Kernel Density Estimation

An Introduction

Justus H. Piater, Universität Innsbruck

Abstract

Probability density functions (PDFs) play a fundamental role in many applications of statistics and probabilistic inference. In most practical cases, these PDFs are not available a priori but must be estimated from observations via density-estimation techniques. This presentation gives a basic introduction to kernel density estimation (KDE), a class of techniques for estimating PDFs from samples without imposing parametric constraints on the general shape of the PDF. It introduces some of the most commonly used techniques for univariate, Euclidean data, presents extensions to more general problems of high practical relevance, and discusses KDE in the context of some basic applications.

Table of Contents

1 1 1 5 7 8 8 13 14 14 17 18 21 21 22 25 27 27 30 30 31 31 32 34 35 36 37 38 39 39 40 49 49 54 56 56 62 63 65 65 67 75 76 79 82 83 84 84 84 90 91 92 93 93 103 103 111 112 119 122 123 126 129 134 144 147 153 163 167 175 228 231 235 242 256 256 257 311 314 322 369 415 573 609 640 737 [Copas and Fryer 1980, Silverman 1986]

1. Preface

1.1. Overview

2. Densities and their Estimation

1 1 1 5 7 8 8 13 14 14 17 18 21 21 22 25 27 27 30 30 31 31 32 34 35 36 37 38 39 39 40 49 49 54 56 56 62 63 65 65 67 75 76 79 82 83 84 84 84 90 91 92 93 93 103 103 111 112 119 122 123 126 129 134 144 147 153 163 167 175 228 231 235 242 256 256 257 311 314 322 369 415 573 609 640 737 [Copas and Fryer 1980, Silverman 1986]

2.1. Can you find structure in these data?

1 1 1 5 7 8 8 13 14 14 17 18 21 21 22 25 27 27 30 30 31 31 32 34 35 36 37 38 39 39 40 49 49 54 56 56 62 63 65 65 67 75 76 79 82 83 84 84 84 90 91 92 93 93 103 103 111 112 119 122 123 126 129 134 144 147 153 163 167 175 228 231 235 242 256 256 257 311 314 322 369 415 573 609 640 737

Lengths of 86 spells of psychiatric treatment undergone by patients used as controls in a study of suicide risks [Copas and Fryer 1980]

2.2. Probability Density Functions

We can assign probabilities to subsets of a continuous sample space Ω by integrating probability density functions (PDFs) p over them:

$$\begin{aligned} P(\text{event}) &= \int_{\text{event}} p(u)\,\mathrm{d}u \\ p(u_1)\,\mathrm{d}u &= P\left(\{u \in [u_1, u_1 + \mathrm{d}u]\}\right) \end{aligned}$$

A PDF p is nonnegative, but can exceed unity, and $P(\Omega) = \int_{\Omega} p(u)\,\mathrm{d}u = 1$.

2.3. Parametric Estimation of Densities

If we have a parametric, generative model of the PDF, we can estimate its parameters from samples.

Example 1. Univariate Normal (Gaussian) Distribution

$$\begin{aligned} p(x; \mu, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ \hat{\mu} &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ \hat{\sigma}^2 &= \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 \end{aligned}$$

2.4. Parametric Estimation of Densities

Example 2. Mixture of Univariate Normal Distributions

$$p(x; \mu_1, \dots, \mu_K, \sigma_1, \dots, \sigma_K, w_1, \dots, w_{K-1}) = \sum_{k=1}^{K} w_k p(x; \mu_k, \sigma_k)$$

with mixture proportions w_k the prior probabilities that a given observation was produced by mixture component k, and $\sum_{k=1}^{K} w_k = 1$:

Expectation-Maximization, gradient descent over the parameters, …

2.5. Nonparametric Estimation of Densities

Without a generative model, all we can use to estimate $\hat{p}(x)$ is our available sample $x_1, \dots, x_N$.

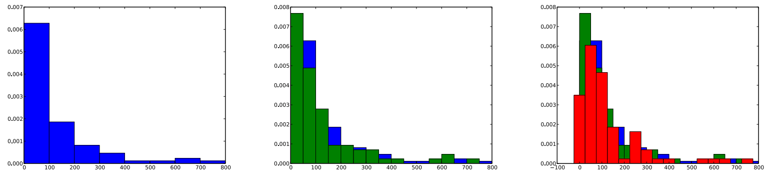
Uses:

- Faithfully model arbitrary distributions from finite samples
- Intuitive presentation and exploration
- Identify multimodality
- Resample

2.6. Histogram

Bin the data into intervals $[x_0 + mh, x_0 + (m+1)h)$ of width h, for $m \in \mathbb{Z}$:

$$\hat{p}(x) = \frac{1}{Nh} |\{x_i \text{ in same bin as } x\}|$$



Issues:

- Sensitivity to x_0, h, and (for multivariate data) the coordinate directions
- Step function

Generalized histogram:

$$\hat{p}(x) = \frac{1}{N} \frac{|\{x_i \text{ in same bin as } x\}|}{\text{width of bin containing } x}$$

2.7. Naive Estimator

From the definition of a PDF: $p(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$

Thus, we can further generalize our histogram:

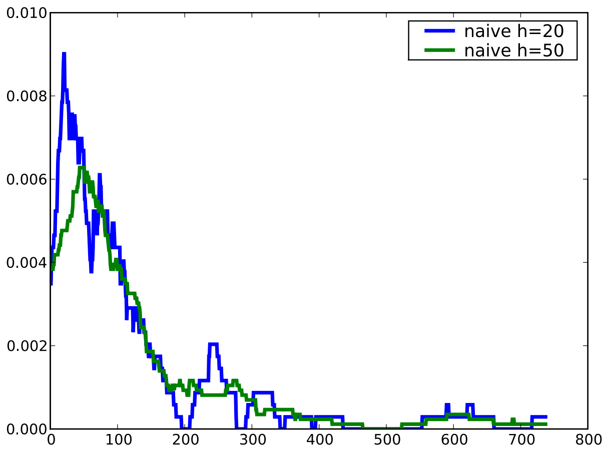
$$\begin{aligned} \hat{p}(x) &= \frac{1}{2Nh} |\{x_i \text{ falling within } (x-h, x+h)\}| \\ &= \frac{1}{Nh} \sum_{i=1}^{N} w\left(\frac{x - x_i}{h}\right) \\ w(u) &= \begin{cases} \frac{1}{2} & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

How many observations fall into a fixed-with box centered at x?

Issue:

- Step function

2.8. Naive Estimators: Illustration



3. Basic Estimators for Univariate KDE

3.1. Kernel Estimator

Replace the weight function w by a kernel function K with $\int_{-\infty}^{\infty} K(x)\mathrm{d}x = 1$.

Then, by analogy with the naive estimator,

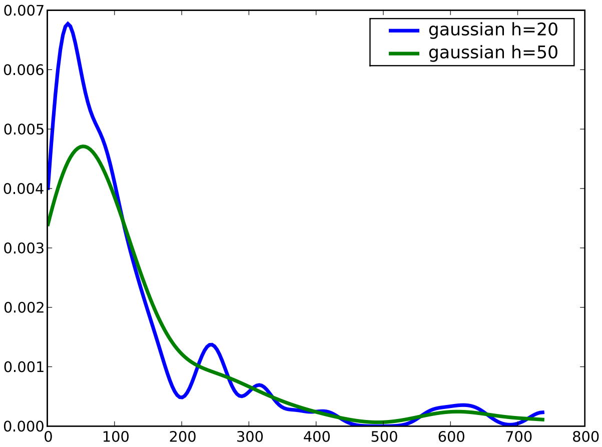
$$\hat{p}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

with *bandwidth* or *smoothing parameter* h.

Properties:

- If K is nonnegative, then it is a PDF, and so is $\hat{p}$.
- $\hat{p}$ inherits all continuity and differentiability properties from K.
- The fixed bandwidth may oversmooth in some regions and undersmooth in others.

3.2. Kernel Estimators: Illustration



3.3. Nearest Neighbor Estimator

Complementary to the naive estimator, *what is the smallest box containing a given number of observations*?

Define $d_1(x) = \|x - x_{n_1}\| \leq \cdots \leq d_N(x) = \|x - x_{n_N}\|$. Then, the ***k*th nearest neighbor density estimate** is defined by

$$\hat{p}(x) = \frac{k}{2Nd_k(x)}.$$

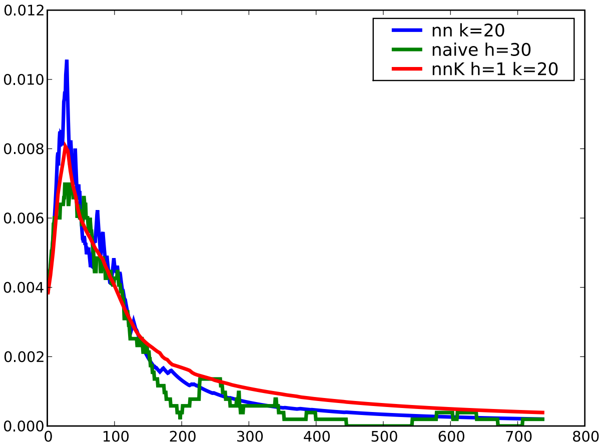
Properties:

- *Bandwidth depends on the density of observations around the query value* ***x***.
- Derivative full of discontinuities
- $\int_{-\infty}^{\infty} \hat{p}(x)\mathrm{d}x = \infty$

This can be generalized by introducing a kernel function:

$$\hat{p}(x) = \frac{1}{Nd_k(x)} \sum_{i=1}^{N} K\left(\frac{x - x_i}{d_k(x)}\right)$$

3.4. Nearest Neighbor Estimators: Illustration

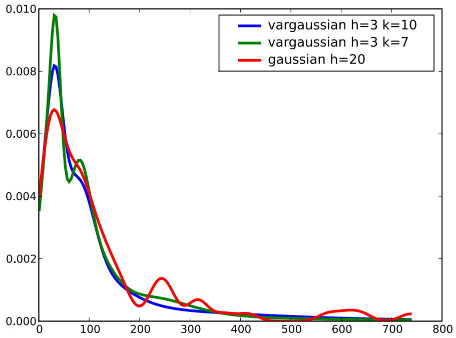


3.5. Variable Kernel Estimator

$$\hat{p}(x) = \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{d_k(x_i)} K\left(\frac{x - x_i}{h d_k(x_i)}\right)$$

The bandwidth depends on the densities around the observations x_i.

Properties: Same as for the basic kernel density estimator.



4. Remarks

4.1. Estimators for Multivariate Densities

All discussed methods extend straightforwardly.

4.2. Sampling from a Kernel Density Estimate

Basically:

1. Choose an observation at random.

2. Sample from the kernel centered at that observation.

Generic methods exist for sampling from arbitrary PDFs (inverse transform sampling, rejection sampling, …).

For many specific kernel functions, more efficient, specialized methods exist (normal distributions: Box-Muller transform, …).

5. Methods for Particular Domains

5.1. Bounded Domains

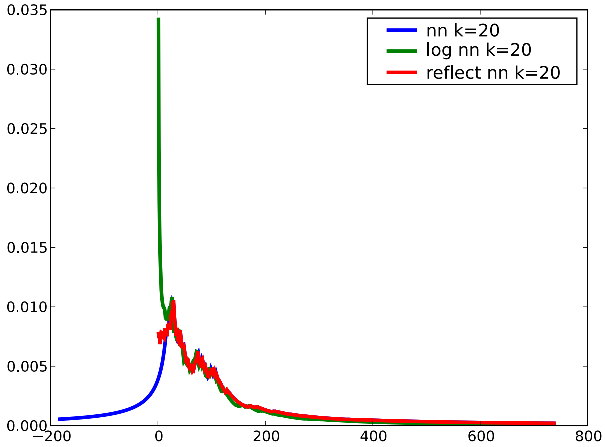
- Outside of the boundaries, set $\hat{p}$ to zero:
  - $\int \hat{p}(x)\mathrm{d}x < 1$
  - Observations near the boundaries contribute less to $\hat{p}$, and $\hat{p}$ is underestimated near the boundaries.
- Take logs: e.g., for positive data, if $\hat{q}$ is the density estimated from the logarithm of the data, then

$$\hat{p}(x) = \frac{1}{x}\hat{q}(\log x).$$

  - There will be a spike at the singularity.
- Reflect the data: e.g., for positive data, if $\hat{q}$ is the density estimated from this augmented data set, then

$$\hat{p}(x) = \begin{cases} 2\hat{q}(x) & \text{for } x \geq 0 \\ 0 & \text{for } x < 0. \end{cases}$$

5.2. Bounded domains: Illustration



5.3. Circular Domains

Use a *von Mises* kernel.

Or, for example, replicate the angular data twice, shifted by $\pm 2\pi$ (or even more copies, if necessary). Then, apply a conventional estimator.

Cylindrical, toroidal etc. domains can be treated similarly.

5.4. Spherical Domains

For example, represent data as unit vectors $\mathbf{x}_i \in \mathbb{R}^d$, and use a ***von Mises–Fisher*** kernel:

$$\hat{p}(\mathbf{x};\kappa) = \frac{1}{N c_d(\kappa)} \sum_{i=1}^{N} \exp(\kappa \mathbf{x}^T \mathbf{x}_i)$$

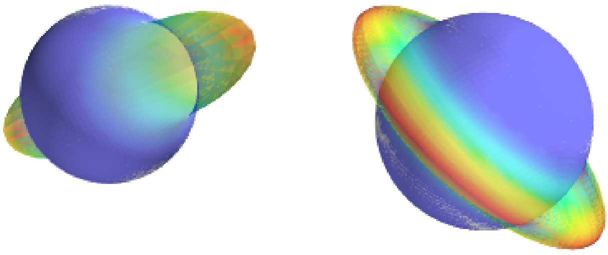
where κ is a concentration parameter. [Mardia and Jupp 1999]

The normalizing constant $c_d(\kappa)$ is messy to compute explicitly and can in practice be determined by numerical integration.

5.5. Axial Domains

Use a ***Dimroth-Watson*** kernel (inner product squared):

- $\kappa > 0$: antipodal
- $\kappa = 0$: uniform
- $\kappa < 0$: girdle

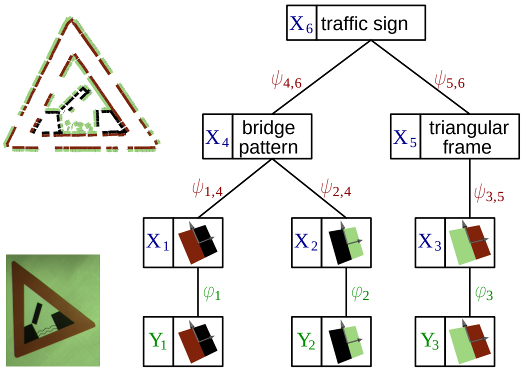


6. Case Study: Pose Estimation

Renaud Detry

[Detry et al. 2009]

6.1. Local Appearance + Spatial Relations



Vertex (random variable): distribution of *positions*

Edge (potential): distribution of *spatial relations* Ψ

Object Model (Markov network): structure; potentials Ψ, Φ

The traffic sign is composed of … etc.

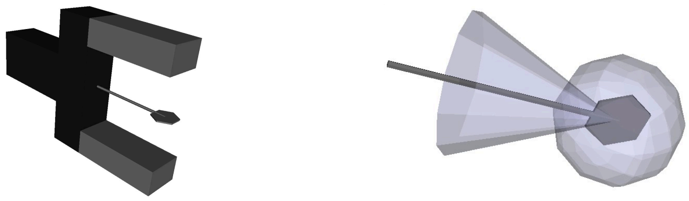
6.2. A Kernel for Pose Densities

- Pose $\mathbf{x} \in SE(3) = \mathbb{R}^3 \times SO(3)$

- Assuming that positions and orientations are independent, use an SE(3) kernel $K(\mathbf{x}; \boldsymbol{\mu}, \sigma)$ that factors into

  - a location kernel $N\left(\mathbf{x}_{\text{pos}}; \boldsymbol{\mu}_{\text{pos}}, \sigma_{\text{pos}}\right)$ (normal)

  - an orientation kernel $c(\sigma_{\text{or}}) \exp\left(\sigma_{\text{or}} \left(\boldsymbol{\mu}_{\text{or}}^{\text{qT}} \mathbf{x}_{\text{or}}^{\text{q}}\right)^2\right)$ (Dimroth-Watson), with $\mathbf{q}$ denoting quaternion representations.



6.3. Uses of KDE

Sampling:

- from weighted-particle representations
- for forming message products
- for stochastic integration of products of densities (cross-correlations)

Finding maximum:

- for determining the maximum-likelihood pose
- to find the best gripper pose for grasping

To sample in SE(*3*)**:**

1. Choose an observation at random.

2. Sample from the position part of the kernel centered at that observation.

3. Sample from the orientation part of the kernel centered at that observation [Wood 1994].

7. References

7.1. References

J. Copas, M. Fryer,[1] “Density estimation and suicide risks in psychiatric treatment ”. *Journal of the Royal Statistical Society (Series A)* 143, pp. 167–176, 1980.

R. Detry, N. Pugeault, J.[2] Piater, “A Probabilistic Framework for 3D Visual Object Representation ”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(10), pp. 1790–1803, 2009.

K. Mardia, P. Jupp, *Directional Statistics*[3] , Wiley 1999.

B. Silverman, *Density Estimation for Statistics and Data Analysis*[4] , Chapman & Hall/CRC 1986.

A. Wood, “Simulation of the von Mises Fisher distribution[5] ”. *Commun. Statist.– Simula.* 23(1), pp. 157–164, 1994.

[1] http://dx.doi.org/10.2307/2981988

[2] https://iis.uibk.ac.at/public/papers/Detry-2009-TPAMI.pdf

[3] http://dx.doi.org/10.1002/9780470316979

[4] https://www.routledge.com/Density-Estimation-for-Statistics-and-Data-Analysis/Silverman/p/book/9780412246203

[5] http://dx.doi.org/10.1080/03610919408813161