V3: Circadian rhythms, time-series analysis (contd‘)

Effects of insufficient sleep on circadian rhythmicity and expression amplitude of the human blood transcriptome

Introduction: 5 paragraphs

(1) Insufficient sleep - Biological/medical relevance

(2) Previous work on effects of insufficient sleep in rodents (dt. *Nagetiere*)

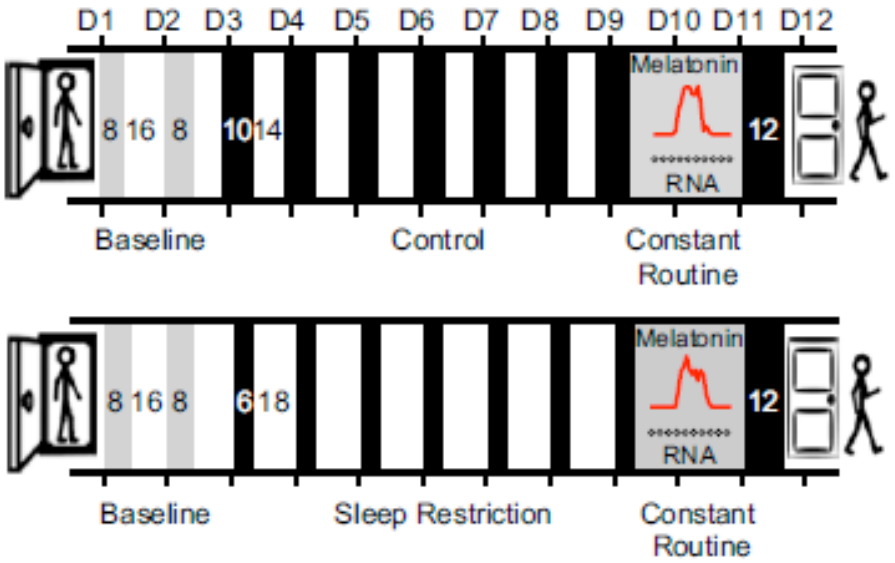
(3) Metabolic effects of 2-week sleep loss.

(4) Lack of understanding of sleep loss in humans. Problem: tissue not accessable, except for blood.

(5) Relationship of sleep loss and circadian rhythms.

Problem with Intro: paragraph 5 seems somehow misplaced.

Cross-over design study



26 participants were first put (top) into sleep-restricted conditions with 6 hours of sleep opportunity per night

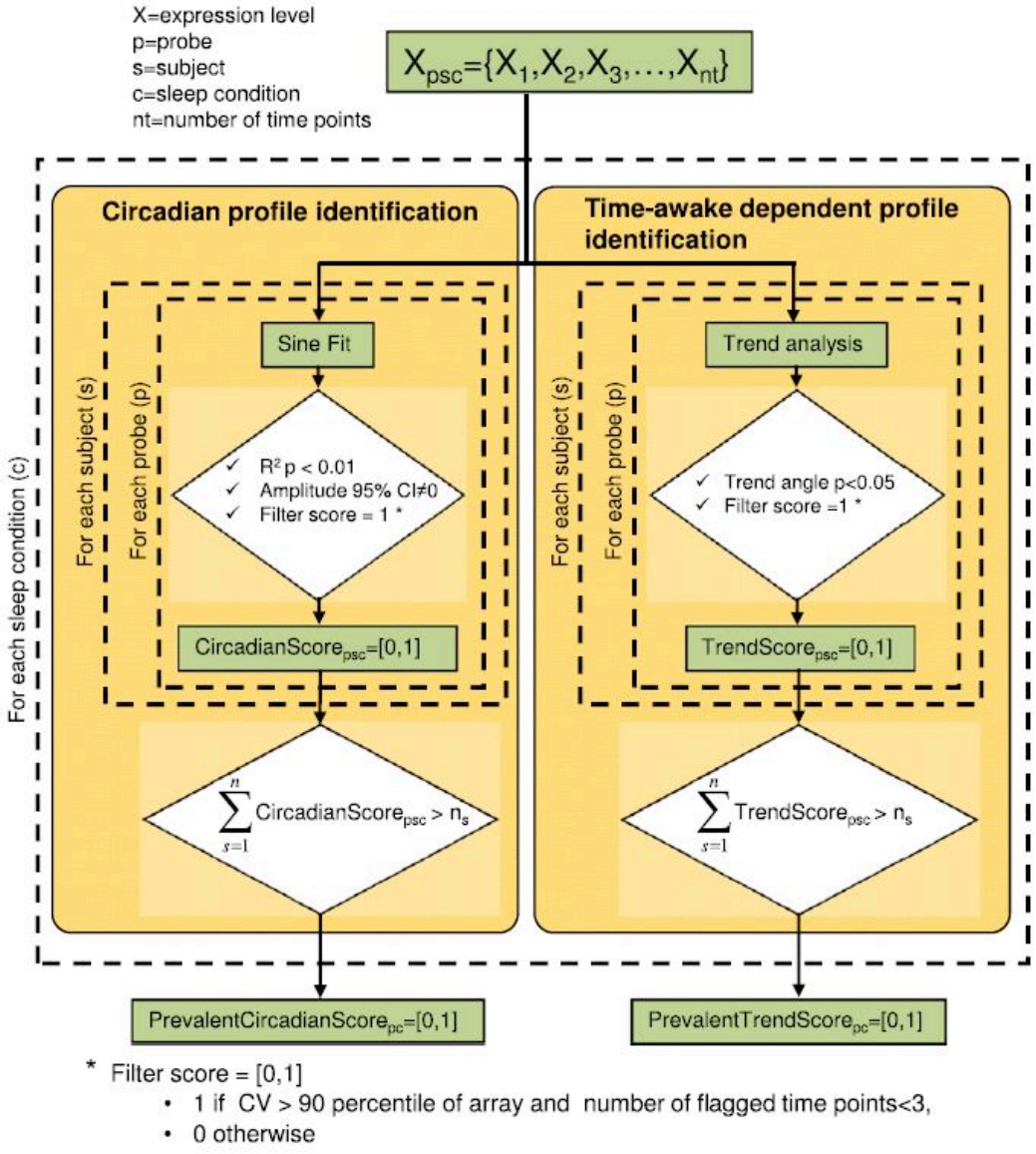
and then into conditions of sufficient sleep with 10 hours of sleep opportunity.

-> effects of genetic pre-disposition are mimimized by using „matched samples“

My problem with study design: no data are collected during phases of „control“ vs. „sleep restriction“, only the behavior during „constant routine“ (week illumination, no sleep)

PNAS 110, E1132 (2013)

Analysis during „constant routine“



Task: identify genes affected by sleep-conditions

2 strategies

Left: identify circadian genes (similar to V2)

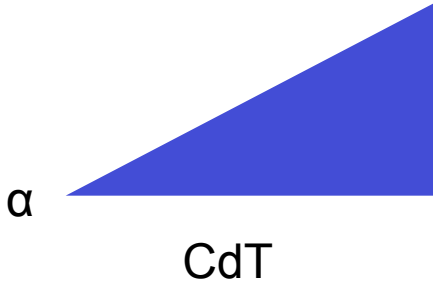
Right: identify time-awake-dependent transcripts

PNAS 110, E1132 (2013)

Trend analysis during „constant routine“

Is there an upward **trend** in gene expression? -> Cumulative upward trend (CuT)

Is there a downward **trend** in gene expression? -> Cumulative downward trend (CdT)



$\tan \alpha = CuT / CdT$

$\arctan(\tan \alpha) = \alpha = \arctan(CuT / CdT)$

Compare arctan (CuT / CdT) for real time series to that of randomly resampled (shuffled) data.

-> p-value

PNAS 110, E1132 (2013)

Standard linear model

Suppose that you observe *n* **data points** $y_1, y_2, \ldots y_n$, and that you want to explain them by using *n* values for each of *p* **explanatory variables** $x_{11}, \ldots, x_{1p}, x_{21}, \ldots, x_{2p}, \ldots, x_{n1}, \ldots, x_{np}$.

The x_{ij} values can be either regression-type continuous variables or dummy variables indicating class membership.

The **standard linear model** for this setup is

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \quad i = 1, \ldots, n$$

where $\beta_1, \ldots, \beta_p$ are unknown **fixed-effects** parameters to be estimated and $\varepsilon_1, \ldots, \varepsilon_n$ are unknown independent and identically distributed normal (Gaussian) random variables with mean 0 and variance σ^2.

SAS/STAT9.2 user guide

Standard linear model

These equations

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \quad i = 1, \ldots, n$$

can also be written using vectors and a matrix, as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

For convenience, simplicity and extendability, this entire system is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where **y** denotes the vector of observed y_i's , **X** is the known matrix of x_{ij}'s, **β** is the unkown fixed-effects parameter vector and **ε** is the unobserved vector of independent and identically distributed Gaussian random errors.

SAS/STAT9.2 user guide

Formulation of the Mixed Model

The general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ certainly useful.

However, often the distributional assumption about **ε** (i.e. indepence) is too restrictive.

The **mixed model** extends the general linear model by allowing a more flexible specification of the covariance matrix of **ε**.

Thus, it allows for both correlation and heterogeneous variances among the elements of **ε**, although you still assume normality (Gaussian distribution).

The mixed model is written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$

Everything is the same as in the general linear model except for the addition of the known **design matrix Z** and the vector of unknown *random-effects parameters γ.*

SAS/STAT9.2 user guide

Formulation of the Mixed Model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

The matrix **Z** can contain either continuous or dummy variables, just like **X**.

The name *mixed model* comes from the fact that the model contains both fixed-effects parameters **β** and random-effects parameters ***γ***.

A key assumption is that ***γ*** and **ε** are normally distributed with

$$\mathrm{E}\begin{bmatrix} \gamma \\ \epsilon \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

$$\mathrm{Var}\begin{bmatrix} \gamma \\ \epsilon \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}$$

SAS/STAT9.2 user guide

Formulation of the Mixed Model

The variance of the observed data points **y** is therefore

$$\mathbf{V} = \mathbf{Z}\,\mathbf{G}\,\mathbf{Z}' + \mathbf{R}$$

You can model **V** by setting up the random-effects design matrix **Z** and by specifying covariance structures for **G** and **R**.

Estimating parameters is more difficult in the mixed model than in the general linear model. Not only do you have **β** as in the general linear model, you have unknown parameters in ***γ*** , **G** and **R** as well.

Least squares is not longer the best method for parameter estimation.

Generalized least squares (GLS) is more appropriate, minimizing

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

SAS/STAT9.2 user guide

ANOVA = Analysis of variance

ANOVA is a collection of statistical models that analyze the differences between group means and the variation between and among groups.

ANOVA somehow generalizes the t-test to more than two groups.

wikipedia.org

Results … more in V4

(1) Main effect of sleep condition („sleep restricted“ vs „control“)

On 711 genes. 444 were down-regulated, 267 were upregulated.

(2) Circadian rhythm

Given sufficient sleep, **1855** (8.8%) genes are classified as circadian.

After sleep restriction, this number declined to **1481** (6.9%).

(3) Response of gene expression to acute sleep loss

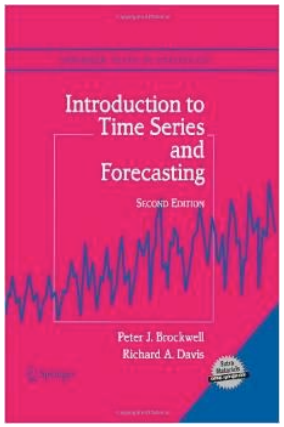
Given sufficient sleep, **122** genes were classified as „time-awake genes“.
After sleep restriction, this number increased to **856** genes.

In both cases, more genes have downward trends than upward trends.

wikipedia.org

Introduction to time series analysis (2)

After book of Peter Brockwell &
Richard A. Davis



End of lecture V2 …

Instead of **complete probabilistic time series models,** we often specify only the **first- and second-order moments** of the joint distributions, i.e., the expected values EX_t and the expected products E($X_{t+h}X_t$), t = 1, 2, . . ., h = 0, 1, 2, . . ., focusing on properties of the sequence {X_t} that depend only on these.

Some zero-mean models: iid noise

Perhaps the simplest model for a time series is one in which there is no trend or seasonal component and in which the observations are simply independent and identically distributed (iid) random variables with zero mean.

We refer to such a sequence of random variables X_1 , X_2 , . . . as **iid noise**.

By definition we can write, for any positive integer n and real numbers x_1 , . . . , x_n,

$$P[X_1 \leq x_1, \ldots, X_n \leq x_n] = P[X_1 \leq x_1] \cdots P[X_n \leq x_n] = F(x_1) \cdots F(x_n),$$

where F (·) is the cumulative distribution function of each of the identically distributed random variables X_1 , X_2 , . . .

(due to the indepence, the joint probability (left) is equal to the product of the individual probabilities.)

(Note that the **distribution function** F of a random variable X is defined by

$$F(x) = P[X \leq x]$$

for all real x.)

iid noise

In the iid model there is no dependence between observations.

In particular, for all $h \geq 1$ and all $x_1, \ldots, x_n$,

$$P[X_{n+h} \leq x | X_1 = x_1, \ldots, X_n = x_n] = P[X_{n+h} \leq x],$$

showing that knowledge of $X_1, X_2, \ldots X_n$, is of no value for predicting the behavior of X_{n+h}.

Given the values of $X_1, X_2, \ldots X_n$, the function f that minimizes the mean squared error $E\left[(X_{n+h} - f(X_1, \ldots, X_n))^2\right]$ is in fact identically zero.

Although this means that iid noise is a rather uninteresting process for forecasters, it plays an important role as a building block for more complicated time series models.

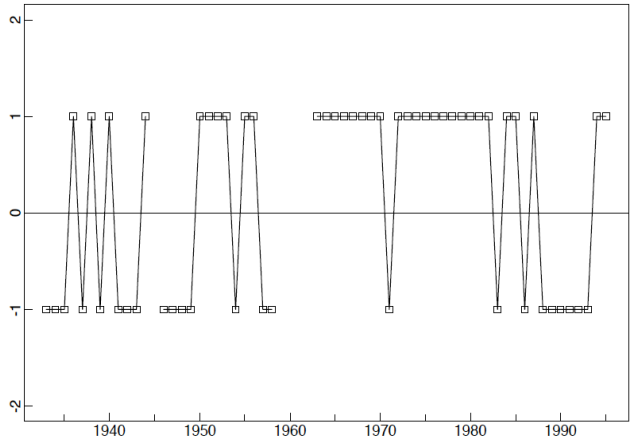
A binary process

As an example of iid noise, consider the sequence of iid random variables $\{X_t, t = 1, 2, \ldots, \}$

with $P[X_t = 1] = p, \quad P[X_t = -1] = 1 - p,$ where $p = 0.5$.

The time series obtained by tossing a penny repeatedly and scoring + 1 for each head and − 1 for each tail is usually modeled as a realization of this process.

A priori we might well consider the same process as a model for the all-star baseball games.



However, even a cursory inspection of the results from 1963–1982, which show the National League winning 19 of 20 games, casts serious doubt on the hypothesis $P[X_t = 1] = \frac{1}{2}$. in this case.

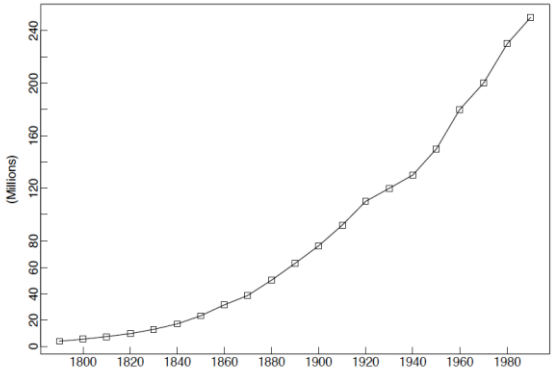
Models with Trend and Seasonality

In several of the previous time series examples there was a clear **trend** in the data.

E.g. an increasing trend was apparent in both the Australian red wine sales and the population of the U.S.A..

In both cases a zero-mean model for the data is clearly inappropriate.

The graph of the population data, which contains no apparent periodic component, suggests trying a model of the form



$$X_t = m_t + Y_t,$$

where m_t is a slowly changing function known as the **trend component** and Y_t has zero mean.

A useful technique for estimating m_t is the method of **least squares.**

Least Squares Method

In the least squares procedure we attempt to fit a parametric family of functions,

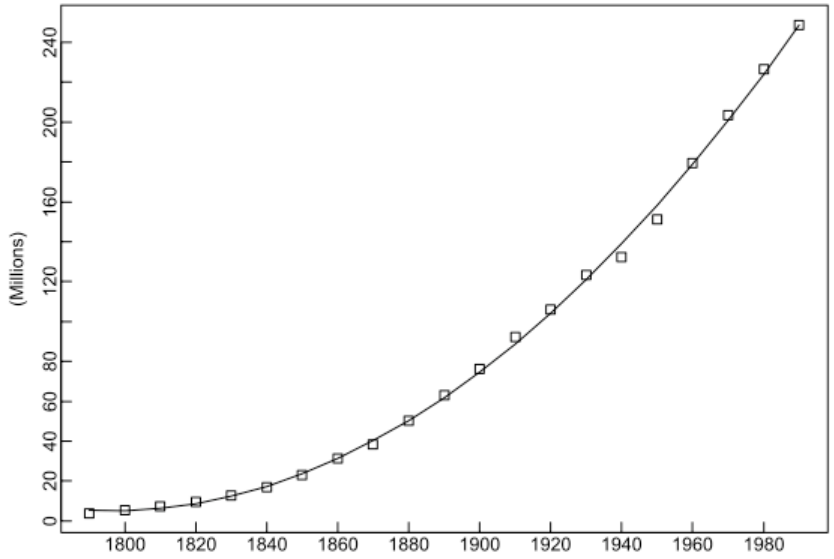
e.g., $m_t = a_0 + a_1 t + a_2 t^2,$ to the data $\{x_1, \ldots, x_n\}$

by choosing the parameters a_0, a_1, and a_2, to minimize

$$\sum_{t=1}^{n} (x_t - m_t)^2.$$

To fit a function of this form to the US population data we relabel the time axis so that $t = 1$ corresponds to 1790 and $t = 21$ corresponds to 1990.

This gives the fitting curve shown on the right.



Harmonic Regression

Many time series are influenced by **seasonally varying factors**, the effect of which can be modeled by a periodic component with fixed known period.

E.g. the accidental deaths series showed a repeating annual pattern with peaks in July and troughs in February, strongly suggesting a seasonal factor with period 12.

In order to represent such a seasonal effect, allowing for noise but assuming no trend, we can use the simple model, $X_t = s_t + Y_t$,
where s_t is a periodic function of t with **period *d*** ($s_{t-d} = s_t$).

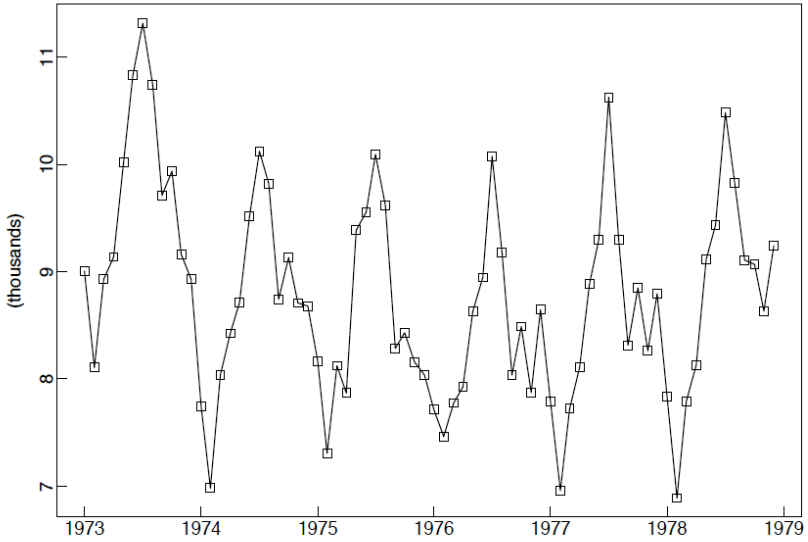
A convenient choice for s_t is a sum of harmonics (or sine waves) given by

$$s_t = a_0 + \sum_{j=1}^{k} (a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t)),$$

where a_0 , a_1 , ..., a_k and b_1 , ... , b_k are unknown parameters and λ_1, ... , λ_k are fixed frequencies, each being some integer multiple of $2\pi/d$.

Harmonic Regression

E.g. the accidental deaths



To fit a sum of two harmonics with periods 12 months and 6 months to the monthly accidental deaths data $x_1, \ldots, x_n$ with $n = 72$,
we choose $k = 2$,
$f_1 = n/12 = 6$, and
$f_2 = n/6 = 12$.

Shown is the fitted curve with optimized parameters.

General Approach to Time Series Modeling

The previous examples illustrate a general approach to time series analysis.

• Plot the series and examine the main features of the graph, checking in particular whether there is

(a) a trend,

(b) a seasonal component,

(c) any apparent sharp changes in behavior,

(d) any outlying observations..

General Approach to Time Series Modeling

• Remove the trend and seasonal components to get stationary residuals.

To achieve this goal it may sometimes be necessary to apply a preliminary transformation to the data.

E.g. if the magnitude of the fluctuations appears to grow roughly linearly with the level of the series, then the transformed series { ln X_1 , . . . , ln X_n } will have fluctuations of more constant magnitude.

Whichever method is used, the aim is to produce a stationary series, whose values we shall refer to as **residuals**.

General Approach to Time Series Modeling

• Choose a model to fit the residuals, making use of various sample statistics including the sample autocorrelation function.

• Forecasting will be achieved by forecasting the residuals and then inverting the transformations described above to arrive at forecasts of the original series $\{X_t\}$.

Stationary Time Series

Loosely speaking, a time series $\{X_t\,,\ t = 0\,,\ \pm\ 1\,,\ .\ .\ .\}$ is said to be **stationary** if it has statistical properties similar to those of the "time-shifted" series $\{X_{t+h}\,,\ t = 0\,,\ \pm\ 1\,,\ .\ .\ .\}$, for each integer h .

Restricting attention to those properties that depend only on the first- and second-order moments of $\{X_t\}$, we can make this idea precise with the following definitions.

Let $\{X_t\}$ be a time series with $E(X_t^2) < \infty$. The **mean function** of $\{X_t\}$ is

$$\mu_X(t) = E(X_t).$$

The **covariance function** of $\{X_t\}$ is

$$\gamma_X(r, s) = \mathrm{Cov}(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

for all integers r and s.

Stationary Time Series

$\{X_t\}$ is **(weakly) stationary** if

(i) $\mu_X(t)$ is independent of t,

and

(ii) $\gamma_X(t+h, t)$ is independent of t for each h.

Let $\{X_t\}$ be a stationary time series. The **autocovariance function** (ACVF) of $\{X_t\}$ at lag h is

$$\gamma_X(h) = \mathrm{Cov}(X_{t+h}, X_t).$$

The **autocorrelation function** (ACF) of $\{X_t\}$ at lag h is

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = \mathrm{Cor}(X_{t+h}, X_t).$$

Discuss as example: autocorrelation of a water dipole moment

As preparation of assignment 1: discuss fitting of sine-waves to data on blackboard

For V4

- In V4, we will focus on the results part of the sleep loss paper
- How is **enrichment** of gene ontology terms computed?
- What is the role of taking the **melatonin profiles**?