

Kernel Density Estimation

The posterior density (or marginals thereof) may be estimated using *kernel density estimation*.

Let X_1, \dots, X_n be independent and identically distributed (scalar) observations from density f . We may estimate $f(x)$ by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right),$$

where K is a function called the *kernel* and h is a positive number called the *bandwidth*.

A kernel density estimator is a more sophisticated version of a histogram.

The kernel is usually a probability density function with finite variance and mean 0. A popular choice for K is a standard normal density.

At a given x , the kernel estimator in essence counts how many values are in a small neighborhood centered at x .

Rather than using a “straight” count, it assigns a weight to each data point with the weights being (roughly) inversely proportional to how far points are from x .

The bandwidth h controls the size of the neighborhood centered at x .

- Small values of h produce very “wiggly” density estimates, while
- large h 's produce very smooth estimates, i.e., ones with few bumps.

The choice of kernel is less important than choice of h . The Gaussian kernel is nearly optimal, and any number of other kernels is as well.

However, for a given data set, bandwidths at the two extremes ($h = 0$ and $h = \infty$) produce (in general) very poor estimates, and so it is important to choose h wisely.

An often used optimality criterion is *mean squared error*. Consider

$$MSE(x; h) = E \left[\hat{f}_h(x) - f(x) \right]^2 .$$

It would be sensible to choose h in such a way that $MSE(x; h)$ is minimized.

One may show that, when n is large, the h that minimizes $MSE(x; h)$ is approximately

$$h_n(x) = \left[\frac{f(x)}{(f''(x))^2} \cdot C_K \right]^{1/5} n^{-1/5}, \quad (K1)$$

where

$$C_K = \frac{\int_{-\infty}^{\infty} K^2(y) dy}{\sigma_K^4}$$

and

$$\sigma_K^2 = \int_{-\infty}^{\infty} y^2 K(y) dy.$$

There are two practical problems with (K1). First, it depends on f , which is unknown, and second it implies that we use a different bandwidth at every x .

The second problem can be avoided by using the criterion of *integrated* mean squared error.

Integrated mean squared error is

$$IMSE(h) = \int_{-\infty}^{\infty} MSE(x; h) dx.$$

The asymptotic minimizer of $IMSE(h)$ is

$$h_n = \left[\frac{C_K}{\int_{-\infty}^{\infty} (f''(x))^2 dx} \right]^{1/5} n^{-1/5}. \quad (K2)$$

What about the quantity $I_f = \int_{-\infty}^{\infty} (f''(x))^2 dx$?

In Bayesian applications, when f is a posterior density, we know that it is often reasonable to assume that the posterior is approximately normal.

If

$$f(x) = \frac{1}{\sigma} \phi \left(\frac{x - \mu}{\sigma} \right),$$

then

$$f''(x) = \frac{1}{\sigma^3} \phi'' \left(\frac{x - \mu}{\sigma} \right).$$

So,

$$\begin{aligned} I_f &= \frac{1}{\sigma^5} \int_{-\infty}^{\infty} (\phi''(y))^2 dy \\ &= \frac{1}{\sigma^5} \int_{-\infty}^{\infty} (y^2 - 1)^2 \phi^2(y) dy \\ &= \frac{3}{8\sqrt{\pi}} \sigma^{-5}. \end{aligned}$$

If we use $K \equiv \phi$, it is easy to verify that

$$C_K = \frac{1}{2\sqrt{\pi}},$$

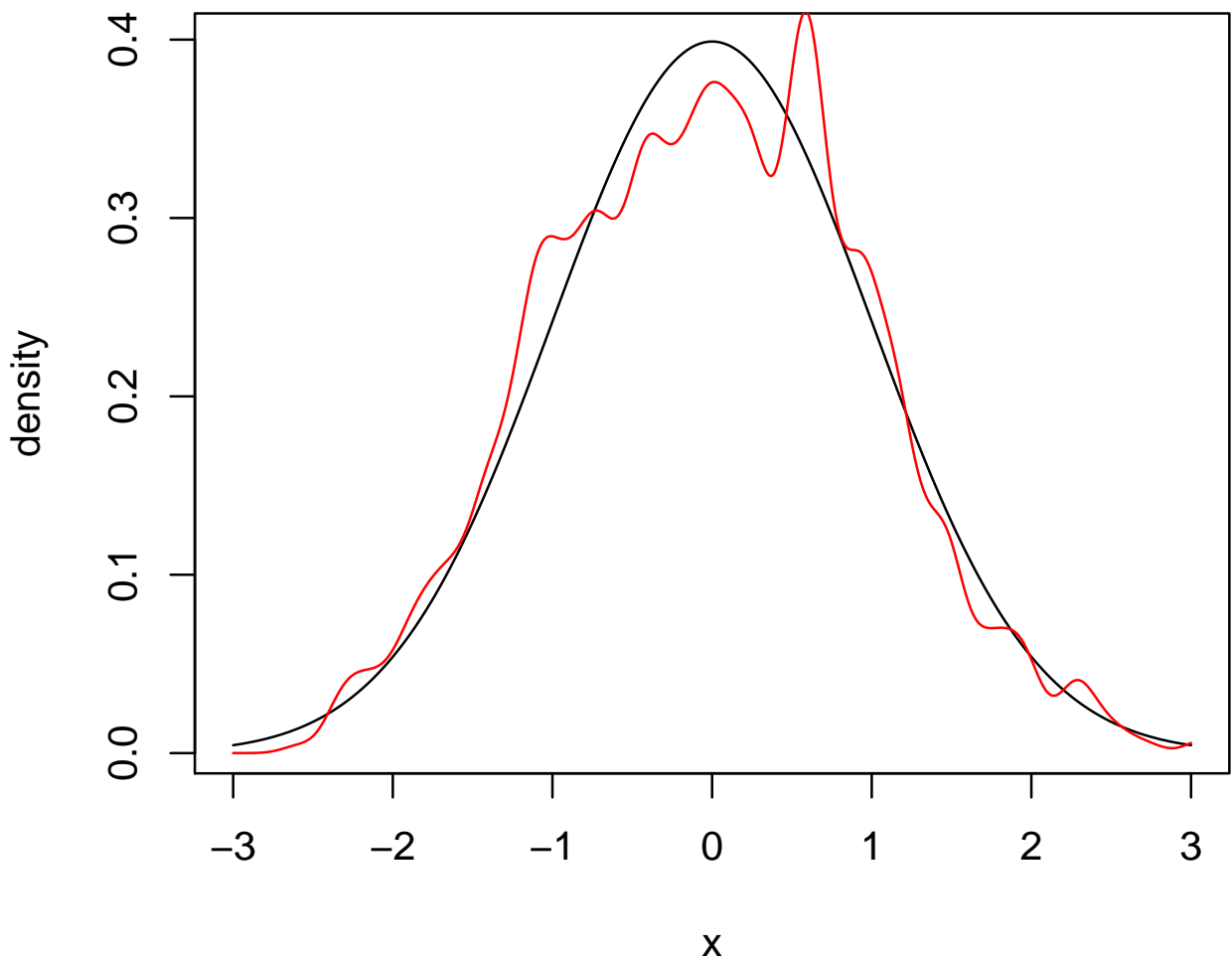
and hence when the underlying density is Gaussian and we use a Gaussian kernel, the asymptotically optimal bandwidth is

$$h_n = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5} \approx 1.06 \sigma n^{-1/5} \approx \sigma n^{-1/5}.$$

This makes for a very simple bandwidth selection rule. Simply estimate σ by the sample standard deviation s , and then use

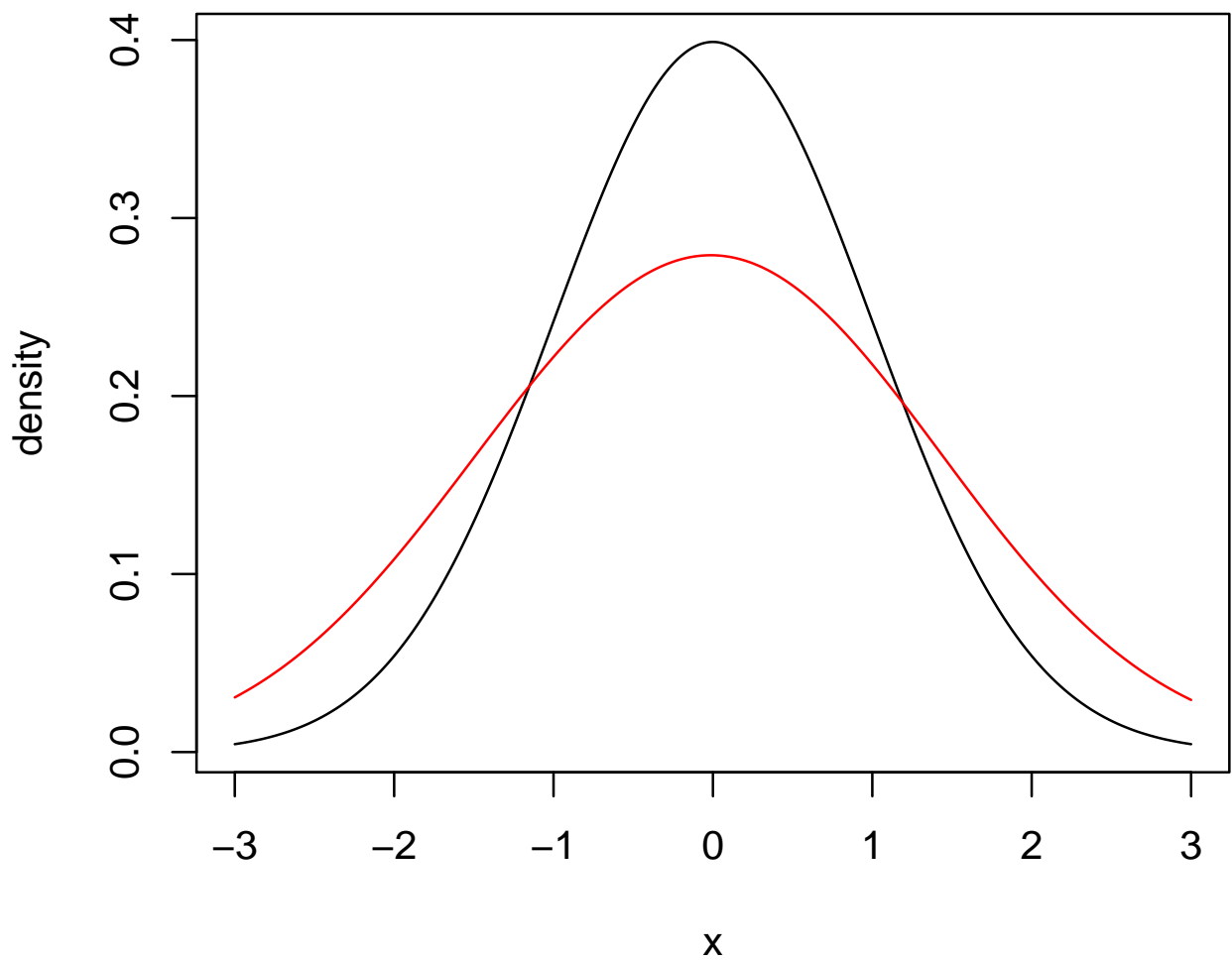
$$\hat{h}_n = s n^{-1/5}.$$

*Kernel density estimate based on a
random sample of size 1000
from the standard normal*



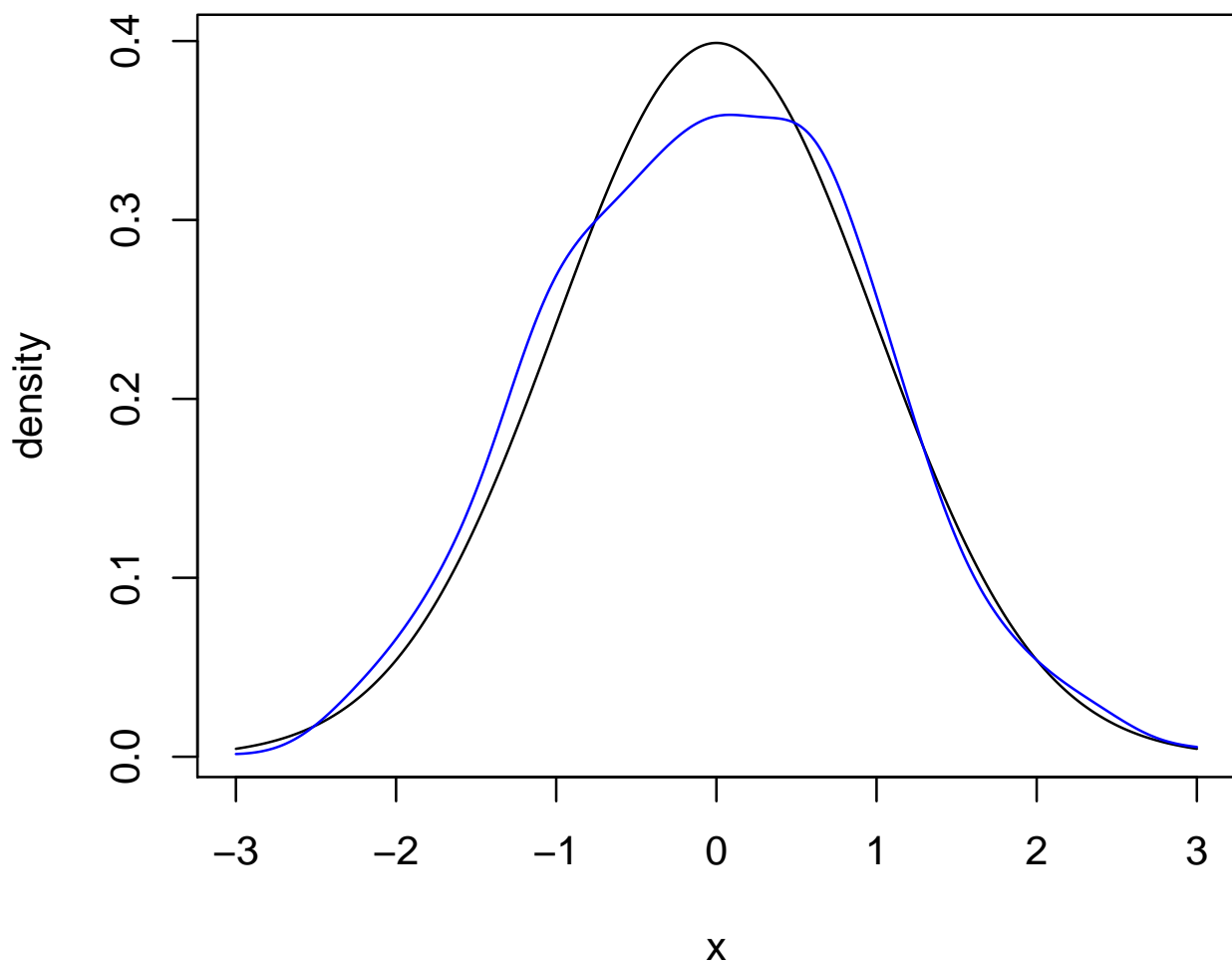
The kernel is Gaussian and $h = 0.1$.

*Kernel estimate for the same
data as on p. 316N*



The kernel is Gaussian and $h = 1$.

*Kernel estimate for the same
data as on p. 316N*



The kernel is Gaussian and
 $h = sn^{-1/5} = 0.25$.

When the data are generated using MCMC, they are not independent. What effect does this have on the conclusions of pp. 310-318N?

The answer is *almost none* if the *chain mixes even moderately well*.

The reason for this is a principle I call *whitening by windowing*. (See Hart, (1996), *Journal of Nonparametric Statistics*, pp. 115-142.)

Whitening means operating upon the data in such a way that the transformed data are uncorrelated.

Windowing means considering only the data that are in a relatively small window (or band) centered at x .

When the data are rapidly mixing, the observations that end up in this small window come from widely different segments of the chain, and hence are approximately uncorrelated.

Another way to think of this is that *the data used by a kernel density estimator are, in essence, automatically thinned.*

Theoretically, it can be shown that, when the chain is rapidly mixing, the expressions obtained previously for $MSE(x; h)$, $h_n(x)$ and h_n are still valid.

Likewise, the bandwidth selection rule (based on approximately normal observations) is still valid under rapid mixing.

Multivariate density estimation

Density estimation for multivariate data is based on the same principles as for univariate data. The main difficulty encountered is the so-called *curse of dimensionality*.

Even if the dimension is fairly small (two or three), enormous sample sizes are needed to avoid an extremely sparse data distribution.

Estimating bivariate or trivariate distributions is probably feasible with MCMC since we can generate large numbers of observations.

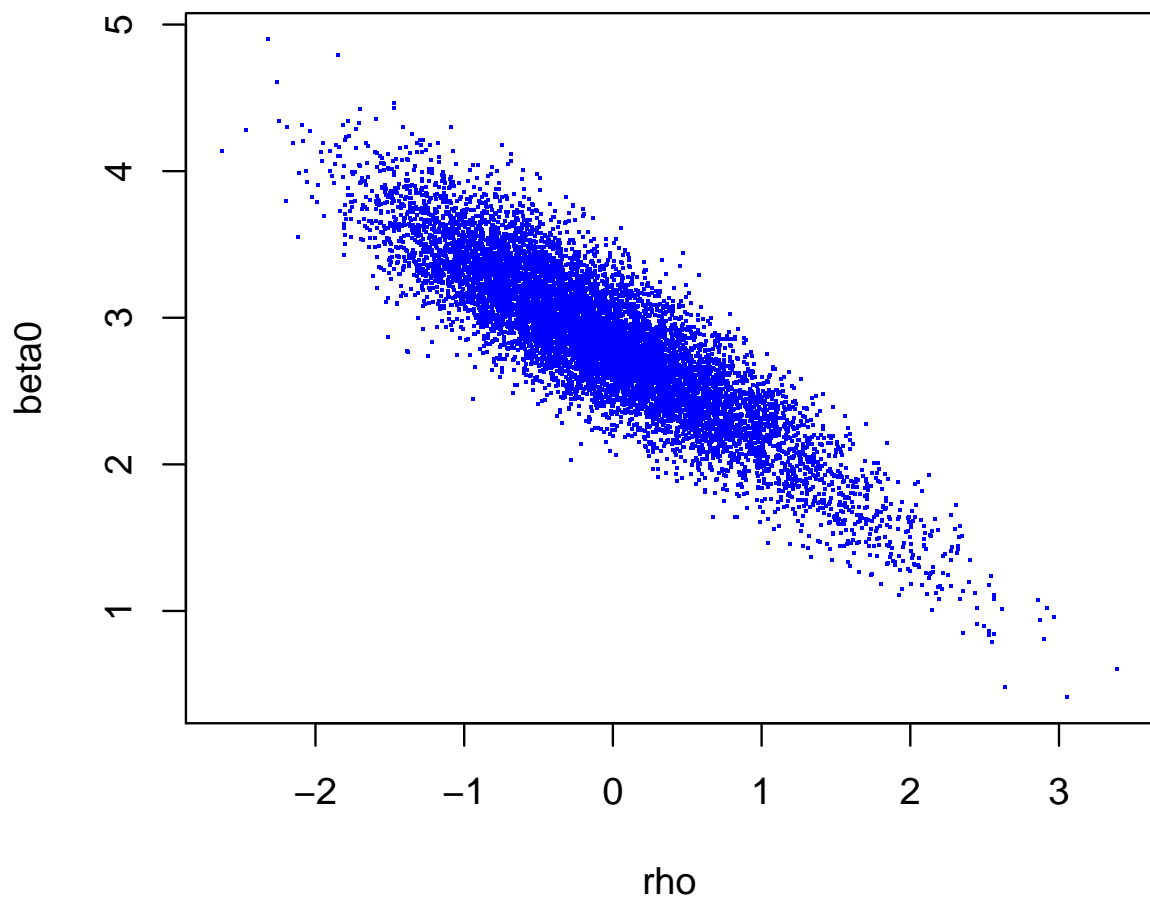
Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be a random sample from a bivariate density f . We may estimate $f(x_1, x_2)$ by

$$\hat{f}_h(x_1, x_2) = \frac{1}{nh_1h_2} \sum_{i=1}^n K\left(\frac{x_1 - X_{1i}}{h_1}\right) \times K\left(\frac{x_2 - X_{2i}}{h_2}\right),$$

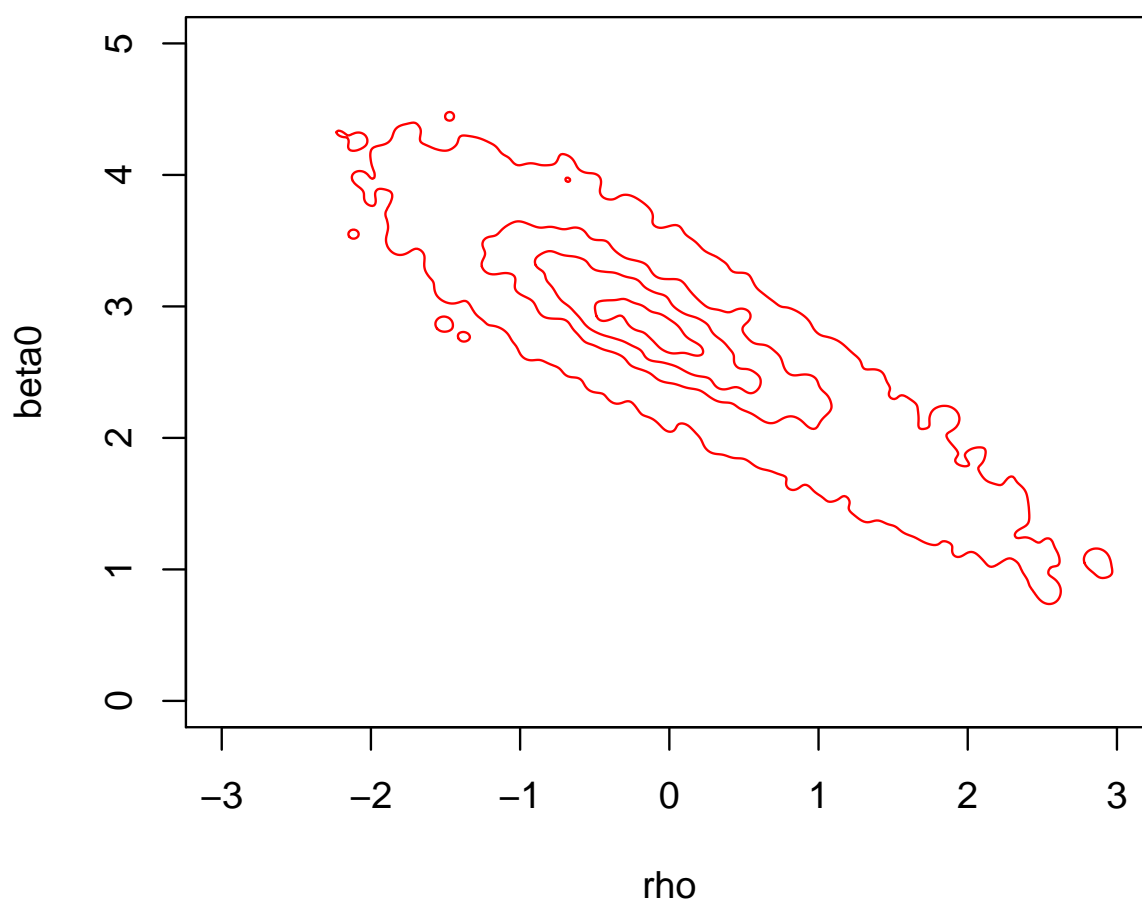
where K is a density satisfying the same properties as in the univariate case and h_1 and h_2 are bandwidths.

We will apply this method to estimate the joint posterior of ρ and β_0 in Example 20. The posterior is graphed in terms of contours.

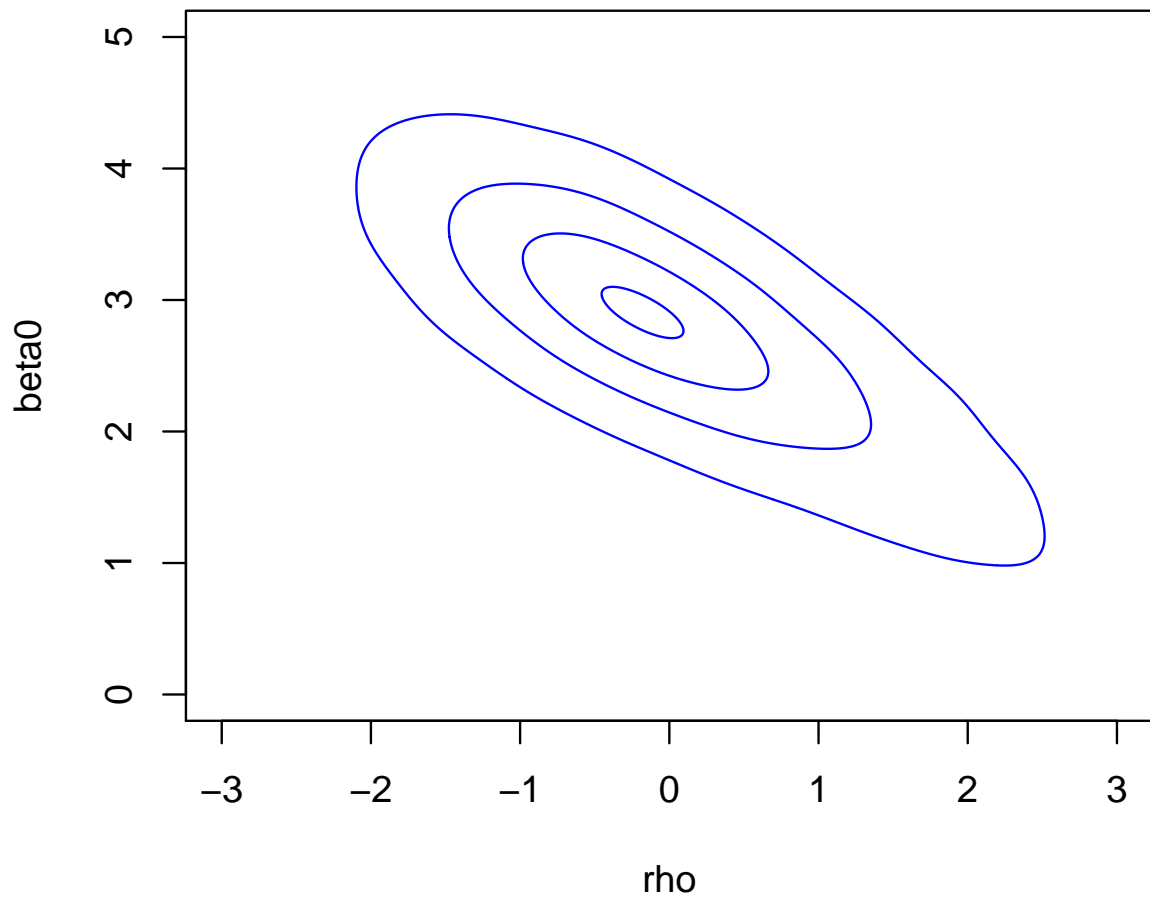
Scatterplot of β_0 vs. ρ for the output generated in Example 20



Below are contours of a kernel estimate for the data on p. 323N. The estimate is as defined on p. 322N with $K \equiv \phi$ and $h_1 = h_2 = 0.05$.



Here the bandwidths are $h_1 = h_2 = 0.25$.



Here the bandwidths are $h_1 = h_2 = 0.5$.

