Estimating and visualizing conditional densities

Rob J Hyndman, David M Bashtannyk, Gary K Grunwald

December 1996

Abstract We consider the kernel estimator of conditional density and derive its asymptotic bias, variance and mean-square error. Optimal bandwidths (with respect to integrated mean-square error) are found and it is shown that the convergence rate of the density estimator is order $n^{-2/3}$. We also note that the conditional mean function obtained from the estimator is equivalent to a kernel smoother. Given the undesirable bias properties of kernel smoothers, we seek a modified conditional density estimator which has mean equivalent to some other nonparametric regression smoother with better bias properties. It is also shown that our modified estimator has smaller mean square error than the standard estimator in some commonly occurring situations. Finally, three graphical methods for visualizing conditional density estimators are discussed and applied to a data set consisting of maximum daily temperatures in Melbourne, Australia.

1 Introduction

In this paper we consider the problem of estimating and visualizing the conditional density of $Y \mid \boldsymbol{X}$ where Y is defined on $\mathbb{R}$ and $\boldsymbol{X}$ is a vector defined on $\mathbb{R}^m$. If we were to assume that the conditional density is normal with constant variance and mean linear in $\boldsymbol{X}$, then we would have a standard multiple regression problem. Allowing the mean of Y to vary flexibly with $\boldsymbol{X}$ leads to non-parametric regression methods such as generalized additive models (Hastie and Tibshirani, 1990) or local regression surfaces (Cleveland et al., 1992). However, even non-parametric regression methods usually assume the conditional density does not change over the domain of $\boldsymbol{X}$ apart from changes in mean. We are interested, here, in the situation where the shape of the densities may change with $\boldsymbol{X}$.

To motivate ideas and as a vehicle of illustration, we shall use daily maximum temperatures in Melbourne, Australia over the ten year period 1981–1990. The scatterplot in Figure 1 shows each day's maximum temperature, Y, plotted against the previous day's maximum temperature, X. There is a suggestion of two 'arms' on the right of the plot indicating that a hot day is likely to be followed by either an equally hot day or one much cooler. This indicates the conditional density of Y given a high value of X will be bimodal whereas the conditional density of Y given a low value of X will be unimodal. Figure 2 shows estimates of some of these conditional densities stacked side by side. The bimodal structure is even clearer here than in the scatterplot of Figure 1. We do not suggest that the AR(1) model implied by these plots is a useful model for the time series. The serial dependency in the data is longer than is captured in an AR(1) model and there are many other explanatory variables which need to be used in formulating a realistic meteorological model. However, the example serves as an interesting illustration and demonstrates the need for estimation of the conditional density rather than just the conditional mean and variance.

Surprising little work on conditional density estimation has been published. One notable exception is Stone (1994) who considered using tensor products of polynomial splines to obtain conditional log density estimates. An alternative approach is to estimate a conditional density from estimates of the conditional quantiles. Recent papers on nonparametric conditional quantiles include Chaudhuri (1991), Welsh et al. (1994) as well as those which appeared in the special issue of *J. Nonparametric Statistics* (Saleh, 1994) on regression quantiles.

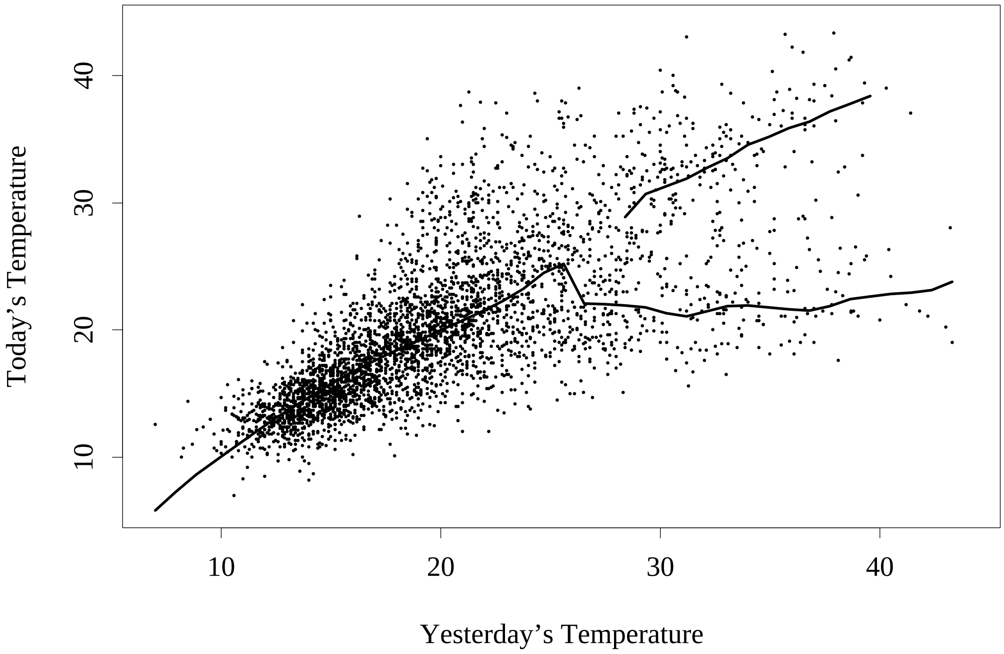

Figure 1: *A lagged scatterplot of each day's temperature against the previous day's temperature. Note the two 'arms' on the right of the plot. The lines shown are from a modal regression discussed in Section 5.3.*

A more direct approach is followed here in which we consider kernel estimators for conditional densities. Although this is probably the most obvious estimator of the conditional density, it does not appear to have received much attention. We know of no other published work which directly applies this estimator to data. Some of its theoretical properties have been considered in the broader context of conditional functional estimation; see Härdle et al. (1988) and Falk (1993) for recent contributions in this area. However, specific properties of the kernel conditional density estimator (such as the mean square error, bias and variance) do not appear to have been previously considered.

The estimator is introduced in Section 2 and in Section 3 we derive its bias, variance and

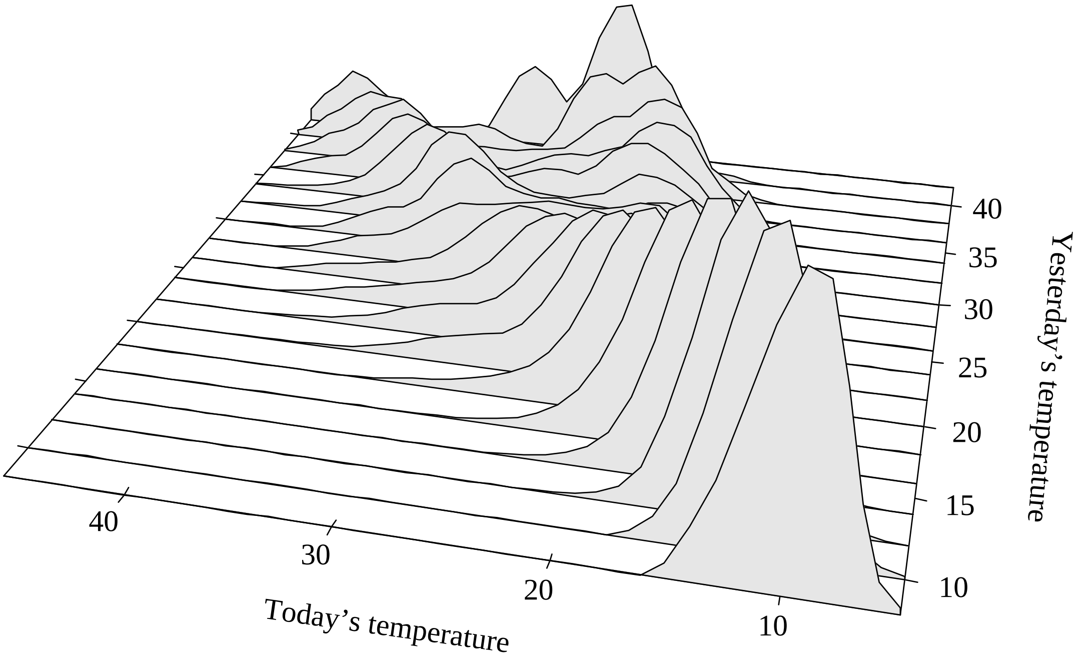

Figure 2: *Stacked conditional density plot of temperature conditional on the previous day's temperature. The bimodality of the distribution of temperature following a hot day is seen more clearly here than in Figure 1.*

mean square error. It is noted in Section 4 that the kernel estimator yields a conditional mean function which is identical to the Nadaraya-Watson kernel smoother. In Section 5 we modify the standard kernel density estimator to obtain conditional densities with mean functions equivalent to other smoothers with better properties than the Nadaraya-Watson smoother. For example, the densities shown in Figure 2 have conditional mean equivalent to a loess (locally linear) regression. Figure 2 shows one of the graphical methods we discuss in Section 6. We also describe a second graphical method based on highest density regions which is more suitable when conditioning over more than one dimension.

2 Kernel estimation of conditional densities

For simplicity, we shall assume that the explanatory variable, X, is univariate and random. The sample shall be denoted by $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ and the observations by $\{(x_1, y_1), \dots, (x_n, y_n)\}$. We shall assume that the bivariate observations are independent. Denoting the conditional mean by $r(x) = \mathsf{E}[Y \mid X = x]$ we can write $Y_j \,|(X_j = x_j) = r(x_j) + \varepsilon_j$ where $\mathsf{E}(\varepsilon_j) = 0$ and the ε_j are independent but not necessarily identically distributed.

We wish to estimate the density of Y conditional on $X = x$. Let $g(x, y)$ be the joint density of (X, Y), $h(x)$ be the marginal density of X and $f(y \,|\, x) = g(x, y)/h(x)$ be the conditional density of $Y \,|(X = x)$. We shall assume that $f(y \,|\, x)$ and $h(x)$ are such that their second derivatives are continuous and square integrable and that $r(x)$ has continuous second derivative.

The natural kernel estimator of $f(y \,|\, x)$ is (e.g., Scott, 1992, p.220)

$$\hat{f}(y \,|\, x) = \frac{\hat{g}(x, y)}{\hat{h}(x)} \tag{1}$$

where

$$\hat{g}(x, y) = \frac{1}{nab} \sum_{j=1}^{n} K\left(\frac{\|x - X_j\|_x}{a}\right) K\left(\frac{\|y - Y_j\|_y}{b}\right)$$

is the kernel estimator of $g(x, y)$ and

$$\hat{h}(x) = \frac{1}{na} \sum_{j=1}^{n} K\left(\frac{\|x - X_j\|_x}{a}\right)$$

is the kernel estimator of $h(x)$. Here, $\|\cdot\|_x$ and $\|\cdot\|_y$ are distance metrics on the spaces of X and Y respectively. A multivariate kernel other than the product kernel could have been used to define $\hat{g}(x, y)$. But the product kernel is simpler to work with, leads to conditional density estimators with several nice properties and is only slightly less efficient than other kernels (Wand and Jones, 1995). In this paper we shall use the Euclidean distance for all numerical examples, except where otherwise stated. The kernel function, $K(x)$, is assumed to be a real, integrable, non-negative, even function on $\mathrm{I\!R}$ concentrated at the origin such that

$$\int_{\mathrm{I\!R}} K(x)dx = 1, \qquad \int_{\mathrm{I\!R}} xK(x)dx = 0 \qquad \text{and} \qquad \sigma_K^2 = \int_{\mathrm{I\!R}} x^2 K(x)dx < \infty. \tag{2}$$

Popular choices for $K(x)$ are defined in terms of univariate and unimodal probability density functions. In this paper, for all numerical examples we use the Epanechnikov kernel,

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{for } |x| < 1; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We shall rewrite (1) as

$$\hat{f}(y \,|\, x) = \sum_{j=1}^{n} w_j(x) \frac{1}{b} K\left(\frac{\|y - Y_j\|_y}{b}\right) \tag{4}$$

where

$$w_j(x) = K\left(\frac{\|x - X_j\|}{a}\right) \Bigg/ \sum_{i=1}^{n} K\left(\frac{\|x - X_i\|}{a}\right). \tag{5}$$

Here, a controls the smoothness between conditional densities in the x direction and b controls the smoothness of each conditional density in the y direction.

3 Asymptotic properties of the kernel estimator

Mean square error and convergence

The asymptotic bias and variance of the density estimator are derived in the appendix, and are shown to be (as $a \to 0$, $b \to 0$ and $n \to \infty$)

$$\begin{aligned} \mathsf{E}\hat{f}(y \,|\, x) - f(y \,|\, x) &= \frac{a^2\sigma_K^2}{2}\left\{2\frac{h'(x)}{h(x)}\frac{\partial f(y \,|\, x)}{\partial x} + \frac{\partial^2 f(y \,|\, x)}{\partial x^2} + \frac{b^2}{a^2}\frac{\partial^2 f(y \,|\, x)}{\partial y^2}\right\} \\ &\qquad + O(a^4) + O(b^4) + O(a^2b^2) + O(\tfrac{1}{na}) \end{aligned} \tag{6}$$

and

$$\mathsf{Var}[\hat{f}(y \,|\, x)] = \frac{R(K)f(y \,|\, x)}{nabh(x)}[R(K) - bf(y \,|\, x)] + O(\tfrac{1}{n}) + O(\tfrac{b}{an}) + O(\tfrac{a}{bn}) \tag{7}$$

where $R(K) = \int K^2(w)dw$.

Adding the variance (7) to the squared bias (6) gives the asymptotic mean square error

$$\begin{aligned} \text{AMSE} &= \frac{a^4\sigma_K^4}{4}\left\{2\frac{h'(x)}{h(x)}\frac{\partial f(y \,|\, x)}{\partial x} + \frac{\partial^2 f(y \,|\, x)}{\partial x^2} + \frac{b^2}{a^2}\frac{\partial^2 f(y \,|\, x)}{\partial y^2}\right\}^2 \\ &\qquad + \frac{R(K)f(y|x)}{nabh_X(x)}[R(K) - bf(y \,|\, x)] \\ &\qquad + O(\tfrac{1}{n}) + O(\tfrac{b}{an}) + O(\tfrac{a}{bn}) + O(a^6) + O(b^6) + O(a^2b^4) + O(a^4b^2). \end{aligned} \tag{8}$$

Thus the estimator is consistent provided $a \to 0$, $b \to 0$ and $nab \to \infty$ as $n \to \infty$. As with many smoothing problems, small bandwidths give small bias and large variance whereas large bandwidths give large bias and small variance. Bandwidths chosen to minimize (8) give a trade-off between bias and variance.

The integrated asymptotic mean square error (IMSE) is obtained by taking the integral with respect to both x and y of the weighted AMSE formed by the product of (8) with $h(x)$. This weighting provides a bounded global accuracy measure with more emphasis on those regions with more data. Similar weighting is used in regression smoothing (e.g. Wand and Jones, 1995). The resulting expression is of the form

$$\text{IMSE} \approx \frac{c_1}{nab} - \frac{c_2}{na} + c_3 a^4 + c_4 b^4 + c_5 a^2 b^2. \tag{9}$$

where the constants c_1, c_2, c_3, c_4 and c_5 depend on the kernel K, the conditional density $f(y \,|\, x)$ and the marginal density $h_X(x)$.

The optimal bandwidths can be derived by differentiating (9) with respect to a, b and setting the derivatives to 0. Taking these derivatives and simplifying we obtain the following expressions

$$\begin{aligned} -\frac{c_1}{n} + \frac{c_2 b}{n} + 4c_3 a^5 b + 2c_5 a^3 b^3 &= 0 && (10) \\ -\frac{c_1}{n} + 4c_4 a b^5 + 2c_5 a^3 b^3 &= 0. && (11) \end{aligned}$$

Subtracting (10) from (11) and solving for b gives

$$b = \left\{ \frac{c_2}{4c_4 a n} + \frac{c_3 a^4}{c_4} \right\}^{1/4}. \tag{12}$$

Substituting b into equation (11) and expanding in a series about $(na)^{-1}$ using a symbolic algebra package we obtain

$$\frac{-c_1}{n} + \sum_{i=0}^{\infty} k_i n^{-i} a^{6-5i} = 0 \tag{13}$$

where each k_i is a function of c_1, c_2, c_3, c_4 and c_5. In particular

$$k_0 = \left\{ 4\Big(\frac{c_3^5}{c_4}\Big)^{1/4} + 2\Big(\frac{c_3}{c_4}\Big)^{3/4} c_5 \right\} \text{ and } k_1 = c_2 \left\{ \frac{5}{4}\Big(\frac{c_3}{c_4}\Big)^{-1/4} + \frac{3}{8}\Big(\frac{1}{c_3 c_4^3}\Big)^{1/4} c_5 \right\}.$$

For (13) to converge to 0 we require a to be of order n^s where $\frac{-1}{5} < s < 0$. Taking the two most dominant terms from the series and solving for a we obtain the optimal value of a:

$$a^* = \left(\frac{c_1}{nk_0}\right)^{1/6} = c_1^{1/6} \left\{ 4\left(\frac{c_3^5}{c_4}\right)^{1/4} + 2c_5 \left(\frac{c_3}{c_4}\right)^{3/4} \right\}^{-1/6} n^{-1/6}.$$

Substituting a^* into (12), expanding about n^{-1} and taking the most dominant term we find the optimal value of b:

$$b^* = \left(\frac{c_3}{c_4}\right)^{1/4} a^* = c_1^{1/6} \left\{ 4\left(\frac{c_4^5}{c_3}\right)^{1/4} + 2c_5 \left(\frac{c_4}{c_3}\right)^{3/4} \right\}^{-1/6} n^{-1/6}.$$

The above two equations show that both a^* and b^* are of order $n^{-1/6}$. Substituting a^* and b^* into (8) shows the IMSE has order $n^{-2/3}$ — the same as for a bivariate kernel estimator (Scott, 1992).

Compare these results to those obtained for a univariate kernel density estimator in which b^* is of order $n^{-1/5}$ and the IMSE is of order $n^{-4/5}$ (see, for example, Scott, 1992). It is to be expected that the convergence properties are better in the univariate case, since we are effectively reducing the number of points used in the estimates when we condition on the value of X.

Of course, a^* and b^* are not practical bandwidth selection rules since they are functions of the unknown density $h(x)$ and conditional density $f(y\,|\,x)$. But they serve as useful benchmarks for what is possible. A rough rule of thumb is obtained by assuming the conditional and marginal densities are normal or some other parametric form. Alternatively, the bandwidths optimal for bivariate density estimation (see Wand and Jones, 1994) may provide a guide for use in conditional density estimation, although the optimality criterion is different.

A special case

Some insight into the bias and AMSE expressions is possible by considering the special case (shown on the left of Figure 3) where the design points are locally uniform near x, the conditional densities near x are identical apart from a shift in location and $r(x)$ is locally linear near x. Hence, $h'(x) \approx 0$, $h''(x) \approx 0$, $f(y\,|\,x) = p(y - r(x))$ where $\int up(u)du = 0$ and $r''(x) \approx 0$. Then,

$$\begin{aligned} \frac{\partial f(y\,|\,x)}{\partial x} &\approx -p'(y-r(x))r'(x), \\ \frac{\partial^2 f(y\,|\,x)}{\partial x^2} &\approx p''(y-r(x))[r'(x)]^2 \\ \text{and} \qquad \frac{\partial^2 f(y\,|\,x)}{\partial y^2} &\approx p''(y-r(x)). \end{aligned}$$

Therefore,

$$\mathsf{E}\hat{f}(y\,|\,x) - f(y\,|\,x) \approx \tfrac{1}{2}\sigma_K^2 p''(y-r(x))\left\{a^2[r'(x)]^2 + b^2\right\}. \tag{14}$$

The variance is unchanged under these conditions so that

$$\text{IMSE} \approx \frac{\sigma_K^4 R(p'')}{4}\int\left\{a^2[r'(x)]^2 + b^2\right\}^2 h(x)dx + \frac{cR(K)}{nab}\left[R(K) - bR(p)\right] \tag{15}$$

where $R(p) = \int p^2(w)dw$, $R(p'') = \int [p''(w)]^2 dw$ and c is the range of X.

Note that $\hat{f}(y\,|\,x)$ will have greater bias when the slope in the mean, $r'(x)$, is greater. But if $r(x)$ is constant, (14) reduces to $\mathsf{E}\hat{f}(y\,|\,x) - f(y\,|\,x) = \frac{1}{2}b^2\sigma_K^2 h_Y''(y)$, which is the bias of a univariate kernel estimator of the marginal density of Y (Scott, 1992).

This case is not of great intrinsic interest since we would normally wish to use conditional density estimation when the densities are changing shape with x. However, it does show that, in this case, the conditional density estimator will have greater asymptotic bias than a univariate density estimator unless $r(x)$ is constant and the IMSE is reduced if $r(x)$ is constant.

4 Conditional moments

The mean of the conditional density estimator $\hat{f}(y \,|\, x)$ provides an estimator of the conditional mean $r(x)$, namely

$$\hat{m}(x) \;:=\; \int y\hat{f}(y \,|\, x)dy \;=\; \sum_{j=1}^{n} w_j(x)Y_j. \tag{16}$$

This is identical to the kernel regression function of Nadaraya (1964) and Watson (1964). In fact, this is how the Nadaraya–Watson smoother is often derived (e.g. Scott, 1992 and Härdle, 1991). Note that $\hat{m}(x)$ depends on a, the smoothing parameter in the x direction, but not on b, the smoothing parameter in the y direction.

With a little more algebra, we obtain an estimator of the conditional variance $\mathsf{Var}[Y \,|\, X = x]$ in a convenient form:

$$\hat{v}(x) \;:=\; \int [y - \hat{m}(x)]^2 \hat{f}(y \,|\, x)dy \;=\; b^2\sigma_K^2 + \sum_{j=1}^{n} w_j(x)[Y_j - \hat{m}(x)]^2. \tag{17}$$

Note that this variance consists of two terms, one proportional to b^2 and the other a weighted sum of squares of the differences about the estimated mean, $\hat{m}(x)$. The first term depends only on b, the smoothing parameter in the y direction, and the second term depends only on a, the smoothing parameter in the x direction. For $b = 0$, we obtain a weighted sum of the squared residuals which is commonly used as a local estimate of the conditional variance. (e.g. Hall and Carroll, 1989).

Bias in the conditional mean estimator

While a kernel regression provides an intuitive and simple estimate of the conditional mean, it can have large bias. Conditional on the observed values of $X_1, \dots, X_n$, the bias of $\hat{m}(x)$ is

$$\begin{aligned} &\mathsf{E}[\hat{m}(x) \,|\, X_1 = x_1, \dots, X_n = x_n] - r(x) \\ &\qquad = \; r'(x)\sum_{j=1}^{n} w_j(x)(x_j - x) + \frac{r''(x)}{2}\sum_{j=1}^{n} w_j(x)(x_j - x)^2 + R \end{aligned} \tag{18}$$

where the remainder R is small under some regularity conditions due to the locality of the kernel (see, for example, Hastie and Loader, 1993). Hence, there may be substantial bias on the boundary of the predictor space because the asymmetry of the kernel neighbourhood causes the first term to be large when $r'(x)$ is large. This is seen in the plot on the left of Figure 3 which shows a data set with linear mean and uniformly distributed x_j. The kernel neighbourhood around $x = x_0$ is marked by the shaded region. Bias can also be a problem in the interior if the true mean function has substantial curvature (if $|r''(x)|$ is large then the second term in (18) is large) or if the design points are very irregularly spaced (again giving some asymmetric neighbourhoods). Such problems are largely eliminated with some other smoothing methods. In Section 5 we modify this kernel conditional density estimator to allow the conditional mean function to be specified or estimated using a smoother with better properties.

We can take the expectation of (18) with respect to $X_1, \ldots, X_n$ to obtain the unconditional bias (e.g. Scott, 1992)

$$\mathsf{E}[\hat{m}(x)] - r(x) = \frac{1}{2}a^2\sigma_K^2 \left\{ r''(x) + 2r'(x)\frac{h'(x)}{h(x)} \right\} + O(a^4). \tag{19}$$

Note that both (18) and (19) show there is approximately zero bias if $r(x) = c$ is constant for all x. In fact, the bias is exactly zero in this case since then $\mathsf{E}(Y_j) = c$ and so $\mathsf{E}[\hat{m}(x) \,|\, X_1 = x_1, \ldots, X_n = x_n] = \sum_j w_j(x)\mathsf{E}(Y_j) = c$.

We shall call the bias in the estimated mean, given by (18) or (19), the *mean-bias* of $\hat{f}(y \,|\, x)$ to distinguish it from the bias in the estimator itself.

5 Modified kernel estimator

5.1 An improved conditional density estimator

From the preceding discussion, we have seen that the standard kernel estimator of a conditional density suffers from many of the same problems as kernel regression. We wish to modify $\hat{f}(y \,|\, x)$ to obtain a new conditional density estimator which has a mean function corresponding to a smoother with better bias properties than kernel smoothing.

As a first step, we assume the true mean $r(x)$ is known and consider possible conditional density estimators which have either mean identical to $r(x)$ or mean equal to $r(x)$ in expected value. In Section 5.2 we will replace $r(x)$ by an estimate which has better properties than a kernel smoother.

Our first approach is motivated by the fact that the error, ε_j, has the same distribution as y_j except for a shift in conditional mean. So we could estimate the conditional density of $Y - r(x) \,|(X = x)$ by applying the standard kernel density estimator (4) to the points $\{(x_j, \varepsilon_j)\}$. Figure 3 shows the transformation graphically. We shall denote the mean of the estimated density of $\varepsilon \,|(X = x)$ by $\hat{m}_1(x)$ which will be approximately zero for all x. Then $r(x)$ can be added to these estimated conditional densities to obtain an estimate of the density of $Y \,|(X = x)$. The advantage of this approach is that the conditional density of $\varepsilon \,|(X = x)$ has mean function which is constant (see the right hand graph in Figure 3). Hence, the mean-bias (19) is zero and the IMSE is reduced under the conditions of (15).

More formally, we define the new conditional density estimator as

$$\hat{f}_1(y \,|\, x) = \frac{1}{b}\sum_{j=1}^{n} w_j(x) K\left(\frac{\|y - Y_j^{(1)}(x)\|_y}{b} \right)$$

where $Y_j^{(1)}(x) = \varepsilon_j + r(x) = Y_j - r(x_j) + r(x)$. This estimator has mean function

$$\hat{r}_1(x) \;:=\; \int y\hat{f}_1(y \,|\, x)dy \;=\; r(x) + \hat{m}_1(x) \;=\; r(x) + \sum_{j=1}^{n} w_j(x)\varepsilon_j.$$

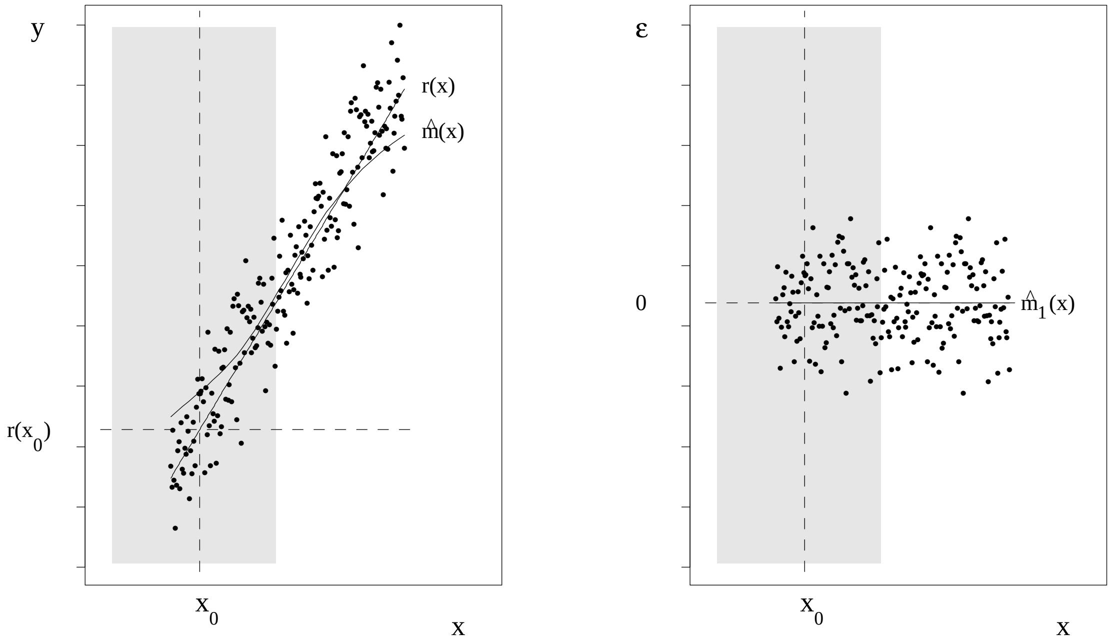

Figure 3: *Graphical representation of the transformation of y_j to ε_j.*

Then $\mathsf{E}[\hat{r}_1(x)] = r(x)$ (but $\hat{r}_1(x)$ is not identical to $r(x)$ since $\hat{m}_1(x)$ is not identical to zero). Also, since the conditional mean of $\varepsilon\,|(X = x)$ is a constant, the IMSE of $\hat{f}_1(y\,|\,x)$ is smaller than that of $\hat{f}(y\,|\,x)$ under the conditions of (15).

We can improve this estimator slightly by defining

$$Y_j^{(2)}(x) = Y_j^{(1)}(x) - \hat{m}_1(x) + r(x) = r(x) + \varepsilon_j - \sum_{i=1}^{n} w_i(x)\varepsilon_i.$$

Applying the standard kernel conditional density estimator (4) to the data $(X_j, Y_j^{(2)}(x))$ gives

$$\hat{f}_2(y\,|\,x) = \frac{1}{b}\sum_{j=1}^{n} w_j(x) K\left(\frac{\|y - Y_j^{(2)}(x)\|_y}{b}\right). \tag{20}$$

The mean function of $\hat{f}_2(y\,|\,x)$ is

$$\hat{r}_2(x) \;:=\; \int y\hat{f}_2(y\,|\,x)dy \;=\; \sum_{j=1}^{n} w_j(x)Y_j^{(2)}(x) \;=\; r(x).$$

Hence, not only does this method give zero mean-bias ($\mathsf{E}[\hat{r}_2(x)] - r(x) = 0$), but the mean of the estimated density, $\hat{r}_2(x)$, is identical to $r(x)$. Also, $\hat{f}_2(y\,|\,x)$ inherits the IMSE of $\hat{f}_1(y\,|\,x)$. So this estimator has mean exactly $r(x)$ and its IMSE is small under the conditions of (15).

Using (17), an improved estimate of the conditional variance $\mathsf{Var}[Y \mid X = x]$ can be computed using $\hat{f}_2(y \mid x)$:

$$
\begin{aligned}
\hat{v}_2(x) &:= \int [y - \hat{r}_2(x)]^2 \hat{f}_2(y \mid x) dy \\
&= b^2\sigma_K^2 + \sum_{j=1}^{n} w_j(x)[Y_j^{(2)}(x) - r(x)]^2 \\
&= b^2\sigma_K^2 + \sum_{j=1}^{n} w_j(x)\Big[\varepsilon_j - \sum_{i=1}^{n} w_i(x)\varepsilon_i\Big]^2.
\end{aligned}
$$

5.2 Estimators with mean specified by a smoother

Of course, we will usually not know the true mean, $r(x)$, and so the estimator introduced above cannot be used in the form given. However, there are numerous methods for estimating a mean $\mathsf{E}(Y \mid X = x)$ which have better properties than kernel regression. For example, we could obtain a density estimate with mean equivalent to a linear regression, a local polynomial (e.g. constant, linear, quadratic) regression (Hastie and Loader, 1993) or a cubic smoothing spline (Buja et al., 1989; Silverman, 1984).

If we replace $r(x)$ by any suitable estimate of the mean, $\hat{r}(x)$, in (20), we obtain the following conditional density estimator:

$$
\hat{f}_*(y \mid x) = \frac{1}{b}\sum_{j=1}^{n} w_j(x) K\left(\frac{\|y - Y_j^*(x)\|_y}{b}\right) \tag{21}
$$

where

$$
Y_j^*(x) = \hat{r}(x) + e_j - \sum_{i=1}^{n} w_i(x) e_i
$$

and $E_i = Y_i - \hat{r}(x_i)$. Then, using the results obtained for $\hat{f}_2(y \mid x)$, we find that the mean of $\hat{f}_*(y \mid x)$ is $\hat{r}(x)$ and the variance is

$$
\hat{v}_*(x) := \int [y - \hat{r}(x)]^2 \hat{f}_*(y \mid x)\, dy = b^2\sigma_K^2 + \sum_{j=1}^{n} w_j(x)\Big[E_j - \sum_{i=1}^{n} w_i(x) E_i\Big]^2.
$$

Hence, the mean-bias of $\hat{f}_*(y \mid x)$ is simply the bias of $\hat{r}(x)$. Also, like $\hat{f}_2(y \mid x)$, the IMSE of $\hat{f}_*(y \mid x)$ is smaller than that of $\hat{f}(y \mid x)$ under the conditions of (15). We have done some numerical studies involving more complicated conditional densities and have found that the MSE of $\hat{f}_*(y \mid x)$ is often, but not always, smaller than that of $\hat{f}(y \mid x)$. The results of this comparison will be reported in a later paper.

In replacing $r(x)$ by $\hat{r}(x)$ we often introduce an extra smoothing parameter. In addition to a and b which play the same role as they do with the estimator (4), a smoother specified by $\hat{r}(x)$ usually also has a smoothing parameter which we shall denote by c. Note that both c and a control smoothness in the x direction; a controls how quickly the conditional densities can change in shape and spread while c controls the smoothness of the mean of the conditional densities over x.

Some special cases are worth noting. Suppose $\hat{r}(x)$ is a linear smoother defined by

$$\hat{r}(x) = \sum_{j=1}^{n} l_j(x)Y_j.$$

1. Setting $l_j(x) = w_j(x) = 1/n$ (obtained by letting $c = a \to \infty$) means (21) gives the kernel estimator of the marginal density of Y (e.g. Scott, 1992).
2. A mean function equivalent to local polynomial regression is obtained by setting
$$l_j(x) = b(x)(B^T V(x)B)^{-1}b^T(x_j)v_j(x)$$
where $b(x)$ is a row vector containing an expansion of x into a basis of polynomials, B is a matrix with $b(x_i)$ as its ith row $(i = 1, \dots, n)$, $v_j(x)$ is a weighting function of the same form as $w_j(x)$ defined in (5) but with window width c replacing a, and $V(x) = \text{diag}[v_1(x), \dots, v_n(x)]$. For example, local linear regression is obtained with $b(x) = [1\ x]$. Note that if $b(x) = 1$, $\hat{r}(x)$ is a kernel regression mean but, unless $c = a$, one with different smoothness to $\hat{m}(x)$.
3. Setting $v_j(x) = 1/n$ in the above case (obtained by letting $c \to \infty$) gives $\hat{r}(x)$ equivalent to linear regression.

Bandwidth selection is obviously a crucial issue in using conditional density estimation for data analysis. However, this is beyond the scope of this paper and we plan to consider it elsewhere. In the numerical examples considered here, bandwidths were chosen by trial and error to give estimates which seem reasonable for the data. Figure 2 was created with $a = 3$ and $b = 2.5$ with the mean function specified by estimated on Splus 3.1 using a loess (locally linear) smoother (Cleveland et al., 1992) with span 50%. The particular densities shown are conditional on the previous day's temperature being 8, 10, 12, ..., 42 degrees Celsius. Since the density of the errors changes with X, it would be reasonable to allow b to change with X as well, although this has not be done here.

5.3 Higher order estimates

The extension to allow several explanatory variables is straightforward. Where $\boldsymbol{X}$ is a vector of length m, we replace $w_j(x)$ by

$$w_j(\boldsymbol{x}) = \frac{K_m(\|\boldsymbol{x} - \boldsymbol{x}_j\|)}{\sum\limits_{j=1}^{n} K_m(\|\boldsymbol{x} - \boldsymbol{x}_j\|)}$$

where $K_m(\cdot)$ is a multivariate kernel function. A popular choice is the product kernel of the form

$$K_m(\boldsymbol{x}) = \prod_{k=1}^{m} \frac{1}{a_k} K\left(\frac{x^{(k)}}{a_k}\right)$$

where $K(\cdot)$ is a univariate kernel function, $x^{(k)}$ is the kth component of $\boldsymbol{x}$ and a_k denotes the window width for $x^{(k)}$.

Similarly, replace $\hat{r}(x)$ by a multivariate smoother $\hat{r}(\boldsymbol{x})$. Then the conditional density estimate is given by (21), with x replaced by $\boldsymbol{x}$. The mean and variance results are exactly analogous. However, the MSE in this case is more difficult to assess.

6 Graphical display

6.1 Modal regression

Scott (1992, pp.233–235) considered using the modes of a conditional density estimate as a form of robust nonparametric regression. The modal regression line used by Scott is

$$\hat{a}(x) = \arg\,\max_y \hat{f}(y \,|\, x),$$

that is the value of the argument, y, which maximizes $\hat{f}(y \,|\, x)$. An alternative form is

$$\hat{a}_s(x, \lambda) = \text{args}\,\max_y \hat{f}(y \,|\, x)$$

where args indicates all local maxima greater than $\lambda\hat{a}(x)$ and $0 < \lambda < 1$. This provides a more useful graphical device than the mean or median when the densities are multimodal.

Figure 1 shows $\hat{a}_s(x, 0.4)$ superimposed over the scatterplot of the temperature data. Here the conditional densities were estimated using a loess (locally quadratic) mean with span $c = 0.6$ computed using Splus 3.1 and smoothing parameters $a = 3.5$ and $b = 3.6$. Larger smoothing parameters were used in computing the modal regression lines in Figure 1 than in Figure 2 so as to remove a number of spurious local modes. As with a and b, $\lambda = 0.4$ was chosen by trial and error to allow the display of two modes only. Smaller values of λ result in small, probably spurious, modes appearing.

6.2 Stacked conditional density plots

Figure 2 shows a number of densities plotted side by side in a perspective plot. We call this a 'stacked conditional density plot'. It allows the changes in the shape of the distribution over the range of the conditioning variable to be seen clearly. We have found this plot is much more informative than the traditional displays of three dimensional functions (e.g. contour plots or three-dimensional perspective plots) since it highlights the conditioning. Furthermore, aspects of the traditional graphical forms such as contour lines are very difficult to interpret in this context because their relation to the conditional densities is not clear.

Scott (1992, p.23) argues against displaying the conditional densities and prefers a display of 'slices' of the joint density. If the goal is to understand the joint density (as it was in Scott's example), than taking slices is preferable because it reduces the visual prominence and relative noisiness of the tails. However, in this example, we are more interested in the conditional densities than the joint density. The increasing bimodality in the conditional densities shown in Figure 2 is much less obvious in slices of the joint density.

The conditional densities become rather noisy at the extremes because of the sparsity of data in those regions. This is partly a result of having a fixed bandwidth. An adaptive bandwidth would allow more smoothing in the tails to overcome this problem. However, variable bandwidths add further degrees of complexity to the problem and we have not yet tackled this problem.

6.3 Highest density region plots

An alternative approach is to plot a number of highest density regions (HDRs) against the conditioning variable (Hyndman, 1996). A highest density region is the smallest region of the sample space containing a given probability. Figure 4 shows a plot of the 50% and 99% HDRs for the Melbourne temperature data, computed from the density estimates shown in Figure 2.

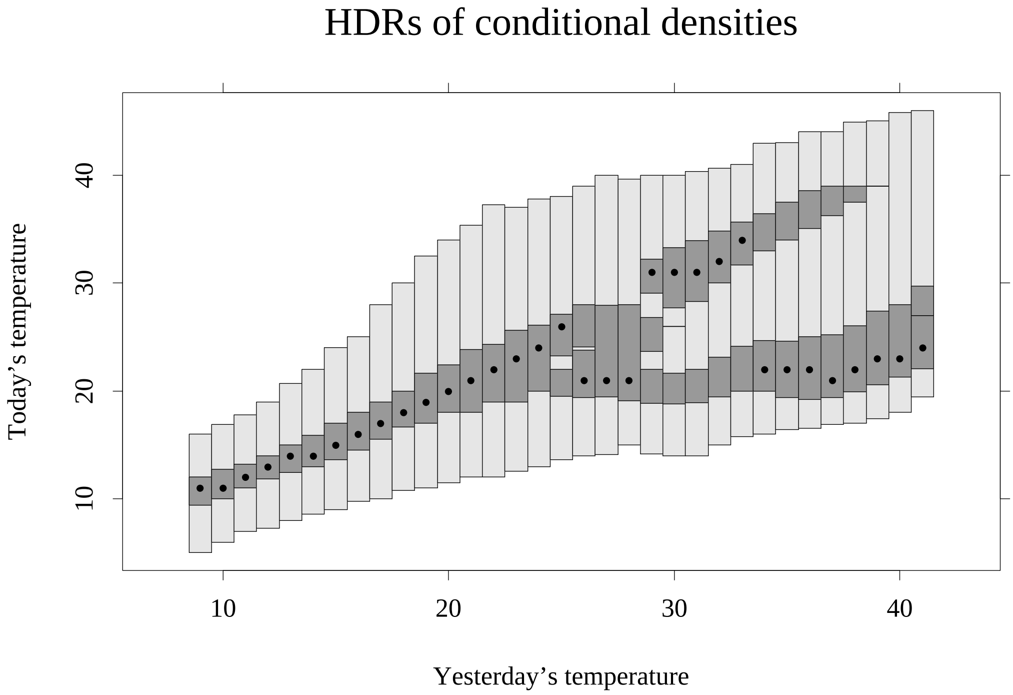

Figure 4: *Highest density regions (50% and 99%) for maximum daily temperature conditional on the previous day's maximum temperature. Conditional modes are also marked (by •) for each x value. Compare this plot with the scatterplot of Figure 1 and the modal regression plot of Figure 5.*

Each vertical strip represents the conditional density for one x value. The x values on which we condition are chosen to lie at 1° intervals. The darker shaded region in each strip is a 50% HDR while the lighter shaded region is a 99% HDR. The mode for each conditional density is also shown as a bullet (•). This plot reveals the bimodality most clearly of all without the distracting smaller bumps that often occur in kernel density estimates. For more details regarding highest density regions and the algorithm used to compute the graph shown here, see Hyndman (1996).

When $\boldsymbol{X}$ is of higher dimension than one, it is necessary to produce several such plots or show them dynamically with one of the conditioning variables changing over time. The following section gives an example of conditional density estimates with two conditioning variables displayed by highest density region plots.

6.4 Application to the Melbourne temperature data

It is clear from Figures 1, 2 and 4 that the mean and variance of today's maximum temperature increase as yesterday's maximum temperature increases, except on very hot days (over 30°C) which are often followed by cooler days. The 50% HDRs consist of two disjoint intervals showing that days of 30–39°C tend to be followed by days of similar temperature or of much lower temperature; they are not generally followed by days with maximum temperature in the high 20s. This occurs because temperatures slowly increase as high pressure systems pass over the city from west to east. At the tail end of a high pressure system, a strong north wind often blows (from off the Australian mainland) bringing high temperatures. A high pressure system is often followed by a cold front causing a rapid drop in temperature. Hence, hot days are generally followed by days of similar or greater temperature or by much cooler days.

It seems reasonable that the temperature distribution may also change with the time of the year. To investigate this idea, we use the approach of Section 5.3 and condition on both the day of the year and the previous temperature. Let $Y_{i,j}$ be the maximum temperature on day i of year j, $i = 1, \ldots, 365$, $j = 1, \ldots, 10$. To simplify the periodic behaviour of the series, we have omitted the two data values observed on 29 February and, for ease of notation, we interpret $Y_{0,j} \equiv Y_{365,j-1}$. The lower case, $y_{i,j}$, shall be used to denote the observed value of $Y_{i,j}$.

Applying the higher dimensional version of (21), we can estimate the density of $Y_{i,j} \mid i, Y_{i-1,j} = x$ by

$$\hat{f}_*(y \mid i, x) = \frac{1}{b} \sum_{j=1}^{10} \sum_{k=1}^{365} w_{k,j}(i,x) K\left(\frac{\|y - y^*_{k,j}(i,x)\|_y}{b} \right) \tag{22}$$

where

$$y^*_{k,j}(i,x) = \hat{r}(i,x) + y_{k,j} - \hat{r}(k, y_{k-1,j}) - \sum_{l=1}^{10} \sum_{m=1}^{365} w_{m,l}(i,x)[y_{m,l} - \hat{r}(m, y_{m-1,l})]$$

and

$$w_{k,j}(i,x) = \frac{K\Big(\frac{\|x - y_{k-1,j}\|_y}{a_1}\Big) K\Big(\frac{\|i-k\|_d}{a_2}\Big)}{\sum_{j=1}^{10} \sum_{k=1}^{365} K\Big(\frac{\|x - y_{k-1,j}\|_y}{a_1}\Big) K\Big(\frac{\|i-k\|_d}{a_2}\Big)}$$

As before, we use the standard Euclidean distance between temperatures, $\|y\|_y = |y|$. However, the distance between days is more difficult as the days at the ends of each year should be close—a fact which is not reflected by taking $\|i-k\|_d = |i-k|$. For example, days 364 and 2 are only three days apart. So we use the metric $\|i-k\|_d = \min(|i-k|, 365-|i-k|)$. The linear surface, $\hat{r}(i,x)$, was specified by a loess surface (Cleveland et al., 1992), modified so that it is periodic in i.

For each value of x we can produce a stacked conditional density plot. Similarly, for each value of i we can produce a stacked conditional density plot. However, it is more revealing to look at HDR plots for different values of x (Figure 5) and different values of i (Figure

6). The lines bounding each shaded region have been removed to reduce visual clutter. Not all densities are shown because of the lack of data for some combinations of x and i. Figures 5 and 6 show the complex interaction between x and i. For example, a quick examination of these graphs reveals the following information.

- In winter (June – August), the conditional mean maximum temperature is almost linear and the distributions are unimodal and symmetric, whereas in the hotter months (December – April) the conditional mean is far from linear and the conditional distributions are bimodal following days over 30°C.
- During those hotter months, the position of the modes varies.
- The temperature distribution following a day of 20°C in June is less skewed, has smaller variance and smaller mean than the temperature distribution following a day of 20°C in January.

These few examples demonstrate the value of conditional density estimation and these graphical displays for data analysis. It is much more difficult to spot these features using other display methods we have seen used.

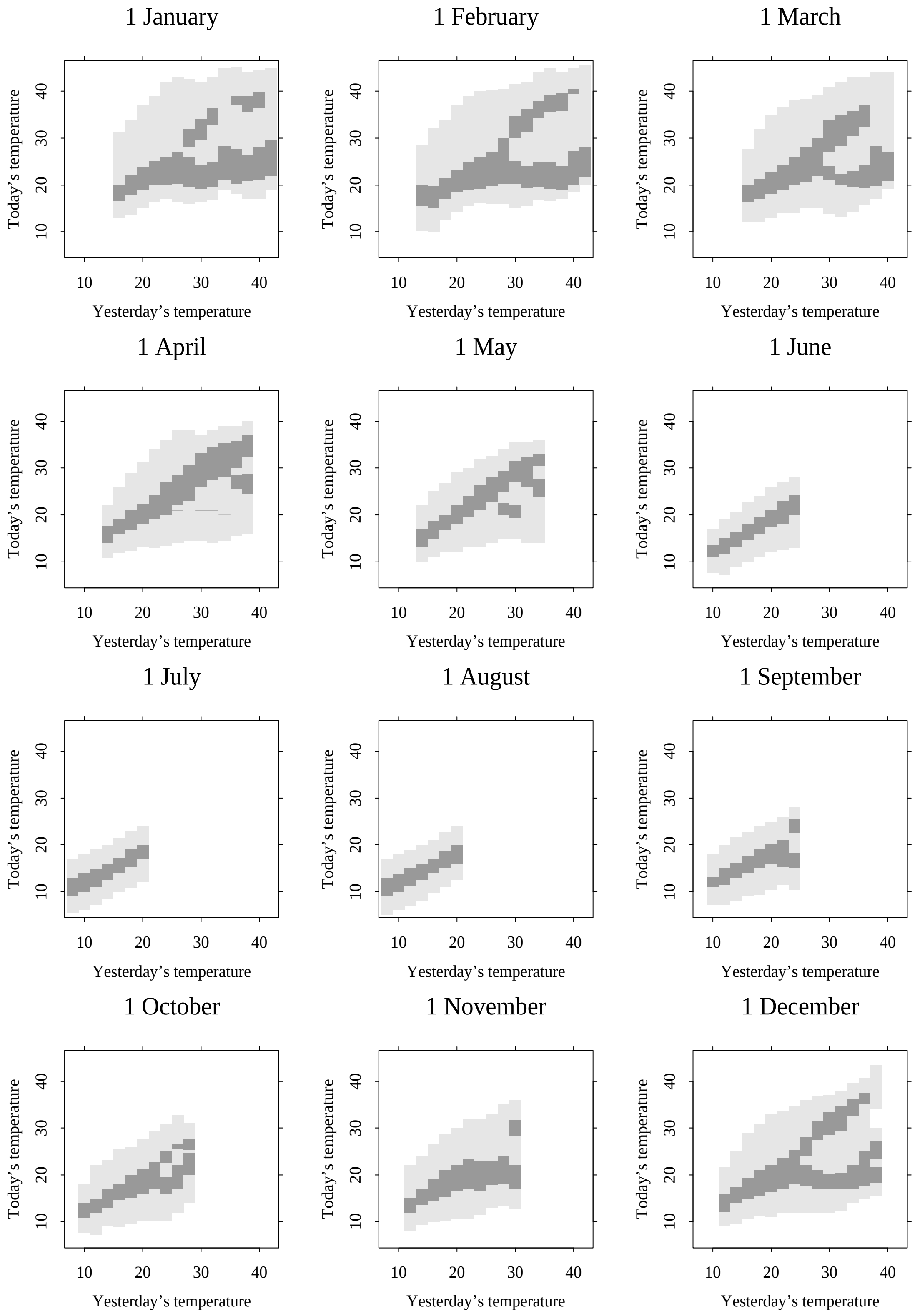

Figure 5: *Highest density regions (50% and 99%) for temperature conditional on the previous day's temperature and the day of the year.*

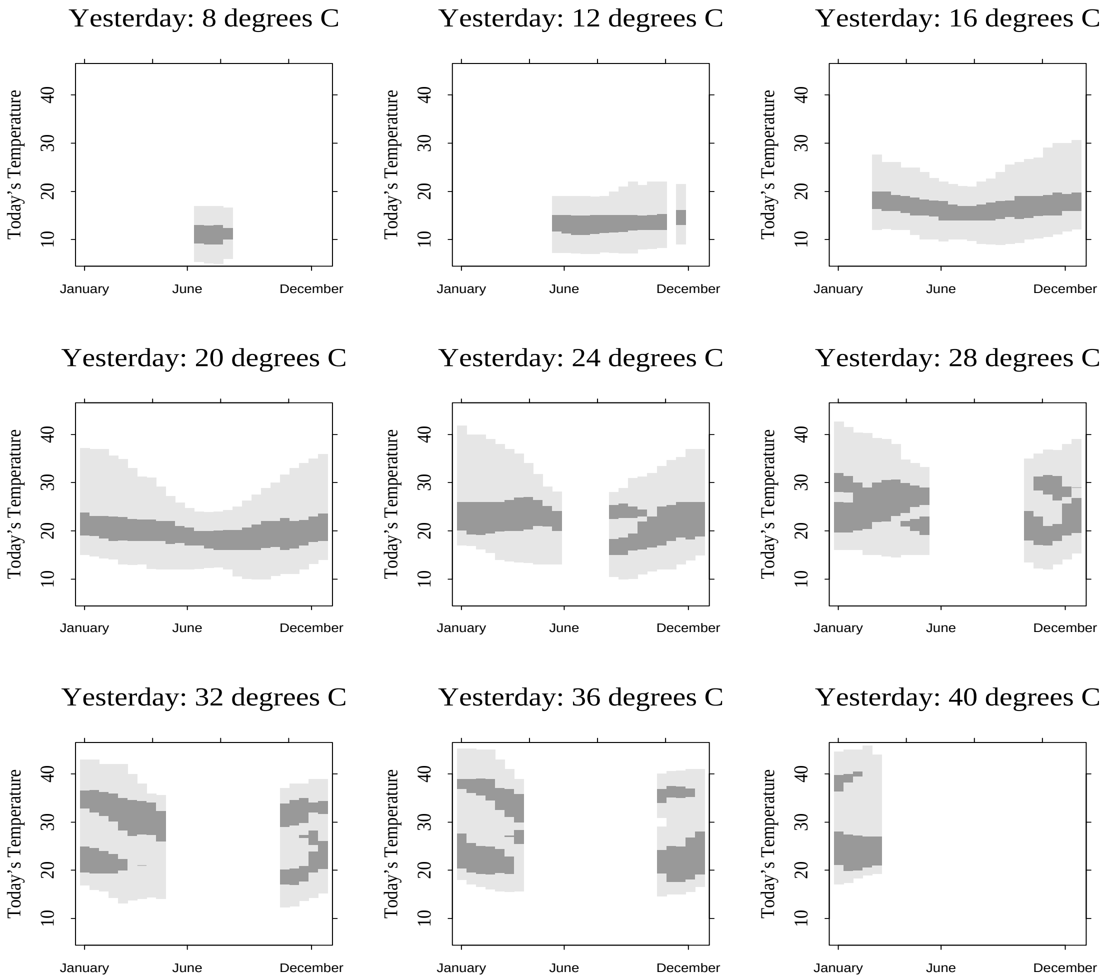

Figure 6: *Highest density regions (50% and 99%) for temperature conditional on the previous day's temperature and the day of the year.*

7 Derivations

Lemma 1 *Let X be a random variable with density $h(x)$ being at least twice continuously differentiable, $K(u)$ be a kernel function satisfying (2), $q(x)$ be at least twice continuously differentiable and defined on the sample space of X, and a be a constant. Then, as $a \to 0$,*

$$\mathsf{E}\left[\frac{1}{a}K\left(\frac{x-X}{a}\right)q(X)\right] = q(x)h(x) + \frac{a^2\sigma_K^2}{2}\frac{d^2}{dx^2}[q(x)h(x)] + O(a^4), \tag{23}$$

$$\mathsf{E}\left[\frac{1}{a^2}K^2\left(\frac{x-X}{a}\right)q(X)\right] = \frac{q(x)h(x)R(K)}{a} + \frac{aG(K)}{2}\frac{d^2}{dx^2}[q(x)h(x)] + O(a^3) \tag{24}$$

and

$$\mathsf{Var}\left[\frac{1}{a}K\left(\frac{x-X}{a}\right)q(X)\right] = q^2(x)h(x)\left(\frac{R(K)}{a} - h(x)\right) + \frac{aG(K)}{2}\frac{d^2}{dx^2}[q^2(x)h(x)] + O(a^2), \tag{25}$$

where

$$\sigma_K^2 = \int w^2K(w)dw, \qquad R(K) = \int K^2(w)dw \qquad \text{and} \qquad G(K) = \int w^2K^2(w)dw.$$

Proof: The first equation is derived as follows.

$$\begin{aligned}
\mathsf{E}\left[\tfrac{1}{a}K\left(\tfrac{x-X}{a}\right)q(X)\right] &= \tfrac{1}{a}\int K\left(\tfrac{x-u}{a}\right)q(u)h(u)du \\
&= \int K(v)[q(x) - vaq'(x) + \tfrac{1}{2}v^2a^2q''(x)][h(x) - vah'(x) + \tfrac{1}{2}v^2a^2h''(x)]dv + O(a^4) \\
&\qquad \text{where } v = (x-u)/a \text{ and using Taylor series expansions about } x \\
&= q(x)h(x) + \tfrac{1}{2}a^2\sigma_K^2[q''(x)h(x) + 2q'(x)h'(x) + q(x)h''(x)] + O(a^4) \\
&= q(x)h(x) + \tfrac{1}{2}a^2\sigma_K^2\tfrac{d^2}{dx^2}[q(x)h(x)] + O(a^4).
\end{aligned}$$

Using a similar argument, we obtain (24). Then (25) follows by noting that

$$\mathsf{Var}\left[\tfrac{1}{a}K\left(\tfrac{x-X}{a}\right)q(X)\right] = \mathsf{E}\left[\tfrac{1}{a^2}K^2\left(\tfrac{x-X}{a}\right)q^2(X)\right] - \left\{\mathsf{E}\left[\tfrac{1}{a}K\left(\tfrac{x-X}{a}\right)q(X)\right]\right\}^2$$

which can be computed from (23) and (24).

□

We shall also use Proposition 31.8 of Port (1994) which is here restated as Lemma 2 for convenience.

Lemma 2 *Let $q_1(X_i)$ and $q_2(X_i)$ be two random variables with means μ_1 and μ_2 and variances v_1 and v_2 respectively and with covariance v_{12}. Let $\{X_1, \dots, X_n\}$ be an iid sequence of random variables and define*

$$\hat{\Sigma}_1 = \frac{1}{n}\sum_{i=1}^{n} q_1(X_i), \qquad \hat{\Sigma}_2 = \frac{1}{n}\sum_{i=1}^{n} q_2(X_i)$$

$$\text{and} \qquad \hat{R} = \hat{\Sigma}_1/\hat{\Sigma}_2.$$

Then the second order approximation of $\mathsf{E}\hat{R}$ *is*

$$\mathsf{E}\hat{R} \;\approx\; \frac{\mu_1}{\mu_2} + \frac{1}{n}\left(\frac{\mu_1 v_2}{\mu_2^3} - \frac{v_{12}}{\mu_2^2}\right) \tag{26}$$

and the first order approximation of $\mathsf{Var}\hat{R}$ *is*

$$\mathsf{Var}\hat{R} \;\approx\; \frac{1}{n\mu_2^2}\left(v_1 + \frac{\mu_1^2 v_2}{\mu_2^2} - 2\frac{\mu_1 v_{12}}{\mu_2}\right). \tag{27}$$

Now the conditional density estimator can be expressed as the ratio of two random variables:

$$\hat{f}(y\,|\,x) \;=\; \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{ab}K\left(\frac{x-X_i}{a}\right)K\left(\frac{y-Y_i}{b}\right)}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{a}K\left(\frac{x-X_i}{a}\right)}. \tag{28}$$

We shall apply Lemma 2 to obtain the bias and variance of $\hat{f}(y\,|\,x)$ given by (6) and (7) respectively.

First note that by applying Lemma 1 we obtain

$$\begin{aligned} \mu_2 = \mathsf{E}\left[\tfrac{1}{a}K\left(\tfrac{x-X_i}{a}\right)\right] &= h(x) + \tfrac{a^2\sigma_K^2}{2}h''(x) + O(a^4) \\ \text{and} \qquad v_2 = \mathsf{Var}\left[\tfrac{1}{a}K\left(\tfrac{x-X_i}{a}\right)\right] &= h(x)\left[\tfrac{R(K)}{a} - h(x)\right] + \tfrac{aG(K)}{2}h''(x) + O(a^2). \end{aligned}$$

Similarly, conditioning on X_i and further applying Lemma 1 gives

$$\mathsf{E}\left[\tfrac{1}{ab}K\left(\tfrac{x-X_i}{a}\right)K\left(\tfrac{y-Y_i}{b}\right)|\,X_i\right] = \tfrac{1}{a}K\left(\tfrac{x-X_i}{a}\right)\left[f(y\,|\,X_i) + \tfrac{b^2\sigma_K^2}{2}\tfrac{\partial^2 f(y\,|\,X_i)}{\partial y^2} + O(b^4)\right] \tag{29}$$

$$\begin{aligned} \mathsf{Var}\left[\tfrac{1}{ab}K\left(\tfrac{x-X_i}{a}\right)K\left(\tfrac{y-Y_i}{b}\right)|\,X_i\right] &= \tfrac{1}{a^2}K^2\left(\tfrac{x-X_i}{a}\right)[f(y\,|\,X_i)\left(\tfrac{R(K)}{b} - f(y\,|\,X_i)\right) \qquad (30) \\ &\qquad + \tfrac{bG(K)}{2}\tfrac{\partial^2 f(y\,|\,X_i)}{\partial y^2} + O(b^2)] \end{aligned}$$

$$\mathsf{E}\left[\tfrac{1}{a^2 b}K^2\left(\tfrac{x-X_i}{a}\right)K\left(\tfrac{y-Y_i}{b}\right)|\,X_i\right] = \tfrac{1}{a^2}K^2\left(\tfrac{x-X_i}{a}\right)\left[f(y\,|\,X_i) + \tfrac{b^2\sigma_K^2}{2}\tfrac{\partial^2 f(y\,|\,X_i)}{\partial y^2} + O(b^4)\right]. \tag{31}$$

Then applying Lemma 1 again to (29) gives the unconditional expectation

$$\begin{aligned} \mu_1 &= \mathsf{E}\left[\tfrac{1}{ab}K\left(\tfrac{x-X_i}{a}\right)K\left(\tfrac{y-Y_i}{b}\right)\right] \\ &= h(x)\left[f(y\,|\,x) + \tfrac{b^2\sigma_K^2}{2}\tfrac{\partial^2 f(y|x)}{\partial y^2}\right] + O(b^4) \\ &\quad + \tfrac{a^2\sigma_K^2}{2}\left[\tfrac{\partial^2 f(y\,|\,x)}{\partial x^2}h(x) + 2\tfrac{\partial f(y\,|\,x)}{\partial x}h'(x) + h''(x)f(y\,|\,x)\right] + O(a^2b^2) + O(a^4). \end{aligned}$$

The variance of (29) and the expectation of (30) can be obtained by further applications of Lemma 1 giving

$$\mathsf{Var}\left\{\mathsf{E}\left[\tfrac{1}{ab}K\left(\tfrac{x-X_i}{a}\right)K\left(\tfrac{y-Y_i}{b}\right)|\,X_i\right]\right\} = f^2(y\,|\,x)h(x)\left(\tfrac{R(K)}{a} - h(x)\right) + O(a) + O(b^2) \tag{32}$$

and

$$\begin{aligned}
&\mathsf{E}\left\{\mathsf{Var}\left[\tfrac{1}{ab}K\left(\tfrac{x-X_i}{a}\right)K\left(\tfrac{y-Y_i}{b}\right)\mid X_i\right]\right\} \\
&\quad = \tfrac{h(x)R(K)}{a}\left[f(y\,|\,x)\left(\tfrac{R(K)}{b}-f(y\,|\,x)\right)+\tfrac{bG(K)}{2}\tfrac{\partial^2 f(y\,|\,x)}{\partial y^2}\right]+O(a)+O(b^2/a). \qquad (33)
\end{aligned}$$

The unconditional variance is the sum of (32) and (33); that is

$$v_1 = \tfrac{h(x)f(y\,|\,x)R^2(K)}{ab} - f^2(y\,|\,x)h^2(x) + \tfrac{bh(x)R(K)G(K)}{2a}\tfrac{\partial^2 f(y\,|\,x)}{\partial y^2} + O(a) + O(b^2/a).$$

Now applying Lemma 1 to (31) gives

$$\mathsf{E}\left[\tfrac{1}{a^2b}K^2\left(\tfrac{x-X_i}{a}\right)K\left(\tfrac{y-Y_i}{b}\right)\right] = \tfrac{h(x)R(K)}{a}\left[f(y\,|\,x)+\tfrac{b^2\sigma_K^2}{2}\tfrac{\partial^2 f(y\,|\,x)}{\partial y^2}\right]+O(b^4/a)+O(a) \quad (34)$$

and subtracting $\mu_1\mu_2$ from (34) we obtain the covariance

$$v_{12} = h(x)f(y\,|\,x)\left[\tfrac{R(K)}{a}-h(x)\right]+O(a)+O(b^2)+O(b^4/a).$$

Now we can apply Lemma 2 to (28) by substituting the above expressions for μ_1, μ_2, v_1, v_2 and v_{12} into (26) and (27). Noting the result $1/(s+\delta) = 1/s - \delta/s^2 + o(\delta)$, we obtain

$$\begin{aligned}
\tfrac{\mu_1}{\mu_2} &= f(y\,|\,x) + \tfrac{b^2\sigma_K^2}{2}\tfrac{\partial^2 f(y\,|\,x)}{\partial y^2} + \tfrac{a^2\sigma_K^2}{2}\left\{2\tfrac{h'(x)}{h(x)}\tfrac{\partial f(y\,|\,x)}{\partial x} + \tfrac{\partial^2 f(y\,|\,x)}{\partial x^2}\right\} \\
&\qquad + O(a^4) + O(b^4) + O(a^2b^2), \\
\tfrac{\mu_1 v_2}{\mu_2^3} &= f(y\,|\,x)\left[\tfrac{R(K)}{ah(x)} - 1\right] + O(a) + O(b^2/a) \\
\text{and} \quad \tfrac{v_{12}}{\mu_2^2} &= \tfrac{f(y\,|\,x)R(K)}{ah(x)} - f(y\,|\,x) + O(a) + O(b^2) + O(b^4/a).
\end{aligned}$$

Hence, from (26), we obtain (6).

Similarly, we obtain

$$\begin{aligned}
\tfrac{v_1}{\mu_2^2} &= \tfrac{f(y\,|\,x)R^2(K)}{abh(x)} - f^2(y\,|\,x) + \tfrac{bR(K)G(K)}{2ah(x)}\tfrac{\partial^2 f(y\,|\,x)}{\partial y^2} - \tfrac{af(y\,|\,x)R^2(K)\sigma_K^2 h''(x)}{2bh^3(x)} \\
&\qquad + O(a) + O(b^2/a) \\
\tfrac{\mu_1^2 v_2}{\mu_2^4} &= \tfrac{f^2(y\,|\,x)R(K)}{ah(x)} - f^2(y\,|\,x) + O(a) + O(b^2) \\
\text{and} \quad \tfrac{\mu_1 v_{12}}{\mu_2^3} &= \tfrac{f^2(y\,|\,x)R(K)}{ah(x)} - f^2(y\,|\,x) + O(a) + O(b^2) + O(b^4/a).
\end{aligned}$$

Hence, from (27), we obtain (7).

Acknowledgments

We wish to thank David Scott (Rice University, Texas) for suggesting that we apply modal regression to the lag plot of the Melbourne temperature data, and the referees for their helpful suggestions which led to several improvements in this paper. Two of the authors were supported in part by Australian Research Council grants.

References

Buja, A., Hastie, T., and Tibshirani, R. (1989), Linear smoothers and additive models (with discussion), *Ann. Statist.*, **17**, 453–555.

Chaudhuri, P. (1991), Nonparametric estimates of regression quantiles and their local bahadur representation, *Ann. Statist.*, **19**, 760–777.

Cleveland, W.S., Grosse, E., and Shyu, W.M. (1992), Local regression models, in *Statistical models in S*, pages 309–376. (Wadsworth & Brooks/Cole: California).

Falk, M. (1993), Asymptotically optimal estimators of general regression functionals, *J. Multivariate Analysis*, **47**, 59–81.

Hall, P. and Carroll, R.J. (1989), Variance function estimation in regression: the effect of estimating the mean, *J. Roy. Statist. Soc. Ser. B*, **51**, 3–14.

Härdle, W., Janssen, P., and Serfling, R (1988), Strong uniform consistency rates for estimators of conditional functionals, *Ann. Statist.*, **16**, 1428–1449.

Härdle, W. (1991), *Smoothing techniques with implementation in S*, (Springer-Verlag: New York).

Hastie, T. and Loader, C. (1993), Local regression: automatic kernel carpentry (with discussion), *Statistical Science*, **8**, 120–143.

Hastie, T. and Tibshirani, R. (1990), *Generalized additive models*, (Chapman and Hall: London).

Hyndman, R.J. (1996), Computing and graphing highest density regions, *Amer. Statist.* To appear, May 1996.

Nadaraya, E.A. (1964), On estimating regression, *Theory Probab. Applic.*, **15**, 134–137.

Port, S.C. (1994), *Theoretical probability for applications*, (Wiley: New York).

Saleh, A.K.Md.E., editor, (1994), Regression quantiles and related topics, *J Nonparametric Statistics*, **3**(3-4), 201–380, special issue.

Scott, D.W. (1992), *Multivariate density estimation: theory, practice, and visualization*, (Wiley: New York).

Silverman, B.W. (1984), A fast and efficient cross-validation method for smoothing parameter choice in spline regression, *J. Amer. Statist. Assoc.*, **79**, 93–99.

Stone, C.J. (1994), The use of polynomial splines and their tensor products in multivariate function estimation, *Ann. Statist.*, **22**, 118–184.

Wand, M.P. and Jones, M.C. (1994), Multivariate plug-in bandwidth selection, *Computational Statistics*, **9**, 97–116.

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

Watson, G.S. (1964), Smooth regression analysis, *Sankhyā A*, **26**, 359–372.

Welsh, A.H., Carroll, R.J., and Ruppert, D. (1994), Fitting heteroscedastic regression models, *J. Amer. Statist. Assoc.*, **89**, 100–116.