Probability Density Function Estimation: Theory and Mathematics CPE 486/586

Instructor: Rahul Bhadani

Sept 14, 2025

Contents

1 Introduction and Problem Formulation **3**

1.1 Problem Statement 3

1.2 Mathematical Framework 3

2 Data Structure and Sampling Considerations **3**

2.1 Temporal Structure 3

2.2 Stationarity Assumptions 4

3 Parametric PDF Estimation **4**

3.1 Normal Distribution Model 4

3.1.1 Parameter Estimation using Maximum Likelihood 4

3.2 Other Parametric Distributions 5

3.2.1 Beta Distribution (for bounded temperature ranges) 5

3.2.2 Gamma Distribution (for positive temperatures in Kelvin) 5

3.3 Goodness-of-Fit Testing 5

3.3.1 Kolmogorov-Smirnov Test 5

3.3.2 Anderson-Darling Test 5

4 Non-Parametric PDF Estimation 5

4.1 Kernel Density Estimation (KDE) . . . 5

4.1.1 Common Kernel Functions . . . 6

4.2 Bandwidth Selection . . . 6

4.2.1 Silverman's Rule of Thumb . . . 6

4.2.2 Cross-Validation Bandwidth . . . 6

4.3 Histogram-Based Estimation . . . 7

4.3.1 Basic Histogram . . . 7

4.3.2 Optimal Bin Width (Sturges' Rule) . . . 7

4.3.3 Freedman-Diaconis Rule . . . 7

5 Time-Varying PDF Models 7

5.1 Conditional PDF Estimation . . . 7

5.1.1 Kernel Regression Approach . . . 7

5.2 Functional Data Analysis . . . 7

5.2.1 Fourier Representation . . . 8

6 Model Selection and Validation 8

6.1 Information Criteria . . . 8

6.1.1 Akaike Information Criterion . . . 8

6.1.2 Bayesian Information Criterion . . . 8

6.2 Cross-Validation . . . 8

6.2.1 K-Fold Cross-Validation . . . 8

6.3 Bootstrap Confidence Intervals . . . 8

6.3.1 Bootstrap Procedure . . . 8

7 Practical Considerations 9

7.1 Sample Size Requirements . . . 9

7.2 Boundary Effects . . . 9

7.3 Computational Complexity . . . 9

8 Implementation Algorithm **9**
8.1 General Workflow 9
8.2 Error Metrics 10
8.2.1 Mean Integrated Squared Error 10
8.2.2 Pointwise Mean Squared Error 10

1 Introduction and Problem Formulation

1.1 Problem Statement

Given a time series of room temperature measurements $X_1, X_2, \ldots, X_n$ collected every 5 minutes from 9 AM to 9 PM, we want to estimate the probability density function $f(x)$ that describes the distribution of temperature values.

1.2 Mathematical Framework

Let X be a continuous random variable representing room temperature. The probability density function $f(x)$ satisfies:

$$f(x) \geq 0 \text{ for all } x \tag{1}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{2}$$

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{3}$$

Our goal is to estimate $f(x)$ from the observed sample $\{x_1, x_2, \ldots, x_n\}$.

2 Data Structure and Sampling Considerations

2.1 Temporal Structure

With measurements every 5 minutes from 9 AM to 9 PM:

- Time points: $t_i = 9:00 + 5i$ minutes, $i = 0, 1, \ldots, 143$
- Daily sample size: $n = 144$ observations
- For k days of data: $N = 144k$ total observations

2.2 Stationarity Assumptions

The basic PDF estimation assumes that observations are independent and identically distributed (i.i.d.). However, room temperature data may exhibit:

- **Temporal dependence**: $X(t)$ may be correlated with $X(t-1)$
- **Diurnal patterns**: $f(x|t)$ may vary with time of day
- **Non-stationarity**: Distribution parameters may change over time

However, in the most simplest case, we will assume i.i.d.

3 Parametric PDF Estimation

3.1 Normal Distribution Model

If we assume $X \sim N(\mu, \sigma^2)$, the PDF is:

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4}$$

3.1.1 Parameter Estimation using Maximum Likelihood

The likelihood function for n observations is:

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i;\mu,\sigma^2) \tag{5}$$

The log-likelihood is:

$$\ell(\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{6}$$

Maximum Likelihood Estimators:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{8}$$

3.2 Other Parametric Distributions

3.2.1 Beta Distribution (for bounded temperature ranges)

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \text{ for } x \in [0,1] \tag{9}$$

After rescaling temperature to $[0,1]$ using: $x' = \frac{x-x_{\min}}{x_{\max}-x_{\min}}$

3.2.2 Gamma Distribution (for positive temperatures in Kelvin)

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \text{ for } x > 0 \tag{10}$$

3.3 Goodness-of-Fit Testing

3.3.1 Kolmogorov-Smirnov Test

Test statistic:

$$D_n = \sup_x |F_n(x) - F_0(x)| \tag{11}$$

Where $F_n(x)$ is the empirical CDF and $F_0(x)$ is the theoretical CDF.

3.3.2 Anderson-Darling Test

$$A = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln F(x_{(i)}) + \ln \left(1 - F(x_{(n+1-i)}) \right) \right] \tag{12}$$

Source: https://www.statsref.com/HTML/anderson-darling.html

4 Non-Parametric PDF Estimation

4.1 Kernel Density Estimation (KDE)

The kernel density estimator is:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \tag{13}$$

Where:

- $K(u)$ is the kernel function
- h is the bandwidth parameter
- n is the sample size

Source: Chapter 4, Harrou, Fouzi, Abdelhafid Zeroual, Mohamad Mazen Hittawe, and Ying Sun. Road traffic modeling and management: Using statistical monitoring and deep learning. Elsevier, 2021.

4.1.1 Common Kernel Functions

Gaussian Kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \tag{14}$$

Epanechnikov Kernel (optimal):

$$K(u) = \frac{3}{4}(1 - u^2) \text{ for } |u| \leq 1, \text{ 0 otherwise} \tag{15}$$

Uniform Kernel:

$$K(u) = \frac{1}{2} \text{ for } |u| \leq 1, \text{ 0 otherwise} \tag{16}$$

4.2 Bandwidth Selection

4.2.1 Silverman's Rule of Thumb

$$h = 0.9 \times \min(\hat{\sigma}, \text{IQR}/1.34) \times n^{-1/5} \tag{17}$$

Where IQR is the interquartile range.

4.2.2 Cross-Validation Bandwidth

Minimize the integrated squared error:

$$\text{ISE}(h) = \int [\hat{f}(x) - f(x)]^2 dx \tag{18}$$

Practical Cross-Validation:

$$h^* = \arg\min \sum_{i=1}^{n} \int [\hat{f}_{-i}(x) - \delta(x - x_i)]^2 dx \tag{19}$$

Where $\hat{f}_{-i}(x)$ is the KDE excluding observation x_i.

4.3 Histogram-Based Estimation

4.3.1 Basic Histogram

$$\hat{f}(x) = \frac{n_j}{n \times \Delta x} \text{ for } x \in [x_j, x_{j+1}) \tag{20}$$

Where n_j is the count in bin j and Δx is the bin width.

4.3.2 Optimal Bin Width (Sturges' Rule)

Number of bins: $k = \lceil 1 + \log_2(n) \rceil$
Bin width: $\Delta x = \frac{x_{\max} - x_{\min}}{k}$

4.3.3 Freedman-Diaconis Rule

$$\Delta x = 2 \times \text{IQR} \times n^{-1/3} \tag{21}$$

5 Time-Varying PDF Models

5.1 Conditional PDF Estimation

For non-stationary data, estimate $f(x|t)$:

5.1.1 Kernel Regression Approach

$$\hat{f}(x|t) = \sum_{i=1}^{n} w_i(t) K\left(\frac{x - x_i}{h}\right) \tag{22}$$

Where $w_i(t)$ are time-dependent weights:

$$w_i(t) = \frac{W\left(\frac{t-t_i}{b}\right)}{\sum_{j=1}^{n} W\left(\frac{t-t_j}{b}\right)} \tag{23}$$

5.2 Functional Data Analysis

Model temperature as a function $T(t)$ and estimate the distribution of functional parameters.

5.2.1 Fourier Representation

$$T(t) = \mu(t) + \sum_{k=1}^{K} [a_k \cos(2\pi kt/P) + b_k \sin(2\pi kt/P)] + \varepsilon(t) \tag{24}$$

Where $P = 12$ hours is the period.

6 Model Selection and Validation

6.1 Information Criteria

6.1.1 Akaike Information Criterion

$$\text{AIC} = -2\ell(\hat{\theta}) + 2p \tag{25}$$

6.1.2 Bayesian Information Criterion

$$\text{BIC} = -2\ell(\hat{\theta}) + p\ln(n) \tag{26}$$

Where p is the number of parameters.

6.2 Cross-Validation

6.2.1 K-Fold Cross-Validation

1. Divide data into K folds
2. For each fold k, estimate $\hat{f}_{-k}(x)$ using remaining data
3. Evaluate log-likelihood on fold k: $\text{LL}_k = \sum_{i \in k} \ln(\hat{f}_{-k}(x_i))$
4. Average: $\text{CV-LL} = \frac{1}{K}\sum_{k=1}^{K} \text{LL}_k$

6.3 Bootstrap Confidence Intervals

6.3.1 Bootstrap Procedure

1. Resample with replacement: $X^* = \{x_1^*, x_2^*, \ldots, x_n^*\}$
2. Compute $\hat{f}^*(x)$

3. Repeat B times

4. Construct pointwise confidence intervals from bootstrap distribution

7 Practical Considerations

7.1 Sample Size Requirements

For reliable PDF estimation:

- Parametric: $n \geq 30$ typically sufficient
- Non-parametric KDE: $n \geq 100$ recommended
- Complex time-varying models: $n \geq 1000$ may be needed

7.2 Boundary Effects

Room temperatures are bounded (e.g., 15-30°C). Use:

- Boundary kernels for KDE
- Transformed variables
- Reflection methods

7.3 Computational Complexity

- Histogram: $O(n)$
- KDE: $O(n^2)$ for evaluation at n points
- Parametric MLE: $O(n)$ per iteration

8 Implementation Algorithm

8.1 General Workflow

Listing 1: General Workflow for PDF Estimation

```
1. Data Preprocessing:
   - Remove outliers (|x - median| > 3xMAD)
   - Check for temporal patterns
   - Test for stationarity

2. Model Selection:
   - Fit parametric candidates
   - Compute non-parametric estimates
   - Use cross-validation for comparison

3. Parameter Estimation:
   - If parametric: MLE or method of moments
   - If KDE: optimize bandwidth
   - If histogram: optimize bin width

4. Validation:
   - Goodness-of-fit tests
   - Cross-validation
   - Bootstrap confidence intervals

5. Final Estimation:
   - Select best model
   - Provide uncertainty quantification
```

8.2 Error Metrics

8.2.1 Mean Integrated Squared Error

$$\text{MISE} = E\left[\int (\hat{f}(x) - f(x))^2 dx\right] \tag{27}$$

8.2.2 Pointwise Mean Squared Error

$$\text{MSE}(x) = E[(\hat{f}(x) - f(x))^2] = \text{Bias}^2(x) + \text{Variance}(x) \tag{28}$$