Lecture 7 Introduction to Time Series

Shiwei Lan[1]

[1]School of Mathematical and Statistical Sciences
Arizona State University

STP598 Spatiotemporal Analysis
Fall 2020

Temporal data and models: challenges

When analyzing time series data, researchers in areas such as economics, climatology, epidemiology, and neuroscience are increasingly faced with challenges:

- highly multivariate, with many important predictors and response variables,
- non-stationary, hard to predict,
- often having single history, or missing data, and
- spatially correlated, as in multi-site signals or other spatially dependent multivariate data.

Time series data

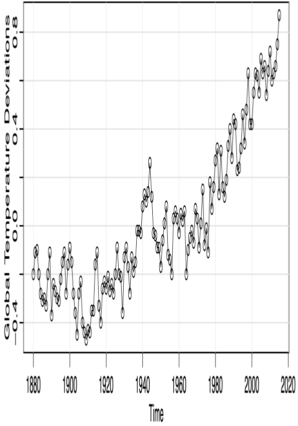

Fig. 1.2. Yearly average global temperature deviations (1880–2015) in degrees centigrade.

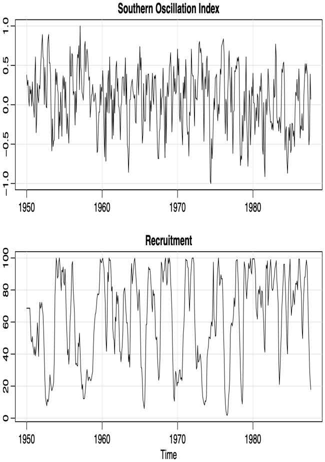

Fig. 1.5. Monthly SOI and Recruitment (estimated new fish), 1950-1987.

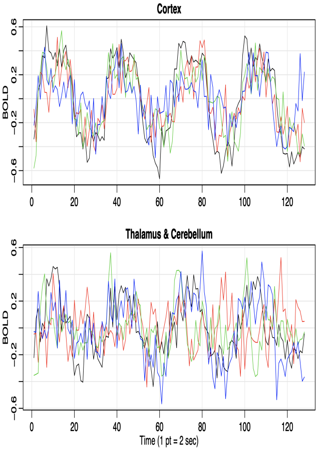

Fig. 1.6. fMRI data from various locations in the cortex, thalamus, and cerebellum; n = 128 points, one observation taken every 2 seconds.

Figure: Time Series Data: Non-stationary (left); cyclic (middle); and multivariate (right)

Time series analysis approaches

There are two separate, but not necessarily mutually exclusive methods for time series analysis:

- *time domain* approach views the investigation of lagged relationships as most important (e.g., how does what happened today affect what will happen tomorrow).
- *frequency domain* approach views the investigation of cycles as most important (e.g., what is the economic cycle through periods of expansion and recession).

In this course, we will focus on *time domain* approach.

Table of Contents

1. Characteristics of Time Series
2. Time Series Regression and Exploratory Data Analysis

Statistical Models

- We consider a time series as a sequence of random variables, $x_1, x_2, x_3, \cdots$, denoted as $\{x_t\}$, indicating random value at time t.
- The collection of random variables $\{x_t\}$ is called a *stochastic process*. The observed values of a stochastic process is termed a *realization*. *Time series* $\{x_t\}$ is generically referred to as the process or a particular realization. How to model it?
- We could model x_t as a linear combination of *white noise* $\{w_t\}$, hence named *moving average model* **MA**(q):

$$x_t = \theta(B) w_t, \quad \theta(B) = \sum_{i=0}^{q} \theta_i B^i, \; w_t \sim wn(0, \sigma_w^2) \tag{1}$$

where B is the backward operator such that $B^i w_t = w_{t-i}$.

Autoregressive Moving Average (ARMA) Models

- Or we could model x_t as a linear combination of of its history, hence named *autoregressive model* **AR**(p):

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t \tag{2}$$

  Therefore it can be written as

$$\phi(B)x_t = w_t, \quad \phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i \tag{3}$$

- Or combining moving average and autoregression to obtain **ARMA**(p, q) model:

$$\phi(B)x_t = \theta(B)w_t \tag{4}$$

Drifted Models

- A model for analyzing trend such as seen in the global temperature data is the *random walk with drift* model

$$x_t = \delta + x_{t-1} + w_t \tag{5}$$

- The constant δ is called the *drift*. When $\delta = 0$, x_t is simply a *random walk*.
- The process can be rewritten as a cumulative sum of white noise variates:

$$x_t = \delta t + \sum_{j=1}^{t} w_j \tag{6}$$

- In general, we might want to write time series x_t in the simple additive format

$$x_t = s_t + v_t \tag{7}$$

where s_t denotes some unknown signal and v_t denotes a time series that may be white or correlated over time.

Measures of Dependence

- The marginal distribution functions of time series

$$F_t(x) = \Pr\{x_t \leq x\} \tag{8}$$

- The corresponding marginal density functions, if exist,

$$f_t(x) = \frac{\partial F_t(x)}{\partial x} \tag{9}$$

- The **mean function** is defined as

$$\mu_{xt} = \mathrm{E}(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx \tag{10}$$

provided it exists. For simplicity we may denote μ_{xt} as μ_t.

Measures of Dependence

- The **autocovariance function** is defined as the second moment product

$$\gamma_x(s,t) = \mathrm{Cov}(x_s, x_t) = \mathrm{E}[(x_s - \mu_s)(x_t - \mu_t)] \tag{11}$$

for all s and t. For simplicity we may denote $\gamma_x(s,t)$ as $\gamma(s,t)$.

Example

Compute the autocovariances of: 1) a moving average $x_t = \frac{1}{3}(w_{t-1} + w_t + w_{t+1})$; 2) a random walk $x_t = \sum_{j=1}^{t} w_j$.

Measures of Dependence

- The **autocorrelation function (ACF)** is defined as

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}} \tag{12}$$

- The ACF measures the linear predictability of the series at time t, say x_t, using only the value x_s.
- The **cross-covariance function** between two series x_t and y_t is

$$\gamma_{xy}(s,t) = \mathrm{Cov}(x_s, y_t) = \mathrm{E}[(x_s - \mu_{xs})(y_t - \mu_{yt})] \tag{13}$$

- There is also a scaled version of the cross-covariance function, **cross-correlation function (CCF)** given by

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\gamma_y(t,t)}} \tag{14}$$

Stationary Time Series

- A **strictly stationary** time series is one for which the probabilistic behavior of every collection of values $\{x_{t_1}, \cdots, x_{t_k}\}$ is identical to that of the time shifted set $\{x_{t_1+h}, \cdots, x_{t_k+h}\}$. That is

$$\Pr\{x_{t_1} \le c_1, \cdots, x_{t_k} \le c_k\} = \Pr\{x_{t_1+h} \le c_1, \cdots, x_{t_k+h} \le c_k\} \tag{15}$$

  for all $k = 1, 2, \cdots$ and all time points $t_1, \cdots, t_k$, all numbers $c_1, \cdots, c_k$ and all time shifts $h = 0, \pm 1, \cdots$.
- A **weakly stationary** time series, x_t, is a finite variance process such that
  1. the mean value function, μ_t, is constant and does not depend on time t and
  2. the autocovariance function, $\gamma(s, t)$, depends on s and t only through their difference $|s - t|$.
- Two time series, x_t and y_t, are said to be **jointly stationary** if they are each stationary, and the cross-covariance function

$$\gamma_{xy}(h) = \mathrm{Cov}(x_{t+h}, y_t) = \mathrm{E}[(x_{t+h} - \mu_x)(y_t - \mu_y)] \tag{16}$$

  is a function only of lag h.

Stationary Time Series

- The **autocovariance function of a stationary time series** will be written as

$$\gamma(h) = \mathrm{Cov}(x_{t+h}, x_t) = \mathrm{E}[(x_{t+h} - \mu)(x_t - \mu)] \tag{17}$$

- The **autocorrelation function (ACF) of a stationary time series** will be written as

$$\rho(h) = \frac{\gamma(t+h, t)}{\sqrt{\gamma(t+h, t+h)\gamma(t, t)}} = \frac{\gamma(h)}{\gamma(0)} \tag{18}$$

- The **cross-correlation function (CCF)** of jointly stationary time series x_t and y_t is defined as

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}} \tag{19}$$

Example

1) Plot ACF of a moving average $x_t = \frac{1}{3}(w_{t-1} + w_t + w_{t+1})$; 2) Is random walk a stationary time series?

Stationary Time Series

- A **linear process**, x_t, is defined to be a linear combination of white noise variates w_t, given by

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty \tag{20}$$

- We may show that the autocovariance function is given by

$$\gamma_x(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j \tag{21}$$

- A process, $\{x_t\}$, is said to be a **Gaussian process** if the n-dimensional vectors $x = (x_{t_1}, \cdots, x_{t_n})'$, for every collection of distinct time points $t_1, \cdots, t_n$, and every positive integer n, have a multivariate normal distribution.

Estimation of Correlation

- If a time series is stationary, the mean function $\mu_t = \mu$ is constant and estimated by the *sample mean*

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t \tag{22}$$

- Its variance can be computed

$$\begin{aligned}
\mathrm{Var}(\bar{x}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{t=1}^{n} x_t\right) = \frac{1}{n^2}\mathrm{Cov}\left(\sum_{t=1}^{n} x_t, \sum_{s=1}^{n} x_s\right) \\
&= \frac{1}{n^2}\left(n\gamma_x(0) + (n-1)\gamma_x(1) + \cdots + \gamma_x(n-1)\right. \\
&\qquad \left. + (n-1)\gamma_x(-1) + \cdots + \gamma_x(1-n)\right) \\
&= \frac{1}{n}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma_x(h)
\end{aligned}$$

Estimation of Correlation

- The **sample autocovariance function** is defined as

$$\hat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x}) \tag{23}$$

  with $\hat{\gamma}(-h) = \hat{\gamma}(h)$ for $h = 0, 1, \cdots, n-1$.

- The variances of linear combinations of the variates x_t can be estimated $\widehat{\mathrm{Var}}(\sum_{j=1}^n a_j x_j) = \sum_{j=1}^n \sum_{k=1}^n a_j a_k \hat{\gamma}(j-k)$.
- The **sample autocorrelation function** is defined as

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \tag{24}$$

- If x_t is iid with finite fourth moment, then $\hat{\rho}_x(h) \xrightarrow{d} N(0, 1/\sqrt{n})$.

Estimation of Correlation

- The estimators of the cross-covariance function, $\gamma_{xy}(h)$ can be given by **sample cross-covariance function**

$$\hat{\gamma}_{xy}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y}) \tag{25}$$

  where $\hat{\gamma}_{xy}(-h) = \hat{\gamma}_{yx}(h)$ deermines the function for negative lags.

- The estimators of the cross-correlation, $\rho_{xy}(h)$ can be given by **sample cross-correlation function**

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}} \tag{26}$$

- For x_t and y_t independent linear processes, we have $\hat{\rho}_{xy}(h) \xrightarrow{d} N(0, 1/\sqrt{n})$ if at least one of the process is independent of white noise.

Vector-Valued Series

- A vector time series $x_t = (x_{t1}, \cdots, x_{tp})'$ contains as its components p univariate time series.
- For the stationary case, we define the mean vector $\mu = (\mu_{t1}, \cdots, \mu_{tp})' = \mathrm{E}(x_t)$.
- the $p \times p$ autocovariance matrix

$$\Gamma(h) = \mathrm{E}[(x_{t+h} - \mu)(x_t - \mu)'] \tag{27}$$

- The elements of the matrix $\Gamma(h)$ are the cross-covariance functions

$$\gamma_{ij}(h) = \mathrm{E}[(x_{t+h,i} - \mu_i)(x_{tj} - \mu_j)] \tag{28}$$

- Their sample estimates are $\bar{x} = n^{-1} \sum_{t=1}^{n} x_t$ and $\hat{\Gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x})'$ respectively.

Multidimensional Series

- The *autocovariance function* of a stationary multidimensional process, x_s, can be defined as a function of the multidimensional lag vector, $h = (h_1, \cdots, h_r)'$

$$\gamma(h) = \mathrm{E}[(x_{s+h} - \mu)(x_s - \mu)'], \quad \mu = \mathrm{E}(x_s) \tag{29}$$

- The *multidimensional sample autocovariance function* is defined as

$$\hat{\gamma}(h) = (S_1 \cdots S_r)^{-1} \sum_{s_1} \cdots \sum_{s_r} (x_{s+h} - \bar{x})(x_s - \bar{x}) \tag{30}$$

  where $s = (s_1, \cdots, s_r)'$ and the range of the summation for each argument is $1 \leq s_i \leq S_i - h_i$ for $i = 1, \cdots, r$.
- The mean is computed over the r-dimensional array

$$\bar{x} = (S_1 \cdots S_r)^{-1} \sum_{s_1} \cdots \sum_{s_r} x_{s_1, \cdots, s_r} \tag{31}$$

- The multidimensional sample autocorrelation function follows $\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$.

Table of Contents

1. Characteristics of Time Series
2. Time Series Regression and Exploratory Data Analysis

In this section, we will discuss:

- classical multiple linear regression in a time series context,
- model selection,
- exploratory data analysis for preprocessing nonstationary time series,
- differencing and the backshift operator,
- variance stabilization,
- nonparametric smoothing of time series.

Classical Regression

- Given *dependent* time series, x_t, for $t = 1, \cdots, n$, we consider the classical regression model with *independent* series, i.e. $z_{t1}, \cdots, z_{tq}$:

$$x_t = \beta_0 + \beta_1 z_{t1} + \cdots + \beta_q z_{tq} + w_t \tag{32}$$

where $\beta_0, \beta_1, \cdots, \beta_q$ are unknown fixed regression coefficients, and $\{w_t\}$ is a random error or noise process consisting of iid $N(0, \sigma_w^2)$ variables.

- Then the classical linear regression theories apply.
- Denote $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_q)'$, and $\mathbf{z}_t = (1, z_{t1}, \cdots, z_{tq})'$. We minimize the error sum of squares $Q = \sum_{t=1}^n w_t^2 = \sum_{t=1}^n (x_t - \boldsymbol{\beta}'\mathbf{z}_t)^2$ to obtain the ordinary least square (OLS) estimate of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^n \mathbf{z}_t \mathbf{z}_t' \right)^{-1} \sum_{t=1}^n \mathbf{z}_t x_t \tag{33}$$

Classical Regression

- The minimized error sum of squares, Q, denoted as SSE, can be written as $\mathrm{SSE} = \sum_{t=1}^{n}(x_t - \hat{\boldsymbol{\beta}}'\mathbf{z}_t)^2$.
- We have

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad \mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma_w^2 C, \quad C = \left(\sum_{t=1}^{n} \mathbf{z}_t \mathbf{z}_t'\right)^{-1} \tag{34}$$

- An unbiased estimator for the variance σ_w^2 is

$$s_w^2 = \mathrm{MSE} = \frac{\mathrm{SSE}}{n-(q+1)} \tag{35}$$

- Under the normal assumption, we have

$$t_i = \frac{\hat{\beta}_i - \beta_i}{s_w\sqrt{c_{ii}}} \sim t(df = n-(q+1)) \tag{36}$$

Classical Regression

- While t_i is often used for individual tests of the null hypothesis $H_0 : \beta_i = 0$ for $i = 1, \cdots, q$, for the joint test $H_0 : \beta_{r+1} = \cdots = \beta_q = 0$ for fixed $r \in \{0, q-1\}$, we consider the following F-statistic

$$F = \frac{(\mathrm{SSE}_r - \mathrm{SSE})/(q-r)}{\mathrm{SSE}/(n-q-1)} = \frac{\mathrm{MSR}}{\mathrm{MSE}} \sim F(df_1 = q-r, df_2 = n-q-1) \quad (37)$$

where SSE_r is the error sum of squares under the reduced model

$$x_t = \beta_0 + \beta_1 z_{t1} + \cdots + \beta_r z_{tr} + w_t \quad (38)$$

Table: Analysis of Variance (ANOVA) for Regression

Source	df	Sum of Squares	Mean Square	F
$z_{t,r+1:q}$	$q-r$	$\mathrm{SSR} = \mathrm{SSE}_r - \mathrm{SSE}$	$\mathrm{MSR} = \mathrm{SSR}/(q-r)$	$F = \frac{\mathrm{MSR}}{\mathrm{MSE}}$
Error	$n-(q+1)$	SSE	$\mathrm{MSE} = \mathrm{SSE}/(n-q-1)$	

Classical Regression

- For a special case $r = 0$, which corresponds to the following trivial model

$$x_t = \beta_0 + w_t \tag{39}$$

$\mathrm{SSE}_0 = \sum_{t=1}^{n}(x_t - \bar{x})^2$ measures the total variation of the time series x_t.

- The proportion accounted by all variables is defined as *coefficient of determination*

$$R^2 = \frac{\mathrm{SSE}_0 - \mathrm{SSE}}{\mathrm{SSE}_0} \tag{40}$$

- We could select models by joint tests in a stepwise manner.

Model Selection

- Alternatively, we could select models based on some criteria involving, e.g. the *maximum likelihood estimator*for the variance

$$\hat{\sigma}_k^2 = \frac{\mathrm{SSE}(k)}{n} \tag{41}$$

where $\mathrm{SSE}(k)$ denotes the residual sum of squares under the model with k regression coefficients.

- Considered **Akaike's Information Criterion (AIC)** (1969, 1973, 1974)

$$\mathrm{AIC} = \log \hat{\sigma}_k^2 + \frac{n+2k}{n} \tag{42}$$

- or **Bias Corrected (AICc) AIC**

$$\mathrm{AICc} = \log \hat{\sigma}_k^2 + \frac{n+k}{n-k-2} \tag{43}$$

- Or **Bayesian Information Criterion (BIC)** (Schwarz, 1978)

$$\mathrm{BIC} = \log \hat{\sigma}_k^2 + \frac{k \log n}{n} \tag{44}$$

Exploratory Data Analysis

- To achieve any meaningful statistical analysis of time series data, the mean and the autocovariance functions should satisfy the conditions of stationarity.
- We can start with the *trend stationary* model wherein the process has stationary behavior around a trend:

$$x_t = \mu_t + y_t \tag{45}$$

where x_t are the observations, μ_t denotes the trend, and y_t is a stationary process.

- Then we need to obtain an estimate of the trend, $\hat{\mu}_t$ and then work with the residuals

$$\hat{y}_t = x_t - \hat{\mu}_t \tag{46}$$

- We could model trend μ_t using a linear model $\mu_t = \beta_0 + \beta_1 t$. Alternatively, We might model trend as a stochastic component using the random walk with drift model

$$\mu_t = \delta + \mu_{t-1} + w_t \tag{47}$$

Differencing

- Based on the trend stationary model, we could obtain a stationary process by *differencing* the data:

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \delta + w_t + y_t - y_{t-1} \tag{48}$$

  where we denote $z_t = y_t - y_{t-1}$ which is also a stationary process. (Why?)
- Pro: no need to estimate any parameters.
- Con: no estimate for the stationary process y_t either.
- What do we get by differencing the data with a linear trend model?

Differencing

- Based on the trend stationary model, we could obtain a stationary process by *differencing* the data:

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \delta + w_t + y_t - y_{t-1} \tag{48}$$

  where we denote $z_t = y_t - y_{t-1}$ which is also a stationary process. (Why?)
- Pro: no need to estimate any parameters.
- Con: no estimate for the stationary process y_t either.
- What do we get by differencing the data with a linear trend model?

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \beta_1 + y_t - y_{t-1} \tag{49}$$

- Because differencing plays a central role in time series analysis, it receives its own notation. The first difference is denoted as $\nabla x_t = x_t - x_{t-1}$.
- The first difference eliminates a linear trend. A second difference can eliminate a quadratic trend. We will have more discussion in ARIMA models.

Backshift operators

- We can define the **backshift operator** by

$$Bx_t = x_{t-1}, \quad B^k x_t = x_{t-k} \tag{50}$$

- Then the inverse B^{-1} is called *forward-shift operator*: $B^{-1}x_t = x_{t+1}$.
- Then we can write the differencing as

$$\nabla x_t = (1 - B)x_t \tag{51}$$

  and extend to the second difference as

$$\nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2} \tag{52}$$

- The **Differences of order** d are defined as

$$\nabla^d = (1 - B)^d \tag{53}$$

Other trend-removing operations

- An alternative to differencing as a less-severe operation that still assumes stationarity is *fractional differencing*. It extends the difference operators to fractional powers $-0.5 < d < 0.5$.
- Granger and Joyeux (1980) and Hosking (1981) introduced long memory time series, which corresponds to $0 < d < 0.5$.
- To nonlinear trends, we can consider *transformations* such as the log-transformation
$$y_t = \log x_t \tag{54}$$
to suppress larger fluctuations.
- Other possibilities are *power transformations* in the Box–Cox family
$$y_t = \begin{cases} (x_t^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log x_t, & \lambda = 0 \end{cases} \tag{55}$$

Cyclic or periodic signals

- To identify cyclic or periodic signals in the time series, we could use the following sinusoidal model

$$x_t = A\cos(2\pi\omega t + \phi) + w_t = \beta_1 \cos(2\pi\omega t) + \beta_2 \sin(2\pi\omega t) + w_t \tag{56}$$

  where $\beta_1 = A\cos(\phi)$ and $\beta_2 = -A\sin(\phi)$.

- We can use frequency method to estimate ω, and then obtain OLS for β_1 and β_2.

Smoothing

- Smoothing is useful in discovering certain traits in a time series, such as long-term trend and seasonal components. In particular, if x_t represents the observations, then

$$m_t = \sum_{j=-k}^{k} a_j x_{t-j} \tag{57}$$

where $a_{-j} = a_j \geq 0$ and $\sum_{j=-k}^{k} a_j = 1$ is a symmetric moving average of the data.

- Alternatively, we could consider *kernel smoothing* that uses a weight function

$$m_t = \sum_{i=1}^{n} w_i(t) x_i, \quad w_i(t) = K\left(\frac{t-i}{b}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{t-j}{b}\right) \tag{58}$$

- $K(\cdot)$ is a kernel function, originally explored by Parzen (1962) and Rosenblatt (1956), which can be chosen as a normal kernel $K(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$, called Nadaraya-Watson estimator (Watson, 1966).

Smoothing

- Another smoother named `lowess`, is based on the basic idea of k-nearest neighbors regression.
- Other smoothers include *cubic splines*

$$m_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 \tag{59}$$

which requires fitted function continuous up to 2nd order at *knots*,

- and *smoothing splines* which minimizes a compromise between the fit and the degree of smoothness given by

$$\sum_{t=1}^{n} [x_t - m_t]^2 + \lambda \int (m_t'')^2 dt \tag{60}$$

where m_t is a cubic spline with a knot at each t.

- The degree of smoothness is controlled by $\lambda > 0$, as a trade-off between linear regression (completely smooth, $\lambda = \infty$) and the data itself (no smoothness $\lambda = 0$).