

ks: Kernel density estimation for bivariate data

Tarn Duong

30 September 2011

Kernel density estimation is a popular tool for visualising the distribution of data. See Simonoff (1996), for example, for an overview. When multivariate kernel density estimation is considered it is usually in the constrained context with diagonal bandwidth matrices, e.g. in the R packages `sm` (Bowman and Azzalini, 2007) and `KernSmooth` (Wand, 2006). We introduce a new R package `ks` which implements diagonal and unconstrained data-driven bandwidth matrices for kernel density estimation, which can also be used for multivariate kernel discriminant analysis. The `ks` package implements selectors for 1- to 6-dimensional data.

This vignette contains only a brief introduction to using `ks` for kernel density estimation for 2-dimensional data. See Duong (2007) for a more detailed account. For a bivariate random sample $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ drawn from a density f , the kernel density estimate is defined by

$$\hat{f}(\mathbf{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$$

where $\mathbf{x} = (x_1, x_2)^T$ and $\mathbf{X}_i = (X_{i1}, X_{i2})^T, i = 1, 2, \dots, n$. Here $K(\mathbf{x})$ is the kernel which is a symmetric probability density function, \mathbf{H} is the bandwidth matrix which is symmetric and positive-definite, and $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2} \mathbf{x})$. The choice of K is not crucial: we take $K(\mathbf{x}) = (2\pi)^{-1} \exp(-\frac{1}{2} \mathbf{x}^T \mathbf{x})$ the standard normal throughout. In contrast, the choice of \mathbf{H} is crucial in determining the performance of \hat{f} . The most common parameterisations of the bandwidth matrix are the diagonal and the general or unconstrained which has no restrictions on \mathbf{H} provided that \mathbf{H} remains positive definite and symmetric, that is

$$\mathbf{H} = \begin{bmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{bmatrix} \text{ or } \mathbf{H} = \begin{bmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{bmatrix}.$$

This latter parameterisation allows kernels to have an arbitrary orientation whereas the former only allows kernels which are oriented to the co-ordinate axes.

For our target density, we use the ‘dumbbell’ density, given by the normal mixture

$$\frac{4}{11} N\left(\begin{bmatrix} -2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) + \frac{3}{11} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.8 & -0.72 \\ -0.72 & 0.8 \end{bmatrix}\right) + \frac{4}{11} N\left(\begin{bmatrix} 2 \\ -2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right),$$

displayed on the left in Figure 1. This density is unimodal. On the right is a sample of 200 data points.

```
> library(ks)
> set.seed(8192)
> samp <- 200
> mus <- rbind(c(-2, 2), c(0, 0), c(2, -2))
> Sigmas <- rbind(diag(2), matrix(c(0.8, -0.72, -0.72, 0.8), nrow = 2),
+   diag(2))
> cwt <- 3/11
> props <- c((1 - cwt)/2, cwt, (1 - cwt)/2)
> x <- rmvnorm.mixt(n = samp, mu = mus, Sigma = Sigmas, props = props)
```

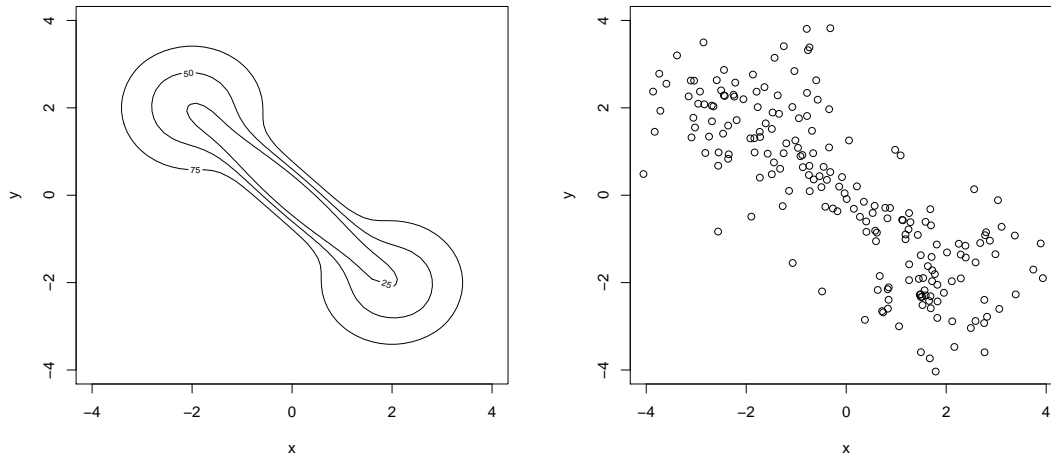


Figure 1: Target ‘dumbbell’ density. (Left) contour plot. (Right) Scatter plot.

We use `Hpi` for unconstrained plug-in selectors and `Hpi.diag` for diagonal plug-in selectors.

```
> Hpi1 <- Hpi(x = x)

      [,1]      [,2]
[1,] 0.4133914 -0.2843047
[2,] -0.2843047 0.3431130

> Hpi2 <- Hpi.diag(x = x)

      [,1]      [,2]
[1,] 0.226123 0.0000000
[2,] 0.000000 0.1826778
```

To compute a kernel density estimate, the command is `kde`, which creates a `kde` class object

```
> fhat.pi1 <- kde(x = x, H = Hpi1)
> fhat.pi2 <- kde(x = x, H = Hpi2)
```

We use the `plot` method for `kde` objects to display these kernel density estimates. The default is a contour plot with the upper 25%, 50% and 75% contours of the (sample) highest density regions. These regions are also plotted by the `sm` library.

```
> plot(fhat.pi1)
> plot(fhat.pi2)
```

The respective kernel density estimates are produced in Figure 2. The diagonal bandwidth matrix constrains the smoothing to be performed in directions parallel to the co-ordinate axes, so it is not able to apply accurate levels of smoothing to the obliquely oriented central portion. The result is a multimodal density estimate. The unconstrained bandwidth matrix correctly produces a unimodal density estimate.

The unconstrained SCV (Smoothed Cross Validation) selector is `Hscv` and its diagonal version is `Hscv.diag`. In Figure 3, the most reasonable density estimate is from the unconstrained SCV selector.

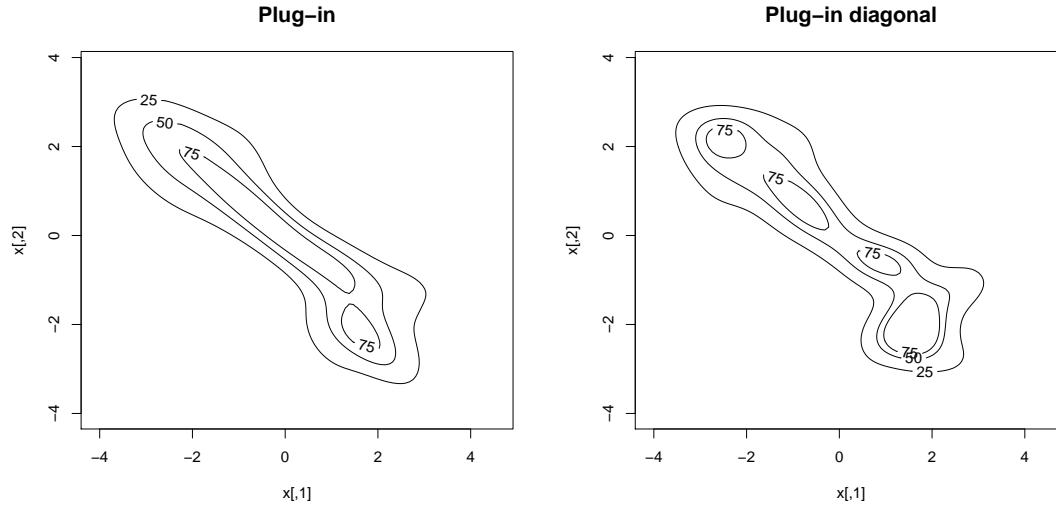


Figure 2: Kernel density estimates with plug-in selectors

```
> Hscv1 <- Hscv(x = x)

      [,1]      [,2]
[1,] 0.5381572 -0.3814680
[2,] -0.3814680 0.4664029

> Hscv2 <- Hscv.diag(x = x)

      [,1]      [,2]
[1,] 0.2783832 0.0000000
[2,] 0.0000000 0.2382577
```

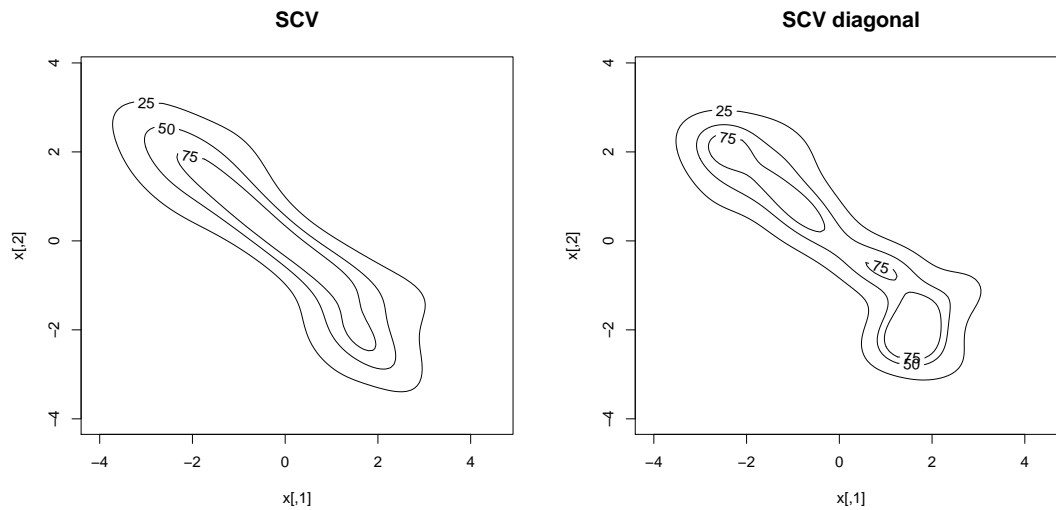


Figure 3: Kernel density estimates with cross validation selectors

The unconstrained bandwidth selectors will be better than their diagonal counterparts when the data have large mass oriented obliquely to the co-ordinate axes, like for the dumbbell data. The unconstrained plug-in and the SCV selectors can be viewed as generally recommended selectors.

References

- Bowman, A. W. and Azzalini, A. (2007). *sm: kernel smoothing methods: Bowman and Azzalini (1997)*. R package version 2.2.
- Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software*. **21** (7), URL <http://www.jstatsoft.org/v21/i07>.
- Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag, New York.
- Wand, M. P. (2006). *KernSmooth: Functions for kernel smoothing for Wand & Jones (1995)*. R package version 2.22-19. R port by Brian Ripley.