Lec 9: Multivariate Kernel Density Estimation and related

Weiping Zhang

November 12, 2020

Multivariate KDE
    Application of KDE

KDE for Conditional density

Multivariate KDE

Consider a d-dimensional data set with sample size n:

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

Goal: Estimate the joint density f of $X = (X_1, \ldots, X_d)'$

$$f(x_1, \ldots, x_d)$$

- From our previous experience with the one-dimensional case we might consider adapting the kernel density estimator to the d-dimensional case, and write

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}\mathcal{K}(\frac{\mathbf{x}-\mathbf{X}_i}{h})$$

$$= \frac{1}{n}\sum_{i=1}^{d}\frac{1}{h^d}\mathcal{K}(\frac{x_1-X_{i1}}{h},\cdots,\frac{x_d-X_{id}}{h})$$

  where $\mathcal{K}$ is a multivariate kernel function with d arguments. Note: h is the same for each components.

- Extension: Bandwidths $h = (h_1, \ldots, h_d)'$

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

Kernel function

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

Multiplicative kernel

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

where K is a univariate kernel function.

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

Note: Contributions to the sum only in the cube:

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

Kernel function

Spherical/radial-symmetric kernel:

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

or

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

where $\|\mathbf{u}\| = \sqrt{\mathbf{u}'\mathbf{u}}$.

- The multivariate Epanechnikov (spherical):

$$\mathcal{K}(\mathbf{u}) \propto (1 - \mathbf{u}'\mathbf{u})I(\mathbf{u}'\mathbf{u} \leq 1)$$

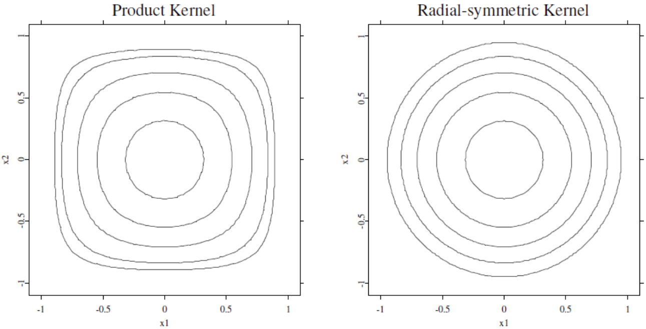
- The multivariate Epanechnikov (multiplicative):

$$\mathcal{K}(\mathbf{u}) = (\frac{3}{4})^d \prod_{j=1}^{d} (1 - u_j^2) I(|u_j| \leq 1)$$

Example

Epanechnikov kernel function Different bandwidth in each direction:

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

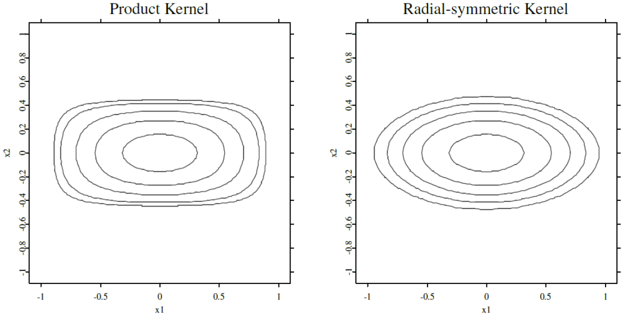


Example

Epanechnikov kernel function Different bandwidth in each direction:

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

$$\text{i.e., } \mathcal{K}_h(u) = \mathcal{K}(u_1/h_1, u_2/h_2)/(h_1 h_2)$$



Product Kernel

Radial-symmetric Kernel

Multivariate kernel density estimator

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

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

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

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 $h_1, \ldots, h_d$:
  $$\mathbf{H} = diag\{h_1, \ldots, h_d\}$$

What effect has the off-diagonal elements?

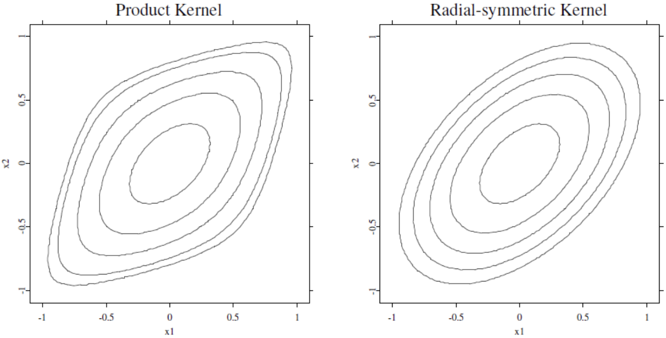
- **Rule-of-Thumb**: Use a bandwidth matrix proportional to $\hat{\Sigma}^{1/2}$, where $\hat{\Sigma}$ is the covariance matrix of the data.
- Such a bandwidth corresponds to a transformation of the data, so that they have an identity covariance matrix, ie. we can use bandwidths matrics to adjust for correlation between the components.

Example

Epanechnikov kernel function Bandwidth matrix:

$$\mathbf{H} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}^{1/2}$$

$$\text{i.e., } \mathcal{K}_{\mathbf{H}}(\mathbf{u}) = \mathcal{K}(\mathbf{H}^{-1}\mathbf{u})/det\mathbf{H}$$



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

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

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

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

- $\mathcal{K}$ has a kernel norm

$$\|\mathcal{K}\|^2 = \int_{\mathbb{R}^d} \mathcal{K}^2(\mathbf{u})d\mathbf{u}$$

- $\mathcal{K}$ 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 **x**:

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

  in probability.

Statistical Properties

- Bias:

$$E\hat{f}_{\mathbf{H}}(\mathbf{x}) - f(\mathbf{x}) \approx \frac{1}{2}\mu_2(\mathcal{K}) tr\{\mathbf{H}'\mathcal{H}_f(\mathbf{x})\mathbf{H}\}$$

- Variance:

$$Var(\hat{f}_{\mathbf{H}}(\mathbf{x})) \approx \frac{1}{n det \mathbf{H}}\|\mathcal{K}\|^2 f(\mathbf{x})$$

- AMISE:

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

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

Univariate case

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

- Bias:

$$E\hat{f}_{\mathbf{H}}(\mathbf{x}) - f(\mathbf{x}) \approx \frac{1}{2}\mu_2(\mathcal{K})h^2 f''(\mathbf{x})$$

- Variance:

$$Var(\hat{f}_{\mathbf{H}}(\mathbf{x})) \approx \frac{1}{nh}\|\mathcal{K}\|^2 f(\mathbf{x})$$

Proof

- We denote with ∇_f the gradient and with $\mathcal{H}_f$ the Hessian matrix of second partial derivatives of a function (here f). Then the Taylor expansion of $f(\cdot)$ around $\mathbf{x}$ is

$$f(\mathbf{x}+\mathbf{u}) = f(\mathbf{x}) + \mathbf{u}'\nabla_f(\mathbf{x}) + \frac{1}{2}\mathbf{u}'\mathcal{H}_f(\mathbf{x})\mathbf{u} + o(\mathbf{u}'\mathbf{u})$$

- This leads to the expectation

$$\begin{aligned}
E\hat{f}_{\mathbf{H}}(\mathbf{x}) &= \int \mathcal{K}_{\mathbf{H}}(\mathbf{u}-\mathbf{x})f(\mathbf{u})d\mathbf{u} \\
&= \int \mathcal{K}(\mathbf{u})f(\mathbf{x}+\mathbf{H}\mathbf{u})d\mathbf{u} \\
&\approx \int \mathcal{K}(\mathbf{u})\{f(\mathbf{x}) + \mathbf{u}'\mathbf{H}'\nabla_f(\mathbf{x}) + \frac{1}{2}\mathbf{u}'\mathbf{H}'\mathcal{H}_f(\mathbf{x})\mathbf{H}\mathbf{u}\}d\mathbf{u} \\
&= f(\mathbf{x}) + \frac{1}{2}\mu_2(\mathcal{K})tr\{\mathbf{H}'\mathcal{H}_f(\mathbf{x})\mathbf{H}\}
\end{aligned}$$

- Variance

$$\begin{aligned} Var(\hat{f}_{\mathbf{H}}(\mathbf{x})) &= \frac{1}{n}\int \{\mathcal{K}_{\mathbf{H}}(\mathbf{u}-\mathbf{x})\}^2 f(\mathbf{u})d\mathbf{u} - \frac{1}{n}\{E\hat{f}_{\mathbf{H}}(\mathbf{x})\}^2 \\ &\approx \frac{1}{ndet\mathbf{H}}\int \mathcal{K}^2(\mathbf{s})f(\mathbf{x}+\mathbf{H}\mathbf{s})d\mathbf{s} \\ &\approx \frac{1}{ndet\mathbf{H}}\int \mathcal{K}^2(\mathbf{s})\{f(\mathbf{x})+\mathbf{s}'\mathbf{H}'\nabla_f(\mathbf{x})\}d\mathbf{s} \\ &\approx \frac{1}{ndet\mathbf{H}}\|\mathcal{K}\|^2 f(\mathbf{x}) \end{aligned}$$

- Denote h a scalar, such that $\mathbf{H} = h\mathbf{H}_0$ and $det(\mathbf{H}_0) = 1$. Then AMISE can be written as

$$AMISE(\mathbf{H}) = \frac{h^4}{4}\mu_2^2(\mathcal{K})\int tr\{\mathbf{H}_0'\mathcal{H}_f(\mathbf{x})\mathbf{H}_0\}^2 d\mathbf{x} + \frac{1}{nh^d}\|\mathcal{K}\|^2$$

Asymptotic Normality

The kernel estimator is the sample average. We can therefore apply the central limit theorem. Thus for multiplicative kernel with $\mathbf{H} = diag(h_1, \ldots, h_q)$, we have

Theorem

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d $d-$vectors with its pdf f having three-times bounded continuous derivatives. Let $\mathbf{x}$ be an interior point of the support of $\mathbf{X}$. If, as $n \to \infty, h_s \to 0$ for all $s = 1, \ldots, d$, $nh_1 \cdots h_d \to \infty$, and $(nh_1 \cdots h_d)(\sum_{s=1}^{d} h_s^4)^2 \to 0$, then

$$\sqrt{nh_1 \cdots h_d}\left[\hat{f}_H(\mathbf{x}) - f(\mathbf{x}) - \frac{\kappa_{21}}{2}\sum_{s=1}^{d} h_s^2 f_{ss}(\mathbf{x})\right] \rightsquigarrow N(0, \kappa_{02}^d f(\mathbf{x}))$$

Proof

since

$$\mu_2(\mathcal{K})tr\{\mathbf{H}'\mathcal{H}_f(\mathbf{x})\mathbf{H}\} = \kappa_{21}\sum_{s=1}^{d} h_s^2 f_{ss}, \quad \|\mathcal{K}\|^2 = \kappa_{02}^d$$

we have

$$\begin{aligned}
&\sqrt{nh_1\cdots h_d}\Big[\hat{f}_H(\mathbf{x}) - f(\mathbf{x}) - \frac{\kappa_{21}}{2}\sum_{s=1}^{d} h_s^2 f_{ss}(\mathbf{x})\Big] \\
&\qquad = \sqrt{nh_1\cdots h_d}\Big[\hat{f}_H(\mathbf{x}) - E\hat{f}_H(\mathbf{x})\Big] \\
&\qquad + \sqrt{nh_1\cdots h_d}\Big[E\hat{f}_H(\mathbf{x}) - f(\mathbf{x}) - \frac{\kappa_{21}}{2}\sum_{s=1}^{d} h_s^2 f_{ss}(\mathbf{x})\Big] \\
&\qquad = \sqrt{nh_1\cdots h_d}\Big[\hat{f}_H(\mathbf{x}) - E\hat{f}_H(\mathbf{x})\Big] \\
&\qquad + o_p(\sqrt{nh_1\cdots h_d}\sum_{s=1}^{d} h_s^2)
\end{aligned}$$

$$= \sum_{i=1}^{n} (nh_1 \cdots h_d)^{-1/2} \Big[K(\frac{\mathbf{X}_i - \mathbf{x}}{h}) - EK(\frac{\mathbf{X}_i - \mathbf{x}}{h}) \Big] + o_p(1)$$

$$= \sum_{i=1}^{n} Z_{n,i} + o_p(1) \rightsquigarrow N(0, \kappa_{02}^d f(\mathbf{x}))$$

where

$$K(\frac{\mathbf{X}_i - \mathbf{x}}{h}) = K(\frac{X_{i1} - x_1}{h_1}) \cdots K(\frac{X_{id} - x_d}{h_d})$$

and

$$Z_{n,i} = (nh_1 \cdots h_d)^{-1/2} \Big[K(\frac{\mathbf{X}_i - \mathbf{x}}{h}) - EK(\frac{\mathbf{X}_i - \mathbf{x}}{h}) \Big]$$

Optimal bandwidth

- If we only allow changes in h the optimal orders for the smoothing parameter h and AMISE are

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

- 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.

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

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

  Then

  $$\int tr\{\mathbf{H}'\mathcal{H}_f(\mathbf{x})\mathbf{H}\}^2 d\mathbf{x}$$
  $$= \frac{1}{2^{d+2}\pi^{d/2}det\Sigma^{1/2}}[2tr(\mathbf{H}'\Sigma^{-1}\mathbf{H})^2 + \{tr(\mathbf{H}'\Sigma^{-1}\mathbf{H}\}^2]$$

  simple case: $\mathbf{H} = diag(h_1, \ldots, h_d)$ and $\Sigma = diag(\sigma_1^2, \ldots, \sigma_d^2)$, then

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

  Silverman's rule-of-thumb $(d = 1)$: $\hat{h}_{opt} = (4/3)^{1/5}\hat{\sigma}n^{-1/5}$

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:

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

It is not possible 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:

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

- **Cross-validation method**

$$
\begin{aligned}
ISE(\mathbf{H}) &= \int (\hat{f}_{\mathbf{H}}(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x} \\
&= \int \hat{f}_{\mathbf{H}}(\mathbf{x})^2 d\mathbf{x} - 2\int \hat{f}_{\mathbf{H}}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} + \int f^2(\mathbf{x}) d\mathbf{x}
\end{aligned}
$$

Estimate of the expectation

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

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

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

Multivariate cross-validation criterion:

$$CV(\mathbf{H}) = \frac{1}{n^2 det\mathbf{H}} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{K} * \mathcal{K}\{\mathbf{H}^{-1}(\mathbf{X}_j - \mathbf{X}_i)\}$$

$$- \frac{2}{n(n-1)} \sum_{j=1, j \neq i}^{n} \mathcal{K}_{\mathbf{H}}(\mathbf{X}_j - \mathbf{X}_i)$$

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

Canonical bandwidth

The canonical bandwidth of kernel j:

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

Therefore,

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

where

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

Example

Adjust from Gaussian to Quartic product kernel

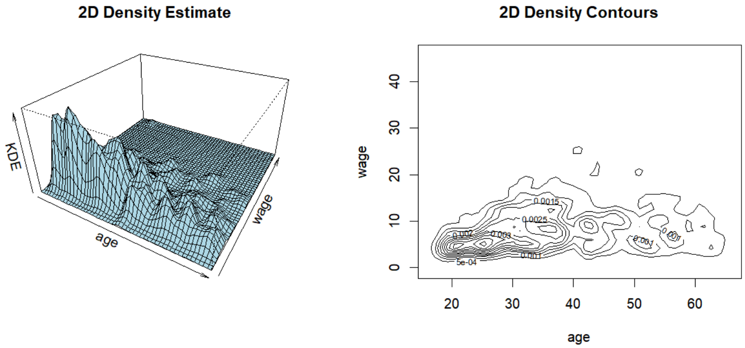
d	δ^G	δ^Q	δ^Q/δ^G
1	0.7764	2.0362	2.6226
2	0.6558	1.7100	2.6073
3	0.5814	1.5095	2.5964
4	0.5311	1.3747	2.5883
5	0.4951	1.2783	2.5820

Graphical representation

Example: Two-dimensions

East-West German migration intention in Spring 1991.

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = \hat{f}_{\mathbf{H}}(x_1, x_2), \quad \mathbf{H} = diag(h_1, h_2)$$



http://www.marlenemueller.de/nspm/SPMdensity2D.R

Example: Three-dimensions

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

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

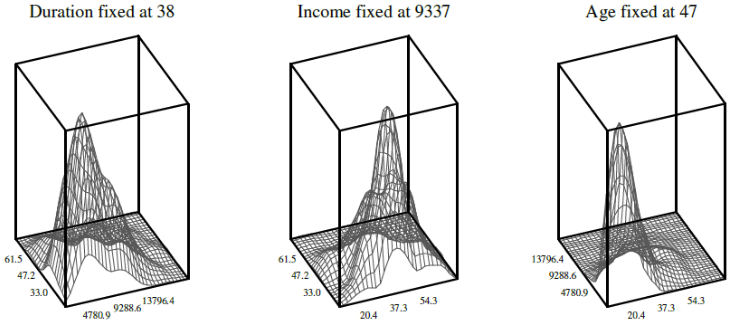
For three-dimensions we have

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

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

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

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



Duration fixed at 38

Income fixed at 9337

Age fixed at 47

Density level-set estimation

- The **level set of the density** f at level $c \geq 0$ is defined as $\mathcal{L}(f;c) := \{\mathbf{x} \in \mathbb{R}^p : f(\mathbf{x}) \geq c\}$
- An estimation of $\mathcal{L}(f;c)$ is useful for visualizing the highest density regions, which give a concise view of the most likely values for $\mathbf{X}$.
- In addition, this is done without the need of restricting to connected sets (such as, e.g, intervals in $\mathbb{R}$) and by allowing to determine the approximate probability contained in them.
- Level sets also are useful for estimating the support of $\mathbf{X}$ and for detecting multivariate outliers without the need of employing the Mahalanobis distance.

- The estimation of the level set, useful for determining high density regions, can be straightforwardly done by plugging-in the kde of f and considering

$$\mathcal{L}(\hat{f}(\cdot;\mathbf{H});c) = \left\{\mathbf{x} \in \mathbb{R}^p : \hat{f}(\mathbf{x};\mathbf{H}) \geq c\right\}$$

- Obtaining the representation of $\mathcal{L}(\hat{f}(;\mathbf{H});c)$ in practice involves the consideration of a grid in $\mathbb{R}^p$ in order to evaluate the condition $\hat{f}(\mathbf{x};\mathbf{H}) \geq c$ and determine the region of $\mathbb{R}^p$ in which it is satisfied.
- The level c in $\mathcal{L}(f;c)$ may be difficult to interpret. It is usually considered the largest c_α such that

$$\int_{\mathcal{L}(f;c_\alpha)} f(\mathbf{x})\mathrm{d}\mathbf{x} \geq 1-\alpha, \quad \alpha \in (0,1)$$

$\mathcal{L}(f,c_\alpha)$ is the smallest region of R^p that contains at least $1-\alpha$ of the probability of X.

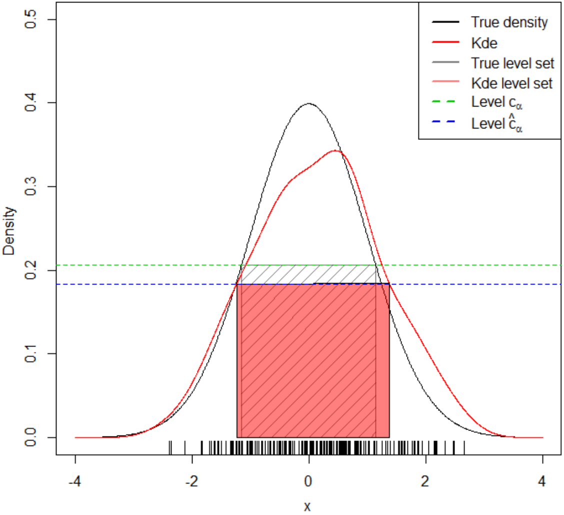


Figure 1: Level set $\mathcal{L}\left(f; c_\alpha\right)$ and its estimation by $\mathcal{L}\left(\hat{f}(\cdot; h); \hat{c}_\alpha\right)$ for $\alpha = 0.25$ and $f = \phi$

Clustering–kernel mean shift clustering

- The general goal of clustering, find clusters of data with low within-cluster variation, can be regarded as the task of determining data-rich regions on the sample. From the density perspective, data-rich regions have a precise definition: modes. Therefore, modes are going to be crucial to define population clusters in the sense introduced by Chacón (2015).
- Given the random vector $\mathbf{X}$ in $\mathbb{R}^p$ with pdf f, we denote the modes of f as $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m$. These are the local maxima of f, i.e. $\mathrm{D}f\left(\boldsymbol{\xi}_j\right) = \mathbf{0}, j = 1, \ldots, m$. Intuitively, we can think about the population clusters as the regions of $\mathbb{R}^p$ that are "associated" with each of the modes of f.

- This "association" can be visualized, for example, by a gravitational analogy: if $\xi_1, \ldots, \xi_m$ denote fixed planets with equal mass distributed on the space, then the population clusters can be thought as the regions that determine the domains of attraction of each planet for an object x in the space that has zero initial speed. If x is attracted by ξ_1, then x belongs to the cluster defined by ξ_1. In our setting, the role of the gravity attraction is played by the gradient of the density $f, Df : \mathbb{R}^p \longrightarrow \mathbb{R}^p$, which is a vector field over $\mathbb{R}^p$.

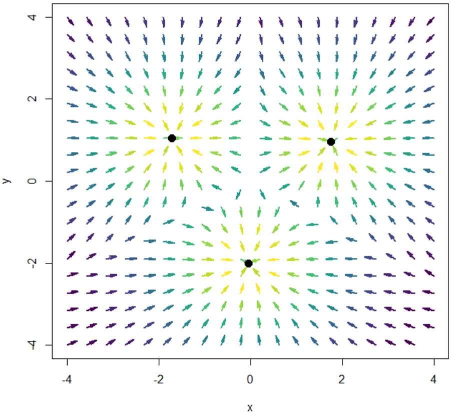


Figure 2: Sketch of the gravity vector field associated. The vector field is computed as the gradient of a mixture of three bivariate normals centered at the black points and whose covariance matrices are $\frac{1}{2}I_2$. The direction of the arrows denotes the direction of the gravity field, and their color the strength of the gravity force.

- The previous idea can be mathematically formalized as follows. We seek to partition in a collection of disjoint subsets $W_{+}^{s}\left(\xi_{1}\right), \ldots, W_{+}^{s}\left(\xi_{m}\right)$ defined[1] as

$$W_{+}^{s}(\xi) := \left\{ \mathbf{x} \in \mathbb{R}^{p} : \lim_{t \to \infty} \phi_{\mathbf{x}}(t) = \xi \right\}$$

  where $\phi_{\mathrm{x}} : \mathbb{R} \longrightarrow \mathbb{R}^{p}$ is a curve in $\mathbb{R}^{p}$ parametrized by $t \in \mathbb{R}$ that satisfies the following Ordinal Differential Equation (ODE) :

$$\frac{\mathrm{d}}{\mathrm{d}t} \phi_{\mathrm{x}}(t) = \mathrm{D} f\left(\phi_{\mathrm{x}}(t)\right), \quad \phi_{\mathrm{x}}(0) = \mathrm{x}$$

- This ODE admits a clear interpretation; the flow curve ϕ_{x} is the path that, originated at x, describes x when reaching ξ_{j} through the direction of maximum ascension.

---

[1] The superscript s stands for stable manifold and the subscript $+$ emphasizes that the positive gradient is considered.

- The ODE can be solved through different numerical schemes. For example, observing that $\frac{d}{dt}\phi_{\mathbf{x}}(t) = \lim_{h\to 0} \frac{\phi_{\mathbf{x}}(t+h)-\phi_{\mathbf{x}}(t)}{h}$, the Euler method considers the approximation

$$\phi_{\mathbf{x}}(t+h) \approx \phi_{\mathbf{x}}(t) + hDf\left(\phi_{\mathbf{x}}(t)\right) \qquad \text{if } h \approx 0.$$

- which motivates the iterative scheme

$$\begin{cases} \mathbf{x}_{t+1} = \mathbf{x}_t + h\mathrm{D}f\left(\mathbf{x}_t\right), \quad t = 0, \ldots, N \\ \mathbf{x}_0 = \mathbf{x} \end{cases}$$

  for a step $h > 0$ and a number of maximum iterations N.

- It is clear how to assign a point $\mathbf{x}$ to a population cluster: compute its associated flow curve and assign the j -th cluster label if $\mathbf{x} \in W_{+}^{s}\left(\boldsymbol{\xi}_j\right)$. In application, this is trivial by replacing f by its kde $\hat{f}(\cdot; \mathbf{H})$.

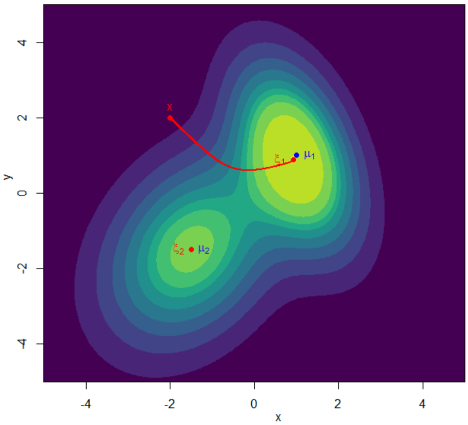


Figure 3: The curve ϕ_x computed by the Euler method, whose path solution is shown in the black curve. The population density is the mixture of bivariate normals $w\phi_{\Sigma_1}(\cdot - \boldsymbol{\mu}_1) + (1-w)\phi_{\Sigma_2}(\cdot - \boldsymbol{\mu}_2)$ where $\boldsymbol{\mu}_1 = (1,1), \boldsymbol{\mu}_2 = (-1.5,-1.5), \boldsymbol{\Sigma}_1 = (1,-0.75;-0.75,3), \boldsymbol{\Sigma}_2 = (2,0.75;0.75,3)$, and $w = 0.45$. The component means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are shown in blue, whereas the two modes $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ of the density are represented in red.

- In application, we consider

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \hat{\boldsymbol{\eta}}\left(\mathbf{x}_t; \mathbf{H}\right), \quad \hat{\boldsymbol{\eta}}(\mathbf{x}; \mathbf{H}) := \frac{\mathbf{H}\mathrm{D}\hat{f}\left(\mathbf{x}_t; \mathbf{H}\right)}{\hat{f}\left(\mathbf{x}_t; \mathbf{H}\right)} \qquad (1)$$

- This is based on the following two tweaks.
    - The first tweak boosts the travel through low density regions by adapting the step size taken at $\mathbf{x}_{t+1}$ by the density at $\mathbf{x}_t$. This amounts to considering the normalized gradient $\boldsymbol{\eta}(\mathbf{x}) = Df(\mathbf{x})/f(\mathbf{x})$.
    - The second tweak multiplies a matrix $\mathbf{A}$ to $\boldsymbol{\eta}(\mathbf{x})$ and allow for more generality. This apparent innocuous change gives a convenient choice for $\mathbf{A}$ and hence for the step in Euler's method: $\mathbf{A} = \mathbf{H}$.

The recipe for clustering a sample $\mathbf{X}_1, \ldots \mathbf{X}_n$ is now simple:

1. Select a "suitable" bandwidth $\hat{\mathbf{H}}$.
2. For each element $\mathbf{X}_i$, iterate the recurrence relation (1) until "convergence" to a given $\mathbf{y}_i, \quad i = 1, \ldots, n$
3. Find the set of "unique" end points $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m\}$ (the modes) among $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$
4. Label $\mathbf{X}_i$ as j if it is associated to the j -th mode $\boldsymbol{\xi}_j$.

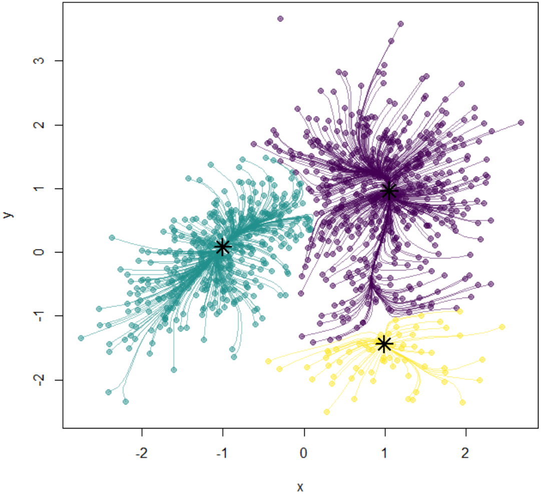


Figure 4: A simulated example for which the population clusters are known

KDE for Conditional density

- The conditional density of y given $\mathbf{X} = \mathbf{x}$ is $f(y|\mathbf{x}) = f(y, \mathbf{x})/f(\mathbf{x})$. An natural estimator is

$$\hat{f}(y|\mathbf{x}) = \frac{\hat{f}(y, \mathbf{x})}{\hat{f}(\mathbf{x})} = \frac{\sum_{i=1}^{n} \mathcal{K}(H^{-1}(\mathbf{X}_i - \mathbf{x}))K_{h_0}(y_i - y)}{\sum_{i=1}^{n} \mathcal{K}(H^{-1}(\mathbf{X}_i - \mathbf{x}))}$$

  where $H = diag(h_1, \ldots, h_d)$ and $y \in R, \mathbf{x} \in R^d$.

- Notice that the conditional expectation of $Z_i = K_{h_0}(y - Y_i)$ given $\mathbf{X}_i = \mathbf{x}$ is

$$\begin{aligned} E(Z_i|\mathbf{X}_i = \mathbf{x}) &= \int \frac{1}{h_0} K(\frac{v - y}{h_0}) f(v|\mathbf{x}) dv \\ &= \int K(u) f(y - uh_0|\mathbf{x}) du \\ &\approx f(y|\mathbf{x}) + \frac{h_0^2 \kappa_{21}}{2} \frac{\partial^2}{\partial y^2} f(y|\mathbf{x}). \end{aligned}$$

- We can view conditional density estimation as a regression.

Bias:

$$\begin{aligned} E\hat{f}(y|\mathbf{x}) &= E(Z_i|\mathbf{X}_i = \mathbf{x}) + \kappa_{21}\sum_{j=1}^{d} h_j^2 B_j(y|\mathbf{x}) \\ &= f(y|\mathbf{x}) + \kappa_{21}\sum_{j=0}^{d} h_j^2 B_j(y|\mathbf{x}) \end{aligned}$$

where

$$B_0(y|\mathbf{x}) = \frac{1}{2}\frac{\partial^2}{\partial y^2} f(y|\mathbf{x}).$$

$$= B_j(y|\mathbf{x}) = \frac{1}{2}\frac{\partial^2}{\partial x_j^2} f(y|\mathbf{x}) + f(\mathbf{x})^{-1}\frac{\partial}{\partial x_j} f(y|\mathbf{x})\frac{\partial}{\partial x_j} f(\mathbf{x}), j > 0$$

Variance Notice that

$$Var(\hat{f}(y|\mathbf{x})) \approx \frac{\kappa_{02}^d}{nh_1 \cdots h_d f(\mathbf{x})} Var(Z_i|\mathbf{X}_i = \mathbf{x})$$

we calculate that

$$\begin{aligned} Var(Z_i|\mathbf{X}_i = \mathbf{x}) &= E(Z_i^2|\mathbf{X}_i = \mathbf{x}) - (E(Z_i|\mathbf{X}_i = \mathbf{x}))^2 \\ &\approx \frac{1}{h_0^2} \int K^2(\frac{v-y}{h_0}) f(v|\mathbf{x}) dv \\ &= \frac{1}{h_0} \int K^2(u) f(y-uh|\mathbf{x}) du \approx \frac{\kappa_{02} f(y|\mathbf{x})}{h_0} \end{aligned}$$

Substituting this into the expression for the estimation variance, we have

$$Var(\hat{f}(y|\mathbf{x})) \approx \frac{\kappa_{02}^{d+1} f(y|\mathbf{x})}{nh_0 h_1 \cdots h_d f(\mathbf{x})}$$

MSE:

$$AMSE(\hat{f}(y|\mathbf{x})) = \kappa_{02}^2\Big(\sum_{j=0}^{d} h_j^2 B_j(y|\mathbf{x})\Big)^2 + \frac{\kappa_{02}^{d+1} f(y|\mathbf{x})}{nh_0h_1\cdots h_d f(\mathbf{x})}$$

Let $h_0 = h_1 = \cdots = h_d = h$, then

$$AMSE(\hat{f}(y|\mathbf{x})) \sim h^4 + \frac{1}{nh^{d+1}}$$

with optimal solution

$$h \sim n^{-1/(d+5)}$$

This is the same rate as for multivariate density estimation (estimation of the joint density $f(y, \mathbf{x})$).

Cross-validation

For an estimator $\hat{f}(y|\mathbf{x})$ of $f(y|\mathbf{x})$ define the weighted integrated squared error

$$\begin{aligned}
I(h) &= \iint \left(\hat{f}(y|\mathbf{x}) - f(y|\mathbf{x})\right)^2 M(\mathbf{x}) f(\mathbf{x}) dy d\mathbf{x} \\
&= \iint \hat{f}^2(y|\mathbf{x}) M(\mathbf{x}) f(\mathbf{x}) dy d\mathbf{x} - 2 \iint \hat{f}(y|\mathbf{x}) M(\mathbf{x}) f(y|\mathbf{x}) dy d\mathbf{x} \\
&\quad + \iint f^2(y|\mathbf{x}) M(\mathbf{x}) f(\mathbf{x}) dy d\mathbf{x} \\
&= E\Big(\int \hat{f}^2(y|\mathbf{X}_i) M(\mathbf{X}_i) \Big) - 2E\Big(\hat{f}(y_i|\mathbf{X}_i) M(\mathbf{X}_i) \Big) \\
&\quad + E\Big(\int f^2(y|\mathbf{X}_i) M(\mathbf{X}_i) dy \Big) \\
&= I_1(h) - 2I_2(h) + I_3
\end{aligned}$$

Let $\hat{f}_{-i}(y|\mathbf{X}_i)$ denote the estimator $\hat{f}(y|\mathbf{x})$ at $\mathbf{x} = \mathbf{X}_i$ with observation i omitted. that is

$$\hat{f}_{-i}(y|\mathbf{X}_i) = \frac{\sum_{j \neq i}^{n} \mathcal{K}(H^{-1}(\mathbf{X}_j - \mathbf{X}_i))K_{h_0}(y_j - y)}{\sum_{j \neq i}^{n} \mathcal{K}(H^{-1}(\mathbf{X}_j - \mathbf{X}_i))}$$

The cross-validation estimators of I_1 and I_2 are

$$\hat{I}_1(h) = \frac{1}{n} \sum_{i=1}^{n} M(\mathbf{X}_i) \int \hat{f}_{-i}^2(y|\mathbf{X}_i)dy$$

$$\hat{I}_2(h) = \frac{1}{n} \sum_{i=1}^{n} M(\mathbf{X}_i) \hat{f}_{-i}(Y_i|\mathbf{X}_i)dy$$

The cross-validation criterion is

$$CV(h) = \hat{I}_1(h) - 2\hat{I}_2(h)$$

Conditional Distribution Estimation

The conditional distribution (CDF) of Y given $\mathbf{X} = x$ is

$$F(y|\mathbf{x}) = E(I(Y \leq y)|\mathbf{X} = \mathbf{x}) = \int I(u \leq y) f(u|\mathbf{x}) du$$

Thus the CDF is a regression, and can be estimated using regression methods. An natural estimator is

$$\hat{F}(y|\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathcal{K}(H^{-1}(\mathbf{X}_i - \mathbf{x})) G(\frac{y_i - y}{h_0})}{\sum_{i=1}^{n} \mathcal{K}(H^{-1}(\mathbf{X}_i - \mathbf{x}))}$$

- Bias is

$$Bias(\hat{F}(y|\mathbf{x})) \approx \kappa_{02} \sum_{j=0}^{d} h_j^2 B_j(y|\mathbf{x})$$

where for $j \geq 1$ the $B_j(y|\mathbf{x})$ are the same as before, and for $j = 0$,

$$B_0(y|\mathbf{x}) = \frac{1}{2} \frac{\partial^2}{\partial y^2} F(y|\mathbf{x})$$

- Variance is

$$Var(\hat{F}(y|\mathbf{x})) \approx \frac{\kappa_{02}^d [F(y|\mathbf{x})(1 - F(y|\mathbf{x})) - h_0 \alpha(k) f(y|\mathbf{x})]}{f(\mathbf{x}) n det(H)}$$

- In sum, the MSE is

$$MSE(\hat{F}(y|\mathbf{x})) = (Bias(\hat{F}(y|\mathbf{x})))^2 + Var(\hat{F}(y|\mathbf{x}))$$

Bandwidth selection via cross-validation

Define the CV criterion as

$$CV(y,h) = \frac{1}{n}\sum_{i=1}^{n}\Big(I(y_i \le y) - \hat{F}_{-i}(y|\mathbf{X}_i)\Big)^2 M(\mathbf{X}_i)$$

and

$$CV(h) = \int CV(y,h)dy$$

where $h = (h_0, h_1, \ldots, h_d)$, $\hat{F}_{-i}$ is is the smooth leave-one-out estimator of F.