Olli-Pekka Kahilakoski

Bayesian Regression Analysis of Sickness Absence

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 18.5.2011

Thesis supervisor:

Prof. Jouko Lampinen

Thesis instructor:

D.Sc. (Tech.) Aki Vehtari

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Olli-Pekka Kahilakoski

Title: Bayesian Regression Analysis of Sickness Absence

Date: 18.5.2011 Language: English Number of pages:8+48

Department of Biomedical Engineering and Computational Science

Professorship: Computational and Cognitive Biosciences Code: S-114

Supervisor: Prof. Jouko Lampinen

Instructor: D.Sc. (Tech.) Aki Vehtari

Individual factors associated with sickness absence have previously been studied with generalized linear models. Using Bayesian methods, we compare generalized linear models to Gaussian process models, which are flexible non-linear regression models that allow local changes in the response surface structure. We find Gaussian process models superior for predicting sickness absence with health questionnaire data in a sample of employees of a Finnish company.
We also do variable selection for Gaussian process models using Bayesian multiple comparisons. In agreement with previous studies, we find that depression and pain-related impairment at work are associated with increased sickness absence, with a possible saturation effect for depression.

Keywords: Bayesian, regression, modeling, Gaussian process,
generalized linear model, sickness absence

AALTO-YLIOPISTO
PERUSTIETEIDEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Olli-Pekka Kahilakoski

Työn nimi: Sairauspoissaolojen bayesilainen regressioanalyysi

Päivämäärä: 18.5.2011 Kieli: Englanti Sivumäärä:8+48

Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Professuuri: Laskennallinen ja kognitiivinen biotiede Koodi: S-114

Valvoja: Prof. Jouko Lampinen

Ohjaaja: TkT Aki Vehtari

Gaussiset prosessit ovat epälineaarisia regressiomalleja, joilla voidaan mallintaa paikallisia muutoksia vastepinnan rakenteessa. Sairauspoissaoloihin yhteydessä olevia yksilötekijöitä on aiemmin tutkittu yleistetyillä lineaarimalleilla. Vertaamme tässä työssä gaussisia prosesseja yleistettyihin lineaarimalleihin bayesilaisilla menetelmillä ja havaitsemme, että gaussiset prosessit ennustavat yleistettyjä lineaarimalleja paremmin sairauspoissaoloja terveyskyselyn avulla.
Teemme myös muuttujanvalinnan gaussisille prosesseille bayesilaisella monivertailumenetelmällä ja havaitsemme, että masennuksella ja kivun aiheuttamalla työhaitalla on yhteys sairauspoissaoloihin. Tulokset ovat linjassa aiempien tutkimusten kanssa. Lisäksi havaitsemme masennuksella ja sairauspoissaoloilla mahdollisen epälineaarisen, saturoituvan yhteyden.

Avainsanat: bayesilainen, regressio, mallintaminen,
gaussinen prosessi, yleistetty lineaarimalli, sairauspoissaolo

Preface

This work was carried out in the Department of Biomedical Engineering and Computational Science at the Helsinki University of Technology.

I would like to thank Doc. Aki Vehtari for guidance, Lic. Tech. Karita Reijonsaari, Prof. Simo Taimela, and the supervising Prof. Jouko Lampinen. Thank you also to my family and friends!

Otaniemi, 16.5.2011

Olli-Pekka Kahilakoski

Contents

Symbols and abbreviations

Symbols

β	regression coefficient
Bernoulli	Bernoulli distribution
ϵ	error term
E	expected value
f	latent function
λ	rate parameter for Poisson or negative binomial distribution
μ	mean of normal distribution
n	the number of observations
N	normal distribution
NB	negative binomial distribution
NB_0	zero-truncated negative binomial distribution
Poisson	Poisson distribution
Poisson_0	zero-truncated Poisson distribution
$p(x)$	probability distribution function
$p(x, y)$	joint probability distribution function
$p(y\|x)$	conditional distribution function
$P(X = k)$	probability mass function
θ	model parameters (also, dispersion parameter)
$\hat{\theta}$	estimator for model parameters
σ	standard deviation for normal distribution
Sd	standard deviation
$U(p(x), x)$	utility function
X	random variable
x	observed predictors
x_*	test datapoint
y	observed response

Abbreviations

AIC	Akaike information criterion
CV	cross-validation
GLM	generalized linear model
MAP	maximum a posteriori
MLE	maximum likelihood estimation
MSE	mean squared error

1 Introduction

The estimated cost of sickness absences to Finnish society is billions of euros annually. Therefore, it is important to study factors that affect the propensity for sickness absence and the lengths of the sickness absence periods. If the factors can be modified, e.g., workplace environment, individual companies and the society at large can obtain considerable savings.

In this Thesis, we use Bayesian regression analysis to study individual factors associated with sickness absence. With regression analysis, statistical relationships between two or more variables are inferred, e.g., the relationship between gender and sickness absence.

Using only regression analysis is usually not sufficient for making causal inferences, but it is the first step in identifying the factors that play a role in sickness absence.

The present study belongs to a larger study, which investigates the effects of a physical activity intervention on a large group of employees. Sickness absence is one of the primary outcomes studied. In this Thesis, we study sickness absence at baseline– during the year before the intervention study–which also helps to identify factors that have changed if the intervention is found to have an effect on sickness absence.

We also use some recent developments in statistical analysis, namely, Gaussian processes for assessing non-linear relationships between variables, cross-validation for estimating predictive performance of a model, Bayesian multiple comparisons for variable selection, and average predictive comparisons for presentation of results.

The structure of the Thesis is as follows:

In Section 2, we present the statistical tools that are needed in later Sections, including an introduction to Bayesian analysis.

In Sections 3, 4, and 5, various regression models are presented. Section 3 introduces the linear model, which is the simplest and most studied regression model. Section 4 presents the generalized linear models, which overcome some limitations of the linear model. In Section 5, we introduce Gaussian process models, which are flexible, non-linear regression models.

In Sections 6 and 7, we introduce methods for selecting a regression model, selecting the variables, and presenting the results of a regression analysis.

In Section 8, we consider data pre-processing, more specifically, recoding, handling missing data, and standardizing data.

Finally, in Section 9, we use methods from the previous Sections to study factors associated with sickness absence. We also compare generalized linear models to Gaussian process models and find Gaussian process models superior for predicting sickness absence.

After comparing models, we do variable selection for the best Gaussian process model and find that depression and pain-related impairment at work are associated with increased sickness absence, with a possible saturation effect for depression, which however requires further study to confirm.

In Section 10, we discuss the results and point out directions for future research.

2 Statistical modeling

In this Section, we introduce the statistical tools for building regression models. The topics covered here are presented in more depth by, e.g., Gelman et al. (1995).

2.1 An introduction to statistical modeling

First, we present some key concepts of statistical modeling.

A statistical model describes an aspect of reality in statistical terms. There is usually a family of statistical models to consider. Of these, the model that best agrees with the observations is selected.

The model parameters are used to select a particular model within a model family.

For example, human male height can be modeled using normal distribution. The model family then consists of all normal distributions, and the model parameters, mean μ and standard deviation σ, specify a single distribution within the family.

The observation model relates the observations to the model parameters. In the above example, observed heights come from a normal distribution with a certain mean and standard deviation. In general, we denote the observation model by $y \sim \text{Distr}(\theta)$, where y is the observed value, θ are the model parameters, and *Distr* stands for a generic distribution. Using normal distribution, $y \sim N(\mu, \sigma^2)$. where N refers to normal distribution, and μ and σ^2 are the model parameters.

A random variable is a numeric variable with random value. A **discrete** random variable can have only specific values, whereas a **continuous** random variable, e.g., height, has a range of possible values.

Random variables are often denoted with uppercase or lowercase letter X. In the case of several random variables, subscripting can be used: X_1, X_2, ...

Probability mass function , $P(X = k)$, denotes the probability that random variable X has value k. For continuous random variables, it is called **probability distribution function**.

As an example, a discrete random variable can be defined according to the outcome of a coin-tossing experiment. For instance, $X = 0$ if the outcome is heads and $X = 1$ if the outcome is tails. With a fair coin, both have equal probabilities, i.e., $P(X = 0) = P(X = 1) = 0.5$.

The expected value (or expectation) of a discrete random variable is defined as

$$E[X] = \sum_k kP(X = k), \tag{1}$$

where the sum is taken over all possible values of X. Intuitively, it is the average value of the random variable over an infinite amount of observations.

For a continuous random variable, the sum transforms to an integral,

$$E[X] = \int_x xP(X = x). \tag{2}$$

Standard deviation is a measure for dispersion of the random values around their expected value. To define it, we first define variance,

$$\mathrm{Var}[X] = E[(X - E[X])^2], \tag{3}$$

that is, variance is the expected value of the squared difference between the random value and its expectation.

Standard deviation is defined as the square root of variance,

$$\mathrm{Sd}[X] = \sqrt{E[(X - E[X])^2]}. \tag{4}$$

Variance and standard deviation both measure dispersion, but we prefer to use standard deviation because it can be interpreted on the same scale with X.

Conditioning can be interpreted as adding information to the model. For example, knowing the weight of a person increases information about the person's height, which is formalized as conditioning height on weight.

A prediction is the model's guess of an unobserved outcome. The prediction can be improved by conditioning on observations.

2.2 Bayesian statistics

2.2.1 The Bayesian probability

Traditionally, there has been two schools of thought about the nature of probability.

The **frequentists** define the probability by considering the outcomes of an experiment that is repeatable, at least in principle. For frequentists, the probability of a certain outcome is the limit of the fraction of that outcome among all possible outcomes when the number of repetitions increases.

According to the **Bayesians**, probability measures the degree of belief in a certain outcome, and is therefore fundamentally subjective. Outcomes that are considered more likely are assigned a greater probability and vice versa.

An essential difference between the two schools is in their position toward model parameters. In frequentists' viewpoint, the parameters have a "real", fixed value that is estimated from the observations. Although Bayesians may also consider the parameters fixed in principle, they use probabilities to model the uncertainty associated with the parameters. Consequently, Bayesian inference results in probability

distributions for the parameters and the model predictions, whereas frequentist inference is concerned with **point estimates**, i.e., single values, for the parameters.

Frequentist inference has also methods to assess uncertainty, namely, **confidence intervals**, but interpreting them correctly is awkward and using them in predicting is difficult.

2.2.2 The Bayes' theorem

The **Bayes' theorem** is the workhorse of Bayesian inference. It arises when the joint probability distribution of two random variables is written in two ways,

$$p(x, y) = p(y|x)p(x) \tag{5}$$

and

$$p(x, y) = p(x|y)p(y). \tag{6}$$

Combining these yields

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \tag{7}$$

which is the Bayes' theorem. In Bayesian inference, x in Equation 7 represents a model parameter and y represents an observation. The model parameters are often denoted with θ, so we can write more conventionally

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \tag{8}$$

The Bayes' theorem applies also to multiple model parameters and observations, so it is written more generally as

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}. \tag{9}$$

The Bayes' theorem is used to update knowledge about the model parameters θ, after observing $\mathbf{y}$. The right-hand side in Equation 9 is usually known, and knowledge updating is done by computing $p(\theta|\mathbf{y})$.

Each part of Equation 9 has a specific name and interpretation.

$p(\theta)$
: The **prior probability distribution** (or **prior**) represents the state of knowledge about the model parameters θ prior to the inference.

$p(\mathbf{y}|\theta)$
: The **likelihood function** relates the observations $\mathbf{y}$ to the model parameters θ. In likelihood function, the observations $\mathbf{y}$ are considered fixed and the model parameters θ vary, so it is not a probability distribution. The likelihood function describes the mechanism that generates the observations, so it is also called the observation model.

$p(\theta|\mathbf{y})$

The **posterior probability distribution** (or **posterior**) represents the state of knowledge about θ after the inference. It is a function of the observations $p(\mathbf{y})$. As seen later, the posterior distribution plays a central role in Bayesian inference.

$p(\mathbf{y})$

The **marginal likelihood** has several interpretations, but most importantly, it is a normalization factor that guarantees that the posterior distribution integrates to 1, which is necessary for any (proper) probability distribution.

2.3 Maximum likelihood estimation

The common non-Bayesian way to infer the model parameters is **maximum likelihood estimation** (MLE). In maximum likelihood estimation, the model parameters are chosen to maximize the probability density of the observations,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(\mathbf{y}|\theta). \tag{10}$$

The maximum likelihood estimation yields a point estimate for $\hat{\theta}$, i.e., a single value for the parameters.

2.4 Bayesian parameter estimation

In Bayesian statistics, the posterior distribution of θ, $p(\theta|\mathbf{y})$, contains all available information about the model parameters θ. Therefore, inferences about the model parameters are made via the posterior distribution.

In its simplest, the posterior can be summarized by computing the most probable value of θ,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|\mathbf{y}), \tag{11}$$

called **maximum a posteriori** (MAP) estimate. It is analogous to frequentist maximum likelihood estimation.

Using the posterior distribution, uncertainty about the model parameters can be summarized with **credible intervals** (or Bayesian confidence intervals). They are parameter intervals that contain the real parameter value with a pre-defined probability, e.g., 95%. Credible intervals are analogous to frequentist confidence intervals, but they have a direct probability interpretation.

Other posterior summary statistics include, e.g., posterior mean and standard deviation, calculated as $E[p(\theta|\mathbf{y})]$ and $Sd[p(\theta|\mathbf{y})]$, respectively.

3 Regression analysis

In this Section, we introduce regression analysis in general and present a specific regression model, the linear model. For a more detailed account of general regression analysis, see Gelman & Hill (2007), or the linear model in particular, see, e.g., Bishop (2006).

3.1 The purpose of regression analysis

Regression analysis refers to a set of statistical techniques used to analyze the relationship between a variable of interest and one or several other variables. Regression analysis answers to questions such as: How does height depend on person's age? How does income vary with gender and the level of education?

In the above examples, height and income are **response variables** (or dependent variables), and age, education, and gender are **predictors** (independent variables, explanatory variables, regressors). In short, regression analysis expresses the response variable in terms of the predictors.

A regression analysis is typically based on a set of measured values of the response variables and the corresponding values of the predictors. The units that are measured, e.g., persons, are called **observational units** (or statistical units).

Regression analysis examines a group of observational units, called a **sample** (or the study population), and uses statistical techniques to infer about the average relationship between the response variable and the predictors.

3.2 The linear model

The simplest regression model is the linear model, which assumes that the average relationship between the response variable and each predictor is linear, i.e., that increasing the predictor by a constant changes the response variable by a constant.

A perfectly linear relationship between two variables is rarely observed. For example, the measurement devices may introduce error. To account for different sources of error, the response variable is assumed to contain random variation.

In mathematical terms, the linear model is written as

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon, \tag{12}$$

where y is the value of the response variable, $x_1, \ldots, x_n$ are the values of the n predictors, and ϵ represents the random variation. The multipliers β_i are the model parameters, called **regression coefficients**.

The predictor values $x_1, \ldots, x_n$ are usually treated as constants and ϵ is treated as a random variable. In Gaussian linear models, which are considered in this Section, we assume ϵ to be normally distributed with zero mean,

$$\epsilon \;\sim\; N(\mathrm{E}[y], \sigma^2). \tag{13}$$

The variation in ϵ can be incorporated directly to y, written as

$$\mu = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (14)$$
$$y \sim N(\mu, \sigma^2). \quad (15)$$

Equation 14 determines the expected value of y, denoted by μ, and the observations y are normally distributed around μ with standard deviation σ, according to Equation 15.

3.2.1 Inference with the linear model

The parameters of the linear model are the regression coefficients $\beta_1, \ldots, \beta_n$ and the standard deviation σ, presented in Equations 14 and 15.

Bayesian inference for the model parameters is done by first constructing the likelihood function, which can be done using Equations 14 and 15. However, we omit the details of the Bayesian treatment of the linear model and present the Bayesian linear model instead as a special case of Gaussian process regression, introduced in Section 5.

4 Regression with generalized linear models

Linear models are limited by their modeling assumptions, i.e., a linear relationship between the predictor and the response variable and the normal distribution of the response variable. The generalized linear models are a class of models used to overcome these limitations.

In this Section, we first present a general form for regression models, which includes generalized linear models and Gaussian process models (Section 5) as special cases. We then introduce the Poisson distribution and the associated negative binomial distribution, which can be used instead of the normal distribution in regression models Finally, we present the generalized linear models and some examples of specific models.

4.1 A general form for regression models

Regression analysis is concerned with the relationship between the response variable y and a set of predictors $\mathbf{x}$, with the regression model f specifying the relationship. In abstract terms, this can be written as

$$y = f(\mathbf{x}). \tag{16}$$

In principle, f can be any computable function. However, f is most often a relatively simple function of $\mathbf{x}$. The function f is typically stochastic, i.e., it contains randomness, so consequently y is a random variable.

With stochastic f, the regression model is usually divided into two parts: A part that describes the expectation of y and a part that describes the actual values of y, written as

$$\begin{aligned} \mathrm{E}[y] &= f(\mathbf{x}) && (17)\\ y &\sim \mathrm{distr}(\mathrm{E}[y]), && (18) \end{aligned}$$

where f is a deterministic function of $\mathbf{x}$, and the randomness is introduced by letting y vary around its mean, as specified in Equation 18. *distr* is an arbitrary distribution, e.g., the normal distribution. In this framework, choosing the regression model amounts to choosing the function f and the distribution of y.

4.2 Choosing the distribution of the response variable for count data

Often, the response variable y is not continuous but discrete, i.e., the possible values of y are a set of isolated values, for instance, $0, 1, 2, 3, \ldots$

In particular, if y is the number of events that are observed in a time interval of certain length, the data are called **count data**. Such data is commonly modeled using the Poisson distribution.

4.2.1 The Poisson distribution

As an introduction to Poisson distribution, we model the number of cars that pass a crossing during an hour. If no additional information related to the crossing is included in the model, we say that the data is modeled without predictors, i.e., we model only the distribution of the response variable.

For example, if we observe $y = 30$ cars during the hour, the simplest model proposes that y is a constant, i.e., we observe 30 cars during any hour.

A more complicated model divides the hour into, e.g., $n = 60$ equally spaced slots, each having the length one minute, so that during each time slot, an average of 0.5 cars pass by. The minutes can be modeled independently from each other as having either 0 or 1 cars, so that with $n = 60$ and $y = 30$, both have the probability $p = y/n = 0.5$. The total number of cars during the hour is then the sum of the number of cars for each minute. This results in a distribution for the total number of cars, called the **binomial distribution**. The model yields an average of $np = 60 \cdot 0.5 = 30$ cars/hour, but, compared to the constant model, the number of cars can range from 0 to 60 – although the limiting cases are extremely rare.

A more fine-grained model divides the hour into $n = 3600$ slots of the length one second, so the probability of a car passing by during each second is $p = 30/n = 1/120$. Like the previous model, this yields the average of $np = 30$ cars/hour, but here the number of cars can vary from 0 to 3600.

The process can be continued so that $n \to \infty$ and $p \to 0$. The limiting distribution is called the **Poisson distribution**.

The Poisson distribution is characterized by its **rate parameter**, λ, which is the average number of events during the observation period. In the above example, $\lambda = 30$. If y follows the Poisson distribution, we denote it as

$$y \sim \text{Poisson}(\lambda). \tag{19}$$

The probability that y attains a particular value is given by the Poisson probability mass function,

$$P(y = k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{20}$$

For $\lambda = 30$, Equation 20 yields $P(20 \leq y \leq 40) = P(y = 20) + \ldots + P(y = 40) \approx 0.95$, i.e., in 19 out of 20 cases, the number of cars observed is between 20 and 40.

The mean and the standard deviation of a Poisson distributed random variable are related to the rate parameter as

$$\begin{aligned} \text{E}[y] &= \lambda & (21) \\ \text{Sd}[y] &= \sqrt{\lambda}. & (22) \end{aligned}$$

That is, the mean and the standard deviation cannot change independently from each other. This disadvantage leads to considering the negative binomial distribution later in this Section.

4.2.2 Modeling sickness absence with Poisson distribution

The Poisson distribution can be used for modeling the number of sickness absence days. In its simplest, the sickness absence days are modeled without predictors, as done in the car counting example. This models the distribution of days when the hypothetical experiment of observing the same individual for the same year is repeated.

However, the events were point-sized in the car counting example, but in sickness absence, the events are days, i.e., each event lasts for a certain time. This modeling inaccuracy is illustrated by the observation that the Poisson distribution allows two or more sickness absences during one day.

4.2.3 Modeling sickness absence with compound Poisson distribution

Another approach for modeling sickness absence days considers the beginning of each sickness absence a point event with variable duration. This yields the **compound Poisson distribution**, written as

$$N \sim \text{Poisson}(\lambda) \tag{23}$$

$$y = \sum_{k=1}^{N} X_k, \tag{24}$$

where N is the number of sickness absence periods, and $X_1, \ldots, X_N$ are independent and identically distributed lengths of the sickness absence periods. The number of the periods is Poisson distributed, and an arbitrary distribution is used for their lengths.

Using the compound Poisson distribution has some drawbacks. Most importantly, the distribution is more complex than the Poisson distribution, and the regression packages of statistical software rarely support compound Poisson distributed response variable. Also, it allows overlapping events, because N, the number of events, is independent of the lengths of the events, X_k.

For these reasons, we do not consider the compound Poisson distribution further in this Thesis, but only introduced it as another example of a distribution for modeling sickness absence.

4.3 Overdispersion in count data

A considerable disadvantage of the Poisson distribution is that it does not allow a standard deviation that is independent from the mean. However, it is common for count data to be **overdispersed**, i.e., have a standard deviation larger than a Poisson model predicts.

Next, we consider causes for overdispersion identified by, e.g., Berk and MacDonald (2008).

4.3.1 The dependence of the events

One cause for overdispersion is the mutual dependence of the events that constitute the counts. For instance, cars that arrive at a crossing with a high rate parameter typically affect each others' movement, so the cars are not independent from each other.

In the sickness absence example, the consecutive days depend on each other, because being absent for a day increases the probability of being absent for the next day.

4.3.2 A missing predictor

Later, we add predictors to the count data model, as was done to linear model in Section 3. It is reasonable to assume that all predictors affecting sickness absence are not included in the model. Consequently, two persons with same predictor values still have differing rates of sickness absence.

When a single Poisson distribution is used to model the two persons with the same predictor values, the distribution has a rate parameter that is between the two. However, depicted in Figure 1, the mixture of two Poisson distributions has a larger standard deviation than a single Poisson distribution with the same mean. Therefore, modeling the mixture with a single Poisson distribution leads to overdispersion.

4.3.3 The negative binomial distribution

The missing predictors in the regression model can be accounted for by adding an error term to Equation 17,

$$\lambda = f(\mathbf{x}) + \epsilon \tag{25}$$
$$y \sim \text{Poisson}(\lambda), \tag{26}$$

where ϵ represents the inter-individual variation that cannot be inferred from the predictors $\mathbf{x}$. Note that the expected value of Poisson distribution is denoted by λ, so $\text{E}[y]$ has been replaced by λ in the notation.

A zero-mean normal distribution is a natural choice for the distribution of ϵ, but it yields no closed-form solution for the distribution of y. If ϵ instead follows a log-gamma distribution, the distribution of y can be equivalently written as (proof omitted)

$$\lambda = f(\mathbf{x}) \tag{27}$$
$$y \sim \text{NB}(\lambda, \theta), \tag{28}$$

where NB denotes the negative binomial distribution.

The **negative binomial distribution** resembles the Poisson distribution, but it has two parameters instead of one: the rate parameter λ controls the mean of the

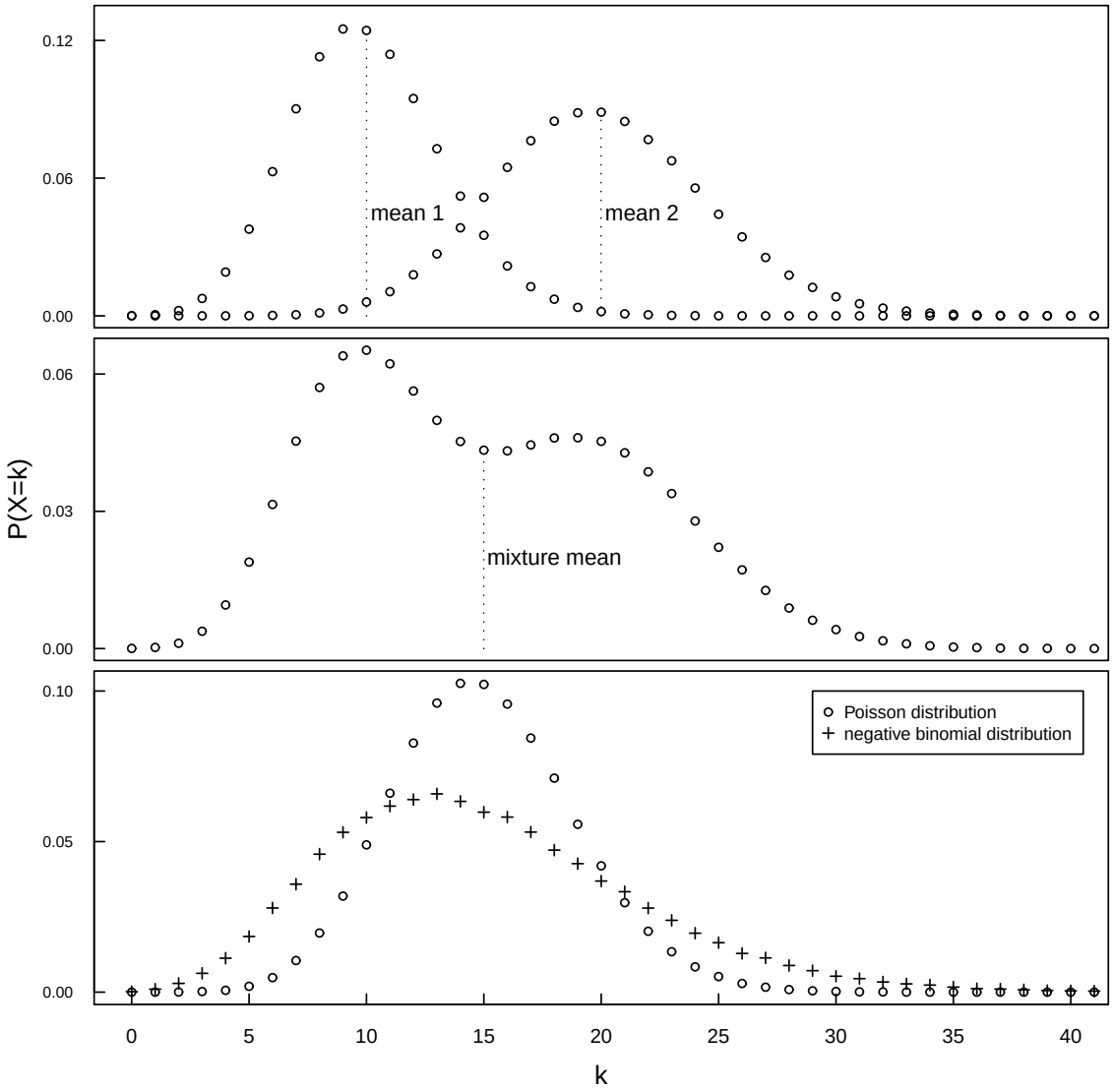

Figure 1: Modeling the mixture of two Poisson distributions with a single Poisson distribution leads to overdispersion. Top panel shows two Poisson distributions with differing means. Middle panel shows their mixture and its mean, calculated using equal weights for both distributions. Bottom panel shows the Poisson and negative binomial distributions fitted to the mixture.

distribution, and the rate parameter λ and the **dispersion parameter** together θ control the standard deviation,

$$\mathrm{E}[y] = \lambda \tag{29}$$

$$\mathrm{Sd}[y] = \sqrt{\lambda + \frac{\lambda^2}{\theta}}. \tag{30}$$

As seen in Equation 30, $\mathrm{Sd}[y] \to \sqrt{\lambda}$ as $\theta \to \infty$, which reduces the negative binomial distribution to a Poisson distribution.

Figure 2 shows the shape of the negative binomial distribution with a fixed value for rate parameter λ and a variable dispersion parameter θ.

4.3.4 The zero-inflated models

The zero-inflated models are a class of models for count data that have excess zeros compared to the Poisson or negative binomial distributions.

Figure 3 presents the distribution of sickness absence days for the data used in Section 9. There is a peak at zero and another peak at 3 days, which makes the distribution **bimodal**, i.e., it has two local maximum values.

To account for a large proportion of zeros and the possible non-zero local maximum, the **zero-inflated models** assume that the population can be divided into two sub-populations: an immune population and a population that has non-zero probability for sickness absence. The immune subpopulation has always zero sickness absences, whereas the non-immune subpopulation is modeled with Poisson or negative binomial distributions.

In mathematical terms, the zero-inflated Poisson model is written as

$$\begin{cases} y = 0, & \text{for } u \in A \\ y \sim \mathrm{Poisson}(\lambda), & \text{for } u \notin A, \end{cases} \tag{31}$$

where $u \in A$ denotes that the observational unit u corresponding to y belongs to the immune sub-population A, and, likewise, $u \notin A$ denotes that u belongs to the non-immune sub-population. The zero-inflated negative binomial model is written similarly.

4.3.5 Hurdle models

Hurdle models are similar to zero-inflated models in that they assume that the population is divided into immune and non-immune sub-populations. However, in hurdle models, the non-immune sub-population always has sickness absences, i.e., the observations with zero sickness absence always belong to the immune subpopulation and the observations with sickness absences always belong to the non-immune sub-population.

In hurdle models, Equation 31 is slightly modified,

$$\begin{cases} y = 0, & \text{for } u \in A \\ y \sim \mathrm{Poisson}_0(\lambda), & \text{for } u \notin A, \end{cases} \tag{32}$$

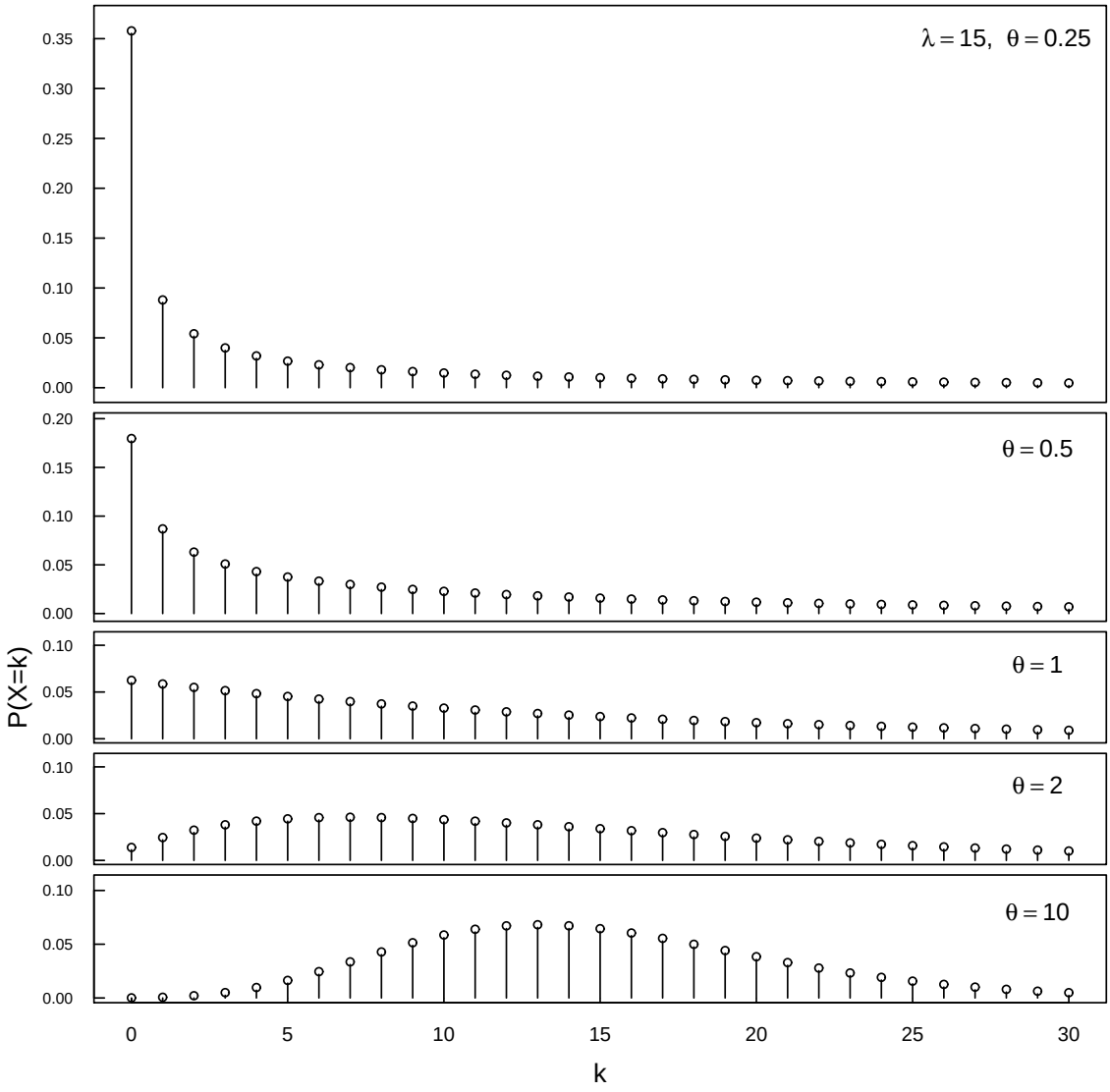

Figure 2: The probability mass function of the negative binomial distribution with fixed rate parameter λ and variable dispersion parameter θ. The shape of the distribution resembles Poisson distribution for large values of θ (bottom-most panel), and as θ decreases, small values obtain more probability mass.

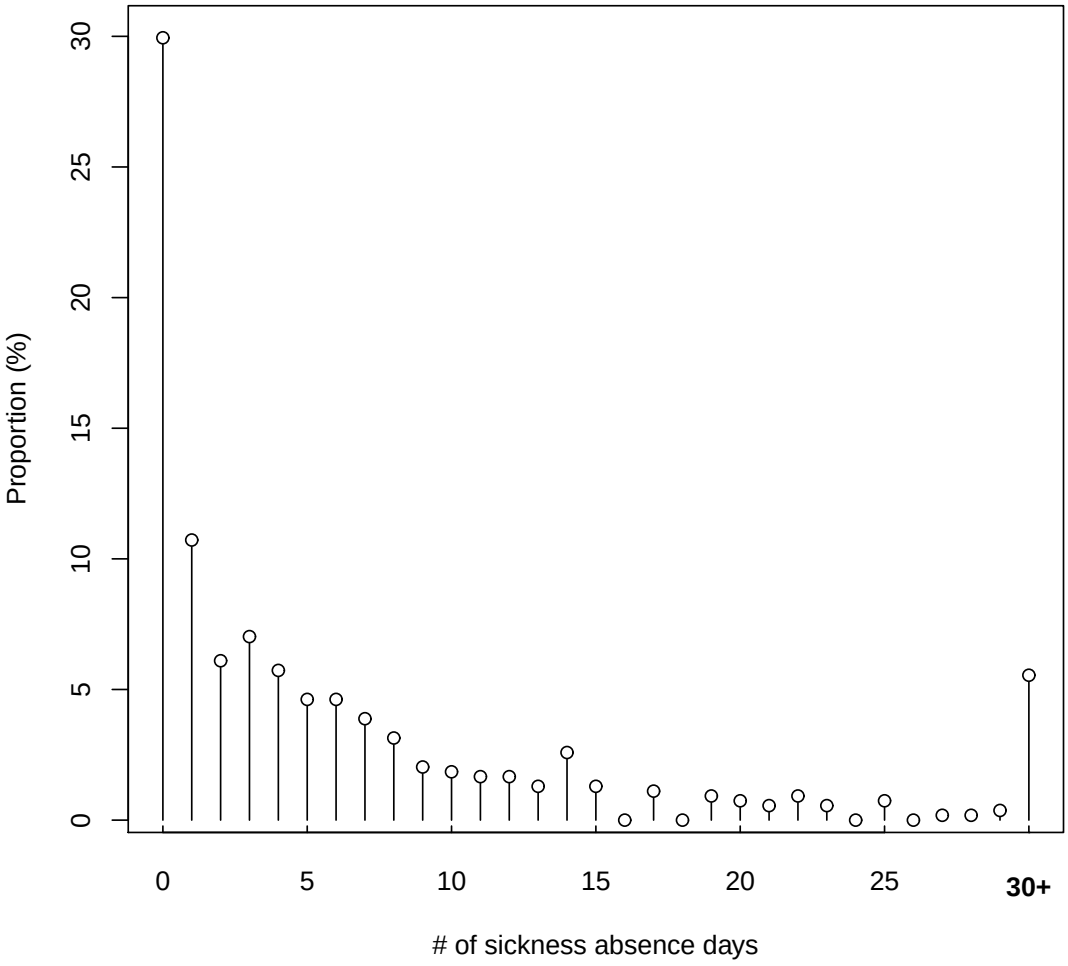

Figure 3: The number of sickness absence days during one-year observation period shown for the population studied in Section 9. Note that all persons with more than 30 sickness absence days are plotted in the right end of the figure, with the corresponding label marked with bold.

where A denotes the immune sub-population, as before. Poisson_0 is the **zero-truncated Poisson distribution**, which has a probability mass function similar to Poisson distribution, but the probability of 0 is set to 0 and the probabilities of $1, 2, \ldots$ are scaled so that they sum to 1. The zero-truncated negative binomial distribution is formed similarly.

4.4 The generalized linear models

So far, we have considered models for the distribution of the response variable, i.e., the stochastic part of the regression model framework of Equations 17 and 18. The deterministic part, or choosing function f in $\mathrm{E}[y] = f(\mathbf{x})$, is the other step in specifying the model.

In Section 3, we introduced the linear model, in which a linear relationship was assumed between $\mathrm{E}[y]$ and $\mathbf{x}$, and y was assumed to be normally distributed. The **generalized linear models** (McCullagh & Nelder, 1989; not to be confused with the general linear model) extend linear regression to non-normal distributions of the observations and to cases where the range of the observations is restricted.

The expected value of a count is always non-negative, so count data serves as an example of restricted range of the observations. With count data, modeling $\mathrm{E}[y]$ as a linear function of $\mathbf{x}$ does not work, because a linear function obtains both negative and positive values.

To account for the restrictions, generalized linear models use a **link function**, which maps the expected value of the model to the range of the linear function. This is denoted as

$$\begin{aligned} g(\mathrm{E}[y]) &= \mathbf{x}^T\beta & (33)\\ y &\sim \mathrm{Distr}(\mathrm{E}[y]), & (34) \end{aligned}$$

where g is the link function and *Distr* is an arbitrary distrition, as before.

The **mean function** is the inverse of link function. Written in terms of the mean function, Equations 33 and 34 become

$$\begin{aligned} \mathrm{E}[y] &= g^{-1}(\mathbf{x}^T\beta) & (35)\\ y &\sim \mathrm{Distr}(\mathrm{E}[y]), & (36) \end{aligned}$$

where g^{-1} denotes the mean function. For clarity, we use only the mean function in the models we present.

4.4.1 Linear model

The generalized linear models contain the linear model introduced in Section 3 as a special case: With mean function $f(x) = x$ and normally distributed y, the model in Equations 35 and 36 reduces to the linear model.

4.4.2 Poisson regression model

The **Poisson regression model** is an example of a generalized linear model, written as

$$\begin{aligned} \mathrm{E}[y] &= \exp(\mathbf{x}^T\beta) & (37)\\ y &\sim \mathrm{Poisson}(\mathrm{E}[y]), & (38) \end{aligned}$$

where the exponential mean function ensures that $\mathrm{E}[y]$ is positive.

4.4.3 Logistic regression model

The **logistic regression model** is used for classification,

$$\mu = \text{logit}^{-1}(\mathbf{x}^T\beta) \quad (39)$$
$$y \sim \text{Bernoulli}(\mu), \quad (40)$$

where the mean function is called the **logistic function**, defined as

$$\text{logit}^{-1}(x) = (1+\exp(x))^{-1}. \quad (41)$$

The logistic function maps the range $(-\infty, \infty)$ to the range $(0, 1)$.

The observation y is Bernoulli(p) distributed, i.e., it attains value 1 with probability p and value 0 with probability $1-p$. In classification, value 1 is interpreted as belonging to class A and value 0 as not belonging to class A.

4.4.4 Negative binomial regression model

Similarly, adding a generalized linear part to the negative binomial model in Equations 27 and 28 yields

$$\lambda = \exp(\mathbf{x}^T\beta) \quad (42)$$
$$y \sim \text{NB}(\lambda, \theta). \quad (43)$$

In the negative binomial model, β and θ are the model parameters. In principle, both θ and λ can depend on the predictors but usually θ is modeled without them.

4.4.5 Zero-inflated and hurdle models

The zero-inflated and hurdle negative binomial models in the generalized linear model framework consist of modeling λ, as in Equation 42, and using a logistic regression model for the classification,

$$\begin{cases} y = 0, & \text{for } u \in A \\ y \sim \text{NB}(\lambda, \theta), & \text{for } u \notin A, \end{cases} \quad (44)$$

where $P(u \in A) = \mu$, and

$$\lambda = \exp(\mathbf{x}^T\beta_{\mathbf{m}}) \quad (45)$$
$$\mu = \text{logit}^{-1}(\mathbf{x}^T\beta_{\mathbf{c}}). \quad (46)$$

In hurdle models, the negative binomial distribution NB is replaced by the truncated negative binomial distribution, denoted by NB_0.

Note that two groups of regression coefficients are needed for the zero-inflated model, $\beta_{\mathbf{m}}$ for modeling the mean λ, and $\beta_{\mathbf{c}}$ for classification.

5 Regression with Gaussian processes

This Section is based on Rasmussen & Williams (2006) and Vanhatalo et al. (2011), who consider many issues ignored here, e.g., detailed algorithms for computation with Gaussian process models.

In this Section, we introduce Gaussian process models, which are flexible non-linear regression models. They can be used to find non-linear relationships between predictors and the response variable. Gaussian process models also automatically assess interactions between the predictors.

5.1 An introduction to Gaussian process models

In Equation 16, we wrote y as a function f of the predictors $\mathbf{x}$. Unlike generalized linear models, Gaussian process models assume a random f so that the values of f follow a zero-mean Gaussian distribution.

Specifically, at each data point $\mathbf{x_i}$,

$$\mathrm{E}[y_i] = f(\mathbf{x_i}) \sim N(0, \sigma_i^2). \tag{47}$$

The function f is called the **latent function**, as it is not observed directly, but only through the observations y_i. As before, the values of y follow an arbitrary distribution around their mean.

Similarly to generalized linear models, we can restrict range of $E[y]$ by using a mean function. For instance, if y_i follows a Poisson distribution, an exponential mean function can be used to ensure that $E[y]$ is non-negative.

Using the short-hand notation $f_i = f(\mathbf{x_i})$,

$$\begin{aligned} \mathrm{E}[y_i] &= g^{-1}(f_i) && (48) \\ f_i &\sim N(0, \sigma_i^2), && (49) \end{aligned}$$

where g^{-1} is the mean function. Note that *mean function* is also used in non-zero-mean Gaussian processes, where it refers to the function that specifies the mean of the process. However, in this text, we only consider zero-mean Gaussian processes, avoiding the risk of confusion.

The next step in formulating the Gaussian process model is adding covariance between data points. This is done using a **covariance function**, $\mathrm{Cov}(f_i, f_j)$, which specifies the covariance as a function of two latent variables, f_i and f_j. Covariance is added because if the distributions of f_i for different data points $\mathbf{x_i}$ were mutually independent, the data $(\mathbf{x_i}, y_i), 1 \leq i \leq n$, do not affect the distribution of f_* for a new data point $\mathbf{x}_*$, and therefore does not affect predictions for x_*.

When $\mathbf{x_i}$ and $\mathbf{x_j}$ are close to each other, it is a reasonable modeling assumption that the corresponding values of f covary more than when they are farther apart. This gives rise to a class of covariance functions called the radial basis covariance functions, which depend only on the distance between $\mathbf{x_i}$ and $\mathbf{x_j}$, i.e.,

$$\mathrm{Cov}(f_i, f_j) = \phi(||\mathbf{x_i} - \mathbf{x_j}||) = \phi(r). \tag{50}$$

A commonly used covariance function is the squared exponential function,

$$\mathrm{Cov}(f_i, f_j) = \sigma^2 \exp\left[-\left(\frac{||\mathbf{x_i} - \mathbf{x_j}||}{l}\right)^2\right], \tag{51}$$

where σ^2 is the variance of each $\mathbf{x_i}$. l is called the **length scale**, and it specifies how quickly the covariance decreases when the distance between two data points increases. l and σ^2 are the parameters of the covariance function, and for now, we consider them fixed. Later, we present methods for optimizing them as a part of the modeling.

The covariances between pairs of the observed data points $\mathbf{x_i}, \mathbf{x_j}, 1 \leq i, j \leq n$ form the covariance matrix K. In Gaussian processes, the data points $\mathbf{x_i}$ reduce to the covariance matrix K, i.e., we do not the data points themselves beyond computing K.

Gaussian processes are non-parametric models, so the function $f(\mathbf{x_i})$ itself does not have parameters. This is in contrast to the generalized linear models, where the inference consists of optimizing the parameters of f. However, if we consider f as a random vector of n dimensions and a Gaussian distribution specified by K, i.e.,

$$\mathbf{f} \sim N(0, K). \tag{52}$$

we can condition the latent values $\mathbf{f}$ on the observations $\mathbf{y}$. Informally, this is interpreted as drawing samples for $\mathbf{f}$ and discarding those that disagree with the observations $\mathbf{y}$. Formally, we do Bayesian inference for $\mathbf{f}$,

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})}. \tag{53}$$

Note that the posterior distribution $p(\mathbf{f}|\mathbf{y})$ is not necessarily Gaussian, although $p(\mathbf{f})$ is. Also, in contrast to the usual application of the Bayes' theorem to compute the parameter posterior, here the inference is done for the latent function $\mathbf{f}$. In a sense, the latent function values f_i at the points $\mathbf{x_i}$ can be interpreted as the model parameters, as they and the parameters of the covariance function govern the model behaviour.

$p(\mathbf{f})$ is often called a **Gaussian process prior** for the latent values f, which is consistent with the above interpretation and Bayesian terminology. The parameters of the covariance function are called the model **hyperparameters**.

5.2 Prediction with Gaussian processes

Contrary to parametric models, e.g., the linear model, the inferred latent posterior $p(\mathbf{f}|\mathbf{y})$ is usually not of primary interest in Gaussian process models. However, the latent posterior provides updated knowledge about the latent function f, which can be used to update knowledge about f at points other than $\mathbf{x_i}$, and ultimately make predictions about y.

For prediction, we examine an arbitrary **test point**, denoted by $\mathbf{x}_*$. The corresponding latent function value is denoted by $\mathbf{f}_*$. The latent posterior at the test point $\mathbf{x}_*$, i.e., $p(f_*|\mathbf{y})$, is calculated by integrating over the latent posterior $p(\mathbf{f}|\mathbf{y})$,

$$p(f_*|\mathbf{y}) = \int p(f_*|\mathbf{f})p(\mathbf{f}|\mathbf{y})d\mathbf{f}. \tag{54}$$

The uncertainty about f_* can be further propagated to compute the **posterior predictive distribution** for y_*,

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{y})df_*. \tag{55}$$

The posterior predictive distribution is a distribution for the predictions of y, i.e., it can be used to draw predictions or compute statistics of interest, for example posterior predictive mean.

5.3 The optimization of the hyperparameters

So far, the parameters of the covariance function, i.e., the hyperparameters, have been fixed. However, an essential part of Gaussian process modeling is to find reasonable values for the hyperparameters.

If hyperparameters are allowed to vary, the prior $p(f)$ becomes conditional on the hyperparameters, denoted by $p(\mathbf{f}|\theta)$, where θ is the hyperparameter vector. For full Bayesian inference, a prior $p(\theta)$ is also needed for the hyperparameters.

Earlier, we computed the latent posterior $p(\mathbf{f}|\mathbf{y})$ with regard to the observations $\mathbf{y}$. We now do the same for hyperparameters,

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}. \tag{56}$$

$p(\mathbf{y}|\theta)$ is related to the value of θ only through the latent function $\mathbf{f}$, so it cannot be calculated directly. Therefore, we compute $p(\mathbf{y}|\theta)$ by integrating over $\mathbf{f}$:

$$p(\mathbf{y}|\theta) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\theta)d\mathbf{f}. \tag{57}$$

The hyperparameter posterior $p(\theta|\mathbf{y})$ can be used in two ways. First, a point estimate for θ can be computed, in which case we maximize $p(\theta|\mathbf{y})$ with regard to θ. This yields the maximum a posteriori (MAP) estimate for θ,

$$\hat{\theta} = \operatorname*{argmax}_{\theta} p(\theta|\mathbf{y}). \tag{58}$$

Once a value for θ is set, we can proceed with the inference as shown previously.

Second, we can take into account the whole hyperparameter posterior $p(\theta|\mathbf{y})$ and propagate the uncertainty in θ to the latent prior $p(\mathbf{f}|\theta)$ and on to the posterior predictive distribution $p(y_*|\mathbf{y}, \theta)$. This is done by first sampling the hyperparameter

posterior $p(\theta|\mathbf{y})$, doing the inference individually for each sample, and then combining the obtained posterior predictive distributions. For that, the **law of total variance** and the **law of total expectation** are used,

$$\begin{aligned} \mathrm{E}[y_*|\mathbf{y}] &= \mathrm{E}_\theta[\mathrm{E}[y_*|\mathbf{y},\theta]] && (59) \\ \mathrm{Var}[y_*|\mathbf{y}] &= \mathrm{E}_\theta[\mathrm{Var}[y_*|\mathbf{y},\theta]] + \mathrm{Var}_\theta[\mathrm{E}[y_*|\mathbf{y},\theta]]. && (60) \end{aligned}$$

Using Equations 59 and 60, the expectations and variances of the posterior predictive distribution, $p(y_*|\mathbf{y})$, can be combined for different values of θ to obtain the expected value and variance of the posterior predictive distribution, to which the uncertainty in θ has been propagated.

5.4 Computation with Gaussian processes

Once the Gaussian process prior $p(\mathbf{f})$ and the likelihood function $p(\mathbf{y}|\mathbf{f})$ are selected, the posterior predictive distribution $p(y_*|\mathbf{y})$ can be computed for the test point $\mathbf{x}_*$ by combining Equations 53, 54, and 55. However, the actual computation of $p(y_*|\mathbf{y})$ is often difficult.

In the simplest case, we assume a normal observation model. Using a Gaussian process prior $p(\mathbf{f})$ and a normal likelihood $p(\mathbf{y}|\mathbf{f})$, the posterior $p(\mathbf{f}|\mathbf{y})$ in Equation 53 is normally distributed. This results in normally distributed posterior predictive distribution $p(y_*|\mathbf{y})$, the parameters of which can be analytically solved.

However, if the likelihood is not Gaussian, the latent posterior $p(\mathbf{f}|\mathbf{y})$ is not Gaussian either, and the integral in 54 becomes analytically intractable. We can still solve the integral by using Monte Carlo sampling or analytic approximations to the latent posterior. Monte Carlo sampling will not be considered here, but next we present a method for approximating the latent posterior, namely, the Laplace approximation.

5.4.1 Laplace approximation

The Laplace approximation is a general method for approximating the posterior distribution in Bayesian inference. It approximates the posterior by a normal distribution, which is a good approximation because the posterior often resembles the normal distribution. The normal distribution is also analytically well-behaved, which makes the approximation particularly useful.

The normal distribution in multiple dimensions, called the **multivariate normal distribution**, has two parameters, the mean and the **covariance matrix** In Laplace approximation, the values of the parameters are obtained from the second-order Taylor polynomial of the logarithm of the approximating distribution around its mode,

$$\log(p(x)) \approx \log(p(\hat{x})) + \nabla \log(p(\hat{x})(x - \hat{x}) + \frac{1}{2}(x - \hat{x})^T \nabla\nabla \log(p(\hat{x}))(x - \hat{x}), \quad (61)$$

where $p(x)$ is the approximated distribution, $\hat{x}$ is the mode of $p(x)$, and $\nabla\nabla$ denotes the Hessian.

By definition, $\nabla \log(p(x)) = 0$ at the mode of $\log(p(x))$, so Equation 61 reduces to

$$\log(p(x)) \approx \log(p(\hat{x})) + \frac{1}{2}(x - \hat{x})^T \nabla\nabla \log(p(\hat{x}))(x - \hat{x}). \tag{62}$$

From that, we obtain an approximation for $p(x)$ by exponentiating,

$$p(x) \approx p(\hat{x}) \exp\left(\frac{1}{2}(x - \hat{x})^T \nabla\nabla \log(p(\hat{x}))(x - \hat{x})\right). \tag{63}$$

The right-hand side of Equation 63 is identified as the unnormalised multivariate normal distribution with the mean $\hat{x}$ and the covariance matrix $\Sigma = -\nabla\nabla \log(p(\hat{x}))$.

Although there are posterior approximations better than the Laplace approximation, namely, expectation propagation (Minka, 2001), we use Laplace approximation in modeling sickness absence in Section 9 because it was found more stable with complex Gaussian process models.

5.5 The connection between Gaussian process models and generalized linear models

Using a Gaussian process model with the linear covariance function, laitakaava!, reduces the model to the generalized linear model. We present this result without justification, but an interested reader is recommended to turn to Rasmussen & Williams (2006) for a discussion of the weight space view and the function space view of Gaussian processes.

We use this result in Section 9 for comparing Gaussian process models to generalized linear models directly, using the same toolbox and the same algorithms for modeling.

5.6 Specifying the Gaussian process model

A Gaussian process model is specified by setting the functional form of the covariance function and specifying the relationship between the response variable y and the latent function f.

Analogously to the generalized linear models, the Gaussian process model is written as

$$f \sim GP(0, \mathrm{Cov}(\mathbf{x}, \mathbf{x}')) \tag{64}$$

$$y \sim \mathrm{Distr}(g(f)). \tag{65}$$

Equation 64 specifies that the latent function f is a zero-mean Gaussian process with the covariance function $\mathrm{Cov}(\mathbf{x}, \mathbf{x}')$, and Equation 65 links the latent value f to the mean of the distribution Distr by the mean function g.

We use the Poisson model, negative binomial model, and hurdle negative binomial model in the comparison of Gaussian process models in Section 9, so they are presented in detail next.

5.6.1 Poisson and negative binomial Gaussian process models

The Poisson Gaussian process model is written as

$$\begin{aligned} f &\sim GP(0, \mathrm{Cov}(\mathbf{x}, \mathbf{x}')) && (66)\\ y &\sim z \cdot \mathrm{Poisson}(\exp(f)). && (67) \end{aligned}$$

The coefficient z is used to match the Poisson distribution to the average values of y because the latent function f is zero-mean.

The coefficient z is often calculated as the average over the observations, but it can also be specified pointwise, e.g., in spatial epidemiology when considering areas of different sizes.

The negative binomial Gaussian process model is written similarly as

$$\begin{aligned} f &\sim GP(0, \mathrm{Cov}(\mathbf{x}, \mathbf{x}')) && (68)\\ y &\sim z \cdot \mathrm{NB}(\exp(f), \theta). && (69) \end{aligned}$$

and a logistic Gaussian process model for classification as

$$\begin{aligned} f &\sim GP(0, \mathrm{Cov}(\mathbf{x}, \mathbf{x}')) && (70)\\ y &\sim \mathrm{logit}^{-1}(\exp(f)). && (71) \end{aligned}$$

5.6.2 Zero-inflated and hurdle Gaussian process models

The zero-inflated Poisson and negative binomial models are written in the Gaussian process framework as

$$\begin{aligned} f_1 &\sim GP(0, \mathrm{Cov}(\mathbf{x}, \mathbf{x}'|\theta_1)) && (72)\\ f_2 &\sim GP(0, \mathrm{Cov}(\mathbf{x}, \mathbf{x}'|\theta_2)) && (73)\\ \mu &= \mathrm{logit}^{-1}(f_1) && (74)\\ \lambda &= \exp(f_2), && (75) \end{aligned}$$

where μ is the probability that the observation belongs to the immune sub-population and λ is the expected number of sickness absence days if the observation belongs to the non-immune sub-population.

As seen in Equations 72 and 73, the zero-inflated Gaussian process models use two latent functions, one for the zero part and one for count part. Zeros in the data can be generated by either part, so the likelihood of zero depends on both parts, which makes the modeling more difficult than with either part alone.

In hurdle models, zeros and non-zeros can only be generated by the zero part and the count part, respectively, which makes hurdle models simpler to implement with Gaussian processes. For a zero observation, we use the latent function f_1,

$$\begin{aligned} f_1 &\sim GP(0, \mathrm{Cov}(\mathbf{x}, \mathbf{x}'|\theta_1)) && (76)\\ \mu &= \mathrm{logit}^{-1}(f_1) && (77)\\ p(y) &= \mu, && (78) \end{aligned}$$

and for non-zero observation, we use f_2,

$$\begin{aligned} f_2 &\sim GP(0, \mathrm{Cov}(\mathbf{x}, \mathbf{x}'|\theta_2)) && (79)\\ \lambda &= \exp(f_2) && (80)\\ y &\sim \mathrm{NB}_0(\mu). && (81) \end{aligned}$$

That is, the two parts can be modeled independently from each other.

5.7 Advantages of using Gaussian process regression

There are several advantages in using Gaussian process regression compared to the generalized linear models. These can be summarized by considering the type of relationships that can be modeled with each.

In generalized linear models, as described in Section 4, the relationship between each predictor and the response variable is linear, apart from the mean function, which introduces nonlinearity to the relationship. However, the mean function is usually monotonic, i.e., it either increases or decreases. Therefore, the relationship as a whole is monotonic.

With Gaussian processes, the predictor and the response variable may depend on each other non-linearly. The difference between a linear and a non-linear relationship is depicted with simulated data in Figure[??]. As seen, the generalized linear model does not detect the nonlinearity present in the relationship.

In practice, nonlinear relationships arise often. For instance, instead of a low or high value, an average predictor value might minimize or maximize the response. When studying factors associated with sickness absence or health in general, e.g., the body-mass index is a good candidate for such predictor.

5.8 The presentation of results

In contrast to the generalized linear models, which can be summarized using regression coefficients, the primary output of Gaussian process regression is the posterior predictive distribution of the response variable at a test point $\mathbf{x}_*$. Therefore, the Gaussian process regression has to be summarized in another way.

One summary can be obtained by considering what happens to the response variable when one of the predictors changes while the others remain fixed at their average values. This is done by computing the posterior predictive distributions at the corresponding test points, and summarizing them, e.g., by calculating mean.

Another summary, namely, average predictive comparisons, presented in Chapter 7, is based on evaluating the average change in the response variable for the average change of the predictor. Average predictive comparisons can be used also for Gaussian processes, but non-linear relationships are not explicitly seen in it.

6 Model selection

In this Section, we present two methods for selecting between different regression models, namely, Akaike information criterion (AIC) and cross-validation. The methods are not specific to regression models, so they are presented in their general form.

6.1 Akaike information criterion

The Akaike information criterion (AIC, see, e.g., Bishop, 2006) is a model selection criterion based on information theory. AIC is a fundamentally a frequentist criterion, for example, it does not take into account prior knowledge. However, it is commonly used to compare generalized linear models, which is why we use it as reference for model selection.

AIC is defined as

$$\mathrm{AIC} = 2k - 2\ln(L), \tag{82}$$

where k is the number of model parameters and L is the likelihood of the data with the current model and its parameters.

The lower the AIC, the better the model is. As seen in Equation 82, if the likelihood does not change, increasing the number of parameters deteriorates the model, whereas increasing likelihood while the number of parameters is fixed improves the model.

The number of model parameters is a measure of the model complexity, so the first term in Equation 82 can be interpreted as a penalty term for too complex models.

6.2 Holdout validation

Before cross-validation, we present a simpler variant, called **holdout validation**. In holdout validation, a part of the data, called the **training dataset**, is used to infer the model, and the rest of the data, the **test dataset**, is used to evaluate the model. The test dataset is not used in training the model, so it can be interpreted as unseen or future data. Therefore, the model performance on the test dataset measures the **predictive performance** of the model.

6.3 Cross-validation

In cross-validation, a part of the data is repeatedly held out to be used as the test dataset, until all data are tested once. The test datasets are usually non-overlapping, i.e., each data point is used exactly once for testing the model.

The predictive performance of the model is then combined over the test datasets, e.g., by mean of the performance criterion.

If the number of repetitions in cross-validation is a constant k, it is called **k-fold cross-validation**. The special case where the number of repetitions equals the number of observations is called **leave-one-out cross-validation**.

6.4 Cross-validation using log predictive densities

Evaluating the model on the test dataset requires a measure of performance. In cross-validation, the actual observations at the test datapoints are known, so the performance measure can compare the model predictions to the observations.

In Bayesian analysis, the model yields predictive distributions instead of point predictions. The predictive distributions can be assessed with an utility function (Vehtari & Lampinen, 2002).

An utility function is defined such that its parameters are a predictive distribution and an observation, and the function returns a value that reflects the discrepancy between the distribution and the observation. An example of a discrepancy measure is the negative mean squared error (MSE) between the observation and the expected value of the distribution. This can be written as

$$U(p(x), x_{\text{obs}}) = -\left|E[p(x)] - x_{\text{obs}}\right|^2 . \tag{83}$$

Often, the negative MSE is not a good utility function, as it compares the observation only to the expected value of the distribution, ignoring most of the information of the distribution. For instance, two very differently shaped distributions with equal means always yield the same MSE.

Another choice for the utility function is the logarithm of the predictive density of the observation (LPD), defined as

$$U(p(x), x_{\text{obs}}) = \log\left(p(x_{\text{obs}})\right) . \tag{84}$$

In the case of many observations, we average the utility function over the observations. This yields an overall estimate for the predictive performance of the model.

In studying sickness absence in Section 9, we use the LPD averaged over the observations, which is called the **mean log predictive density** (MLPD), together with leave-one-out cross-validation.

6.5 Comparing cross-validated models using Bayesian bootstrap

6.5.1 Bootstrap and Bayesian bootstrap

Bootstrap (Efron, 1979) is a non-parametric statistical procedure for estimating sampling distributions, i.e., the distribution of an estimator over repeated sampling of the population. It is often used when minimal distribution assumptions are preferred or when computing standard errors for estimators for which theoretical results of the sampling distribution are not available, e.g., sample standard deviation.

Bootstrap is based on generating additional samples, called **bootstrap samples**, by sampling the observed data with replacement. This is thought to approximate sampling the population. The samples are then treated as if they were independent samples from the population.

In Bayesian bootstrap (Rubin, 1981), the observed data is sampled with non-uniform distribution, i.e., the probability that an observation is picked for the sample is not $1/n$. Instead, the probabilities are assigned by drawing $n-1$ independent random numbers uniformly from the interval $[0, 1]$, ordering them, and calculating the gaps between successive numbers. For the smallest and the largest number, gaps to 0 and 1, respectively, are calculated. The n gaps are then used as the probabilities for the n observations.

Bayesian bootstrap allows the interpretation of the bootstrap samples as samples generated by draws from the posterior of the estimated statistic, whereas ordinary bootstrap has no such Bayesian interpretation.

6.5.2 Comparing models using Bayesian bootstrap

Cross-validation yields an estimate of the predictive performance of the model at each test datapoint. Pairwise comparison of the predictions of two models can then be done using, e.g., a paired t-test. However, Bayesian bootstrap yields samples from the posterior of the paired difference, which allow a wider range of tests, e.g., based on correctly inferring the sign of the difference. Also, Bayesian bootstrap is a non-parametric method, so it does not depend on distribution assumptions.

To keep with the Bayesian paradigm, we use Bayesian bootstrap to compare cross-validated models in Section 9.

7 Summarizing regression analysis

Commonly, a summary of a regression analysis consists of the credible intervals of the regression coefficients (confidence intervals in the case of a frequentist analysis) and a corresponding point estimate, e.g., the posterior mean.

Here, we present another scheme for summarizing the results, average predictive comparisons (Gelman & Pardoe, 2007). We also present a method for variable selection, based on the probability of knowing the sign of the effect of a predictor (Gelman & Tuerlinckx, 2000).

7.1 Average predictive comparisons

In a linear regression model, the regression coefficients can be interpreted as the average change in the response variable when the predictor changes by one. However, in a regression model with nonlinearities, e.g., Gaussian process regression, the change depends on the original predictor value, as well as the values of other predictors.

We may still summarize the change in the response variable, assuming a certain initial combination of the predictor values. For this, the **predictive comparison** is defined as

$$\delta(u_1 \to u_2, v, \theta) = \frac{E[y|u_2, v, \theta] - E[y|u_1, v, \theta]}{u_2 - u_1}, \tag{85}$$

where u_1 is the initial value of the predictor of interest, u_2 is the changed value, and v consists of all other predictors. θ are the parameters of the model.

The predictive comparison can be interpreted as the expected change in the response variable per unit change in the predictor of interest, when the predictor of interest changes from u_1 to u_2 and all other predictors are held constant.

We are interested in computing the average change in the response variable over all increasing transitions $u_1 \to u_2$ and over every v, weighted by the probability of each combination of predictors occurring in the data. This can be estimated by

$$\Delta = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(E[y|u_j, v_i, \theta] - E[y|u_i, v_i, \theta]\right)\operatorname{sign}(u_j - u_i)}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(u_2 - u_1)\operatorname{sign}(u_j - u_i)}, \tag{86}$$

where the sums are taken over all observations, and the weights w_{ij} reflect the probability that u_j is observed together with v_i. The sign function ensures that only increasing transitions are considered. Equation 86 is called the **average predictive comparison**, calculated for the predictor of interest u.

The weights are needed because, in contrast to the predictor combinations u_i, v_i and u_j, v_j, the probability of observing u_i, v_j cannot be calculated directly. Instead, the weights can be estimated using Mahalanobis distance,

$$w_{ij} = \frac{1}{1 + (v_i - v_j)^T \Sigma_v^{-1} (v_i - v_j)}, \tag{87}$$

where Σ_v is the covariance matrix of the components of v.

Credible intervals or other statistics of interest for the average predictive comparisons can be calculated by propagating the uncertainty from the parameters θ to the average predictive comparison Δ in Equation 86. This is typically done by sampling the posterior of θ and calculating the average predictive comparison for each sample, which samples the posterior of the average predictive comparison.

Both Gaussian process regression and generalized linear models are used in studying sickness absence in Section 9, so we summarize the results with average predictive comparisons.

7.2 Variable selection

Traditionally, predictors in regression are considered relevant based on whether the sign of the regression coefficient reliably differs from zero, which is formalised using null hypotheses and p-values.

Gelman and Tuerlinckx (2000) argue that in social sciences, the regression coefficients in principle always differ from zero, which makes the comparison meaningless. Instead, they propose an alternative method, called the **Bayesian multiple comparisons**, based on whether the sign of the regression coefficient can be reliably inferred.

Specifically, they suggest that those variables are considered relevant for which the joint posterior probability of correctly inferring the sign of the effect exceeds a pre-specified threshold, e.g., 95%. These variables can be found using a stepwise procedure, in which the variable that least decreases the probability is added at each step until the probability becomes less than the threshold.

In studying sickness absence in Section 9, we compare Bayesian multiple comparisons to the traditional method that selects the predictors whose marginal credible intervals (which are analogous to confidence intervals in non-Bayesian analysis) excludes zero.

7.3 Disadvantages of using complex models

Using a more complex model is not always preferable, even if it has a higher predictive performance. The increase in performance can be marginal, albeit statistically significant, and a complex model is more difficult to present and interpret.

For example, zero-inflated models yield two sets of regression coefficients. The interpretation of the coefficients is confounded by the model assumptions: The zero part describes the probability that the individual belongs to the immune subpopulation, but the immune sub-population itself is a modeling assumption, so it is possible that the coefficients of both parts describe the same phenomenon.

For these considerations, we also present the results for a relatively simple model, namely, the negative binomial model, in studying sickness absences in Section 9.

8 Considerations for pre-processing the data

8.1 Recoding

To treat different predictors in a consistent manner, we code them as numerical variables. In some environments, e.g., in R, explicit recoding of non-numerical variables is not necessary for regression analysis, but it might still be helpful for other pre-processing steps, e.g., imputation, which is considered in Subsection 8.2.

In general, there are two types of non-numerical variables. A categorical variable has a discrete set of levels that the variable can attain. The levels are not ordered, so, for instance, nationality can be coded as a categorical variable.

In categorical variables, each level has its own indicator variable, i.e., a variable that is is given the value 1 if the original variable is at the indicated level, and 0 otherwise.

In a linear model with a constant term, one of the levels is chosen as the "baseline", which is coded with each indicator at 0. Otherwise, the constant term will be linearly dependent on the combination of the indicators, which causes the model inference to fail.

For example, a variable with the levels "yes", "no", "maybe" can be coded as

$$\begin{aligned} \text{“}yes'' &\rightarrow (1,0) \\ \text{“}no'' &\rightarrow (0,1) \\ \text{“}maybe'' &\rightarrow (0,0). \end{aligned}$$

An ordinal variable also has discrete levels, but the levels are ordered. The order can be inherent to the scale, as in Likert scale ("Strongly disagree", ..., "Strongly agree"), or it can be an interpretation of the levels (e.g., the above levels "yes", "maybe", and "no").

Typically, an ordinal variable is coded with consecutive natural numbers such that the order is retained, e.g., "$yes'' \rightarrow 2$, "$maybe'' \rightarrow 1$, "$no'' \rightarrow 0$. However, it is possible to use any monotonically increasing function of the ordered category labels, e.g., to reflect the gap sizes between adjacent categories.

8.2 Handling missing data

Missing data is a problem in most data analysis done on surveys or questionnaires.

Imputation, i.e., replacing missing values with their statistical estimates, is commonly used for handling missing data. Many authors advocate using multiple imputation, that is, generating several imputed datasets and combining the estimates of interest, e.g., regression coefficients.

However, if the rate of missing data is small, as in the study presented in Section 9, satisfactory results can be obtained using single imputation, i.e., doing the analysis with only one imputed dataset.

8.2.1 Gaussian expectation-maximization algorithm

A general method for both single and multiple imputation is to fit a probability distribution to the observed data and using the distribution to draw samples for the missing values.

To fit the distribution, the **expectation-maximization** (EM) algorithm can be used (e.g., Bishop 2007). EM algorithm begins with an initial value for the distribution parameters and consists of iterating two steps,

1. The **expectation step** calculates the expected values of the missing data using the current estimate for the probability distribution.

2. The **maximization step** adjusts the parameters of the distribution to maximize the likelihood of the current data, which includes the missing value estimates calculated in the expectation step.

The steps are repeated until the algorithm converges, i.e., the changes in the parameters become small enough.

Using a multivariate normal distribution for fitting yields the **Gaussian expectation-maximization** algorithm.

8.3 Standardizing the data

Standardization (or normalization) refers to transforming several variables to a common scale. In regression analysis, standardization of numerical predictors is a common pre-processing step. This has the effect of making the model parameters comparable to each other.

Standardization is often done by subtracting the mean and dividing each variable by its standard deviation (sd). This way, the standardized regression coefficients in a linear model are interpreted as the average change in the response variable corresponding to a change of 1 sd in the original scale.

This procedure has a drawback. The regression coefficients of discrete predictors, e.g., gender, become non-interpretable as such, as there is no "1 sd change in gender." Various modifications to the standardization procedure have been proposed to overcome this.

The simplest modification is to standardize only non-binary predictors. This gives two distinct groups of regression coefficients, which are comparable within but not between the groups. In this way, the binary predictors, coded with 0–1, retain the property that the regression coefficient is the average change in the response variable as the predictor value changes from 0 to 1.

Gelman has proposed that in addition to not standardizing the binary predictors, the non-binary predictors are standardized by dividing them by 2 sd's instead of 1 sd, yielding them a sd of 0.5. This reduces the gap between the two predictor groups, because a binary predictor whose both categories are equally likely has a sd of $\sqrt{p(1-p)} = \sqrt{0.5 \cdot 0.5} = 0.5$. After that, the closer the distribution of the two

categories of a binary predictor is to uniform, the more comparable the corresponding regression coefficient is to the regression coefficients of non-binary predictors.

Table 1: A summary of standardization procedures.

Procedure	Effect on regression coefficients
Subtract mean. Divide each predictor by 1 sd.	Coefficients describe the effect of 1 sd change in the predictor and are comparable to each other; coefficients for binary predictors are non-interpretable.
Subtract mean. Divide each predictor by 2 sd's.	Coefficients describe the effect of 2 sd change in the predictor and are comparable to each other; coefficients for uniformly distributed binary predictors describe the difference between the two categories.
Subtract mean. Divide each non-binary predictor by 1 sd.	Coefficients for binary and non-binary predictors are comparable to each other within but not between groups; coefficients for binary predictors remain interpretable.
Subtract mean. Divide each non-binary predictor by 2 sd.	Coefficients for binary and non-binary predictors arecomparable to each other within but not always between groups; coefficients for binary predictors remain interpretable; coefficients for uniformly distributed binary predictors are comparable to coefficients of non-binary predictors.

9 Case study: modeling sickness absence with healthcare questionnaire data

In this Section, we use regression analysis to study factors associated with sickness absence. For that, we have data consisting of health questionnaire answers obtained from employees of a Finnish company (n=541) and the corresponding sickness absence days during the one year period before filling the questionnaire.

First, we review previous research on factors associated with sickness absence. Then, we describe current data in detail. After that, we present the regression models used and finally the results.

9.1 Previous research on factors associated with sickness absence

Of demographic factors, gender has been strongly associated with sickness absence (females have a higher rate of sickness absence, see, e.g., Kivimäki et al., 2001, Laaksonen et al., 2008). High level of education has also been associated with decreased sickness absence (Ala-Mursula et al., 2002, Niedhammer et al., 1998). For age, Duijts et al. (2007) did not find an association to sickness absence in their meta-analysis.

Of factors related to physical health, self-rated poor health, diagnosed chronic illnesses (Kivimäki et al., 2001), smoking (e.g., Niedhammer et al., 1998), and high body-mass index (Ala-Mursula et al., 2002) have been associated with increased sickness absence.

For alcohol use, Niedhammer et al. (1998) found that abstainers have higher rate of sickness absence, whereas Ala-Mursula et al. (2002) did not find association and Kivimäki et al. (2001) found a similar association in one of the two occupational groups studied. Vahtera et al. (2002) have proposed a U-shaped relationship between alcohol intake and sickness absence, i.e., moderate consumption is associated with lower rate of sickness absence than high or low consumption.

Mental health has also been associated with sickness absence. Psychiatric morbidity has been associated especially with long spells of sickness absence (Kivimäki et al., 2001). According to the meta-analysis of Duijts et al. (2007), psychological symptoms and psychosomatic complaints in general are associated with increased sickness absence.

Also, psychosocial factors at work, e.g., lack of social support (Niedhammer et al., 1998) and bullying (Kivimäki et al., 2000), have been associated with increased sickness absence.

Taimela et al. (2007) used models similar to our study and found that having two or more health problems in domains of pain, depression, or insomnia is a strong predictor for sickness absence.

Interactions between gender and other predictors have been found. For example, feeling overloaded at work has been associated with increased sickness absence in males but not in females (Kivimäki et al., 2000).

Also, predictors that were not used in the current study have been found relevant for sickness absence, e.g., being unmarried has been strongly associated with increased sickness absence especially in males (Niedhammer et al., 1998, Kivimäki et al., 2001).

9.2 Data characteristics

The questionnaire contained validated items about, e.g., physical activity, work-related stress, depression, and pain. It was administered as a part of the employees' occupational healthcare.

The response variable consisted of the total number of sickness absence days during the one-year period before the administration of the questionnaire–from the beginning of September in 2008 to the end of August in 2009–gathered from the employer's registry.

Previously, sickness absence has been studied using generalized linear models, e.g., in Taimela et al., 2007. We follow the same approach, but in addition, we try a novel method, namely, the Gaussian process regression, introduced in Section 5.

The present study belongs to a larger study (Reijonsaari et al., 2009), which investigates the effects of a physical activity intervention on a large group of employees. Therefore, the study population consisted of only those employees who were willing to participate in the one-year-long intervention study and were not excluded.

The exclusion criteria for the intervention study consisted of medical reasons, e.g., pregnancy or disorders that make physical activity dangerous. The complete list and a detailed description of the questionnaire items is found in Reijonsaari et al., 2009.

9.3 Data pre-processing

One participant did not return the questionnaire, and two participants denied the use of their sickness absence data. After these removals, the study population consisted of 541 employees.

The missing data in the questionnaire was imputed using the expectation-maximization (EM) algorithm with a multivariate Gaussian distribution. The imputation was done using *pmtk3* (Murphy & Dunham, 2008) package in MATLAB. The missing data rate was low (1.1%), so single imputation was used.

There were in total 106 questionnaire items. Of these, 24 items were pre-selected with expert help, considering them potentially related to sickness absence. The items were then divided into four groups, which consisted of items related to different areas of life. The items and the corresponding groups are listed in Table 2,

We transformed the sickness absence days to make them comparable between subjects. First, we calculated the potential working days for each subject, in which maternal leaves and other non-sickness related absences were excluded. Then, we scaled the sickness absence days for each subject by the ratio of their potential working days and the maximum number of potential working days in a year, typically 220

days in Finland. The scaled number was then rounded to the nearest integer value. The sickness absence data were also checked for inconsistencies, e.g., duplicates.

Table 2: For model selection, predictors were divided into four groups. The predictors marked with * are three-class ordinal variables.

Individual factors	**Health-related factors**
Age	Leisure-time PA
Sex	Smoking
	Alcohol consumption
	Diabetes score
	Body-mass index
	The use of intoxicants*

Symptoms	**Work-related factors**
Cardiovascular disease	Work status
Musculoskeletal disorder	Assessment on future disability to work
Pain-related impairment at work	Perceived stress at work
Depression score (DEPS)	Feeling of being in control at work
Daytime sleepiness (Epworth Sleepiness Scale, ESS)	Time for family and friends
Insomnia*	Work community
Sleep deprivation*	Workplace bullying
	Contentment

9.4 Non-Bayesian regression using generalized linear models

For reference, we did non-Bayesian regression analysis using generalized linear models. We tested six model types: Poisson, hurdle Poisson, zero-inflated Poisson, negative binomial, hurdle negative binomial, and zero-inflated negative binomial. For each model type, the four predictor groups (Table 1) were used incrementally, i.e., individual factors only, then individual factors and health-related factors, etc. This was done because including more predictors may deteriorate the model performance with complex model types.

The model fitting was done using R (R Development Core Team, 2009, http://www.R-project.org/) and its toolboxes *MASS* (Venables & Ripley, 2002) and *pscl* (Zeileis et al., 2008). All models except two were fitted using the default optimization algorithm. The two models were fitted using simulated annealing because the default algorithm did not convergence.

Table 3 presents the AIC values for each regression model and predictor group. The negative binomial model with all predictor groups has the smallest AIC. The corresponding regression coefficients are presented in Figure 4.

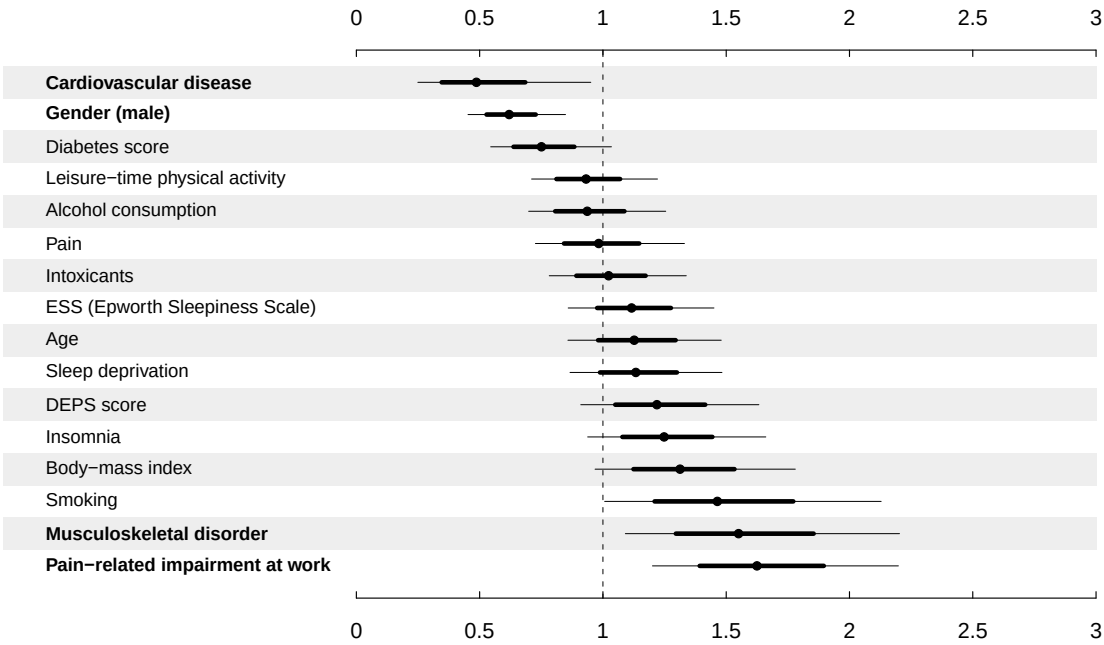

Figure 4: The regression coefficients for the best generalized linear model: the negative binomial model with individual factors, symptoms, and health factors as the predictors. The non-Bayesian 68% and 95% confidence intervals are represented by thick and thin lines, respectively. The coefficients are presented as mean ratios, e.g., a ratio 2 on average doubles the sickness absence days compared to the reference. The predictors whose mean ratios differ statistically significantly from zero are marked with bold. Note that all predictor values are self-reported, including cardiovascular disease and musculoskeletal disorder.

Table 3: AIC values for each model and predictor group. Smallest AIC is marked with bold. I, H, S, and W denote that individual factors, health factors, symptoms, and work factors were used as predictors, respectively. - denotes modeling without predictors. The models marked with * were fitted using simulated annealing.

		-	I	I H	I H S	I H S W
Poisson	ordinary	9686	9266	8929	7818	7529
	hurdle	7458	7301	7041	6318	6040
	zero-inflated	7458	7301	7041	6318	6040
Negative binomial	ordinary	3133	3112	3111	**3079**	3083
	hurdle	3132	3100	3111	3081	3087
	zero-inflated	3133	3103	3105	3172*	3179*

9.5 Bayesian regression using Gaussian process models

We also did Bayesian regression using Gaussian process models to discover non-linear relationships between the predictors and sickness absence. The analysis was done in two stages: selecting the regression model and selecting the predictors.

The stages are described in detail below.

In the first stage, the regression model was selected among the combinations of three model types–Poisson, negative binomial, and hurdle negative binomial– and four covariance functions–linear, squared exponential, Matérn 5/2, and the neural network covariance function. Each covariance function except the neural network was used together with the linear covariance function by summing them. We also tested two different hyperparameter priors for the length scale of the squared exponential and Matérn 5/2 covariance functions–a hierarchical prior and a log-Gaussian prior. A detailed description of the priors used for each covariance function is in Appendix A.

Mean log predictive density (MLPD) was used as the model selection criterion, evaluated with leave-one-out cross-validation. The model with the smallest MLPD was tested against the other models using Bayesian bootstrap for pairwise comparison of the log predictive densities of the models. The Bayesian modeling and the cross-validation were done using the MATLAB toolbox *GPstuff* (Vanhatalo et al., 2011).

For efficiency, the model hyperparameters were optimized for the full dataset, and the same hyperparameters were used for each fold of the cross-validation. This was considered to cause negligible error. We also used Laplace approximation instead of the more accurate expectation propagation for approximating the latent posterior distribution because Laplace approximation was found more stable.

For the first stage model, we present the results using average predictive comparisons.

In the second stage, we did variable selection using Bayesian multiple compar-

isons and a threshold of obtaining at least 95% posterior probability for correctly inferring the signs of the predictors. After variable selection, the Gaussian process model was refitted using the selected predictors only.

For the second stage model, we present both average predictive comparisons and the latent function of the Gaussian process for qualitative assessment of the relationship between the selected variables and sickness absence.

9.5.1 Model selection

In the first stage, we did model selection among various Gaussian process models.

Table 4 shows the MLPD values for each model and predictor group. The model with Matérn 5/2 covariance function and the log-Gaussian prior with all predictors included had the largest MLPD, so it was used for second stage analysis.

Table 4: MLPD values for each model and predictor group. I, H, S, and W denote that individual factors, health factors, symptoms, and work factors were used as predictors, respectively. Largest MLPD is marked with bold. The model with no MLPD value did not convergence. NB stands for negative binomial, prior I is the log-Gaussian prior, and prior II is the hierarchical prior.

		I	I H	I H S	I H S W
Linear	Poisson	-4.71	-4.68	-4.36	-4.33
	NB	-2.87	-2.86	-2.83	-2.82
	hurdle NB	-2.86	-2.86	-2.81	-2.81
Neural network	Poisson	-4.72	-3.98	-3.53	-3.41
	NB	-2.87	-2.87	-2.84	-2.83
	hurdle NB	-2.86	-2.85	-2.82	-2.83
Matérn 5/2 (prior I)	Poisson	-4.73	-2.86	-2.80	-2.80
	NB	-2.87	-2.86	-2.82	-2.81
	hurdle NB	-2.86	-2.84	-2.79	**-2.75**
Matérn 5/2 (prior II)	Poisson	-4.61	-2.86	-2.80	-2.80
	NB	-2.87	-2.84	-2.81	-2.80
	hurdle NB	-2.86	-2.83	-2.77	-
SE (prior I)	Poisson	-4.71	-2.86	-2.81	-2.79
	NB	-2.87	-2.86	-2.82	-2.81
	hurdle NB	-2.86	-2.84	-2.79	-2.81
SE (prior II)	Poisson	-4.71	-2.86	-2.81	-2.80
	NB	-2.87	-2.84	-2.81	-2.80
	hurdle NB	-2.86	-2.83	-2.77	-2.78

We also compared the model with largest MLPD to other models using Bayesian bootstrap. Comparing the best model that used Matérn 5/2 covariance function with log-Gaussian prior to the best model that used Matérn 5/2 with hierarchical prior, squared exponential with log-Gaussian prior, and squared exponential with

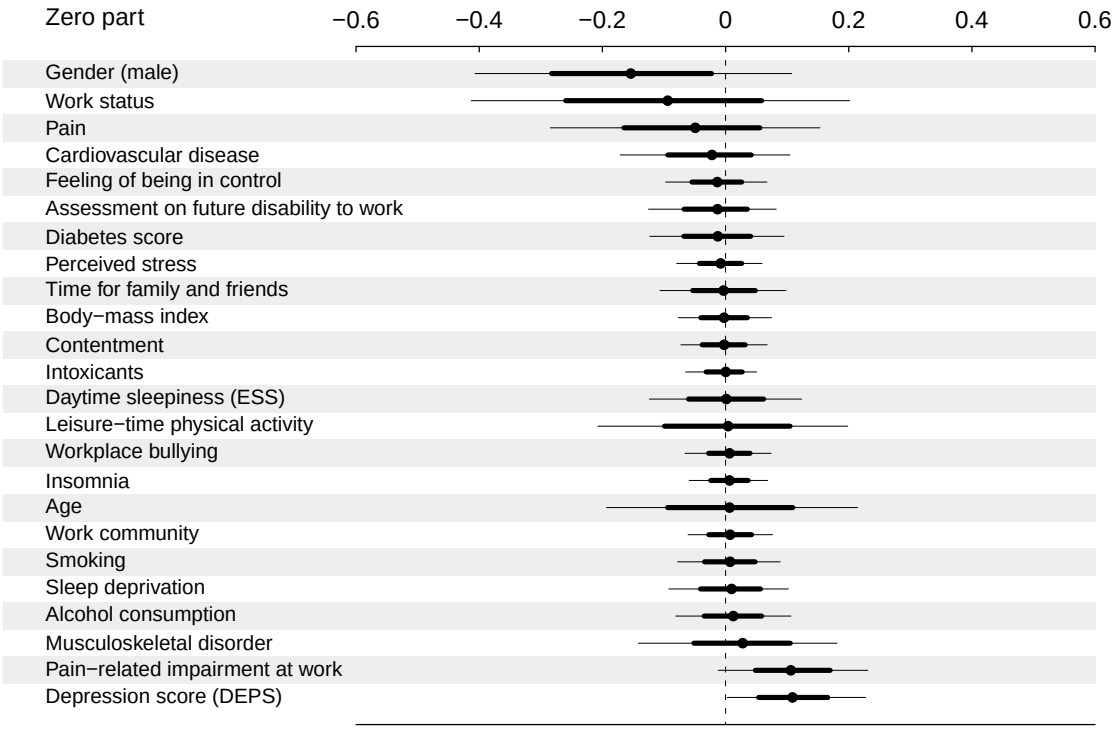

Figure 5: The average predictive comparisons for the zero part of the best Gaussian process model, i.e., the model with all predictors and Matérn 5/2 covariance function with a log-Gaussian prior. For zero part, the average predictive comparison is interpreted as the change in the probability of non-zero sickness absence corresponding to the average change in the predictor. The 50% and 95% credible intervals are represented by thick and thin lines, respectively.

hierarchical prior yielded probabilities 0.92, 0.93, and 0.75, respectively, that the model with largest MLPD was better. The best Gaussian process model was also found better than the best generalized linear model or the best neural network model with a probability > 0.9999.

The average predictive comparisons for the first stage model, i.e., the Gaussian process model with Matérn 5/2 covariance function and the log-Gaussian prior with all predictors included, are shown in Figures 5 and 6.

9.5.2 Variable selection

After choosing the initial model, we did variable selection using Bayesian multiple comparisons.

For the first stage model, i.e., the Gaussian process model with Matérn 5/2

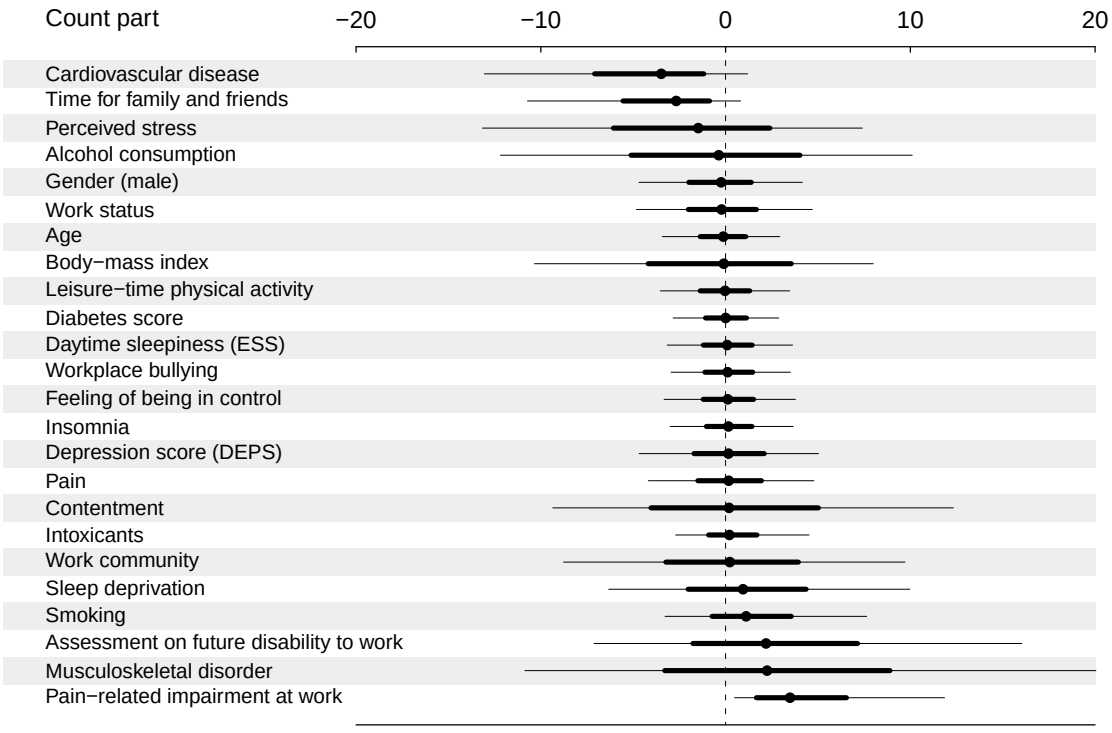

Figure 6: The average predictive comparisons for the count part of the best Gaussian process model, i.e., the model with all predictors and Matérn 5/2 covariance function with a log-Gaussian prior. For count part, the average predictive comparison is interpreted as the change in the expected sickness absence days corresponding to the average change in the predictor. The 50% and 95% credible intervals are represented by thick and thin lines, respectively.

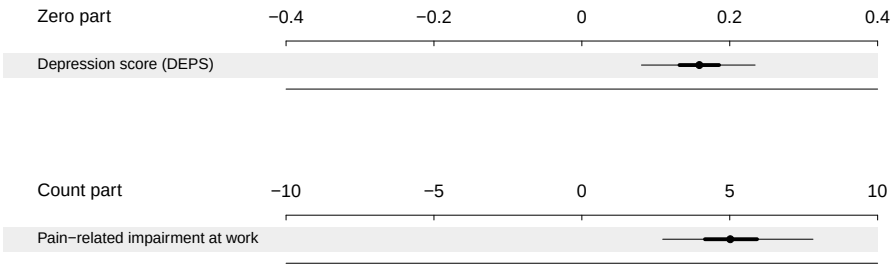

Figure 7: The average predictive comparisons for the zero part and count part of the best Gaussian process model after variable selection with Bayesian multiple comparisons and refitting. The average predictive comparison of zero part and count part are interpreted as the change in the probability of non-zero sickness absence and the change in the expected sickness absence days, respectively, corresponding to the average change in the predictor. The 50% and 95% credible intervals are represented by thick and thin lines, respectively.

covariance function and the log-Gaussian prior with all predictors included, pain-related impairment at work was selected in the count part and depression score (DEPS) was selected in the zero part. The posterior probability of correctly inferring the signs of their effects was 97%. The first variable that was not included was pain-related impairment at work in the zero part, which would have dropped the posterior probability to 93%.

The predictive performance (MLPD) for the model after variable selection was -2.88, which is considerably lower than most models in Table 4. Compared to the model before variable selection, the model after variable selection is worse with probability ¿ 0.9999.

Figure 8 shows the latent function values for the model refitted using the selected variables only, and Figure 7 presents the average predictive comparisons.

9.6 Conclusions

The non-Bayesian analysis using generalized linear models yielded self-reported cardiovascular disease and male gender as factors associated with lower than average sickness absence days, and self-reported musculoskeletal disorder and pain-related impairment at work as factors associated with higher than average sickness absence (Subsection 9.4).

Similarly, the Bayesian analysis using Gaussian process models yielded depression score as a factor associated with higher than average probability of non-zero sickness absence, and pain-related impairment at work as a factor associated with higher than average number of sickness absence days (Subsection 9.5).

The latent function values (Figure 8) show a saturation effect for depression score, i.e., changes in sickness absence become smaller for high depression scores.

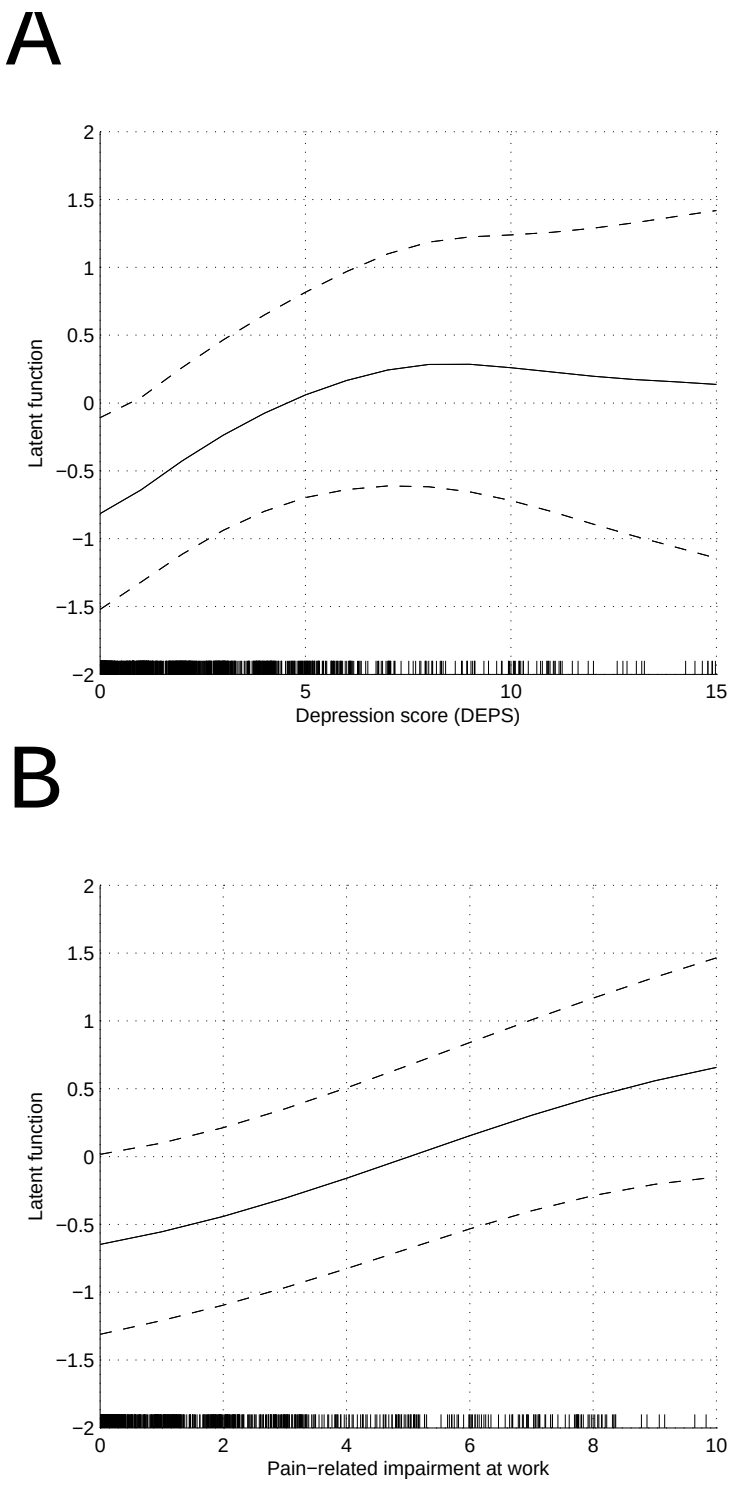

Figure 8: The latent function values for the best Gaussian process model after variable selection and refitting for (A) zero part and (B) count part. The 95% credible intervals are marked with dashed line. The bottom of each plot shows the observed predictor values with jitter added. Note that the range of the DEPS scale is 0–30, but because only one subject had a score higher than 15, the lower half of the scale is presented.

However, the effect might not reflect the actual relationship between depression and sickness absence, but instead be an artifact of too little data.

Predictors not included in the model after variable selection, presented in Figures 7 and 8 may contain predictive information, but the data does not have enough information to assess the sign of the effects with high certainty (probability higher than 95%). At least some predictors that were not included in the model contain useful predictive information, as the predictive performance dropped from -2.75 to -2.88 when variable selection was done.

Also, using average predictive comparisons for variable selection may hide non-linear effects because average predictive comparisons considers only average changes over the whole population.

Bayesian multiple comparisons seems conservative in inferring the relevance of predictors. However, the threshold 95% used in Bayesian multiple comparisons is arbitrary, and lowering it would have included more predictors in the model.

10 Discussion

The predictors from both Gaussian process regression and generalized linear models generally agree with previous research. However, the association between cardiovascular disease and a low rate of sickness absence found with the non-Bayesian generalized linear model has not been found in previous studies.

The current finding may be an artifact of a sample selection bias, which is caused by non-random participation to the underlying intervention study. Cardiovascular disease was self-reported, so biased reporting might also contribute to the finding. Vascular diseases also include varices, which is not assumed to increase sickness absence, although not lower them either, but which might still confound the results.

Our Gaussian process models can probably be improved. For example, zero-inflated Gaussian process models might yield better predictive performance than the hurdle models, but we did not use them because of stability issues. Also, approximating the latent posterior using expectation propagation instead of Laplace approximation is likely to yield better results. The forthcoming version of *GPstuff* toolbox has a more stable implementation of the expectation propagation algorithm, but it was not yet available during the writing of this Thesis.

Many predictors were partly redundant or overlapped, e.g., daytime sleepiness, insomnia, and sleep deprivation. For overlapping predictors, the total effect they have on sickness absence is distributed between them, which can cause each to be considered non-relevant separately and thus not included in the variable selection, although their common effect could be relevant. A factor analysis to discover latent predictors or a careful preselection of the predictors lessen the problem but can hide unexpected effects, for example, in the hypothetical case that daytime sleepiness has an effect on sickness absence that insomnia or sleep deprivation do not convey.

As noted in Section 5, Gaussian process models are well-suited for finding non-linear relationships between the predictors and the response variable and also interactions between the predictors. Based on Figure 8, it seems that there is a saturation effect for depression. On the other hand, we did not find the specific non-linear relationships suggested, e.g., U-shaped relationship between alcohol use and sickness absence, suggested by Vahtera et al. (2002), or the interaction of gender and feeling overloaded at work, found by Kivimäki et al. (2000). This may be due to them using different models, or differences in the study population.

We did the data cleaning and checking painstakingly, so the data quality was good. However, modeling sickness absence is inherently difficult because the distribution of sickness absence days is skewed and long-tailed. Future research could try new models to overcome these difficulties, e.g., as suggested by Taimela et al. (2007), a model with distinct parts for zero sickness absence, low rate of sickness absence, and high rate of sickness absence.

References

[1] Ala-Mursula, L. et al., 2002. Employee control over working times: associations with subjective health and sickness absences. *Journal of Epidemiology & Community Health*, **56**(4), 272–278.

[2] Bishop, C. M., 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

[3] Berk, R., MacDonald, J. M., 2008. Overdispersion and Poisson Regression. *Journal of Quantitative Criminology*, **24**(3), 269–284.

[4] Efron, B., 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26.

[5] Duijts, S. F. et al., 2007. A meta-analysis of observational studies identifies predictors of sickness absence. *Journal of Clinical Epidemiology*, **60**(11), 1105–1115.

[6] Gelman, A., Carlin, J., Stern, H., Rubin, D., 1995. *Bayesian Data Analysis*. 2nd ed. London: Chapman & Hall/CRC.

[7] Gelman, A., Hill, J., 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

[8] Gelman, A., Pardoe, I., 2007. Average predictive comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology*, **37**(1), 23–51.

[9] Gelman, A., Tuerlinckx, F., 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, **15**(3), 373–390.

[10] Gelman, A., 2008. Scaling regression inputs by dividing by two standard deviations *Statistics in Medicine*, **27**(15), 2865–2873.

[11] Kivimäki, M., Elovainio, M., Vahtera, J., 2000. Workplace bullying and sickness absence in hospital staff. *Occupational and Environmental Medicine*, **57**(10), 656–660.

[12] Kivimäki, M., et al., 2001. Sickness absence in hospital physicians: 2 year follow up study on determinants. *Occupational and Environmental Medicine*, **58**(6), 361–366.

[13] Laaksonen, M., Martikainen, P., Rahkonen, O., Lahelma, E., 2008. Explanations for gender differences in sickness absence: evidence from middle-aged municipal employees from Finland. **65**(5), 325–330.

[14] Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1), 1–14.

[15] McCullagh, P., Nelder, J., 1989. *Generalized Linear Models*. 2nd ed. London: Chapman and Hall/CRC.

[16] Minka, T. P., 2001. Expectation propagation for approximate Bayesian inference. *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 362–369.

[17] Murphy, K., Dunham, M., 2008. PMTK: Probabilistic modeling toolkit. *Neural Information Processing Systems (NIPS) Workshop on Probabilistic Programming*.

[18] Niedhammer, I. et al., 1998. Psychosocial factors at work and sickness absence in the Gazel cohort: a prospective study. *Occupational and Environmental Medicine*, **55**(11), 735–741.

[19] R Development Core Team, 2009. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. URL http://www.R-project.org/.

[20] Rasmussen, C. E., Williams, C. K. I., 2006. *Gaussian Processes for Machine Learning*. 2nd ed. Cambridge: MIT Press.

[21] Reijonsaari, K. et al., 2009. The effectiveness of physical activity monitoring and distance counselling in an occupational health setting - a research protocol for a randomised controlled trial (CoAct). *BMC Public Health*, **9**(1), p.494.

[22] Rubin, D. B., 1981. The Bayesian bootstrap. *The Annals of Statistics 1981*, **9**(1), 130–134.

[23] Taimela, S. et al., 2007. Self-reported health problems and sickness absence in different age groups predominantly engaged in physical work. *Occupational and Environmental Medicine*, **64**(11), 739–746.

[24] Vanhatalo, J., Riihimäki, J., Hartikainen, J., Vehtari, A., 2011. Bayesian modeling with Gaussian processes using the MATLAB toolbox GPstuff, *submitted*.

[25] Vahtera, J. et al., 2002. Alcohol intake and sickness absence: a curvilinear relation. *American Journal of Epidemiology*, **156**(10), 969–976.

[26] Vehtari, A., Lampinen, J., 2002. Bayesian model assessment and comparison using cross-validation predictive densities.

[27] Venables, W. N. & Ripley, B. D., 2002. *Modern Applied Statistics with S*. 4th ed. New York: Springer. *Neural Computation*, **14**(10), 2439–2468.

[28] Virtanen, M. et al., 2007. Job strain and psychologic distress influence on sickness absence among Finnish employees. *American Journal of Preventive Medicine*, **33**(3), 182–7.

[29] Zeileis, A., Kleiber, C., Jackman, S., 2008. Regression models for count data in R. *Journal of Statistical Software*, **27**(8), 1–25. URL http://www.jstatsoft.org/v27/i08/.

A Priors

We present here a detailed account of the priors used for Gaussian process models in Section 9 and the lines of code used to implement them in MATLAB toolbox *GPstuff* (Vanhatalo et al., 2011).

For linear covariance function, we used a hierarchical prior of inverse chi-squared distributions,

```
pc_s2 = prior_sinvchi2('s2', 0.01, 'nu', 1);
pc = prior_sinvchi2('s2', 0.01, 'nu', .1, 's2_prior', pc_s2);
```

The corresponding covariance function structure was initialized by

```
gp_cf = gpcf_linear('coeffSigma2', 0.02*ones(1,k), 'coeffSigma2_prior', pc);
```

where k is the number of inputs to the regression.

For neural network covariance function, a Student's t distribution was used as the prior for the square root of the weights,

```
pnn = prior_sqrtt('s2', 10^2);
gp_cf = gpcf_neuralnetwork('weightSigma2', ones(1,k), 'biasSigma2', 1, ...
  'weightSigma2_prior', pnn);
```

Matérn 5/2 and squared exponential covariance functions with the hierarchical prior used inverse chi-squared distributions,

```
pl_s2 = prior_sinvchi2('s2', 1, 'nu', 1);
pl = prior_sinvchi2('s2', 1, 'nu', .2, 's2_prior', pl_s2);
```

The log-Gaussian prior was initialized by

```
pl = prior_loggaussian('s2', 2, 'mu', 1.6);
```

For both priors, a prior for the magnitude term was the Student's t distribution for the square root of the parameter,

```
pm = prior_sqrtt('s2', 4, 'nu', 40);
```

The Matérn 5/2 and squared exponential covariance function structures were initialized by

```
gp_cf = gpcf_matern52('lengthScale', ones(1,k), 'magnSigma2', .5, ...
  'lengthScale_prior', pl, 'magnSigma2_prior', pm);
```

and

```
gp_cf = gpcf_sexp('lengthScale', ones(1,k), 'magnSigma2', .2, ...
  'lengthScale_prior', pl, 'magnSigma2_prior', pm);
```