Introduction to Time Series Analysis

Jussi Klemelä
University of Oulu

September 4, 2014

Contents

1) Introduction: Basic concepts (stationarity, deterministic trend, stochastic trend, autocovariance, white noise)

2) Confidence intervals for correlated data (central limit theorem and asymptotic variance, estimating the variance of the mean)

3) Prediction: Model free prediction (moving average, kernel estimator, time space smoothing and state space smoothing)

4) Time series models (prediction in ARMA and GARCH models)

5) Spatial statistics (basic concepts)

- The lectures are available at http://cc.oulu.fi/~jklemela/talks/exactus2014.pdf The exercises are available at http://cc.oulu.fi/~jklemela/talks/exactus2014-tsexercises.pdf

Part I: Introduction: Basic Concepts

- Stochastic process is a sequence of random variables index by time:

$$\ldots, Y_{-1}, Y_0, Y_1, \ldots, \qquad Y_0, Y_1, \ldots,$$

  where $Y_t \in \mathbf{R}$. We denote stochastic processes as

$$\{Y_t\}, \qquad \{Y_t\}_{t\in\mathbf{Z}}, \qquad \{Y_t\}_{t\in\mathbf{N}_0}, \qquad \{Y_t\}_{t\in\mathbf{R}}, \qquad \{Y_t\}_{t\in[0,\infty)}.$$

- Time series is a collection of observed values $y_1, \ldots, y_T \in \mathbf{R}$.

- Often the term "time series" means both the underlying process and the observed sequence.

Spatial Statistics

- Time series data have a temporal ordering. Time series models take into account that observations close together in time will be more closely related than distant observations.

- Spatial data are associated with geographical locations. Spatial models take into account the closeness in space.

Purposes of Time Series Analysis

- Unveiling the underlying probability law that governs the observed time series leads to
  (1) understanding the underlying dynamics,
  (2) forecasting future events,
  (3) controlling future events via intervention. (Fan and Zao, 2005)

- Prediction is an activity which is specific to time series analysis. We concentrate on prediction in these lectures.

- Note that regression analysis is related with prediction and control. In regression analysis current values of explanatory variables are used to explain current values of the response variable.

Frequency Domain and Time Domain

- Time series methods can be divided into frequency domain methods and into time domain methods.
- Frequency domain methods include estimation of the spectral density.
- Time domain methods include autocorrelation analysis in time space, prediction with moving averages,. . .

Strict Stationarity

- Time series models (ARMA and GARCH) are defined for stationary time series.

- Time series $\{Y_t\}$ is called strictly stationary, if $(Y_1, \ldots, Y_t)$ and $(Y_{1+k}, \ldots, Y_{t+k})$ are identically distributed for all $t, k \in \{0, \pm 1, \pm 2, \ldots\}$.

- The distribution of a stationary time series can be described by describing all finite dimensional marginal distribution: Distributions of

$$Y_1, (Y_1, Y_2), (Y_1, Y_2, Y_3), \ldots$$

Removing Trend

- Time series $Y_t = f(t) + \epsilon_t$ is not stationary, when $f(t)$ is a deterministic trend and ϵ_t is a stationary stochastic noise.

- A trend can be removed, for example, by
  (1) differencing: $\Delta Y_t = Y_t - Y_{t-1}$, or by
  (2) estimating the trend with $\hat{f}(t)$ and subtracting the estimated trend.

Two-sided Moving Average

- We observe time series $Y_1, \ldots, Y_T$. The two-sided moving average at t is

$$\hat{f}(t) = \frac{1}{2h+1} \sum_{i=t-h}^{t+h} Y_i, \qquad h = 0, 1, 2, \ldots, T-t.$$

- To get a more flexible class of moving averages we define

$$\hat{f}(t) = \sum_{i=1}^{T} p_i(t)\, Y_i,$$

where $p_i(t) = K((t-i)/h) / \sum_{j=1}^{T} K((t-j)/h)$, $K : \mathbf{R} \to \mathbf{R}$ is a kernel function, and $h > 0$ is smoothing parameter. We can take $K(x) = \exp\{-x^2\}$, for example.

Removing Trend: S&P 500

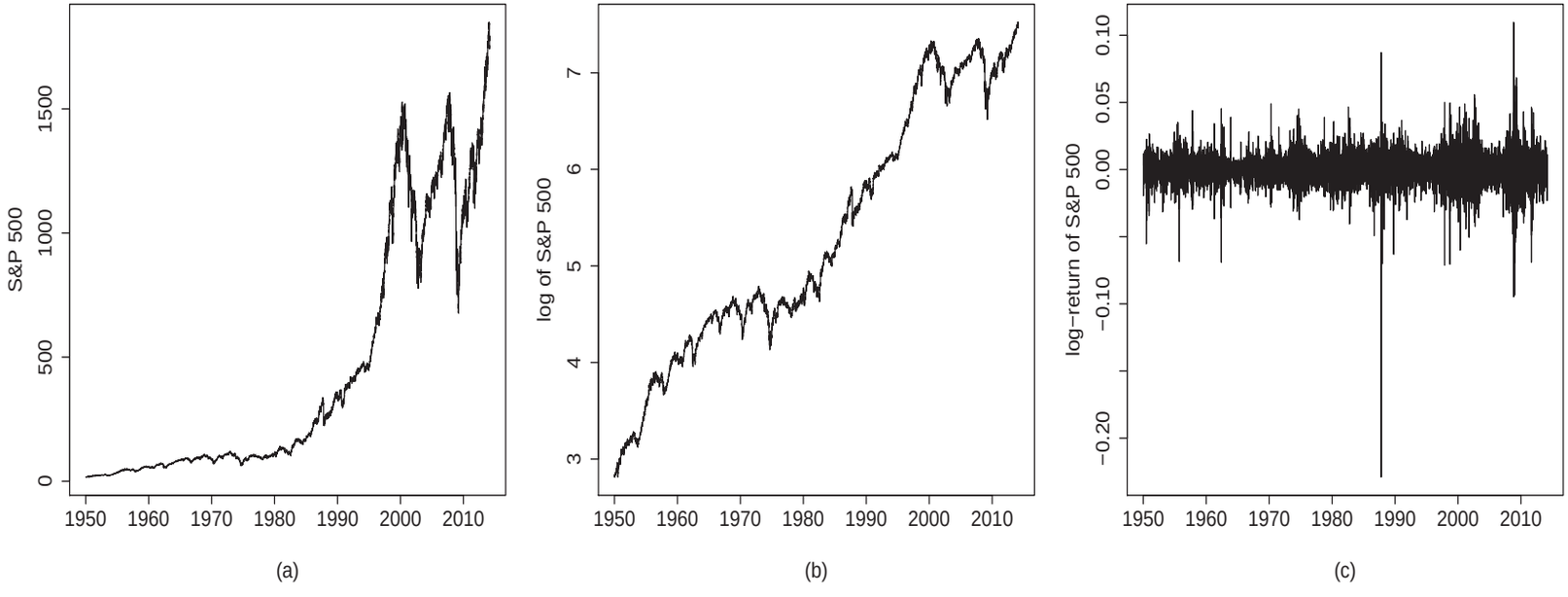

Figure 1: (a) S&P 500 prices S_t; (b) logarithms of S&P 500 prices $\log(S_t)$; (c) differences of the logarithmic prices $\log(S_t/S_{t-1})$.

Removing Trend: S&P 500 Differences

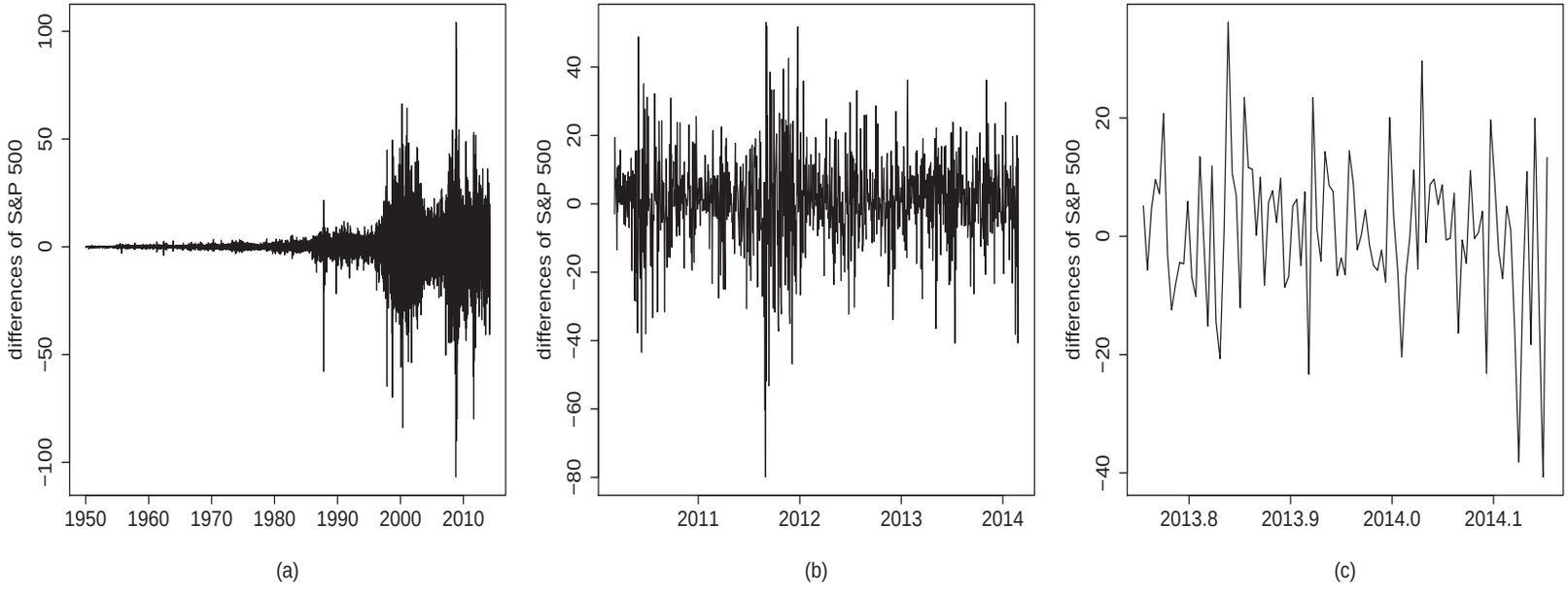

Figure 2: (a) Differences $S_t - S_{t-1}$ of S&P 500 prices over 65 years; (b) differences over 4 years: (c) differences over 100 days.

Econophysics

- Econophysics studies economical and financial phenomena using tools familiar from physics. Sometimes statistical finance is used as a synonym for econophysics. (Compare to statistical physics.)

- Econophysics started in the 1980's when large amounts of financial data became available. Econophysics sometimes applies statistical mechanics to economic analysis (interaction among many heterogeneous agents). For example, kinetic exchange models of markets (kinetic theory of gas), chaotic models, models with self-organizing criticality.

- Random walk of small particles in suspension in a fluid was discovered in 1827 by Robert Brown. Brownian motion has been used in option pricing to model stock prices by Bachelier (1900), Black and Scholes (1972). The theory of the Brownian motion and the atomistic backgrounds of diffusion were developed by Albert Einstein (1905)

Random Walk: Stochastic Trend

- Random walk is defined by $Y_t = Y_{t-1} + \epsilon_t$, where $t = 1, 2, \ldots$, ϵ_t is white noise, and Y_0 is a random variable or a constant.
- We have that $Y_t = Y_0 + \sum_{i=1}^{t} \epsilon_i$, because $Y_1 = Y_0 + \epsilon_1$, $Y_2 = Y_0 + \epsilon_2 + \epsilon_1, \ldots$, $Y_t = Y_0 + \epsilon_t + \cdots + \epsilon_1$.
- Random walk is not a stationary process, because $\text{Var}(Y_t) = t\text{Var}(\epsilon_1)$, when Y_0 is constant.
- Differenced random walk $\Delta Y_t = Y_t - Y_{t-1}$ is stationary because $\Delta Y_t = \epsilon_t$.

S&P 500 and Random Walk

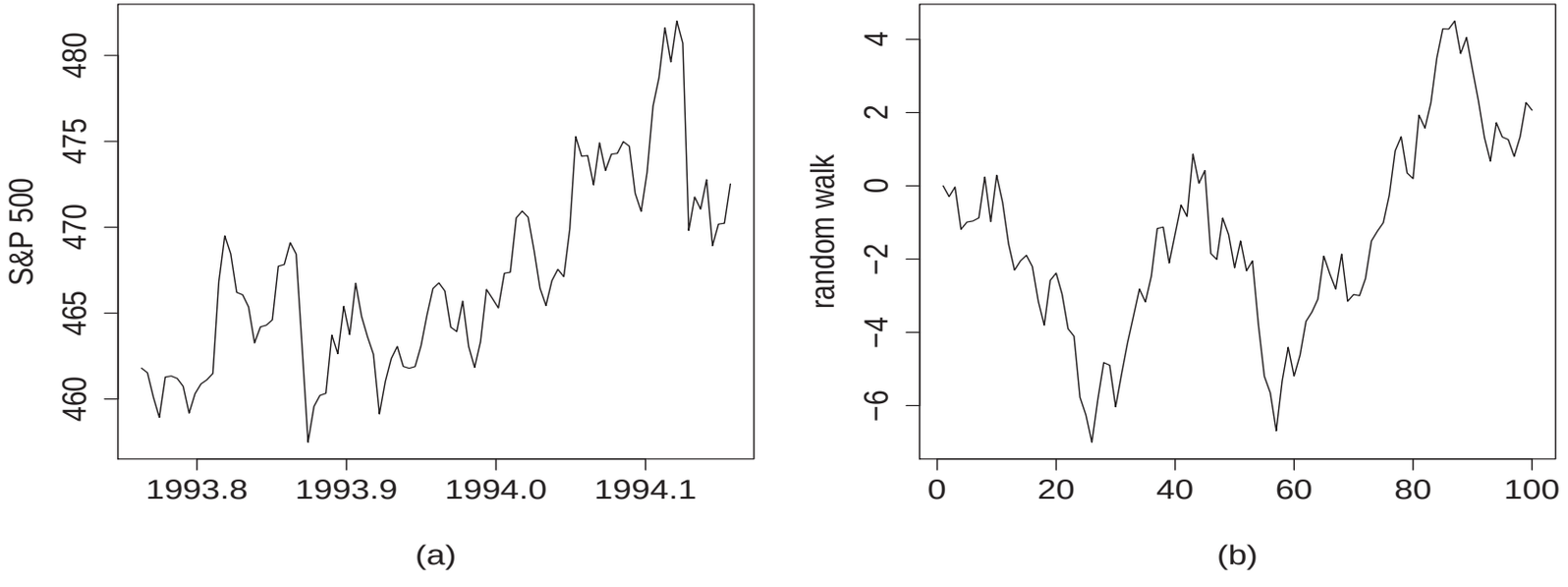

Figure 3: S&P 500 prices over 100 trading days and a simulated random walk with 100 steps.

Covariance Stationarity

- A time series $\{Y_t\}$ is covariance stationary if EY_t is constant, not depending on t, and $\text{Cov}(Y_t, Y_{t+k})$ depends only on $k \geq 1$ but not on t.

- If $EY_t^2 < \infty$, then strict stationarity implies covariance stationarity. Covariance stationarity does not imply strict stationarity.

- For a covariance stationary time series the autocovariance function is defined by $\gamma(k) = \text{Cov}(Y_t, Y_{t-k})$, where $k = 0, 1, \ldots$. Covariance stationarity implies that $\gamma(k)$ depends only on k and not on t.

- The autocorrelation function is defined as

$$\rho(k) = \text{Cor}(Y_t, Y_{t-k}) = \frac{\gamma(k)}{\gamma(0)},$$

  where $k = 0, 1, 2, \ldots$.

White Noise and IID Processes

- White noise and IID processes are used as innovation processes, to define ARMA and GARCH models.

- Time series $\{\epsilon_t\}$ is white noise if (1) $E\epsilon_t = 0$, (2) $E\epsilon_t^2 = \sigma^2$, (3) $E\epsilon_t\epsilon_{t+k} = 0$ for $k \neq 0$, where σ^2 is a constant.

- Time series $\{\epsilon_t\}$ is an IID process (independent identically distributed) if (1) $E\epsilon_t = 0$, (2) $E\epsilon_t^2 = \sigma^2$, (3) ϵ_t and ϵ_{t+k} are independent for $k \neq 0$.

- White noise process are used to define ARMA processes and IID processes are used to define GARCH processes.

PART II: Confidence Intervals for Correlated Data

Central Limit Theorem: I.I.D. Case

- Let $Y_1, Y_2, \ldots$ be a sequence of real-valued i.i.d. random variables with $\text{Var}(Y_i) = \sigma^2$, where $0 < \sigma^2 < \infty$. According to the central limit theorem we have

$$T^{-1/2} \sum_{i=1}^{T} (Y_i - \mu) \xrightarrow{d} N\left(0, \sigma^2\right),$$

  as $T \to \infty$, where $\mu = EY_i$.

- We have approximately $T^{-1} \sum_{i=1}^{T} Y_i - \mu \sim T^{-1/2} \sigma Z$, where $Z \sim N(0,1)$. Thus,

$$P\left(\left|\frac{1}{T} \sum_{i=1}^{T} Y_i - \mu\right| \leq x\right) \approx P\left(|Z| \leq T^{1/2} \sigma^{-1} x\right).$$

  (Then solve x from $1 - p = P(|Z| \leq T^{1/2}\sigma^{-1}x)$ to get $x = T^{-1/2}\sigma\Phi^{-1}(1 - p/2)$, where Φ is the distribution function of Z.)

Central Limit Theorem: Dependent Observations

- Let $(Y_t)_{t\in\mathbf{Z}}$ be a strictly stationary time series. Let $E|Y_t|^\delta < \infty$ and $\sum_{j=1}^{\infty} \alpha_j^{1-2/\delta} < \infty$ for some constant $\delta > 2$, where α_j are the α-mixing coefficients. Then,

$$T^{-1/2} \sum_{i=1}^{T} (Y_i - EY_i) \xrightarrow{d} N(0, \sigma^2),$$

as $T \to \infty$, where

$$\sigma^2 = \sum_{j=-\infty}^{\infty} \gamma(j) = \gamma(0) + 2 \sum_{j=1}^{\infty} \gamma(j),$$

$\gamma(j) = \text{Cov}(X_t, X_{t+j})$, and we assume that $\sigma^2 > 0$. See Ibragimov and Linnik (1971), Peligrad (1986).

α-Mixing Coefficients

- We define the weak dependence in terms of a condition on the α-mixing coefficients.

- Let $\mathcal{F}_i^j$ denote the sigma algebra generated by random variables $Y_i, \ldots, Y_j$. The α-mixing coefficient is defined as

$$\alpha_n = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_n^\infty} |P(A \cap B) - P(A)P(B)|,$$

  where $n = 1, 2, \ldots$.

Estimating the Asymptotic Variance

- To make a confidence interval we have to estimate the asymptotic variance σ^2.
- For i.i.d. data we can use the sample variance

$$\frac{1}{T}\sum_{i=1}^{T}(Y_i - \bar{Y})^2,$$

  where $\bar{Y} = T^{-1}\sum_{i=1}^{T} Y_i$.

- For dependent data we have to estimate autocovariances, because the asymptotic variance is

$$\sigma^2 = \gamma(0) + 2\sum_{j=1}^{\infty}\gamma(j).$$

Estimating the Asymptotic Variance, cont.

- We estimate σ^2 using the observations $Y_1, \ldots, Y_T$. An application of the sample covariances would lead to the estimator

$$\tilde{\sigma}^2 = \hat{\gamma}(0) + 2 \sum_{j=1}^{T-1} \hat{\gamma}(j),$$

where

$$\hat{\gamma}(j) = \frac{1}{T} \sum_{i=1}^{T-j} (Y_i - \bar{Y})(Y_{i+j} - \bar{Y}),$$

for $j = 0, \ldots, T-1$. Note that for large j only few observations are used in the estimator $\hat{\gamma}(j)$. For example, when $j = T-1$ the estimator uses only one observation: $\hat{\gamma}(T-1) = Y_1 Y_T / T$, which is an imprecise estimator.

Estimating the Asymptotic Variance, cont. 2

- We can use weighting to remove the imprecise estimators and define

$$\hat{\sigma}^2 = \hat{\gamma}(0) + 2\sum_{j=1}^{T-1} w(j)\,\hat{\gamma}(j),$$

where

$$w(j) = \max\left\{1 - \frac{j}{h}, 0\right\},$$

and $1 \leq h \leq T-1$ is a smoothing parameter (Newey-West estimator).

- We can generalize the estimator to other weights and define

$$w(j) = K(j/h),$$

where $K : \mathbf{R} \to \mathbf{R}$ is a kernel function satisfying $K(x) = K(-x)$, $K(0) = 1$, $|K(x)| \leq 1$ for all x, and $K(x) = 0$ for $|x| > 1$.

Spectral Density

- The weighting we have used is related to the smoothing in the estimation of the spectral density.

- The unnormalized spectral density function of a weakly stationary time series, having autocorrelation coefficients $\gamma(k)$ with $\sum_{j=-\infty}^{\infty} |\gamma(j)| < \infty$, is defined by

$$g(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(j)\, e^{-ij\omega},$$

  where $\omega \in [-\pi, \pi]$.

Estimation of Spectral Density

- The lag window spectral density estimator, based on data $Y_1, \ldots, Y_T$, is defined by

$$\hat{g}(\omega) = \frac{1}{2\pi} \sum_{|j| \le h} K(j/h)\, \hat{\gamma}(j)\, e^{-ij\omega},$$

  where $\hat{\gamma}(j)$ are the sample autocorrelation coefficients, $h = 1, 2, \ldots, T-1$.

- Now we have

$$\hat{g}(0) = \frac{1}{2\pi} \sum_{|j| \le h} K(j/h)\, \hat{\gamma}(j) = \frac{1}{2\pi}\, \hat{\sigma}^2.$$

Spectral Density: Motivation

- The periodic process is

$$Y_t = \sum_{j=1}^{k} A_j \cos(\omega_j t + \phi_j),$$

where $\phi_j \sim$ Uniform$[-\pi, \pi]$ are i.i.d, $0 \leq \omega_1 < \ldots < \omega_k \leq \pi$, and A_j are constants.

- The spectral distribution of $\{Y_t\}$ is the discrete distribution on $\{\omega_1, \ldots, \omega_k\}$ with

$$P(\omega_j) = \frac{A_j^2}{\sum_{i=1}^{k} A_i^2}, \qquad j = 1, \ldots, k.$$

PART III: Prediction: Model Free Predictors

Prediction

- Let us have observations $Y_1, \ldots, Y_t$. We want to predict the future value $Y_{t+\eta}$ for some prediction horizont $\eta \geq 1$.

- The best prediction in the mean squared error sense for $Y_{t+\eta}$ is $E(Y_{t+\eta} \,|\, Y_t, Y_{t-1}, \ldots, Y_1)$: The function $f(Y_t, \ldots, Y_1)$ minimizing $E(f(Y_t, \ldots, Y_1) - Y_{t+\eta})^2$ is equal to $f(Y_t, \ldots, Y_1) = E(Y_{t+\eta} \,|\, Y_t, Y_{t-1}, \ldots, Y_1)$.

- $E(Y_{t+\eta} \,|\, Y_t, \ldots, Y_1)$ is approximately equal to $E(Y_{t+\eta} \,|\, Y_t, Y_{t-1}, \ldots)$.

- Assume that we have a time series model for $\{Y_t\}$, like an ARMA or GARCH model. To get a predictor, we can estimate the parameters of the model, derive the formula for the conditional expectation, and plug-in the estimated parameters into the formula.

- We can also use nonparametric predictors, like moving averages.

Comparing Predictors

- Let us have data $Y_1, \ldots, Y_T$. Let $\hat{f}(t)$ be a predictor of $Y_{t+\eta}$, constructed using data $Y_1, \ldots, Y_t$.

- The mean squared prediction error, with starting point t_0, is

$$\text{MSPE}(\hat{f}) = \frac{1}{T - \eta - t_0 + 1} \sum_{t=t_0}^{T-\eta} \left(\hat{f}(t) - Y_{t+\eta} \right)^2 .$$

- We choose the predictor with the smallest mean squared prediction error. This is called cross-validation.

Moving Average Predictors

- We observe time series $Y_1, \ldots, Y_t$. We define the moving average prediction of $Y_{t+\eta}$ for $\eta \geq 1$ as

$$\hat{f}(t) = \frac{1}{h+1} \sum_{i=t-h}^{t} Y_i, \qquad h = 0, 1, 2, \ldots, t-1.$$

- To get a more flexible class of moving averages we define the one-sided moving average

$$\hat{f}(t) = \sum_{i=1}^{t} p_i(t)\, Y_i,$$

where $p_i(t) = K((t-i)/h) / \sum_{j=1}^{t} K((t-j)/h)$, we use a general kernel function $K : [0, \infty) \to \mathbf{R}$, and smoothing parameter $h > 0$. We can take $K(x) = \exp(-x) I_{[0,\infty)}(x)$, for example.

Moving Average Predictors, cont.

- Note that the prediction step $\eta \geq 1$ does not show up in the definition of the one-sided moving average. The prediction step η can affect the choice of the smoothing parameter h. It is natural to choose a large smoothing parameter h when the prediction step η is large. Then a longe horizon predictor would be close to the artihmetic mean $t^{-1} \sum_{i=1}^{t} Y_i$.

Exponential Moving Average

- The exponential moving average is a one-sided moving average obtained by taking $K(x) = \exp(-x)\, I_{[0,\infty)}(x)$ and $h = -1/\log\gamma$, where $0 < \gamma < 1$. Now the one-sided moving average is equal to

$$\hat{f}(t) = \frac{1-\gamma}{1-\gamma^t} \sum_{i=1}^{t} \gamma^{t-i} Y_i.$$

- We get a slightly different exponential moving average by making the recursive definition

$$\mathrm{ma}(t) = (1-\gamma)Y_t + \gamma \mathrm{ma}(t-1),$$

where $0 \le \gamma \le 1$. This leads to $\mathrm{ma}(t) = (1-\gamma)\sum_{i=1}^{t} \gamma^{t-i} Y_i$, when we choose the initial value $\mathrm{ma}(1) = (1-\gamma)Y_1$.

State Space Predictors

- A state space predictor is a predictor which is obtained from a regression function estimator. A regression function $f : \mathbf{R}^d \to \mathbf{R}$ is the conditional expectation $f(x) = E(Z \mid X = x)$, where $Z \in \mathbf{R}$ is the response variable and $X \in \mathbf{R}^d$ is the explanatory variable. A regression function estimator is a function $\hat{f} : \mathbf{R}^d \to \mathbf{R}$, which is calculated from regression data $(X_1, Z_1), \ldots, (X_n, Z_n)$, consisting of identically distributed observations.

- In the case of time series we take the regression data to be the sequence

$$(X_1, Y_{1+\eta}), \ldots, (X_{T-\eta}, Y_T),$$

  where $X_t \in \mathbf{R}^d$ contains information which is available at time t, and $\eta \geq 1$ is the prediction horizont.

State Space Predictors (continued)

- For example, the state variable X_t can be the sequence of the previous observations of Y_t, so that we define

$$X_t = (Y_{t-d+1}, \ldots, Y_t),$$

where $d \geq 1$.

- Let us denote $n = T - \eta$ and $Z_t = Y_{t+\eta}$, so that the previously defined regression data can be written as

$$(X_1, Z_1), \ldots, (X_n, Z_n),$$

where $X_t \in \mathbf{R}^d$ are observations from d explanatory variables and $Z_t \in \mathbf{R}$ are observations from the response variable. If $\hat{f}$ is a regression function estimator, we predict the value $Y_{T+\eta}$ by $\hat{f}(X_T)$, using the value X_T of the state variable observed at time T.

Linear Least Squares Regression

- Linear least squares regression function estimator is

$$\hat{f}(x) = \hat{\alpha} + \hat{\beta}' x,$$

where $x \in \mathbf{R}^d$ and $(\hat{\alpha}, \hat{\beta})$ is obtained as the minimizer of

$$\sum_{i=1}^{n} (Z_i - \alpha - \beta' X_i)^2 .$$

Kernel Regression

- Kernel regression estimator is

$$\hat{f}(x) = \sum_{i=1}^{n} p_i(x)\, Z_i,$$

where

$$p_i(x) = \frac{K((x - Z_i)/h)}{\sum_{j=1}^{n} K((x - Z_j)/h)},$$

$K : \mathbf{R}^d \to \mathbf{R}$ is a kernel function and $h > 0$ is the smoothing parameter.

Local Likelihood

- Let $Y_1, \ldots, Y_T$ be identically distributed with density f_θ, where $\theta \in \Theta \subset \mathbf{R}^p$. Thus the distribution of Y_t is modeled with a collection $(f_\theta, \theta \in \Theta)$ of densities.

- If $Y_1, \ldots, Y_T$ are independent, then the density of $(Y_1, \ldots, Y_T)$ is

$$f(y_1, \ldots, y_T) = \prod_{i=1}^{T} f_\theta(y_i).$$

- The maximum likelihood estimator of θ is the value $\hat{\theta}$ maximizing $\sum_{i=1}^{T} \log f_\theta(Y_i)$ over $\theta \in \Theta$.

- We can find a time varying estimator $\hat{\theta}_t$ using either time space or state space smoothing.

Time Space Localization

- Let $p_i(t)$ be the kernel weights. Let $\hat{\theta}_t$ be the value maximizing

$$\sum_{i=1}^{t} p_i(t) \log f_\theta(Y_i)$$

  over $\theta \in \Theta$.

- For example, let

$$Y_t = \mu_t + \sigma_t \epsilon_t,$$

  where $\epsilon_t \sim N(0,1)$ are i.i.d. Let $\theta = (\mu, \sigma^2)$ and $f_\theta(y) = \phi((y-\mu)/\sigma)/\sigma$, where $\phi(y) = (2\pi)^{-1/2} \exp\{-y^2/2\}$ is the density of the standard normal distribution. Now $\hat{\theta}_t = (\hat{\mu}_t, \hat{\sigma}_t^2)$, where $\hat{\mu}_t = \sum_{i=1}^t p_i(t) Y_i$ and $\hat{\sigma}_t^2 = \sum_{i=1}^t p_i(t) Y_i^2 - \hat{\mu}_t^2$.

State Space Localization

- In addition to the time series $Y_1, \ldots, Y_T$, let us observe the state variables $X_1, \ldots, X_T$. Let $p_i(x)$ be the weight defined by $p_i(x) = K((x - X_i)/h)$, where $K : \mathbf{R}^d \to \mathbf{R}$ is a kernel function and $h > 0$ is the smoothing parameter.

- Let $\hat{\theta}_t$ be the value maximizing $\sum_{i=1}^{t} p_i(X_t) \log f_\theta(Y_i)$ over $\theta \in \Theta$.

- Let

$$Y_t = \mu(X_t) + \sigma(X_t)\epsilon_t,$$

where $\epsilon_t \sim N(0, 1)$. (We have $Y \,|\, X = x \sim N(\mu(x), \sigma(x))$.) Denote $\theta = (\mu, \sigma^2)$ and $f_\theta(y) = \phi((y - \mu)/\sigma)/\sigma$, where ϕ is the density of the standard normal distribution. Then, $\hat{\theta}_t = (\hat{\mu}_t, \hat{\sigma}_t^2)$, where

$$\hat{\mu}_t = \sum_{i=1}^{t} p_i(X_t) Y_i, \qquad \hat{\sigma}_t^2 = \sum_{i=1}^{t} p_i(X_t) Y_i^2 - \hat{\mu}_t^2.$$

Local Least Squares

- Time space smoothing: Define the time varying regression coefficients as the values $\hat{\alpha}_t$ and $\hat{\beta}_t$ minimizing

$$\sum_{i=1}^{t} (Y_i - \alpha - \beta' Z_i)^2 \, p_i(t),$$

where $p_i(t) = K((t-i)/h)$ are the kernel weights.

- State space smoothing: Let $\hat{\alpha}_t$ and $\hat{\beta}_t$ be the values minimizing

$$\sum_{i=1}^{t} (Y_i - \alpha - \beta' Z_i)^2 \, p_i(X_t),$$

where $p_i(x) = K((x - X_i)/h)$, where $K : \mathbf{R}^d \to \mathbf{R}$ is a kernel function and $h > 0$ is the smoothing parameter.

PART IV: TIME SERIES MODELS

MA(q) Process

- We define a moving average process $\{Y_t\}$ of order $q \geq 0$ as a process satisfying

$$Y_t = \mu + b_0\epsilon_t + b_1\epsilon_{t-1} + \cdots + b_q\epsilon_{t-q},$$

where $\mu, b_0, \ldots, b_q \in \mathbf{R}$ and $\{\epsilon_t\}$ is a white noise process WN$(0, \sigma^2)$.

Properties of MA(q) Process

- We have that

$$EY_t = \mu, \qquad \text{Var}(Y_t) = \sigma^2 \left(b_0^2 + b_1^2 + \cdots + b_q^2 \right),$$

and

$$EY_t Y_{t+k} = \begin{cases} \sigma^2 \sum_{j=0}^{q-k} b_j b_{k+j}, & k = 1, \ldots, q, \\ 0, & k > q. \end{cases}$$

- Thus MA(q) process is such that the correlation exists between Y_t and Y_{t+k} only if $|k| \leq q$. Previous two equations show that MA(q) process is covariance stationary.

Prediction for MA(q) Processes

- The conditional expectation $E\left(Y_{t+\eta} \,|\, Y_t, Y_{t-1}, \ldots\right)$ is the best prediction of $Y_{t+\eta}$, $\eta \geq 1$, given the infinite past $Y_t, Y_{t-1}, \ldots$, in the sense of the mean squared prediction error.

- For the MA(1) process $Y_t = \epsilon_t + b\epsilon_{t-1}$ we have

$$E\left(Y_{t+\eta} \,|\, Y_t, Y_{t-1}, \ldots\right) = \begin{cases} b\sum_{k=0}^{\infty}(-1)^k b^k Y_{t-k}, & \eta = 1, \\ 0, & \eta \geq 2. \end{cases}$$

- The prediction formula for prediction step $\eta = 1$ is a version of exponential moving average.

Autoregressive Processes

- An autoregressive process $\{Y_t\}$ of order $p \geq 1$ is a process satisfying

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_p Y_{t-p} + \epsilon_t,$$

where $a_1, \ldots, a_p \in \mathbf{R}$, $\{\epsilon_t\}$ is a white noise process, and $t = 0, \pm 1, \pm 2, \ldots$. We assume that ϵ_t is uncorrelated with $Y_{t-1}, Y_{t-2}, \ldots$. We use AR(p) as a short hand notation for an autoregressive process of order p.

Prediction of Autoregressive Processes AR(p)

- Let us consider the prediction of $Y_{t+\eta}$ for $\eta \geq 1$.
- The best prediction of Y_{t+1}, given the observations $Y_t, Y_{t-1}, \ldots$, is

$$E(Y_{t+1} \,|\, \mathcal{F}_t) = a_1 Y_t + a_2 Y_{t-1} + \cdots + a_p Y_{t-p+1},$$

because $E(\epsilon_{t+1} \,|\, \mathcal{F}_t) = 0$.

- For the two step prediction the best predictor is

$$\begin{aligned} E(Y_{t+2} \,|\, \mathcal{F}_t) &= E[E(Y_{t+2} \,|\, \mathcal{F}_{t+1}) \,|\, \mathcal{F}_t] \\ &= E[a_1 Y_{t+1} + a_2 Y_t + \cdots + a_p Y_{t-p+2} \,|\, \mathcal{F}_t] \\ &= a_1 \text{pred}_t(1) + a_2 Y_t + \cdots + a_p Y_{t-p+2}, \end{aligned}$$

where $\text{pred}_t(1) = E(Y_{t+1} \,|\, \mathcal{F}_t)$.

Prediction of Autoregressive Processes AR(p), cont.

- The general prediction formula is

$$E(Y_{t+\eta} \mid \mathcal{F}_t) = a_1 \text{pred}_t(\eta - 1) + a_2 \text{pred}_t(\eta - 2) + \cdots + a_p \text{pred}_t(\eta - p),$$

  where $\text{pred}_t(\eta - k) = E(Y_{t+\eta-k} \mid \mathcal{F}_t)$.

- In particular, for the AR(1) process $Y_t = aY_{t-1} + \epsilon_t$ we have

$$E\left(Y_{t+\eta} \mid Y_t, Y_{t-1}, \ldots\right) = a^\eta Y_t.$$

ARMA Processes

- We define an autoregressive moving average process $\{Y_t\}$, of order (p, q), $p, q \geq 0$, as a process satisfying

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_p Y_{t-p} + u_t,$$

where $a_1, \ldots, a_p \in \mathbf{R}$ and $\{u_t\}_{t \in \mathbf{Z}}$ is a MA(q) process. We use ARMA(p, q) as a short hand notation for an autoregressive moving average process of order (p, q).

Prediction for ARMA Processes

- For the ARMA(1,1) process $Y_t = aY_{t-1} + \epsilon_t + b\epsilon_{t-1}$ we have

$$E\left(Y_{t+\eta} \mid Y_t, Y_{t-1}, \ldots\right) = a^{\eta-1}(a+b)\sum_{k=0}^{\infty}(-1)^k b^k Y_{t-k},$$

  where $\eta \geq 1$.

- In particular, for the AR(1) process $Y_t = aY_{t-1} + \epsilon_t$ we have

$$E\left(Y_{t+\eta} \mid Y_t, Y_{t-1}, \ldots\right) = a^{\eta} Y_t.$$

Stationarity in ARMA Processes

- Covariance stationarity does not guarantee that the ARMA(p, q) process would be suitable for modeling.

- Let us consider the AR(1) model

$$Y_t = aY_{t-1} + \epsilon_t,$$

  where $\{\epsilon_t\} \sim \text{WN}(0, \sigma^2)$.

- The AR(1) model is covariance stationary if and only if $|a| \neq 1$ but we require that $|a| < 1$.

Simulating AR(1) Process

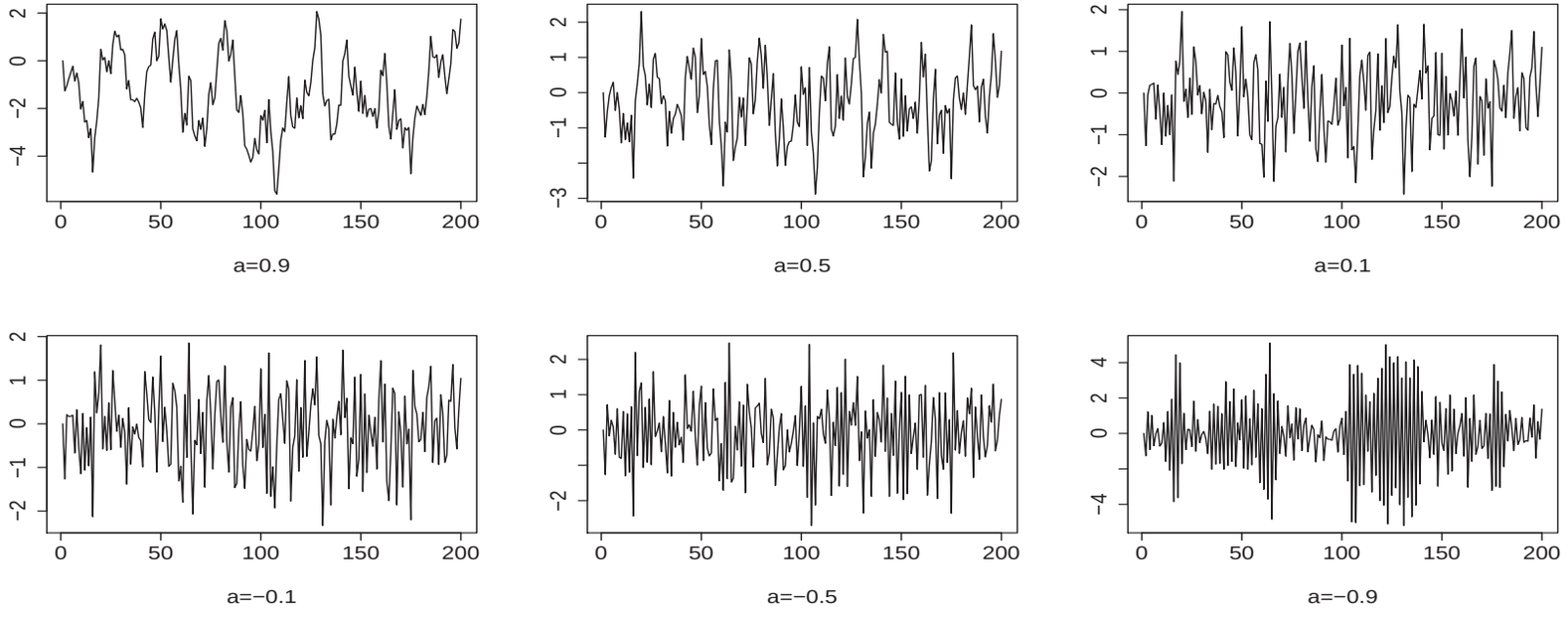

Figure 4: Simulating $Y_t = aY_{t-1} + \epsilon_t$.

Stationarity in ARMA Processes, cont.

- Let us consider first the case $|a| < 1$. We can write recursively

$$\begin{aligned} Y_t &= aY_{t-1} + \epsilon_t \\ &= a^2 Y_{t-2} + a\epsilon_{t-1} + \epsilon_t \\ &\vdots \\ &= a^{k+1} Y_{t-k-1} + a^k \epsilon_{t-k} + \cdots + a\epsilon_{t-1} + \epsilon_t, \end{aligned}$$

where $k \geq 0$. Since $|a| < 1$, we get the MA(∞) representation

$$Y_t = \sum_{j=0}^{\infty} a^j \epsilon_{t-j},$$

which implies that $\{Y_t\}$ is covariance stationary.

Stationarity in ARMA Processes, cont. 2

- Let us then consider the case $|a| > 1$. Since $Y_{t+1} = aY_t + \epsilon_{t+1}$, we have

$$\begin{aligned} Y_t &= a^{-1}Y_{t+1} - a^{-1}\epsilon_{t+1} \\ &= a^{-2}Y_{t+2} - a^{-2}\epsilon_{t+2} - a^{-1}\epsilon_{t+1} \\ &\vdots \\ &= a^{-k-1}Y_{t+k+1} - a^{-k-1}\epsilon_{t+k+1} - \cdots - a^{-1}\epsilon_{t+1}, \end{aligned}$$

  where $k \geq 0$. Since $|a| > 1$, we get the MA(∞) representation

$$Y_t = -\sum_{j=1}^{\infty} a^{-j}\epsilon_{t+j},$$

  which implies that $\{Y_t\}$ is covariance stationary. The latter case $|a| > 1$ is not suitable for modeling because Y_t is a function of future innovations ϵ_{t+j} with $j \geq 1$.

Stationarity in ARMA Processes, cont. 3

- We define causality of the process to exclude examples like the AR(1) model with $|a| > 1$.

- An ARMA(p, q) process $\{Y_t\}$ is called causal if there exists constants $\{\psi_j\}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and

$$Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}, \qquad t = 0, \pm 1, \pm 2, \ldots$$

- Let $Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_p Y_{t-p} + b_0 \epsilon_t + b_1 \epsilon_{t-1} + \cdots + b_q \epsilon_{t-q}$. Denote

$$a(z) = 1 - a_1 z - a_2 z^2 - \cdots - a_p z^p, \qquad b(z) = b_0 + b_1 z + b_2 z^2 + \cdots + b_q z^q,$$

where $z \in C$. Let the polynomials $a(z)$ and $b(z)$ have no common zeroes. Then $\{Y_t\}$ is causal if and only if $a(z) \neq 0$ for all $z \in C$ such that $|z| \leq 1$.

Dependent Noncorrelated Process

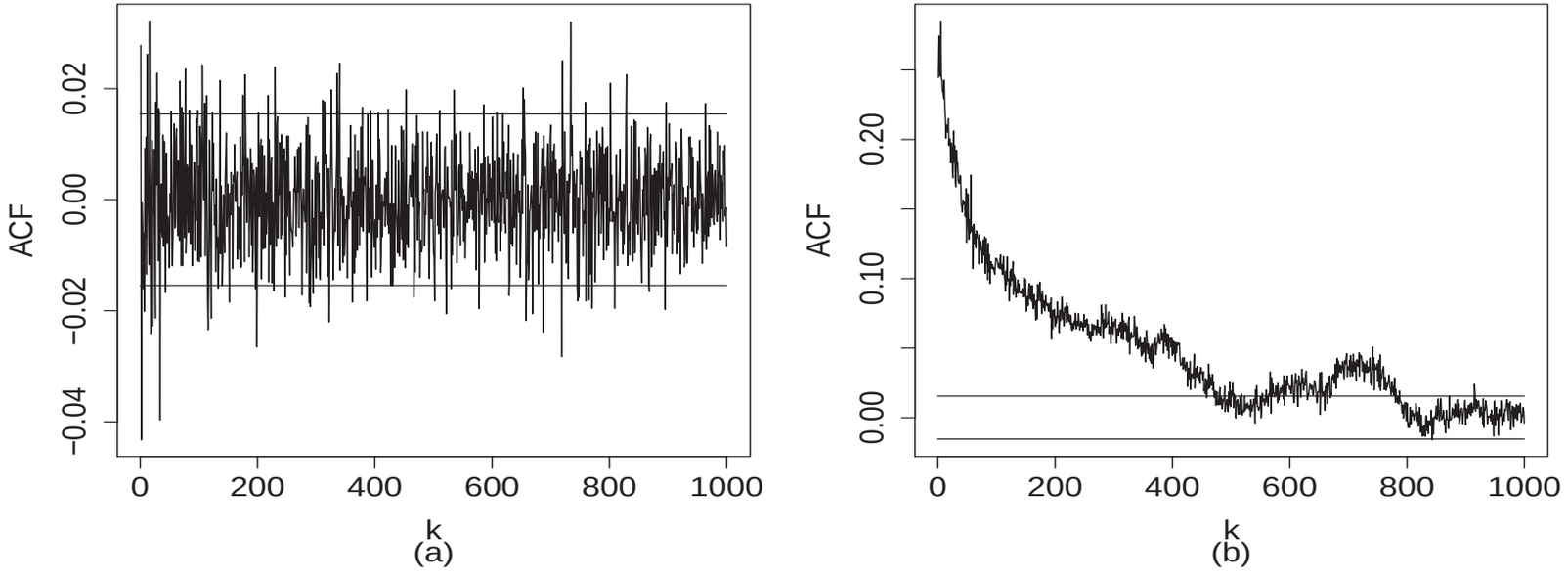

Figure 5: (a) Autocorrelation function $\rho(k) = \text{Cor}(Y_t, Y_{t+k})$ of S&P 500 returns Y_t; (b) autocorrelation function $\rho(k) = \text{Cor}(|Y_t|, |Y_{t+k}|)$ of absolute S&P 500 returns $|Y_t|$.

Conditional Heteroskedasticity Models

- Time series $\{Y_t\}$ satisfies the conditional heteroskedasticity assumption if

$$Y_t = \sigma_t \epsilon_t, \qquad t = 0, \pm 1, \pm 2, \ldots,$$

  where $\{\epsilon_t\}$ is an IID$(0, 1)$ process and $\{\sigma_t\}$ is the volatility process.

- The innovation process ϵ_t is independent of $Y_{t-1}, Y_{t-2}, \ldots$. The volatility process σ_t is a function of $Y_{t-1}, Y_{t-2}, \ldots$. (σ_t is measurable with respect to the sigma-field $\mathcal{F}_{t-1}$ generated by the variables $Y_{t-1}, Y_{t-2}, \ldots$).

Prediction under Conditional Heteroskedasticity

- Denote $E(Y_{t+1}^2 \,|\, \mathcal{F}_t) = E(Y_{t+1}^2 \,|\, Y_t, Y_{t-1}, \ldots)$. Under the conditional heteroskedasticity model,

$$E(Y_{t+1}^2 \,|\, \mathcal{F}_t) = E(\sigma_{t+1}^2 \epsilon_{t+1}^2 \,|\, \mathcal{F}_t) = \sigma_{t+1}^2 E(\epsilon_{t+1}^2 \,|\, \mathcal{F}_t) = \sigma_{t+1}^2 E(\epsilon_{t+1}^2) = \sigma_{t+1}^2.$$

  Thus, the best prediction of Y_{t+1}^2, given $Y_t, Y_{t-1}, \ldots$, is equal to σ_{t+1}^2, in the mean squared error sense.

- Also, for $\eta \geq 1$,

$$\begin{aligned} E(Y_{t+\eta}^2 \,|\, \mathcal{F}_t) &= E(\sigma_{t+\eta}^2 \epsilon_{t+\eta}^2 \,|\, \mathcal{F}_t) = E\left[E(\sigma_{t+\eta}^2 \epsilon_{t+\eta}^2 \,|\, \mathcal{F}_{t+\eta-1}) \,|\, \mathcal{F}_t\right] \\ &= E\left[\sigma_{t+\eta}^2 E(\epsilon_{t+\eta}^2 \,|\, \mathcal{F}_{t+\eta-1}) \,|\, \mathcal{F}_t\right] = E\left[\sigma_{t+\eta}^2 \,|\, \mathcal{F}_t\right]. \end{aligned}$$

  Thus, the best prediction the best prediction of $Y_{t+\eta}^2$, given $Y_t, Y_{t-1}, \ldots$ is equal to the best prediction of $\sigma_{t+\eta}^2$.

ARCH Processes

- Process $\{Y_t\}$ is an ARCH(q) process (autoregressive conditional heteroscedasticity process of order $p \geq 0$), if

$$Y_t = \epsilon_t \sigma_t, \qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i}^2,$$

  where $\{\epsilon_t\}$ is an IID$(0, 1)$ process and $\alpha_0 > 0$ and $\alpha_1, \ldots, \alpha_p \geq 0$.

- The ARCH model was introduced by Engle (1982) for modeling U.K. inflation rates.

- The ARCH(p) process is strictly stationary if $\sum_{i=1}^{p} \alpha_i < 1$.

Prediction of ARCH Processes

- Let us consider prediction in the ARCH(p) model. The best one step prediction of the squared value is obtained from as

$$E\left(Y_{t+1}^2 \,|\, \mathcal{F}_t\right) = \sigma_{t+1}^2 = \alpha_0 + \alpha_1 Y_t^2 + \cdots + \alpha_p Y_{t-p+1}^2.$$

- The best η-step prediction of Y_t^2 in the ARCH(1) model is

$$E\left(Y_{t+\eta}^2 \,|\, \mathcal{F}_t\right) = \bar{\sigma}^2 + \alpha_1^{\eta-1}\left(\sigma_{t+1}^2 - \bar{\sigma}^2\right) = \alpha_0 \frac{1-\alpha_1^\eta}{1-\alpha_1} + \alpha_1^\eta Y_t^2,$$

where $\eta \geq 1$, we assumed condition $\alpha_1 < 1$, which guarantees stationarity, and we denote $\bar{\sigma}^2 = EY_t^2 = \alpha_0/(1-\alpha_1)$.

GARCH Processes

- Process $\{Y_t\}$ is a GARCH(p, q) process (generalized autoregressive conditional heteroskedasticity process of order $p \geq 0$ and $q \geq 0$), if

$$Y_t = \epsilon_t \sigma_t, \qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2,$$

  where $\{\epsilon_t\}$ is an IID$(0, 1)$ process, $\alpha_0 > 0$, $\alpha_1, \ldots, \alpha_p \geq 0$, and $\beta_1, \ldots, \beta_q \geq 0$.

- As a special case we get the GARCH$(1, 1)$ model, where $\sigma_t^2 = \alpha_0 + \alpha Y_{t-1}^2 + \beta \sigma_{t-1}^2$.

- The GARCH model was introduced in Bollerslev (1986). The GARCH(p, q) process is strictly stationary if

$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1.$$

Prediction in GARCH Processes

- The best one step prediction of the squared value is

$$E\left(Y_{t+1}^2 \,|\, \mathcal{F}_t\right) = \sigma_{t+1}^2.$$

- In the GARCH$(1,1)$ model the best η-step prediction of the squared value, in the mean squared error sense, is

$$E\left(Y_{t+\eta}^2 \,|\, \mathcal{F}_t\right) = \bar{\sigma}^2 + (\alpha_1 + \beta)^{\eta-1}\left(\sigma_{t+1}^2 - \bar{\sigma}^2\right), \qquad \eta \geq 1,$$

  where we assumed condition $\alpha_1 + \beta < 1$, which guarantees stationarity, and we denote $\bar{\sigma}^2 = EY_t^2 = \alpha_0/(1-\alpha_1-\beta)$. In fact, $\bar{\sigma}^2$ is the unconditional variance.

- We have

$$\sigma_{t+1}^2 = \frac{\alpha_0}{1-\beta} + \alpha_1 \sum_{k=0}^{\infty} \beta^k Y_{t-k}^2.$$

PART V: Spatial Statistics

Spatial Data

- Random field underlying spatial data is a collection of random variables index by location:

$$\{Y_t\}_{t \in \mathbf{T}},$$

where

$$\mathbf{T} = \mathbf{R}^2, \qquad \mathbf{T} = \mathbf{R}^3, \qquad \mathbf{T} = \{x \in \mathbf{R}^3 : \|x\| = 1\}.$$

- Spatial data is a collection of observed values $y_{t_1}, \ldots, y_{t_n} \in \mathbf{R}$.

Moving Average

- In time series analysis we defined two-sided and one-sided moving average.
- We observe spatial data $Y_1, \ldots, Y_T$. The moving average at t is

$$\hat{f}(t) = \frac{1}{\#\mathcal{N}(t)} \sum_{i \in \mathcal{N}(t)} Y_i,$$

where $\mathcal{N}(t)$ is a neigborhood of $t \in \{1, \ldots, T\}$.

Neigborhoods

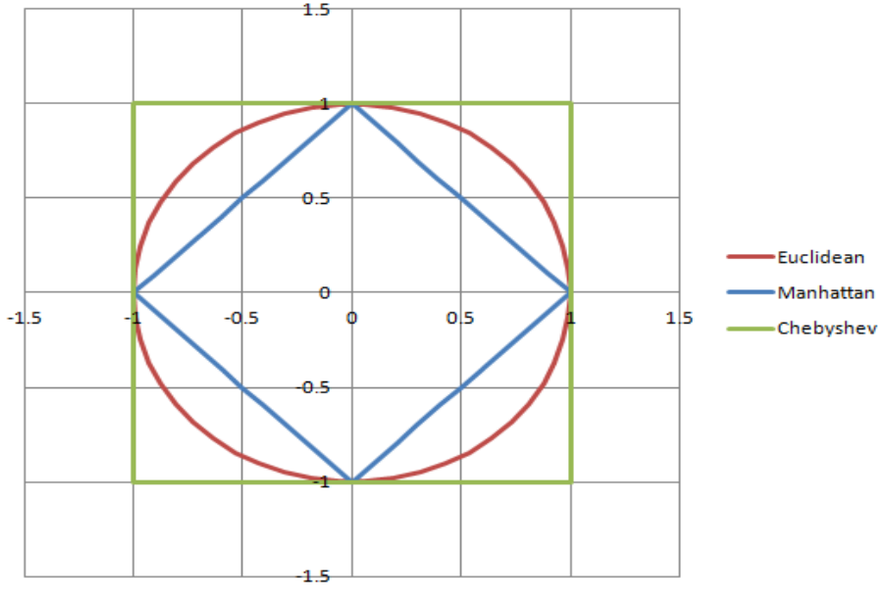

Figure 6: *Euclidean, Chebyshev, and Manhattan distances*

Source: http://gamedev.stackexchange.com/questions/75702/eculidean-space-and-vector-magnitude. Take $||x|| = (\sum_{i=1}^{d} = x_i^2)^{1/2}$, $||x|| = \max_{i=1,\ldots,d} |x_i|$, or $||x|| = \sum_{i=1}^{d} |x_i|$.

Kernel Estimator

- To get a more flexible class of moving averages we define

$$\hat{f}(t) = \sum_{i=1}^{T} p_i(t)\, Y_i,$$

where $p_i(t) = K((t-i)/h) / \sum_{j=1}^{T} K((t-j)/h)$, $K : \mathbf{T} \to \mathbf{R}$ is a kernel function, and $h > 0$ is smoothing parameter. We can take $K(x) = \exp\{-\|x\|^2\}$, for example.

Markov Random Fields

- The distribution of a stationary time series can be described by describing all finite dimensional marginal distribution: Distributions of $Y_1, (Y_1, Y_2), (Y_1, Y_2, Y_3), \ldots$ For a Markov time series it enough to define the distribution of $Y_1, (Y_1, Y_2)$, because

$$Y_t \,|\, Y_{t-1}, Y_{t-2}, \ldots \sim Y_t \,|\, Y_{t-1}.$$

- Markov random field is such that

$$Y_t \,|\, (Y_s, s \neq t) \sim Y_t \,|\, (Y_s, s \in \mathcal{N}(t), s \neq t),$$

where $\mathcal{N}(t)$ is a neigborhood of t.

Covariance Function and Stationarity

- Let $\{Y_t\}_{t\in\mathbf{T}}$ be random field. The covariance function is $C : \mathbf{T}\times\mathbf{T} \to \mathbf{R}$ with

$$C(t,s) = \mathrm{Cov}(Y_t, Y_s) = E(Y_tY_s) - EY_tEY_s.$$

- A random field is covariance stationary if EY_t is a constant, not depending on t, and

$$C(t,s) = C(t+\tau, s+\tau)$$

for $t, s, \tau \in \mathbf{T}$. Now we can denote $\gamma(\tau) = C(s,t)$, when $\tau = s - t$.

- A random field $\{Y_t\}_{t\in\mathbf{T}}$ is strictly stationary if

$$(Y_{t_1+\tau}, \ldots, Y_{t_k+\tau}) \sim (Y_{t_1}, \ldots, Y_{t_k})$$

for all $\tau \in \mathbf{T}$ and all $t_1, \ldots, t_k \in \mathbf{T}$.

Central Limit Theorem for Random Fields

- Let $\{Y_t\}_{t \in \mathbf{Z}^d}$ be random field. Let Λ_n be a sequence of "increasing" finite subsets of $\mathbf{Z}^d$. Under "mixing conditions" for the random field,

$$(\#\Lambda_n)^{-1/2} \sum_{t \in \Lambda_n} (Y_t - \mu) \xrightarrow{d} N(0, \sigma^2),$$

  where $EY_t = \mu$, $\sigma^2 = \sum_{t \in \mathbf{Z}^d} \mathrm{Cov}(Y_0, Y_t)$, and $\#\Lambda_n$ is the number of elements in Λ_n; see Bolthausen (1982).

- Compare to the time series case: $T^{-1/2} \sum_{i=1}^{T} (Y_i - \mu) \xrightarrow{d} N(0, \sigma^2)$, where $\sigma^2 = \sum_{j=-\infty}^{\infty} \mathrm{Cov}(Y_0, Y_j)$.

Spectral Density

- In time series analysis covariance function can be analyzed with the help of a spectral density: we defined $g(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(j)\, e^{-ij\omega}$, where $\omega \in [-\pi, \pi]$, which implies $\gamma(\tau) = \int_{-\pi}^{\pi} e^{i\omega\tau} g(\omega)\, d\omega$.

- Let $\{Y_t\}_{t \in \mathbf{T}}$ be random field with $\mathbf{T} = \mathbf{R}^d$. The covariance function is $C : \mathbf{T} \times \mathbf{T} \to \mathbf{R}$ with $C(t, s) = \mathrm{Cov}(Y_t, Y_s) = E(Y_t Y_s) - EY_t EY_s$. Let $\gamma(\tau) = C(t, s)$ for $\tau = t - s$. Function $f : \mathbf{T} \to \mathbf{R}$ is spectral density if

$$\gamma(\tau) = \int_{\mathbf{T}} e^{i\omega'\tau} f(x)\, d\omega.$$