1 General definitions

1.1 Basic

- Sample variance

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1.1}$$

- Sample correlation coefficient

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)\sqrt{(S_X^2 S_Y^2)}} \tag{1.2}$$

- QQ-plot for cumulative distribution function F is the set of points $\left(q_F\left(\frac{i}{n+1}\right), x_{(i)}\right)$, where $q_F(\cdot)$ is the quantile function for the distribution.

- Mean Squared Error (MSE)

$$\mathrm{MSE}(\theta; T(X), g(\theta)) = \mathbb{E}_\theta\left(T(X) - g(\theta)\right)^2 \tag{1.3}$$

- Bias-variance decomposition

$$\mathrm{MSE}(\theta; T(X)) = \mathrm{var}_\theta T + \left(\mathbb{E}_\theta T(X) - g(\theta)\right)^2 \tag{1.4}$$

- Empirical distribution function

$$\hat{F}_n(x) = \sum_{i=1}^{n}\mathbb{I}(X_i \leq x) \tag{1.5}$$

1.2 k-th order statistic $X_{(k)}$

$X_{(k)}$ — $k-th$ order statistic distribution for n i.i.d. variables from continuous distribution F.

$$f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!}F(x)^{k-1}(1-F(x))^{n-k}f(x) \tag{1.6}$$

$$F_{(k)}(x) = \sum_{j=k}^{n}\binom{n}{j}F(x)^j(1-F(x))^{n-j} \tag{1.7}$$

$$\mathbb{E}F(X_{(k)}) = \frac{k}{n+1} \tag{1.8}$$

1.3 Time Series

Time series below are assumed to be *weakly stationary* in the following sense:

- Stochastic time series $X_t(\omega)$ are called *weakly stationary*, if $\mathbb{E}X_t$ and $\mathbf{cov}(X_t, X_s)$ are inependent of time shifts; in particular, it first and second moments exist.
- For a weakly stationary time series X_t, the following functios are defined:
  - *Autocovariance funciton*:
  $$\gamma_X(h) = \mathbb{E}(X_{t+h}, X_t) \tag{1.9}$$
  - *Autocorrelation function*
  $$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} \tag{1.10}$$
  - *Partial autocorrelation function* $\varphi(h)$ is defined as the coefficient the regression of X_{t+h} on X_t when controlled for constant and $X_{t+1}, \ldots X_{t+h-1}$ (see Proposition 3.23). In particular, $\varphi(0) = 1$ and $\varphi(1) = \rho(1)$

2 Important distributions

- Student's t-distribution t_ν, $\nu \in \mathbb{R}_{>0}$
  - pdf
  $$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{2.1}$$
  - cdf
  $$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \frac{{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}, -\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma(\frac{\nu}{2})} \tag{2.2}$$
  - t-distribution with $n \in \mathbb{N}$ degrees of freedom arises from the ratio of independent $N(0,1)$- and χ^2_n-distributions

- Poisson distribution Poisson(λ), $\lambda > 0$
  - λ is the average number of events per interval
  - pdf
$$p_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{2.3}$$
- Geometric distribution $G(\theta)$, $0 \le \theta \le 1$
  - pdf
$$f_\theta(k) = (1-\theta)^{1-k}\theta \tag{2.4}$$
  - cdf
$$F_\theta(k) = 1-(1-\theta)^k \tag{2.5}$$
- Exponential distribution $F(x;\lambda)$
  - pdf
$$f_\lambda(x) = \lambda e^{-\lambda x} \tag{2.6}$$
  - cdf
$$F_\lambda(x) = 1 - e^{-\lambda x} \tag{2.7}$$
  - $\mathbb{E}_\lambda X = 1/\lambda$
- Beta distribution $B(\alpha,\beta)$, $\alpha,\beta > 0$
  - pdf
$$f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},\ B(\alpha,\beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{2.8}$$
- Weibull distribution
  - α and λ are the "shape" and 'inverse scale" parameters.
  - pdf
$$f_{\lambda,\alpha}(x) = \lambda^\alpha \alpha x^{\alpha-1}e^{-(\lambda x)^\alpha} \tag{2.9}$$
  - cdf
$$F_{\lambda,\alpha}(x) = 1 - e^{-(\lambda x)^\alpha} \tag{2.10}$$
- Gamma distribution $\Gamma(\alpha,\lambda)$, $\alpha > 0, \lambda > 0$

- α and λ are known as "shape" and "inverse scale" parameters.
- pdf
$$f_{\alpha,\lambda}(x) = \frac{x^{\alpha-1}\lambda^{\alpha}e^{-\lambda x}}{\Gamma(\alpha)} \tag{2.11}$$
- cdf (where $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt$ — is the "incomplete gamma function")
$$F_{\alpha,\lambda}(x) = \frac{\gamma(\alpha, x\beta)}{\Gamma(\alpha)} \tag{2.12}$$

Definition 2.1. A family of probability densities p_θ that depends on a parameter θ is called a k-dimensional exponential family if there exist functions $c(\theta)$, $h(x)$, $Q_j(\theta)$, and $V_j(x)$ such that

$$p_\theta(x) = c(\theta)h(x)e^{\sum_{j=1}^{k} Q_j(\theta)V_j(x)}$$

3 Fundamental results

3.1 Basic statistics

Theorem 3.1. *Let $X_1, \dots X_n$ be an i.i.d. random variables from the $N(\mu, \sigma^2)$ distribution, then*

1. *$\bar{X}$ is $N(\mu, \sigma^2/n)$ distributed;*
2. *$(n-1)S_X^2/\sigma^2$ is χ^2_{n-1}-distributed (see 1.1);*
3. *$\bar{X}$ and S_X^2 are independent;*
4. *$\sqrt{n}(\bar{X}-\mu)/\sqrt{S_X^2}$ has the t_{n-1}- distribution.*

Proof. $\|X\|^2 - n\bar{X}^2 = (n-1)S_X^2$

Definition 3.2. Let $A \in \mathbb{C}^{m\times n}$ and $A^{\sim} \in \mathbb{C}^{n\times m}$. Consider a list of conditions:

1. $AA^{\sim}A = A$
2. $A^{\sim}AA^{\sim} = A^{\sim}$
3. $(AA^{\sim})^* = AA^{\sim}$

4. $(A^{\sim}A)^* = A^{\sim}A$

If $A^{\sim}$ satisfies the first condition, then it is a *generalized inverese* of A. If it satisfies the first two conditions, then it is a *reflexive generalized inverese* of A. If A satisfies all four conditions the it is called the *Moore-Penrose inverse.*

Proposition 3.3. *The space of generalized inverse matrices to a given matrix has dimension* $2 \operatorname{rk} A \operatorname{corank} A$. *It is characterized by two properties:*

- *The restriction of* $A^{\sim}$ *on* $\operatorname{Im} A$ *is an arbitrary lift of the inverse of induced isomorphism on the quotient by* $\operatorname{Ker} A$.
- *The image of* $A^{\sim}$ *equals* $A^{\sim} \operatorname{Im} A$.

Theorem 3.4. *Let* $X \sim N(\mu_X, \mathbb{V}_X)$ *be a* k*-dimensional random vector, where* $\mathbb{V}_X$ *might be singular. Then for any linear condition of the form* $BX = b$ *and any linear transformation* AX, *conditional distributions of* X *and* AX *are given by the following formulae:*

$$\begin{aligned} f(X|BX) &\sim N(\mu_X + \mathbb{V}_X B^T (B\mathbb{V}_X B^T)^{\sim}(BX - B\mu_X), \\ &\qquad \mathbb{V}_X - \mathbb{V}_X B^T (B\mathbb{V}_X B^T)^{\sim} B\mathbb{V}_X) \\ f(AX|BX) &\sim N(A\mu_X + A\mathbb{V}_X B^T (B\mathbb{V}_X B^T)^{\sim}(BX - B\mu_X), \\ &\qquad A\mathbb{V}_X A^T - A\mathbb{V}_X B^T (B\mathbb{V}_X B^T)^{\sim} B\mathbb{V}_X A^T) \end{aligned}$$

where $A^{\sim}$ *denotes reflexive generalized inverse of matrix* A.

Proof. Follows from the proposition below.

Proposition 3.5. *Let* Y *be a Gaussian random vecor such that*

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \right)$$

Then Y_1 *conditional on* Y_2 *is Gaussian:*

$$f(Y_1|Y_2) \sim N(V_{12}V_{22}^{\sim}Y_2, V_{11} - V_{12}V_{22}^{\sim}V_{21})$$

Example 3.6. Let

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Then

$$f(y|x) \sim N(\rho x, 1 - \rho^2)$$

Proof of Proposition 3.5. We only consider the case with $\mathbb{V}(Y)$ nonsingular. It is generalized in a straightforward way, but does not admit straightforward matrix notation for Gaussian distributions.

By definition,

$$f(Y_1|Y_2 = y_2) = \frac{f(Y_1, y_2)}{\int_{Y_2=y_2} f(Y_1, y_2)dY_1} \tag{3.1}$$

It is enough to show that

$$Y^T\mathbb{V}(Y)^{-1}Y = (Y_1 - V_{12}V_{22}^{-1})^T(V_{11} - V_{12}V_{22}^{-1}V_{21})^{-1}(Y_1 - V_{12}V_{22}^{-1}) + C(Y_2) \tag{3.2}$$

where $C(Y_2)$ is a term that does not depend on Y_1. Note that in (3.1) the part of the exponential in Gaussian density depending on $C(Y_2)$ cancels out. After the cancellation, we are left with the required density up to a multiplicative constant. The latter has to have correct magnitude so that the whole expression corresponds to a density function; alternatively it can be computed directly.

The equation (3.2) is a direct consequence of the inversion formula for 2×2 block matrix.

3.2 Reminder on different convergence types

Definition 3.7. Let X_n be a sequence of random variables defined on the probability space $(\Omega, \mathbb{P})$:

- X_n is said to converge to X *almost surely* if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$
- convergence in probability
- weak convergence
- L_p-convergence

Theorem 3.8. *If a sequence of random variables converges almost surely then it converges in probability.*

Proposition 3.9. *Assume $\sum_{n\in\mathbb{Z}} \mathbb{E}|X_n| < \infty$, then $\sum_{n\in\mathbb{Z}} X_n$ is defined as an almost sure limit and:*

$$\mathbb{E}\sum_{n\in\mathbb{Z}} X_n = \sum_{n\in\mathbb{Z}} \mathbb{E}X_n$$

Proof. Established using Monotone and Dominated Convergence theorems applied to partial sums of random variables.

Proposition 3.10. *Let $f(x)$ be a function of period T that is absolutely integrable on the interval $[-T/2, T/2]$. Define its Fourier coefficients as follows:*

$$a_k = \frac{2}{T}\int_{-T/2}^{T/2} f(x)\cos\frac{2\pi kx}{T}dx,\ \ k = 0, 1, \dots$$
$$b_k = \frac{2}{T}\int_{-T/2}^{T/2} f(x)\sin\frac{2\pi kx}{T}dx,\ \ k = 1, 2, \dots$$

Then the partial sum $S_n(x) \stackrel{\text{def}}{=} a_0/2 + \sum_{k=1}^{n}\left(a_k \cos\frac{2\pi kx}{T} + b_k \sin\frac{2\pi kx}{T}\right)$ of the Fourier series admits an integral presentation:

$$S_n(x) = \frac{2}{T}\int_{-T/2}^{T/2} f(x+u)\frac{\sin(n+\frac{1}{2})\frac{2\pi u}{T}}{\sin\frac{\pi u}{T}}dx$$

Remark: Note that the coefficient a_0 defined by the formulas above is twice the coefficient to of the constant function in Fourier series (as the norm of a sine wave is twice less).

Theorem 3.11. *Let $f(x)$ be a function of period T that is absolutely integrable on the interval $[-T/2, T/2]$.*

1. *At a point of continuity where $f(x)$ has a right and a left derivative, the Fourier series coverge absolutely to the value $f(x)$*

2. *If $f(x)$ is continuous and its derivative $f'(x)$ is square integrable, then the Fourier series converge to $f(x)$ absolutely and uniformly*

3. *If $f(x)$ is continuous, then the Fourier series are uniformally summable to $f(x)$ by the method of Cesàro*

Theorem 3.12. *If series $\sum_{k=1}^{\infty}(|a_k| + |b_k|)$ are absolutely summable then the associated trigonometric series*

$$\frac{a_0}{2} + \sum_{k=1}^{\infty}\left(a_k \cos\frac{2\pi kx}{T} + b_k \sin\frac{2\pi kx}{T}\right)$$

converges absolutely and uniformly to a continuous periodic function of period T of which it is a Fourier series

Proposition 3.13. *Let $(a_n)_{n\in\mathbb{Z}}$ be an element of $l^1(\mathbb{Z})$, and let $(Z_t)_{t\in\mathbb{Z}}$ be a sequence of random variables satisfying $\mathbb{E}|Z_t| < C_1$ $\forall t$ for some constant C_1. Then the convolution*

$$X_t = \sum_{n\in\mathbb{Z}} a_n Z_{t-n}$$

is defined almost surely $\forall t \in \mathbb{Z}$.

Moreover, if there exists a constant such that $\mathbb{E}Z_t^2 < C_2$ $\forall t$, then X_t is also a limit in L^2-norm and $\mathbb{E}X_t^2 \le |a_n|_1^2 C_2$.

Theorem 3.14. *Let $(X_t)_{t\in\mathbb{Z}} \in L^2\left(\Omega, \mathbb{P}, \mathbb{C}^{d_X}\right)$ and $(Y_t)_{t\in\mathbb{Z}} \in L^2\left(\Omega, \mathbb{P}, \mathbb{C}^{d_Y}\right)$. Let $(a_m)_{m\in\mathbb{Z}} \in l^1(\mathbb{Z}, \mathbb{C}^{d'_X \times d_X})$ and $(b_n)_{n\in\mathbb{Z}} \in l^1(\mathbb{Z}, \mathbb{C}^{d'_Y \times d_Y})$. Then*

1. *Convolutions $(a * X)_t$ and $(b * Y)_t$ are well-defined elements of $L^2\left(\Omega, \mathbb{P}, \mathbb{C}^{d'_X}\right)$ and $L^2\left(\Omega, \mathbb{P}, \mathbb{C}^{d'_Y}\right)$.*

2. $$\gamma_{a*X,b*Y}(h) = \sum_{m,n\in\mathbb{Z}} a_m \gamma_{X,Y}(h+n-m)\bar{b}_n^T$$

Corollary 3.15. *Let (a_m), $(b_n) \in l^1(\mathbb{Z})$ and $(e_t)_{t\in\mathbb{Z}}$ be a sequence of uncorrelated $(0, \sigma^2)$ random variables. Denote Let $X_t = (a * e)_t$, $Y_t = (b * X)_t$. Then*

1. $$\gamma_X(h) = \sum_{n\in\mathbb{Z}} a_m a_{m-h} \sigma^2$$

2. $$\gamma_Y(h) = \sum_{n\in\mathbb{Z}} c_n c_{n-h} \sigma^2, \tag{3.3}$$

   where $c_n = (a * b)_n$.

Theorem 3.16. *Let $(a_m) \in l^1(\mathbb{Z})$, $(b_n) \in l^2(\mathbb{Z})$ and $(e_t)_{t\in\mathbb{Z}}$ be a sequence of uncorrelated $(0, \sigma^2)$ random variables. Denote $X_t = (a * e)_t$, then*

1. *For each $t \in \mathbb{Z}$ random variable Y_t defined as L_2-limit of the sequence $\sum_{n=-N}^{N} b_n X_{t-n}$ is well defined.*

2. *Formula (3.3) holds for Y_t.*

Corollary 3.17. *For $(b_n) \in l^2(\mathbb{Z})$ the sequence $(b * e)_t$ is defined in $L^2(\Omega, \mathbb{P})$*

Theorem 3.18. *Conclusions of Theorem 3.16 hold for $(a_m) \in l^2(\mathbb{Z})$, $(b_n) \in l^1(\mathbb{Z})$*

3.3 Time Series

Proposition 3.19. *Given the difference equation of order n:*

$$y_t + a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_n y_{t-n} = r_t,\ t = n, n+1, \ldots \tag{3.4}$$

The solution $(y_n)_{n=0}^{\infty}$ *can be expressed in the form:*

$$y_t = \sum_{i=0}^{t} w_i r_{t-i},\ t = 0, 1, \ldots$$

with $w_i \equiv 0$ *for* $i < 0$*, and satisfying the homogeneous differrence equaition of* (y_t)*:*

$$\sum_{i=0}^{n} a_i w_{t-i} = 0,\ i = 1, 2, \ldots \tag{3.5}$$

Proof. Using recursive formula (3.4) for y_t, any element of the sequence can be written in the form:

$$y_t = \sum_{i=0}^{t} w_i^{(t)} r_{t-i}$$

for some $w_i^{(t)}$ with $t \in \mathbb{Z}_{\geq 0},\ 0 \leq i \leq t$. Using inductive argument, we show that

1. $w_i^{(t)} \equiv w_i$ for some constant depending on i only
2. The sequence w_i satisfies equation (3.5)

Indeed, from the equation 3.4 it is immediate to see that $w_0 = 1$ and $w_1 = -a_1$. Assuming that the pair of statements above is shown for $s < t$, write:

$$y_t = r_t - \sum_{k=0}^{n} a_k^{(t)} y_{t-k} = r_t - \sum_{k=1}^{t} a_k \sum_{j=0}^{t-k} w_j r_{t-k-j} \tag{3.6}$$

By definition, we set w_t to be the coefficient of r_0, that is, $w_t = -\sum_{i=1}^{n} a_i w_{t-i}$. Moreover, by rearranging terms in 3.6, we get:

$$y_t = -\sum_{j=0}^{t} r_j \sum_{k=1}^{t-j} a_k w_{t-j-k},$$

it follows that the coefficient $w_{t-j}^{(t)}$ of r_j, $j = 1, \dots t$, is equal to w_{t-j} by induction assumpiton and it concludes the proof.

Theorem 3.20. *1. Let X_t be an AR(p) process given by*

$$X_t + \sum_{j=0}^{p} a_j X_{t-j} = e_t,$$

where $(e_t)_{t\in\mathbb{Z}}$ is a sequence of uncorrelated $(0, \sigma^2)$ random variables. Suppose that all roots of the polynomial

$$m^p + \sum_{j=0}^{p} a_j m^{p-j}$$

have magnitude less than 1. Then X_t admits an infinite MA-presentation $X_t = \sum_{j=0}^{\infty} w_j e_{t-j}$, where

$$\begin{aligned} &w_0 = 1 \\ &w_j + \sum_{i=1}^{j} a_i w_{j-i} = 0, \ j = 1, \dots p-1 \\ &w_j + \sum_{i=1}^{p} a_i w_{j-i} = 0, \ j = p, p+1, \dots \end{aligned}$$

2. Let X_t be an $MA(q)$ process given by

$$X_t = e_t + \sum_{j=0}^{q} b_j e_{t-j}$$

where $(e_t)_{t\in\mathbb{Z}}$ is a sequence of uncorrelated $(0, \sigma^2)$ random variables. Suppose that all roots of the polynomial

$$m^p + \sum_{j=0}^{q} b_j m^{q-j}$$

have magnitude less than 1. Then X_t admits an infinite AR-presentation

$\sum_{j=0}^{\infty} c_j X_{t-j} = e_t$, *where*

$$\begin{aligned}
&c_0 = 1 \\
&c_j + \sum_{i=1}^{j} b_i c_{j-i} = 0,\ j = 1, \ldots q-1 \\
&c_j + \sum_{i=1}^{p} b_i c_{j-i} = 0,\ j = q, q+1, \ldots
\end{aligned}$$

Proposition 3.21. *Let X_t be an AR(p) process satisfying conditions of Theorem 3.20. Then the partial autocorrelation function $\varphi(h) = 0$ for $h > p$*

Proof. By definition, $\varphi(h)$ is the correlation between the X_{t-h} and the residual obtained from the regression of X_t on $X_{t-1}, \ldots X_{t-h+1}$. The latter is e_t and the result follows.

Proposition 3.22. *Let X_t be an AR(p) process satisfying conditions of Theorem 3.20. Then autocovariance function $\gamma(h)$ satisfies:*

$$\begin{gathered}
\gamma(0) + a_1\gamma(1) + \ldots + a_p\gamma(p) = \sigma^2 \\
\gamma(h) + a_1\gamma(h-1) + \ldots + a_p\gamma(h-p) = 0,\ h > 0
\end{gathered}$$

Remark: this result allows one to express $\gamma(0), \ldots \gamma(p)$ through the coefficients $a_1, \ldots a_p$ and vice versa.

Proposition 3.23. *Let X_t be weakly stationary time series:*

1. *The partial autocorrelation coefficient $\varphi(h)$ equals θ_{hh} in the linear regression:*
$$X_{t+h} = \theta_{0h} + \theta_{1h} X_{t+h-1} + \ldots + \theta_{hh} X_t + a_{ht}$$

2. *Let $\rho_{t+h,t\cdot(t+1,\ldots t+h-1)}$ denote the partial correlation of X_{t+h} and X_t when controlled for $X_{t+1}, \ldots X_{t+h-1}$. The squared norm of the residual term in the regression above equals:*
$$\mathbb{E}(a_{ht}^2) = \gamma(0) \prod_{i=1}^{h} (1 - \rho^2_{t+h,t+i-1\cdot(t+i,\ldots t+h-1)})$$

Proof. These statements follow from basic Eucledian geometry. Denote by P the projection onto the subspace spanned by $X_{t+i}, \ldots X_{t+h-1}$. Now consider sequentially orthogonal decompositions $X_{t+h} = P(X_{t+h}) + (1-P)(X_{t+h})$ and look at the component of the second summand along $(1-P)(X_{t+i})$.

Proposition 3.24. *Let $(Y_t)_{t\in\mathbb{Z}}$ be a sequence of elements in $L^2(\Omega, \mathbb{P}, \mathbb{C})$, denote $\boldsymbol{Y}_n \equiv (Y_1, \ldots, Y_n)$. Let $\hat{Y}_{n+s}(Y_1, \ldots Y_n) \equiv \boldsymbol{Y}_n b_{n,s},\ b_{n,s} \in \mathbb{C}^n$ be a linear predictor minimizing mean squared error*

$$\tau^2_{n,s} \stackrel{\text{def}}{=} MSE\left(Y_{n+s}, \hat{Y}_{n+s}(Y_1, \ldots Y_n)\right) \equiv \mathbb{E}\left\{\|Y_{n+s} - \hat{Y}_{n+s}\|^2\right\}$$

Then $b_{n,s} = (\mathbb{V}_{n,n})^+ V_{n,s}$ is a solution, where

$$V_{n,s} \stackrel{\text{def}}{=} \mathbb{E}\left(\boldsymbol{Y}_n^T Y_{n+s}\right)$$
$$\mathbb{V}_{n,n} \stackrel{\text{def}}{=} \mathbb{E}\left(\boldsymbol{Y}_n^T \boldsymbol{Y}_n\right)$$

Furthermore, MSE is given by:

$$\tau^2_{n,s} = \mathbb{V}(Y_{n+s}) - b^T_{n,s} V_{n,s} = \mathbb{V}(Y_{n+s}) - V^T_{n,s} \mathbb{V}^+_{n,n} V_{n,s}$$

Proof. Follows from standard linear regression theory.

Definition 3.25. A time-series is *nonsingular (regular, nondeterministic)* if the sequence of mean squared errors of one-preriod prediction $\tau^2_{n,1}$ is bounded away from zero. A time series is *singular (deterministic)* if

$$\lim_{n\to\infty} \tau_{n,1} = 0$$

Theorem 3.26. *Let $Y_t, b_{n,s}$ and $\tau_{n,s}$ be as defined in Proposition 3.24 and assume that Y_t is weakly stationary, nondeterministic. Denote the components of $b_{n,s}$ by $b_{n,s,i},\ i = 1, \ldots n$ so that*

$$\hat{Y}_{n+s}(Y_1, \ldots Y_n) = \sum_{i=1}^{n} b_{n,s,i} Y_i$$

Then the following recursive relations take place:

1. $$b_{n,s,1} = \tau^{-2}_{n-1,1}\left(\gamma(n+s-1) - \sum_{i=1}^{n} b_{n-1,s,i}\gamma(n+s-1-i)\right)$$

2. $\tau^2_{n,s} = \tau^2_{n-1,s} - b_{n,s,1}\tau^2_{n-1,1}$

3. $\begin{pmatrix} b_{n,s,2} \\ b_{n,s,3} \\ \vdots \\ b_{n,s,n} \end{pmatrix} = \begin{pmatrix} b_{n-1,s,2} \\ b_{n-1,s,3} \\ \vdots \\ b_{n-1,s,n} \end{pmatrix} - b_{n,s,1} \begin{pmatrix} b_{n-1,1,n-1} \\ b_{n-1,1,n-2} \\ \vdots \\ b_{n-1,1,1} \end{pmatrix}$

Remark. Note that one-step prediction terms, $\tau_{n-1,1}$ and $b_{n-1,1}$, appear in the recursion, and the components of the last vector are reversed.

Theorem 3.27. (Gram-Schmidt) *Let Let $(Y_t)_{t\in\mathbb{Z}}$ from $L^2(\Omega, \mathbb{P}, \mathbb{C})$ be a zero-mean, stationary, nondeterministic time series.*

Then one can write $Y_t = \sum_{i=1}^{t} c_{t,i} Z_i$ where

$$\begin{aligned}
\mathbb{E} Z_{t,i} &= 0 \\
\mathbb{E}\left(Z_{t,i} Z_{t,j}\right) &= \delta_{ij} \kappa_i^2 \\
c_{t,1} &= \kappa_1^{-2} \gamma_Y(t-1) \\
c_{t,i} &= \kappa_i^{-2} \left(\gamma_Y(t-i) - \sum_{j<i} c_{t,j} c_{i,j} \kappa_i^2 \right) \\
\kappa_t^2 &= \gamma_Y(0) - \sum_{i=1}^{t-1} c_{t,i}^2 \kappa_i^2
\end{aligned}$$

Proof. This is a result of direct application of Gram-Schmidt orthogonalization algorythm to the sequence $Y_1, \ldots, Y_t$. Note that the process can be generalized to vector-valued processes.

Theorem 3.28. *The real valued function $\rho(h)$ is the correlation function of a real valued stationary time series $X_t(\omega)$ with index set $t \in \mathbb{Z}$ if and only if it is representable in the form*

$$\rho(h) = \int_{-\pi}^{\pi} e^{ihx} dG(x),$$

where $G(x)$ is a symmetric distribution function

Theorem 3.29. *Let the correlation funciton $\rho(h)$ of a stationary time series be abcolutely summable. Then there exists a continuous function $f(\omega)$ such that.*

1. $\rho(h) = \int_{-\pi}^{\pi} f(\omega) \cos \omega h d\omega$

2. $\int_{-\pi}^{\pi} f(\omega)d\omega = 1$

3. $f(\omega) \geq 0$

4. $f(\omega)$ *is an even function*

3.4 Estimator convergence via information matrix

Definition 3.30. Let X be a random variable defined on probability space $(\Omega, \mathbb{P}_\theta)$, $\theta \in \Theta$. Suppose that the likelihood function $\theta \mapsto \ell_\theta \stackrel{\text{def}}{=} \log p_\theta$ is differentiable for all $x \in \Omega$. The gradient

$$\dot{\ell}_\theta(x) = \frac{\partial}{\partial\theta}\ell_\theta(x)$$

is called the *score function.* The *Fisher information* is defined as the matrix

$$i_\theta = \mathbb{V}_\theta\left(\dot{\ell}_\theta(X)\right)$$

Theorem 3.31. *Suppose that Θ is compact and convex and that θ is identifiable, and let $\hat{\theta}_n$ be the maximum likelihood estimator based on a sample of size n from the distribution with (marginal) probability density p_θ. Suppose, furthermore, that the map $\vartheta \mapsto \log p_\vartheta(x)$ is continuously differentiable for all x, with derivative $\dot{\ell}_\vartheta(x)$, such that $\|\dot{\ell}_\vartheta(x)\| \leq L(x)$ for every $\vartheta \in \Theta$, where $L(x)$ is a function with $\mathbb{E}_\theta L^2(X) < \infty$. If θ is an interior point of Θ and the function $\theta \mapsto i_\theta$ is continuous and positive, then under θ, $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to a normal distribution with expectation 0 and variance i_θ^{-1}. Therefore, under θ, as $n \to \infty$, we have*

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, i_\theta^{-1})$$

Theorem 3.32. (Cramer-Rao) *Suppose $\theta \mapsto p_\theta(x)$ is differentiable for every x. Then under certain regularity conditions any unbiased estimator T for $g(\theta)$ satisfies:*

$$\mathbb{V}_\theta(T) \geq g'(\theta)I_\theta^{-1}g'(\theta)^T,$$

where I_θ denotes the full *information matrix.*

3.5 Sufficient Statistics and UMVU estimators

Definition 3.33. For a statistical model $(\Omega, \mathbb{P}_\theta),\ \theta \in \Theta$, a statistic $V(x)$ is called *sufficient* (for X) if conditional distribution $f(x|V = v)$ is independent of V.

Theorem 3.34. *A statistic $V(x)$ is sufficient if there exist functions $h(x)$ and $g(v, \theta)$ such that*

$$p_\theta(x) = h(x)g(V(x), \theta)$$

Theorem 3.35. (Rao-Blackwell) *Let $V = V(X)$ be a sufficient statistic, and let $T = T(X)$ be an arbitrary real-valued estimator for $g(\theta)$. Then there exists an estimator $T^* = T^*(V)$ for $g(\theta)$ that depends only on V, such that $\mathbb{E}_\theta T^* = \mathbb{E}_\theta T$ and $\mathbb{V}_\theta T^* \leq \mathbb{V}_\theta T$ for all θ. In particular, we have $MSE(\theta; T^*) \leq MSE(\theta; T)$. This inequality is strict unless $\mathbb{P}_\theta(T^* = T) = 1$.*

Definition 3.36. For a statistical model $(\Omega, \mathbb{P}_\theta),\ \theta \in \Theta$, a statistic $V(x)$ is called complete if $\mathbb{E}_\theta(f(V)) = 0,\ \forall\theta \in \Theta$ implies $f(V) = 0$ a.s.

Theorem 3.37. *Let $V(x)$ be sufficient and complete, and $T(V)$ be an unbiased estimator for $g(\theta)$. Then $T(V)$ is UMVU estimator (i.e. has smallest variance among all unbiased estimators $\forall\theta \in \Theta$).*

Theorem 3.38. *Suppose that for a k-dimensional exponential family (2.1) the set below contains an interior point:*

$$(Q_1(\theta), \dots Q_k(\theta)),\ \theta \in \Theta$$

Then the random vector $(V_1(x), \dots V_n(x))$ is sufficient and complete.

4 Estimators

4.1 Maximum of n uniformally distributed statistics

Set up: $X_1, X_2, \dots X_n$ i.i.d. drown from $U[0, \theta]$, where θ is the parameter of interest.

- $\hat{\theta} = 2\bar{X_n}$
  - method of moments estimator
  - *unbiased*

- $\mathrm{MSE}(\theta, \hat{\theta}) = \frac{\theta^2}{3n}$, see (1.6)

- $X_{(n)}$ — n-th order statistic, i.e. maximum.
  - $\mathbb{E}_\theta X_{(n)} = \frac{n}{n+1}\theta$, see (1.6)
  - $\mathrm{MSE}(\theta, X_{(n)}) = \frac{2\theta^2}{(n+2)(n+1)}$

- $\frac{n+2}{n+1}X_{(n)}$
  - best estimator of the form $cX_{(n)}$
  - $\mathrm{MSE}(\theta, \frac{n+2}{n+1}X_{(n)}) = \frac{\theta^2}{(n+1)^2}$

4.2 Univariate normal distribution

- $(\hat{\mu}, \hat{\sigma}^2) = \left(\bar{X}_n, \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)\right) = \left(\bar{X}_n, \frac{n-1}{n}S_X^2\right)$
  - maximum likelihood estimator
  - method of moments estimator
  - $\hat{\mu}$ is *unbiased*
  - $\mathbb{E}_{(\mu,\sigma^2)}\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2$

4.3 Empirical distribution function

Let $X_1, \ldots X_n$ be an i.i.d. sample drawn from the distribution F.

- The empirical distribution function (ecdf) $\hat{F}(x) = \sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$ (see 1.1)
  - *unbiased*
  - $\mathbf{cov}_F\left(\hat{F}(u), \hat{F}(v)\right) = n^{-1}(F(\min(u,v)) - F(u)F(v))$ – positively correlated

4.4 Linear Regression

Theorem 4.1. (Ordinary Least Squares)
(i) In one-factor setting, maximum likelihood estimators for slope, intercept and variance are given by (see (1.1, 1.2)):

$$\hat{\beta} = \frac{S_Y r_{X,Y}}{S_X}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} = (Y_i - \hat{\beta}X_i - \hat{\alpha})^2$$

(ii) If the design matrix X in a multiple linear regression has full rank, then the maximum likelihood estimators are given by:

$$\hat{\beta} = (X^T X)^{-1}(X^T Y), \quad \hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n}$$

Theorem 4.2. (Weighted Least Squares/Heteroscedacity)
(i) Assume that error terms ε_i are have variance as $\sigma_i^2 \equiv z_i\sigma^2$ for known constants z_i. Let $w_i \stackrel{\text{def}}{=} (z_i\sigma^2)^{-1}$, then maximum likelihood estimators for slope, intercept and variance are given by (see (1.1, 1.2)):

$$\begin{aligned} \tilde{\beta} = \frac{\sum w_i(x - \tilde{x})(y - \tilde{y})}{\sum w_i(x - \tilde{x})^2} &= \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2} \\ \tilde{\alpha} &= \tilde{y} - \tilde{\beta}\tilde{x} \\ \hat{\sigma}^2 &= n^{-1} \sum \frac{1}{z_i}(y_i - \tilde{\beta}x_i - \tilde{\alpha})^2 \end{aligned}$$

(ii) For the multi-factor model, maximum likelihood estimators can be written in the form:

$$\tilde{\beta} = (X^T W X)^{-1}(X^T W Y)$$

Theorem 4.3. *Let $V \stackrel{\text{def}}{=} \operatorname{span}(X)$, and $V_0 \subset V$. Denote the projection onto V by P_V.*

1. *The likelihood ratio statistic for $H_0 : X\beta_0 \in V_0$ equals*

$$2\log\lambda_n(X, Y) = n \log \frac{\|(E - P_{V_0})Y\|^2}{\|(E - P_V)Y\|^2},$$

2. *Under the null hypothesis, the following quantity has $F_{n-p,p-p_0}$ distribution:*

$$\frac{\|(P_V - P_{V_0})Y\|^2/(p - p_0)}{\|(E - P_V)Y\|^2/(n - p)}$$

5 Statistical tests

5.1 t-tests

5.1.1 One-sample t-test

Let $X_1, X_2, \dots X_n$ be an i.i.d. sample from the $N(\mu, \sigma^2)$-distribution with μ and σ^2 unknown. Given $\mu_0 \in \mathbb{R}$ we test:

$$H_0 : \mu \leq \mu_0 \text{ against } H_1 : \mu > \mu_0 \tag{5.1}$$

Test statistic:

$$T = \sqrt{n}\frac{\bar{X}_n - \mu_0}{S_X} \tag{5.2}$$

By Theorem 3.1, under $\mu = \mu_0$ the statistic has Student's t_{n-1} distribution, consequently we can use.

$$\sup_{\mu \leq \mu_0} \mathbb{P}\left(T \geq t_{n-1,1-\alpha}\right) \leq \alpha \tag{5.3}$$

5.1.2 t-Test for paired observations

Let $(X_1, Y_1), (X_2, Y_2), \dots (X_n, Y_n)$ be the i.i.d. sample of paired observations. We assume that $Z_i \overset{\text{def}}{=} X_i - Y_i$ is $N(\Delta, \sigma^2)$ is normally distributed, and the ordinary One-sample t-test can be used to test the null hypotheses $H_0 : \Delta \geq 0$. Note that if X_i and Y_i are strongly correlated then variance of Z_i decreases and this improves the power of the t-test.

5.1.3 Two-sample t-test

Let $X_1, X_2, \dots X_n$ and $Y_1, Y_2, \dots Y_m$ be two mutually independent i.i.d samples from $N(\mu, \sigma^2)$ and $N(\nu, \sigma^2)$. The test checks

$$H_0 : \mu - \nu \leq 0 \text{ against } H_1 : \mu - \nu > 0 \tag{5.4}$$

Test statistic:

$$T = \frac{\bar{X} - \bar{Y}}{S_{X,Y}\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{5.5}$$

$$S_{X,Y}^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2\right) \tag{5.6}$$

Theorem 3.1 implies that $S_{X,Y}^2$ follows $\sigma^2 \cdot \chi^2_{m+n-2}$ distribution.

5.2 Kolmogorov-Smirnov test

Given and i.i.d. sample $X_1, \dots X_n$ from some unknown distribution F, we want to test:

$$H_0 : F = F_0 \text{ against } H_1 : F \neq F_0 \tag{5.7}$$

The test statistic is given by

$$T = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|, \tag{5.8}$$

where $\hat{F}_n(x)$ stands for the empirical distribution function (see 1.1). The distribution of T is the same for every continuous cdf F_0. The following limit establishes the test:

$$\lim_{n \to \infty} \mathbb{P}_{F_0} \left(T > \frac{z}{n} \right) = 2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-j^2 z^2} \tag{5.9}$$