis measures of g, namely

$$b_1 = \sigma^{-3} E_g(X-\mu)^3$$

and

$$b_2 = \sigma^{-4} E_g(X-\mu)^4 - 3.$$

Note that if $g = f$ then $b_1 = b_2 = 0$ and so to this order of approximation $E_f(\hat{\theta}_t) = E_f(\hat{\theta}) = \theta$, as expected.

Much the same method can be used to study the variance of $\hat{\theta}_t$. We find that

$$n \operatorname{Var}_f(\hat{\theta}_t) = n \operatorname{Var}_f(\hat{\theta}) + O(h^{-4}),$$

but in general

$$n \operatorname{Var}_g(\hat{\theta}_t) = n \operatorname{Var}_g(\hat{\theta}) + O(h^{-2}).$$

3. Estimating the true density $g(t)$

In this section we study the estimation of $g(t)$ by the model density $f(t,\theta)$ with θ taken as the maximum likelihood estimate with and without local adaption. To simplify the notation, let $\alpha_t = g(t)$, $\tilde{\alpha}_t = f(t,\hat{\theta}_t)$ and $\hat{\alpha}_t = f(t,\hat{\theta})$.

First note that

$$\tilde{\alpha}_t - \hat{\alpha}_t \simeq f(t,\theta) u^T(t,\theta)(\hat{\theta}_t - \hat{\theta}),$$

and so we can use (2.6) and (2.12) to obtain an approximation (accurate to $O(h^{-4})$) to the difference between the asymptotic biases of the two estimates. This comes to

$$\begin{aligned} &E_g(\tilde{\alpha}_t - \hat{\alpha}_t) \\ &\quad = -\frac{1}{4h^2\sigma} f(t,\theta)(2b_1(t-\mu)(2\sigma^2 - (t-\mu)^2) + \sigma b_2((t-\mu)^2 - \sigma^2)). \end{aligned} \tag{3.1}$$

The difference in the mean squared errors,

$$E_g(\tilde{\alpha}_t - \alpha_t)^2 - E_g(\hat{\alpha}_t - \alpha_t)^2$$

is

$$(\operatorname{Var}_g(\tilde{\alpha}_t) - \operatorname{Var}_g(\hat{\alpha}_t)) + 2E_g(\tilde{\alpha}_t - \hat{\alpha}_t) E_g(\hat{\alpha}_t - \alpha_t) + O(h^{-4}).$$

The first two parts of this expression have the orders of magnitude of $n^{-1}h^{-2}$ and h^{-2} respectively, and so for large n the bias term is dominant, and $\tilde{\alpha}_t$ is better than $\hat{\alpha}_t$ if the bias difference $E_g(\tilde{\alpha}_t - \hat{\alpha}_t)$ has the opposite sign to that of the bias of $\hat{\alpha}_t$. Asymptotic approximations to these quantities are

$$E_g(\hat{\alpha}_t - \alpha_t) = f(t,\theta) - g(t)$$

and $E_g(\tilde{\alpha}_t - \hat{\alpha}_t)$ given by equation (3.1).

Even if the model is misspecified $\hat{\alpha}_t$ will be approximately unbiased for at least one value of t (where the density curves $f(t,\theta)$ and $g(t)$ cross) and so it is not surprising that $\tilde{\alpha}_t$ cannot be uniformly better than $\hat{\alpha}_t$ for all t. To see an example, suppose $g(t)$ is a gamma distribution with shape parameter 2. Without loss of generality we suppose it is shifted and scaled to have $\mu = 0$ and $\sigma = 1$, so that

$$g(t) = 2(t + \sqrt{2})\exp(-\sqrt{2}(\sqrt{2} + t))$$

for $t > -\sqrt{2}$ and zero otherwise. For this distribution $b_1 = \sqrt{2}$ and $b_2 = 3$. Figure 1 shows the asymptotic bias $E_g(\hat{\alpha}_t - \alpha_t)$ and a suitable positive multiple of $E_g(\tilde{\alpha}_t - \hat{\alpha}_t)$. These functions have opposite signs for almost all values of t except for values just above 0, where f and g happen to be close together.

For discrete data the fit of a distribution is usually measured by the chi-squared statistic, the sum of squared differences between observed and expected frequencies weighted inversely with the expected frequencies. Motivated by this, define the risk function

$$\begin{aligned} R_1 &= E_g \int \frac{(\tilde{\alpha}_t - \alpha_t)^2 - (\hat{\alpha}_t - \alpha_t)^2}{\hat{\alpha}_t} dt \\ &\simeq 2 \int E_g(\tilde{\alpha}_t - \hat{\alpha}_t) \frac{f(t,\theta) - g(t)}{f(t,\theta)} dt. \end{aligned}$$

From (3.1) this is approximately

$$\begin{aligned} &-\frac{1}{2h^2\sigma} \int (2b_1(t-\mu)(2\sigma^2 - (t-\mu)^2) + \sigma b_2((t-\mu)^2 - \sigma^2))(f(t,\theta) - g(t))dt \\ &\quad = -h^{-2}\sigma^2 b_1^2. \end{aligned}$$

Hence if $g(t)$ is skewed so that $b_1 \neq 0$, R_1 is negative, indicating that $\tilde{\alpha}_t$ is on average "closer" to α_t than is $\hat{\alpha}_t$.

The same result also holds if we use the Kullback-Leibler distance (or relative entropy) to give the risk function

$$R_2 = -E_g \left(\int \log \frac{\tilde{\alpha}_t}{\alpha_t} \alpha_t dt - \int \log \frac{\hat{\alpha}_t}{\alpha_t} \alpha_t dt \right).$$

To the same approximation this gives

$$\begin{aligned} R_2 &= -E_g \int \frac{\tilde{\alpha}_t - \hat{\alpha}_t}{\hat{\alpha}_t} \alpha_t dt \\ &\simeq \int E_g(\tilde{\alpha}_t - \hat{\alpha}_t) \frac{f(t,\theta) - g(t)}{f(t,\theta)} dt. \end{aligned}$$

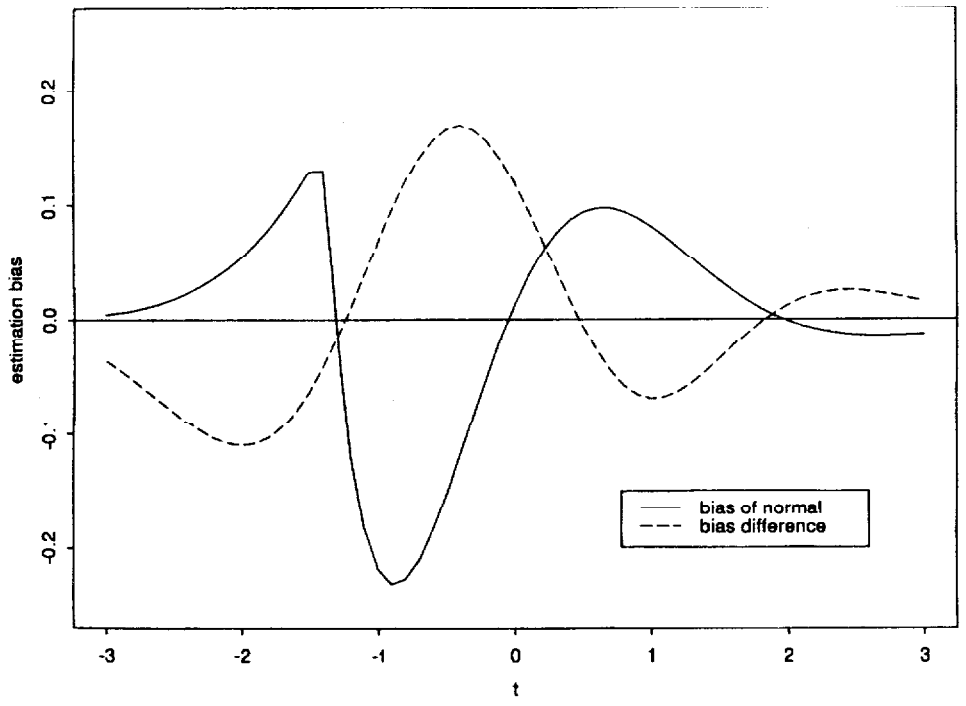

Fig. 1. Estimating gamma density.

Here we have used the fact that the integral of $E_g(\bar{\alpha}_t - \dot{\alpha}_t)$ is of the order $O(h^{-4})$, since the integral of (3.1) over t is zero. Hence

$$R_2 \simeq -\frac{1}{2}h^{-2}\sigma^2 b_1^2,$$

again always negative when $b_1 \neq 0$.

Clearly R_1 and $R_2 \to 0$ as $h \to \infty$, both becoming increasingly negative as h decreases. But these approximations are only valid for large h—the improvement in bias gained by small h has to be offset against the deterioration in variance, which would be substantial if h is small even with large sample sizes. As mentioned above, a parallel asymptotic theory for small h is developed in Copas(1996).

4. Estimating generalized moments

Suppose we now wish to estimate the parameter α taking the form of a "generalized moment"

$$\alpha = \int s(t)g(t)dt,$$

the expectation of a given function s. Our estimates with and without local adaption are

$$\tilde{\alpha} = \int s(t)f(t, \hat{\theta}_t)dt$$

and

$$\hat{\alpha} = \int s(t) f(t, \hat{\theta}) dt.$$

Using the same argument as at the beginning of Section 3 we have

$$E_g(\tilde{\alpha} - \alpha)^2 - E_g(\hat{\alpha} - \alpha)^2 = \mathrm{Var}_g(\tilde{\alpha}) - \mathrm{Var}_g(\hat{\alpha}) + 2E_g(\tilde{\alpha} - \hat{\alpha})E_g(\hat{\alpha} - \alpha) + O(h^{-4}). \tag{4.1}$$

As before this is dominated by the bias terms.

As $f(t, \hat{\theta}_t)$ does not necessarily integrate to one, a more careful definition of $\tilde{\alpha}$ might be

$$\tilde{\alpha}' = \frac{\int s(t) f(t, \hat{\theta}_t) dt}{\int f(t, \hat{\theta}_t) dt}$$

From Section 2 we have that $f(t, \hat{\theta}_t) - f(t, \hat{\theta})$ is of the order of h^{-2} and so we can write $\tilde{\alpha} - \hat{\alpha} = Ah^{-2}$ and $\int f(t, \hat{\theta}_t) dt = 1 + Bh^{-2}$. Then

$$\tilde{\alpha}' - \hat{\alpha} = (A - B\hat{\alpha})h^{-2} + O(h^{-4}).$$

But from (3.1) $E_g(B) = O(h^{-2})$ and so to this approximation $\tilde{\alpha}'$ has the same mean squared error as $\tilde{\alpha}$ and so we can keep to the simpler definition.

The bias terms in (4.1) are obtained from the formulae in Sections 2 and 3. Equation (3.1) implies

$$E_g \int s(t)(\tilde{\alpha}_t - \hat{\alpha}_t) dt = -\frac{1}{4h^2\sigma}(2b_1E_1 + \sigma b_2 E_2), \tag{4.2}$$

where

$$E_1 = E_f(s(X)(X - \mu)(2\sigma^2 - (X - \mu)^2)) \tag{4.3}$$

and

$$E_2 = E_f(s(X)((X - \mu)^2 - \sigma^2)). \tag{4.4}$$

Note that the special case of $s(t) = 1$ gives both E_1 and E_2 to be 0, and so, since $f(t, \theta)$ integrates to one for all θ,

$$E_g \int f(t, \hat{\theta}_t) dt - 1 + O(h^{-4}), \tag{4.5}$$

as already noted in Section 3. This extends equation (25) in Copas (1995) to the case when $g \neq f$. We now use (4.2) to give

$$E_g(\tilde{\alpha} - \hat{\alpha})E_g(\hat{\alpha} - \alpha) = -\frac{1}{4h^2\sigma}E_3(2b_1E_1 + \sigma b_2 E_2) \tag{4.6}$$

where

$$E_3 = E_f(s(X)) - E_g(s(X)).$$

Since the mean and variance of g are matched exactly by f, E_3, and hence (4.6), is 0 when $s(t)$ is any quadratic function of t. Suppose that $s(t) = (t-\mu)^3$. Then $E_1 = -9\sigma^6$, $E_2 = 0$ and $E_3 = -\sigma^3 b_1$. Hence (4.6) is then

$$-\frac{9}{2h^2}\sigma^8 b_1^2$$

which is negative whenever $b_1 \neq 0$.

To add a quartic term we need to allow for the "correlation" between $(t-\mu)^4$ and $(t-\mu)^2$. Let

$$s(t) = \beta_3(t-\mu)^3 + \beta_4((t-\mu)^4 - 15\sigma^2(t-\mu)^2). \tag{4.7}$$

Then $E_1 = -9\beta_3\sigma^6$, $E_2 = -18\beta_4\sigma^6$ and $E_3 = -\beta_3 b_1\sigma^3 - \beta_4 b_2\sigma^4$. Hence (4.6) becomes

$$-\frac{9\sigma^8}{2h^2}(\beta_3 b_1 + \sigma\beta_4 b_2)^2$$

which is also less than or equal to zero.

Any quartic polynomial can be written in the form

$$\beta_0 + \beta_1(t-\mu) + \beta_2(t-\mu)^2 + s(t) \tag{4.8}$$

with $s(t)$ in (4.7). Let

$$\hat{\alpha}^* = \beta_0 + \beta_2\sigma^2 + \hat{\alpha},$$

which is just the usual estimate of the expectation of (4.8) obtained from the fitted normal distribution. The estimate with local adaption is

$$\tilde{\alpha}^* = \beta_0 + \beta_2\sigma^2 + \tilde{\alpha},$$

and so $\tilde{\alpha}^* - \tilde{\alpha} = \hat{\alpha}^* - \hat{\alpha}$. Hence the difference between the mean squared errors of $\tilde{\alpha}^*$ and $\hat{\alpha}^*$ will be the same as the difference between the mean squared errors of $\tilde{\alpha}$ and $\hat{\alpha}$, and so is also less than or equal to zero. In practice the coefficients of this polynomial will be calculated using the sample rather than the population values of μ and σ, but this will not effect the asymptotic comparisons being made.

Section 3 remarks that $f(t, \hat{\theta}_t)$ cannot be a uniformly better estimate of $g(t)$ than $f(t, \hat{\theta})$ for all t, and likewise $\tilde{\alpha}$ cannot be better than $\hat{\alpha}$ for all possible functions $s(t)$. However the above suggests that local adaption will give better estimates for smooth functions s which can be approximated by low order polynomials, and examples suggest that this is so for a wide variety of generalized moments. For instance let $s(t)$ equal 1 if $t \leq a$ and 0 if $t > a$. Then α is the cumulative distribution function of $g(t)$ at $t = a$ and $\hat{\alpha}$ is the corresponding cumulative probability for the normal distribution with the same mean and variance. Here

$$E_3 = \Phi\left(\frac{a-\mu}{\sigma}\right) - \int_{-\infty}^{a} g(t)dt, \tag{4.9}$$

where Φ is the standard normal distribution function. Evaluating the incomplete moments of the normal needed for E_1 and E_2 in (4.3) and (4.4) leads to

$$E_g(\tilde{\alpha} - \hat{\alpha}) - -\frac{1}{4h^2}(a-\mu)\psi\left(\frac{a-\mu}{\sigma}\right)(2b_1(a-\mu) - b_2\upsilon). \tag{4.10}$$

Note that this is zero at $a = \mu$, and so local adaption has no effect (to this approximation) at the mean. It is also zero in both tails ($a \to \pm\infty$), which must be so as the limiting values of $\tilde{\alpha}$ and $\hat{\alpha}$ are either 0 or 1. If g is skewed only ($b_2 = 0$), the bias difference has the same sign in the two tails, but if g is symmetrical ($b_1 = 0$), the bias difference has opposite signs in the two tails, both as expected. More generally, if $(a - \mu)$ changes sign, and b_1 changes sign, then (4.10) also changes sign but retains the same magnitude, again as one would expect from considerations of symmetry.

Equation (4.9) and a suitably scaled version of (4.10) are illustrated in Fig. 2 for the same gamma distribution which was used in Section 3. The two functions have opposite signs (and hence $\tilde{\alpha}$ is better than $\hat{\alpha}$) for almost all values of a except those just less than zero, and local adaption is beneficial for estimating both tails.

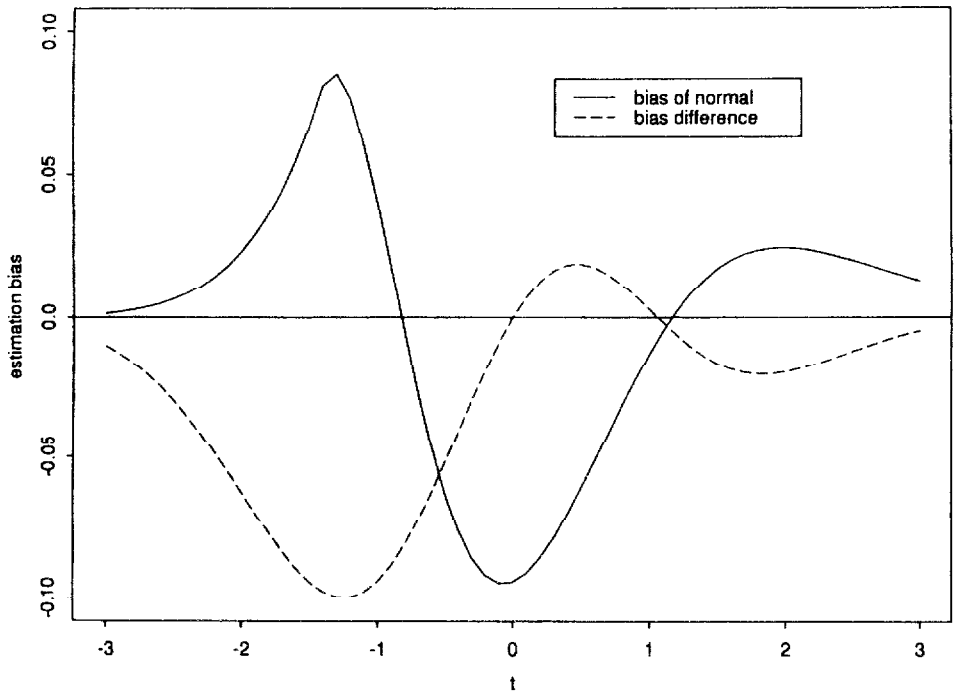

Fig. 2. Estimating cdf of gamma.

5. Application: percentiles of hearing threshold

Davis (1991) presents extensive tables of percentiles of the distribution of hearing threshold, the distribution being broken down by age, sex and occupational group. Briefly, a person's hearing threshold is the lowest volume of sound (measured in decibels, dB) which can just be heard. The upper percentiles of the distribution are of particular interest in assessing hearing disability (King *et al.* (1992)).

Davis's tables are based on a sample survey of the adult population of the U.K. (Davis (1989)), and report case-weighted empirical percentiles calculated from appropriate subsets of the data. An obvious difficulty is that there may be insufficient data for estimation in the less commonly occurring combinations of covariate levels, particularly for the more extreme percentiles. An alternative is to fit a statistical model to the data. Bowater *et al.* (1996) suggest a three parameter lognormal distribution, noting that these data are typically skewed to the right. Much of the literature in this area, however, uses standard statistical techniques and so is tacitly assuming a normal distribution. An example in the book by Longford ((1993), Section 6.8) assumes a normal distribution for such hearing threshold data, but presents versions with and without a logarithmic transformation. Our approach here is to use a normal model but to allow for some local adaption for model misspecification. The large-h theory is relevant in this application since the data are only moderately skewed (the dB scale is already logarithmic), and so only small departures between g and f are of interest.

We illustrate the method by taking Davis's data for male manual workers aged over 60 years, defining hearing threshold X dB to be the average of the three readings obtained for the better ear at the frequencies 1, 2 and 3 kHz. There are 244 observations for which sample estimates of μ and σ are 37.3 and 20.9 respectively. Estimated skewness and kurtosis measures are $b_1 = 0.98$ (± 0.16) and $b_2 = 1.36$ (± 0.31), indicating rather longer tails than the normal. Interest is in the upper tail of the distribution, so we estimate percentiles from about the 90th percentile upwards.

For each of a fine grid of values of t, $\hat{\theta}_t$ was found by numerical maximisation of (1.1), noting that for the normal kernel $K(u) = \exp(-u^2/2)$ the integral needed in (1.2) is available explicitly as

$$\int K\left(\frac{x-t}{h}\right) f(x,\theta)dx = \frac{h}{(h^2+\sigma^2)^{1/2}} \exp\left(-\frac{(t-\mu)^2}{2(h^2+\sigma^2)}\right). \tag{5.1}$$

Values of the cumulative probabilities based on $f(t, \hat{\theta}_t)$ were estimated directly by taking partial sums over the grid of t, normalized so that the value is 1 when t is above the range of the data. The corresponding estimates of the percentiles of the distribution are shown in Fig. 3, using three different values of h as indicated. Also shown in the graph are the actual data (the dots corresponding to the inverse of the empirical distribution function), and the fitted normal model (the solid curve being $\mu + \sigma\Phi^{-1}(\%/100)$). The choice $h = 50$ (two and a half standard deviations) seems about right, $h = 100$ showing very little adaption and $h = 10$ too much in the sense that it is strongly influenced by the pattern of the three highest

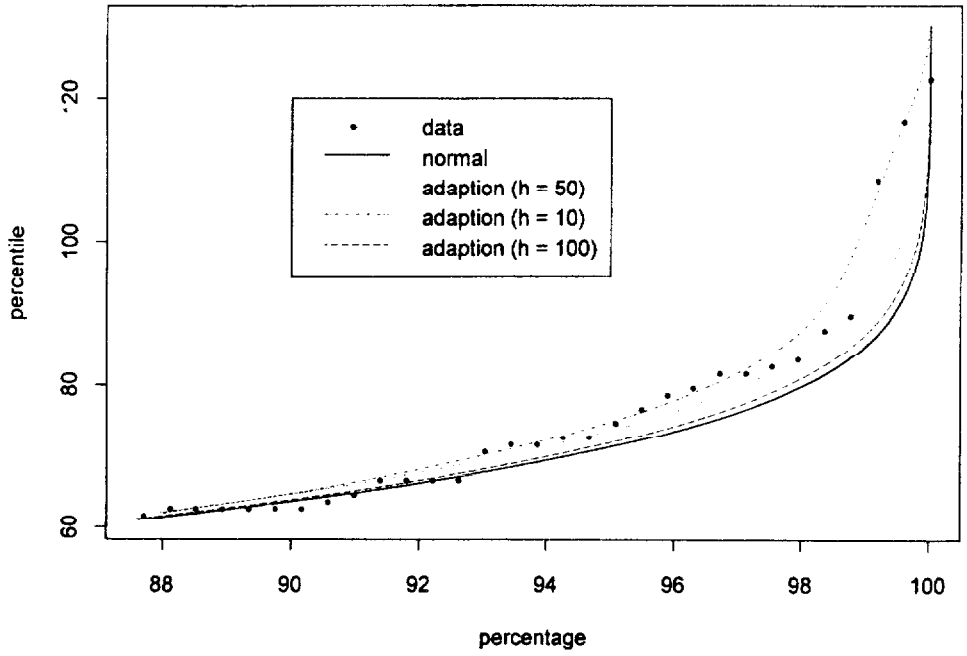

Fig. 3. Upper percentiles of hearing threshold.

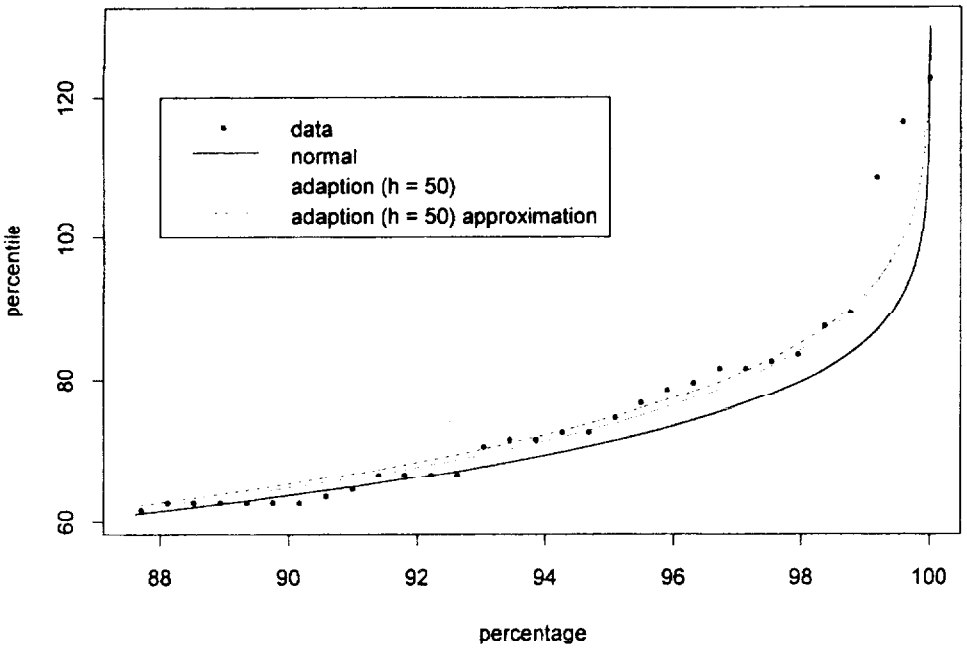

Fig. 4. Checking asymptotic approximation.

observations. Note that the longer right tail of the data is quite well reflected by the percentiles estimated with local adaption.

To check the accuracy of the asymptotic approximations used in the paper, expression (4.10) was evaluated using the sample statistics b_1 and b_2, and taking a to be each value of t in the grid. These were then added to the estimated normal cumulative probabilities to give the large-h approximations to the corresponding estimates with local adaption. Figure 4 compares this approximation for $h = 50$ with the estimates obtained by direct numerical maximisation of the local likelihood, indicating quite good agreement.

6. Generalization

The only property of normality which is important for the theory in the previous sections is the matching of mean and variance in equation (2.9). This leads to the formulae (2.10) and (2.11) and thence to the relatively simple expressions involving the third and fourth moments of g. The normal is not the only distribution having this property—a more general class is the quadratic exponential family

$$f(x, \theta) = \exp(\theta_1 x + \theta_2 x^2 + k(x) + \psi(\theta)) \tag{6.1}$$

where $k(x)$ is a given function of x and $\psi(\theta)$ is the normalizing constant

$$\psi(\theta) = -\log \int \exp(\theta_1 x + \theta_2 x^2 + k(x)) dx,$$

assuming this integral exists. The normal distribution is of course is a special case of (6.1).

All the results of this paper can be generalized to cover this distribution. The role of b_1 and b_2 is now taken by the *difference* between the respective moments of f and g. The two risk differences studied in Section 3 are again less than or equal to zero.

The function $k(x)$ in (6.1) is assumed given. A further generalization is to allow $k(x)$ to depend on additional unknown parameters.

Acknowledgements

We are grateful to Dr. Adrian Davis for access to the data used in Section 5, and to a referee for helpful comments.

References

Bowater, R. J., Copas, J. B., Machado, O. A. and Davis, A. C. (1996). Hearing impairment and the log-normal distribution, *Applied Statistics*, **45**, 203–217.

Copas, J. B. (1995). Local likelihood based on kernel censoring, *J. Roy. Statist. Soc. Ser. B*, **57**, 221–235.

Copas, J. B. (1996). Semi-parametric density estimation by likelihood, *Probability Theory and Mathematical Statistics—Proceedings of the Euler Institute Seminars dedicated to the memory of Kolmogorov* (eds. I. Ibragimov and A. Y. Zaitsev), Gordan and Breach Publishers, London.

Davis, A. C. (1989). The prevalence of hearing impairment and reported hearing disability amongst adults in Great Britain, *International Journal of Epidemiology*, **18**, 911–917.

Davis, A. C. (1991). Means and percentiles of hearing threshold for the adult population aged 18-80, by age group, sex, and occupational group, Institute of Hearing Research Report, Series A, No. 10, Volume 1, Nottingham.

Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation, *Ann. Statist.*, **24**, 1619–1947.

King, P. F., Coles, R. R. A., Lutman, M. E. and Robinson, D. W. (1992). *Assessment of Hearing Disability*, Whurr, London.

Longford, N. T. (1993). *Random Coefficient Models*, Oxford University Press, Oxford.

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