# Week 6

## 6.1 Classification problems

An important application of density estimates is in classification problems. Suppose that we observe a random variable $Z$ that is distributed according to one of two densities, $p_0(\cdot)$ or $p_1(\cdot)$. Based on the observed value of $Z$, our goal to determine if it is more likely that $Z$ is distributed according to $p_0(\cdot)$ or according to $p_1(\cdot)$.

For now, we will assume that the density functions $p_0(\cdot)$ and $p_1(\cdot)$ are known; later in this section, we will extend the methodology to the case in which $p_0(\cdot)$ and $p_1(\cdot)$ must be estimated.

One approach to this problem is to use maximum likelihood. Note that we may view $Z$ as being distributed according to the density function $p(\cdot;\theta)$, where $\theta$ is a parameter, taking values in the set $\{0,1\}$; the density $p(\cdot;\theta)$ is given by

$$p(z;\theta) = p_\theta(z).$$

Hence, from this perspective, choosing between the two density functions is simply a parametric estimation problem, albeit a slightly unusual one, since the parameter $\theta$ can take only 2 possible values.

For a given observation $Z = z$, the likelihood function for $\theta$ is given by the usual expression

$$L(\theta) = p(z;\theta), \quad \theta = 0, 1$$

so that, in the present context,

$$L(\theta) = p_\theta(z), \quad \theta = 0, 1.$$

Thus, the relative likelihood of $\theta = 1$ versus $\theta = 0$ is given by

$$\frac{L(1)}{L(0)} = \frac{p_1(z)}{p_0(z)}.$$

It follows that the maximum likelihood estimate of $\theta$ is 1 if $L(1) > L(0)$ and it is 0 if $L(0) < L(1)$; that is, the maximum likelihood estimate of $\theta$ is 1 if

$$p_1(z) > p_0(z)$$

and it is 0 if

$$p_0(z) > p_1(z).$$

If $p_1(z) = p_0(z)$ then both $\theta = 0, 1$ maximize the likelihood function.

Thus, according to the maximum likelihood approach, if $Z = z_0$ is observed, we conclude that $Z$ is distributed according to $p_1(\cdot)$ if

$$p_1(z_0) > p_0(z_0)$$

and we conclude that $Z$ is distributed according to $p_0(\cdot)$ if

$$p_0(z_0) > p_1(z_0).$$

An alternative approach to the classification problem is to use Bayesian reasoning. In this approach, we model $\theta$ as a random variable; suppose that, before $Z$ is observed, we consider the two density functions to be equally likely to have generated $Z$ so that

$$\mathrm{P}(\theta = 0) = \mathrm{P}(\theta = 1) = \frac{1}{2}.$$

Based on the observation $Z = z$, we evaluate the likelihood that $Z$ is distributed according to $p_1(\cdot)$ by calculating $\mathrm{P}(\theta = 1|Z = z)$ using Bayes' Theorem. The only slight complication here is that $\theta$ has a discrete distribution while $Z$ is continuously distributed. However, we can use the same basic approach used in the discrete case by interpreting the probability that $Z = z$ in terms of its density function; this can be justified using a technical argument that won't be given here.

Thus,

$$\mathrm{P}(\theta = 1|Z = z) = \frac{\mathrm{P}(Z = z|\theta = 1)\mathrm{P}(\theta = 1)}{\mathrm{P}(Z = z)}.$$

The probability $\mathrm{P}(Z = z|\theta = 1)$ is given by $p_1(z)$, $\mathrm{P}(\theta = 1) = 1/2$, and using a result analogous to the "law of total probability",

$$\begin{aligned}\mathrm{P}(Z = z) &= \mathrm{P}(Z = z|\theta = 0)\mathrm{P}(\theta = 0) + \mathrm{P}(Z = z|\theta = 1)\mathrm{P}(\theta = 1) \\ &= p_0(z)\frac{1}{2} + p_1(z)\frac{1}{2}\end{aligned}$$

so that

$$\mathrm{P}(\theta = 1|Z = z) = \frac{p_1(z)}{p_0(z) + p_1(z)};$$

it follows that

$$\mathrm{P}(\theta = 0|Z = z) = \frac{p_0(z)}{p_0(z) + p_1(z)}.$$

Note that $\mathrm{P}(\theta = 1|Z = z) > \mathrm{P}(\theta = 0|Z = z)$ provided that $p_1(z) > p_0(z)$ so that the Bayesian approach leads to the same decision rule as the likelihood approach. Furthermore, the relative likelihood of $\theta = 1$ versus $\theta = 0$, $L(1)/L(0)$, is equal to

$$\frac{\mathrm{P}(\theta = 1|Z = z)}{\mathrm{P}(\theta = 0|Z = z)}$$

so that the Bayesian approach may be viewed as an extension of the likelihood approach that produces of a measure of the probabililty that the observation is from each distribution to go along with the choice of the "most likely" distribution.

Of course, these methods require knowledge of the density functions $p_0(\cdot)$ and $p_1(\cdot)$. Now suppose that these density functions are unknown but we have a "training samples" from each distribution. Specifically, let $Y_1, Y_2, \ldots, Y_n$ denote i.i.d. random variables, each distributed according to the distribution with density $p_0(\cdot)$ and let $X_1, X_2, \ldots, X_m$ denote i.i.d. random variables, each distributed according to the distribution with density $p_1(\cdot)$.

For $j = 1, 2$, let $\widehat{p}_j(\cdot)$ denote a kernel estimator of $p_j(\cdot)$. Then, classification of an observation $Z$ may be based on

$$\frac{\widehat{p}_1(Z)}{\widehat{p}_0(Z) + \widehat{p}_1(Z)}$$

which may interpreted either in terms of the Bayesian posterior probability that $Z$ is distributed according to $p_1(\cdot)$ or in terms of the likelihood values $L(0)$ and $L(1)$.

**Example 6.1** A study was conducted on the relationship between certain anthropometric measurements on ancient Etruscans and those on modern Italians. The variable `skull.et` contains the maximum head breadth (in mm) taken from the skulls of 84 Etruscan males and `skull.it` contains similar measurements from 70 modern Italian males. These data are available in the dataset "skulls". Figure 6.1 contains plots of kernel density estimates based on each set of data, with the smoothing parameter chosen by the Sheather-Jones method in each case.

Consider the problem of trying to classify a given skull as either ancient Etruscan or modern Italian based on this measurement. Let $\widehat{p}_0(\cdot)$ denote the kernel estimate of the density of the skull measurements for Italians and let $\widehat{p}_1(\cdot)$ be the kernel estimate for Etruscans. The following R commands may be used to estimate the quantity

$$\frac{\widehat{p}_1(\cdot)}{\widehat{p}_0(\cdot) + \widehat{p}_1(\cdot)}.$$

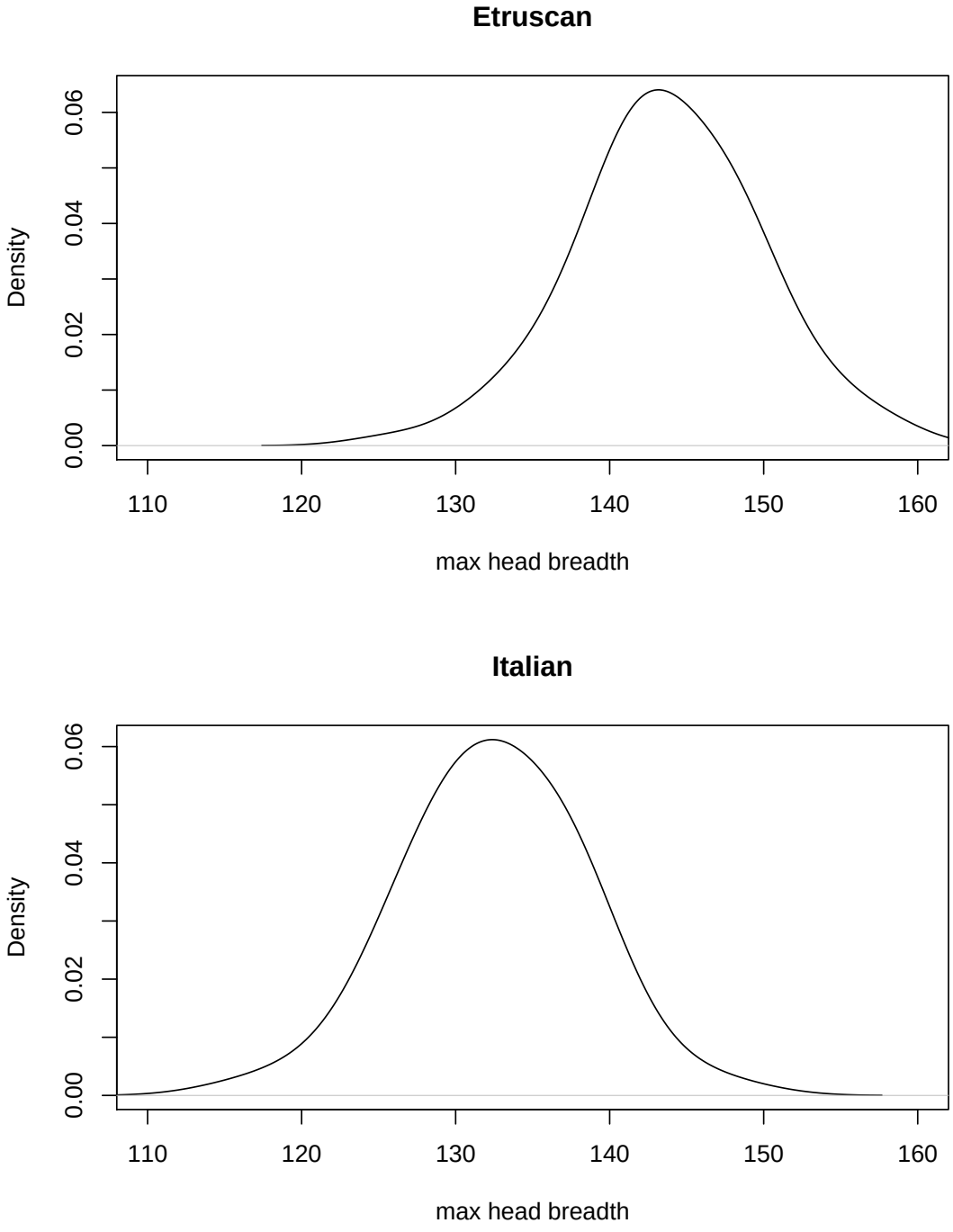


Figure 6.1: Kernel Estimates of the Densities of the Maximum Head Breadth for Etruscan and Italian Skulls

```
> den.et<-density(skull.et, bw=bw.SJ(skull.et), from=110, to=160)
> den.it<-density(skull.it, bw=bw.SJ(skull.it), from=110, to=160)
> prob.et<-den.et$y/(den.et$y + den.it$y)
```

The variables `den.et` and `den.it` contain the results of the function `density` applied to the data in `skull.et` and `skull.it`, respectively. The results of `density` include two components, `$x`, which gives the values at which the density is estimated, and `$y`, which gives the values of the estimate. Thus, choosing the same `from` and `to` values in the function `density` ensures that the same `$x` values will be used in each of `den.it` and `den.et`. It follows that `prob.et` gives the values of

$$\frac{\widehat{p}_1(\cdot)}{\widehat{p}_0(\cdot) + \widehat{p}_1(\cdot)}$$

for each of the $x$-values specified in `den.it$x` (and in `den.et$x`).

Figure 6.2 gives a plot of the probability function

$$\frac{\widehat{p}_1(\cdot)}{\widehat{p}_0(\cdot) + \widehat{p}_1(\cdot)}$$

which, in the present context, gives an estimate of the probability that a given observation is an Etruscan skull, based on the value of its maximum head breadth, using the Bayesian approach discussed above.

To find the value of this probability function corresponding to a given maximum head breadth value, we can use the `approx` function, which performs linear interpolation. E.g., to find the estimated probability that a skull with maximum head breadth 145 is an Etruscan skull, we can use the command

```
> approx(den.et$x, prob.et, xout=145)
$x
[1] 145

$y
[1] 0.903
```

Thus, the estimated probability is 0.903. □

## 6.2 Multivariate kernel density estimators

The kernel method of estimating the density of a real-valued random variable (i.e., a one-dimensional random variable) $Y$ can be adapted to estimate the density of a $p$-dimensional random vector $(Y_1, Y_2, \ldots, Y_p)$. Here we consider the method for the case $p = 2$; the same approach can be used for larger values of $p$ as well.

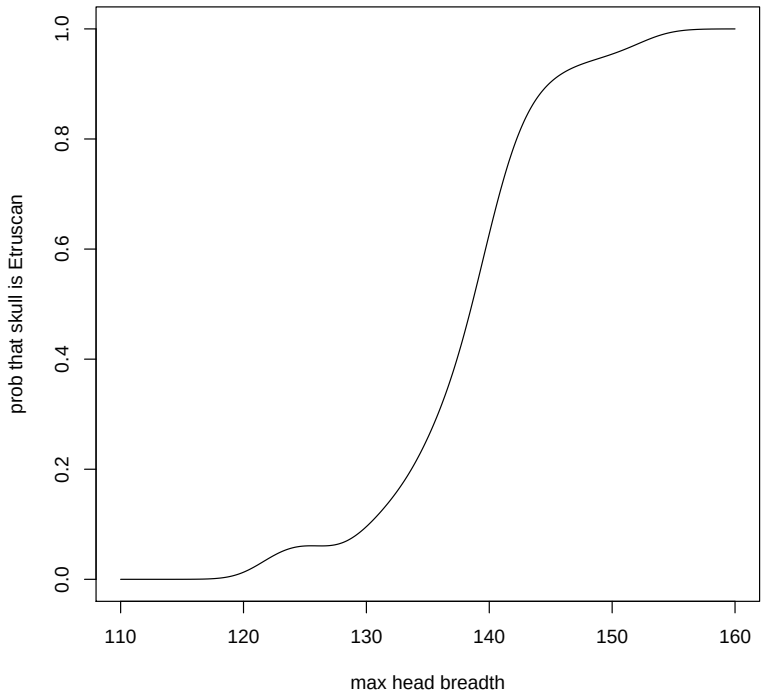


Figure 6.2: Kernel Estimate of the Probability that a Given Skull is Etruscan as a Function of Its Maximum Head Breadth

Hence, the data consist of pairs

$$(Y_{11}, Y_{21}),\ (Y_{12}, Y_{22}),\ \ldots,\ (Y_{1n}, Y_{2n}),$$

which are assumed to be independent, identically distributed random vectors with a continuous distribution with density $p(y_1, y_2)$. Note that, although $(Y_{1i}, Y_{2i})$ and $(Y_{1j}, Y_{2j})$ are assumed to be independent for $i \neq j$, $Y_{1i}$ and $Y_{2i}$ are not independent, in general.

Let $K(\cdot)$ denote a kernel function. Then a two-dimensional kernel estimator of $p$ is given by

$$\widehat{p}(y_1, y_2) = \frac{1}{nh_1h_2} \sum_{j=1}^{n} K(\frac{y_1 - Y_{1j}}{h_1}) K(\frac{y_2 - Y_{2j}}{h_2}), \quad -\infty < y_1 < \infty; \quad -\infty < y_2 < \infty.$$

Thus, the kernel used here is of the form

$$K(u_1)K(u_2),$$

known as a *product kernel*; the smoothing parameter $(h_1, h_2)$ is two-dimensional. As in the univariate kernel density estimation, we assume that $K(\cdot)$ is a symmetric density function satisfying

$$\int_{-\infty}^{\infty} uK(u)du = 0 \quad \text{and} \quad \int_{-\infty}^{\infty} u^2K(u)du = 1.$$

Note that $\widehat{p}(\cdot,\cdot)$ is a genuine density function, in the sense that it is non-negative (because $K(\cdot)$ is non-negative) and it integrates to 1:

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\widehat{p}(y_1,y_2)dy_1dy_2 &= \frac{1}{nh_1h_2}\sum_{j=1}^{n}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K(\frac{y_1-Y_{1j}}{h_1})K(\frac{y_2-Y_{2j}}{h_2})dy_1dy_2\\ &=\frac{1}{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K(u_1)K(u_2)du_1du_2\\ &\qquad \text{where}\ \ u_1=(y_1-Y_{1j})/h_1\ \ \text{and}\ \ u_2=(y_2-Y_{2j})/h_2\\ &=\frac{1}{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty}K(u_1)du_1\ \int_{-\infty}^{\infty}K(u_2)du_2\\ &=\frac{1}{n}\sum_{j=1}^{n}1=n.\end{aligned}$$

In practice, the values of $h_1, h_2$ to use in forming a kernel estimate can be chosen using the same approaches used in univariate kernel estimation:

- cross-validation
- choosing the optimal values for a reference distribution, typically the bivariate normal
- using a two-stage method in which a preliminary estimate of $p$ is used to estimate the optimal values of $h_1, h_2$.

The methodology is illustrated in the following example.

**Example 6.2** The dataset "tests", and the R data frame `tests`, contain data on two tests given to a number of preschool children. The first column (`lets`) is the score from a test of the child's knowledge of letters; the second column (`vocab`) is the results of the Peabody Picture Vocabulary Test.

```
> head(tests)
     lets vocab
[1,]   30    62
[2,]   37    80
[3,]   46    32
[4,]   14    27
[5,]   63    71
[6,]   36    32
```

To compute a bivariate kernel density estimate for the density of these two variables, we can use the function `sm.density` in the R package "sm". The arguments to `sm` include `x`, the matrix or data frame with the two variables (this is the first argument and, hence, `x=` can be omitted), `method`, the method used to determine the values of the smoothing parameters – "cv" for cross-validation or "normal" (the default) for the normal-distribution reference method, and `display`, which indicates how the density estimate is to be displayed (the default is a perspective plot).

For instance, the commands

```
> library(sm)
> sm.density(tests)
```

produces the perspective plot given in Figure 6.3.

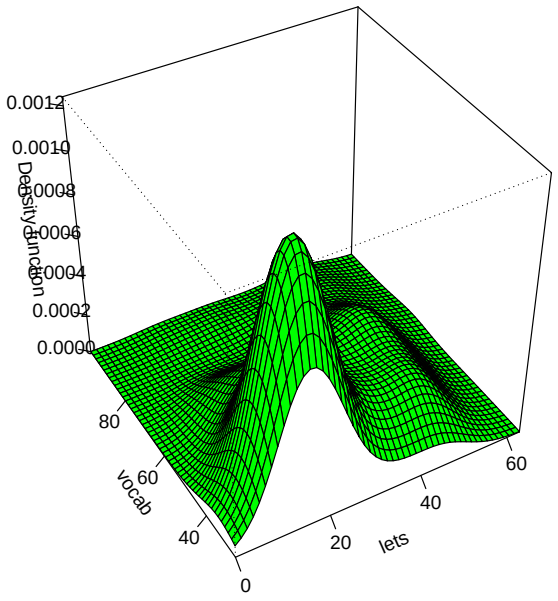


Figure 6.3: Perspective Plot Bivariate Kernel Estimate of the Density of the Test Variables using the Default Region

Note that, using the default values for ranges of the variables, the $xz$-plane in the plot slices off part of the relevant information. In cases like this, the plot can be improved by specifying the ranges for the variables; for example,

```
> sm.density(tests, ylim=c(10, 100), xlim=c(0, 60))
```

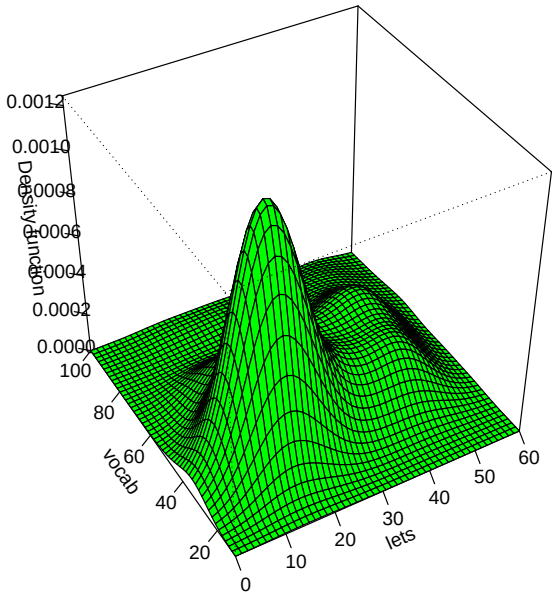


Figure 6.4: Alternative Perspective Plot of the Bivariate Kernel Estimate of the Density of the Test Variables

produces the plot in Figure 6.4.

Although the default density plot produced by `sm.density` is a perspective plot, in many cases, a contour plot is more useful. The command

```
> out<-sm.density(tests, ylim=c(10, 100), xlim=c(0, 60), display="contour",
+ props=seq(10, 90, 20))
```

produces the contour plot given in Figure 6.5. Here the argument `props` specifies how the contours are chosen, specifically it specifies the proportions of the data to be included within each contour. Hence, in the present example, contours are constructed that include $10\%, 30\%, 50\%, 70\%$ and $90\%$ of the data.

Saving the result of `sm.density` to a variable provides useful information regarding the kernel estimate and the estimation procedure. For instance, the component `eval.points` gives a matrix, the columns of which contain vectors of $x$ (column 1) and $y$ (column 2) values. These vectors define a grid for which density estimates are computed; the component `estimate` contains a matrix of density estimates for that grid. For instance, suppose that output from `sm.density` is saved in the variable `out`:

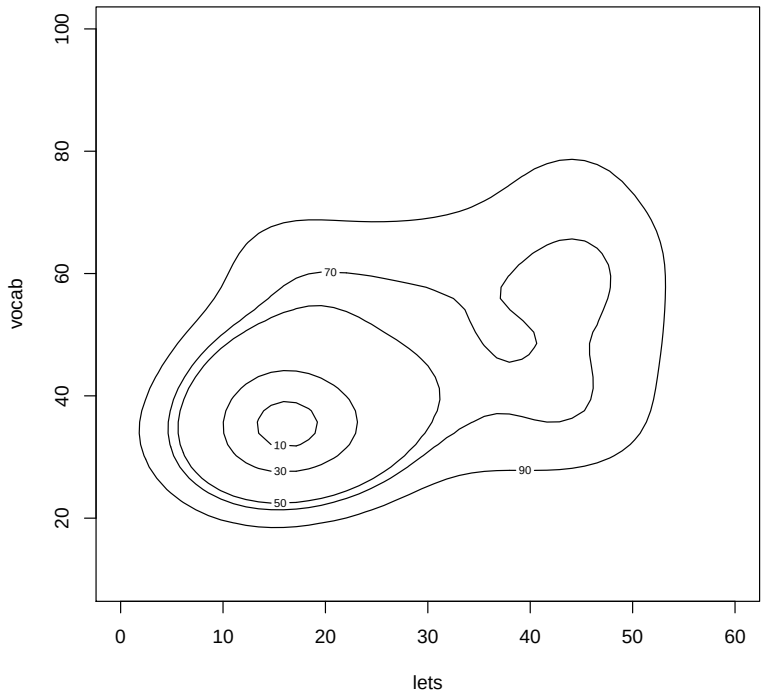


Figure 6.5: Contour Plot of the Bivariate Kernel Estimate of the Density of the Test Variables

```
> out<-sm.density(tests, ylim=c(10, 100), xlim=c(0, 60), display="contour",
+ props=seq(10, 90, 20))
> head(out$eval.points)
     xgrid ygrid
[1,] 0.000 10.00
[2,] 1.224 11.84
[3,] 2.449 13.67
[4,] 3.673 15.51
[5,] 4.898 17.35
[6,] 6.122 19.18
```

For instance, the density estimate corresponding to `lets = 2.449` and `vocab=19.18` is element $(3, 6)$ of the matrix `out$estimate`:

```
> out$estimate[3, 6]
[1] 3.062e-05
```

The default is for the matrix given in `estimate` to be $50 \times 50$; this can be changed by including the argument `ngrid`. Also, the argument `eval.points` can be used to specify

the evaluation points to be used; its format is the same as that used in the component `eval.points` of the output from `sm.density`.

The component `h` of the output from `sm.density` gives the values used for the smoothing parameters:

```
> out$h
 lets vocab
5.365 6.414
```

Here the default method of choosing $h_1, h_2$, which uses a normal reference distribution, was used. Note that the ratio of the smoothing parameters, 1.195 matches the ratio of the variable standard deviations,

```
> sd(tests$vocab)/sd(tests$lets)
[1] 1.195
```

Alternatively, cross-validation can be used to choose the smoothing parameters, by including the argument `method="cv"` in the function. Here the cross-validation values of $h_1$ and $h_2$ are 4.054 and 4.846, respectively; see Figure 6.6 for a contour plot of the resulting estimate. As with univariate kernel estimation, larger values of $h_1, h_2$ yield smooth density estimates and smaller values yield estimates with more local variation; see Figure 6.7. Note that, in that figure, the densities shown in the top two plots are multimodal, while those shown in the bottom two plots are unimodal.

□

One difficulty with using bivariate density estimates is that we do not have much experience with plots of bivariate densities; furthermore, there are few "standard" bivariate densities to use as references.

An exception is the bivariate normal density. Figure 6.8 contains contour plots of bivariate normal densities in which both variables have mean 0 and standard deviation 1, for different values of the correlation. Note that the contours are elliptical, with the orientation of the ellipses indicating the correlation between the variables.

Nonzero means for the variables simply changes the center of the ellipses. If the standard deviations are equal, but not 1, the only the scale of the plot changes. However, if the standard deviations are not equal, then the contours extend farther in one direction than in the other. See Figure 6.9 for examples.

It is important to note that not all bivariate densities with elliptical contours are bivariate normal, just like not all symmetric univariate densities with smooth, decreasing tails are normal. However, elliptical contours suggest a distribution that has a shape similar to that of the bivariate normal distribution, roughly speaking.

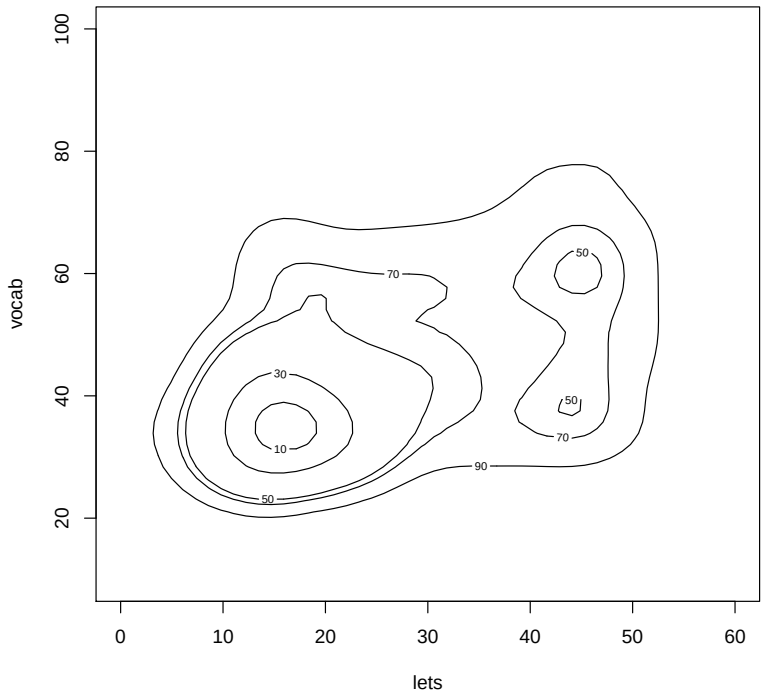


Figure 6.6: Contour Plot of the Bivariate Kernel Estimate with the Smoothing Parameters Chosen by Cross-Validation

### Accuracy of a bivariate kernel estimate

The properties of $\widehat{p}$ can be derived using same basic approach we used in the univariate case. For instance,

$$\begin{aligned}\mathrm{E}\left(\widehat{p}(y_1, y_2)\right) &= \frac{1}{h_1 h_2}\mathrm{E}\left(K(\frac{y_1 - Y_1}{h_1})K(\frac{y_2 - Y_2}{h_2})\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{h_1 h_2}K(\frac{y_1 - t_1}{h_1})K(\frac{y_2 - t_2}{h_2})p(t_1, t_2)dt_1 dt_2.\end{aligned}$$

As in the univariate case, we approximate these integrals by using the change-of-variable

$$u_1 = \frac{y_1 - t_1}{h_1} \quad \text{and} \quad u_2 = \frac{y_2 - t_2}{h_2},$$

so that

$$\mathrm{E}\left(\widehat{p}(y_1, y_2)\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K(u_1)K(u_2)p(y_1 - u_1 h_1, y_2 - u_2 h_2)du_1 du_2$$

and then using a two-dimensional Taylor's series expansion to approximate

$$p(y_1 - u_1 h_1, y_2 - u_2 h_2).$$

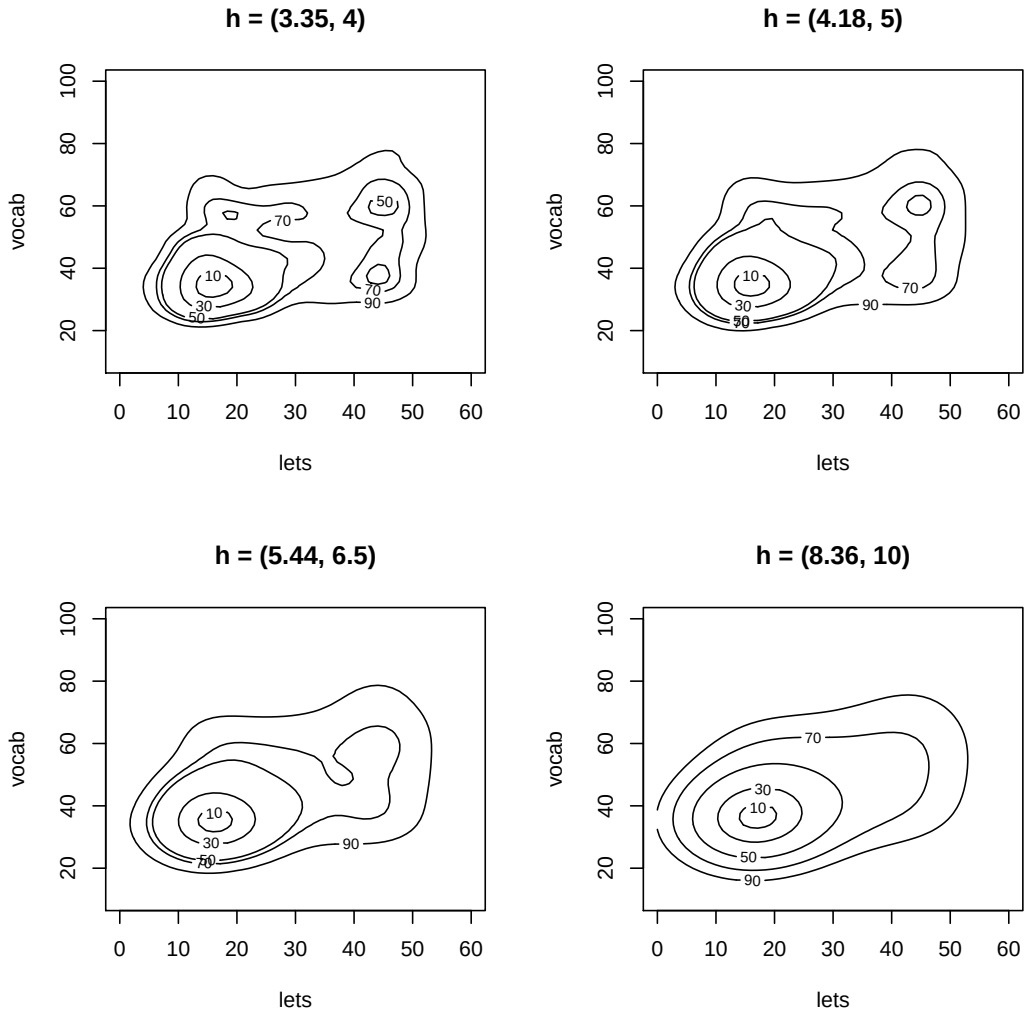


Figure 6.7: Contour Plots of the Kernel Estimates for Different Values of the Smoothing Parameter

The result is of the form

$$\mathrm{E}\left(\widehat{p}(y_1, y_2)\right) = p(y_1, y_2) + \frac{1}{2} d^2 p(y_1, y_2; h_1, h_2) + O(h_1^4) + O(h_2^4)$$

where $d^2 p(y_1, y_2; \delta_1, \delta_2)$ is the two-dimensional versions of $p''(y)\delta^2$. Specifically,

$$d^2 p(y_1, y_2; \delta_1, \delta_2) = \frac{\partial^2}{\partial y_1^2} p(y_1, y_2)\delta_1^2 + 2\frac{\partial^2}{\partial y_1 y_2} p(y_1, y_2)\delta_1\delta_2 + \frac{\partial^2}{\partial y_2^2} p(y_1, y_2)\delta_2^2.$$

It follows that the bias of the bivariate kernel estimator $\widehat{p}(y_1, y_2)$ can be expanded

$$\frac{1}{2} d^2 p(y_1, y_2; h_1, h_2) + O(h_1^4) + O(h_2^4).$$

Note, as in the case for the bias of the univariate kernel estimator, the leading term in this expansion involves second-degree powers of the smoothing parameter. The main difference

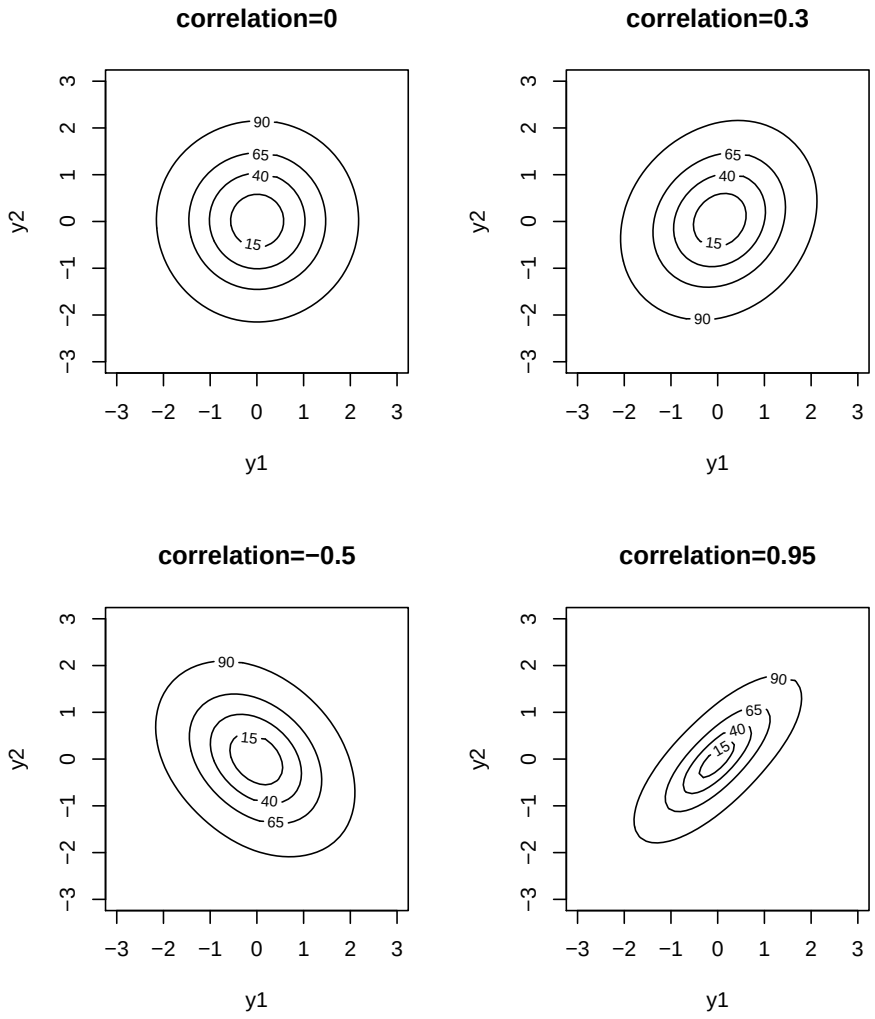


Figure 6.8: Bivariate Normal Densities for Random Variables with Mean 0 and SD 1

is that, in the univariate case, that term involves $h^2$, while in the bivariate case it is the sum of three terms, with factors $h_1^2$, $h_1 h_2$, and $h_2^2$.

The expansion for the variance of $\widehat{p}(y_1, y_2)$ also follows the argument used in the univariate case. An exact expression for the variance is given by

$$\begin{aligned}\text{Var}\left(\widehat{p}(y_1,y_2)\right) &= \frac{1}{nh_1^2h_2^2}\text{Var}\left(K(\frac{y_1-Y_1}{h_1})K(\frac{y_2-Y_2}{h_2})\right)\\ &= \frac{1}{nh_1h_2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{h_1}K(\frac{y_1-t_1}{h_1})^2\frac{1}{h_2}K(\frac{y_2-t_2}{h_2})^2 p(t_1,t_2)dt_1dt_2\\ &\quad -\frac{1}{n}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{h_1}K(\frac{y_1-t_1}{h_1})\frac{1}{h_2}K(\frac{y_2-t_2}{h_2})p(t_1,t_2)dt_1dt_2\right)^2.\end{aligned}$$

Using the usual change-of-variable for these integrals and expanding as we did for the ex-

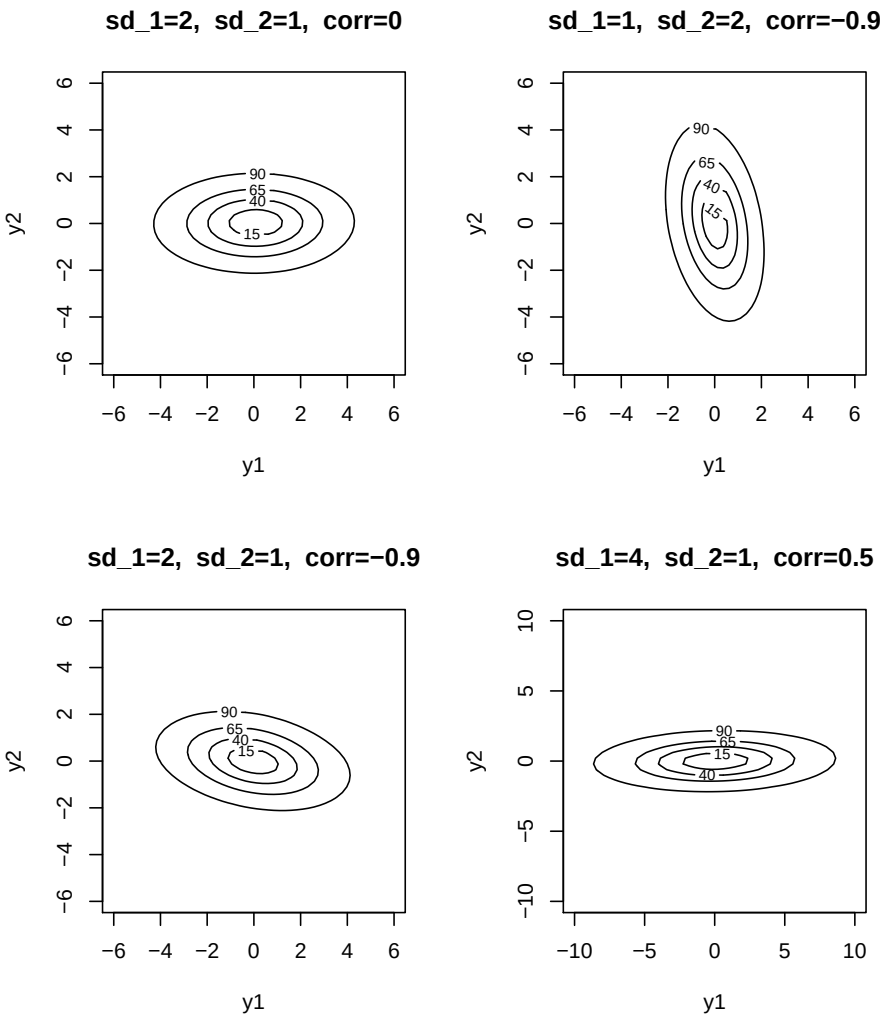


Figure 6.9: Bivariate Normal Densities

pected value yields the result

$$\text{Var}\,(\widehat{p}(y_1, y_2)) = \frac{1}{nh_1h_2} K_2^2\, p(y_1, y_2) + O(\frac{1}{n})$$

where

$$K_2 = \int_{-\infty}^{\infty} K(u)^2 du.$$

Recall that the variance of the univariate kernel estimator is of the form

$$\frac{1}{nh} K_2\, p(y) + O(\frac{1}{n}).$$

Thus, there is an important difference between the expansions for the variance of the bivariate and univariate kernel estimators: in the bivariate case, the leading term in the expansion is of order $O(1/(nh_1h_2))$, with a second-degree power in $(h_1, h_2)$, while in the univariate case, the leading term in the expansion includes only the first-degree power in $h$.

This is relevant because we know that the smoothing parameters must be small, in order for the bias of the estimators to be small. This suggests that the variance of the bivariate kernel estimator tends to be larger (relative to the sample size $n$) than the variance of the univariate kernel estimator.

These approximations to the bias and variance can be combined in the usual way to form the mean squared error and then integrated to give the following expression for the AIMSE:

$$\frac{1}{4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(d^2p(y_1,y_2;h_1,h_2)\right)^2 dy_1dy_2 + \frac{1}{nh_1h_2}K_2^2. \tag{6.1}$$

Consider minimizing (6.1) to find the optimal choices for $h_1$ and $h_2$. Although the actual minimizing values are complicated functions of the density $p(y_1, y_2)$ and its derivatives, it is not difficult to determine some basic properties of these optimal values.

Note that $d^2p(y_1, y_1; h_1, h_2)$ is a second-degree polynomial in $(h_1, h_2)$ so that squaring this expression and integrating yields a fourth-degree polynomial in $(h_1, h_2)$. Therefore, the equation obtained by taking the derivatives of (6.1) with respect to $h_1, h_2$ and setting the result equal to 0 has the form of sixth-degree polynomial in $(h_1, h_2)$ equal to a term that is of order $O(1/n)$. It follows that the optimal values of $h_1$ and $h_2$ are both of order $O(n^{-\frac{1}{6}})$.

Recall that for univariate kernel density estimators, the optimal choice for the smoothing parameter $h$ is of order $O(n^{-\frac{1}{5}})$. Thus, we tend to use less smoothing in bivariate density estimation than in univariate density estimation. This is a consequence of the higher variance of the bivariate kernel estimator.

Furthermore, using the fact that $h_1$ and $h_2$ are of order $O(n^{-\frac{1}{6}})$ in (6.1) (a fourth-degree polynomial in $(h_1, h_2)$) shows that the AIMSE corresponding to the optimal choices of $h_1, h_2$ approaches 0 at the rate $O(n^{-\frac{2}{3}})$.

This can be compared to the univariate case, in which the AIMSE corresponding to the optimal choice of $h$ approaches 0 at the rate $O(n^{-\frac{4}{5}})$. Thus, bivariate density estimation is a "more difficult" estimation problem than is univariate density estimation.

In general, for estimating the density of a $p$-dimensional random vector, the AIMSE corresponding to the optimal choices of the smoothing parameter converges to 0 at the rate

$$O(\frac{1}{n^{\frac{4}{p+4}}}).$$

This is an example of the *curse of dimensionality* – in many respects, high-dimensional problems are inherently more difficult than the corresponding low-dimensional problem. In this case, the rate at which the AIMSE approaches 0 decreases as the dimension of the random variable increases.

**Application to classification problems**

In Section 6.1, the use of (univariate) density estimators in classification problems was discussed. Here we consider the use of bivariate density estimators in classification.

The same basic argument used in the univariate case applies here. Let $Z = (Z_1, Z_2)$ denote a bivariate random vector that is distributed according to one of two densities, $p_0(\cdot)$ and $p_1(\cdot)$.

Based on the observed value of $Z$, our goal to determine if it is more likely that $Z$ is distributed according to $p_0(\cdot)$ or according to $p_1(\cdot)$. We may view $Z$ as being distributed according to the density function $p(\cdot;\theta) = p_\theta(\cdot)$, where $\theta$ is a parameter, taking values in the set $(0, 1)$. Thus, $p(z; 1) = p_1(z)$ can be viewed as the conditional density function of $Z$ given that $\theta = 1$; $p(z; 0)$ has a similar interpretation.

Suppose that, before $Z$ is observed, we consider the two density functions to be equally likely to have generated $Z$ so that

$$\mathrm{P}(\theta = 0) = \mathrm{P}(\theta = 1) = \frac{1}{2}.$$

Then

$$\mathrm{P}(\theta = 1|Z = z) = \frac{p_1(z)}{p_0(z) + p_1(z)}$$

and

$$\mathrm{P}(\theta = 0|Z = z) = \frac{p_0(z)}{p_0(z) + p_1(z)}.$$

To implement this approach, we estimate $p_0$ and $p_1$ using samples from each distribution. Thus, for $j = 1, 2$, let $\widehat{p}_j(\cdot)$ denote a kernel estimator of $p_j(\cdot)$. Then, classification of an observation $Z$ may be based on

$$\frac{\widehat{p}_1(Z)}{\widehat{p}_0(Z) + \widehat{p}_1(Z)}.$$

**Example 6.3** To regulate the salmon industry, it is important to identify fish caught in the ocean as being from either Alaskan or Canadian waters. It has been found that the growth rings on the scales of a fish are useful for identifying its source. The dataset "salmon" contains two variables describing these growth rings, for samples of fish whose origin is known: `fresh`, the diameter of growth rings for the first-year of freshwater growth and `marine`, the diameter of growth rings for the first-year of marine growth, both measurements are in hundredths of an inch.

The data frame `salmonA` contains the data on the Alaskan salmon and the data frame `salmonC` contains data on the Canadian salmon. The results of bivariate kernel density estimation, choosing the smoothing parameters by cross-validation, are stored in the variables `denA` and `denC`:

```
> denA<-sm.density(salmonA, method="cv",  xlim=c(53, 179), ylim=c(301, 511),
+       display="contour", xlab="fresh", ylab="marine")
> denC<-sm.density(salmonC, method="cv", xlim=c(53, 179), ylim=c(301, 511),
+        display="contour", xlab="fresh", ylab="marine")
```

As with classification based on univariate kernel estimates, the estimates need to be given for the same values of the variables; hence we use `xlim` and `ylim` to specify the range for each variable.

The function `sm.density` has a useful feature for including two (or more) sets of density contours on the same plot by adding the argument `add=T` to the second use of `sm.density`. For instance,

```
> denA<-sm.density(salmonA, method="cv",  xlim=c(53, 179), ylim=c(301, 511),
+       display="contour", xlab="fresh", ylab="marine")
> denC<-sm.density(salmonC, method="cv", xlim=c(53, 179), ylim=c(301, 511),
+        display="contour", xlab="fresh", ylab="marine", add=T, lty=5)
```

produces the plot in Figure 6.10; the argument `lty=5` displays the second set of contours as dashed lines. Note that the estimates are fairly different; hence, we expect that the growth ring measurements should be useful in determining the origin of a fish.

Let $\widehat{p}_A$ denote the density estimate for the Alaskan salmon and let $\widehat{p}_B$ denote the density estimate for the Canadian salmon. Then values of the function

$$\frac{\widehat{p}_A(\cdot)}{\widehat{p}_A(\cdot) + \widehat{p}_B(\cdot)}$$

can be calculated using

```
probA<-denA$estimate/(denA$estimate + denC$estimate)
```

To plot the result, we can use the function `contour`, which produces a contour plot. The main arguments are `x` and `y`, the vectors of values for the variables on the $x$-axis and $y$-axis, respectively, and `z`, a matrix of values to be represented by the contours. The values in `z` correspond to a grid formed from `x` and `y`, with the rows of `z` corresponding to the values in `x` and the columns of `y` corresponding to the values in `y`. There are a number of other arguments to `contour` that can be used to customize the plot.

The command

```
> contour(x=denA$eval.points[,1], y=denA$eval.points[,2], z=probA,
+      ylab="marine", xlab="fresh")
```

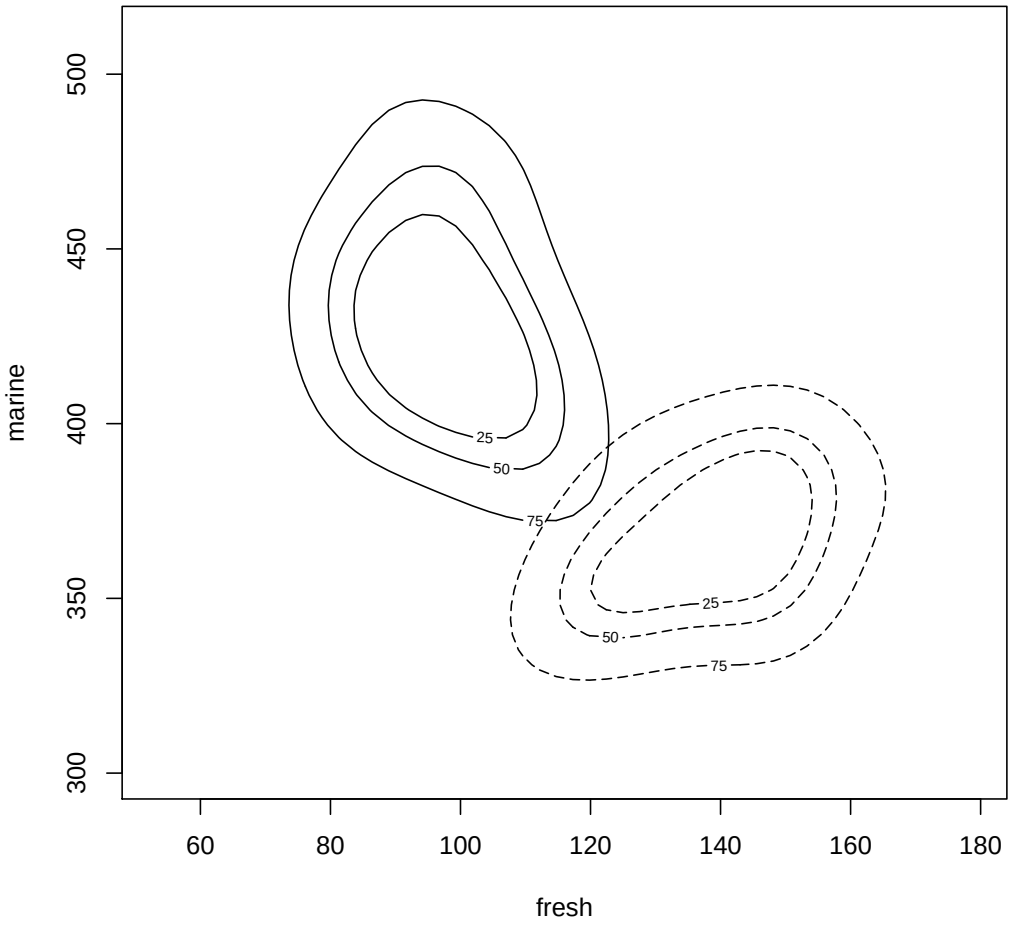


Figure 6.10: Density Estimates for Samples of Alaskan (solid line) and Canadian Salmon

produces the plot in Figure 6.11.

Estimates of the probability that a fish is Alaskan can be obtained for specific values of `fresh` and `marine` using (two-dimensional) linear interpolation. In R, this can be performed using the function `interp2` in the package "pracma". For instance, suppose we would like to estimate the probability that a given fish is Alaskan when its value of `fresh` is 130 and its value of `marine` is 390. The estimate can be obtained using the commands

```
> library(pracma)
> interp2(x=denA$eval.points[,1], y=denA$eval.points[,2], Z=t(probA),
+             xp=130, yp=390,  method="linear")
[1] 0.2627
```

The arguments of `interp2` are similar to those of `contour`: `x` and `y` describe a grid and `Z` is a matrix of function values for that grid. The arguments `xp` and `yp` (which may be

vectors) specify the values for which the value of function is desired. One important difference between `interp2` and `contour` is that, in `interp2`, the rows in the matrix `Z` correspond to `y` and the columns correspond to `x`; hence, in the commands given above, `Z` is taken to be the transpose of the matrix `probA`. □

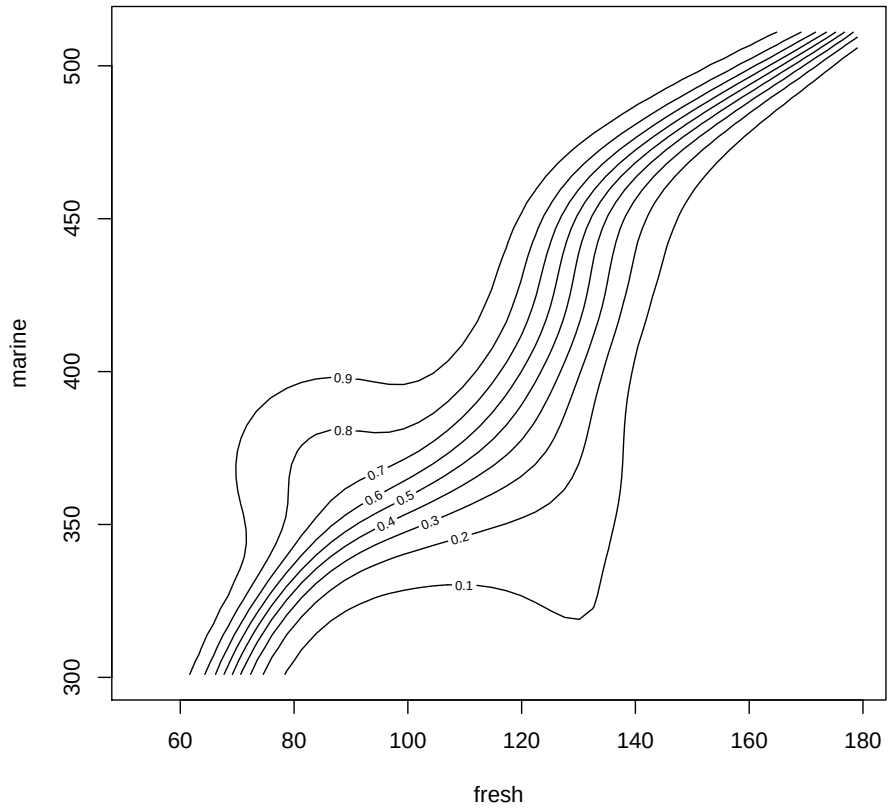


Figure 6.11: Contour Plot of the Estimated Probability that a Salmon is Alaskan

## 6.3 Exercises

**6.1.** In Example 5.8 in the Week 5 notes data are on the level of plasma triglycerides for patients with and without heart disease were analyzed and it was shown that there is evidence that the distribution of plasma triglycerides is different for these two groups. These data are in the dataset "blood".

(a) Using the method described in Section 6.1, estimate the probability that a patient has heart disease as a function of the patient's plasma triglycerides level; present the result as a plot of the estimated probability function. Choose the smoothing parameters needed using the Sheather-Jones method. Restrict the analysis to triglyceride levels in the range 0 to 300; to do this, include the arguments `from=0` and `to=300` in the function `density`.

(b) Using the result given in part (a), estimate the probability that a patient has heart disease given a triglyceride level of $y$, for $y = 100, 200, 250$.

(c) It is sometimes to desirable to use additional smoothing to obtain a more regular estimate of the probability function. Hence, repeat parts (a) and (b), including the argument `adjust = 3` in the function `density`. This uses the value of the smoothing parameter given by 3 times the Sheather-Jones value.

**6.2.** Suppose we observe a random variable $Z$ that is distributed according to one of two densities, $p_0(\cdot)$ and $p_1(\cdot)$, and on the basis of the observed value of $Z$, we wish to determine the probability that $Z$ is distributed according to $p_0(\cdot)$.
We have seen that, using Bayesian reasoning, the probability that $Z$ is distributed according to $p_0(\cdot)$ based on an observation $Z = z$ is given by

$$\frac{p_0(z)}{p_0(z) + p_1(z)}.$$

This result is based on the assumption that the two distributions are considered to be equally likely; that is, it is based on

$$\Pr(\theta = 0) = \Pr(\theta = 1) = \frac{1}{2},$$

where $\theta$ indicates the density that applies to $Z$ (that is, $Z$ is distributed according to $p_\theta(\cdot)$).
Now suppose that the two distributions are not considered to be equally likely and let

$$\alpha = \Pr(\theta = 0)$$

where $0 < \alpha < 1$.
Find an expressions for $\Pr(\theta = 0|Z = z)$ and

$$\frac{\Pr(\theta = 0|Z = z)}{\Pr(\theta = 1|Z = z)}.$$

**6.3.** The dataset "couples" contains the heights of married couples.

(a) Plot two bivariate kernel estimates of the data, one using the normal reference method of choosing the smoothing parameter and one using cross-validation. Plot each density twice, once as a perspective plot and once as a contour plot.

(b) Comment on the relationship between the estimates based on the different smoothing parameters and on the relative usefulness of perspective and contour plots for displaying the estimates.

(c) Based on the results, does the distribution appear to be approximately bivariate normal? Are there any interesting features?

**6.4.** The dataset "geyser" contains data on the time between eruptions, along with the length of the eruptions, of the Old Faithful geyser at Yellowstone National Park, over the period from August 1 to August 15, 1985.
Repeat Exercise 6.3 using the geyser data.

**6.5.** The dataset "hemophilia" has data collected in order to study the detection of hemophilia A carriers. Blood samples from two groups of women were analyzed. The first group did not carry the hemophilia gene; this is known as the "non-carrier" group. The second group consists of known hemophilia carriers; this is known as the "carrier" group. Two characteristics of each subject's blood, denoted act and ant, were measured; see the pdf file for the dataset for further details.

(a) Plot estimates of the densities of $(\text{act}, \text{ant})$ for the non-carrier and carrier groups. Briefly summarize the differences between the estimates for the two groups?

(b) Using the method described in Example 6.3, estimate the probability that a subject is a hemophilia A carrier, as a function of her values of act and ant; present the result as a plot of the estimated probability function. Choose the smoothing parameters needed using the Sheather-Jones method. Restrict the analysis to values of ant and act in the range $(-0.4, 0.2)$ for each variable; to do this, include the arguments `xlim=c(-0.4, 0.2)` and `ylim=c(-0.4, 0.2) in the function \verb`sm.density—.

(c) Using the result given in part (a), estimate the probability that a subject is a hemophilia carrier given values of $(\text{act}, \text{ant})$ of $(-0.15, 0.15)$, $(0.1, 0.1)$ and $(-0.3, 0.1)$.