Analyzing Periodic Data: Statistical Perspective

C.N.R. Rao Lecture

By
Debasis Kundu
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur

March 06, 2013

Outline

1. Introduction
2. Basic Formulation
3. Preliminaries
4. Different Estimation Procedures
5. Higher Dimensional Model

Outline

1. Introduction
2. Basic Formulation
3. Preliminaries
4. Different Estimation Procedures
5. Higher Dimensional Model

Introduction

We observe periodic phenomena everyday in our lives. For example the number of tourists visiting the famous Taj Mahal, the daily temperature of Delhi or the ECG data of a normal human being clearly follow periodic pattern. Sometimes the data may not be exactly periodic but it is nearly periodic.
Our aim is to analyze such periodic/ nearly periodic data.

Question?

1. What is a periodic data?
2. Why do we care to analyze?

What is a periodic data?

We do not give the formal definition. But informally speaking

What is a periodic data?

We do not give the formal definition. But informally speaking

1. it shows a repeated (periodic) pattern in one dimension.

What is a periodic data?

We do not give the formal definition. But informally speaking

1. it shows a repeated (periodic) pattern in one dimension.
2. it shows a symmetric (periodic) pattern in higher dimension.

Why do we want to analyze?

Why do we want to analyze?

1. Theoretical reason.

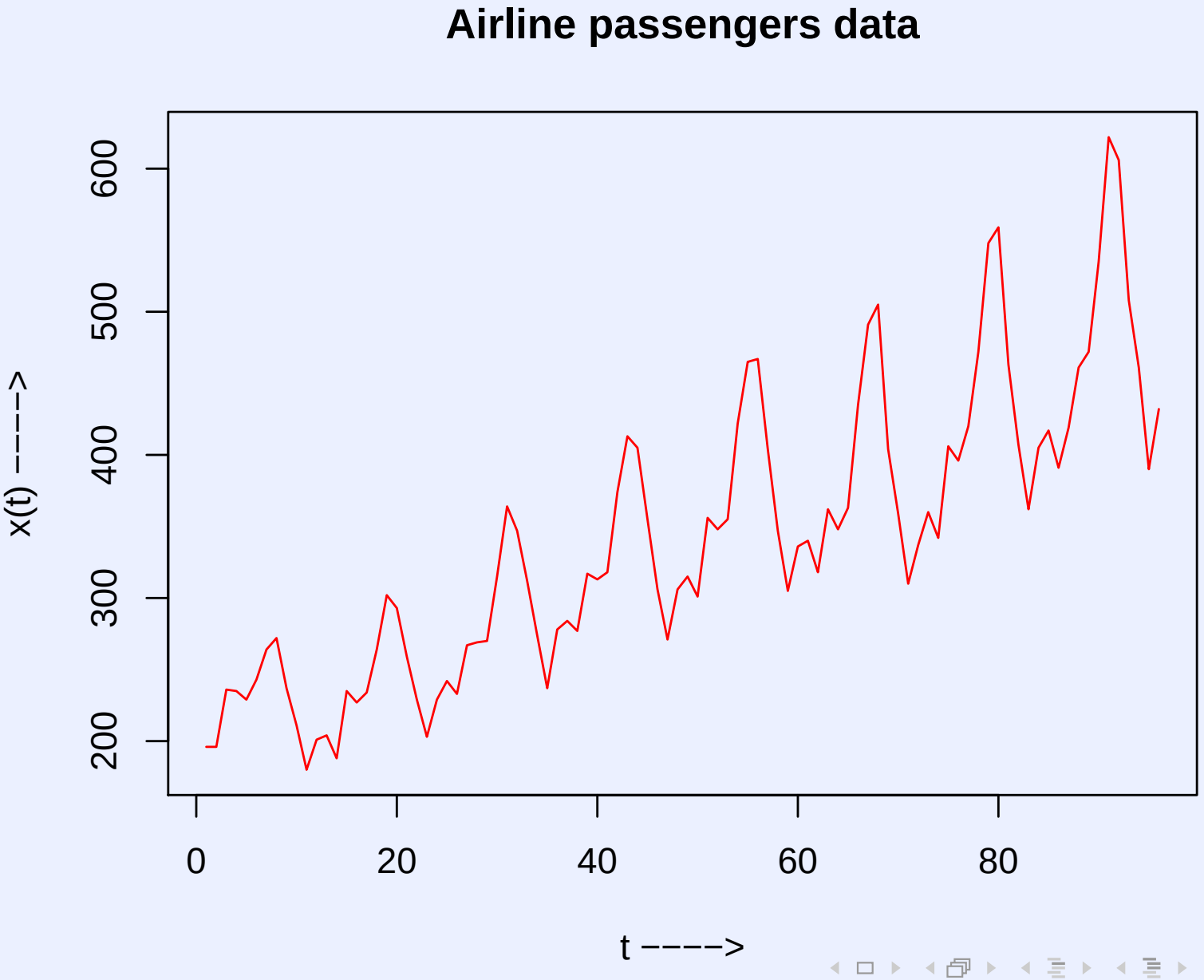
Why do we want to analyze?

1. Theoretical reason.
2. Prediction purposes.

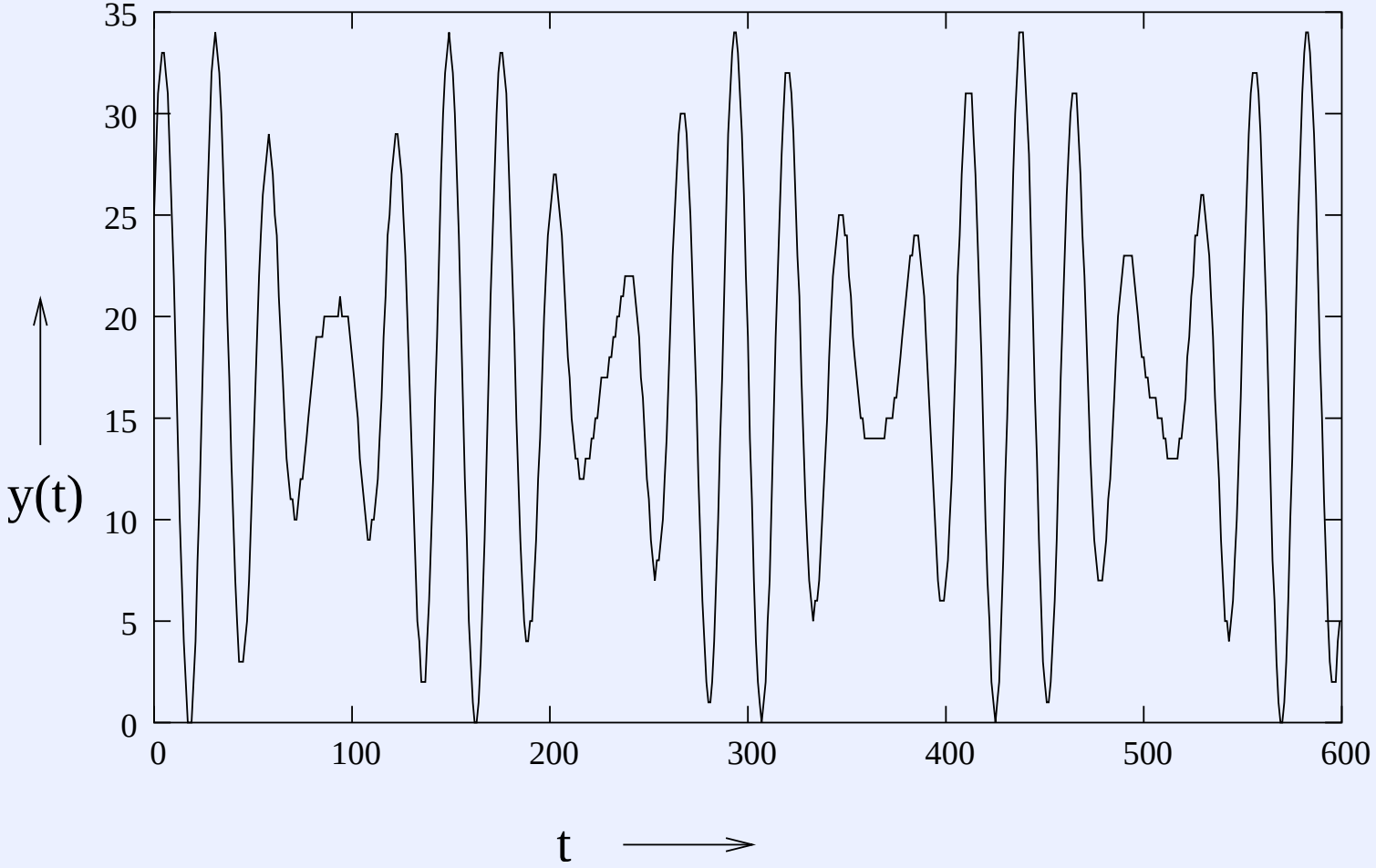
Why do we want to analyze?

1. Theoretical reason.
2. Prediction purposes.
3. Compression purposes.

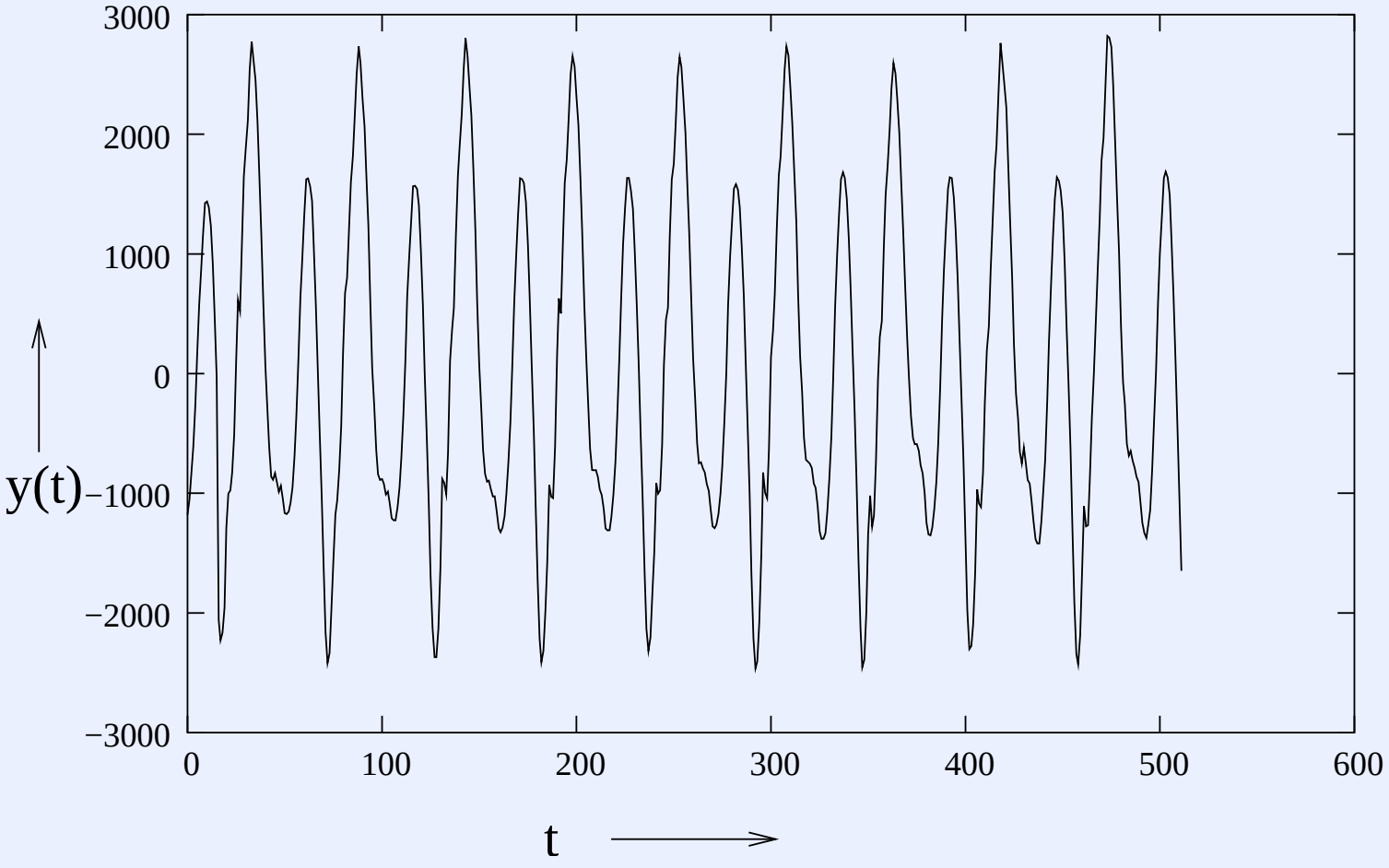
Example: Airlines Passenger Data



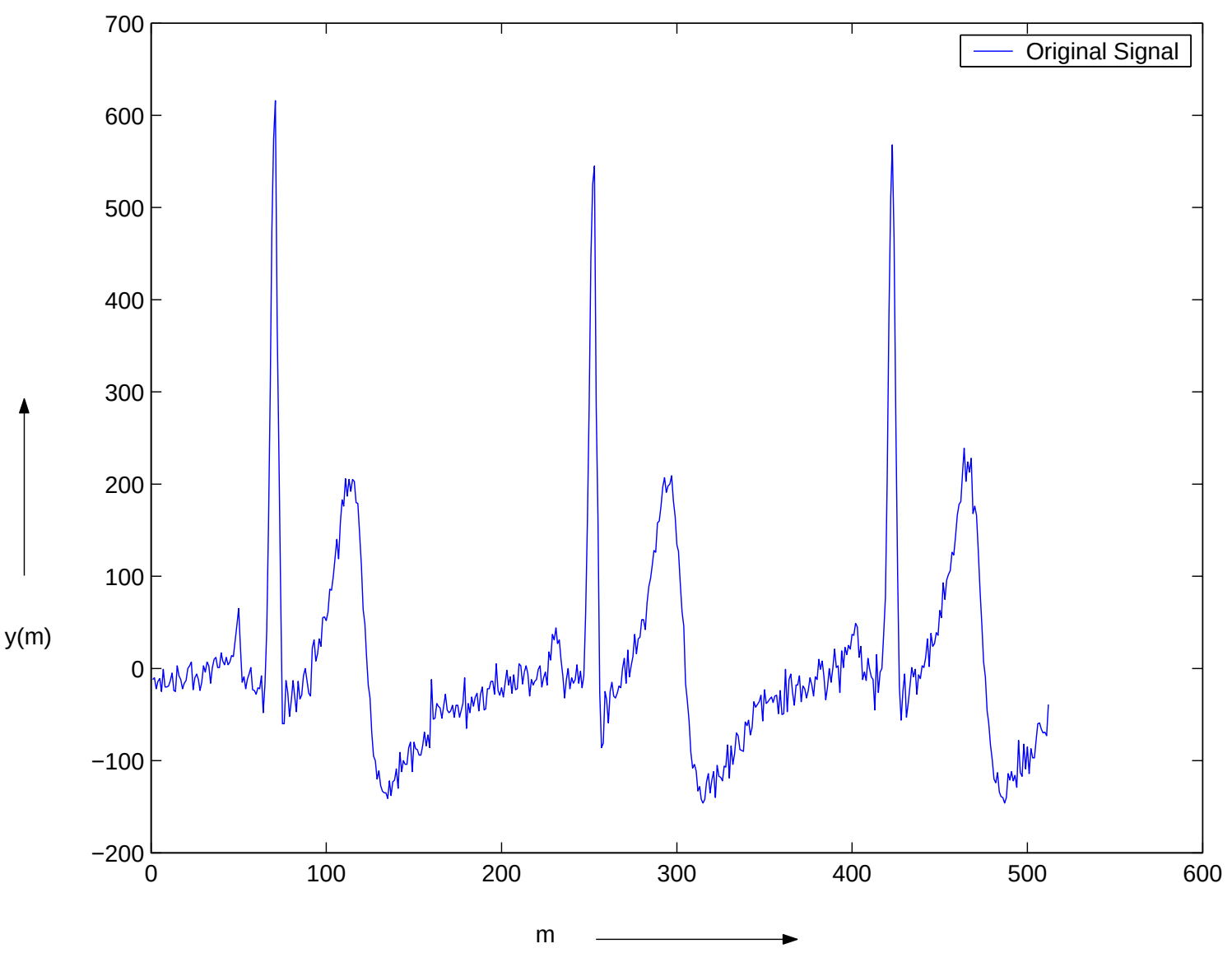
Example: Brightness of Variable Star Data



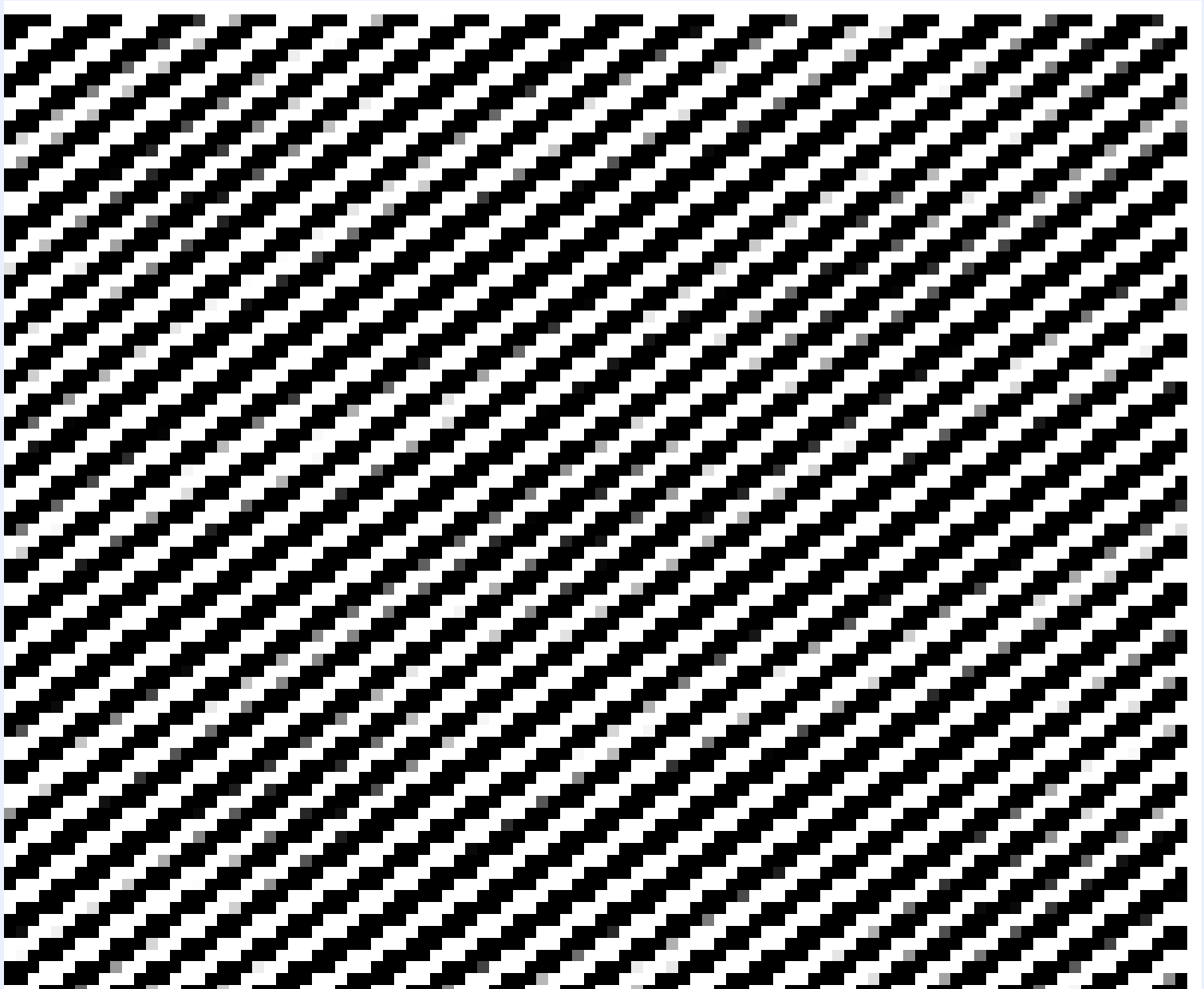
Example: Vowel Sound Data 'uuu'



Example: ECG Data of a Normal Human

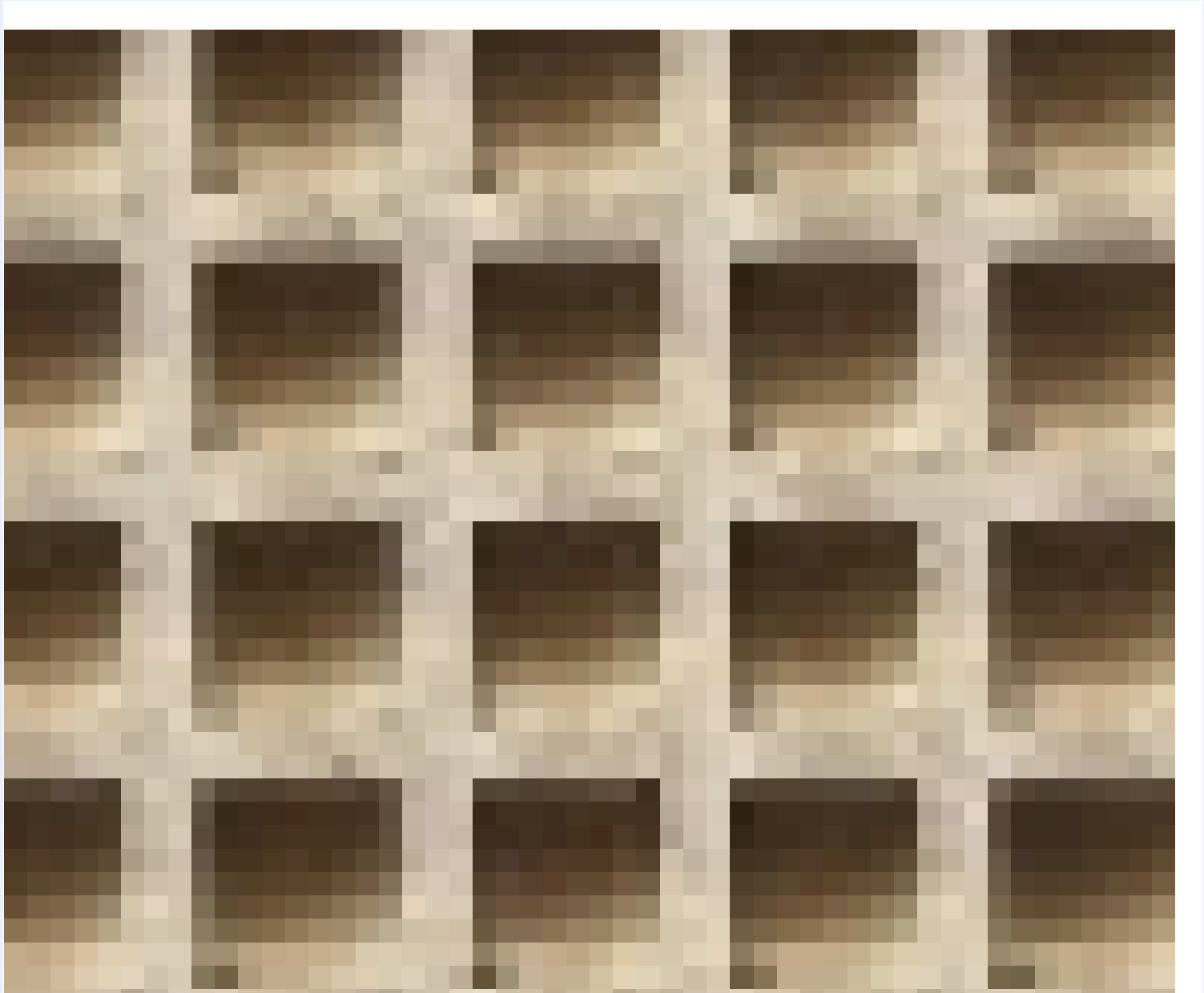


Example: Two Dimension Periodic Data



Example: Three Dimension Periodic Data

Example: Three Dimension Periodic Data



Outline

1 Introduction

2 Basic Formulation

3 Preliminaries

4 Different Estimation Procedures

5 Higher Dimensional Model

Simplest Periodic Function

The simplest periodic function is the sinusoidal function, and it can be written in the following form:

$$y(t) = A\cos(\omega t) + B\sin(\omega t)$$

The period of $y(t)$ is the shortest time taken for $y(t)$ to repeat itself, and it is $2\pi/\omega$.

Smooth Periodic Function

In general a smooth periodic function (mean adjusted) with period $2\pi/\omega$, can be written in the form:

$$y(t) = \sum_{k=1}^{\infty} \left[A_k \cos(\omega k t) + B_k \sin(\omega k t) \right],$$

and it is well known as the Fourier expansion of $y(t)$.

Extracting Parameters

From $y(t)$, A_k and B_k can be obtained uniquely.

$$\int_0^{2\pi/\omega} \cos(j\omega t) y(t) dt = \begin{cases} \frac{\pi A_j}{\omega} & \text{if} \quad j \geq 1 \\ \frac{2\pi A_0}{\omega} & \text{if} \quad j = 0 \end{cases}$$

and

$$\int_0^{2\pi/\omega} \sin(j\omega t) y(t) dt = \frac{\pi B_j}{\omega}.$$

Noisy Periodic Function

Most of the times $y(t)$ is corrupted with noise, so we observe the following:

$$y(t) = \sum_{k=1}^{\infty} [A_k \cos(\omega t) + B_k \sin(\omega t)] + X(t),$$

where $X(t)$ is the noise component.

Practical Model

It is impossible to estimate infinite number of parameters. Hence the model is approximated by the following model:

$$y(t) = \sum_{k=1}^{p} [A_k \cos(\omega_k t) + B_k \sin(\omega_k t)] + X(t),$$

for some $p < \infty$.

Model

The model has two components,

Model

The model has two components,

1. Deterministic component

Model

The model has two components,

1. Deterministic component
2. Random component

Aim

The aim is to extract (estimate) the deterministic component $\mu(t)$, where

$$\mu(t) = \sum_{k=1}^{p} \left[A_k \cos(\omega_k t) + B_k \sin(\omega_k t) \right],$$

in presence of the random error component $X(t)$, based on the available data $y(t), t = 1, \ldots, N$.

Problem Formulation

Based on the available data $\{y(t); t = 1, \ldots, N\}$,

Problem Formulation

Based on the available data $\{y(t); t = 1, \ldots, N\}$,

1. Deterministic Component
   - Determine (estimate) p
   - Determine (estimate) $A_1, \ldots, A_p$, $B_1, \ldots, B_p$
   - Determine (estimate) $\omega_1, \ldots, \omega_p$.

Problem Formulation

Based on the available data $\{y(t); t = 1, \ldots, N\}$,

1. Deterministic Component
   - Determine (estimate) p
   - Determine (estimate) $A_1, \ldots, A_p$, $B_1, \ldots, B_p$
   - Determine (estimate) $\omega_1, \ldots, \omega_p$.
2. Random Component
   - Estimate $X(t)$

Procedure

Procedure

1. Assume certain structure on $X(t)$

Procedure

1. Assume certain structure on $X(t)$
2. Estimate the deterministic component $\mu(t)$

Procedure

1. Assume certain structure on $X(t)$
2. Estimate the deterministic component $\mu(t)$
3. Estimate the error $X(t)$

Procedure

1. Assume certain structure on $X(t)$
2. Estimate the deterministic component $\mu(t)$
3. Estimate the error $X(t)$
4. Verify the assumption.

Procedure

1. Assume certain structure on $X(t)$
2. Estimate the deterministic component $\mu(t)$
3. Estimate the error $X(t)$
4. Verify the assumption.
5. If the assumption is satisfied then stop the process, otherwise go back to step 1.

Outline

1 Introduction

2 Basic Formulation

3 Preliminaries

4 Different Estimation Procedures

5 Higher Dimensional Model

Random Variable

X is called a random variable if it takes certain values with a given probability. It can be written as follows:

$$P(a < X < b) = \int_a^b f(x)dx.$$

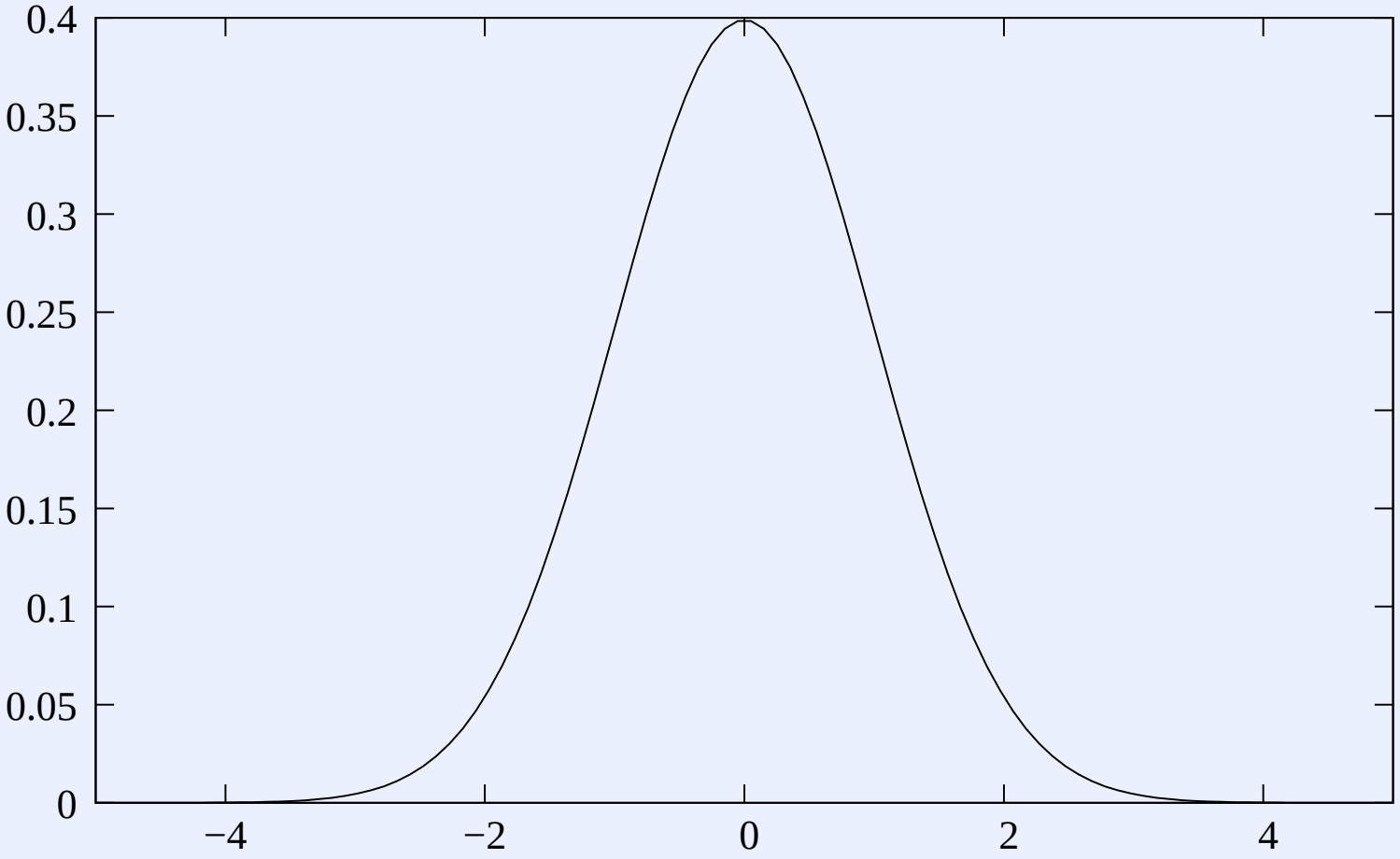
The function $f(x)$ is known as the probability density function of X

Gaussian Random Variable

X is called a Gaussian random variable with mean 0 and variance 1, if

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}; \qquad -\infty < x < \infty$$

Gaussian PDF



Linear Equations

Suppose we want to solve the following linear equation:

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

Suppose $\mathbf{A}$ is an $m \times n$ (for $m > n$) matrix, $\mathbf{x}$ is a $n \times 1$ and $\mathbf{b}$ is a $m \times 1$ vector. The least squares solution of $\mathbf{x}$ is

$$\widehat{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}.$$

Non-Linear Equation: Newton-Raphson

We want to solve the following linear equation:

$$f(x) = 0,$$

Suppose $\widehat{x}$ is a solution of the equation $f(\widehat{x}) = 0$.

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}$$

Outline

1. Introduction
2. Basic Formulation
3. Preliminaries
4. Different Estimation Procedures
5. Higher Dimensional Model

Periodogram Estimators

The most used and popular estimation procedure is the periodogram estimators. The periodogram at a particular frequency is defined as

$$I(\theta) = \left(\frac{1}{N}\sum_{t=1}^{N} y(t)\cos(\theta t)\right)^2 + \left(\frac{1}{N}\sum_{t=1}^{N} y(t)\sin(\theta t)\right)^2$$

$$\approx \left(\frac{1}{N}\sum_{t=1}^{N} \mu(t)\cos(\theta t)\right)^2 + \left(\frac{1}{N}\sum_{t=1}^{N} \mu(t)\sin(\theta t)\right)^2$$

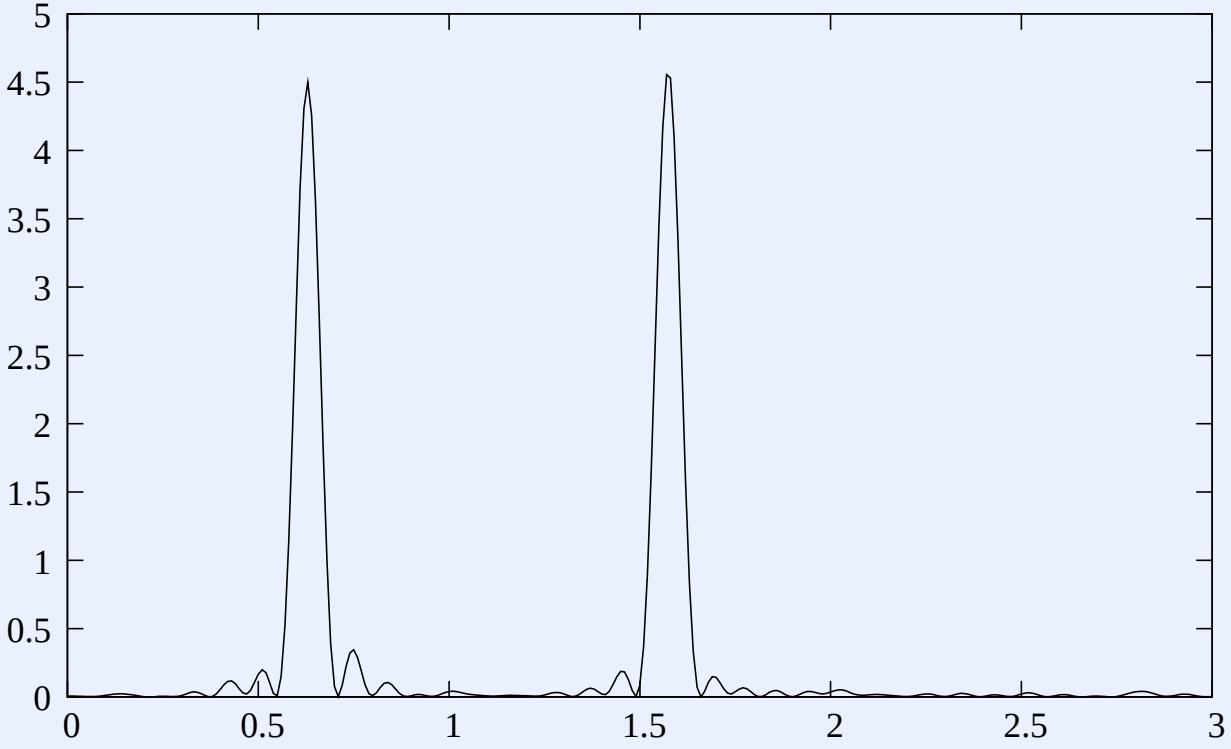
Periodogram Estimator

Consider the following sinusoidal signal: Sinusoidal Example 1:

$$y(t) = 3.0(\cos(0.2\pi t)+\sin(0.2\pi t))+3.0(\cos(0.5\pi t)+\sin(0.5\pi t))+X(t)$$

Here $X(t)$'s are i.i.d. N(0,0.5)

Examples: Sinusoidal Signal



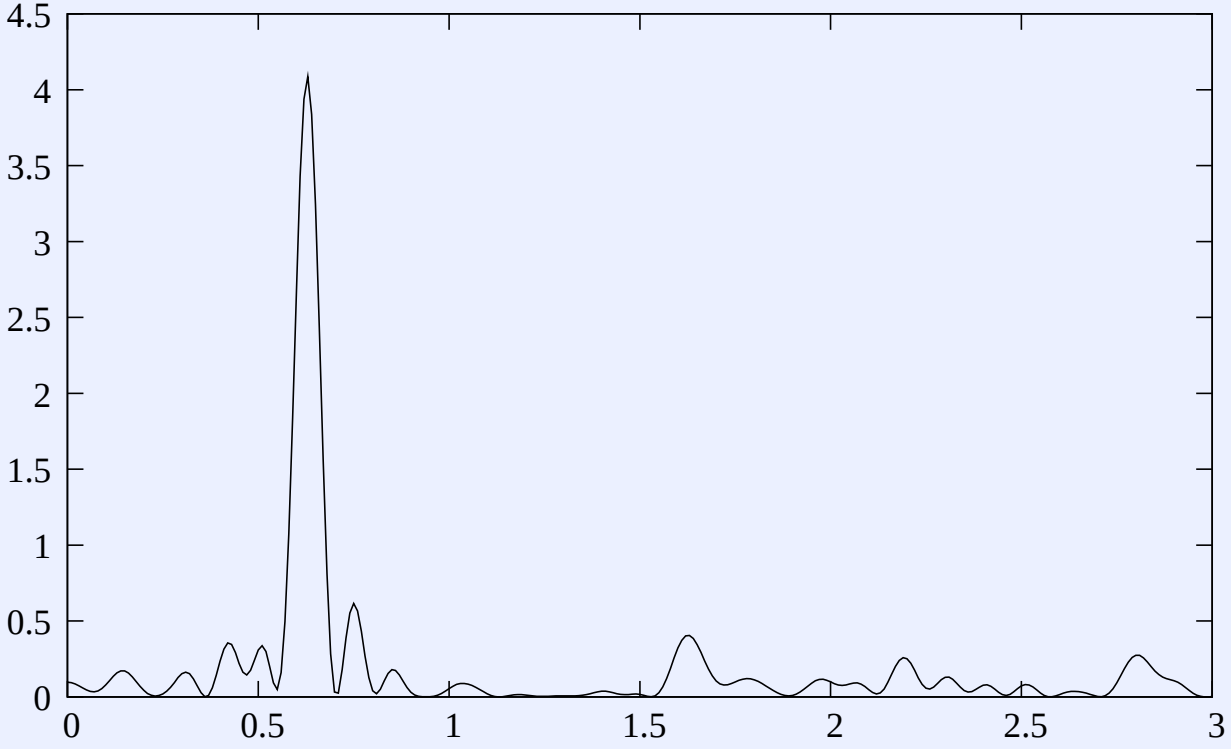
Periodogram Estimator

Consider the following sinusoidal signal: Sinusoidal Example 2:

$$y(t) = 3.0(\cos(0.2\pi t)+\sin(0.2\pi t))+0.25(\cos(0.5\pi t)+\sin(0.5\pi t))+X(t)$$

Here $X(t)$'s are i.i.d. N(0,2.0)

Examples: Sinusoidal Signal



Least Squares Estimators

Assuming p is known, the most natural estimators will be the least squares estimators and they can be obtained as follows:

$$\sum_{t=1}^{n}\left(y(t)-\left[\sum_{k=1}^{p}A_k\cos(\omega_k t)+B_k\sin(\omega_k t)\right]\right)^2$$

Numerical Issues

1. It is a highly non-linear problem. The least squares surface has several local minima.
2. Most of the time the standard Newton-Raphson algorithm may not converge.
3. Even if they converge, often it converges to the local minimum rather than the global minimum.
4. If p is large, it becomes a higher dimensional optimization problem, extremely accurate initial guesses are required for any iterative procedure to work well.

Least Absolute Deviation Estimators

Assuming p is known, the LAD estimators can be obtained by minimizing:

$$\sum_{t=1}^{n} \left| y(t) - \left[\sum_{k=1}^{p} A_k \cos(\omega_k t) + B_k \sin(\omega_k t) \right] \right|$$

Sequential Estimation Procedures

It is based on the facts that the components are orthogonal and it works like this
First minimize

$$\sum_{t=1}^{n} \left(y(t) - A\cos(\omega t) - B\sin(\omega t)\right)^2$$

with respect to A, B and ω.

Take out their effect from $y(t)$, *i.e.* consider

$$\tilde{y}(t) = y(t) - \widehat{A}\cos(\widehat{\omega} t) - \widehat{B}\sin(\widehat{\omega} t)$$

Repeat the procedure p times.

Advantage

It reduces the computational burden significantly. For example if $p = 25$, instead of solving a 25 dimensional optimization problem, we need to solve 25 one dimensional optimization problems. It does not have any problem about initial guess or convergence.

It produces the same accuracy as the least squares estimators.

Super Efficient Estimators

When $p = 1$, the Newton-Raphson algorithm will be of the following form:

$$\omega^{(j+1)} = \omega^{(j)} - \frac{Q'(\omega)}{Q''(\omega)}$$

It has been suggested

$$\omega^{(j+1)} = \omega^{(j)} - \frac{1}{4}\frac{Q'(\omega)}{Q''(\omega)}$$

It not only converges, it produces estimators which are better than the least squares estimators.

Main Theoretical Results

1. Least squares estimators are consistent under mild assumptions on the errors.
2. Least squares estimators have the convergence rate $N^{-3/2}$.
3. Sequential estimators have the same convergence rate as the least squares estimators.
4. Asymptotic variances of the super efficient estimators are smaller than the least squares estimators.
5. Periodogram estimators are consistent, but it has the convergence rate $N^{-1/2}$.

Estimation of p

It is a difficult problem. We still do not have a satisfactory solution.

Estimation of p

1. Consider the number of peaks of the periodogram function.
2. It can be very misleading.
3. In the least squares procedure, consider residual sums of squares.
4. It can be very misleading too.
5. Information theoretic criterion.

Information Theoretic Criterion

$$AIC(k) = n \ln R_k + 2(3k)$$

$$BIC(k) = n \ln R_k + \frac{1}{2} \ln n(3k)$$

$$EDC(k) = n \ln R_k + C_n k.$$

Here

$$R_k = \sum_{t=1}^{n} \left[y(t) - \sum_{j=1}^{k} \left(\widehat{A}_j \cos(\widehat{\omega}_j t) + \widehat{B}_j \sin(\widehat{\omega}_j t) \right) \right]^2$$

Choose that model for which $AIC(k)$, $BIC(k)$ or $EDC(k)$ is minimum

Outline

1. Introduction
2. Basic Formulation
3. Preliminaries
4. Different Estimation Procedures
5. Higher Dimensional Model

Two-Dimensional Model

$$y(m,n) = \sum_{k=1}^{p} [A_k \cos(\theta_k m + \omega_k n) + B_k \sin(\theta_k m + \omega_k n)] + X(m,n),$$

for some $p < \infty$.

Three-Dimensional Model

$$y(m,n,s) = \sum_{k=1}^{p} \mu_k(m,n,s) + X(m,n,s),$$

for some $p < \infty$, where

$$\mu_k(m,n,s) = A_k \cos(\theta_k m + \omega_k n + \lambda_k s) + B_k \sin(\theta_k m + \omega_k n + \lambda_k s).$$

Thank You