

Using ks for bivariate kernel density estimation

Tarn Duong

Department of Statistics, University of New South Wales
Sydney Australia

27 September 2006

1 Introduction

Kernel density estimation has become 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. the R packages `sm` and `KernSmooth`. We introduce a new package `ks` for kernel smoothing with unconstrained bandwidth matrices. Currently kernel density estimation and kernel discriminant analysis for 1- to 6-dimensional data are available. This vignette focuses on kernel density estimation for the 2-dimensional case.

2 Kernel density estimation

For a bivariate sample $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ drawn from a density f , the kernel density estimate is

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

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 bivariate kernel (which we assume to be a probability density function); $\mathbf{H} = \begin{bmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{bmatrix}$ 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. On the other hand, the choice of \mathbf{H} is crucial in determining the performance of \hat{f} .

We measure the performance of \hat{f} (in common with the majority of researchers in this field) using the Mean Integrated Squared Error (MISE) criterion,

$$\text{MISE}(\mathbf{H}) = \mathbb{E} \int_{\mathbb{R}^2} [\hat{f}(\mathbf{x}; \mathbf{H}) - f(\mathbf{x})]^2 d\mathbf{x}.$$

Our aim in bandwidth selection is to estimate

$$\mathbf{H}_{\text{MISE}} = \underset{\mathbf{H}}{\text{argmin}} \text{MISE}(\mathbf{H}),$$

over the space of all symmetric, positive definite 2×2 matrices. It is well known that the optimal bandwidth \mathbf{H}_{MISE} does not have a closed form. To make progress it is usual to employ

an asymptotic approximation, known as the AMISE (Asymptotic MISE):

$$\text{AMISE}(\mathbf{H}) = (4\pi)^{-1}n^{-1}|\mathbf{H}|^{-1/2} + \frac{1}{4}(\text{vech}^T \mathbf{H})\Psi_4(\text{vech } \mathbf{H}) \quad (2)$$

where vech is the vector half operator i.e.

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

We can show that

$$\Psi_4 = \begin{bmatrix} \psi_{40} & 2\psi_{31} & \psi_{22} \\ 2\psi_{31} & 4\psi_{22} & 2\psi_{13} \\ \psi_{22} & 2\psi_{13} & \psi_{04} \end{bmatrix} \quad (3)$$

where the integrated density derivative functional is

$$\psi_{r_1, r_2} = \int_{\mathbb{R}^2} f^{(r_1, r_2)}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

and the partial derivatives of f are

$$f^{(r_1, r_2)}(\mathbf{x}) = \frac{\partial^4}{\partial x_1^{r_1} \partial x_2^{r_2}} f(\mathbf{x}).$$

(Note that the subscript 4 on Ψ relates to the order of the derivatives involved).

We make use of the tractability of AMISE by seeking

$$\mathbf{H}_{\text{AMISE}} = \underset{\mathbf{H}}{\text{argmin}} \text{AMISE}(\mathbf{H})$$

rather than \mathbf{H}_{MISE} . For the next step we estimate the AMISE. The minimiser of the $\widehat{\text{AMISE}}$ is then a data-driven bandwidth selector i.e.

$$\hat{\mathbf{H}} = \underset{\mathbf{H}}{\text{argmin}} \widehat{\text{AMISE}}(\mathbf{H}). \quad (4)$$

3 Plug-in bandwidth selectors

The most well-known univariate plug-in selector is due to Sheather and Jones (1991). Plug-in selectors require pilot estimates of the ψ_{r, r_2} functionals which comprise Ψ_4 . This in turn produces an estimate of the AMISE

$$\text{PI}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + \frac{1}{4}(\text{vech}^T \mathbf{H})\hat{\Psi}_4(\text{vech } \mathbf{H}) \quad (5)$$

that can be numerically minimised to give the plug-in bandwidth matrix, $\hat{\mathbf{H}}_{\text{PI}}$. To compute $\hat{\Psi}_4$ we need to use a helper or ‘pilot’ bandwidth matrix \mathbf{G} . Like \mathbf{H} , we need to choose a sensible value for \mathbf{G} . If we note that $\psi_{r_1, r_2} = \mathbb{E} f^{(r_1, r_2)}(\mathbf{X})$ where $\mathbf{X} \sim f$, then a natural estimator of ψ_{r_1, r_2} is

$$\hat{\psi}_{r_1, r_2}(\mathbf{G}) = n^{-1} \sum_{i=1}^n \hat{f}^{(r_1, r_2)}(\mathbf{X}_i; \mathbf{G}) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n K_{\mathbf{G}}^{(r_1, r_2)}(\mathbf{X}_i - \mathbf{X}_j). \quad (6)$$

3.1 AMSE pilot bandwidth selectors

We consider pilot bandwidth matrices of the form $\mathbf{G} = g^2 \mathbf{I}$. It may appear that we are contradicting ourselves since we advocate using unconstrained matrices for \mathbf{H} . It turns out that a restricted form on \mathbf{G} has less effect on the performance of \hat{f} than a restricted form on \mathbf{H} . We can reduce the effect of the restricted form of \mathbf{G} by using an appropriate pre-transformation (we return to this topic in Section 6).

The AMSE-optimal pilot bandwidths (i.e. optimal for the Asymptotic Mean Squared Error) are (a) if both r_1 and r_2 are even then

$$g_{r_1, r_2, \text{AMSE}} = \left[\frac{A_{r_1, r_2}}{n} \right]^{1/8} \quad (7)$$

and (b) if at least one of r_1 and r_2 is odd then

$$g_{r_1, r_2, \text{AMSE}} = \left[\frac{B_{r_1, r_2}}{n} \right]^{1/7}. \quad (8)$$

See Wand and Jones (1994) for explicit expressions for A_{r_1, r_2} and B_{r_1, r_2} . We can use 1 or 2 stages to estimation these pilot bandwidths. Intuitively we can think of 2 estimation stages as a strategy to reduce the dependence on the normality assumptions for 1 stage estimation.

The $\hat{\Psi}_4$ estimated in this element-wise way uses a different pilot bandwidth for each unique element (5 in total). Then $\hat{\Psi}_4$ is not guaranteed to be positive definite. Hence using appropriate estimators of each element of a matrix will not necessarily lead to an appropriate estimator of the matrix as a whole. Positive-definiteness can be guaranteed by using a single, common pilot bandwidth.

3.2 SAMSE pilot bandwidth selectors

Modifying AMSE pilot selectors, we derive a SAMSE (Sum of Asymptotic Mean Squared Error) pilot selector. This type of selector has been specially devised to maintain the positive definiteness of $\hat{\Psi}_4$ which is crucial to the numerical minimisation of the plug-in criterion PI. This method is also simpler and more parsimonious than AMSE selectors.

Let the fourth order SAMSE be

$$\text{SAMSE}_4(g) = n^{-2}g^{-12}A_1 + n^{-1}g^{-4}A_2 + \frac{1}{4}g^4A_3 \quad (9)$$

where A_1 and A_3 are positive constants and A_2 is a negative constant. This has a minimum at

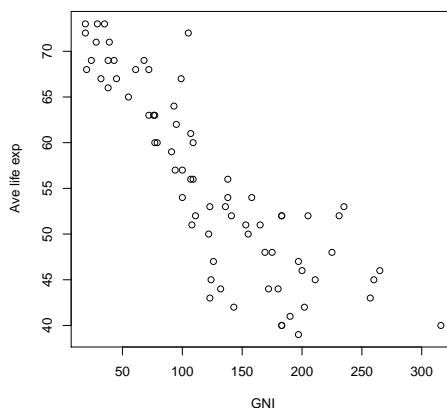
$$g_{4, \text{SAMSE}} = \left[\frac{6A_2}{(-A_2 + \sqrt{A_2^2 + 3A_1A_3})n} \right]^{1/8}. \quad (10)$$

This is the SAMSE optimal pilot bandwidth. Note that only one pilot bandwidth is required to compute $\hat{\Psi}_4$. Explicit expressions for A_1, A_2 and A_3 can be found in Duong and Hazelton (2003). Again we have the choice of either 1 or 2 stages for estimating the pilot bandwidth.

3.3 R examples

The `unicef` data set contains the number of deaths of children under 5 years of age per 1000 live births and the average life expectancy (in years) at birth for 73 countries with GNI (Gross National Income) less than 1000 US dollars per annum per capita. The aim of kernel density estimation is to find a smooth estimate of the distribution of this data.

```
> library(ks)
> data(unicef)
> plot(unicef)
```



Use `Hpi` for full plug-in selectors and `Hpi.diag` for diagonal plug-in selectors. There are two arguments which further specify the plug-in selector used: `nstage` is the number of pilot estimation stages (1 or 2) and `pilot` is the type of pilot estimation (“amse” or “samse”). The other argument `pre` involves the pre-transformations outlined in Section 6.

```
> Hpi1 <- Hpi(x = unicef, nstage = 1, pilot = "amse", pre = "scale")
```

```
      [,1]      [,2]
[1,] 391.03446 -34.733738
[2,] -34.73374   9.898164
```

```
> Hpi2 <- Hpi(x = unicef, nstage = 2, pilot = "samse", pre = "sphere")
```

```
      [,1]      [,2]
[1,]  797.5755 -106.63338
[2,] -106.6334   19.56761
```

```
> Hpi3 <- Hpi.diag(x = unicef, nstage = 2, pilot = "amse", pre = "scale")
```

```
      [,1]      [,2]
[1,] 201.5118 0.000000
[2,]   0.0000 6.242821
```

```
> Hpi4 <- Hpi.diag(x = unicef, nstage = 2, pilot = "samse", pre = "scale")
```

```
      [,1]      [,2]
[1,] 227.0192 0.000000
[2,]   0.0000 6.179491
```

To compute a kernel density estimate (Equation (1)), the command is `kde`.

```

> fhat1 <- kde(x = unicef, H = Hpi1)
> fhat2 <- kde(x = unicef, H = Hpi2)
> fhat3 <- kde(x = unicef, H = Hpi3)
> fhat4 <- kde(x = unicef, H = Hpi4)

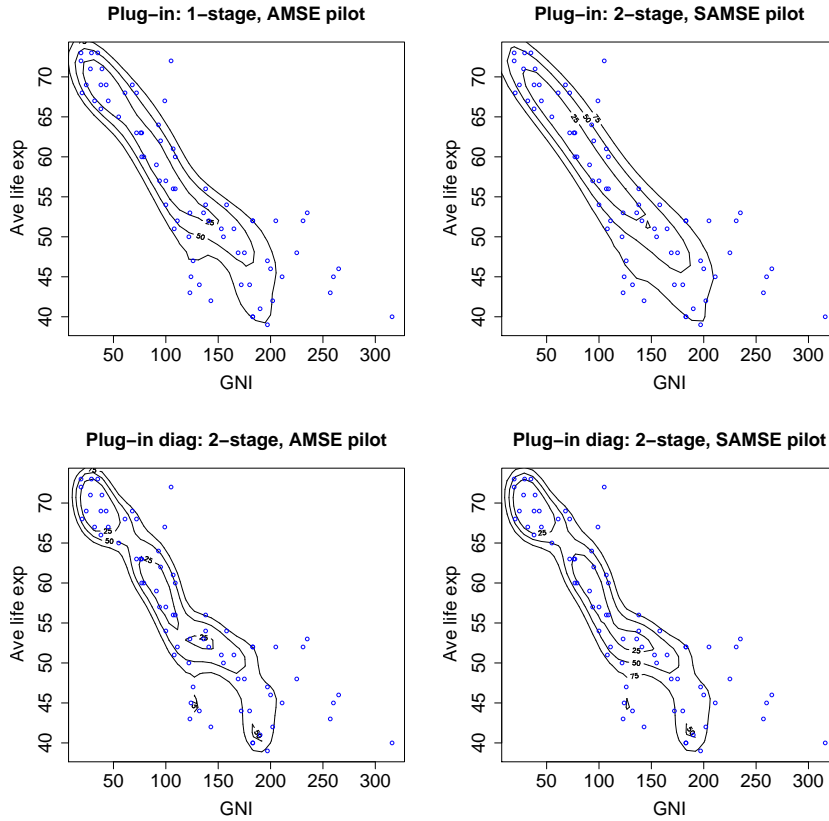
```

We can use the `plot` command to display these kernel density estimates.

```

> plot(fhat1, main = "Plug-in: 1-stage, AMSE pilot")
> plot(fhat2, main = "Plug-in: 2-stage, SAMSE pilot")
> plot(fhat3, main = "Plug-in diag: 2-stage, AMSE pilot")
> plot(fhat4, main = "Plug-in diag: 2-stage, SAMSE pilot")

```



4 Cross validation bandwidth selectors

Cross-validation selectors are the main alternative to plug-in selectors. There are three main types of cross-validation selectors: least squares, biased and smoothed.

4.1 Least squares cross validation

The bivariate version of the least squares cross validation (LSCV) criterion of Rudemo (1982) and Bowman (1984) is

$$\text{LSCV}(\mathbf{H}) = \int_{\mathbb{R}^2} \hat{f}(\mathbf{x}; \mathbf{H})^2 d\mathbf{x} - 2n^{-1} \sum_{i=1}^n \hat{f}_{-i}(\mathbf{X}_i; \mathbf{H}),$$

where the leave-one-out estimator is

$$\hat{f}_{-i}(\mathbf{x}; \mathbf{H}) = (n-1)^{-1} \sum_{\substack{j=1 \\ j \neq i}}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_j).$$

The LSCV selector $\hat{\mathbf{H}}_{\text{LSCV}}$ is the minimiser of $\text{LSCV}(\mathbf{H})$. We can rewrite LSCV as

$$\text{LSCV}(\mathbf{H}) = (4\pi)^{-1} n^{-1} |\mathbf{H}|^{-1/2} + n^{-1} (n-1)^{-1} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (K_{2\mathbf{H}} - 2K_{\mathbf{H}})(\mathbf{X}_i - \mathbf{X}_j). \quad (11)$$

4.2 Biased cross validation

Plug-in methods use a pilot bandwidth matrix \mathbf{G} which is independent of \mathbf{H} to Ψ_4 . For BCV, we set $\mathbf{G} = \mathbf{H}$ and use slightly different estimators. There are two versions of BCV, depending on the estimator of ψ_{r_1, r_2} with $r_1 + r_2 = 4$, see Sain et al. (1994). We can use

$$\check{\psi}_{r_1, r_2}(\mathbf{H}) = n^{-2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K_{2\mathbf{H}}^{(r_1, r_2)}(\mathbf{X}_i - \mathbf{X}_j) \quad (12)$$

or we could use

$$\tilde{\psi}_{r_1, r_2}(\mathbf{H}) = n^{-1} \sum_{i=1}^n \hat{f}_{-i}^{(r_1, r_2)}(\mathbf{X}_i; \mathbf{H}) = n^{-1} (n-1)^{-1} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K_{\mathbf{H}}^{(r_1, r_2)}(\mathbf{X}_i - \mathbf{X}_j). \quad (13)$$

The estimates $\check{\Psi}_4$ and $\tilde{\Psi}_4$ are obtained from Ψ_4 by substituting $\check{\psi}_{r_1, r_2}$ and $\tilde{\psi}_{r_1, r_2}$ for ψ_{r_1, r_2} . The BCV1 function is the version of BCV with $\check{\Psi}_4$

$$\text{BCV1}(\mathbf{H}) = (4\pi)^{-1} n^{-1} |\mathbf{H}|^{-1/2} + \frac{1}{4} (\text{vech}^T \mathbf{H}) \check{\Psi}_4 (\text{vech } \mathbf{H}) \quad (14)$$

and the BCV2 function is the version with $\tilde{\Psi}_4$

$$\text{BCV2}(\mathbf{H}) = (4\pi)^{-1} n^{-1} |\mathbf{H}|^{-1/2} + \frac{1}{4} (\text{vech}^T \mathbf{H}) \tilde{\Psi}_4 (\text{vech } \mathbf{H}). \quad (15)$$

The BCV selectors $\hat{\mathbf{H}}_{\text{BCV}}$ are the minimisers of the appropriate BCV function.

4.3 Smoothed cross validation

Smoothed cross validation (SCV), introduced by Hall et al. (1992), can be thought of as a hybrid of LSCV and BCV. The SCV criterion takes the asymptotic integrated variance from the BCV but attempts to estimate the integrated squared bias exactly rather than using its asymptotic form

$$\text{SCV}(\mathbf{H}) = (4\pi)^{-1} n^{-1} |\mathbf{H}|^{-1/2} + n^{-2} \sum_{i=1}^n \sum_{j=1}^n (K_{2\mathbf{H}+2\mathbf{G}} - 2K_{\mathbf{H}+2\mathbf{G}} + K_{2\mathbf{G}})(\mathbf{X}_i - \mathbf{X}_j). \quad (16)$$

The SCV selector $\hat{\mathbf{H}}_{\text{SCV}}$ is the minimiser of $\text{SCV}(\mathbf{H})$.

Again, we consider pilot bandwidth matrices of the form $\mathbf{G} = g^2 \mathbf{I}$. We generalise the process of Jones et al. (1991) to find an optimal pilot bandwidth selector. We can show that the AMSE optimal pilot selector is

$$g_{\text{AMSE}} = \left[\frac{12C_{\mu_2}}{(-4C_{\mu_1} + \sqrt{C_{\mu_0}})n} \right]^{1/8} \quad (17)$$

where $C_{\mu_0}, C_{\mu_1}, C_{\mu_2}$ are constants. The details of the derivation of this pilot bandwidth is available in Duong and Hazelton (2005).

4.4 R examples

We continue with the `unicef` data. `Hlscv` and `Hlscv.diag` are the full and diagonal LSCV selectors. (The `Hstart` argument in the below code specifies the initial value for the numerical minimisation - it can be specified for any of the bandwidth selector functions). `Hbcv` implements both BCV1 and BCV2. The default is BCV1; set `whichbcv=2` to use BCV2. Their diagonal counterpart is `Hbcv.diag`. `Hscv` is the full SCV selector (no diagonal version is available). Its argument `pre` is the same as for `Hpi` and `Hpi.diag` in Section 4.4.

```
> Hlscv1 <- Hlscv(unicef)

      [,1]      [,2]
[1,] 388.18250 -83.34084
[2,] -83.34084  25.12909

> Hlscv2 <- Hlscv.diag(unicef, Hstart = Hlscv1)

      [,1]      [,2]
[1,] 194.4292  0.00000
[2,]  0.0000 11.11751

> Hbcv1 <- Hbcv(unicef)

      [,1]      [,2]
[1,] 1087.0682 135.33067
[2,] 135.3307  23.58613

> Hbcv2 <- Hbcv.diag(unicef, whichbcv = 2)

      [,1]      [,2]
[1,] 1072.781 0.000000
[2,]  0.000 9.298466

> Hscv1 <- Hscv(unicef, pre = "sphere")

      [,1]      [,2]
[1,] 1322.3429 -191.76583
[2,] -191.7658  34.98503

> Hscv2 <- Hscv(unicef, pre = "scale")
```

```

      [,1]      [,2]
[1,] 694.06056 -73.07412
[2,] -73.07412  17.49643

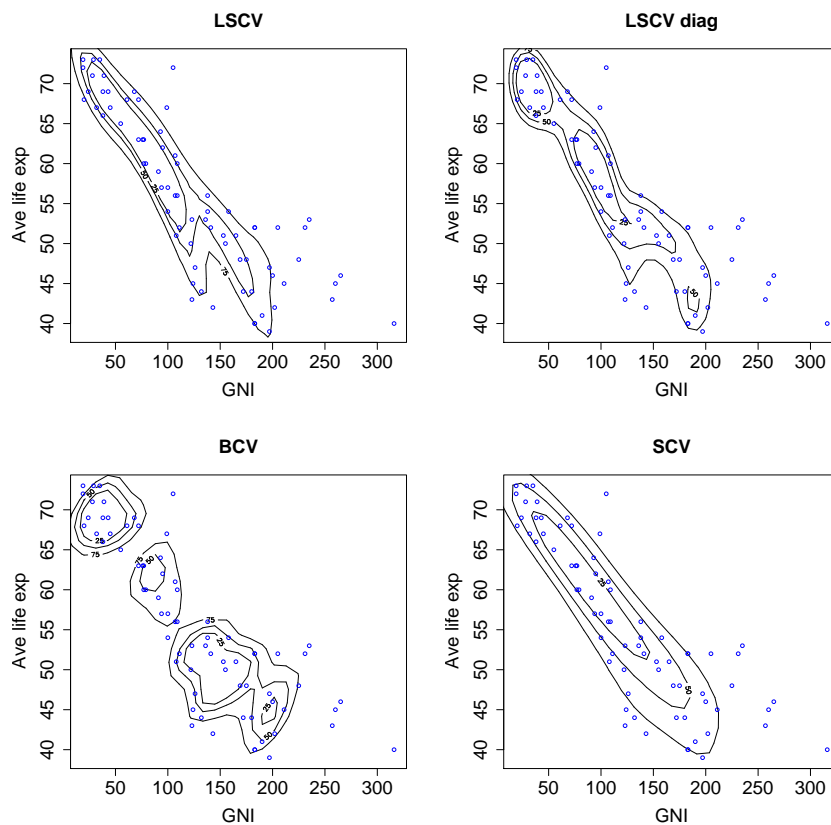
```

```

> fhat1 <- kde(unicef, Hlscv1)
> fhat2 <- kde(unicef, Hlscv2)
> fhat3 <- kde(unicef, Hbcv1)
> fhat4 <- kde(unicef, Hscv1)

> plot(fhat1, main = "LSCV")
> plot(fhat2, main = "LSCV diag")
> plot(fhat3, main = "BCV")
> plot(fhat4, main = "SCV")

```



5 More graphics

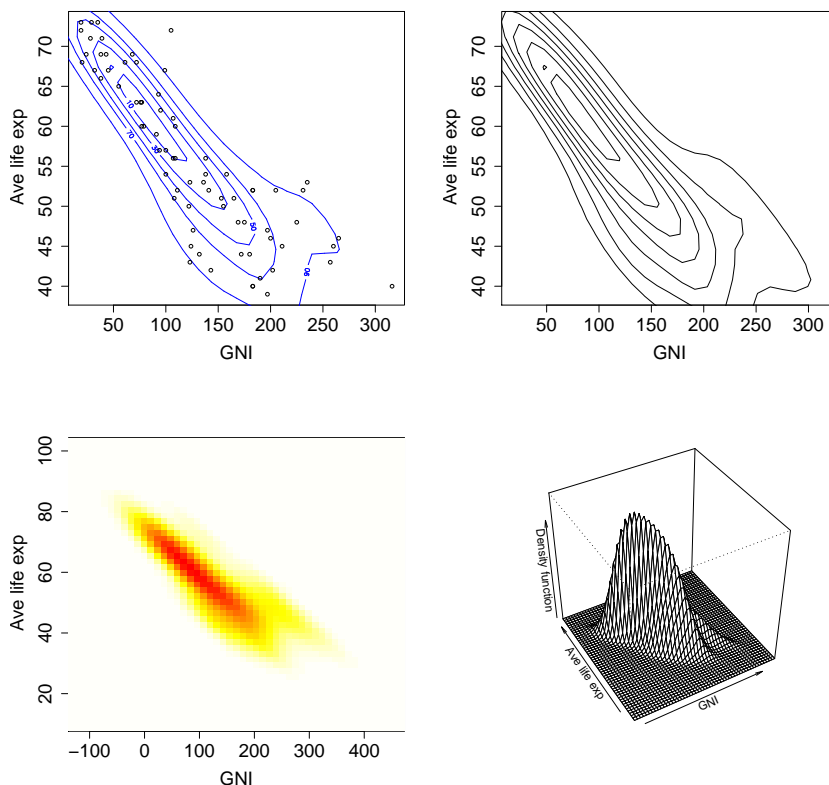
There are many options with plotting kernel density estimates. Slice or contour plots are the default or can be explicitly called by setting `display="slice"`. The way the contours are plotted is controlled by `cont` or `ncont`: only one of these needs to be set. The argument `cont` takes a vector of percentages and produces a set of contours at the levels corresponding to the percentages of the maximum height of the density estimate. The argument `ncont` takes a number and R tries to compute a pretty set of `ncont` contours. The colour(s) of the contour

lines is `lcol` and the colour(s) of the plotting symbols is `ptcol`. The logical flags `drawlabels` and `drawpoints` indicate whether to draw the labels for the contours levels or the data points.

An alternative to contour plots are the image or heat plots, called by `display="image"`. These are similar to contour plots except that the heights of the density estimate are represented by different colours rather than with different level curves. The default colours is `heat.colors` but we use instead `rev(heat.colors)` which gives us the high values as red and the low values (close to zero) as white with yellow/orange as intermediate.

The other alternative is a perspective or wire-frame plot, called by `display="persp"`. This is an attempt to capture the three-dimensional structure more directly than the image or contour plots.

```
> plot(fhat4, lcol = "blue", ptcol = "black", cont = seq(10, 90,
+       by = 20))
> plot(fhat4, ncont = 8, drawlabels = FALSE, drawpoints = FALSE)
> plot(fhat4, display = "image", col = rev(heat.colors(100)))
> plot(fhat4, display = "persp")
```



6 Pre-transformations

The plug-in and SCV selector functions use pilot bandwidths of the form $\mathbf{G} = g^2 \mathbf{I}$ which will be inappropriate if the dispersion of the data differs markedly between the two coordinate

directions. We estimate using the transformed data $\mathbf{X}_1^*, \mathbf{X}_2^*, \dots, \mathbf{X}_n^*$, where the transformation is either *sphering*

$$\mathbf{X}^* = \mathbf{S}^{-1/2} \mathbf{X}$$

where \mathbf{S} is the sample covariance matrix of the untransformed data; or *scaling*

$$\mathbf{X}^* = \mathbf{S}_D^{-1/2} \mathbf{X}$$

where $\mathbf{S}_D = \text{diag}(s_1^2, s_2^2)$ and s_1^2, s_2^2 are the marginal sample variances. The bandwidth matrix $\hat{\mathbf{H}}^*$ for the sphered or scaled data can be back transformed to the original scale by $\hat{\mathbf{H}} = \mathbf{S}^{1/2} \hat{\mathbf{H}}^* \mathbf{S}^{1/2}$ or $\hat{\mathbf{H}} = \mathbf{S}_D^{1/2} \hat{\mathbf{H}}^* \mathbf{S}_D^{1/2}$, as appropriate.

7 General recommendations

It is generally advisable to try a few different selectors and visually examine the resulting density estimates. One of the primary motivations for the author to write the `ks` package is to make visualising kernel density estimates easy. There are many different bandwidth selectors available in `ks`: they may now pose a problem of *too* much choice. The full bandwidth selectors will be better than their diagonal counterparts when the data have large mass oriented obliquely to the co-ordinate axes (as is the case for the `unicef` data). Amongst the full selectors, we advise against using the BCV selector. The LSCV selector is useful in some cases though its performance is known to be highly variable. The 2-stage plug-in and the SCV selectors can be viewed as generally recommended bandwidth selectors.

References

- Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–360.
- Duong, T. and Hazelton, M. L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, **15**, 17–30.
- Duong, T. and Hazelton, M. L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, **32**, 485–506.
- Hall, P., Marron, J. S., and Park, B. U. (1992). Smoothed cross-validation. *Probability Theory and Related Fields*, **92**, 1–20.
- Jones, M. C., Marron, J. S., and Park, B. U. (1991). A simple root n bandwidth selector. *The Annals of Statistics*, **19**, 1919–1932.
- Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics. Theory and Applications*, **9**, 65–78.
- Sain, S. R., Baggerly, K. A., and Scott, D. W. (1994). Cross-validation of multivariate densities. *Journal of the American Statistical Association*, **89**, 807–817.
- Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B. Methodological*, **53**, 683–690.

- Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag, New York.
- Wand, M. P. and Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, **9**, 97–116.