Nonparametric Density Estimation (Multidimension)

Härdle, Müller, Sperlich, Werwarz, 1995, *Nonparametric and Semiparametric Models, An Introduction*

Nonparametric kernel density estimation

Tine Buch-Kromann

February 19, 2007

Setup

One-dimensional estimation $\longrightarrow$ Multivariate estimation

Consider a d-dimensional data set with sample size n

$$\mathbf{X_i} = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{id} \end{pmatrix}, \qquad i = 1, ..., n.$$

Goal: Estimate the density f of $\mathbf{X} = (X_1, ..., X_d)^T$

$$f(\mathbf{x}) = f(x_1, ..., x_d)$$

Multivariate kernel density estimator

Kernel density estimator in d-dimensions

$$
\begin{aligned}
\hat{f}_h(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{X_i}}{h}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}\mathcal{K}\left(\frac{x_1-X_{i1}}{h},\ldots,\frac{x_d-X_{id}}{h}\right)
\end{aligned}
$$

where $\mathcal{K}$ is a multivariate kernel function with d arguments.

Note: h is the same for each components.

Multivariate kernel density estimator

Extension:

Bandwidths: $h = (h_1, ..., h_d)^T$

$$\hat{f}_h(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1 ... h_d} \mathcal{K}\left(\frac{x_1 - X_{i1}}{h_1}, ..., \frac{x_d - X_{id}}{h_d} \right)$$

Kernel function

What form should the multidim. kernel $\mathcal{K}(\mathbf{u}) = \mathcal{K}(u_1, ..., u_d)$ take?

Multiplicative kernel:

$$\mathcal{K}(\mathbf{u}) = K(u_1) \cdot ... \cdot K(u_d)$$

where K is a univariate kernel function.

$$\begin{aligned} \hat{f}_h(\mathbf{x}) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1 ... h_d} \mathcal{K}\left(\frac{x_1 - X_{i1}}{h_1}, ..., \frac{x_d - X_{id}}{h_d} \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{x_j - X_{ij}}{h_j} \right) \right\} \end{aligned}$$

Note: Contributions to the sum only in the cube:

$$X_{i1} \in [x_1 - h_1, x_1 + h_1), ..., X_{id} \in [x_d - h_d, x_d + h_d)$$

Kernel function

Spherical/radial-symmetric kernel:

$$\mathcal{K}(\mathbf{u}) \propto K(||\mathbf{u}||)$$

or

$$\mathcal{K}(\mathbf{u}) = \frac{K(||\mathbf{u}||)}{\int_{\mathbb{R}^d} K(||\mathbf{u}||)}$$

where $||\mathbf{u}|| = \sqrt{\mathbf{u}^T\mathbf{u}}$. (Exercise 3.13)

The multivariate Epanechnikov (spherical):

$$\mathcal{K}(\mathbf{u}) \propto (1 - \mathbf{u}^T\mathbf{u})1_{(\mathbf{u}^T\mathbf{u} \leq 1)}$$

The multivariate Epanechnikov (multiplicative):

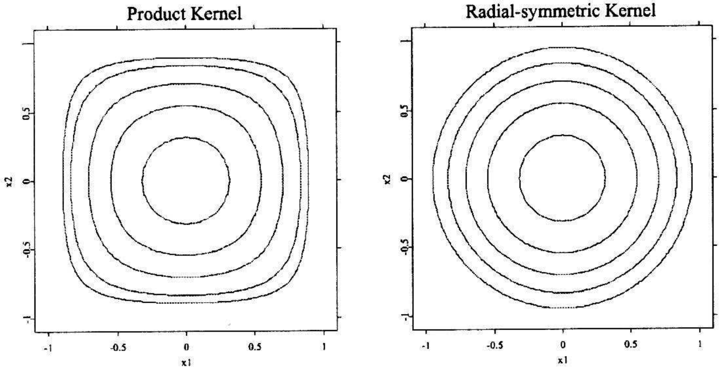
$$\mathcal{K}(\mathbf{u}) = \left(\frac{3}{4}\right)^d (1 - u_1^2)1_{(|u_1| \leq 1)} \dots (1 - u_d^2)1_{(|u_d| \leq 1)}$$

Kernel function

Epanechnikov kernel function

Equal bandwidth in each direction:

$$\mathbf{h} = (h_1, h_2)^T = (1, 1)^T$$

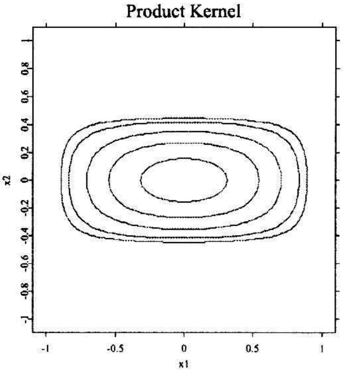


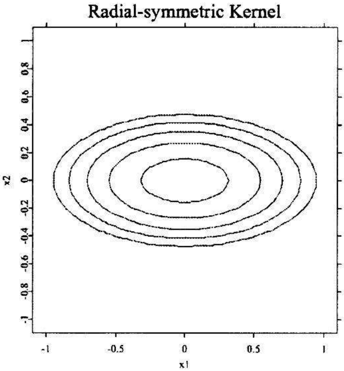
Kernel function

Epanechnikov kernel function

Different bandwidth in each direction:

$$\mathbf{h} = (h_1, h_2)^T = (1, 0.5)^T$$





Multivariate kernel density estimator

The general form for the multivariate density estimator with bandwidth matrix **H** (nonsingular)

$$
\begin{aligned}
\hat{f}_{\mathbf{H}}(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\det(\mathbf{H})}\mathcal{K}\left(\mathbf{H}^{-1}(\mathbf{x}-\mathbf{X_i})\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{\mathbf{H}}(\mathbf{x}-\mathbf{X_i})
\end{aligned}
$$

where $\mathcal{K}_{\mathbf{H}}(\cdot) = \frac{1}{\det(\mathbf{H})}\mathcal{K}(\mathbf{H}^{-1}\cdot)$

Multivariate kernel density estimator

The bandwidth matrix includes all simpler cases.

Equal bandwidth h:

$$\mathbf{H} = h\mathbf{I}_d$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix.

Different bandwidths $h_1, ..., h_d$:

$$\mathbf{H} = \mathrm{diag}(h_1, ..., h_d)$$

Multivariate kernel density estimator

What effect has the off-diagonal elements?

Rule-of-Thumb:
Use a bandwidth matrix proportional to $\hat{\Sigma}^{-\frac{1}{2}}$, where $\hat{\Sigma}$ is the covariance matrix of the data.

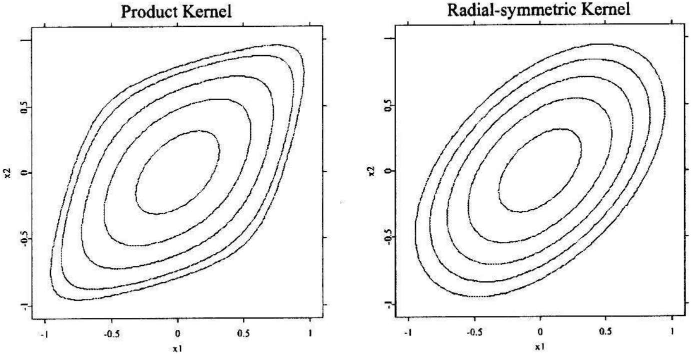
Such a bandwidth corresponds to a transformation of the data, so that they have an identity covariance matrix, ie. we can use bandwidths matrics to adjust for correlation between the components.

Kernel function

Epanechnikov kernel function

Bandwidth matrix:

$$\mathbf{H} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$



Properties of the kernel function

- $\mathcal{K}$ is a density function

$$\int_{\mathbb{R}^d} \mathcal{K}(\mathbf{u})\, d\mathbf{u} = 1 \quad \text{and} \quad \mathcal{K}(\mathbf{u}) \geq 0$$

- $\mathcal{K}$ is symmetric

$$\int_{\mathbb{R}^d} \mathbf{u}\mathcal{K}(\mathbf{u})\, d\mathbf{u} = 0_d$$

- $\mathcal{K}$ has a second moment (matrix)

$$\int_{\mathbb{R}^d} \mathbf{u}\mathbf{u}^T \mathcal{K}(\mathbf{u})\, d\mathbf{u} = \mu_2(\mathcal{K})\mathbf{I}_d$$

  where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix
- $\mathcal{K}$ has a kernel norm

$$||K||_2^2 = \int \mathcal{K}^2(\mathbf{u})\, d\mathbf{u}$$

Properties of the kernel function

$\mathcal{K}$ is a density function. Therefore is also $\hat{f}_{\mathbf{H}}$ a density function

$$\int \hat{f}_{\mathbf{H}}(\mathbf{x})\, d\mathbf{x} = 1$$

The estimate is consistent in any point $\mathbf{x}$

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) \xrightarrow{P} f(\mathbf{x})$$

Statistical Properties

Bias:

$$\mathbb{E}\left(\hat{f}_{\mathbf{H}}(\mathbf{x})\right) - f(x) \approx \frac{1}{2}\mu_2(\mathcal{K})\mathrm{tr}\{\mathbf{H}^T\mathcal{H}_f(x)\mathbf{H}\}$$

Variance:

$$\mathbb{V}\left(\hat{f}_{\mathbf{H}}(\mathbf{x})\right) \approx \frac{1}{n\det(\mathbf{H})}||\mathcal{K}||_2^2 f(\mathbf{x})$$

AMISE:

$$\mathrm{AMISE}(\mathbf{H}) = \frac{1}{4}\mu_2^2(\mathcal{K})\int \mathrm{tr}\{\mathbf{H}^T\mathcal{H}_f(x)\mathbf{H}\}^2\,d\mathbf{x} + \frac{1}{n\det(\mathbf{H})}||\mathcal{K}||_2^2$$

where $\mathcal{H}_f$ is the Hessian matrix and $||\mathcal{K}||_2^2$ is the $d-$dimensional squared L_2-norm af $\mathcal{K}$.

Special case

Univariate case:

For $d = 1$ we obtain $\mathbf{H} = h, \mathcal{K} = K, \mathcal{H}_f(x) = f''(x)$

Bias:

$$\begin{aligned} \mathbb{E}\left(\hat{f}_{\mathbf{H}}(\mathbf{x})\right) - f(x) &\approx \frac{1}{2}\mu_2(\mathcal{K})\mathrm{tr}\{\mathbf{H}^T\mathcal{H}_f(x)\mathbf{H}\} \\ &\approx \frac{1}{2}\mu_2(K)h^2 f''(x) \end{aligned}$$

Variance:

$$\begin{aligned} \mathbb{V}\left(\hat{f}_{\mathbf{H}}(\mathbf{x})\right) &\approx \frac{1}{n\,\det(\mathbf{H})}||\mathcal{K}||_2^2 f(\mathbf{x}) \\ &\approx \frac{1}{nh}||K||_2^2 f(x) \end{aligned}$$

Bandwidth selection

AMISE optimal bandwidth:

We have a bias-variance trade-off which is solved in the AMISE optimal bandwidth.

h is a scalar, $\mathbf{H} = h\mathbf{H}_0$ and $\det(\mathbf{H}_0) = 1$, then

$$\mathrm{AMISE}(\mathbf{H}) = \frac{1}{4} h^4 \mu_2^2(\mathcal{K}) \int \left[\mathrm{tr}\{\mathbf{H}_0^T \mathcal{H}_f(\mathbf{x}) \mathbf{H}_0\} \right]^2 d\mathbf{x} + \frac{1}{nh^d} ||K||_2^2$$

Then the optimal bandwidth and the optimal AMISE are

$$h_{opt} \sim n^{-1/(4+d)}, \qquad \mathrm{AMISE}(h_{opt}\mathbf{H}_0) \sim n^{-4/(4+d)}$$

Note: The multivariate density estimator has a slower rate of convergens compared to the univariate one.

$\mathbf{H} = h\mathbf{I}_d$ and fix sample size n: The AMISE optimal bandwidth larger in higher dimensions.

Bandwidth selection

Bandwidth selection:

- Plug-in method (rule-of-thumb, generalized Silvermann rule-of-thumb)
- Cross-validation method

Plug-in method

Idea: Optimize AMISE under the assumption that f is multivariate normal distribution $N_d(\mu, \mathbf{\Sigma})$ and $\mathcal{K}$ is a multivariate Gaussian, ie. $N_d(0, \mathbf{I})$, then

$$\mu_2(\mathcal{K}) = 1 \qquad ||\mathcal{K}||_2^2 = 2^{-d}\pi^{-d/2}$$

Then

$$\begin{aligned} & \int \mathrm{tr}\{\mathbf{H}^T \mathcal{H}_f(x) \mathbf{H}\}^2 \, d\mathbf{x} \\ = \quad & \frac{1}{2^{d+2}\pi^{d/2}\det(\mathbf{\Sigma})^{1/2}} [2\mathrm{tr}(\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H})^2 + \{\mathrm{tr}(\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H})\}^2] \end{aligned}$$

Bandwidth selection

Simple case:

$\mathbf{H} = \mathrm{diag}(h_1, ..., h_d)$ and $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_d)$, then

$$\tilde{h}_j = \underbrace{\left(\frac{4}{d+2}\right)^{1/(d+4)}}_{C} n^{-1/(d+4)} \sigma_j$$

Silverman's rule-of-thumb $(d = 1)$:

$$\hat{h}_{rot} = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5}$$

Bandwidth selection

Replace σ_j with $\hat{\sigma}_j$ and notice that C always is between 0.924 ($d = 11$) and 1.059 ($d = 1$):

Scott's rule

$$\hat{h}_j = n^{-1/(d+4)} \hat{\sigma}_j$$

It is not possibel to derive the rule-of-thumb in the general case, but it might be a good idea to choose the bandwidth matrix proportional to the covariance matrix.

Generalization of Scott's rule:

$$\hat{\mathbf{H}} = n^{-1/(d+4)} \hat{\mathbf{\Sigma}}^{1/2}$$

Bandwidth selection

Cross-validation:

$$
\begin{aligned}
\mathrm{ISE}(\mathbf{H}) &= \int \left(\hat{f}_{\mathbf{H}}(x) - f(\mathbf{x})\right)^2 d\mathbf{x} \\
&= \underbrace{\int \hat{f}^2_{\mathbf{H}}(\mathbf{x})\, d\mathbf{x}}_{\textit{Cal. from data}} + \underbrace{\int f^2(\mathbf{x})\, dx}_{\textit{Ignore}} - 2 \underbrace{\int \left(\hat{f}_{\mathbf{H}} f\right)(\mathbf{x})\, d\mathbf{x}}_{=\mathbb{E}\hat{f}_{\mathbf{H}}(\mathbf{X})}
\end{aligned}
$$

Estimate of the expectation

$$
\widehat{\mathbb{E}\hat{f}_{\mathbf{H}}(\mathbf{X})} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{\mathbf{H},-i}(\mathbf{X}_i)
$$

where the multivariate version of the leave-one-out estimator is

$$
\hat{f}_{\mathbf{H},-i}(\mathbf{x}) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \mathcal{K}_{\mathbf{H}}(\mathbf{X}_j - \mathbf{x})
$$

Bandwidth selection

Multivariate cross-validation criterion:

$$\begin{aligned}\mathrm{CV}(\mathbf{H}) &= \frac{1}{n^2 \det(\mathbf{H})} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{K} \star \mathcal{K} \left\{ \mathbf{H}^{-1}(\mathbf{X}_j - \mathbf{X}_i) \right\} \\ &\quad - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathcal{K}_{\mathbf{H}}(\mathbf{X}_j - \mathbf{X}_i)\end{aligned}$$

Note: The bandwidths is a $d \times d$ matrix $\mathbf{H}$ which means we have to minimize over $\frac{d(d+1)}{2}$ parameters.
Even if $\mathbf{H}$ is diagonal matrix, we have a d-dimensional optimization problem.

Canonical bandwidths

The canonical bandwidth of kernel j

$$\delta^j = \left\{ \frac{||\mathcal{K}||_2^2}{\mu_2(\mathcal{K}^{|})^2} \right\}^{1/(d+4)}$$

Therefore

$$\mathrm{AMISE}(\mathbf{H}^j, \mathcal{K}^j) = \mathrm{AMISE}(\mathbf{H}^i, \mathcal{K}^i)$$

where

$$\mathbf{H}^i = \frac{\delta^i}{\delta^j}\mathbf{H}^j$$

Canonical bandwidths

Example:

Adjust from Gaussian to Quartic product kernel

d	δ^G	δ^Q	δ^Q/δ^G
1	0.7764	2.0362	2.6226
2	0.6558	1.7100	2.6073
3	0.5814	1.5095	2.5964
4	0.5311	1.3747	2.5883
5	0.4951	1.2783	2.5820

Graphical representation

Example: Two-dimensions

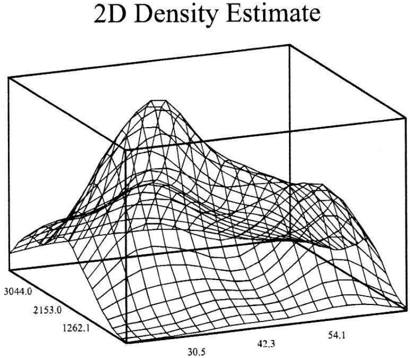
Est-West German migration intention in Spring 1991.

Explanatory variables: Age and household income

Two-dimensional nonparametric density estimate

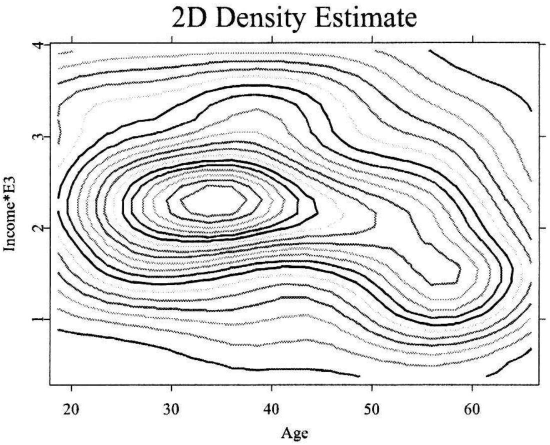
$$\hat{f}_{\mathbf{h}}(\mathbf{x}) = \hat{f}_{\mathbf{h}}(x_1, x_2)$$

where the bandwidth matrix $\mathbf{H} = \mathrm{diag}(\mathbf{h})$



Graphical representation

Contour plot



Graphical representation

Example: Three-dimensions

How can we display three- or even higher diemsional density estimates?

Hold one variable fix and plot the density function depending on the other variables.

For three-dimensions we have

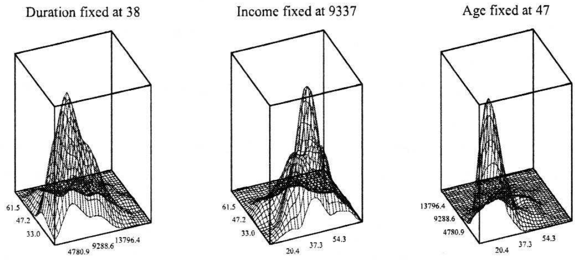
- x_1, x_2 vs. $\hat{f}_{\mathbf{h}}(x_1, x_2, x_3)$
- x_1, x_3 vs. $\hat{f}_{\mathbf{h}}(x_1, x_2, x_3)$
- x_2, x_3 vs. $\hat{f}_{\mathbf{h}}(x_1, x_2, x_3)$

Graphical representation

Example: Three-dimensions

Credit scoring sample.

Explanatory variables: Duration of the credit, household income and age.



Contour plot