Introduction to Time Series Analysis: Lecture 1

Lecture by: Prof. Illia Horenko
Notes by: Lars Putzig

version of June 20, 2008

Definition 1.1:

1. *Let $[0,T]$ be an observation intervall, then we call*

$$x(t) : [0,T] \to \Psi \subset \mathbb{R}^n \tag{1.1}$$

a **time series**.

2. *Let $\mathcal{F}([0,T],\Psi)$ be the space of all functions from $[0,T]$ to Ψ, then we call*

$$g(x,\Theta) : \mathcal{F}([0,T],\Psi) \times \Omega \to [0,\infty) \tag{1.2}$$

the **model functional**. *$\Theta \in \Omega, \left(\Omega \subset \mathbb{R}^d\right)$ are the* **modell parameters**.

3. *We denote by "problem of* **time series analysis***" the solution of:*

$$\Theta^* = \arg\min_{\Theta\in\Omega} g(x,\Theta) \tag{1.3}$$

for given $x(t), t \in [0,T]$ and formaly given $g(x,\Theta)$.

These general definitions will give us the possibility to solve a lot of problems

Example 1.1: (analysis of a time series using a gaussian model) *Let $t \in \{0,1,\ldots,T\}$ and $x(t) = x_t$ be normaly distributed with mean $\bar{\mu}$ and variance $\bar{\sigma}^2$. Moreover, let $x_0, x_1, \ldots, x_T$ be independend. Our model parameters are choosen to be $\Theta = (\mu, \sigma^2)$. We look for the best parameters $\bar{\mu}$ and $\bar{\sigma}^2$ for the given dataset. Since x_t are normaly distributed, the density of the x_t is*

$$f_t = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_t-\mu)^2}{2\sigma^2}\right). \tag{1.4}$$

Given the independence of the x_t we can calculate the joint distribution of $(x_t), t \in \{0,\ldots,T\}$:

$$f_x = \prod_{t=0}^{T} f_t = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{T+1} \exp\left[-\frac{\sum_{t=0}^{T}(x_t-\bar{x})^2 + (T+1)(\bar{x}-\mu)^2}{2\sigma^2}\right], \tag{1.5}$$

where $\bar{x} = E[x_t] = \frac{1}{T+1}\sum_{t=0}^{T} x_t$. If we define the Gaussian **model functional** *as*

$$g := -\ln f_x = \frac{T+1}{2}\ln(2\pi\sigma^2) + \frac{\sum_{t=0}^{T}(x_t-\bar{x})^2 + (T+1)(\bar{x}-\mu)^2}{2\sigma^2} \tag{1.6}$$

we can solve the problem of the time series analysis analyticaly by solving the system of equations:

$$\left(\frac{\partial g}{\partial \mu}, \frac{\partial g}{\partial \sigma^2}\right) = 0 \tag{1.7}$$

thus

$$\frac{\partial g}{\partial \mu} = -\frac{2(T+1)(\bar{x}-\mu)}{2\sigma^2} = 0, \tag{1.8}$$

$$\frac{\partial g}{\partial \sigma^2} = \frac{T+1}{2\sigma^2} - \frac{\sum_{t=0}^{T}(x_t-\bar{x})^2 + (T+1)(\bar{x}-\mu)^2}{2\sigma^4} = 0. \tag{1.9}$$

The solution is:

$$\mu = \bar{x}, \tag{1.10}$$

$$\sigma^2 = \frac{1}{T+1}\sum_{t=0}^{T}(x_t-\bar{x})^2. \tag{1.11}$$

These are the known formulars for mean and variance from statistics. ◇

Example 1.2: (K-Means clustering) *Given a set of points $x(t)$ for $t \in \mathcal{T} \subset [0,T]$, we look for K clusters $C_1,\dots,C_K$. Therefore let $\gamma_i(t)$ for $i=1,\dots,K$ and $t \in \mathcal{T}$ be the probabilities for point $x(t)$ to belong to cluster C_i. Moreover let*

$$\sum_{i=1}^{K}\gamma_i(t) = 1, \forall t \in \mathcal{T} \tag{1.12}$$

$$\gamma_i(t) \geq 0, \forall t \in \mathcal{T}, i = 1,\dots,K. \tag{1.13}$$

Now let C_i be the central point of the ith cluster and a point should belong to this cluster, if the function of distance $d(x(t),C_i) \leq d(x(t),C_j)$ for all $j=1,\dots,K$. Then we can choose the model parameters according to

$$\Theta(t) = (C_1,\dots,C_K,\gamma_1(t),\dots,\gamma_K(t)). \tag{1.14}$$

The model function should be

$$g = \sum_{i=1}^{K}\sum_{t\in\mathcal{T}}\gamma_i(t)||x(t)-C_i||_2. \tag{1.15}$$

For given $C_1,\dots,C_K$ the $\gamma_i^*(t)$ can be calculated easily: Choose $\gamma_i(t)=1$ if $x(t)$ belongs to cluster C_i and else $\gamma_i(t)=0$. On the other hand, for given γ_i, the C_i^* can be choosen optimaly by setting C_i^* to be the mean of all points belonging to the ith cluster. This leads to the following algorithm:

1. *Choose the initial $C_1^0,\dots,C_K^0$ randomly.*
2. *Repeat for $j=1,2,\dots$ unitl some stopping criteria is satisfied:*
   (a) *Choose $\gamma_i^j(t)=1$ if $x(t)$ belongs to cluster C_i, else $\gamma_i^j(t)=0$.*
   (b) *Choose C_i^j by calculation the mean of all points belonging to the ith cluster.*

One can prove: Every step of this algorithm improves the result (in sense of making g smaller). ◇

Introduction to Time Series Analysis: Lecture 2

Lecture by: Prof. Illia Horenko
Notes by: Lars Putzig

version of June 24, 2008

Example 2.1: (Markov chain) *In this example we will have a closer look at the discrete time series that satisfy certain independency conditions. Let* $x = x(t), t \in \{t_0, t_1, \ldots, t_N\} \subset [0, T]$ *with equidistant timesteps* $t_i = t_0 + i\Delta_t$ *a time series. Moreover, the set of states shall be* $\Psi^N \in \mathbb{R}^N$ *with* $x(t) \in \Psi = \{1, \ldots, K\}$*. We abreviate* $x_k = x(t_k)$*.*

Definition 2.1: *A process* x *is called* **Markovian** *or a* **Markov process**, *if* x *satisfies*

$$\mathbb{P}[x_{k+1} = j | x_k = i_k, \ldots, x_0 = i_0] = \mathbb{P}[x_{k+1} = j | x_k = i_k]. \tag{2.1}$$

If the term above does not depend on the time t_k*, the Markov process is called* **homogeneous**. *For a homogeneous Markov processes is the* **transition matrix** P *with* $P = [p_{ij}], p_{ij} = \mathbb{P}[x_{k+1} = j | x_k = i]$ *time-invariant. The transition matrix satisfies the following conditions:*

1. $0 \leq p_{ij} \leq 1$
2. $\sum_j p_{ij} = 1, \forall i$
3. $P\mathbb{1}^K = \mathbb{1}^K$, *where* $\mathbb{1}^K = (1, 1, \ldots, 1)^T$
4. $|\lambda| \leq 1$ *where* λ *is any eigenvalue of* P*.*
5. *If the process is reversible, thus* $\pi_i p_{ij} = \pi_j p_{ji}$ *for some* π *we get:* $\lambda \in \mathbb{R}$*.*

Now we want to solve the minimization problem:

$$\Theta^* = \arg\min_{\Theta} g(x, \Theta). \tag{2.2}$$

Using the homogeneous Markovian property of x *we can state:*

$$\mathbb{P}[x = (x_0, x_1, \ldots, x_T) | P] = \mathbb{P}[x(0) = x_0] \prod_{t=1}^{T} p_{x_{t-1}, x_t}. \tag{2.3}$$

Remark: *We start by expressing this probabilities depending on the whole series to time* t*. Then we use the Markov property to truncate all but the last dependency and in the end we use the homogeneous property to express the probabilities by the components of* P*.*

Now we denote by N_{ij} *the number of jumps from* i *to* j *during the lifetime of* x*. Then we can write:*

$$\mathbb{P}[x = (x_0, x_1, \ldots, x_T) | P] = \mathbb{P}[x(0) = x_0] \prod_{i=1}^{K} \prod_{j=1}^{K} p_{ij}^{N_{ij}}. \tag{2.4}$$

Using $\tilde{g}(\cdot) = -\ln(\cdot)$ *we get*

$$\tilde{g}(P) = -\ln(\mathbb{P}[x = (x_0, \ldots, x_T) | P]) = -\ln(\mathbb{P}[x(0) = x_0]) - \sum_{i=1}^{K} \sum_{j=1}^{K} N_{ij} \ln p_{ij}. \tag{2.5}$$

We use only the condition $\forall i : \sum_j p_{ij} = 1$, the other ones will be satisfied automaticaly (has to checked afterwards). Using the Lagrange technic we get the extended problem:

$$g(P, \mu) = -\log \mathbb{P}[x(0) = x_0] - \sum_{i=1}^{K} \sum_{j=1}^{K} N_{ij} \ln p_{ij} + \sum_{i=1}^{K} \mu_i \left(\sum_{j=1}^{K} p_{ij} - 1 \right). \tag{2.6}$$

Now we can write the derivatives:

$$\frac{\partial g}{\partial p_{ij}} = -\frac{N_{ij}}{p_{ij}} + \mu_i. \tag{2.7}$$

Setting this equal to zero we get:

$$p^*_{ij} = \frac{N_{ij}}{\mu_i}. \tag{2.8}$$

Now we put this into the condition to get:

$$1 = \sum_j p^*_{ij} = \frac{1}{\mu^*_i} \sum_j N_{ij}, \mu^*_i = \sum_j N_{ij} \, \forall i. \tag{2.9}$$

Using this for p^*_{ij} we get:

$$p^*_{ij} = \frac{N_{ij}}{\sum_l N_{il}}. \tag{2.10}$$

This is a good point to introduce the robustness:

Definition 2.2: Robustness *is a measure of the influence of small perturbations on the optimal solution.*

Since $\tilde{g}$ as a function of Θ is convex, we have a closer look at the second derivative:

$$\frac{\partial^2 g}{\partial p^2_{ij}}(p^*_{ij}) = \frac{N_{ij}}{p^2_{ij}}(p^*_{ij}) = \frac{\left(\sum_l N_{il}\right)^2}{N_{ij}}. \tag{2.11}$$

Thus we can state: The more information we have (thus N_{ij} is large) the more stability we have in the solution. ◇

We will need some convergence for stochastic objects, so we define:

Definition 2.3:

1. *X_n converges* **almost surely (a.s.)** *to X if $\mathbb{P}[X_n \to X] = 1$.*
2. *X_n converges* **in probability** *to X if $\mathbb{P}[|X_n - X| > \varepsilon] \to 0$.*
3. *X_n converges* **in distribution** *to X if $\mathbb{E}[X_n] \to \mathbb{E}[X]$.*

Introduction to Time Series Analysis: Lecture 3

Lecture by: Prof. Illia Horenko
Notes by: Lars Putzig

version of July 1, 2008

We start by bringing some definitions back in mind:

Definition 3.1: *The* **expectation** $\mathbb{E}_\omega[X(\omega)]$ *can be calculated by*

$$\mathbb{E}_\omega[X(\omega)] = \int_\Omega X(\omega) P(d\omega). \tag{3.1}$$

The **variance** *of a random variable is defined by*

$$Var_\omega[X(\omega)] = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]. \tag{3.2}$$

This is a good point, to repeat the Central Limit Theorem:

Theorem 3.1 *Let X_n, $X_n : \Omega \to \mathbb{R}$ be a sequenze of iid random variables with $\mathbb{E}[X_n] = \mu$, $Var[X_n] = \sigma^2$ for all n. Moreover, let*

$$S_N = \frac{\frac{1}{N}\sum_{i=1}^N X_i - \mu}{\frac{1}{\sqrt{N}}\sigma}. \tag{3.3}$$

Then $S_n \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$ in distribution. Moreover, the probability for S_∞ to be in an intervall $[a,b)$ is given by

$$\mathbb{P}[a \le S_\infty < b] = \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(-\frac{x^2}{2}\right) dx. \tag{3.4}$$

The second important theorem is the Law of Large Numbers:

Theorem 3.2 *Let X_n iid with $\mathbb{E}[X_n] = \mu < \infty$ and $Var[X] = \sigma^2 < \infty$. Then*

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{l=1}^N X_l - \mu\right| \ge \varepsilon\right] \le \frac{\sigma^2}{N\varepsilon^2}. \tag{3.5}$$

Now let $\Psi = \{1, 2, 3, \ldots, K\}$ be the **state space** and P the transition matrix for a Markovian process. Moreover $v \in \mathbb{R}^K$ with $0 \le v_i \le 1$ and $\sum_i v_i = 1$ should be a vector of state probabilites, i.e. this is a distribution. Then we have the sequence of distributions starting by some v^0 given by

$$v^{t+1} = P^T v^t. \tag{3.6}$$

Translating the fixed point theorem of Banach to our setting, we get:

Theorem 3.3 *If the eigenvalue $\lambda = 1$ has a geometric size of 1, thus the sub space of the eigenvectors has dimension 1, then $v^t \overset{t\to\infty}{\longrightarrow} \pi$. Where π is the stationary distribution.*

This leads us to the Law of Large Numbers for Markovian processes

Theorem 3.4 *Let π be unique and $f : \Psi \to \mathbb{R}$. Then*

$$\mathbb{E}_\pi[f] = \sum_{j\in\Psi} \pi_j f(j). \tag{3.7}$$

If $\mathbb{E}_\pi[|f|] < \infty$, then

$$\frac{1}{N}\sum_{j=1}^N f(x_j) \overset{N\to\infty}{\longrightarrow} \mathbb{E}_\pi[f] \tag{3.8}$$

almost surely.

We might be interested in the leaving time, this is the average time we stay in a specific state:

$$\tau(i) = \min\{t \geq 0 : X_{t_0} = i, X_{t_0+t} \neq i\} \tag{3.9}$$

The average leaving time would be:

$$\mathbb{E}[\tau(i)] = \Delta t \sum_{k=1}^{\infty} k(1 - p_{ii}) p_{ii}^{k-1} \tag{3.10}$$

where Δt is the step size we use during the time. We can calculate this by using:

$$\mathbb{E}[\tau(i)] = \Delta t \sum_{k=0}^{\infty} p_{ii}^{k} = \frac{\Delta t}{1 - p_{ii}} \tag{3.11}$$

if and only if $p_{ii} < 1$. This is a measure for the metastability of the time series.

Definition 3.2: *For $x \in \Psi, y \in \Psi$ we write $x \mapsto y$ if $P[X_t = y | X_0 = x] > 0$ for some t. Then y is called* **attainable from** *x. We write $x \mapsto y \mapsto x$ as $x \leftrightarrow y$.*

We can use this to define communication classes:

Theorem 3.5 *Let $\mathbb{K}$ be the set of close disjunct communication classes (thus $x, y \in K \in \mathbb{K} \Leftrightarrow x \leftrightarrow y$), then $\lambda = 1$ is $\#\mathbb{K}$ times an eigenvalue.*

Theorem 3.6 *Let C be a close, disjunct communication class and P_C with $p_{C,kl} = p_{kl}$ for $k, l \in C$. Then there exists an irreducible Markovian process with transition matrix P_C.*

Introduction to Time Series Analysis: Lecture 4

Lecture by: Prof. Illia Horenko
Notes by: Lars Putzig

version of July 1, 2008

Definition 4.1: *$E_1, \ldots, E_d$ are called* **cyclic or periodic** *classes, if*

$$\mathbb{P}[x_1 \in E_{k+1} | x_0 \in E_k] = 1 \tag{4.1}$$

and $E_{d+1} = E_1$.

Example 4.1: (deterministic chain) *We have a markov chain with transistion matrix*

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.2}$$

Then this chain has two communication classes $C_1 = \{1, 2\}$ and $C_2 = \{3\}$. C_1 can be split in two periodic classes. ◇

Definition 4.2: *If for all cyclic classes $E_1, \ldots, E_d$ of a Markov chain $d = 1$ then the chain is called aperiodic.*

Theorem 4.1 *The following statements are equivalent:*

- *P has K eigenvalues with $|\lambda| = 1$*
- *$d = K$*

Theorem 4.2 Frobenius-Perron *Let a Markov chain be irreducible and aperiodic, then $\lambda_1 = 1$ is a unique eigenvalue and $|\lambda_i| < 1$ for $i > 1$ and there exists a unique stationary distribution $\pi^T = \pi^T P$.*

If $A, B \subset S$ where S is the state space of a Markov chain, then

$$\mathbb{P}[x_1 \in B | x_0 \in A] = \sum_{l \in A} \mathbb{P}[x_1 \in B | x_0 = l] \mathbb{P}[x_0 = l | x_0 \in A] \tag{4.3}$$

$$= \sum_{\substack{l \in A \\ m \in B}} \underbrace{\mathbb{P}[x_1 = m | x_0 = l]}_{p_{lm}} \mathbb{P}[x_0 = l | x_0 \in A] \tag{4.4}$$

Example 4.2: (stochastic periodic markov chain) *Let P be the transition matrix for a Markov chain with*

$$P = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}. \tag{4.5}$$

$\lambda = 1$ is two times an eigenvalue of this matrix, we have $C_1 = \{2\}$ and $C_2 = \{1, 3\}$, thus by exchanging the states 2 and 3 we get the block structure. For this blocks the assumptions of Frobenius-Perron are satisfied and indeed: $\lambda_1 = 1$ is unique and $\pi^1 = 1$ respectively $\pi^2 = \left(\frac{1}{2}, \frac{1}{2}\right)^T$ are the unique stationary distributions. $\diamond$

Definition 4.3: *Let $(X_n)_n$ be a sequence of outputs of a Markov chain, with $X_n \in \{1, 2\}$ and (μ_1, σ_1) and (μ_2, σ_2) two parameter sets. Now let $O_i \sim \mathcal{N}(\mu_{X_i}, \sigma^2_{X_i})$. This is called a (Gaussian)* **Hidden Markov Modell (HMM)**.

Introduction to Time Series Analysis: Lecture 5

Lecture by: Prof. Illia Horenko
Notes by: Lars Putzig

version of July 7, 2008

This time, we will have a closer look to so called **Gaussian Mixture Models** (GMMs) and **Finite Mixture Models** (FMMs). Let μ_i, θ_i be sequences of mean and variance for some gaussian processes. Let f_i be the density of $\mathcal{N}(\mu_i, \theta_i)$. Now we construct the density function

$$f = \sum_y \alpha_y f_y \tag{5.1}$$

with $\sum_y \alpha_y = 1$ and $0 \leq \alpha_y \leq 1$. Where f_y and α_y depend on the random variable Y. For the Likelyhood operator $\mathbb{L}$ we can define the problem as

$$\tilde{\Theta} = \arg\max_{\Theta} \mathbb{L}[\Theta | x_0, \ldots, x_T] = \arg\max_{\Theta} \prod_{t=0}^{T} \mathbb{P}[X_t = x_t | \Theta] \tag{5.2}$$

Given this we could use the Loglikelihood $\mathcal{L}$, thus:

$$\tilde{\Theta} = \arg\max_{\Theta} \sum_{t=0}^{T} \log \left(\sum_y \mathbb{P}[X_t = x_t | Y_t = y, \Theta] \mathbb{P}[Y_t = y] \right) \tag{5.3}$$

Now we reduce this to the problem with marginal distribution:

$$\hat{\Theta} = \arg\max_{\Theta} \log \mathbb{L}[\Theta | X] = \arg\max_{\Theta} \log \left(\int_{\mathcal{Y}} \mathbb{P}[x_0, \ldots, x_T, y | \Theta] dy \right) \tag{5.4}$$

where $Y : \Omega \to \mathcal{Y}$. Defining a family of lower bounds $B(\Theta)$ with

$$B(\Theta) \leq \log \mathbb{L}[\Theta, x], \; \forall \Theta \tag{5.5}$$

we let Θ^i be the first i elements of this family with

$$\log \mathbb{L}[\Theta | x] = \log \mathbb{P}[x_0, \ldots, x_T | \Theta] = \log \int_{\mathcal{Y}} \mathbb{P}[x, y | \Theta] \, dy \tag{5.6}$$

$$= \log \int_{\mathcal{Y}} \frac{\mathbb{P}[X, Y]}{f(y)} f(y) \, dy \tag{5.7}$$

$$= \log \mathbb{E}_{f(Y)} \left[\frac{\mathbb{P}[X, Y]}{f(Y)} \right]. \tag{5.8}$$

Now we use the Jensen inequation to get

$$\log \mathbb{L}[\Theta | x] \geq \mathbb{E}_{f(Y)} \left[\log \frac{\mathbb{P}[X, Y]}{f(y)} \right] \tag{5.9}$$

Theorem 5.1 Jensen inequation *Let $u : D \subset \mathbb{R} \to \mathbb{R}$ be a convex function then:*

$$\mathbb{E}[u(X)] \geq u(\mathbb{E}[X]) \tag{5.10}$$

This leads us to the formal equation:

$$LB = \left\{ B_f(\Theta) = \mathbb{E}_{f(Y)} \left[\log \frac{\mathbb{P}[x, Y | \Theta]}{f(Y)} \right] \right\} \tag{5.11}$$

where f is a density function with $f : \mathcal{Y} \to \mathbb{R}$. This leads to the Expectation-Maximization (EM) algoritm. This algorithm consists of two steps: Expectation and Maximization. We have a look at the first step. Here we try to solve the following problem:

$$B_{\max} = \arg \max_{B_f \in LB} B_f(\Theta^i) \tag{5.12}$$

With $\int_{\mathcal{Y}} f(y)\, dy = 1$, therefore we define the Lagrangian

$$J(f, \Theta^i) = B_f(\Theta^i) + \lambda \left(1 - \int_{\mathcal{Y}} f(y)\, dy\right). \tag{5.13}$$

with

$$\frac{dJ}{df} = \lim_{\varepsilon \to 0} \frac{J(f + \varepsilon g) - J(f)}{\varepsilon} \tag{5.14}$$

For some testfunction g. We will have a look at the result during the next lecture.

Introduction to Time Series Analysis: Lecture 6

Lecture by: Prof. Illia Horenko
Notes by: Lars Putzig

version of July 7, 2008

Last time we defined a family of lower bounds to use the EM-Algorithm. In step one (the expectation step) we need to solve the following problem:

$$\max_f B_f(\Theta) \text{ with } \int_{\mathcal{Y}} f(y)\,dy = 1 \tag{6.1}$$

Thus we use the Lagrangian J to get

$$J(f,\Theta) = B_f(\Theta) + \lambda\left(\int_{\mathcal{Y}} f(y)\,dy - 1\right) \tag{6.2}$$

$$\frac{\partial J}{\partial f} = 0 \text{ (How do we solve this?)} \tag{6.3}$$

$$\frac{\partial J}{\partial \lambda} = 0 \tag{6.4}$$

Equation (6.3) is not a typical problem, here we have to use a variational differential:

$$\frac{\partial J}{\partial f} = \lim_{\varepsilon\to 0}\frac{J(f+\varepsilon g) - J(f)}{\varepsilon} = 0 \tag{6.5}$$

$$\Rightarrow 0 = \lambda + \log \mathbb{P}(x,y|\Theta^{(i)}) - (1+\log f(Y)) \tag{6.6}$$

To get the following solution

$$f^*(y) = \frac{\mathbb{P}(x,y|\Theta(i))}{\mathbb{P}(x|\Theta^{(i)})} = \mathbb{P}(y|x,\Theta^{(i)}) \tag{6.7}$$

Now we use this within our family to get

$$B_{f^*}(\Theta) = \int_{\mathcal{Y}} \mathbb{P}(y|x,\Theta^{(i)}) \log\frac{\mathbb{P}(x,y|\Theta)}{\mathbb{P}(y|x,\Theta^{(i)})}\,dy \tag{6.8}$$

And thus

$$B_{f^*}(\Theta^{(i)}) = \int_{\mathcal{Y}} \mathbb{P}(y|x,\Theta^{(i)}) \log\frac{\mathbb{P}(x,y|\Theta^{(i)})}{\mathbb{P}(y|x,\Theta^{(i)})}\,dy \tag{6.9}$$

$$= \int_{\mathcal{Y}} \mathbb{P}(y|x,\Theta^{(i)}) \log \mathbb{P}(x|\Theta^{(i)})\,dy \tag{6.10}$$

$$= \int_{\mathcal{Y}} \mathbb{P}(y|x,\Theta^{(i)})\,dy \log \mathbb{P}(x|\Theta^{(i)}) \tag{6.11}$$

$$= \log \mathbb{P}(x|\Theta^{(i)}) = \log \mathbb{L}(\Theta^{(i)}|x) \tag{6.12}$$

Thus, since $B_f \le \log \mathbb{L}$ we have the maximum property. Now we can proceed to the second step: Our new problem is

$$\arg\max_\Theta B_{f^*}(\Theta) = \arg\max_\Theta \int_y \mathbb{P}(y|x,\Theta^{(i)}) \log\frac{\mathbb{P}(x,y|\Theta)}{\mathbb{P}(y|x,\Theta^{(i)})}\,dy \tag{6.13}$$

$$= \arg\max_\Theta \int_y \mathbb{P}(y|x,\Theta^{(i)}) \log \mathbb{P}(x,y|\Theta)\,dy =: \arg\max_\Theta Q^{(i)}(\Theta) = \Theta^{(i+1)} \tag{6.14}$$

Theorem 6.1 *Let $\Theta^{(i)}$ be output of the algorithm above. Then $\mathbb{L}(\Theta^{(i+1)}|x) \ge \mathbb{L}(\Theta^{(i)}|x)$.*

Proof:

$$\log \mathbb{L}(\Theta^{i+1}|x) \geq B_f(\Theta^{(i+1)}) = \max_{\Theta} B_f(\Theta) \geq B_f(\Theta^{(i)}) = \log \mathbb{L}(\Theta^{(i)}|x) \tag{6.15}$$

□

Nevertheless, solving these problems is no trivial task. Let's have a closer look to the solutions: If the timesteps are statistical independend we get

$$f^* = \mathbb{P}(y|x, \Theta^{(i)}) = \prod_{t=1}^{T} \mathbb{P}(y_t|x, \Theta^{(i)}) = \prod_{t=1}^{T} \gamma_t(y_t) \tag{6.16}$$

with

$$\mathbb{P}(y|x_t, \Theta^{(i)}) = \frac{P(x_t, y|\Theta^{(i)})}{\sum_{y=1}^{K} P(x_t, y|\Theta^{(i)})} = \gamma_t(y) \tag{6.17}$$

then

$$f^* = \gamma_t(y). \tag{6.18}$$

Now we might use this for $Q^{(i)}$:

$$Q^{(i)}(\Theta) = \sum_{y=1}^{K} \gamma_t(y) \log \mathbb{P}(x_t, Y_t = y|\Theta) \tag{6.19}$$

Introduction to Time Series Analysis: Lecture 7

Lecture by: Prof. Illia Horenko
Notes by: Lars Putzig

version of July 7, 2008

While the theory of the last lectures was for general Mixture Models, we will now have a special look at Gaussian Mixture Models. Let $G(\cdot, \mu_y, \Sigma_y)$ be the density function of $\mathcal{N}(\mu_y, \Sigma_y)$, thus

$$G(x, \mu_y, \Sigma_y) = \frac{1}{z} \exp\left(-(x-\mu_y)^T \Sigma_y^{-1} (x-\mu_y)\right) \tag{7.1}$$

where z is used to normalize the measure. Additionally let

$$\alpha_y = P(Y = y|\Theta) \tag{7.2}$$

with $\Theta = (\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \alpha_1, \ldots, \alpha_K)$. Then we write a Lagrangian:

$$J(\Theta) = Q^{(i)}(\Theta) + \lambda \left(\sum_{y=1}^{K} \alpha_y - 1 \right) \tag{7.3}$$

and differentiate with respect to Θ and set those to zero.

$$\partial_{\Sigma_y^{-1}} J(\Theta) = \sum_{t=0}^{T} \gamma_t(y)(-(x_t - \mu_y)(x_t - \mu_y)^T + \partial_{\Sigma_y^{-1}} \log(z(\Sigma_y))) = 0 \tag{7.4}$$

$$\partial_{\mu_y} J(\Theta) = -2 \sum_{t=0}^{T} \gamma_t(y)(-(x_t - \mu_y)) = 0 \tag{7.5}$$

$$\partial_{\lambda} J(\Theta) = \sum_{y=1}^{K} \alpha_y - 1 = 0 \tag{7.6}$$

$$\partial_{\alpha_y} J(\Theta) = \sum_{t=0}^{T} \frac{\gamma_t(y)}{\alpha_y} - \lambda = 0 \tag{7.7}$$

Now we start to solve these equations

$$\mu_y^* = \frac{\sum_t \gamma_t x_t}{\sum_t \gamma_t} \tag{7.8}$$

$$\Sigma_y^* = \frac{1}{\sum_t \gamma_t} \sum_t \gamma_t (x_t - \mu_y^*)(x_t - \mu_y^*)^T \tag{7.9}$$

$$\alpha_y^* = \frac{1}{\sum_y \gamma(y)} \sum_t \gamma_t, \gamma_y = \sum_{t=0}^{T} \gamma_t(y). \tag{7.10}$$

At this point, the lecture was continued by a computational and graphical example, this will not be included in this file.

Introduction to Time Series Analysis: Lecture 8

Lecture by: Prof. Illia Horenko
Notes by: Lars Putzig

version of July 7, 2008

During this lecture, we will have a look at high-dimensional problems. Therefore we assume there are some correlations between the different dimensions. Now we want to keep as much information as possible while reducing the dimensions. Thus let T be a matrix where the columns define a basis of some submanifold. Moreover, the colums should be orthonormal. Then we have a look at

$$\sum_t \|(x_t - \mu) - TT^T(x_t - \mu)\|_2^2 \to \min_T, T^T T = \mathbb{1}^{m\times m} \tag{8.1}$$

for some μ. To solve this, we use once more the Lagrangian principle:

$$L := \sum_t \|(x_t - \mu) - TT^T(x_t - \mu)\|_2^2 + \sum_{i,j=1}^{m} \lambda_{ij}(T^T T - \mathbb{1}^{m\times m}) \tag{8.2}$$

$$= \sum_t \|(x_t - \mu) - TT^T(x_t - \mu)\|_2^2 + e^T\lambda \odot (T^T T - \mathbb{1})e \tag{8.3}$$

Where e is a vector full of ones. Let's have a look at the first part of the L:

$$\|(x_t - \mu) - TT^T(x_t - \mu)\|_2^2 \tag{8.4}$$

$$=((x_t - \mu) - TT^T(x_t - \mu))^T((x_t - \mu) - TT^T(x_t - \mu)) \tag{8.5}$$

$$=((x_t - \mu)^T(x_t - \mu) - (x_t - \mu)^T TT^T(x_t - \mu) - (x_t - \mu)^T TT^T(x_t - \mu) \tag{8.6}$$

$$+ (x_t - \mu)^T TT^T TT^T(x_t - \mu)) \tag{8.7}$$

$$=((x_t - \mu)^T(x_t - \mu) - (x_t - \mu)^T TT^T(x_t - \mu) \tag{8.8}$$

$$=(x_t - \mu)^T \underbrace{(\mathbb{1} - TT^T)}_{=:Q}(x_t - \mu) \tag{8.9}$$

Thus

$$\frac{\partial L}{\partial \mu} = -2Q\sum_t (x_t - \mu) \tag{8.10}$$

By setting this to zero, we get

$$\sum_t (x_t - \mu) \in \ \mathrm{Kern}(Q) \ni \xi, \bar{x} = \frac{1}{N}\sum_t x_t, \mu = \bar{x} + \xi. \tag{8.11}$$

We need one more derivative:

$$\frac{\partial L}{\partial T} = -2\sum_t (x_t - \mu)(x_t - \mu)^T T - 2T\lambda \tag{8.12}$$

Now we set this to zero, too:

$$\sum_t (x_t - \mu)(x_t - \mu)^T T = -T\lambda \tag{8.13}$$

For $m = 1$ this is an eigenvalue problem. Thus we can solve this problem step by step by solving m eigenvalue problems. Therefore, this approach is called "principle components

analysis". Let us have a closer look at this stuff:

$$\sum_t ((x_t - \mu) - TT^T(x_t - \mu))^T((x_t - \mu) - TT^T(x_t - \mu)) \tag{8.14}$$

$$= \sum_t (x_t - \mu)^T(x_t - \mu) - \sum_t \text{trace}(T^T(x_t - \mu)(x_t - \mu)^T T) \tag{8.15}$$

$$= \sum_t (x_t - \mu)^T(x_t - \mu) - \text{trace}(T^T \underbrace{\sum_t (x_t - \mu)(x_t - \mu)^T}_{=\text{Cov}} T) \tag{8.16}$$

$$= \sum_t (x_t - \mu)^T(x_t - \mu) - \text{trace}(T^T \Lambda T) \tag{8.17}$$

$$= \sum_t (x_t - \mu)^T(x_t - \mu) - \text{trace}(\Lambda) \tag{8.18}$$

$$= \sum_{i=m+1}^{n} \lambda_i \tag{8.19}$$

Introduction to Time Series Analysis: Lecture 9

Lecture by: Prof. Illia Horenko
deputized by: Eike Meerbach
Notes by: Lars Putzig

version of July 8, 2008

This time we have a look at autoregressive processes. Therefore let $Z = \{z_1, \dots, z_t\}$ with $z_i \in \mathbb{R}^d$ be a time series. Normally, we are looking for some $f : \mathbb{R}^d \to \mathbb{R}^d$ such that $\hat{Z}_{k+1} = f(Z_k)$ and $||\hat{Z}_{k+1} - Z_{k+1}||$ should be as small as possible. Now we add additional information to this function, thus $\hat{Z}_{k+1} = f(Z_k, Z_{k-1}, \dots, Z_{k-p})$. We will use some general assumptions throughout this lecture:

1. f should be affin multilinear, thus $\hat{Z}_t = \nu + A_1 z_{t-1} + A_2 z_{t-2} + \cdots + A_p z_{t-p}$ with $\nu \in \mathbb{R}^d$ and $A_i \in \mathbb{R}^{d \times d}$.

2. The error $\varepsilon_t = z_t - \hat{Z}_t \in \mathbb{R}^d$ should be white noise, thus ε_t is a stochastic process with $E[\varepsilon_t] = 0 \, \forall t$ and $E[\varepsilon_t \varepsilon_t^T] = R \, \forall t$ and $E[\varepsilon_t \varepsilon_s^T] = 0 \, \forall, s \neq t$.

Definition 9.1: *A stochastic process $Z = \{z_t\}$ is called a* **VAR(p)-process** *(Vector AutoRegressive), if a process $\varepsilon = \{\varepsilon_t\}$ exists such that*

$$z_{t+1} = \nu + A_1 z_t + A_2 z_{t-1} + \cdots + A_p z_{t-p+1} + \varepsilon_{t+1} \tag{9.1}$$

with $\varepsilon_t, \nu \in \mathbb{R}^d$ and $A_i \in \mathbb{R}^{d \times d}$.

There are two different ways to initialize VAR(p)-processes:

1. We assume Z to be stationary (e.g. we started the process at $t = -\infty$).

2. We initialize Z with arbitrary values $(z_1, \dots, z_p)$.

For now, let $p = 1$, thus

$$Z_t = \nu + A_1 z_{t-1} + \varepsilon_t \tag{9.2}$$

For some given z_0 we get

$$z_1 = \nu + A_1 z_0 + \varepsilon_1 \tag{9.3}$$

$$z_2 = \nu + A_1(\nu + A_1 + z_0 + \varepsilon_1) + \varepsilon_2 = (I + A_1)\nu + A_1^2 z_0 + A_1 \varepsilon_1 + \varepsilon_2 \tag{9.4}$$

$$z_t = (I + A_1 + \cdots + A_1^{t-1})\nu + A_1^t z_0 + \sum_{k=0}^{t-1} A_1^k \varepsilon_{t-k} \tag{9.5}$$

$$z_\infty = \lim_{t \to \infty} z_t = (I - A_1)^{-1}\nu + \sum_{k=0}^{\infty} A_1^k \varepsilon_{t-k} \tag{9.6}$$

if and only if $\sigma(A_1) < 1$, where $\sigma(A_1)$ is the spectral radius of A_1.

Definition 9.2: *A VAR(1)-process is called* **stable**, *if* $\sigma(A_1) < 1$. *This is equivalent to* $\det(I - sA) \neq 0$ *for* $s \leq 1$.

Now we expand this theory to VAR(p)-processes by splitting these into VAR(1)-processes:

$$\bar{Z}_t = [z_t, z_{t-1}, \ldots, z_{t-p+1}]^T, \bar{\nu} = [\nu, 0, \ldots, 0]^T \tag{9.7}$$

$$\bar{A} = \begin{pmatrix} A_1 & \ldots & A_p & \\ I & & & 0 \\ & \ddots & & \vdots \\ 0 & & I & 0 \\ & & & 0 \end{pmatrix}, \bar{\varepsilon}_t = [\varepsilon_t, 0, \ldots,]^T \tag{9.8}$$

$$\bar{Z}_t = \bar{\nu} + \bar{A}\bar{z}_{t-1} + \bar{\varepsilon}_t \tag{9.9}$$

Thus a VAR(p)-process could be called stable, if $\sigma(\bar{A}) < 1$ or $\det(I - A_1 s - A_2 s^2 - \cdots - A_p s^p) \neq 0$ for all $s \leq 1$. A stable VAR(p)-process started by $t = -\infty$ will be stationary. How about prediction?

$$\bar{z}_{t+1} = \mathbb{E}[z_{t+1}|z_t, \ldots, z_{t-p+1}] = \nu + A_1 z_t + \cdots + A_p z_{t-p+1} \tag{9.10}$$

Multi-step prediction works the same way, we simply use the predicted values to calculate the missing ones recursively. The error is ε_{t+1}. For multi-step predictions the errors sum up, thus:

$$z_{t+j} - \bar{z}_{t+j} = \sum_{i=1}^{j} A_i^{j-1} \varepsilon_{t+i} \tag{9.11}$$

The expectation of the error is always zero, thus the guess is unbiased. Moreover, this is a minimal MSE-estimator. As a last point, we try to estimate the parameters of the VAR(p) model. Let $Z = \{z_1, \ldots, z_T\}$ where the first p elements are used as initial values. We define

$$\Phi := (\nu, A_1, A_2, \ldots, A_p) \in \mathbb{R}^{d \times (dp+1)} \tag{9.12}$$

$$Y := (z_{p+1}, z_{p+2}, \ldots, z_T) \in \mathbb{R}^{d \times (T-p)} \tag{9.13}$$

$$X_t := \begin{pmatrix} 1 \\ z_t \\ \vdots \\ z_{t-p+1} \end{pmatrix} \in \mathbb{R}^{dp+1} \tag{9.14}$$

$$X := (X_p, X_{p+1}, \ldots, X_T) \in \mathbb{R}^{(dp+1) \times (T-p+1)} \tag{9.15}$$

Thus $z_t = \Phi X_{t-1} + \varepsilon_t$ for $t > p$. Then we get

$$z_t X_{t-1}^T = \Phi X_{t-1} X_{t-1}^T + \varepsilon_t X_{t-1} \tag{9.16}$$

$$\mathbb{E}[z_t X_{t-1}^T] = \mathbb{E}[\Phi X_{t-1} X_{t-1}^T] \tag{9.17}$$

The lefthand side of this equations could be estimated by

$$\mathbb{E}[z_t X_{t-1}^T] \approx \frac{1}{T-p} \sum_{i=p+1}^{T} z_i X_{i-1}^T = \frac{1}{T-p} Y X^T, \tag{9.18}$$

while the righthand side is approximatly

$$\mathbb{E}[\Phi X_{t-1} X_{t-1}^T] \approx \frac{1}{T-p} \sum_{i=p+1}^{T} X_{i-1} X_{i-1}^T = \frac{1}{T-p} X X^T. \tag{9.19}$$

Putting these estimators together, we get:

$$Y X^T = \Phi X X^T \Rightarrow \hat{\Phi} = Y X^T (X X^T)^{-1} \tag{9.20}$$

Introduction to Time Series Analysis: Lecture 10

Lecture by: Prof. Illia Horenko
Notes by: Lars Putzig

July 8, 2008

Let $Z_t \in \mathbb{R}^n$ with

$$Z_{t+\tau} = \sum_{i=0}^{p} A_i(\tau) Z_{t-i\tau} + B(\tau)\varepsilon_{t+\tau} \tag{10.1}$$

with $A_i(\tau), B(\tau) \in \mathbb{R}^{n\times n}$, $\mathbb{E}[\varepsilon_t] = 0$ and $\mathbb{E}[\varepsilon_{t_1}^T \varepsilon_{t_2}] = \delta(t_1 - t_2)$. Then we get

$$\mathbb{E}[Z_{t+\tau}] = \sum_{i=0}^{p} A_i(\tau)\mathbb{E}[Z_{t-i\tau}] \tag{10.2}$$

Now we multiply the $Z_{t+\tau}$ by $Z_{t-j\tau}^T$ for $j = 0, 1, \ldots$ to see

$$\mathbb{E}[Z_{t+\tau} Z_{t-j\tau}^T] = \sum_{i=0}^{p} A_i(\tau)\mathbb{E}[Z_{t-i\tau} Z_{t-j\tau}^T]. \tag{10.3}$$

Definition 10.1: *We denote by $\rho(t-\tau)$ the* **auto-covariance**. *We define*

$$\rho(t-\tau) = \mathbb{E}[(Z_t - \mathbb{E}[Z_t])(Z_\tau - \mathbb{E}[Z_\tau])^T]. \tag{10.4}$$

A special property of the auto-covariance is $\rho(t-\tau) = \rho(\tau - t)$.

This allows us to write the equation (10.3) as

$$\rho((1-j)\tau) = \sum_{i=0}^{p} A_i(\tau)\rho((j-i)\tau). \tag{10.5}$$

This is the *Jule-Walker-equation.* Now let Z_t be an auto-regressive process of order p. Therefore let

$$\alpha = [A_0, A_1, \ldots, A_p] \in \mathbb{R}^{n\times(p+1)n} \tag{10.6}$$

and

$$Y^T = [Z_{p+1}, \ldots, Z_T], \tag{10.7}$$

$$X^T = \begin{bmatrix} Z_p & Z_{p+1} & \ldots & Z_T \\ \vdots & \vdots & \ddots & \vdots \\ Z_0 & Z_1 & \ldots & Z_{T-p} \end{bmatrix}, \tag{10.8}$$

$$\epsilon^T = (\varepsilon_{p+1} \ldots \varepsilon_T) \tag{10.9}$$

Then we can write

$$Y = X\alpha + \epsilon B^T \tag{10.10}$$

and thus

$$\mathbb{E}[Y - X\alpha] = \mathbb{E}[\epsilon B^T] = 0 \tag{10.11}$$

so we try to solve

$$||Y - X\alpha|| \to \min. \tag{10.12}$$

We denote by $\hat{\alpha}$ so solution of this problem, therefore we get

$$\hat{\alpha} = (X^T X)^{-1} X^T Y \tag{10.13}$$

For ergodic Z_t the $(X^T X)$ is just the empiric auto-covariance matrix. If Z is periodic, then we might get an eigenvalue of zero, thus the problem is no longer solveable. By changing the problem to

$$||Y - X\alpha|| + \Gamma||\alpha|| \to \min, \tag{10.14}$$

we might get better results. Thus we define:

$$\tilde{Y} = [Y\, 0] \tag{10.15}$$

$$\tilde{X} = [X\, \Gamma] \tag{10.16}$$

Then we can solve the problem $||\tilde{Y} - \tilde{X}\alpha||$ by using

$$\widehat{\tilde{\alpha}}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y} = (X^T X + \Gamma^T \Gamma)^{-1} X^T Y \tag{10.17}$$