



Swiss Federal Institute of Technology Zurich

Seminar for  
Statistics

Department of Mathematics

---

Master Thesis

Summer 2023

---

Student Muster

**Time Series Analysis  
for Irregularly Sampled Data**

---

Submission Date: 13 March 2023

---

Co-Advisor    Your co-supervisor  
Advisor:     Prof. Dr. Your supervisor



To some special person



# Preface

First words and acknowledgements.



# Abstract

Short summary of my thesis.

## Contents

|  |           |
|--|-----------|
| Notation   | <b>xi</b> |
| <b>1 Introduction</b>  | <b>1</b>  |
| 1.1 Thesis Objective . . . . .   | 1         |
| 1.2 Thesis Outline . . . . .   | 2         |
| <b>2 Characteristics of Time Series</b>  | <b>3</b>  |
| 2.1 Stationarity . . . . .   | 3         |
| 2.2 ARMA Model . . . . .   | 3         |
| 2.3 Characteristics of the Blood Pressure Time Series . . . . .                                | 3         |
| 2.4 Epistemic and Aleatoric Uncertainty . . . . .  | 3         |
| <b>3 Time Series Decomposition and Linear Regression</b>                                       | <b>5</b>  |
| 3.1 Introduction . . . . .   | 5         |
| 3.2 Regression with Correlated Errors . . . . .  | 6         |
| 3.2.1 Maximum-Likelihood Estimation . . . . .  | 6         |
| 3.2.2 Sandwich Estimation . . . . .  | 7         |
| 3.3 Conclusion . . . . .   | 7         |
| 3.3.1 Extension to Irregularly Spaced Time Series . . . . .                                    | 7         |
| 3.3.2 Prediction and Epistemic Uncertainty . . . . .   | 7         |
| <b>4 First Chapter</b>   | <b>9</b>  |
| 4.1 To include a picture . . . . .   | 9         |
| 4.2 To make a proof . . . . .  | 10        |
| 4.3 To include R code . . . . .  | 10        |
| 4.4 Other information . . . . .  | 10        |
| <b>5 Summary</b>   | <b>11</b> |
| 5.1 Future Work . . . . .  | 11        |
| <b>Bibliography</b>  | <b>13</b> |
| <b>A Complementary information</b>   | <b>15</b> |
| A.1 Including R code with verbatim . . . . .   | 15        |
| A.2 Including R code with the <i>listings</i> package . . . . .                                | 16        |
| A.3 Using <i>Sweave</i> (or <i>knitr</i> ) to include R code (and more) in your report . . . . | 17        |
| <b>B Yet another appendix....</b>  | <b>19</b> |
| B.1 Description . . . . .  | 19        |
| B.2 Tables . . . . .   | 19        |
| <b>C 2nd Appendix: More sophisticated R code listing</b>                                       | <b>21</b> |
| C.1 Chapter 5 . . . . .  | 21        |
| <b>Epilogue</b>  | <b>23</b> |



**List of Figures**

- 4.1 Geyser data: binned histogram, Silverman's and another kernel . . . . . 9
- 4.2 Geyser data: binned histogram, Silverman's and another kernel . . . . . 9

**List of Tables**

|                            |    |
|----------------------------|----|
| B.1 Test results . . . . . | 19 |
|----------------------------|----|

# Notation

BP: blood pressure

CI: Confidence Intervals

OLS: Ordinary Least Squares

Prediction: -

Forecasting: -

Filtering: -

Smoothing: -

$\mathcal{N}(\mu, \sigma^2)$  : Normal distribution with mean  $\mu$  and standard deviation  $\sigma$



# Chapter 1

## Introduction

### 1.1 Thesis Objective

The thesis aims at giving an overview of time series analysis methods for irregularly sampled data.

The standard time series analysis methods usually assume discrete equispaced time and introductory textbooks on time series analysis either completely omit the irregularly spaced case or they only dedicate a very small section to continuous time models or to state-space models with missing observations ([Brockwell and Davis](#), [Brockwell and Davis](#), [Cryer and Chan](#), [Chatfield](#)).

I will thereafter present the most important concepts and what I have identified to be the basic methods for the analysis of irregularly spaced time series.

The topic is motivated by a "real world" problem from medicine. The problem at hand is the one of extracting time series characteristics from a dataset featuring blood pressure (BP) measurements sampled at irregularly spaced time points. High BP is known to be a risk factor for cardiovascular disease. A person's BP level is generally summarized using the average BP value over available measurements within a given time range. A novel monitoring device already allows to collect BP estimates round the clock. The device is collecting photoplethysmography (PPG) signals and converting them into BP measurements. Typically, the system will yield approximately 1.5 BP measurements per hour, but depending on the quality of the PPG signal and some additional external factors, this sampling frequency can widely vary and the expected range lies roughly between 0 and 5 measurements per hour. Having good estimates of the true BP values at any, potentially not observed, time would allow for a better estimation of the person's cardiovascular risk, and enable the development of novel valuable metrics. The thesis will focus on a set of time series characteristics, which have been considered most relevant for estimating the person's cardiovascular risk. The characteristics of interest are.

- the mean function of the BP time series
- the one-week mean BP value
- any "long-term" trends
- characteristics of the circadian cycle, such as the mean day and night BP

Besides the point estimates also their CI are of interest. Importantly, the CI should be able to capture the uncertainty due to the lack of data in the proximity of the point of prediction. This means the width of the CI intervals around the mean function will not be constant over time but depend on how much data is available in the proximity of a given time point. The described endpoints are all based on prediction at the not observed passed time points however not on forecasting at new time points in the future. Hence, the thesis will only focus on the task of reconstructing BP values between the first and last time point in the dataset.

This "real world" problem will serve as a running example throughout the Thesis. Although the topic is motivated by a real dataset we will restrict ourselves to simulated data, which will mimic the most important characteristics of BP time series data.

## 1.2 Thesis Outline

This happens here and that happens there

## Chapter 2

# Characteristics of Time Series

A **time series**  $(x_t : t \in T_0)$  is a collection of observations  $x_t$ , each one being recorded at a specific time  $t$ .  $T_0$  is the set of times at which observations are made. In case of discrete time series  $T_0$  is a discrete set, e.g. for the equispaced case  $T_0 = \{1, 2, \dots, T\}$  and for the unequally spaced case  $T_0 = \{t_1, t_2, \dots, t_n\}$  with  $t_1 < t_2 < \dots < t_n$ . For continuous time series  $T_0$  is an interval, e.g.  $T_0 = (0, T]$ .

A **time series model** for the observed data  $(x_t : t \in T_0)$  is specified by the collection of random variables  $(X_t : t \in T_0)$  of which  $(x_t : t \in T_0)$  is thought to be a realization. Alternatively the time series model can also be considered a random function  $f : T_0 \rightarrow \mathbb{R}$ .

Throughout the thesis the term time series is used both refer to the data and the process from which it is generated.

[Brockwell and Davis](#)

TODO Notation should be adapted/extended to unequally spaced case.

**mean function**

**autocovariance function**

### 2.1 Stationarity

Stationarity is needed for being able to statistically learn from time series data.

### 2.2 ARMA Model

Autoregressive Process Moving Average Process

### 2.3 Characteristics of the Blood Pressure Time Series

circadian cycle

### 2.4 Epistemic and Aleatoric Uncertainty





## Chapter 3

# Time Series Decomposition and Linear Regression

### 3.1 Introduction

As most time series, the mean function of the BP time series is not constant in time and hence it is not stationary. One can try to decompose the time series  $Y(t)$  into a deterministic component, the mean function  $\mu(t)$  and a zero mean stationary process  $E(t)$ :

$$Y(t) = \mu(t) + E(t)$$

This decomposition allows to extract a stationary residual  $E(t)$ , for which we can find a probabilistic model using the theory of such stationary time series processes. The idea is to then use this model in combination with  $\mu(t)$  to obtain a probability distribution of  $Y^*$  at some time  $t^*$ .

The task of time series decomposition is hence to estimate  $\mu(t)$ , which might be an arbitrary function of  $t$ , from the data. Given our knowledge of the BP time series we will start with a simple linear model that features a linear trend and a seasonal component with known frequency  $f$ . If the seasonal component is represented by a cosine  $\alpha \cos(2\pi ft - \phi)$  with phase shift  $\phi$  we obtain the following liner model for the BP time series  $Y(t)$

$$Y(t) = \beta_0 + \beta_1 t + \beta_2 \cos(2\pi ft) + \beta_3 \sin(2\pi ft) + E(t),$$

where based on the trigonometric angle sum identity we know:  $\beta_2 = \alpha \cos(\phi)$  and  $\beta_3 = \alpha \sin(\phi)$

Based on the trigonometric angle sum identities we know that  $\beta_2 = \alpha \cos(\phi)$  and  $\beta_3 = \alpha \sin(\phi)$

If we assume to observe the BP time series at potentially unequally spaced time points  $t_1, t_2 \dots t_n$  and  $t_1 < t_2 < \dots t_n$ , we can write in matrix notation:

$$\mathbf{Y} = X\beta + \mathbf{E}$$

Where  $\mathbf{Y} = [Y_{t_1}, \dots Y_{t_n}]^\top$  is the observed time series,  $X = [x_{t_1}, \dots x_{t_n}]^\top \in \mathbb{R}^{n \times 4}$  is the design matrix with  $i$ -th row  $x_{t_i} = [1, t_i, \cos(2\pi ft_i), \sin(2\pi ft_i)]^\top$  and  $\mathbf{E} = [E_{t_1}, \dots E_{t_n}]^\top$  the zero-mean stationary time series, which we will call errors.

We can use ordinary least squares to find unbiased and asymptotically normal estimates  $\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top Y$  for the regression coefficients  $\beta$ , without the requirement of regularly spaced data points or uncorrelated errors  $E_{t_1}, \dots, E_{t_n}$  (White). In the case of uncorrelated errors with constant variance  $\sigma^2$  we have  $Var(\mathbf{E}) = \sigma^2 I_n$  and an unbiased and consistent estimator for  $\Psi = Var(\hat{\beta}_{OLS})$  is given by:

$$\hat{\Psi} = \hat{\sigma}^2 (X^\top X)^{-1}$$

where  $\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n (y_{t_i} - x_{t_i}^\top \hat{\beta}_{OLS})^2$  and  $p = 4$  in our example

Since  $\mathbf{E}$  is a time series, the assumption of uncorrelated errors is usually violated and the covariance matrix  $\hat{\Psi}$  is thus no longer unbiased (Brockwell and Davis).

## 3.2 Regression with Correlated Errors

The argument presented in this section is based on the textbook of Brockwell and Davis.

If the covariance matrix of the errors  $Var(\mathbf{E}) = \Sigma$  is known, we can use generalized least squares to obtain a unbiased, consistent and efficient coefficient estimate:

$$\hat{\beta}_{GLS} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} Y$$

with unbiased and consistent covariance matrix estimate:

$$Var(\hat{\beta}_{GLS}) = (X^\top \Sigma^{-1} X)^{-1}$$

If  $\Sigma$  is unknown one can exploit the knowledge we have about the stationary time series process  $\mathbf{E}$  to estimate it. In the following subsections will present two approaches to estimate  $\Sigma$ ,  $\hat{\beta}$  and its covariance matrix. Both methods assume an ARMA(p,q) process for  $\mathbf{E}$  and equispaced time points, hence  $\mathbf{E} = (E_t : t \in \{1, 2, \dots, n\})$  and:

$$\Phi(B)E_t = \Theta(B)W_t, \text{ where } W_t \sim WN(0, \sigma_w^2)$$

### 3.2.1 Maximum-Likelihood Estimation

If we additionally assume  $W_t \sim N(0, \sigma_w^2)$ , we can simultaneously estimate the regression coefficients and  $\Sigma$  by maximizing the Gaussian likelihood:

$$L(\beta, \phi, \theta, \sigma_w^2) = (2\pi)^{-\frac{n}{2}} (\det(\Sigma_n))^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{Y} - X\beta)^\top \Sigma_n^{-1} (\mathbf{Y} - X\beta)\right)$$

Where the covariance matrix  $\Sigma_n(\theta, \phi, \sigma_w^2)$  is parametrized by the coefficients  $\theta, \phi, \sigma_w^2$ , which define the ARMA process assumed for  $(E_t : t \in \{1, 2, \dots, n\})$ . Assuming an ARMA(2,3) process we can implement this approach in R using the nlme library (Box, Jenkins, and Reinsel) :

```
library(nlme)
cs <- corARMA(from = ~t, p=2, q=3)
fit.gls <- gls(y ~ t + cos(2 * pi * f * t) + sin(2 * pi * f * t), corr=cs)
```

### 3.2.2 Sandwich Estimation

The second approach to fit an OLS regression first and correct the estimated covariance matrix of the regression coefficients  $\Psi$  with a sandwich estimator. In the presence of autocorrelation one usually estimates  $\Phi = \frac{1}{n}X^\top \Sigma X$ , the covariance matrix of the estimating functions  $V_i(\beta) = x_{t_i}(y_{t_i} - x_{t_i}^\top \beta)$ , which can then be used to derive  $\Psi$ :

$$\Psi = \text{Var}(\hat{\beta}_{OLS}) = (X^\top X)^{-1} X^\top \Sigma X (X^\top X)^{-1} = \left(\frac{1}{n}X^\top X\right)^{-1} \frac{1}{n} \Phi \left(\frac{1}{n}X^\top X\right)^{-1} \quad (3.2.2.1)$$

The general form of the estimators for  $\Phi$  is:

$$\hat{\Phi} = \frac{1}{n} \sum_{i,j=1}^n w_{|i-j|} \hat{V}_i \hat{V}_j^\top \quad (3.2.2.2)$$

where  $w = [w_0, \dots, w_{n-1}]^\top$  is a weight vector and  $\hat{V}_i = V_i(\hat{\beta}_{OLS})$ .

Plugging  $\hat{\Phi}$  into the equation 3.2.2.1 one obtains the heteroskedasticity and autocorrelation consistent (HAC) covariance estimate  $\hat{\Psi}_{HAC}$ .

Newey and West, Andrews and others have suggested different approaches for calculating the weights  $w$ . They all yield decreasing weights with increasing lag  $l = |i - j|$ . The R sandwich package implements some of these methods to estimate  $\hat{\Psi}_{HAC}$ . An introduction to the sandwich package and how it can be used for inference is described by Zeileis.

## 3.3 Conclusion

### 3.3.1 Extension to Irregularly Spaced Time Series

Although literature and "ready to use" implementations only seem to exist for the equispaced case, both of the approaches described above could probably be extended to the case of irregularly spaced time series. For the Maximum-Likelihood approach the parametrization of the covariance matrix  $\Sigma_n$  as described in 3.2.1 would need to be adapted, such that the covariance of the errors at different time points depends on the absolute time difference rather than the lag. Similarly for the sandwich estimator, the weights in 3.2.2.2 should depend on the absolute time difference rather than on the lag.

### 3.3.2 Prediction and Epistemic Uncertainty



## Chapter 4

# First Chapter

### 4.1 To include a picture

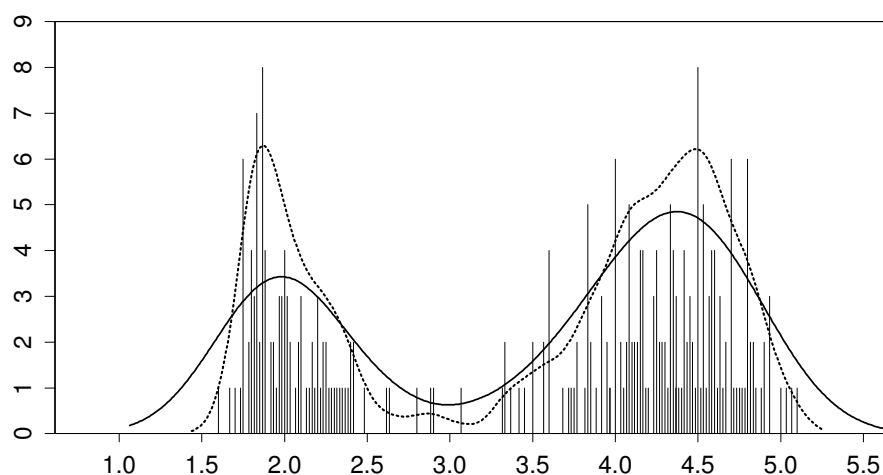


Figure 4.1: Old Faithful Geyser eruption lengths,  $n = 272$ ; binned data and two (Gaussian) kernel density estimates ( $\times 10$ ) with  $h = h^* = .3348$  and  $h = .1$  (dotted).

Or also with `includegraphics`:

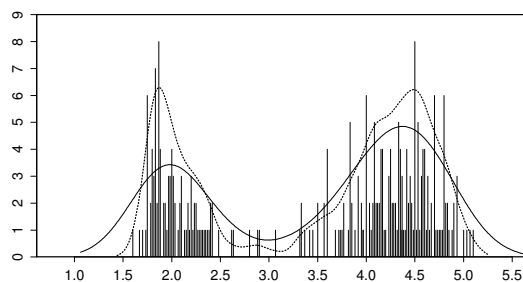


Figure 4.2: Old Faithful Geyser eruption lengths,  $n = 272$ ; binned data and two (Gaussian) kernel density estimates ( $\times 10$ ) with  $h = h^* = .3348$  and  $h = .1$  (dotted).

## 4.2 To make a proof

*Proof.*  $1 + 1 = 2$

□

## 4.3 To include R code

See information in Appendix [A](#).

## 4.4 Other information

Put a text between quotes: make sure to use nice quotes, such as “quote”.

Cite an article or book you refer shortly here, and then listed in the bibliography. Or mention that [Robinson](#) (a person) (two persons) have already done quite a bit work.

[Marvasti and Wolf](#)

Referencing a different part of your work: please refer to Appendix [A](#).

## Chapter 5

# Summary

Summarize the presented work. Why is it useful to the research field or institute?

### 5.1 Future Work

Possible ways to extend the work.





# Bibliography

- Andrews, D. W. K. (1991, May). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica* 59(3), 817.
- Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994). *Time series analysis: forecasting and control* (3rd ed ed.). Englewood Cliffs, N.J: Prentice Hall.
- Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods*. Springer Series in Statistics. New York, NY: Springer New York.
- Brockwell, P. J. and R. A. Davis (2016). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Cham: Springer International Publishing.
- Chatfield, C. (2003, July). *The Analysis of Time Series* (0 ed.). Chapman and Hall/CRC.
- Cryer, J. D. and K.-s. Chan (2008). *Time series analysis: with applications in R* (2nd ed ed.). Springer texts in statistics. New York: Springer. OCLC: ocn191760003.
- Marvasti, F. and J. K. Wolf (Eds.) (2001). *Nonuniform Sampling*. Information Technology: Transmission, Processing, and Storage. Boston, MA: Springer US.
- Newey, W. K. and K. D. West (1994, October). Automatic Lag Selection in Covariance Matrix Estimation. *The Review of Economic Studies* 61(4), 631–653.
- Robinson, P. (1977, November). Estimation of a time series model from unequally spaced data. *Stochastic Processes and their Applications* 6(1), 9–24.
- White, H. (2001). *Asymptotic theory for econometricians* (Rev. ed ed.). San Diego: Academic Press.
- Zeileis, A. (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software* 11(10).



## Appendix A

# Complementary information

Additional material. For example long mathematical derivations could be given in the appendix. Or you could include part of your code that is needed in printed form. You can add several Appendices to your thesis (as you can include several chapters in the main part of your work).

### A.1 Including R code with verbatim

A simple (rather too simple, see [A.2](#)) way to include code or *R* output is to use `verbatim`. It just prints the text however it is (including all spaces, “strange” symbols,...) in a slightly different font.

```
## loading packages
library(RBGL)
library(Rgraphviz)
library(boot)
```

```
## global variables
X_MAX <- 150
```

```
    This allows me to put as many s p a c e s as I want.
I can also use \ and ' and & and all the rest that is usually only
accepted in the math mode.
```

```
I can also make as
                many
            line
        breaks as
I want... and
                where I want.
```

But really recommended, much better is the following:

However, it is much nicer to use the *listings* package to include R code in your report. It allows you to number the lines, color the comments differently than the code, and so on. All the following is produced by simply writing `\lstinputlisting{Pictures/picture.R}` in your L<sup>A</sup>T<sub>E</sub>X “code”:

```
or \lstinputlisting{/u/maechler/R/Pkgs/sfsmisc/R/ellipse.R} :
```

[illegible]

### A.3 Using Sweave (or knitr) to include R code (and more) in your report

The easiest (and most elegant) way to include R code and its output (and have all your figures up to date with your report) is to use Sweave—or the **knitr** R package with even more possibilities.

Search the web to find lots of intro material on how to use Sweave or **knitr** ([on Wikipedia](#)).



## Appendix B

# Yet another appendix....

### B.1 Description

**Something** details.

**Something else** other definition.

### B.2 Tables

Refer to Table [B.1](#) to see a left justified table with caption on top.

Table B.1: Results.

| <b>Student</b> | <b>Grade</b> |
|----------------|--------------|
| Marie          | 6            |
| Alain          | 5.5          |
| Josette        | 4.5          |
| Pierre         | 5            |





## Appendix C

# 2nd Appendix: More sophisticated R code listing

Chapter-wise listing of parts of R code, using

- `firstline=n1`
- `lastline=n2`
- `title=<text>`

e.g., for the first example below

```
\lstinputlisting[firstline=1,lastline=20,  
                  title= \texttt{ellipse.R}]{ellipse.R}
```

and the second example

```
\lstinputlisting[firstline=20,lastline=40,  
                  title=\texttt{ellipse.R}]{ellipse.R}
```

### C.1 Chapter 5

```
1 ellipsePoints ← function(a,b, alpha = 0, loc = c(0,0), n = 201,  
2                       keep.ab.order = FALSE)  
3 {  
4   ## Purpose: ellipse points, radially equispaced, given geometric par.s  
5   ## -----  
6   ## Arguments: a, b : length of half axes in (x,y) direction  
7   ##             alpha: angle (in degrees) for rotation  
8   ##             loc  : center of ellipse  
9   ##             n    : number of points  
10  ## -----  
11  ## Author: Martin Maechler, Date: 19 Mar 2002  
12  
13  stopifnot(is.numeric(a), is.numeric(b))  
14  reorder ← a < b && keep.ab.order  
15  B ← min(a,b)  
16  A ← max(a,b)  
17  ## B <= A  
18  d2 ← (A-B)*(A+B) ## = A^2 - B^2  
19  phi ← 2*pi*seq(0,1, len = n)  
20  sp ← sin(phi)
```

---

ellipse.R

```
1  sp <- sin(phi)
2  cp <- cos(phi)
3  r <- a*b / sqrt(B^2 + d2 * sp^2)
4  xy <- r * if(reorder) cbind(sp, cp) else cbind(cp, sp)
5  ## xy are the ellipse points for alpha = 0 and loc = (0,0)
6  al <- alpha * pi/180
7  ca <- cos(al)
8  sa <- sin(al)
9  xy %*% rbind(c(ca, sa), c(-sa, ca)) + cbind(rep(loc[1],n),
10                                              rep(loc[2],n))
11 }
```

ellipse.R

# Epilogue

A few final words.



# Declaration of Originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor .

**Title of work** (in block letters):

**Authored by** (in block letters):

*For papers written by groups the names of all authors are required.*

**Name(s):**

**First name(s):**

|        |         |
|--------|---------|
| Muster | Student |
|        |         |
|        |         |
|        |         |
|        |         |

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the Citation etiquette information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work .
- I am aware that the work may be screened electronically for plagiarism.
- I have understood and followed the guidelines in the document *Scientific Works in Mathematics*.

**Place, date:**

**Signature(s):**

|                         |     |
|-------------------------|-----|
| Zurich August 19th 2009 | bla |
|                         |     |
|                         |     |
|                         |     |
|                         |     |

*For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.*