UPPSALA DISSERTATIONS IN MATHEMATICS

111

Gaussian process models of social change

Björn Rune Helmer Blomqvist

Department of Mathematics
Uppsala University
UPPSALA 2018

Dissertation presented at Uppsala University to be publicly examined in Polhemsalen, Ångströmlaboratoriet, Lägerhyddsvägen 1, Uppsala, Friday, 21 December 2018 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Dr. Ian Vernon (Durham University).

Abstract
Blomqvist, B. R. H. 2018. Gaussian process models of social change. *Uppsala Dissertations in Mathematics* 111. 51 pp. Uppsala: Department of Mathematics. ISBN 978-91-506-2734-3.

Social systems produce complex and nonlinear relationships in the indicator variables that describe them. Traditional statistical regression techniques are commonly used in the social sciences to study such systems. These techniques, such as standard linear regression, can prevent the discovery of the complex underlying mechanisms and rely too much on the expertise and prior beliefs of the data analyst. In this thesis, we present two methodologies that are designed to allow the data to inform us about these complex relations and provide us with interpretable models of the dynamics.

The first methodology is a Bayesian approach to analysing the relationship between indicator variables by finding the parametric functions that best describe their interactions. The parametric functions with the highest model evidence are found by fitting a large number of potential models to the data using Bayesian linear regression and comparing their respective model evidence. The methodology is computationally fast due to the use of conjugate priors, and this allows for inference on large sets of models. The second methodology is based on a Gaussian processes framework and is designed to overcome the limitations of the first modelling approach. This approach balances the interpretability of more traditional parametric statistical methods with the predictability and flexibility of non-parametric Gaussian processes.

This thesis contains four papers where we apply the methodologies to both real-life problems in the social sciences as well as on synthetic data sets. In paper I, the first methodology (Bayesian linear regression) is applied to the classic problem of how democracy and economic development interact. In paper II and IV, we apply the second methodology (Gaussian processes) to study changes in the political landscape and demographic shifts in Sweden in the last decades. In paper III, we apply the second methodology on a synthetic data set to perform parameter estimation on complex dynamical systems.

Keywords: Gaussian processes, Bayesian statistics, Dynamical systems, Social sciences

Björn Rune Helmer Blomqvist, Applied Mathematics and Statistics, Box 480, Uppsala University, SE-75106 Uppsala, Sweden.

ISSN 1401-2049
ISBN 978-91-506-2734-3
urn:nbn:se:uu:diva-364656 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-364656)

Till mina syskon,
Hannes och Lisa

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Björn R. H. Blomqvist, Richard P. Mann & David J. T. Sumpter
Using Bayesian dynamical systems, model averaging and neural networks to determine interactions between socio-economic indicators , PLoS ONE 13(5): e0196355. https://doi.org/10.1371/journal.pone.0196355 (2018).

II Björn R. H. Blomqvist, David J. T. Sumpter & Richard P. Mann
Explaining the rise of a radical right-wing party using Gaussian processes
Submitted (2018).

III Björn R. H. Blomqvist
Model selection and parameter estimation of complex dynamical systems using semi-parametric Gaussian processes
Manuscript (2018).

IV Ranjula Bali Swain, Björn R. H. Blomqvist & David J. T. Sumpter
Last night in Sweden? Using Gaussian processes to study changing demographics at the level of municipalities
Submitted (2018).

Contents

Part I:
Introduction & preliminaries

In this first part (out of two) of this thesis, I will present the motivation for this thesis, explain the main theory and provide a brief summary of the appended papers - that later appear in part II.

1. Introduction

This thesis develops and applies statistical models to explain and predict change over time in social systems. The methods explained in this thesis are based on a Bayesian paradigm for inference. This framework is used to provide us with probabilistic models [1] for model selection and hypothesis testing. Mathematical modelling in the social sciences is, unlike in the physical sciences, not aimed to capture any fundamental relationships, like the orbits of planets. Instead, it is used to conceptualize the interactions between humans - something that is continuously changing over time.

In conventional empirical social science, the researchers put forward theories about how social systems work and then use standard regression techniques to test their hypotheses. The conclusions are considered supported if they are pass two tests. Firstly, if there is a significant correlation between a set of indicator variables — describing phenomena in a social system — and a set of explanatory variables, that is, those that are thought to cause the phenomena. Secondly, if the correlations are still present and significant when additional variables, called confounding variables, are introduced. Confounding variables are variables that may affect the dependent variable but are not the variables we are interested in investigating specifically. For example, if we are interested in finding the correlation effect of income levels on unemployment without using confounding variables, such as education level, etc., the resulting found effect may be biased. In the last decade, there has been an increase in criticism of this theoretical approach because it does not focus on the micro-level mechanisms behind the correlations [2, 3, 4].

It has also been argued that the simple additive linear models used in these traditional approaches ignore the complex interactions and nonlinearities in social systems and thereby cannot capture the mechanistic explanations [2]. Another criticism of this approach is that it sometimes consider correlations as evidence of causality. In a traditional linear regression approach, it is impossible to isolate all of the conceivable underlying mechanisms [5]. In mechanistic modelling [6], the assumption is that the social system can be understood by investigating the mechanics of its parts and how they are coupled [7, 8]. Mechanistic models are tested by first constructing models of plausible mechanistic relations and then testing these models against observational data.

Today, computational social science is a growing field and has helped mechanistic social scientists to better combine qualitative assumptions

with quantitative methods [9, 10]. New methodologies together with an explosion of publicly available data sets on all levels — micro, meso, macro — has made this an interesting field with much potential. Additionally, many advances have been made in machine-learning and other computational methods for analysis of complex data structures [11]. These, often 'black-box', methodologies are suited for making predictions, but they often lack in interpretability. Conversely, the mechanistic models often have low predictability, but provide lots of interpretability.

In this thesis, I balance both of the above-mentioned approaches. I do this by developing methodologies to expand the traditionally used 'simple additive linear models' to include both non-linearities and interactional terms and provide a model selection framework to find the most plausible mechanisms to describe observed social dynamics. Using two different probabilistic frameworks, I let the data inform us about the non-linear and interactional dependencies that drive the development of the system. By doing so, I can both test existing theories and use our method as an exploratory tool to test the different hypothesis of the dynamics of social systems.

The first methodological advance in this thesis uses Bayesian linear regression (paper I) that allows us to extend the model space to test a large set of interactions and nonlinearities in the data and perform automatic model selection using parametric modelling. This methodology is an extension of the Bayesian dynamical system modelling approach introduced in [12]. This modelling approach has been used in many applications, and some examples are: how democracy relates to cultural values [13]; studying the sustainable development goals [14, 15, 16]; and school segregation [17]. Our main extension is that our choice of prior information into the models allows for analytical, read fast, computations and thereby allows us to, not only perform model selection to find the 'best' mechanistic model, but to also perform inference on a large set of good models. Even though this methodology provides a step towards bridging the gap between traditional methodologies used in the social sciences and modern statistical techniques, this methodology has some limitations. The two most noticeable are (1) that the number of variables and terms one uses as explanatory terms need to be limited because of the computational costs and (2) that the polynomial form of our model is limited to our choices of allowed linear, non-linear and interactional terms.

To overcome these two limitations, I focus the remaining three papers (II - IV) on a Gaussian process framework. The Gaussian process framework allows us to include many more variables and to break free from the parametric limitation used in paper I. The Gaussian processes framework is a Bayesian statistics technique, heavily used in machine learning, that allows us to perform flexible non-parametric regression

and thereby provide far better predictions. These non-parametric models can be used to get an insight into the dynamics and relevance of the variables. However, they are still a bit like 'black-box' models. To combine the interpretability of the parametric model with the high predictability and flexibility of the non-parametric Gaussian processes, I introduce semi-parametric models. In our semi-parametric models, one can choose to find the 'best' parametric model for some chosen explanatory variables, while controlling for confounding variables using a fully flexible non-parametric component. By using all of these three Gaussian process approaches (parametric, non-parametric and semi-parametric) together, I can get a far better understanding of the mechanisms governing the dynamics of the investigated system, by utilizing their different strengths.

The focus of this thesis is on methodology, but I show how our methodologies can be used by applying them to three different specific problems in social science. The first is related to human development, more specifically the frequently studied relation between democracy and economic development, using global longitudinal data. The second and third application is more recent and highly debated developments regarding Sweden - explaining and predicting the growth in support of the Swedish radical right-wing party 'the Sweden Democrats' and the demographics compositions in Swedish municipalities and their contributions to the changing levels of reported criminality. In paper III, I prove the accuracy of the Gaussian processes methodology using synthetic data.

In chapter 2, I provide a brief overview of the basics of the methodologies used in this thesis. In chapter 3, I describe the main methodological findings in the papers. In chapter 4, I provide a brief summary of the appended papers. In chapter 5, I correct some typos and respond to a few comments on paper I, an already published paper, that appeared in the review process of this thesis.

2. Preliminaries: From standard linear regression to Gaussian process regression

In order to make robust inference models for the development of social systems — that is, to fairly compare different hypothetical models, assess their accuracy, and to produce accurate forecasts — we need a suitable framework to perform model selection. In this section we will look at the two main frameworks we will utilize to analyze social systems. The two approaches are (1) Bayesian linear regression and (2) Gaussian processes. Bayesian linear regression is - as the name indicates - a linear approach within a Bayesian context, where prior information about the system can easily be incorporated. Gaussian processes are a more flexible approach to regression. Rather than assuming that the relations between the response variable y and the predictor variable x takes a linear (parametric) form $f(\mathbf{x})$, (e.g., $y = 1 + x + x^2 + x^3$), the Gaussian process approach allows for a more flexible, non-parametric approach where the correlations in the data are defined by the choice of mean function and kernel (covariance function). These two approaches both have their advantages and disadvantages, but in this thesis, we will show how we can retain the best aspects of Bayesian linear regression in a Gaussian process framework and combine them in order to make better models for social systems.

In paper I, we use Bayesian linear regression, and in papers II-IV, we use Gaussian processes. In this section, we will first briefly go through both of these two approaches in turn. This chronology is natural for this thesis, not only because it is natural to introduce Gaussian processes using a technique known by a broader audience, but it is also the order in which the papers were written. The reader should view this section as an overview of the techniques we deal with in this thesis. For a more detailed look, I highly recommend the book *Gaussian processes for machine learning* by Rasmussen and Williams (2006) [18].

Throughout the upcoming sections, we assume n observations of response variables y, a scalar, and D dimensional explanatory variables $\mathbf{x} = [x_1, x_2, ..., x_D]^\top$, i.e., we consider the data $\{(y_i, \mathbf{x}_i)|i = 1, ..., n\}$. The aggregated data where all n observations are included for the response variable is the 1 times n dimensional vector $\mathbf{y} = [y_1, y_2, ..., y_n]^\top$. The aggregated form of the explanatory variables is a D times n dimensional design matrix X,

$$X = \begin{pmatrix} x_{1,1} & x_{1,2}, & \dots & x_{1,n} \\ x_{2,1} & x_{2,2}, & \dots & x_{2,n} \\ \vdots & \vdots, & \ddots & \vdots \\ x_{D,1} & x_{D,2}, & \dots & x_{D,n} \end{pmatrix}. \tag{2.1}$$

Sometimes we will also refer to test data, intended for predictions using our models, these test data will be either $\mathbf{x}_*$, a D dimensional test points or we aggregate multiple test points, we use X_*, which is a D times n_* dimensional matrix.

2.1 Parametric regression

Classical regression is one of the most widely used methods for statistical analysis. In this section, I will give a short overview of the Bayesian treatment of linear regression. This technique is mainly used in paper I (with a slightly more complicated structure) and for the parametric models in paper II and paper III.

2.1.1 Parametric approach to regression

In a standard approach to linear regression, the model assumes that a response variable y (also known as the dependent variable, or target), is explained by the function $f(\mathbf{x})$, a linear combinations of a set of explanatory variables (also known as predictor variables, or covariates) multiplied with the parameters (or weights) $\beta = [\beta_0, \beta_1, \beta_2, ...]^\top$. This approach is also known as a weight space approach to regression and has the following form,

$$y = f(\mathbf{x}) + \epsilon \tag{2.2}$$

In the model, we have also included a normally distributed error term ϵ to take random sampling noise into account. For example, a model with three explanatory variables $\mathbf{x} = [x_1, x_2, x_3]^\top$, can be represented as:

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}_{f(\mathbf{x})} + \epsilon \tag{2.3}$$

Where the noise ϵ is assumed to be additative, independent, identically distributed (i.i.d.) Gaussian distribution with zero mean and variance σ_n^2, i.e, $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$. In this equation, the response variable y is explained by the weights: β_0 a constant intercept variable and $\beta_1, \beta_2, \beta_3$, assigning weights to explanatory variables x_1, x_2 and x_3. Linear regression is an elementary and highly interpretable model. The weights β are

often called the parameters of the model — hence the name parametric approach to regression — shows the effect on the response variable when the explanatory variable increases or decreases. For example, if the weight β_1 is 2, then the response variable increases by two for every unit increase in x_1.

The linear model can be generalized to multivariate matrix formulation with any number of explanatory variables. We can add a vector of ones to the predictor matrix X to account for the intercept β_0, but for simplicity we chose to remove the intercept. In matrix notation, the linear model is then,

$$\mathbf{y} = X^\top \beta + \epsilon \tag{2.4}$$

Now when we have a linear model for the response variables, the next step is to train our model, i.e., finding the parameter values of the weights β that best explain the data. In traditional linear regression, the parameter values are most often found by identifying the parameter values that minimize the sum of the residual squared, S (minimizes the sum of the squared difference between the actual observations and the predictions made by the linear model) [19]. So we look for the β that minimize,

$$S(\beta) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{D} x_{j,i}\beta_j)^2. \tag{2.5}$$

For the ordinary least squares estimate of the parameters $\hat{\beta}$, the closed form solution is,

$$\hat{\beta}_{OLS} = (XX^\top)^{-1} X\mathbf{y} \tag{2.6}$$

Observe that the parameter estimates in the OLS estimate are dependent only on the information given in the data. For Gaussian distributed noise in our models, like we have in this case, the OLS estimate is equivalent of the maximum likelihood estimate - where we want to maximize the likelihood function, $p(\mathbf{y}|X, \beta)$, i.e., finding the parameters β that maximizes the probability of producing the response $\mathbf{y}$ given the data X. From a Bayesian viewpoint, which we will discuss in the next section, it is worth noting that the maximum likelihood estimate is a special case of the maximum a posteriori estimate where the prior is uniform.

2.1.2 Bayesian linear regression

In a Bayesian way of looking at the linear model, we are interested in probability distributions of the parameter values instead of point estimates - where our prior beliefs of the parameters is incorporated. In a Bayesian way of looking at the model, we thereby use a probabilistic

approach of modelling. Instead of finding one estimate of the parameter values of β, that best fits the data, we are interested in finding the posterior distribution of the parameters of the model. The posterior of the model parameters is the conditional probability assigned after the relevant evidence is taken into account, and is calculated using the likelihood $p(\mathbf{y}|X,\beta)$, prior $p(\beta)$ and marginalized likelihood $p(y|X)$ using Bayes theorem,

$$p(\beta|\mathbf{y},X) = \frac{p(\mathbf{y}|X,\beta)\cdot p(\beta)}{p(\mathbf{y}|X)} \tag{2.7}$$

or in text form,

$$\text{posterior} = \frac{\text{likelihood}\cdot\text{prior}}{\text{marginal likelihood}} \tag{2.8}$$

The probability density of the observations given the parameters is called the *likelihood* and often plays a big role in inference also where no priors are incorporated. It is the probability of the observed response given the explanatory variables and the parameter values. For the model with Gaussian noise, the likelihood is given by,

$$\begin{aligned} p(\mathbf{y}|X,\beta) &= \prod_{i=1}^{N} p(y_i|\mathbf{x}_i,\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_n} exp(\frac{(y_i-\mathbf{x}_i^\top\beta)^2}{2\sigma_n^2}) \\ &= \frac{1}{(2\pi\sigma_n^2)^{n/2}} exp(-\frac{1}{2\sigma_n^2}|\mathbf{y}-X^\top\beta|^2) \end{aligned} \tag{2.9}$$

where we use the notation: $|\mathbf{w}|$ for the euclidian length of the vector $\mathbf{w}$. Hence, the response variable is assumed to have the following distribution,

$$\mathbf{y}|X,\beta \sim \mathcal{N}(X^\top\beta, \sigma_n^2\mathbf{I}) \tag{2.10}$$

In Bayesian formalism, the prior $p(\beta)$ represents the prior beliefs of the parameter β that go into the model, before any data is introduced. If we assume a zero mean Gaussian prior with covariance matrix V the prior of β is,

$$\beta \sim \mathcal{N}(\mathbf{0}, V). \tag{2.11}$$

The normalizing constant $p(\mathbf{y}|X)$ in the denominator Eq. 2.7 is the marginalized likelihood, where we have marginalized by integrating the likelihood times the prior over the space of all possible parameter values for β. So the marginalized likelihood is given by,

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|X,\beta)p(\beta)d\beta \tag{2.12}$$

The posterior distribution $p(\mathbf{y}|X,\beta)$ of the parameter combines all our knowledge about the parameter by using the likelihood and the prior, i.e.,

is a 'compromise' between the likelihood and the prior. After combining the likelihood and the prior - and some rearranging (See Gelman [20]), we can see that the posterior is distributed as,

$$\beta|X,\mathbf{y} \sim \mathcal{N}\Big(\sigma_n^{-2}(\sigma_n^{-2}XX^\top + V^{-1})^{-1}X\mathbf{y}, (\sigma_n^{-2}XX^\top + V^{-1})^{-1}\Big) \quad (2.13)$$

where $\sigma_n^{-2}(\sigma_n^{-2}XX^\top + V^{-1})^{-1}X\mathbf{y}$ is the maximum a posteriori $\bar{\beta}$ of the parameters of the model and $(\sigma_n^{-2}XX^\top + V^{-1})^{-1}$ is the covariance matrix.

So the first good aspect of using a Bayesian formalism is that we can get a probability distribution of the parameter values of our model. The second related positive aspect is that we can incorporate previous knowledge into our models using priors. If we wish to produce predictions using this model, then we can use the predictive distribution f_*, the prediction for f at test locations X_*. The predictive distribution is also normally distributed as,

$$f_*|X_*.X,\mathbf{y} \sim \mathcal{N}\Big(\bar{f}_*, \mathrm{cov}(f_*)\Big) \quad (2.14)$$

where the mean and the covariance of the predictor distribution is,

$$\begin{aligned} \bar{f}_* = &\quad \sigma_n^{-2}X_*^\top(\sigma_n^{-2}XX^\top + V^{-1})^{-1}X\mathbf{y} \\ \mathrm{cov}(f_*) = &\quad X_*^\top(\sigma_n^{-2}XX^\top + V^{-1})^{-1}X_*. \end{aligned} \quad (2.15)$$

The predictive distribution is the average of the outputs of all the possible models with respect to the Gaussian posterior [18].

Now we have expressions to understand the posterior distribution of the parameter β and have formulas to get the predictive distributions for our function f. This combined helps us to understand how sure we are about our estimates.

Kernel formulation

To be able to easily compare these results with the next section about Gaussian process regression, we use the 'kernel trick' to rewrite our expression using some matrix algebra into a kernel formulation. In kernel formulation, the expression for the mean $\bar{f}_*$ and variance $\mathrm{cov}(f_*)$ is,

$$\begin{aligned} \bar{f}_* = &\quad K_{\mathrm{dot}}(X_*, X)(K_{\mathrm{dot}} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} \\ \mathrm{cov}(f_*) = &\quad K_{\mathrm{dot}}(X_*, X_*) - K_{\mathrm{dot}}(X_*, X)(K_{\mathrm{dot}} + \sigma_n^2\mathbf{I})^{-1}K_{\mathrm{dot}}(X, X_*) \end{aligned} \quad (2.16)$$

where,

$$\begin{aligned} K_{\text{dot}} &= X^\top V X \\ K_{\text{dot}}(X_*, X) &= X_*^\top V X \\ K_{\text{dot}}(X, X_*) &= X^\top V X_* \\ K_{\text{dot}}(X_*, X_*) &= X_*^\top V X_*. \end{aligned} \tag{2.17}$$

We call the kernels 'dot' because this is known as a dot product kernel in a Gaussian process . To show that the expressions in Eq. 2.16 are equivalent to the expressions in Eq. 2.15 we show them in turn.

We begin by showing it for the mean $\bar{f}_*$. If we start with the following equality,

$$\begin{aligned} &(\sigma_n^{-2} X X^\top + V^{-1}) V X \\ &= \sigma_n^{-2} (X X^\top V + \sigma_n^2 \mathbf{I}) X \\ &= \sigma_n^{-2} X (X^\top V X + \sigma_n^2 \mathbf{I}). \end{aligned} \tag{2.18}$$

If we then multiply both sides from the left with $(\sigma_n^{-2} X X^\top + V^{-1})^{-1}$ we get,

$$V X = \sigma_n^{-2} (\sigma_n^{-2} X X^\top + V^{-1})^{-1} X (X^\top V X + \sigma_n^2 \mathbf{I}). \tag{2.19}$$

If we then multiply both sides with $(X^\top V X + \sigma_n^2 \mathbf{I})^{-1}$ from the right, we then get,

$$V X (X^\top V X + \sigma_n^2 \mathbf{I})^{-1} = \sigma_n^{-2} (\sigma_n^{-2} X X^\top + V^{-1})^{-1} X \tag{2.20}$$

and then by finally multiplying both sides with $X_*^\top$ from the left, and $\mathbf{y}$ from the right gives us,

$$\underbrace{X_*^\top V X}_{K_{\text{dot}}(X_*, X)} (\underbrace{X^\top V X}_{K_{\text{dot}}} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} = \sigma_n^{-2} X_*^\top (\sigma_n^{-2} X X^\top + V^{-1})^{-1} X \mathbf{y} \tag{2.21}$$

which show that the expressions for the mean is equivalent.

To show the variances of Eq. 2.15 and Eq. 2.16 are equal we start with,

$$\text{cov}(f_*) = \quad X_*^\top \underbrace{(\sigma_n^{-2} X X^\top + V^{-1})^{-1}}_{\oplus} X_* \tag{2.22}$$

and then use the matrix inversion lemma (e.g., A.9 in [18]), on $\oplus$,

$$(\sigma_n^{-2} X X^\top + V^{-1})^{-1} = \quad V - V X (\sigma_n^2 \mathbf{I} + X^\top V X)^{-1} X^\top V \tag{2.23}$$

hence, we get the following expression for the variance,

$$X_*^\top (V - VX(\sigma_n^2 \mathbf{I} + X^\top VX)^{-1} X^\top V) X_* =$$
$$= \underbrace{X_*^\top V X_*}_{K_{\text{dot}}(X_*, X_*)} - \underbrace{X_*^\top V X}_{K_{\text{dot}}(X_*, X)} (\underbrace{X^\top V X}_{K_{\text{dot}}} + \sigma_n^2 \mathbf{I})^{-1} \underbrace{X^\top V X_*}_{K_{\text{dot}}(X, X_*)}. \tag{2.24}$$

Model selection

The marginal log likelihood or 'model evidence' of a Bayesian linear regression model is a built-in, and often used tool of measuring the fit of a model. The marginal likelihood refers to the marginalization over the parameter values of our model, which we have already seen in Eq. 2.12. In our Bayesian linear model, the marginalized likelihood is distributed according to, $\mathbf{y}|X \sim \mathcal{N}(\mathbf{0}, K_{\text{dot}} + \sigma_n^2 \mathbf{I})$, where $\mathbf{I}$ is the identity matrix, and then it directly follows that the logarithmized marginal likelihood (LogML) is [18],

$$\begin{aligned} \log p(\mathbf{y}|X, \theta) = \quad & \frac{1}{2} \mathbf{y}^\top (K_{\text{dot}} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} \\ & - \frac{1}{2} \log |K_{\text{dot}} + \sigma_n^2 \mathbf{I}| \\ & - \frac{n}{2} \log(2\pi) \end{aligned} \tag{2.25}$$

where θ is the set of parameters of the model. Please note that the expression of the model evidence gets slightly different if we do not assume β to have a zero mean prior (see Eq. 2.43 in [18].

In Eq. 2.25, the top quadratic term in $\mathbf{y}$ gets bigger when there is greater model fit, the (middle) negative log determinant term punishes over-structured (too detailed) models by giving a negative contribution, and then the last term, proportional to the number of observations n is a normalization. So by altering the hyperparameters, we can balance the model fit and over-complexity, to find our models. In a Bayesian linear regression model, the complexity can be increased by adding more explanatory variables or by introducing more complicated terms in the polynomial form.

Dealing with non-linearities

Previously, we presented linear regression as weighted combinations of explanatory variables that together explain the response variable. To overcome the limitation in the expressiveness of this approach, we can project the input variables into a higher dimensional space using basis functions and then perform regression on this new space instead. The higher dimensional space can be specified as we wish, and a typical

example of this approach is called polynomial regression. The typical example would be to project the scalar x_1 into the space of powers of x_1: $\mathbf{b}(x_1) = (1, x_1, x_1^2, x_1^3...)^\top$ — e.g., $f(x_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3$ — to be able to capture more complicated relationships in the data. This is the same thing as performing polynomial regression and we thereby, with ease, use $\mathbf{b}$ instead of $\mathbf{x}$ and the aggregated basis function B instead of X. Also, interactional terms between multiple input variables are possible. As long as the projection into the polynomial space is independent of the weight parameters β, then the model is still linear in the parameters, and the regression is still analytically tractable [18].

Choosing a prior

In a Bayesian approach to inference, it is useful to accept that we often have some knowledge about the possible values of the parameter of our model. The prior is how we quantify this knowledge, and how we can incorporate it into the model. There is no limitation to what prior distribution we can assign to our model parameters. There are several different priors we can choose, for example, uniform prior, where the probability distribution is constant over some range; Jeffreys' priors [21]; reference priors [22, 23]; maximum entropy priors [24]; and weakly informative priors, where some partial information is expressed. The prior distribution can hence be chosen in line with our current beliefs of the distribution, such as the results of previous studies or the subjective beliefs of practitioners in the field, or even to have some more specific mathematical property. In general, we can divide the choice of priors into two main groups [25], informative and uninformative priors. An informative prior (or subjective prior) gives a specific distribution based on previous information. For example, if we would want to express our beliefs about a particular political party's current voter support rate, a naive guess would be to assume a normal distribution and to take the average of the latest polls (or last election result) to be the mean and sample variance of the polls to be the variance of the parameter. An uninformative prior (or objective prior) expresses some general knowledge about the parameters. For example, an uninformative prior could be to put sharp limits for possible parameter values, e.g., that the temperature where some bacteria lives, or that it the temperature has to be positive.

The prior should somehow reflect the prior information or guess, but sometimes we want the prior to express how little we know about the parameter values. For example, we may not wish to assume any form of the prior, i.e., letting the data fully inform us about the relationships hidden within. By assigning a uniformly flat prior ranging from $-\infty$ to $+\infty$, we have assigned equal weight to all possible parameter values, i.e., all parameter values are equally probable. This choice of prior is the same as arguing that the posterior is proportional to the likelihood,

with only a normalization constant. However, this assumption does not necessarily imply that the normalization constant in the calculation of the posterior converges [26]. It is common in applications [12] to choose a uniform prior distribution but to have cut-offs by only allowing a specific plausible range for the parameter values. By doing so we have set limits to our ignorance, but there are other limitations to this choice.

Depending on our choices we need to use different techniques to calculate the posterior distribution. For an arbitrary prior distribution, an analytical solution might not exist, and this is the case with the uniform prior. If we can not calculate the posterior distribution analytically, we need to use numerical sampling techniques to approximate the posterior distribution by drawing samples from it. There are many different ways of sampling from the posterior distribution, but it is common to use Monte Carlo sampling technique.

However, using a conjugate prior, a prior that is conjugate to the likelihood function if the functional form is the same can give a posterior distribution that can be derived analytically. This is why we, in this thesis, have chosen to work with either Normal Inverse Gamma (NIG) priors or just simply Gaussian priors. We choose these conjugate priors, together with high prior variances to express our ignorance. By having this choice, that allows for all possible parameter values, even though we assume that the most likely value is zero, the more data we include, the less weight is on the prior and more on the actual evidence. These approach does not put much weight on the prior, i.e., a vague prior, and thereby allowing the data 'do all the talking'.

2.1.3 An example

To show how a Bayesian formalism can be used for model selection, we present an example where we compare the model fit on some data for three different model set-ups. We generate 200 data points using the polynomial generating function $y = x + 0.3x^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$, i.e., a linear and a quadratic component. We then fit three different models, first a simple linear model $y = \beta_1 x$, then a linear plus quadratic model $y = \beta_1 x + \beta_2 x^2$, called the quadratic model, due to the highest order term, and a cubic model, $y = \beta_1 x + \beta_2 x^2 + \beta_3 x^3$. We assume

Table 2.1. *Model evidence for the linear, quadratic, and cubic models*

Model	Fitted model	Model evidence (LogML)
Linear	$y = 1.12x$	-1539.26
Quadratic	$y = 1.04x + 0.29x^2$	-297.32
Cubic	$y = 1.05x + 0.29x^2; +0.007x^3$	-307.05

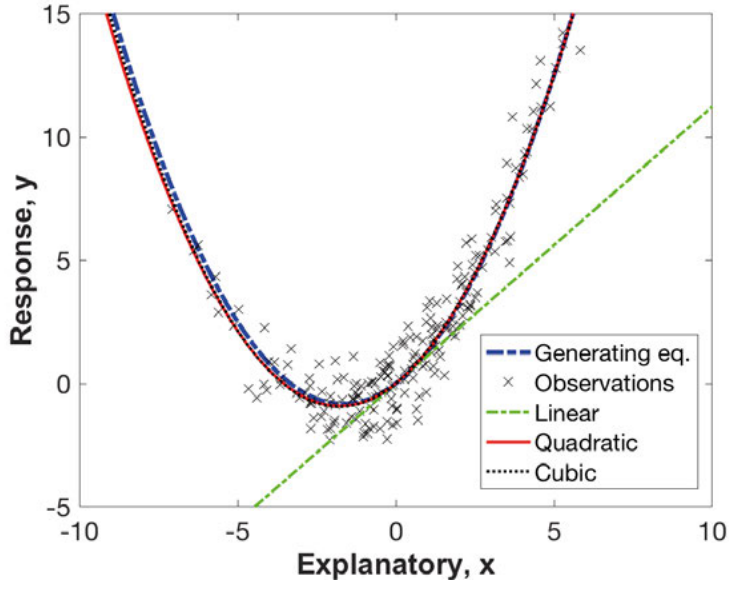


Figure 2.1. **Model fit of the three different models: linear (green), quadratic (red), and cubic (black)**. The data is generated by the function, $y = x + 0.3x^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$.

the following prior distribution for the parameters $\beta = \mathcal{N}(\mathbf{0}, 100 \cdot \mathbf{I})$, a vague prior. After fitting these models to the observed data, we find the following means of the posterior distribution of the parameters of the three different models together with their model evidence (LogML),

As we can see, the model evidence is the highest for the quadratic model. In Fig. 2.1 we can see that the linear model does not provide enough complexity to capture the generating function, and both the quadratic and cubic models seem to fit the data well. The higher model evidence for the quadratic model tells us that the cubic model is too complex and we do not need the extra information that an x^3-term gives.

2.1.4 Transition to Gaussian processes

In linear (polynomial) regression it is fundamental to the model that the observable data are noisy observations of regressors that are latent linear functions. Clearly, a lot of other functional relationships than the ones used in linear regression are possible. The goal with the next section, Gaussian processes are introduced to loosen up this restriction in the functional form of f in Eq. 2.2. By using Gaussian processes, we are not restricted to polynomial functions. Instead, we are free to define the how covariances in the data should be specified.

2.2 Gaussian processes

Gaussian processes [18, 27, 28, 29, 30, 31, 32] is a generic method for supervised learning used for both regression and classification. Gaussian processes provide us with a powerful and flexible framework to perform Bayesian inference. In a Gaussian process viewpoint, we can dispense the parametric model (Bayesian linear regression) and instead of putting priors on the weights, we can now define prior probabilities over the functions directly, i.e., performing inference on the function space instead of on the weigh space [32]. Rasmussen and Williams (2006) [18], define a Gaussian process as:

Definition: A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

A Gaussian process is thereby a collection of random variables, and these random variables are in our case the function $g(\mathbf{x})$ we wish to learn about at positions $\mathbf{x}$. Gaussian processes are entirely specified by their second-order statistics, i.e., the two functions: (1) the mean function μ,

$$\mu(\mathbf{x}) = \mathbb{E}[g(\mathbf{x})] \tag{2.26}$$

and (2) the covariance function k (often called the kernel),

$$k(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[g(\mathbf{x}), g(\mathbf{x}')] = \mathbb{E}[\big(g(\mathbf{x}) - \mu(\mathbf{x})\big)\big(g(\mathbf{x}') - \mu(\mathbf{x}')\big)]. \tag{2.27}$$

In supervised learning approaches, it is often assumed that input data $\mathbf{x}$ that are located close to each other are likely to have similar responses y. The mean function expresses our beliefs about the expected values of the random variable $g(\mathbf{x})$ and in the covariance function we specify how we choose to defines to define the similarity of the function values of $g(\mathbf{x})$ at data points $\mathbf{x}$ and $\mathbf{x}$. [32, 33].

Gaussian processes and models equivalent to Gaussian processes have been extensively studied in a multitude of fields. For example, the Kalman filter [34], radial basis function networks [35] (an artificial neural network using radial basis activation functions) and ARMA [36] (autoregressive moving average) models, can all be seen as forms of Gaussian processes. Also, in geostatistics Gaussian processes regression is known as Kriging [37].

2.2.1 The model

We now choose to approximate the response variable y with the function $g(\mathbf{x})$, of some unspecified functional form, and some additative i.i.d. noise $\epsilon \sim \mathcal{N}(0, \sigma_n^2\mathbf{I})$. We choose to call the function that encodes the

dependency of variables $\mathbf{x}$, $g(\mathbf{x})$, instead of $f(\mathbf{x})$, to distinguish this approach from the Bayesian linear regression approach. The model is then,

$$y = g(\mathbf{x}) + \epsilon \tag{2.28}$$

In this approach to regression we specify the function g, by it's covariance structure instead of it's basis functions. In many applications, the mean μ is set to be zero, which implies that we assume that the function takes values to be zero, unless the data tells us otherwise. In this section, as well as in the papers, we also make this assumption. In our model setup (Eq. 2.28) we assumed noisy observations, and those can also be modelled as a Gaussian process. The sum of two Gaussian processes is also a Gaussian process, and thereby we can combine the two Gaussian processes,

$$\begin{aligned} g(\mathbf{x}) \sim &\quad \mathcal{GP}\big(0, k(\mathbf{x}, \mathbf{x}')\big) \\ \epsilon \sim &\quad \mathcal{GP}\big(0, \sigma_n^2 \delta_{\mathbf{x},\mathbf{x}'}\big) \end{aligned} \tag{2.29}$$

to form another Gaussian process. The $\delta_{\mathbf{x},\mathbf{x}'}$ is the Kronecker delta, i.e., $\delta_{\mathbf{x},\mathbf{x}'} = \mathbb{I}(\mathbf{x} = \mathbf{x}')$. Since we have added the two functions in $g(\mathbf{x})$ and ϵ in Eq. 2.28 we can present the model as the Gaussian process,

$$y(\mathbf{x}) \sim \mathcal{GP}\big(0, k(\mathbf{x}, \mathbf{x}') + \sigma_n^2 \delta_{\mathbf{x},\mathbf{x}'}\big) \tag{2.30}$$

So the mean function of the new Gaussian process is zero and, in aggregated form, the covariance of the response variable $\mathbf{y}$ is then $\text{cov}(\mathbf{y}) = K(X, X) + \sigma_n^2 \mathbf{I}$, where $K(X, X)$ is the generic covariance matrix, a $n \times n$ matrix of the covariances evaluated for every pair of the input data. The joint distribution of the training data from the observations $\mathbf{y}$ and the response output g_* for some test points $\mathbf{x}_*$ is,

$$\begin{bmatrix} \mathbf{y} \\ g_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \quad \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$

where $K(X, X_*)$ denotes the n times n_* matrix and $K(X_*, X)$ denotes the n_* times n matrix of the covariances evaluated at every single pair of training points and test points. We are interested in finding the posterior distribution over function to be able to get predictions of our model. The posterior distribution is obtained by restricting the joint prior distribution only to contain functions that behave similarly and according to the observed data. By conditioning the joint Gaussian prior distribution on the observed data we obtain the predictive distribution,

$$g_* | X, X_*, \mathbf{y} \sim \mathcal{N}(\bar{g}_*, \text{cov}(g_*)) \tag{2.31}$$

where the expected predictive equations $\bar{g}_*$ and the covariance of the prediction $\text{cov}(g_*)$ are [18],

$$\begin{aligned} \bar{g}_* = &\quad K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y} \\ \text{cov}(g_*) = &\quad K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, X_*) \end{aligned} \tag{2.32}$$

Observe the similarity in predictive distribution with the ones in the Bayesian linear regression case. The only thing that is different is that the Bayesian linear regression is a particular case of covariance function.

2.2.2 Covariance functions

The covariance function encodes our assumptions about the function we want to learn. The covariance function holds the covariance or correlation of pair of the output variables, $g(x)$ and $g(x')$ as a function of the input variables, x and x'. Depending on our choices of covariance function, it is possible to reproduce many different types of regression, e.g. Bayesian linear regression and neural networks, and this is one of the reasons why Gaussian processes are so flexible and powerful. For example, to see that we can obtain Bayesian linear regression using a Gaussian process framework we can plug in the standard Bayesian linear regression model $f(x) = \mathbf{x}^\top \beta$, which prior $\beta \sim \mathcal{N}(\mathbf{0}, V)$ into Eq. 2.26 and Eq. 2.27. We then get the mean function,

$$\mu(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] = \mathbf{x}^\top \mathbb{E}[\beta] = 0 \tag{2.33}$$

and the covariance function,

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] = \mathbf{x}^\top \mathbb{E}[\beta\beta^\top]\mathbf{x}' = \mathbf{x}^\top V \mathbf{x}' \tag{2.34}$$

or in aggregated form, $K(X, X) = X^\top V X$, just like in Eq. 2.17.

In this thesis we will focus on the most commonly used covariance function in Gaussian process regression, namely the squared exponential covariance function $k_{\text{SE}}(\mathbf{x}, \mathbf{x}')$. This covariance function can be seen as a parametric polynomial regression with infinitely many basis functions and is infinitely differentiable, which means that it is very smooth. This covariance function specifies the covariance and correlations between pairs of random variables to be,

$$k_{\text{SE}}(\mathbf{x}, \mathbf{x}') = \quad \sigma_g^2 \exp\left(-\sum_{i=1}^{D} \frac{(x_i - x_i')^2}{2l_i^2}\right). \tag{2.35}$$

where the hyperparameters l_i defines the characteristic length-scale of variable i and σ_g^2 is the signal variance. The characteristic length-scale l

can be seen as the distance between inputs x and x' before the value of the function change significantly, i.e., how 'wiggly' the function is allowed to be. The signal variance σ_g^2 is a scaling factor that determines how free the function are to deviate from the mean. A small signal variance gives us functions that are close to the mean, and a large signal variance allow the function to adapt more freely to the data. Hyperparameters is a name for parameters before the optimization process, to find the appropriate values, has started.

Even though the squared exponential covariance function is one of the most commonly used covariance functions in applications, it has some potentially problematic properties. Since it is so smooth, it can have problems fitting some real physical systems where there are sharp 'kinks' in the approximating function, and the characteristic length-scale can be set too short to fit the 'kinks' resulting in overfitting in other regions. In section 2.2.5, we will show what happens when we change the hyperparameters of the model. To overcome the problems that might arise using a too smooth covariance function, other covariance functions can be used, such as the less smooth Matérn covariance function,

$$k_{\text{Matérn}}(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)}\Big(\frac{\sqrt{2\nu}}{l}|\mathbf{x}-\mathbf{x}'|\Big)^{\nu} K_{\nu}\Big(\frac{\sqrt{2\nu}}{l}|\mathbf{x}-\mathbf{x}'|\Big) \tag{2.36}$$

where, K_ν is a modified Bessel function of order ν, Γ is the gamma function, and ν and the characteristic length-scale l are positive parameters. If we take sample functions of Matérn form, then these are $(\nu - 1)$ times differentiable. If the parameter $\nu \to \infty$ then we obtain the squared exponential covariance function. There are many different covariance functions we can choose for our models. We can also combine and change existing covariance functions to make new, more specialized covariance functions. For example, the sum, and the product of two covariance functions is also a covariance function. In applications of Gaussian process, the hard thing is often to choose appropriate covariance function form and combinations.

For the purposes in this thesis, where we study social systems where the data is often very noisy, and the objects we study tend to not have sharp 'kinks' according to previous studies and theory. We use the squared exponential covariance function because we want a model that provides lots of flexibility and is easy to grasp - since it is equivalent to a polynomial model with infinitely many terms.

As a prior, the user of this covariance function set the hyperparameters to some arbitrary choice - commonly all set to one. The goal now is to determine the values of the hyperparameters that best fits the data.

2.2.3 Model selection

In Bayesian inference, the logarithm of the marginalized likelihood is in general used as a measure for how well the model fits the data. The log marginal likelihood is the logarithmized probability of getting the data response variables $\mathbf{y}$ given the explanatory variables $\mathbf{x}$ and the model. There are many ways of using the log marginal likelihood to perform model selection. One way is to try different types of covariance functions and then pick the one that has the highest model evidence, and the other way is to change within one choice of covariance function, i.e., the hyperparameters.

The squared exponential covariance function has the log marginal likelihood,

$$\begin{aligned} \log p(\mathbf{y}|X,\theta) = \quad & \frac{1}{2}\mathbf{y}^\top (K_{\mathrm{SE}} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} \\ & -\frac{1}{2}\log|K_{\mathrm{SE}} + \sigma_n^2\mathbf{I}| \\ & -\frac{n}{2}\log(2\pi) \end{aligned} \tag{2.37}$$

where $K_{\mathrm{SE}} = K_{\mathrm{SE}}(X,X)$ is the squared exponential covariance function, and $\theta = (l, \sigma_g, \sigma_n)$ are the hyperparameters. One nice thing with the log marginal likelihood is that it has a built-in way of punishing over complicated models. In Eq. 2.37, the top quadratic term in $\mathbf{y}$ gets bigger when there is greater model fit, the (middle) negative log determinant term punishes over-structured (too detailed) models by giving a negative contribution, and the last term, proportional to the number of observations n is a normalization. So by altering the hyperparameters, we can balance the model fit and complexity of our models. Other well known criteria for balancing model fit with model parsimony are e.g., the Akaike information criterion (AIC) [38], Bayesian information criterion (BIC) [39] and the widely applicable information criterion (WAIC) [40].

2.2.4 Optimizing hyperparameters

By changing the hyperparameters — l, σ_g, σ_n — we can obtain different model evidence and model fit when we introduce some data. One way of finding the hyperparameters that provide the highest model evidence is to use a gradient-based optimizer. The gradient-based optimizer we use throughout this thesis is a part of the gpml toolbox [41] developed by Rasmussen and Hannes Nickisch. This method uses the partial derivatives of the model evidence with respect to the hyperparameters and updates the hyperparameters in the direction to increase the model evidence. This method is relatively computationally cheap to use, and the

complexity of the optimization is dominated by the cost of inverting the matrix K_{SE}. For more details about the optimization procedure, see [18].

When we search for the optimal choice of hyperparameters, we are not guaranteed to find the global maxima of the marginal likelihood. It is possible that the optimization procedure gets stuck on some local maxima. One way of trying to overcoming this problem is to do multiple random restarts, i.e., restart the optimization with some random initial hyperparameters and the pick the model with the highest model evidence. According to [18], for simple choices of covariance functions — like the ones we are working with — this is not a big problem, but we should always be aware that this potential problem exists.

2.2.5 An example

A good way of understanding how Gaussian process regression can be used in regression is by demonstrating it using an example. We generate response variables y using the same polynomial generating function as in the previous example in section 2.1.3 with only one explanatory variable x. We then use the squared exponential covariance function with a zero mean to approximate the generating function. Initially we set the hyperparameters $\theta = (l, \sigma_g, \sigma_n)$ to all be set to one, $(l, \sigma_g, \sigma_n) = (1, 1, 1)$. The model fit, using these initial values before optimization, has model evi-

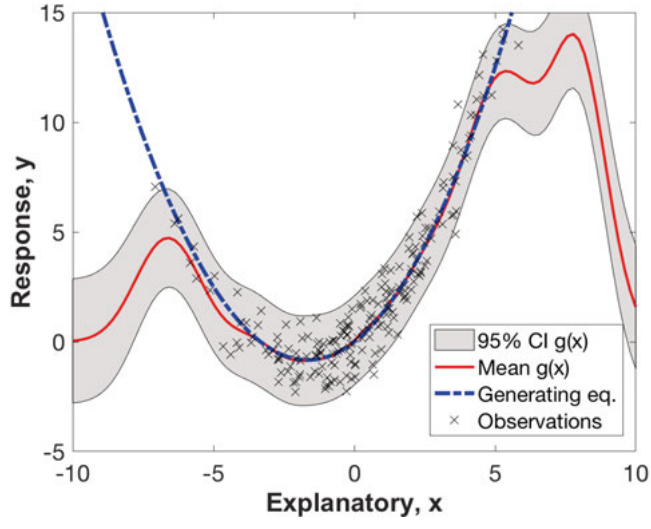


Figure 2.2. **Model fit of the model with a squared exponential covariance function before the optimization of hyperparameters**. The length-scales are set to $(l, \sigma_g, \sigma_n) = (1, 1, 1)$. The model evidence of this model is, logML $= -583.9572$.

Table 2.2. *Model evidence and parameter values during the process of optimizing the hyperparameters of the model.*

Iteration	Hyperparameters l	σ_g	σ_n	LogML
initial	1	1	1	−583.96
1	1	1	0.3679	−1122.86
5	1.0022	1.0099	0.3736	−1096.27
10	1.7024	6.2738	0.7048	−345.40
20	10.6084	51.9238	1.0654	−301.88
30	19.7598	88.9712	1.0026	−299.70
40	17.3968	67.2153	0.9986	−299.63
50	17.2998	67.6596	0.9992	−299.63
100	17.1404	66.0101	0.9992	−299.62
200	17.1405	66.0100	0.9992	−299.62

dence, $LogML = -583.96$, and can be seen in Fig. 2.2. We can see that the model fits well where the data lies and is 'pushed' towards zero where there is no data - however it seems like we have a too short characteristic length-scale. The shaded area in gray, is a 95% confidence region, i.e.,

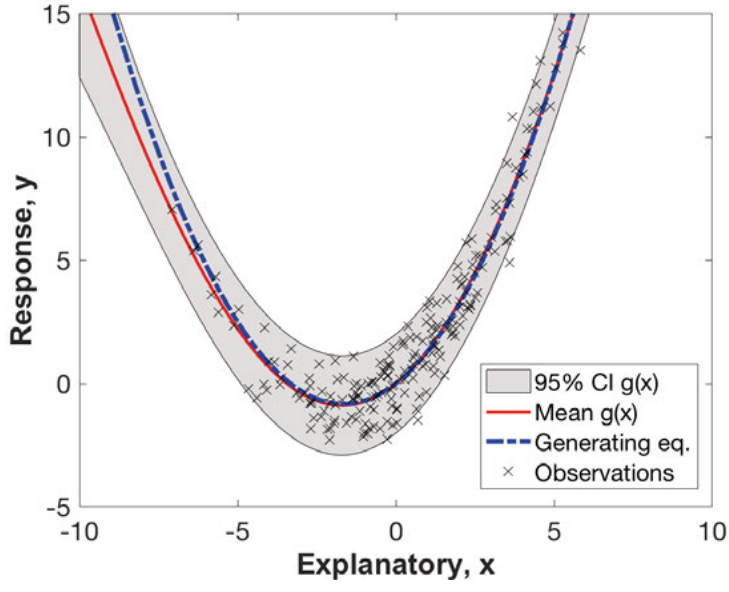
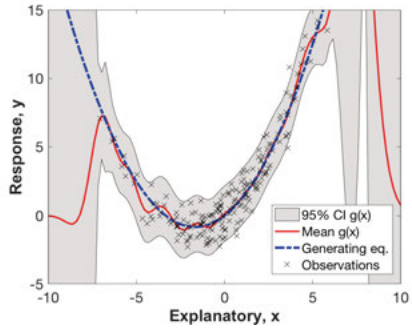
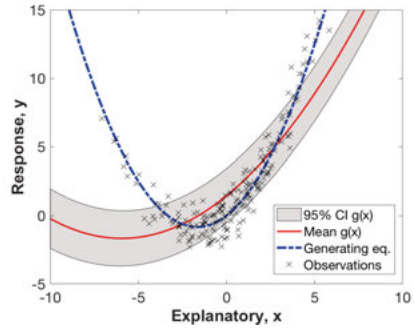


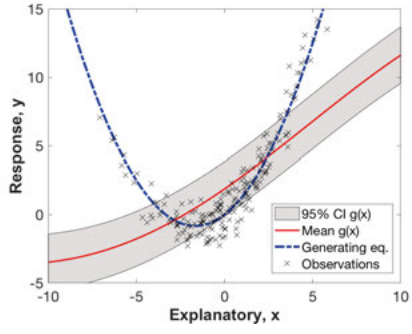
Figure 2.3. **Model fit of squared exponential covariance function after 200 iterations in the optimization of hyperparameters**. The length-scales that maximizes the model evidence was found to be $\theta = (l, \sigma_g, \sigma_n) = (17.14, 66.01, 0.99)$ and logML $= -299.62$.



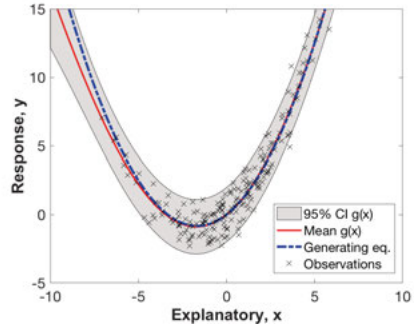
A: $\theta = (1, 66.01, 0.99)$, LogML$= -362.4$



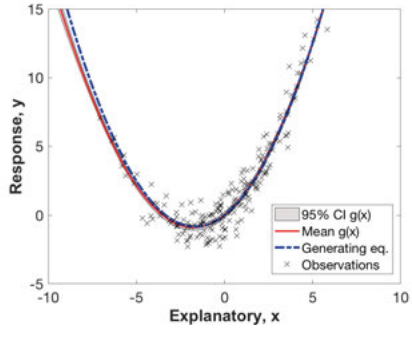
B: $\theta = (100, 66.01, 0.99)$, LogML$= -874.3$



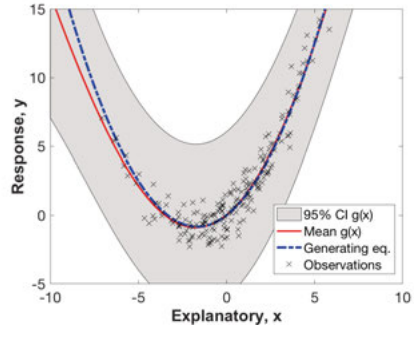
C: $\theta = (17.14, 1, 0.99)$, LogML$= -1199.1$



D: $\theta = (17.14, 100, 0.99)$, LogML$= -300.2$



E: $\theta = (17.14, 66.01, 0.1)$, LogML$= -9514.6$



F: $\theta = (17.14, 66.01, 3)$, LogML$= -428.2$

Figure 2.5. **The effect of changing the hyperparameters.** In all of these figures, we show what happens if one of the three hyperparameters are changed from the optimized hyperparameter values. In Fig. A, the length-scale is set to be 1 (short), and in Fig. B, the length-scale is set to be 100 (long). In Fig. C the signal variance is set to be small, and in Fig. D, the signal variance is set to be large. In Fig. E, the noise variance is set to be small, and in Fig. 2.4f, the noise variance is set to be large.

the region where 95% of the predictive functions goes within. The model evidence is however not even close to the polynomial Bayesian linear re-

gression with the highest model evidence in 2.1.3 ($LogML = -297.32$). To find parameters that better fits the model to the data, we use the gradient-based optimizer to iteratively update the hyperparameters until the model evidence cease to increase. In table 2.2, I show the updated hyperparameters, with corresponding model evidence, throughout the optimization process. As we can see, after the first iteration, the model evidence is even lower (LogML $= -1122.86$), compared to the initial set of hyperparameters. This is because the optimizer looked in a direction where the signal noise parameter σ_n is too small. After five iterations the model evidence is still low (logML$= -1096$), but has started to increase. After ten iterations the model evidence has got considerably higher (LogML$= 355.40$) — after the characteristic length-scale l and signal variance σ_g^2 have started to increase and the noise variance σ_n^2 went up. After 200 iterations we end up with significantly higher model evidence, LogML$= -299.62$. At this point, the characteristic length-scale is considerably larger, $(l, \sigma_g, \sigma_n) = (17.14, 66.01, 0.99)$. Now we can see in Fig. 2.3 that the model fit is less 'wiggly', due to the longer characteristic length-scale, and can now be more suited for extrapolation. The model evidence is now very close to the polynomial Bayesian linear regression model with the highest model evidence in 2.1.3 (LogML $= -297.32$).

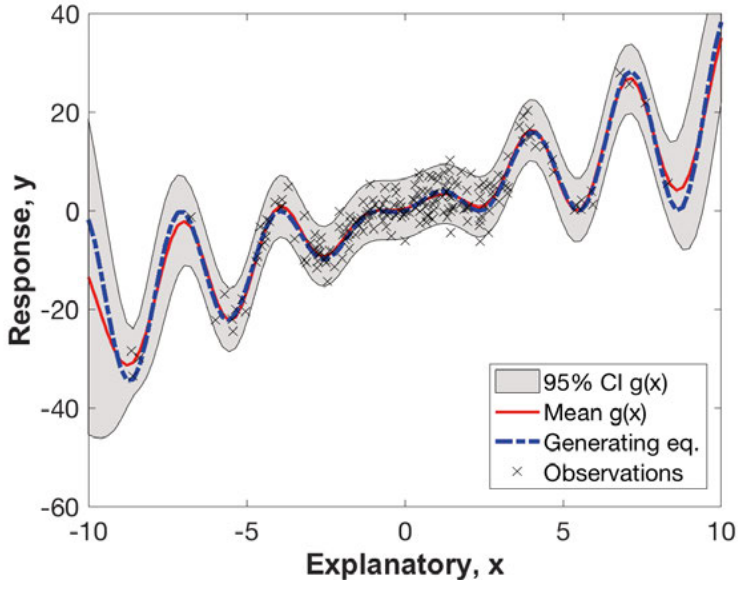


Figure 2.6. **Model fit, using a squared exponential covariance function, of a slightly more complicated generating function** The generating function is $y = 2x + 2x \cdot \sin 2x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 3)$.

To show how the hyperparameters change both the model evidence as well as model fit, we show in Fig. 2.5, what happens if the hyperparameters are altered after the optimization. Fig. 2.4a and Fig. 2.4b, show what happens if the length-scale is altered. Clearly, if the length-scale is short, then there are more 'wiggles' in the prediction function, and the confidence region gets wide faster where there are less data. In Fig. 2.4c and Fig. 2.4d we see what happens if the signal variance is changed. The left figure (Fig. 2.4c) shows that if the signal variance σ_g^2 is set to be small, giving us a function more similar to the mean. Fig. 2.4e and Fig. 2.4f shows what happens if we change the noise variance σ_n^2. The left figure has a much more narrow confidence region than the right figure where the noise variance is set to be large. So the larger the noise variance hyperparameter is, the larger the confidence region gets.

We have now seen that a Gaussian process framework with squared exponential covariance can be used to fit the polynomial generating function with ease. The great advantage with this Gaussian process approach is that it is extremely flexible and can, with good choices of covariance functions, provide us with models with a fantastic model fit, even when the underlying dynamics are complicated. This is visualized in Fig. 2.6, where we have fitted an optimized Gaussian process to the somewhat more complicated generating function, $y = 2x + 2x \cdot \sin 2x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 3)$. We can see that the Gaussian process captures the sine dynamics and provides a good model for predicting the system in regions where there are data available. However, how interpretable would a non-parametric Gaussian process be when we move up in dimensionality and cannot easily visualize our model, and what do the hyperparameters tell us that we can understand? Clearly, there are positive aspects of the non-parametric regression using Gaussian processes when it comes to predictability and model fit, but there are clear limitations when it comes to the interpretability provided compared to Bayesian linear regression models.

3. Main methodological contributions

The overall aim of this thesis is to identify the mechanisms that govern the dynamics of social systems. I want to identify interpretable models that are not too complex but still have much predictability, i.e., perform model selection where we balance between interpretability and predictability.

We have seen that we can use Bayesian inference to perform model selection using two different, but connected frameworks - Bayesian linear regression and Gaussian processes. Both frameworks have their advantages, and how they can be used and interplay is more or less what this thesis is about. Remember that, Bayesian linear regression is a particular case of Gaussian processes where we have a chosen a dot product covariance function built up by the choices of basis functions. The fact that we specify the basis functions (terms) in a Bayesian linear regression model either subjectively, from existing theory, or by testing a vast number of potential terms, i.e. models — linear, non-linear and interactional — to find a well-fitting model, means that we might not find it because it wasn't in the set of chosen basis functions. This gets more and more complicated, the higher dimensional our data becomes. This makes us want to lean towards a non-parametric Gaussian process model with more flexible covariance function — but this approach is not as interpretable and easy to communicate as explicit functions. So, how do we balance between these two approaches?

In this thesis, there are two main methodology contributions. The first contribution is an improved model selection methodology, based on Bayesian linear regression, built to study social systems, and this methodology is presented in paper I. In this paper, we identify the mechanisms in the relationships in the interactions between two indicator variables (economic development and democracy) by testing a vast number of models and then comparing their model evidence to find the model that best explains the mechanistic development. We do this by fitting different polynomial models $\mathcal{M}_i$, i.e., different polynomials $f_i(\mathbf{x})$ of linear, non-linear and interactional terms of the indicator variables $\mathbf{x}$, to the change $dy(\mathbf{x})$ between two times in the response variable, i.e.,

$$dy(\mathbf{x}) = f_i(\mathbf{x}) + \epsilon \tag{3.1}$$

The main differences to existing models are that our choice of normal inverse gamma NIG priors (normal prior on the slope parameters β, and

inverse gamma distribution on the noise in the model) we obtain faster computations that are analytically tractable. This allows us to compare all potential models and thereby check the robustness of the best models by investigating how often specific terms appear among the models with the highest model evidence. This is a leap forward compared to the previous methodology presented in [12], where only the one model for each number of terms was chosen to calculate the model evidence. This choice was made because it is computationally expensive to calculate the marginal likelihood of a uniform prior to the parameters β. In this paper, we also benchmark our models to more predictive-oriented feed-forward artificial neural networks (ANNs) [42] and Bayesian model averaging [43]. In paper I, we use methods of Bayesian model averaging and neural networks to approximate the dynamics of the system. These methods are intended to provide a more flexible estimate than the parametric models.

The second main contribution is that I incorporate and change focus towards Gaussian processes. The transition to Gaussian processes was primarily made to overcome the four main shortcomings of the Bayesian linear regression approach; That we are (1) limited to either a few variables or a low number of potential terms in the basis functions. (2) We have to choose the basis function ourselves and are thereby not that flexible, (3) that we do not get measures of the relative importance of the different indicator variables and (4) that it does not provide benchmark models.

Gaussian process models and their opportunities with flexible covariance functions are a good choice because they do not have to be limited to any polynomial form - still compensating for overfitting. We use Gaussian processes with squared exponential (SE) covariance functions as benchmark models to provide 'lower' (where some variables are omitted) and 'upper' (using all variables) bounds for polynomial models. Fitting SE models, where the input variables have been normalized, and using automatic relevance detection (ARD) for the covariance function provides us with a guiding 'measure' on how relevant the variables are. To overcome the tradeoff between interpretability and predictability, we developed a new semi-parametric modelling approach. Semi-parametric models have both a parametric and a non-parametric component and have been studied previously by, e.g., [44], however, the methods were intended to model the residuals of the polynomial model with the GP, not for model selection, which is our aim (more on this in paper III). A semi-parametric approach allows us to combine high interpretability by fitting the model with both a parametric component and a non-parametric component for predictability. This is studied in this thesis and is probably where the most significant contribution lies.

In the semi-parametric modelling approach, we model the response dy using the following setup,

$$dy(\mathbf{x}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z}) + \epsilon \tag{3.2}$$

where we have chosen to model the dependency in certain explanatory variables $\mathbf{x}$, using Bayesian linear regression $f(\mathbf{x})$, to allow for interpretability, and then control for confounding variables $\mathbf{z}$ using a non-linear Gaussian process $g(\mathbf{z})$. In applications, we can add or remove any variables in $\mathbf{x}$ and $\mathbf{z}$, depending on what we wish to investigate. To accomplish this, we combine two Gaussian processes $f(x)$ and $g(\mathbf{z})$. So the new Gaussian process is then,

$$y(\mathbf{x}, \mathbf{z}) \sim \mathcal{GP}(0, k_{\text{dot}}(\mathbf{x}, \mathbf{x}') + k_{\text{SE}}(\mathbf{z}, \mathbf{z}') + \sigma_n^2 \delta(\mathbf{x}, \mathbf{x}')) \tag{3.3}$$

and is built out of the GPs,

$$\begin{aligned} f(\mathbf{x}) \sim &\quad \mathcal{GP}(0, k_{\text{dot}}(\mathbf{x}, \mathbf{x}')) \\ g(\mathbf{z}) \sim &\quad \mathcal{GP}\big(0, k_{\text{SE}}(\mathbf{z}, \mathbf{z}')\big) \end{aligned} \tag{3.4}$$

where the covariance function of the square exponential component is,

$$k_{\text{SE}}(\mathbf{z}, \mathbf{z}') = \quad \sigma^2 \exp\left(-\sum_{i=1}^{D_*} \frac{(z_i - z_i')^2}{2 l_i^2} \right) \tag{3.5}$$

and the Bayesian linear regression (parametric) component, has the covariance function,

$$k_{\text{dot}}(\mathbf{x}, \mathbf{x}') = \mathbf{b}(\mathbf{x})^\top V \mathbf{b}(\mathbf{x}') \tag{3.6}$$

Where $\mathbf{b}(\mathbf{x})$ is a set of predefined basis functions consisting of linear, nonlinear, and interactional polynomials of $\mathbf{x}$ and we have assumed that $\beta \sim \mathcal{N}(0, V)$ is a parameter with mean 0 and variance V are the weights of the parameters.

To be able to test a large set of potential polynomial models $f(\mathbf{x})$, to see which one best fits to the data, we first optimize the hyperparameters to another model, namely,

$$dy(\mathbf{x}, \mathbf{z}) = g_1(\mathbf{x}) + g_2(\mathbf{z}) + \epsilon \tag{3.7}$$

where both $g_1(\mathbf{x})$ and $g_2(\mathbf{z})$ are Gaussian processes with squared exponential covariance functions. After we have found the hyperparameters that produce the model with the highest model evidence, we fix the hyperparameters of $g_2(\mathbf{z})$ and first remove and then approximate $g_1(\mathbf{x})$ with all potential models $f_i(\mathbf{x})$, and then picking the one with the highest model evidence. By this procedure, we thereby can approximate the very flexible $g_1(\mathbf{x})$ with the best fitting polynomial model $f(\mathbf{x})$.

In Paper II, we present a early version of the estimation of the hyperparameters for the parameters of $g_2(\mathbf{z})$ where we first fit the Gaussian process $g(\mathbf{x}, \mathbf{z})$ and then picking the hyperparameters corresponding to $\mathbf{z}$ to build an approximation of $g_2(\mathbf{z})$. The updated approach is used in papers III and IV. For more details about how to fit the hyperparameters and how to then perform model selection for the parametric component, please read these papers.

4. Paper summary

This chapter contains short summaries of the four papers in this thesis. The first paper explains the main contributions and findings in the first methodology, based on Bayesian linear regression. The second to the fourth papers is about the methodology introduced in this thesis, based on a Gaussian process framework. The reader is referred to the papers themselves for details of the findings in the papers.

4.1 Paper I

Björn R. H. Blomqvist, Richard P. Mann & David J. T. Sumpter
Using Bayesian dynamical systems, model averaging and neural networks to determine interactions between socio-economic indicators, PLoS ONE 13(5): e0196355. https://doi.org/ 10.1371/journal.pone.0196355 (2018).

Understanding the social and economic development of the nations of the world is of high importance if we want to to make useful predictions about how the world will develop in the future. The well-studied relationship between economic development (GDP per capita) and level of democracy in governments is of particular interest. Are democratic reforms a step towards economic growth or is democracy just a symptom of an economically highly developed society?

In this paper, we present a data-driven methodology that use large sets of longitudinal data to inform and build up a dynamical system model of the relationship between GDP per capita and a democracy index. We use a Bayesian methodology that allows us to find a dynamical relationship between indicator variables by identifying the non-linear and interaction functions that best describe the interactions between the indicator variables, out of a large set of potential polynomial models. Having too many terms in our polynomial models can provide good explanation power but makes the models harder to interpret. We use the built-in feature in the Bayesian framework to perform model selection by punishing overcomplicated functional forms (with too many additional terms) to obtain a good trade-off between explanatory power and interpretability. We use conjugate priors to allow fast and tractable computations. This choice of prior both makes it possible to perform inference on a broader

set of good models, it also and makes robustness checks easier. We also benchmark our methodology with two additional methods, namely model averaging (the average prediction of multiple models) and artificial neural networks. We find that our methodology performs similarly to the benchmark modelling approaches when it comes to predictability, but provide far better when it comes to interpretability. We find no robust model for economic development, but for democracy, we find that a long-term democratic development increases after the economic situation gets better.

My contribution:
I developed much of the methodology, analyzed the data and was responsible for writing the paper.

4.2 Paper II

Björn R. H. Blomqvist, David J.T. Sumpter & Richard P. Mann
Explaining the rise of a radical right-wing party using Gaussian processes
Submitted (2018).

Throughout western Europe, radical right-wing parties with an anti-immigration agenda have increased in electoral support in the last decades. Sweden is no exception with 12.86% of the electorate supporting the radical right-wing party the “Sweden Democrats", in the 2014 election and was consistently polling as the second or third largest party before the Swedish election in September 2018. Understanding the underlying mechanisms behind this radical transition of the political landscape of Sweden - where this change has been exceptional rapid - is not only exciting and essential for the academic discourse, it is also of broad interest for the population of Sweden. In this manuscript, we analyze the rise of the Swedish radical right-wing populist party from the first time its support could be measured in national elections in 2002 until the election in 2014 using a new Gaussian process framework. Our modeling approach combines the interpretability of standard regression with the flexibility and precise model fitting capabilities and flexibility of Gaussian processes. More specifically, we use a Gaussian process framework to be able to perform Bayesian linear regression to identify the best explicit model in some explanatory variables, the variables we wish to investigate in detail (e.g., education levels, and unemployment) how they relate to change in the electoral support, together with a fully non-parametric statistical control for confounding variables, providing more than twice the model evidence compared to traditional methods. This combination

of interpretability and the precise model fit is allowed by smart choices of covariance functions in the Gaussian process setup. We also present multiple statistical ways to obtain a guiding measure the relevance of different variables adequately. We use this new methodology to test the two prevailing theories widely used to explain radical right-wing support: (1) the ethnic conflict theory and (2) social marginality theory and find results conflicting with the current body of knowledge. We found no significant support for the ethnic conflict theory, i.e., we find no significant correlation between immigrant density in the different municipalities and radical right-wing support on the municipality level. The only socioeconomic variable we found to be significant was education level. Instead, we found that the spatial location (longitude and latitude), population density and time (together with education levels) to be the only significant variables and all other effects such as unemployment levels, median income, mean age, and criminality levels are found to be redundant.

This paper was written and submitted in February 2018, eight months before the Swedish election - with some small corrections later afterwards in the review process of this thesis. In the Swedish election in September 2018, the “Sweden Democrats" got 17.53% of the votes.

My contribution:
I developed the methodology together with Richard P. Mann, analyzed the data and was responsible for writing the paper.

4.3 Paper III

Björn R. H. Blomqvist
Model selection and parameter estimation of complex dynamical systems using semi-parametric Gaussian processes
Submitted (2018).

In this paper, I show the capability of the methodology developed in paper II using synthetic longitudinal data from two dynamical systems — the first is constructed to imitate the dynamics found in social systems and the second system is a chaotic Lorenz-type system. We compare how the three different Gaussian process approaches: (1) parametric, (2) non-parametric and (3) semi-parametric can be used to capture the dynamics of the systems. We investigate the circumstances in which each is advantageous when it comes to interpretability (understandable models) and predictability (model fit). We show that the parametric approach gives interpretable models, but have some limitations. They are computational heavy if we wish to test all combinations of potential

model for many variables and terms in the models. Also, If the relevant variables are not included in the set of tested models, they can result in the selection of misleading models. Non-parametric models provide more flexibility and predictability, but they often lack interpretability. Despite this limitation, a method where variables are omitted from the model one at a time can be used to provide a guiding measure of the relevance of the different variables. The semi-parametric model provides a combination of the interpretability of the parametric approach and a model fit approximately equal to that of the non-parametric approach. Our semi-parametric Gaussian process approach accurately captures the dynamics of both dynamical systems, even in the presence of noisy data and chaos, and allows us to identify the correct model even when the model is high dimensional.

4.4 Paper IV

Ranjula Bali Swain, Björn R. H. Blomqvist & David J. T. Sumpter
Last night in Sweden? Using Gaussian processes to study changing demographics at the level of municipalities
Submitted (2018).

In this manuscript, we analyse the relation between immigration proportions in the population to three different types of reported criminality — rape, burglary, and assault — using Swedish municipality level data from the year 2002 until 2014, years where there have been significant changes in these crime types. More specifically, we use a semi-parametric Gaussian process framework to perform Bayesian linear regression to find the best explicit model in explanatory variables, together with a fully non-parametric statistical control for confounding variables. This combination of interpretability and the flexible model fit is allowed by smart choices of covariance functions in our semi-parametric Gaussian process setup. We also present multiple statistical ways to obtain guiding measures of the relevance of different variables for explaining the changes in criminality.

We find that foreign-born percentage in a municipality has no significant connection to changes in reported rape. Instead, there are clear signs that the substantial increase in reported rape is due to new and modern laws and regulations. Reported burglary rates have decreased over the period and are slightly higher in municipalities with high immigration levels. Assault rates are primarily self-limiting but are positively correlated to the proportion of foreign-born residents in a municipality. The found positive correlation to reported burglary and assault does not necessarily tell us that the foreign-born population is committing more

crimes than the native-born population. One thing that we can state with more confidence is that there is an inequality in exposure to some types of criminality between the foreign-born population and the native-born population, where the foreign-born population is more exposed to burglary and assault than native Swedes.

My contribution:
I developed the methods, analyzed the data and participated in writing the paper.

5. Errata and comments on paper I

In this chapter, I will correct some of the errors pointed out in the review process of this thesis. Most of the typos and comments on the papers in this thesis are fixed, in the papers themselves, however for Paper I, that is already published, I will present corrections in this chapter instead.

In paper I, the following typos needs to be adressed:

- In equation (17) there is a typo in the paper that does not effect any calculations nor subsequant analysis. In the paper, we incorrectly write the equation of the coefficient of determination R^2 to be,

$$R^2 = 1 - \frac{\sum_i (f_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}. \tag{5.1}$$

  The correct way of writing this equation is,

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}. \tag{5.2}$$

- The last sentence of section 2.5 (Surrogate data testing) is grammatically incorrect. Instead of being written as: “the changes the investigated indicator variables, producing data for an additional 25 time-steps.", it should read: “changes in the investigated indicator variables, producing data for an additional 25 time-steps."
- In table 2, I misspelled “Percent". So the “Procent" (Swedish) should be “Percent".
- In table 2, we write: “furthest to the left shows the most eight most frequently", instead it should read: “furthest to the left shows the eight most frequently".

There are also some things that needs be commented:

- In this paper we use terms of the form $\frac{1}{1+x_i}$ in the regression and are generalised with both interactions and powers. The variables x_i are scaled to lie within the range, $0 < x_i < 1$, making the function gently decreasing. This choice is made to avoid the singularity and dramatic effects as $x_i \to 0$. We have no reason to assume that the behaviour should be extreme when x_i is close to zero, and we wanted the terms to be more symmetric.

- The Bayesian priors in our regression model are indeed chosen to provide a 'trick' to reduce collinearity, but we also want our priors to reflect that we want the data to inform us about the best models without influencing it too much, as an exploratory tool. It is also why we choosed a zero mean for β.

6. Sammanfattning på svenska

I denna avhandling både utvecklas och tillämpas nya statistiska modeller för att förklara och förutsäga hur sociala system förändras över tid. Det som menas med ett socialt system är relationerna inbördes mellan olika individer eller grupper av människor. Att studera hur sådana sociala system utvecklas kan göras genom att analysera förändringen över tid i olika tillståndsvariabler som till exempel, inkomst, utbildningsnivå, medellivslängd, hälsotal. Sociala system karaktäriseras ofta av komplexa och ickelinjära dynamiska interaktioner. Ett klassiskt exempel, som jag även studerar i denna avhandling, är hur ekonomisk utveckling och demokratiutveckling relaterar till varandra. Är det så att det krävs en stark ekonomisk utveckling innan samhällen utvecklas demokratiskt eller är det tvärtom? Sker utvecklingen linjärt eller är dynamiken ännu mer komplicerad?

Traditionella statistiska metoder, såsom linjär regression, är fortfarande det sätt som väldigt många använder sig av för att studera sociala system. Huvudanledningen till detta är att dessa metoder ger väldigt enkla och tydliga resultat. Under de senaste decennierna har väldigt stora datamängder av tillståndsvariabler gjorts tillgängliga för både forskare och allmänheten. Detta tillsammans med nya statistiska metoder har gjort att vi kan studera dessa system på helt nya sätt och få djupare förståelse om världen omkring oss. Det brukar kunna vara ett problem att dessa nya moderna statistiska metoder kan uppfattas lite som 'svarta lådor', där stora datamängder stoppas in, och ut kommer prognoser och samband som det inte alltid är så lätt att förstå hur den 'svarta lådan' lyckats producera.

I denna avhandling presenterar jag två olika statistiska metoder för att analysera sociala system, inspirerade av maskininlärning, men som producerar lättförståeliga funktionsuttryck. Båda dessa metoder bygger på Bayesiansk statistik och låter stora datamängder helt bestämma och forma de resulterande funktionsuttrycken som beskriver de studerade systemen. På så sätt kan dessa metoder användas för att testa befintliga teorier eller för att ge nya insikter som kan användas för att bygga upp nya teorier.

Den första metoden bygger på Bayesiansk linjär regression. Denna metod testar ett stort antal linjära och ickelinjära funktionsuttryck för att undersöka vilken av dessa modeller som passar bäst för att beskriva dynamiken i systemet. Den andra metoden utvecklades för att kunna

hantera några av bristerna i den första parametriska metoden, som att den till exempel blev väldigt beräkningsintensiv när vi ville inkludera många variabler (fler än 4-5 stycken) eller tillåta många potentiella funktionsuttryck. Den nya metoden som bygger på Gaussiska processer och låter oss kombinera det lättförståeliga med parametriska funktionsuttryck med flexibiliteten hos icke-parametriska regressionsmetoder.

Denna avhandling är en sammanslagning av fyra delarbeten. I artikel I introducerar vi den första metodiken, Bayesiansk parametrisk regression, och tillämpar den genom att studera hur ekonomisk utveckling och demokratiutveckling interagerar med hjälp av longitudinell data från hela världen. Vi kan visa att det är först efter att den ekonomiska utvecklingen gått framåt som det kan bli en långsiktig demokratisk utveckling. I artikel II introducerar vi den andra metodiken, som bygger på Gaussiska processer, för första gången och applicerar den genom att studera vad som kännetecknat de regioner i Sverige där det nationalistiska och socialkonservativa politiska partiet Sverigedemokraterna vuxit fram starkast mellan åren 2002 och 2014. I artikel III testar jag tre olika sätt — parametriskt, icke-parametriskt och semi-parametriskt — att använda sig av Gaussiska processer. Detta genom att studera två komplexa dynamiska system av syntetisk data. I denna artiklel visar jag styrkorna och svagheterna för dessa olika angreppssätt. I artikel IV används Gaussiska processer ytterligare en gång för att studera, det i media mycket debatterade ämnet, vad det är som orsakat förändringarna i brottstatistiken de senaste åren i Sverige.

7. Acknowledgements

David, you are the main reason behind why I made it all the way to the end of this long and challenging journey. Without your guidance, encouragement, support, positivity, and confidence in me, this would not have been possible. Thank you for encouraging me to explore new ideas as well as teaching me when to stop. I will miss our daily lunches at Rullan and all your life advice. You have been the best academic supervisor I could ever wish for, but most of all, a fantastic friend. I will always be grateful for everything you have done for me.

Ranjula, I have so many things to thank you for. Working together with you is always fun and inspiring. I am amazed by how full of energy and how supportive you are. Every single time I left your office, I felt completely reenergized. Thank you for all of your patience and your trust in me — I will never forget it.

Richard, the time I spent visiting you in Leeds in 2017 was, without any doubt, the most fun I had during all my years of studies and was a turning point for me as a graduate student. Not only did you introduce me to Gaussian processes, but you have also been a fantastic role model. It has been incredibly inspiring to work with such a fantastic researcher and friend. Because of you, your family, and your friends, Leeds will always have a special place in my heart. I will never ever forget your mild reaction when 'water' started dripping from the ceiling during one of many amazing nights at the Victoria hotel. Best wishes to you, Viktoria, and Gabriel.

Alex, thank you for being the best friend I could ever wish for! I can't imagine my time here in Uppsala without you! You have always been there to give support and to listen to all of my ideas and problems - big and small. Thank you for being amazingly kind and nonjudgmental, for all your feedback on my manuscripts, and all the late night jam sessions.

I want to thank every single one in David's research group over the years. In particular: Teddy, thank you for always being so incredibly fun to be around and your hospitality. Ernest, thank you for being so funny and interesting to talk to. Arianna, thanks for 'tricking' me into taking tango lessons and for helping me with proofreading. Sebastian, thanks for your hospitality and for always being friendly and polite. Emri, thank you for being my close friend and for all the adventures we have been on

so far — hopefully, you will come back soon for more moose and reindeer spotting up in Vilhelmina. I also want to give my thanks to all the other Ph.D. students and supporting staff at the Mathematics department. I especially want to thank you, Erik, for being my good friend and for always being a few steps ahead of me in Squash. Jakob, thank you for always helping me find motivation and for your kindness. I also want to thank you, Cecilia, Olow, Ove, Filipe, Kostas, Tilo, Benny, Anna, and Martin for your friendship and support over the years.

I also want to thank all my friends from the time before I started this Ph.D. Thank you, John, Jens, Samuel, and Marcus for providing a 'slipstream' when I needed it. Thank you, Sebastian, Jonathan, and Erik for being fantastic friends during my years studying in Stockholm. I also want to give a special thanks to my two oldest friends, Erik, and Klas for always being there for me and for all the big dreams.

I want to thank the School of Mathematics, University of Leeds, for giving me the opportunity to come on a research visit in 2017. It was a truly inspiring visit that helped me a lot in the development of this thesis. Thanks for all the interesting seminars and for the excellent research and work environment.

Mest av allt vill jag tacka min familj och släkt hemma i Umeå. Tack mamma och pappa för att ni alltid tar er tid och stöttar mig oavsett vad jag tar mig för. Tack att ni gett mig världens bästa syskon, Hannes och Lisa, och för tryggheten att våga prova på nya saker.

Björn Blomqvist
Uppsala, November 2018

8. References

[1] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452, 2015.

[2] Petter Holme and Fredrik Liljeros. Mechanistic models in computational social science. *Frontiers in Physics*, 3:78, 2015.

[3] Wesley C Salmon. *Causality and explanation*. Oxford University Press, 1998.

[4] Peter Hedström and Petri Ylikoski. Causal mechanisms in the social sciences. *Annual review of sociology*, 36, 2010.

[5] Andrew Sayer. *Realism and social science*. Sage, 1999.

[6] Peter Hedström, Richard Swedberg, and Gudmund Hernes. *Social mechanisms: An analytical approach to social theory*. Cambridge University Press, 1998.

[7] Stuart Glennan. Rethinking mechanistic explanation. *Philosophy of science*, 69(S3):S342–S353, 2002.

[8] Peter Hedström. Dissecting the social: On the principles of analytical sociology. 2005.

[9] Adam Mann. Core concept: Computational social science. *Proceedings of the National Academy of Sciences*, 113(3):468–470, 2016.

[10] Marc Keuschnigg, Niclas Lovsjö, and Peter Hedström. Analytical sociology and computational social science. *Journal of Computational Social Science*, 1(1):3–14, 2018.

[11] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[12] Shyam Ranganathan, Viktoria Spaiser, Richard P Mann, and David JT Sumpter. Bayesian dynamical systems modelling in the social sciences. *PloS one*, 9(1):e86468, 2014.

[13] Viktoria Spaiser, Shyam Ranganathan, Richard P Mann, and David JT Sumpter. The dynamics of democracy, development and cultural values. *PloS one*, 9(6):e97856, 2014.

[14] Shyam Ranganathan, Stamatios C Nicolis, Viktoria Spaiser, and David JT Sumpter. Understanding democracy and development traps using a data-driven approach. *Big data*, 3(1):22–33, 2015.

[15] Viktoria Spaiser, Shyam Ranganathan, Ranjula Bali Swain, and David JT Sumpter. The sustainable development oxymoron: quantifying and modelling the incompatibility of sustainable development goals. *International Journal of Sustainable Development & World Ecology*, 24(6):457–470, 2017.

[16] Shyam Ranganathan, Stamatios C Nicolis, Ranjula Bali Swain, and David JT Sumpter. Setting development goals using stochastic dynamical system models. *PloS one*, 12(2):e0171560, 2017.

[17] Viktoria Spaiser, Peter Hedström, Shyam Ranganathan, Kim Jansson, Monica K Nordvik, and David JT Sumpter. Identifying complex dynamics in social systems: A new methodological approach applied to study school segregation. *Sociological Methods & Research*, 47(2):103–135, 2018.
[18] Carl Edward Rasmussen and Cristopher K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.
[19] William M Bolstad and James M Curran. *Introduction to Bayesian statistics*. John Wiley & Sons, 2016.
[20] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.
[21] Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.
[22] James O Berger and José M Bernardo. Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association*, 84(405):200–207, 1989.
[23] James O Berger, José M Bernardo, Dongchu Sun, et al. The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938, 2009.
[24] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.
[25] Stefan Van Dongen. Prior specification in bayesian statistics: three cautionary tales. *Journal of Theoretical Biology*, 242(1):90–100, 2006.
[26] Richard Mann. *Prediction of homing pigeon flight paths using Gaussian processes*. PhD thesis, University of Oxford, 2010.
[27] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for regression. In *Advances in neural information processing systems*, pages 514–520, 1996.
[28] David JC MacKay. Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998.
[29] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.
[30] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
[31] Christian Robert. Machine learning, a probabilistic perspective, 2014.
[32] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.
[33] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.
[34] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.
[35] David S Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.
[36] George C Tiao and George EP Box. Modeling multiple time series with applications. *journal of the American Statistical Association*,

76(376):802–816, 1981.
[37] Noel Cressie. *Statistics for spatial data*. Wiley, 1993.
[38] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.
[39] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.
[40] Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897, 2013.
[41] Carl Edward Rasmussen and Hannes Nickisch. GPML Toolbox: Matlab software version 4.1.
[42] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
[43] Adrian E Raftery, David Madigan, and Jennifer A Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.
[44] BJN Blight and L Ott. A bayesian approach to model inadequacy for polynomial regression. *Biometrika*, 62(1):79–88, 1975.
[45] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.