

02427 Advanced Time Series Analysis

Lecture 1

Tobias K. S. Ritschel

Department of Applied Mathematics and Computer Science,
Technical University of Denmark

September 2nd, 2025

Outline

Kernel methods

Kernel estimator

Non-parametric least-squares

Conditional mean and covariance estimators

Locally weighted polynomial regression

Conditional parametric models

Mixture approximations

Lag-dependence functions

Kernel methods

Problem statement

Objective: Estimate

$$\mathbb{E}[Y_{t+1}|Y_t, Y_{t-1}, \dots] \quad (1)$$

More general objective: Estimate

$$\mathbb{E}[Y_{t_3}, \dots, Y_{t_2}|Y_{t_1}, \dots, Y_{t_0}] \quad (2)$$

Another variant: Estimate

$$\mathbb{E}[G(Z_t)|Y_t] \quad (3)$$

Approximate probability density function

Definition of probability density

$$\begin{aligned} f(y) &= \frac{dF}{dy}(y) = \lim_{h \rightarrow 0} \frac{F(y+h) - F(y-h)}{2h} \\ &= \lim_{h \rightarrow 0} \frac{P(y-h \leq Y < y+h)}{2h} \end{aligned} \quad (4)$$

Let $\{Y_t\}_{t=1}^N$ be a time series

Approximate probability for finite h

$$P(y-h \leq Y < y+h) \approx \frac{N_h(y)}{N} \quad (5)$$

$N_h(y)$ is the number of $t \in \{1, \dots, N\}$ for which

$$y-h \leq Y_t < y+h \quad (6)$$

Approximate probability density function

Indicator function

$$\chi_h(Y_t; y) = \begin{cases} 1, & y - h \leq Y_t < y + h, \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

Rewrite approximation

$$P(y - h \leq Y_t < y + h) \approx \frac{1}{N} \sum_{t=1}^N \chi_h(Y_t; y) \quad (8)$$

Approximate probability density

$$f(y) \approx \frac{1}{2Nh} \sum_{t=1}^N \chi_h(Y_t; y) \quad (9)$$

Exercise: lecture02_kernel_methods_approximate_pdf

Approximate probability density function

Weight function

$$k(u) = \begin{cases} \frac{1}{2}, & -1 < u \leq 1, \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

Approximate probability density

$$f(y) \approx \frac{1}{Nh} \sum_{t=1}^N k\left(\frac{y - Y_t}{h}\right) \quad (11)$$

This approximation applies for any kernel, k , for which

$$\int_{-\infty}^{\infty} k(u) \, du = 1, \quad \int_{-\infty}^{\infty} uk(u) \, du = 0 \quad (12)$$

Approximate conditional mean

Approximate probability densities

$$f(y) \approx \frac{1}{Nh} \sum_{t=1}^N k\left(\frac{y - Y_t}{h}\right), \quad (13a)$$

$$f(y, z) \approx \frac{1}{Nh^2} \sum_{t=1}^N k\left(\frac{y - Y_t}{h}\right) k\left(\frac{z - Z_t}{h}\right) \quad (13b)$$

Approximate conditional mean

$$\begin{aligned} \mathbb{E}[Z_t | Y_t = y] &= \int_{-\infty}^{\infty} z f(z|y) dz = \int_{-\infty}^{\infty} z \frac{f(y, z)}{f(y)} dz \\ &= \int_{-\infty}^{\infty} z \frac{\frac{1}{Nh^2} \sum_{s=1}^N k\left(\frac{y - Y_s}{h}\right) k\left(\frac{z - Z_s}{h}\right)}{\frac{1}{Nh} \sum_{s=1}^N k\left(\frac{y - Y_s}{h}\right)} dz \\ &= \frac{\sum_{s=1}^N k\left(\frac{y - Y_s}{h}\right) \int_{-\infty}^{\infty} \frac{z}{h} k\left(\frac{z - Z_s}{h}\right) dz}{\sum_{s=1}^N k\left(\frac{y - Y_s}{h}\right)} \end{aligned} \quad (14)$$

Approximate conditional mean

Change integration variable to $w = z - Z_s$

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{z}{h} k\left(\frac{z - Z_s}{h}\right) dz &= \int_{-\infty}^{\infty} \frac{w + Z_s}{h} k\left(\frac{w}{h}\right) dw \\ &= \underbrace{\int_{-\infty}^{\infty} \frac{w}{h} k\left(\frac{w}{h}\right) dw}_{=0} + Z_s \underbrace{\int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{w}{h}\right) dw}_{=1} = Z_s\end{aligned}\quad (15)$$

Approximate conditional expectation

$$\mathbb{E}[Z_t | Y_t = y] \approx \frac{\sum_{s=1}^N k\left(\frac{y - Y_s}{h}\right) Z_s}{\sum_{s=1}^N k\left(\frac{y - Y_s}{h}\right)} \quad (16)$$

Approximate conditional expectation

$$\mathbb{E}[G(Z_t) | Y_t = y] \approx \frac{\sum_{s=1}^N k\left(\frac{y - Y_s}{h}\right) G(Z_s)}{\sum_{s=1}^N k\left(\frac{y - Y_s}{h}\right)} \quad (17)$$

Exercise: lecture02_kernel_methods_conditional_mean

Kernel estimator

Kernel estimation

Model

$$Y = g(X) + e, \quad \mathbb{E}[e] = 0 \quad (18)$$

X is known, but g is unknown

Conditional mean

$$\mathbb{E}[Y|X = x] = g(x) \quad (19)$$

Use the approximate conditional expectation
(X_s and Y_s for $s = 1, \dots, N$ is “training data”)

$$\mathbb{E}[Y|X = x] \approx \frac{\sum_{s=1}^N k\left(\frac{x-X_s}{h}\right) Y_s}{\sum_{s=1}^N k\left(\frac{x-X_s}{h}\right)} \quad (20)$$

Kernel functions

Reparametrization

$$K_h(u) = \frac{1}{h} k\left(\frac{u}{h}\right)$$

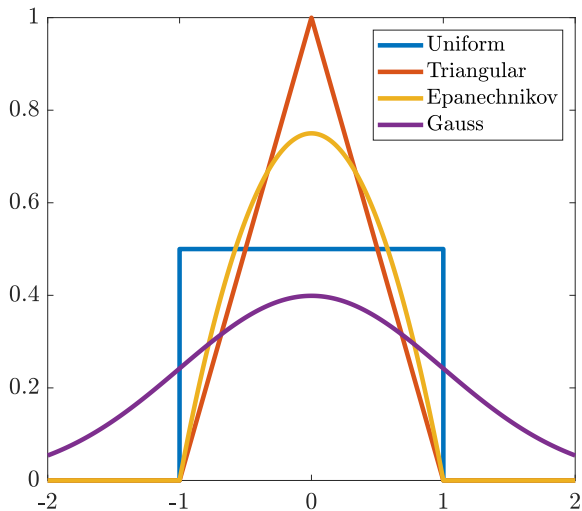
Epanechnikov kernel with bandwidth h

$$K_h^{Epa}(u) = \frac{3}{4h} \left(1 - \frac{u^2}{h^2}\right) I_{\{|u| \leq h\}} \quad (21)$$

Gaussian kernel with bandwidth h

$$K_h^{Gau}(u) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{u^2}{2h^2}\right) \quad (22)$$

Kernel functions - illustration



Nearest Neighbor (NN) estimation

- ▶ Previously: h was fixed
- ▶ Another (and often better) alternative is
 - * k -NN (k Nearest Neighbor) (Note: here, k is a natural number - not the kernel).

Since it is reasonable to link the value of k to N we often consider a fraction (or percentage) of the data. Often the symbol, $\tilde{k}$, is used in this case.

Multivariate kernels

If we allow $q > 1$, a kernel of q 'th order have to be used, i.e. a real bounded function $k_q(u_1, \dots, u_q)$, where $\int k_q(u) du = 1$

By a generalization of the bandwidth h to the quadratic matrix h with the dimension $q \times q$ the more general kernel estimator for $g(X^{(1)}, \dots, X^{(q)})$ is

$$\hat{g}(x) = \frac{\frac{1}{N} \sum_{s=1}^N Y_s k_q[(x - X_s)h^{-1}]}{\frac{1}{N} \sum_{s=1}^N k_q[(x - X_s)h^{-1}]} \quad (23)$$

where $x = (x_1, \dots, x_q)$ and $X_s = (X_s^{(1)}, \dots, X_s^{(q)})$. This is called the *Nadaraya-Watson estimator*.

Product kernel

In practice product kernels are used, i.e.

$$k_q(xh^{-1}) = \prod_{i=1}^q k(x_i/h_i)$$

where k is a kernel of order 1

Assuming that $h_i = h$, $i = 1, \dots, q$, the following generalization is obtained

$$\hat{g}(x) = \frac{\frac{1}{N} \sum_{s=1}^N Y_s \prod_{i=1}^q k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{N} \sum_{s=1}^N \prod_{i=1}^q k\{h^{-1}(x_i - X_s^{(i)})\}}$$

Non-parametric least-squares

Non-parametric LS

Define the weight

$$w_s(x) = \frac{\prod_{i=1}^q k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{N} \sum_{s=1}^N \prod_{i=1}^q k\{h^{-1}(x_i - X_s^{(i)})\}}$$

the estimator corresponds to the local average:

$$\hat{g}(x) = \frac{1}{N} \sum_{s=1}^N w_s(x) Y_s$$

Non-parametric LS - continued

The estimate is actually a *non-parametric least squares estimate* at the point x

This is recognized from the fact that the solution to the least squares problem

$$\arg \min_{\theta} \frac{1}{N} \sum_{s=1}^N w_s(x) (Y_s - \theta)^2$$

is given by

$$\hat{\theta}_{LS}(x) = \frac{\sum_{s=1}^N w_s(x) Y_s}{\sum_{s=1}^N w_s(x)}$$

Exercise: lecture02_nonparametric_least_squares

Non-parametric LS - continued

At each x , $\hat{g}$ is a weighted least-squares estimate

$$\hat{g}(x) = \hat{\theta}_{LS}(x) \frac{1}{N} \sum_{s=1}^N w_s(x) = \hat{\theta}_{LS}(x)$$

Variance of the curve estimate at point x

$$\hat{\sigma}^2(x) = \frac{1}{N} \sum_{s=1}^N w_s(x) (Y_s - \hat{g}(X_s))^2$$

Weights (as shown previously)

$$w_s(x) = \frac{\prod_{i=1}^q k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{N} \sum_{s=1}^N \prod_{i=1}^q k\{h^{-1}(x_i - X_s^{(i)})\}}$$

The bandwidth

The bandwidth, h , determines the smoothness of $\hat{g}$

- ▶ If h is large the variance is small but the bias is large.
- ▶ If h is small the variance is large but the bias is small.
- ▶ As $h \rightarrow \infty$,

$$\hat{g}(x) \rightarrow \bar{Y} \quad (24)$$

- ▶ As $h \rightarrow 0$,

$$\hat{g}(x) \rightarrow \begin{cases} Y_i, & x = (X_i^{(1)}, \dots, X_i^{(q)}), \\ 0, & \text{otherwise} \end{cases} \quad (25)$$

Selection of bandwidth - Cross Validation

Goal: Select h that minimizes

$$\text{MSE1}(h) = \frac{1}{N} \sum_{i=1}^N [\hat{g}(X_i) - g(X_i)]^2 \pi(X_i)$$

The weight function $\pi(\dots)$ screens off extreme values.

The function g is unknown

Approximate $g(X_i)$ by Y_i

$$\text{MSE2}(h) = \frac{1}{N} \sum_{i=1}^N [\hat{g}(X_i) - Y_i]^2 \pi(X_i)$$

As $h \rightarrow 0$, $\widehat{\text{MSE2}}(h) \rightarrow 0$

The “Leave one out” method

Goal: Prevent that X_i is used in the estimate for Y_i

For all (X_i, Y_i) , define an estimator $\hat{g}_{(i)}$ for Y_i based on all data except (X_i)

N “leave-one-out” estimators $\hat{g}_{(1)}, \dots, \hat{g}_{(N)}$

$$\hat{g}_{(j)}(x) = \frac{\frac{1}{N-1} \sum_{s \neq j} Y_s \prod_{i=1}^q k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{N-1} \sum_{s \neq j} \prod_{i=1}^q k\{h^{-1}(x_i - X_s^{(i)})\}}$$

The “Leave one out” method

Cross-Validation criterion based on the N “leave-one-out” estimates $\hat{g}_{(1)}, \dots, \hat{g}_{(N)}$

$$CV(h) = \frac{1}{N} \sum_{i=1}^N [Y_i - \hat{g}_{(i)}(X_i)]^2 \pi(X_i)$$

Goal: Find h that minimizes $CV(h)$

Conditional mean and covariance estimators

Applications of kernel estimators

Given: Realizations $X_1, \dots, X_N$ of a stochastic process $\{X_t\}$

Goal: Estimate the functions $g(\cdot)$ and $h(\cdot)$ in the model

$$X_t = g(X_{t-1}, \dots, X_{t-p}) + h(X_{t-1}, \dots, X_{t-q})e_t$$

Estimators

$$M(x_{i_1}, \dots, x_{i_d}) = \mathbb{E}[X_t | X_{t-i_1} = x_{i_1}, \dots, X_{t-i_d} = x_{i_d}]$$

$$V(x_{i_1}, \dots, x_{i_d}) = \text{Var}[X_t | X_{t-i_1} = x_{i_1}, \dots, X_{t-i_d} = x_{i_d}]$$

- ▶ Kernel estimator
- ▶ Splines
- ▶ Nearest neighbor
- ▶ Neural network estimators

Applications of kernel estimators - continued

Using a product kernel we obtain

$$\hat{M}(x_{i_1}, \dots, x_{i_d}) = \frac{\frac{1}{N-i_d} \sum_{s=i_d+1}^N X_s \prod_{j=1}^d k\{h^{-1}(x_{i_j} - X_{s-i_j})\}}{\frac{1}{N-i_d+i_1} \sum_{s=i_d+1}^{N+i_1} \prod_{j=1}^d k\{h^{-1}(x_{i_j} - X_{s-i_j})\}}$$

Assuming that $E(X_t) = 0$ the estimator for $V(\cdot)$ is

$$\hat{V}(x_{i_1}, \dots, x_{i_d}) = \frac{\frac{1}{N-i_d} \sum_{s=i_d+1}^N X_s^2 \prod_{j=1}^d k\{h^{-1}(x_{i_j} - X_{s-i_j})\}}{\frac{1}{N-i_d+i_1} \sum_{s=i_d+1}^{N+i_1} \prod_{j=1}^d k\{h^{-1}(x_{i_j} - X_{s-i_j})\}}$$

Applications of kernel estimators - continued

If the mean is not zero

$$\hat{V}(x_{i_1}, \dots, x_{i_d}) = \frac{\frac{1}{N-i_d} \sum_{s=i_d+1}^N X_s^2 \prod_{j=1}^d k\{h^{-1}(x_{i_j} - X_{s-i_j})\}}{\frac{1}{N-i_d+i_1} \sum_{s=i_d+1}^{N+i_1} \prod_{j=1}^d k\{h^{-1}(x_{i_j} - X_{s-i_j})\}} - \hat{M}^2(x_{i_1}, \dots, x_{i_d})$$

Identification of functional relationship

Goal: Estimate conditional mean and variance

We have simulated 10 stochastic independent time series for both of the following non-linear models

$$\text{A : } X_t = \begin{cases} -0.8X_{t-1} + e_t, & X_{t-1} \geq 0 \\ 0.8X_{t-1} + 1.5 + e_t, & X_{t-1} < 0 \end{cases}$$

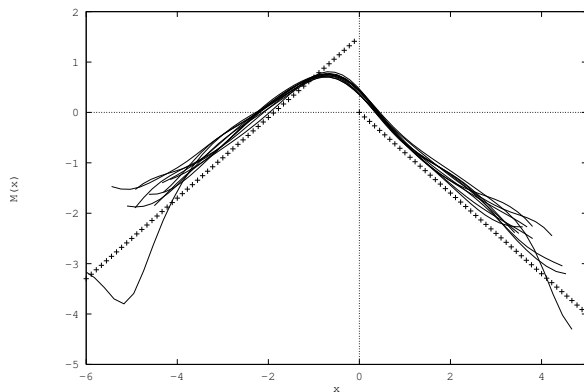
$$\text{B : } X_t = \sqrt{1 + X_{t-1}^2} e_t$$

- ▶ Gaussian kernels with bandwidths, $h = 0.6$, are used
- ▶ Simulation of 10 independent time series

Exercise: lecture02_conditional_mean_and_variance_A

Identification of functional relationship

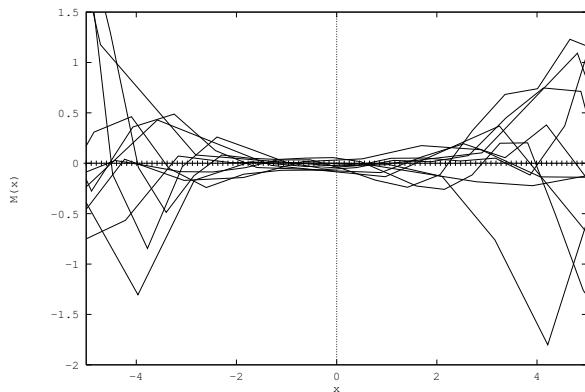
- Estimated $M(x)$ for model A



The '+' shows the theoretical conditional mean and variance

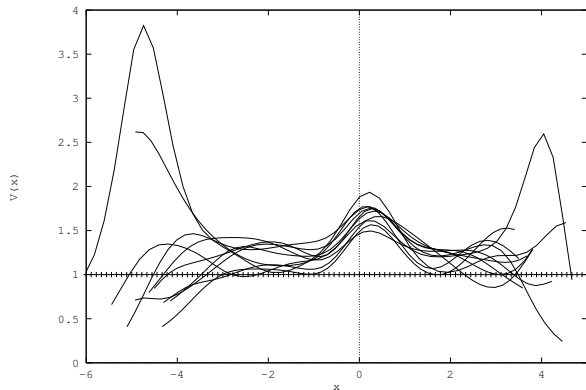
Identification of functional relationship

- Estimated $M(x)$ for model B



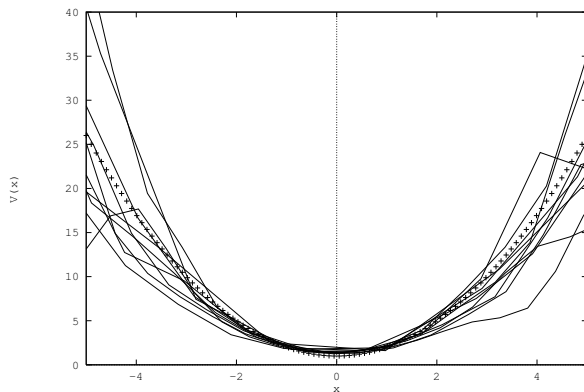
Identification of functional relationship

- Estimated $V(x)$ for model A



Applications of kernel estimators

- Estimated $V(x)$ for model B



Locally weighted polynomial regression

Locally-weighted polynomial regression

- Previously, the local model, $Y_t = \theta + e_t$ was estimated as

$$\arg \min_{\theta} \frac{1}{N} \sum_{s=1}^N w_s(x) (Y_s - \theta)^2 \quad (26)$$

- Extension: Locally weighted least squares

Model

$$Y_t = \mu(X_t) + e_t$$

Polynomial $P(X_t, x)$ such that $\mu(X_t) = P(X_t, x)$

$$\hat{\theta}(x) = \arg \min_{\theta} \frac{1}{N} \sum_{s=1}^N w_s(x) (Y_s - P(X_s, x))^2$$

The local estimate of $\mu(x)$ is $\hat{\mu}(x) = \hat{P}(x, x)$.

Locally-weighted polynomial regression

–Example

$$P(X_t, x) = \theta_0 + \theta_1(X_t - x) + \theta_2(X_t - x)^2 \quad (27)$$

$\theta = (\theta_0, \theta_1, \theta_2)^T$ is estimated by

$$\hat{\theta}(x) = \arg \min_{\theta} \frac{1}{N} \sum_{s=1}^N w_s(x) (Y_s - P(X_s, x))^2 \quad (28)$$

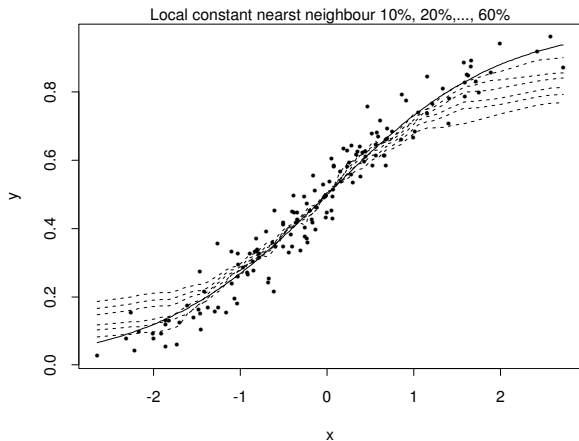
The local estimate of μ is:

$$\hat{\mu}(x) = \hat{\theta}_0(x) \quad (29)$$

Locally-weighted polynomial regression

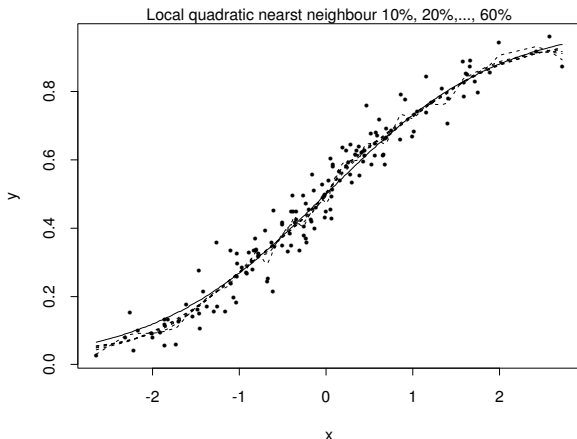
LWLS estimate using nearest neighbor.

First a local constant model is applied.



Locally-weighted polynomial regression

Then using local quadratic approximation.



The bias is much larger for the constant mean model.

Conditional parametric models

Conditional parametric models

Also called Varying-coefficient models

Conditional parametric model

$$Y_t = Z_t^T \theta(X_t) + e_t, \quad t = 1, \dots, N \quad (30)$$

Estimate

$$\arg \min_{\theta} \frac{1}{N} \sum_{s=1}^N w_s(x) \left(Y_s - Z_s^T \theta(x) \right)^2 \quad (31)$$

where $\theta(x)$ are polynomials.

Conditional parametric models – Example

Locally weighted linear model: Assume $Z_t^T = (1, Z_{1t})$ and $\dim(x) = 1$. Then, $\theta^T(x) = (\theta_0(x), \theta_1(x))$

Write out inner product

$$\begin{aligned} Z_t^T \theta(x) &= (1, Z_{1t}) \begin{pmatrix} \theta_0(x) \\ \theta_1(x) \end{pmatrix} \\ &= (1, Z_{1t}) \begin{pmatrix} \theta_{00} + \theta_{01}x + \theta_{02}x^2 \\ \theta_{10} + \theta_{11}x + \theta_{12}x^2 \end{pmatrix} \\ &= \left(1, x, x^2, Z_{1t}, Z_{1t}x, Z_{1t}x^2\right) \begin{pmatrix} \theta_{00} \\ \theta_{01} \\ \theta_{02} \\ \theta_{10} \\ \theta_{11} \\ \theta_{12} \end{pmatrix} \end{aligned}$$

Conditional parametric models – example

Identified model

$$\hat{\theta}_0(x) = \hat{\theta}_{00}(x) + \hat{\theta}_{01}(x)x + \hat{\theta}_{02}(x)x^2$$

$$\hat{\theta}_1(x) = \hat{\theta}_{10}(x) + \hat{\theta}_{11}(x)x + \hat{\theta}_{12}(x)x^2$$

Mixture approximations

Mixture approximations – Problem statement

We have N samples, $\{Y_t\}_{t=1}^N$, sampled from a distribution $y \sim \mathcal{F}(\theta)$ over a domain $\mathcal{D}$

How can we approximate $\mathcal{F}$?

Assumptions: The probability density $p(y)$ is **bounded** and **continuous**

- ▶ $p(y) \leq K < \infty$ for all $y \in \mathcal{D}$
- ▶ For all $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|y - z| < \delta \Rightarrow |p(y) - p(z)| < \epsilon \quad (32)$$

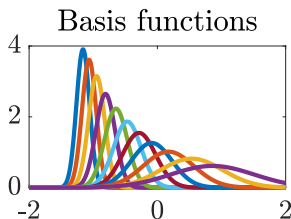
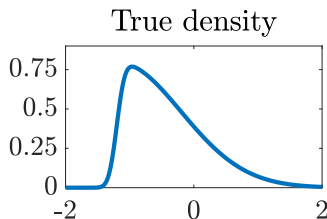
Mixture approximations – Mixture approximations

Approximate the density

$$p(y) \approx \hat{p}(y) = \sum_{m=0}^M c_m \ell_m(y) \quad (33)$$

Task: Determine

- ▶ the order, M
- ▶ the coefficients, c_m
- ▶ maybe some hyperparameters in the basis functions, ℓ_m



Mixture approximations – Different types

Mixture approximation

$$p(y) \approx \hat{p}(y) = \sum_{m=0}^M c_m \ell_m(y), \quad (34)$$

Gauss basis functions (2 hyperparameters: μ and σ)

$$\ell_m(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right), \quad y \in \mathcal{D} = (-\infty, \infty) \quad (35)$$

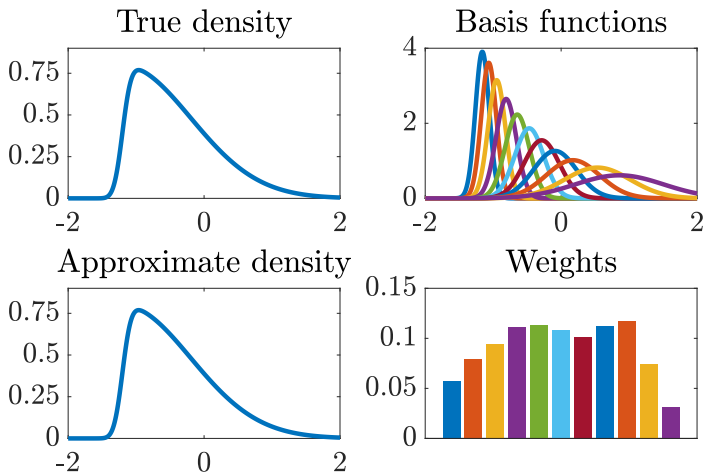
Erlang basis functions (1 hyperparameter: a)

$$\ell_m(y) = \frac{a^{m+1}}{m!} t^m e^{-at}, \quad y \in \mathcal{D} = [0, \infty) \quad (36)$$

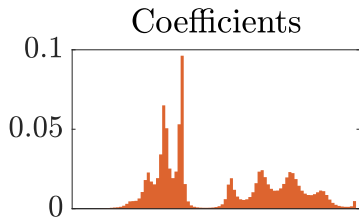
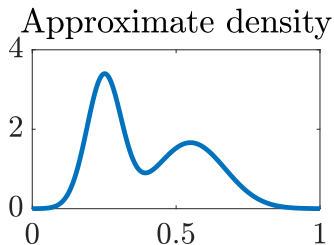
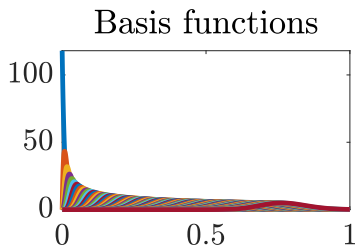
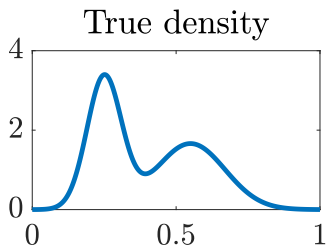
Beta basis functions (0 hyperparameters)

$$\ell_m(y) = \frac{(M+1)}{m!(M-m)!} t^m (1-t)^{M-m}, \quad y \in \mathcal{D} = [0, 1] \quad (37)$$

Gaussian mixture approximation ($M = 10$)



Example: Erlang mixture approximation ($M = 90$)



Estimation methods

Least-squares (based on histogram)

$$\min_{\{c_m\}_{m=0}^M} \frac{1}{2} \sum_{t=1}^N (\hat{p}(Y_t; q) - \tilde{p}(Y_t))^2, \quad (38a)$$

$$\text{subject to} \quad \sum_{m=0}^M c_m = 1 \quad (38b)$$

Estimate both coefficients and hyperparameters

$$\min_{\{c_m\}_{m=0}^M, q} \frac{1}{2} \sum_{t=1}^N (\hat{p}(Y_t; q) - \tilde{p}(Y_t))^2, \quad (39a)$$

$$\text{subject to} \quad \sum_{m=0}^M c_m = 1 \quad (39b)$$

Estimation methods

Estimate both coefficients and individual hyperparameters

$$\min_{\{c_m, q_m\}_{m=0}^M} \frac{1}{2} \sum_{t=1}^N (\hat{p}(Y_t; q) - \tilde{p}(Y_t))^2, \quad (40a)$$

$$\text{subject to } \sum_{m=0}^M c_m = 1 \quad (40b)$$

Estimate both coefficients and individual hyperparameters

+ guarantee non-negativity

$$\min_{\{c_m, q_m\}_{m=0}^M} \frac{1}{2} \sum_{t=1}^N (\hat{p}(Y_t; \{q_m\}_{m=0}^M) - \tilde{p}(Y_t))^2, \quad (41a)$$

$$\text{subject to } \sum_{m=0}^M c_m = 1, \quad (41b)$$

$$c_m \geq 0, \quad m = 0, \dots, M \quad (41c)$$

Estimation methods

Maximum likelihood estimation

$$\max_{\{c_m, q_m\}_{m=0}^M} \mathcal{L}(\{Y_t\}_{t=1}^N) = \prod_{t=1}^N \hat{p}(Y_t; \{q_m\}_{m=0}^M), \quad (42a)$$

$$\text{subject to } \sum_{m=0}^M c_m = 1 \quad (42b)$$

Moment-matching

$$\mathbb{E}[\hat{y}^k] \approx \mathbb{E}[y^k], \quad k = 1, \dots, K \quad (43)$$

See [1, Sec. 1.4] and [2] for more methods

What can be said theoretically?

Types of convergence

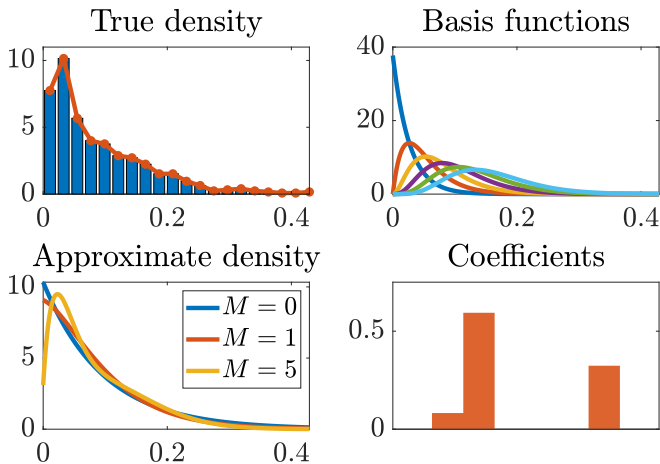
$$\int_0^Y \hat{p}(y) dy \rightarrow \int_0^Y p(y) dy \quad \text{for all } Y \in \mathcal{D} \quad (\text{weak convergence}),$$
$$\hat{p}(y) \rightarrow p(y) \quad \text{for all } y \in \mathcal{D} \quad (\text{pointwise})$$

1. Gauss mixture approximation, $\mathcal{D} = (-\infty, \infty)$
 - ▶ “... any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities with common variance ...” [2, Sec. 1.1.1], [1, Sec. 1.4 and 5.3]
 - ▶ Pointwise convergence [3]
2. Erlang mixture approximation, $\mathcal{D} = [0, \infty)$
 - ▶ Weak convergence [4, 5, 6, 7]
 - ▶ Pointwise convergence [8]
3. Beta mixture approximations, $\mathcal{D} = [0, 1]$
 - ▶ ℓ_m is a Bernstein polynomial scaled by $M + 1$
 - ▶ pointwise convergence is guaranteed by Weierstrass' approximation theorem [9] as $M \rightarrow \infty$

What is the difference between a mixture approximation
and a mixture model?

What can you do with mixture approximations?

Model of concavity from breast cancer data set¹



¹Link to data set on Kaggle.

What about time-varying distributions?

The observations, $\{Y_t\}_{t=1}^N$, are sampled from a time-varying distribution, $y_t \sim \mathcal{F}(\theta_t)$

Time-varying approximation

$$p(t, y) \approx \hat{p}(t, y) = \sum_{m=0}^M c_m(t) \ell_m(y) \quad (44)$$

Use recursive or moving-horizon estimation methods to update estimates of c_m

Relation to kernel estimation

Choose coefficients [1, Sec. 5.3]

$$c_m = \frac{1}{M} \quad (45)$$

Choose basis functions

$$\ell_m = \frac{1}{h} \ell \left(\frac{y - \mu_m}{h} \right) \quad (46)$$

μ_m is a location parameter, and h is a spread parameter. Set the location parameter equal to Y_t and $M = N$

$$\hat{p}(y) = \frac{1}{Nh} \sum_{t=1}^N \ell \left(\frac{y - Y_t}{h} \right) \quad (47)$$

From the first part of the lecture

$$f(y) \approx \frac{1}{Nh} \sum_{t=1}^N k \left(\frac{y - Y_t}{h} \right) \quad (48)$$

Lag-dependence functions

Introduction

- ▶ Autocorrelation functions are used to identify the necessary number of lags
- ▶ Parameter estimation leads to residuals which can be used for model validation and quality assurance
- ▶ In classical tests of residuals, the sample autocorrelation function (ACF) and the sample sample partial autocorrelation function (PACF) are used
- ▶ IMPORTANT: Correlation functions measure **linear** dependencies only
- ▶ In this lecture, we suggest generalizations to nonlinear dependencies

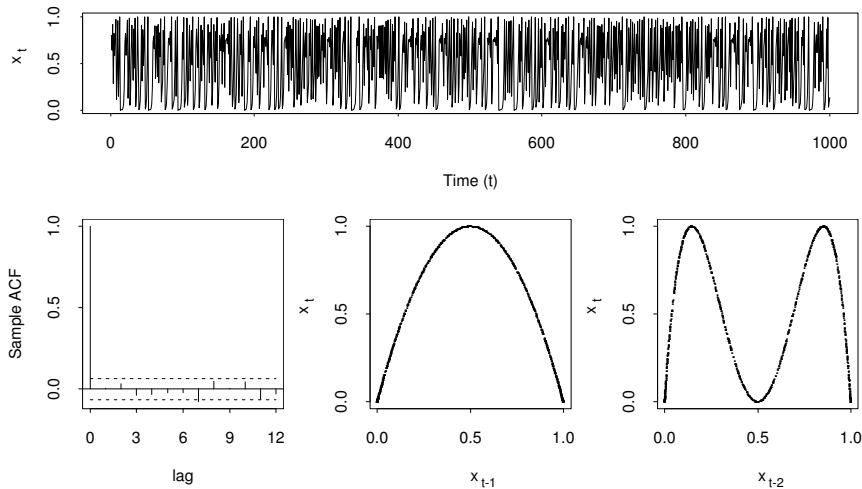
Motivation

- ▶ Consider the nonlinear (deterministic) model:

$$x_t = \alpha x_{t-1}(1 - x_{t-1})$$

- ▶ This model is the so-called *logistic map*.
- ▶ For $\alpha \in [3.56994568, 4]$ a chaotic behaviour is frequently observed.

Motivation ($\alpha = 4, x_1 = 0.8$)



Preliminaries

- ▶ Estimates of correlation and partial correlation are closely related to values of the coefficient of determination (R-squared) obtained using linear models
- ▶ The generalizations are based on definitions of similar R-squared values obtained using non-linear models
- ▶ To avoid model structure limitations **non-parametric models** are used

Autocovariance and autocorrelation

Autocovariance function

$$\gamma(t_1, t_2) = \text{Cov}[Y_{t_1}, Y_{t_2}] \quad (49)$$

Autocovariance function for stationary process

$$\gamma(k) = \gamma(t, t + k) = \text{Cov}[Y_t, Y_{t+k}] \quad (50)$$

Autocorrelation function

$$\rho(t_1, t_2) = \frac{\gamma(t_1, t_2)}{\sqrt{\sigma^2(t_1)\sigma^2(t_2)}} \quad (51)$$

Autocorrelation function for stationary processes

$$\rho(k) = \rho(t, t + k) = \frac{\gamma(k)}{\sigma^2} \quad (52)$$

Squared multiple correlation coefficient

Squared multiple correlation coefficient

$$\rho_{0,(1,\dots,k)}^2 = \frac{\text{Var}[Y] - \text{Var}[Y|X_1, \dots, X_k]}{\text{Var}[Y]} \quad (53)$$

Estimate

$$R_{0,(1,\dots,k)}^2 = \frac{\hat{\sigma}^2(Y) - \hat{\sigma}^2(Y|X_1, \dots, X_k)}{\hat{\sigma}^2(Y)} \quad (54)$$

Estimated variances

$$\hat{\sigma}^2(Y) = \sum_{i=1}^N (y_i - \hat{\mu})^2, \quad \hat{\mu} = \frac{1}{N} \sum_{i=1}^N y_i, \quad (55a)$$

$$\hat{\sigma}^2(Y|X_1, \dots, X_k) = \sum_{i=1}^N \epsilon_i^2, \quad \epsilon_i = y_i - \hat{y}_i^{lin}, \quad (55b)$$

$$\hat{y}_i^{lin} = \sum_{j=1}^k c_j x_{ji}, \quad \min_{\{c_j\}_{j=1}^k} \frac{1}{2} \sum_{i=1}^N (y_i - \hat{y}_i^{lin})^2$$

Lag-dependence function

Lag-dependence function

$$LDF(k) = \text{sign}(\hat{f}_k(b) - \hat{f}_k(a)) \sqrt{[\tilde{R}_{0,(k)}^2]^+} \quad (56)$$

$$[\cdot]^+ = \max\{0, \cdot\}$$

Estimate

$$\tilde{R}_{0,(k)}^2 = \frac{\hat{\sigma}^2(Y) - \hat{\sigma}^2(Y|X_k)}{\hat{\sigma}^2(Y)} \quad (57)$$

$\tilde{R}_{0,(k)}^2$ is computed based on the smooth, $\hat{f}_k(x)$. The $\sim$ just indicates that it is not computed based on a linear model.

The smooth is an approximation of the conditional mean

$$f_k(x) = \mathbb{E}[X_t | X_{t-k} = x] \approx \hat{f}_k(x) \quad (58)$$

Exercise: lecture02_lag_dependence_functions_logistic_map

Partial autocorrelation

Partial autocorrelation (Gaussian process)

$$\begin{aligned}\phi_{kk} &= \text{Corr}[Y_t, Y_{t+k} | Y_{t+1}, \dots, Y_{t+k-1}] \\ &= \frac{\text{Cov}(Y_t, Y_{t+k} | Y_{t+1}, \dots, Y_{t+k-1})}{\sqrt{\text{Var}(Y_t | Y_{t+1}, \dots, Y_{t+k-1}) \text{Var}(Y_{t+k} | Y_{t+1}, \dots, Y_{t+k-1})}}\end{aligned}\quad (59)$$

Yule-Walker equations

$$\begin{bmatrix} 1 & \rho(1) & \cdots & \rho(k-1) \\ \rho(1) & 1 & \cdots & \rho(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(k-1) & \rho(k-2) & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(k) \end{bmatrix} \quad (60)$$

Partial correlation coefficient

Partial correlation coefficient

$$\rho_{(0,k)|(1,\dots,k-1)}^2 = \frac{\text{Var}[Y|X_1, \dots, X_{k-1}] - \text{Var}[Y|X_1, \dots, X_k]}{\text{Var}[Y|X_1, \dots, X_{k-1}]} \quad (61)$$

Estimate

$$R_{(0,k)|(1,\dots,k-1)}^2 = \frac{\hat{\sigma}^2(Y|X_1, \dots, X_{k-1}) - \hat{\sigma}^2(Y|X_1, \dots, X_k)}{\hat{\sigma}^2(Y|X_1, \dots, X_{k-1})} \quad (62)$$

Partial lag dependence function

AR model (PACF)

$$X_t = \phi_{k0} + \phi_{k1}X_{t-1} + \cdots + \phi_{kk}X_{t-k} + e_t \quad (63)$$

Nonlinear AR model (PLDF)

$$X_t = \varphi_{k0} + f_{k1}(X_{t-1}) + \cdots + f_{kk}(X_{t-k}) + e_t \quad (64)$$

Partial lag dependence function

$$PLDF(k) = \text{sign}(\hat{f}_{kk}(b) - \hat{f}_{kk}(a))\sqrt{[\tilde{R}_{(0,k)|(1,\dots,k-1)}]^+} \quad (65)$$

Fitting the additive models

- ▶ Suggestion: Fit models of increasing order, $k = 1, \dots, K$, to compute $PLDF(K)$
- ▶ Book: Smoothing based on local polynomial regression
- ▶ For $k = 1$, the estimation problem reduces to local polynomial regression
- ▶ If convergence is not achieved for any $k = 2, \dots, K$, or if the sum of squared residuals increases compared to the previous lag, set $\hat{f}_{jk}(\cdot) = 0$ for $j = k, \dots, K$ and $\hat{f}_{kj}(\cdot) = \hat{f}_{k-1,j}(\cdot)$ for $j = 1, \dots, k - 1$

Examples

- The nonlinear autoregressive process ($NLAR(1)$)

$$X_t = \frac{1}{1 + \exp(-5X_{t-1} + 2.5)} + e_t, \quad (66)$$

- The nonlinear moving average process ($NLMA(1)$)

$$X_t = e_t + 2 \cos(e_{t-1}), \quad (67)$$

- The nonlinear and deterministic process described in the motivation (the logistic map) is called $DNLAR(1)$ in the following

Example (cont.)

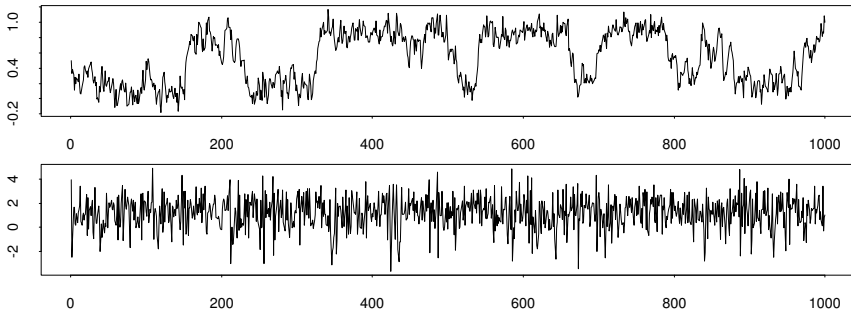
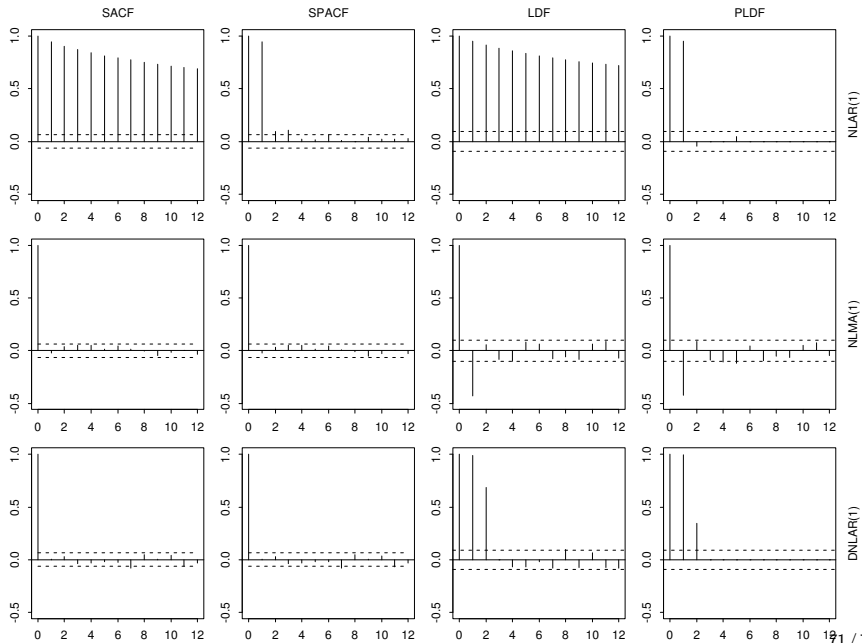


Figure: Plots of the series $NLAR(1)$ (top) and $NLMA(1)$ bottom.



Bibliography I

- [1] B. S. Everitt and D. J. Hand, *Finite mixture distributions*. Monographs on Applied Probability and Statistics, Chapman and Hall, 1981.
- [2] G. McLachlan and D. Peel, *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics Section, Wiley, 2000.
- [3] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, 1971.
- [4] H. C. Tijms, *Stochastic models: An algorithmic approach*. Wiley Series in Probability and Statistics, Wiley, 1995.
- [5] H. Cossette, D. Landriault, E. Marceau, and K. Moutanabbir, "Moment-based approximation with mixed Erlang distributions," *Variance*, vol. 10, no. 1, pp. 166–182, 2016.

Bibliography II

- [6] S. Asmussen, *Applied probability and queues*, vol. 51 of *Applications of Mathematics: Stochastic Modelling and Applied Probability*. Springer, second ed., 2003.
- [7] M. Bladt and B. F. Nielsen, *Matrix-exponential distributions in applied probability*, vol. 81 of *Probability Theory and Stochastic Modelling*. Springer, 2017.
- [8] T. K. S. Ritschel, “On Erlang ODE approximations of differential equations with distributed time delays.” arXiv:2502.12984, 2026. In submission.
- [9] G. G. Lorentz, *Bernstein polynomials*. New York: Chelsea Publishing Company, 2nd ed., 1986.