Bioinformatics, 37(13), 2021, 1860–1867
doi: 10.1093/bioinformatics/btab021
Advance Access Publication Date: 21 January 2021
Original Paper

Genetics and population analysis

lgpr: an interpretable non-parametric method for inferring covariate effects from longitudinal data

Juho Timonen *, Henrik Mannerström, Aki Vehtari and Harri Lähdesmäki*

Department of Computer Science, Aalto University, Espoo 00076, Finland

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on October 2, 2020; revised on December 14, 2020; editorial decision on January 7, 2021; accepted on January 8, 2021

Abstract

Motivation: Longitudinal study designs are indispensable for studying disease progression. Inferring covariate effects from longitudinal data, however, requires interpretable methods that can model complicated covariance structures and detect non-linear effects of both categorical and continuous covariates, as well as their interactions. Detecting disease effects is hindered by the fact that they often occur rapidly near the disease initiation time, and this time point cannot be exactly observed. An additional challenge is that the effect magnitude can be heterogeneous over the subjects.

Results: We present *lgpr*, a widely applicable and interpretable method for non-parametric analysis of longitudinal data using additive Gaussian processes. We demonstrate that it outperforms previous approaches in identifying the relevant categorical and continuous covariates in various settings. Furthermore, it implements important novel features, including the ability to account for the heterogeneity of covariate effects, their temporal uncertainty, and appropriate observation models for different types of biomedical data. The *lgpr* tool is implemented as a comprehensive and user-friendly R-package.

Availability and implementation: *lgpr* is available at jtimonen.github.io/lgpr-usage with documentation, tutorials, test data and code for reproducing the experiments of this article.

Contact: juho.timonen@aalto.fi or harri.lahdesmaki@aalto.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biomedical studies often collect observational longitudinal data, where the same individuals are measured at several time points. This is an important study design for examining disease development and has been extensively leveraged in biomedical studies, including various -omics studies, such as proteomics (Liu *et al.*, 2018), metagenomics (Vatanen et al., 2016) and single-cell transcriptomics (Sharma et al., 2018). The measured response variable of interest can be continuous (such as the abundance of a protein), discrete (such as the number of sequencing reads in a genomic region) or binary (such as patient condition). Often also several additional variables—i.e. covariates—are measured for each subject at each measurement time point. These can be categorical variables (such as sex, location or whether the subject is diagnosed with a disease or not) or continuous (such as age, time from disease initiation or blood pressure). Identifying the relevant covariates that affect the response variable is important for assessing potential risk factors of the disease and for understanding disease pathogenesis.

A large body of literature has focused on the statistical analysis of longitudinal data (Diggle et al., 2002). Observations corresponding to the same individual are intercorrelated, and specialized statistical methods are therefore required. Methods must be able to model both time-dependent and static covariate effects at the same time and handle irregular measurement intervals, missing data and a varying number of measurements for different individuals. Generalized linear mixed models (GLMMs) (Stroup, 2012) have been found to best conform to these challenges, and they have become the standard workhorse for longitudinal data analysis. The R-package *lme4* (Bates et al., 2015) has gained high popularity and become a default choice for fitting GLMMs. These models, however, require specifying a parametric (linear) form for the covariate effects and provide biased inferences when their true effects are non-linear or non-stationary.

GLMMs are an example of additive models, where the modelled function is decomposed as $f = f^{(1)} + \cdots + f^{(J)}$, and each $f^{(j)}$ depends only on a subset of the covariates. The term generalized additive models (Hastie and Tibshirani, 1986) (GAMs) is often used to refer to models where each $f^{(j)}$ depends only on one covariate. GAMs are especially interpretable since the effect of each covariate can be studied independently (Plate, 1999). Examples of non-parametric GAMs are penalized smoothing splines, and their fitting involves penalizing the wiggliness of the functions (Wood, 2006).

A Gaussian process (GP) is a popular Bayesian non-parametric model that is commonly used for time series modelling (Rasmussen and Williams, 2006; Roberts et al., 2013). GPs can model various types of functions, and the prior knowledge about the modelled unknown function is conveniently specified by a covariance (or kernel) function. For example, the exponentiated quadratic (EQ) kernel specifies that the function values are more similar for observations that are closer in time, and the periodic (PER) kernel specifies a repeating pattern.

In additive GPs (Duvenaud et al., 2011; Plate, 1999), the GP prior is defined for several additive components $f^{(j)}$, and they can be tailored also for longitudinal study designs (Cheng et al., 2019; Quintana et al., 2016). *LonGP* (Cheng *et al.*, 2019) is a recent additive GP modelling method that utilizes kernels that can depend not only on time, but possibly categorical factors and other covariates as well. It utilizes the binary mask (BIN) and categorical (CAT) kernel as building blocks, as they can be multiplied with continuous kernels such as EQ or PER to allow modelling effects that are present only for a subgroup of individuals, or effects that are different across individuals or groups. LonGP is specifically designed for detecting relevant covariates, and it employs a two-stage forward search with approximate leave-one-out and stratified cross-validation to add new additive components to an initial model one by one until the model does not improve significantly anymore. Due to computational convenience, GP models such as *LonGP* are often restricted to Gaussian observation model, which is not appropriate for count or proportion data commonly observed in biomedicine. A common approach is to use the Gaussian observation model after first applying a variance-stabilizing transform, such as log-transform, to the response variable, but this is not statistically justified and can lead to biased inferences (O'Hara and Kotze, 2010).

Longitudinal studies often comprise a case and control group, and commonly a clinically determined time of disease initiation for each case individual is marked in the data. To reveal phenomena related to disease progression or to identify biomarkers, statistical modelling can utilize the disease-related age, i.e. time from disease initiation or onset, as one covariate that can explain changes in the response variable. Disease effects can be rapid when compared to other effects and expected to occur near the time of disease initiation, which is another aspect that GLM models cannot capture. In *LonGP*, these effects can be modelled using a non-stationary (NS) kernel. A major challenge, however, is that many diseases, such as Type 1 Diabetes (T1D), are heterogeneous (Pietropaolo et al., 2007), and disease-specific biomarkers are likely detectable only in a subset of the diagnosed individuals. Another problem that can confound the analysis of disease effects, is that the disease initiation (or onset) time is difficult to determine exactly. For example in T1D, the presence of islet cell autoantibodies in the blood is the earliest known marker of disease initiation (Ziegler et al., 2013), but they can only be measured when the subject visits a doctor. In general, the detected disease initiation time can differ from the true initiation time, and the extent of this difference can vary across individuals. To our knowledge, there exist no methods that can model non-stationary disease effects while taking into account the disease heterogeneity and uncertainty of initiation time.

In this work, we propose a longitudinal data analysis method called *lgpr*, designed for revealing general non-linear and non-stationary effects of individual covariates and their interactions (see Fig. 1a). It is based on the additive GP approach similar to *LonGP* but provides several significant improvements that tackle the challenges stated above. We use special interaction kernels that allow separating category effects (e.g. different temporal profiles for male and female subjects) from shared effects. This allows us to develop a straightforward but useful covariate relevance assessment method, which requires fitting only one model and gives estimates of the proportion of variance explained by each signal component and noise. Our package implements additive GP modelling and covariate relevance assessment also in the case of a non-Gaussian observation model and allows incorporating sample normalization factors that account for technical effects commonly present for example in RNA-sequencing data. Additionally, our tool can account for uncertainty in the disease effect time and features a novel kernel that allows identification of heterogeneous effects detectable only in a subset of individuals. For increased interpretability of disease effects, we propose a new variance masking (VM) kernel which separates effects related to disease development from the baseline difference between case and control individuals.

We have implemented *lgpr* as a user-friendly R-package (R Core Team, 2018) that can be used as a plug-in replacement for *lme4*. Under the hood, Bayesian model inference is carried out using the dynamic Hamiltonian Monte Carlo sampler (Betancourt, 2017; Hoffman and Gelman, 2014), as implemented in the high-performance statistical computation framework Stan (Carpenter et al., 2017). The new tool is summarized in Figure 1a, and its improvements over *LonGP* are highlighted in Table 1. More background information and related research can be found in Supplementary Material.

We use simulated data to prove the benefit of each new feature of our method. Additionally, we use *lgpr* to analyse data from two recent T1D studies. The first one is a longitudinal proteomics dataset (Liu *et al.*, 2018) and the second one is RNA-sequencing data from peripheral blood cells (Kallionpää et al., 2019).

2 Materials and methods

2.1 The probabilistic model

We denote a longitudinal dataset with N data points and D covariates by a tuple $(\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X}$ is an $N \times D$ covariate matrix and $\boldsymbol{y}$ is a vector of N response variable measurements. We refer to the ith row of $\boldsymbol{X}$ by $\boldsymbol{x}_i \in \mathcal{X}$, where $\mathcal{X} = \times_{d=1}^{D} \mathcal{X}_d$ and $\mathcal{X}_d$ is the set of possible values for covariate d. In general, $\mathcal{X}_d$ can be discrete, such as the set of individual identifiers, or connected such as $\mathbb{R}$ for (normalized) age.

Our model involves an unobserved signal $f : \mathcal{X} \rightarrow \mathbb{R}$, which is a function of the covariates. The signal is linked to $\boldsymbol{y}$ through a likelihood function, motivated by a statistical observation model for $\boldsymbol{y}$, and uses transformed signal values $g^{-1}\left(f(\boldsymbol{x}_i) + c_i\right)$, where g is a link function and c_i are possible additional scaling factors. We have implemented inference under Gaussian, Poisson, binomial, beta binomial (BB) and negative binomial (NB) likelihoods, and they are defined in detail in Supplementary Material.

The process f is assumed to consist of J low-dimensional additive components, so that $f(\boldsymbol{x}) = f^{(1)}(\boldsymbol{x}) + \cdots + f^{(J)}(\boldsymbol{x})$ (see Fig. 1b and c). Each component j is modelled as a Gaussian process (GP) with zero mean function and kernel function $\alpha_j^2 k_j(\boldsymbol{x}, \boldsymbol{x}')$. This means that the vector of function values $\boldsymbol{f}^{(j)} = [f^{(j)}(\boldsymbol{x}_1), \ldots, f^{(j)}(\boldsymbol{x}_N)]^\top$ has a multivariate normal prior $\boldsymbol{f}^{(j)} \sim \mathcal{N}\left(0, \boldsymbol{K}^{(j)}\right)$ with zero mean vector and $N \times N$ covariance matrix with entries $\{\boldsymbol{K}^{(j)}\}_{ik} = \alpha_j^2 k_j(\boldsymbol{x}_i, \boldsymbol{x}_k)$. Because the components are *a priori* independent, the sum f is also a zero-mean Gaussian process with kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=1}^{J} \alpha_j^2 k_j(\boldsymbol{x}, \boldsymbol{x}\prime)$. See more info about GPs in Supplementary Material or (Rasmussen and Williams, 2006).

The parameter α_j^2 is called the marginal variance of component $f^{(j)}$ and it determines how largely the component varies. The base kernel function $k_j(\boldsymbol{x}, \boldsymbol{x}')$ on the other hand determines the component's shape, as well as covariance structure over individuals or groups (see Fig. 1b and c). The base kernels are constructed, as explained in the next section, so that each $f^{(j)}, j = 1, \ldots, J$ is a function of only one or two covariates. This is a sensible assumption in many real-world applications and apt to learn long-range structures in the data (Duvenaud *et al.*, 2011). Furthermore, this decomposition into additive components allows us to obtain interpretable covariate effects after fitting the model. Duvenaud *et al.* (2011) used also higher-order interaction terms (which we could incorporate into our model as well), but they did not study relevances of individual covariates, as high-order interactions inherently confound their interpretation.

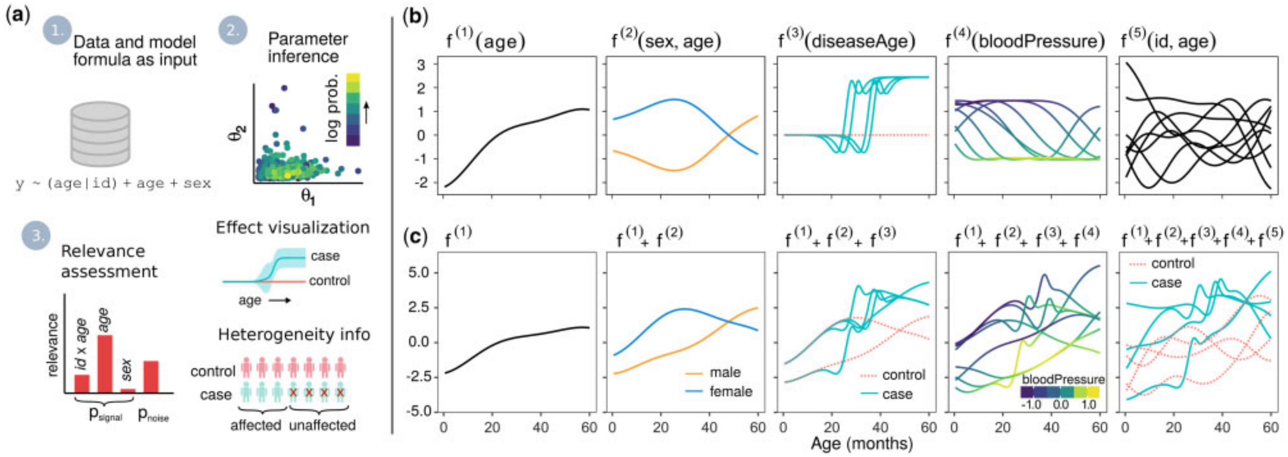

Fig. 1. Overview of additive Gaussian process modelling of longitudinal data using *lgpr*. (**a**) A typical workflow with *lgpr*. *1*. User gives the data and model formula as input, along with possible additional modelling options such as non-default parameter priors or a discrete observation model. *2*. The model is fitted by sampling the posterior distribution of its parameters. *3*. Relevances of different covariates and interaction terms are computed. The inferred signal components can be visualized to study the magnitude and temporal aspects of different covariate effects. If a heterogeneous model component was specified, the results inform how strongly each individual experiences the effect. (**b**) Examples of different types of covariate effects that can be modelled using *lgpr*. The components $f^{(j)}$, $j = 1, \dots, 5$ are draws from different Gaussian process priors. This artificial data comprises 8 individuals (4 male, 4 female), and 2 individuals of each sex are cases. The shown age-dependent components are a shared age effect $f^{(1)}$, a sex-specific deviation $f^{(2)}$ from the shared age effect, a disease-related age (*diseaseAge*) effect $f^{(3)}$ and a subject-specific deviation $f^{(5)}$ from the shared age effect. For each of the diseased individuals, the disease initiation occurs at a slightly different age, between 20 and 40 months. Here, the magnitude of the disease effect is equal for each case individual, but *lgpr* can model also heterogeneous effects. The component $f^{(4)}$ is a function of blood pressure only, but is plotted against age for consistency as the simulated blood pressure variable has a temporal trend. (**c**) The cumulative effect $f = \sum_j f^{(j)}$ is the sum of the low-dimensional components

Table 1. Key differences between *lgpr* and *LonGP*

	lgpr	*LonGP* (Cheng *et al.*, 2019)
Available kernels	BIN, CAT, ZS, EQ, NS (parameterized warping), VM	BIN, CAT, EQ, PER, NS (fixed warping)
Available observation models	Gaussian, Poisson, NB, binomial, BB	Gaussian
Bayesian inference	Dynamic HMC	Slice sampling and CCD (Vanhatalo *et al.*, 2013)
Heterogeneous effects	Available	Not available
Covariate uncertainty	Available	Not available
Covariate relevance assessment	Decomposition of variance	Stepwise model search with crossvalidation

Note: Kernel name abbreviations: BIN, binary mask; CAT, categorical; ZS, zero-sum; EQ, exponentiated quadratic; NS, non-stationary; VM, variance mask; PER, periodic. The input warping steepness (a in Equation 3) is fixed in *LonGP* but sampled in *lgpr*.

2.2 Kernel functions for longitudinal data

2.2.1 Shared effects

Stationary shared effects of continuous covariates are modelled using the exponentiated quadratic (EQ) kernel $k_{\text{eq}}(x, x'|\ell) = \exp\left(-\frac{(x-x\prime)^2}{2\ell^2}\right)$. Here, x refers to a generic continuous covariate, and each shared effect component has its own lengthscale parameter ℓ, which determines how rapidly the component can vary. For example, a shared age effect kernel is $k_{\text{eq}}(x_{\text{age}}, x'_{\text{age}}|\ell_{\text{age}})$.

2.2.2 Category effects

Effects of categorical covariates (such as sex or individual id) can be modelled either as fluctuating category-specific deviations from a shared effect (interaction of a categorical and continuous covariate) or as static category-specific offsets. For a pair of categorical covariate z (with $M \geq 2$ categories) and continuous covariate x, we use the kernel function

$$k_{z \times x}((z, x), (z', x')|\ell) = k_{\text{zerosum}}(z, z') \cdot k_{\text{eq}}(x, x'|\ell), \qquad (1)$$

when modelling the effect of z as deviation from the shared effect of x. The zero-sum kernel $k_{\text{zerosum}}(z, z')$, returns 1 if $z = z'$ and $\frac{1}{1-M}$ otherwise. This is similar to the GP ANOVA approach in (Kaufman and Sain, 2010). If $f : \mathbb{R} \times \{1, \dots, M\} \to \mathbb{R}$ is modelled using the kernel in Equation 1, the sum $\sum_{r=1}^{M} f(t, r)$ is always zero for any t (see proof in Supplementary Material). The fact that the sum over categories equals exactly zero for any t greatly helps model interpretation as this property separates the effect of the categorical covariate from the shared effect (see Supplementary Fig. S1 for illustration). If the effect of z is modelled as a batch or group offset, which does not depend on time or other continuous variables, the corresponding kernel function is just $k_{\text{zerosum}}(z, z')$. Again, z refers to a generic categorical covariate.

We note that the *lgpr* software implementation allows using also the categorical (CAT) kernel in place of k_{zerosum}, when modelling the effects of categorical covariates. This kernel function returns 1 if its arguments belong to the same category and 0 otherwise.

2.2.3 Non-stationary effects

We use the input warping approach (Snoek et al., 2014) to model non-stationary functions $f^{(j)}(x)$, where most variability occurs near the event $x = 0$. The non-stationary kernel is

$$k_{\text{ns}}(x, x'|a, \ell) = k_{\text{eq}}(\omega_a(x), \omega_a(x')|\ell), \qquad (2)$$

where $\omega_a : \mathbb{R} \to]-1, 1[$ is a monotonic non-linear input warping function

$$\omega_a(x) = 2 \cdot \left(\frac{1}{1 + e^{-ax}} - \frac{1}{2}\right), \qquad (3)$$

and the parameter a controls the width of the effect window around $x = 0$.

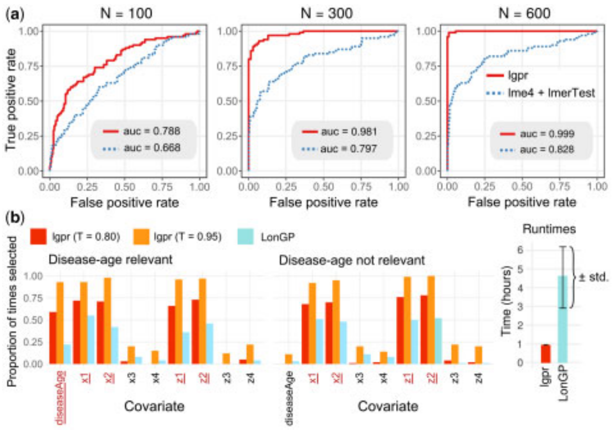

Fig. 2. Covariate relevance assessment comparison with other methods and demonstration of our method's scalability. (**a**) Comparison between *lgpr* and linear mixed effect modelling using the *lme4* and *lmerTest* packages. The panels show ROC curves for the problem of classifying covariates as relevant or irrelevant, when the total number of data points is $N = 100$, 300 and 600, respectively. (**b**) Comparison against *LonGP*. The bar plots show the fraction of times each covariate was chosen in the final model over 100 simulated datasets. The red underlined text indicates the covariates that were relevant in generating the data. The left panel shows results for 100 datasets that includes the disease-related age (*diseaseAge*) as a relevant covariate. The centre panel shows results for 100 simulations where the disease-related age was not a relevant covariate. The right panel shows distribution of runtimes over the total 200 datasets for both methods. The bar lengths are average runtimes, and the turnstiles indicate runtime standard deviations

2.2.4 Disease effects

Cheng *et al.* (2019) modelled disease effects using the kernel in Equation 2 for the disease-related age x_{disAge}, i.e. time from disease initiation or onset of each individual. Note that for the control subjects, x_{disAge} is not observed at all. In general, data for a continuous covariate x can be missing in part of the observations. In such cases, we adopt the approach of (Cheng *et al.*, 2019) and multiply the kernel of x with a binary mask (BIN) kernel which returns 0 if either of its arguments is missing and 1 if they are available.

Whereas this approach can model a non-stationary trend that is only present for the diseased individuals, its drawback is that it can capture effects that are merely a different base level between the diseased and healthy individuals. In order to find effects caused by the disease progression, we design a new kernel

$$k_{\text{vm}}(x, x'|a, \ell) = f^a_{\text{vm}}(x) \cdot f^a_{\text{vm}}(x') \cdot k_{\text{ns}}(x, x'|a, \ell), \tag{4}$$

where $f^a_{\text{vm}}(x) : \mathbb{R} \to]0, 1[$ is a variance mask function that forces the disease component to have zero variance, i.e. the same value for both groups, when $x \to -\infty$. We choose to use $f^a_{\text{vm}}(x) = \frac{1}{1+e^{-a(x-r)}}$, which means that the allowed amount of variance between these groups rises sigmoidally from 0 to the level determined by the marginal variance parameter, so that the midpoint is at $r = \frac{1}{a}\log\left(\frac{h}{1-h}\right)$ and $\omega_a(r) = 2h - 1$. The parameter h therefore determines a connection between the regions where the disease component is allowed to vary between the two groups and where it is allowed to vary over time. In our experiments, we use the value $h = 0.025$. This means, that 95% of the variation in ω_a occurs on the interval $[-r, r]$. The kernels in Equations 2 and 4 combined with the missing value masking, as well as functions drawn from the corresponding GP priors, are illustrated in Supplementary Figure S2.

2.2.5 Heterogeneous effects

To model effects that have the same effect shape but possibly different magnitude for each individual, we define additional parameters $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_Q]$, where Q is the number of individuals and each $\beta_i \in [0, 1]$. Denote $\mathcal{X}_{\text{id}} = \{1, \ldots, Q\}$ and assume two individuals $x_{\text{id}} = q \in \mathcal{X}_{\text{id}}$ and $x'_{\text{id}} = q' \in \mathcal{X}_{\text{id}}$. An effect is made heterogeneous in magnitude by multiplying its kernel by $k_{\text{heter}}(x_{\text{id}}, x'_{\text{id}}|\boldsymbol{\beta}) = \sqrt{\beta_{\text{q}}\beta_{\text{q}'}}$.

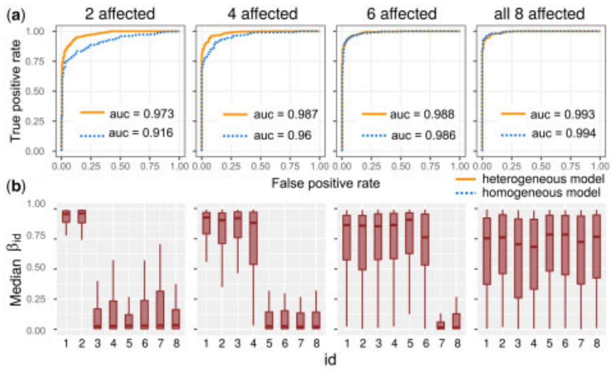

Fig. 3. Heterogeneous disease effect modelling with *lgpr* improves detection of effects that are present only for a subset of case individuals. (**a**) ROC curves for covariate relevance assessment using both a heterogeneous and a homogeneous disease model for simulated data with 2, 4, 6 and 8 out of the 8 case individuals affected, respectively. (**b**) Heterogeneous modelling with *lgpr* can reveal the affected individuals. The boxplots show the distributions of the posterior medians of the individual-specific disease effect magnitude parameters β_{id}, id $= 1, \ldots, 8$ over 100 simulated datasets. The box is the interquantile range (IQR) between the 25th and 75th percentiles, vertical line inside the box is the 50th percentile and the whiskers extend a distance of at most $1.5 \cdot IQR$ from the box boundary. Each panel corresponds to the same experiment as the one above it

For example, to specify a heterogeneous disease effect component, we use the novel kernel

$$k_{\text{heter}}(x_{\text{id}}, x'_{\text{id}}|\boldsymbol{\beta}) \cdot \mathrm{k}_{\text{vm}}(\mathrm{x}_{\text{disAge}}, \mathrm{x}'_{\text{disAge}}|\mathrm{a}, \ell_{\text{disAge}}). \tag{5}$$

For heterogeneous disease effects, the number of needed β parameters equals the number of only the case individuals.

In our implementation, the prior for the unknown parameters $\boldsymbol{\beta}$ is $\beta_i \sim \text{Beta}(b_1, b_2)$, where the shape parameters b_1 and b_2 can be defined by the user. By default, we set $b_1 = b_2 = 0.2$, in which case most of the prior mass is near the extremes 0 and 1 (Supplementary Fig. S3c). This choice is expected to induce sparsity, so that some individuals have close to zero effect magnitude. The posterior distributions of β_i can then be used to make inferences about which case individuals are affected by the disease (β_i close to 1) and which are not (β_i close to 0). The kernel in Equation 5 is illustrated in Supplementary Figure S2c.

We note that the *lgpr* software implementation allows defining also different types of heterogeneous components, by replacing the VM kernel with the EQ or NS kernel in Equation 5, and that multiple heterogeneous components can be included in a model.

2.2.6 Temporally uncertain effects

The presented disease effect modelling approach relies on being able to measure the disease onset or effect time t_{eff} for each case individual, since the disease-related age is defined as $x_{\text{disAge}} = x_{\text{age}} - t_{\text{eff}}$. In Cheng *et al.* (2019), t_{eff} was defined as age on the clinically determined disease initiation date, but in general the effect time can differ from it. Our implementation allows Bayesian inference also for the effect times, and can therefore capture effects that for some or all case individuals occur at a different time point than the clinically determined date. The user can set the prior either directly for the effect times t_{eff}, or for the difference between the effect time and observed initiation time, $\Delta t = t_{\text{obs}} - t_{\text{eff}}$. The first option is suitable if the disease is known to commence at certain age for all individuals. The latter option is useful in a more realistic setting where such information is not available, and it is reasonable to think that the clinically determined initiation time t_{obs} is close to the true effect time.

2.3 Model inference

We collect all marginal variances, lengthscales and other possible kernel hyperparameters in a vector $\boldsymbol{\theta}_{\text{kernel}}$. Parameters of the observation model are denoted by $\boldsymbol{\theta}_{\text{obs}}$ and other parameters such as those related to input uncertainty by $\boldsymbol{\theta}_{\text{other}}$. The collection of all unknown parameters is then $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\text{kernel}}, \boldsymbol{\theta}_{\text{obs}}, \boldsymbol{\theta}_{\text{other}}\}$. Under the hood, *lgpr* uses the dynamic Hamiltonian Monte Carlo sampler with multinomial sampling of dynamic length trajectories (Betancourt, 2017; Hoffman and Gelman, 2014), as implemented in Stan (Carpenter *et al.*, 2017), to obtain S draws from the posterior distribution of $\boldsymbol{\theta}$. The parameters are given robust priors that normalize model fitting (specified in Supplementary Material), and our software includes prior predictive checks that help in prior validation. Our default prior for the steepness parameter a of the input warping function (Equation 3) allows disease effects that occur approximately on a 36 month interval around the disease initiation time. Supplementary Figures S3d and e illustrate the effect of the prior choice for this parameter.

The remaining unknowns of the model are the values of the function components $\boldsymbol{f}^{(j)}$, and their sum $\boldsymbol{f} = \sum_{j=1}^{J} \boldsymbol{f}^{(j)}$. Under the Gaussian observation model, the posterior distributions of $\boldsymbol{f}^{(1)}, \ldots, \boldsymbol{f}^{(J)}$ and $\boldsymbol{f}$ can be derived analytically (see Supplementary Material). With other observation models, we sample the posterior of each $\boldsymbol{f}^{(j)}$ simultaneously with $\boldsymbol{\theta}$.

2.4 Covariate relevance assessment

Our method only requires sampling the posterior of a full model including all covariates. From now on we assume that each continuous covariate can be present in at most one shared effect term and arbitrarily many interactions terms. Its relevance is then interpreted to be the relevance of the shared effect component. The first requirement is not a restriction, and we consider the case of multiple shared effect components in Supplementary Section S2.2.5. We also assume that each categorical covariate can appear only in one term, which can be an interaction or a first-order term, and its relevance is then interpreted to be the relevance of the component where it appears. This way the covariate relevance assessment problem reduces to determining the relevance of each component. In Supplementary Section S2.2.6, we briefly consider also higher-order interaction terms.

2.4.1 Determining the amount of noise

After posterior sampling, we have S parameter draws $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^{S}$ and if using a non-Gaussian observation model, also draws $\{\boldsymbol{f}^{(j,s)}\}_{s=1}^{S}$ of each function component $j = 1, \ldots, J$. For each draw s, our model gives predictions $\boldsymbol{y}_s^* = [y_{1,s}^*, \ldots, y_{N,s}^*] = g^{-1}\left(\boldsymbol{h}^{(s)}\right)$. With the Gaussian observation model, $\boldsymbol{h}^{(s)} = \boldsymbol{\mu}_s$, i.e. the analytically computed posterior mean of $\boldsymbol{f}$ (see Supplementary Material for formula and derivation), and the link function g is identity. With other observation models, $\boldsymbol{h}^{(s)} = \boldsymbol{c} + \sum_{j=1}^{J} \boldsymbol{f}^{(j,s)}$, where $\boldsymbol{c} = [c_1, \ldots, c_N]$ are the scaling factors. Link functions for different observation models are defined in Supplementary Material.

We determine how much of the data variation is explained by noise, using an approach closely related to the Bayesian R^2-statistic (Gelman et al., 2019) (see Supplementary Section 1). The noise proportion in draw s is

$$p_{\text{noise}}^{(s)} = \frac{RSS_s}{ESS_s + RSS_s} \in [0, 1] \tag{6}$$

where $RSS_s = \sum_{i=1}^{N} (y_{i,s}^* - y_i)^2$ and $ESS_s = \sum_{i=1}^{N} (y_{i,s}^* - \overline{y}_s^*)^2$ are the residual and explained sum of squares, respectively, and $\overline{y}_s^* = \frac{1}{N}\sum_{i=1}^{N} y_{i,s}^*$. With this definition, $p_{\text{noise}}^{(s)}$ will be one if the model gives constant predictions and zero if predictions match data exactly. Note that with binomial and BB models, y_i is replaced by y_i/η_i, where η_i is the total the number of trials, as y_i is the number of successes.

2.4.2 Decomposing the explained variance

The proportion of variance that is associated with the actual signal, $p_{\text{signal}}^{(s)} = 1 - p_{\text{noise}}^{(s)}$, can then be seen as explained variance and is further divided between each model component. For cleaner notation, we define the variation of a vector $\boldsymbol{v} = [v_1, \ldots, v_L]$ as a sum of squared differences from the mean, i.e. $SS(\boldsymbol{v}) = \sum_{l=1}^{L} (v_l - v)^2$. The relevance of component j is

$$\text{rel}_j^{(s)} = p_{\text{signal}}^{(s)} \frac{SS_j^{(s)}}{\sum_{j'=1}^{J} SS_{j'}^{(s)}} \tag{7}$$

where $SS_j^{(s)} = SS\left(\boldsymbol{\mu}^{(\mathrm{j,s})}\right)$ with Gaussian observation model and $SS_j^{(s)} = SS\left(\boldsymbol{f}^{(j,s)}\right)$ otherwise. Above we used $\boldsymbol{\mu}^{(\mathrm{j,s})}$ to denote the posterior mean vector of component j, corresponding to draw s (see Supplementary Material). The final component and noise relevances are then

$$\text{rel}_j = \frac{1}{S}\sum_{s=1}^{S} \text{rel}_j^{(s)} \quad \text{and} \quad p_{\text{noise}} = \frac{1}{S}\sum_{s=1}^{S} p_{\text{noise}}^{(s)}, \tag{8}$$

i.e. averages over the S MCMC draws. Our definition has the properties that $\text{rel}_j \in [0, 1]$ for all j, and that we can compute the proportion of variance explained by a subset of components $\mathcal{J} \subseteq \{1, \ldots, J\}$ simply as $\text{rel}_{\mathcal{J}} = \sum_{j \in \mathcal{J}} \text{rel}_j$. Furthermore, $p_{\text{noise}} + \sum_{j=1}^{J} \text{rel}_j = 1$ and for two component subsets $\mathcal{J} \subseteq \mathcal{K}$, it holds that $\text{rel}_{\mathcal{J}} \leq \text{rel}_{\mathcal{K}}$.

2.4.3 Covariate selection

We prefer reporting the numerical relevance values (rel_j) as a summary of how much effect each covariate has on the response variable, instead of classifying each covariate as either relevant or irrelevant. However, we also provide a method for performing covariate selection. The approach is to select the minimal subset of components $\mathcal{J}_{\text{sel}}$ that together with noise explain at least T% of variance. Formally, $\mathcal{J}_{\text{sel}} = \text{argmin}_{\mathcal{J}} |\text{rel}_{\mathcal{J}}|$, subject to $\text{rel}_{\mathcal{J}} + p_{\text{noise}} \geq \frac{T}{100}$ and $T = 95$ by default. In Supplementary Material, we describe also a probabilistic extension of this method. We emphasize that when selecting covariates, we are not testing whether or not a given effect is exactly zero. Therefore we do not perform multiple testing corrections, as in frequentist literature, when analysing multiple response variables (several proteins or genes). See Gelman et al. (2012) for discussion.

A related method, which also relies on selecting a minimal subset of covariates based on inference of a full model with all covariates, is the projection predictive model selection method (Goutis and Robert, 1998). It has been shown to perform well in predictive covariate selection for generalized linear models (Piironen and Vehtari, 2017). However, it still requires comparing lots of alternative submodels to the full model, whereas in our case finding the minimal subset of predictors does not require additional sampling or parameter fitting. Moreover, sequential subset search methods, such as the projection predictive method and *LonGP* (Cheng *et al.*, 2019), are prone to most often selecting the most expressive components. We argue that our method is more suitable for longitudinal GP models that contain components of different complexities. For example, an individual-specific age component is more expressive than a shared age effect component.

3 Results

3.1 Experiments with simulated data

First, we use simulated data to demonstrate the accuracy of covariate relevance assessment and benefits of the novel features of our method. In each experiment, we generate data with different types

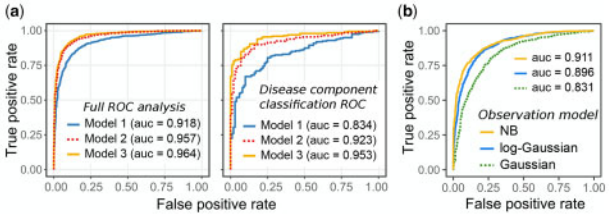

Fig. 4. (**a**) Modelling the uncertainty in the disease effect time enhances covariate relevance assessment accuracy, when data is generated so that the disease effect can occur earlier than the observed disease initiation. The left panel shows ROC curves for covariate relevance assessment with and without modelling the effect time uncertainty. In Model 1, the effect time is fixed to equal the observed initiation time, whereas Models 2-3 account for its uncertainty. Model 2 has an exponential decay prior for the difference between the effect time and observed onset. Model 3 has an oracle prior for the effect time. The right panel shows ROC curves for the same three models, in the task of classifying just the disease component as relevant or irrelevant. (**b**) Using a discrete observation model improves covariate selection accuracy for negative binomially distributed count data

of continuous and categorical covariates (see Supplementary Material for details of data simulation). In order to test the accuracy of our covariate relevance assessment, we simulate noisy measurements of a response variable so that only part of the covariates are relevant. In each experiment we generate several random dataset realizations and measure performance in classifying covariates as relevant or irrelevant using the area under curve (AUC) measure for receiver operating characteristic (ROC) curves. Higher AUC value indicates better performance. The computed covariate relevances (rel_j in Equation 8) are used as a score in the ROC analyses, which are performed using the *pROC* package (Robin et al., 2011).

3.1.1 Comparison with linear mixed effect modelling and LonGP

We first confirm that linear mixed modelling cannot capture the covariate relevances whereas our GP modelling approach can, when the covariate effects are non-linear. We use the *lme4* package (Bates *et al.*, 2015) for fitting linear mixed effect models, and the *lmerTest* package (Kuznetsova et al., 2017) for computing p-values of the linear model components. The p-values are used as the score in ROC analysis. The resulting ROC curves and AUC scores are shown in Figure 2a. It is evident that the linear mixed model approach performs poorly, whereas *lgpr* is consistently more accurate, reaching near-perfect performance when $N = 600$. To test the effect of the amount of noise, we repeat the experiment with $N = 100$, using different signal-to-noise ratios. Results are in Supplementary Figure S6a, and they show that the accuracy of *lgpr* improves consistently as the data is less noisy.

We also compare our method with the additive Gaussian process model selection method *LonGP* (Cheng *et al.*, 2019). Here, we set up a more difficult covariate selection problem with more covariates of different types, and also generate non-stationary disease effects for half of the individuals. Since *LonGP* uses a sequential model search, we cannot compute full ROC curves for it. Therefore we compare performances by counting how often each covariate is selected. *LonGP* tends to select very few covariates, and to have comparable results for *lgpr*, we set a rather low threshold of $T = 80$. Figure 2b shows the number of times each method selected different covariates across the 100 simulated datasets for both the case where the disease effect was and was not relevant. We see that *lgpr* can more clearly distinguish the relevant covariates. Furthermore, the average run time per dataset is approximately five times smaller for *lgpr* (Fig. 2b).

In addition to $T = 80$, we include results with the default *lgpr* threshold of $T = 95$. The total covariate selection accuracies for *lgpr* using both thresholds as well as *LonGP*, are shown in Supplementary Table S1. In additional experiments, we repeat the experiment with a higher signal-to-noise ratio, and test the behaviour of the methods when some of the case individuals are mistakenly modelled as controls. For these additional experiments, the accuracies are reported in Supplementary Table S1 and proportion of times each covariate is selected is reported in Supplementary Figure S6b.

3.1.2 Heterogeneous and temporally uncertain disease effect modelling

To test the heterogeneous disease effect modelling approach, we generate data with 16 individuals out of which 8 are cases, but so that the disease effect is generated for only $N_{\text{affected}} = 2, 4, 6$ or 8 of the case individuals. For each dataset replication, the inference is done using both a heterogeneous and homogeneous model. The results in Figure 3 show that heterogeneous modelling improves covariate selection accuracy, and the improvement is clearest when $N_{\text{affected}} = 2$. Moreover, in heterogeneous modelling, the posterior distribution of the individual-specific disease effect magnitude parameters β_{id} indicates the affected individuals. See Supplementary Figure S4 for a detailed demonstration of heterogeneous model inference.

To test the model where the disease effect time is considered uncertain, we simulated data where the observed disease initiation time is later than the true generated effect of the disease-related age covariate. For each dataset we run the inference first by fixing the effect time to equal the clinically determined onset time (Model 1), and then using two different priors for the effect time uncertainty. The first prior is $\Delta t \sim \text{Exp}(0.05)$, meaning that the observed onset is most likely, and prior mass decays exponentially towards birth (Model 2). An oracle prior, which is exactly the distribution that is used to sample the real effect time, is used for reference (Model 3). The results in Figure 4a show that the uncertainty modelling improves the covariate selection accuracy, and the oracle prior performs best as expected. Especially, we see that detection of the disease-related age covariate is more accurate when the uncertainty is being modelled. See Supplementary Figure S5 for a more specific demonstration of effect time inference.

3.1.3 Non-Gaussian data

To demonstrate the benefit of using a proper observation model for count data, we generate negative binomially distributed data and run the inference using both a Gaussian and NB observation model. For reference, we also run the inference using the Gaussian observation model after transforming the counts through mapping $y \mapsto \log(1 + y)$. Results in Figure 4b confirm that using the correct observation model in *lgpr* for this kind of count data improves covariate selection accuracy compared to the Gaussian or log-Gaussian models. We note, however, that covariate selection performance of the log-Gaussian model improves (relative to that of the NB model) when data has higher count values and dispersion is smaller, i.e. when the NB model is better approximated by the log-Gaussian model.

3.1.4 Experiments under model mismatch

In the previous experiments, the true covariate effects were drawn from the additive GP priors. To validate the modelling approach further, we perform an additional experiment where the true effects are different types of parametric functions (Supplementary Section S3.6, Supplementary Fig. S7a). The experiment setup is otherwise the same as in the comparison against *LonGP* (Section 3.1.1). We observed that the component relevances and amount of noise are captured well also in this case (Supplementary Fig. S7b). We also test the behaviour of the model and covariate relevance assessment method when some of the true relevant covariates are not included in the model. This is a likely scenario in reality, because some true relevant predictor variables are not necessarily even measured, and therefore cannot be included in the model. Results in Supplementary Figure S7b suggest that failure to include relevant components in the model has the effect that more variance is explained by noise and individual-specific variation. This is expected behaviour, and we see that the other true relevant covariates can still be distinguished from the irrelevant ones, but modellers should be aware of this behaviour and interpret results accordingly.

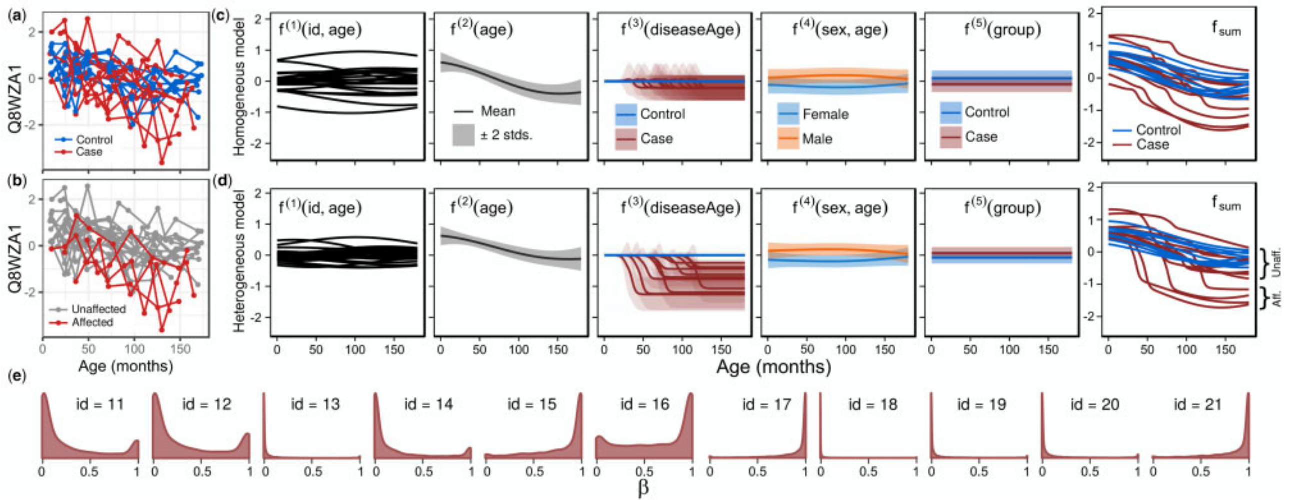

Fig. 5. Results of analysing one example protein from a longitudinal proteomics dataset. (**a**) The normalized measurements for protein Q8WZA1, highlighted based on group (case or control). The lines connect an individual. (**b**) Same data where four case individuals (id=15, 16, 17, 21) are highlighted, based on being determined as affected by the disease in heterogeneous modelling. (**c**) Inferred function components, as well as their sum f (using posterior mean parameters), for Q8WZA1 analysed using the homogeneous and (**d**) heterogeneous model. The component relevances (rel$_j$ in Equation 8) for each $f^{(j)}$, $j = 1, \ldots, 5$ are $0.229, 0.157, 0.03, 0.031, 0.007$ for the homogeneous model and $0.096, 0.116, 0.25, 0.037, 0.004$ for the heterogeneous model, respectively. The heterogeneous model selects the disease component as relevant, whereas the homogeneous model does not. The posterior distributions of the function components and their sum outside observed time points is computed as explained in Supplementary Material. For clarity, standard deviations are not show for $f^{(1)}$ and f_{sum}. (**e**) Kernel density estimates for the posterior distributions of the individual-specific disease effect magnitude parameters of the heterogeneous model

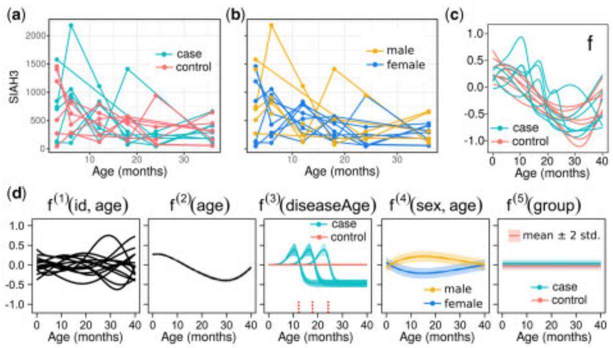

Fig. 6. Data and inferred covariate effects for the SIAH3 gene. (**a**) Raw count data highlighted based on group (case/control) and (**b**) sex. (**c**) Inferred cumulative effect f and (**d**) additive function components. Interpolation outside observed time points is done as explained in Supplementary Material. For clarity, standard deviations are not show for $f^{(1)}$ and f. The seroconversion times of the seven case individuals, i.e. used disease effect times, are 12, 12, 18, 24, 18, 12 and 18 months, indicated by the dashed red vertical lines. Inferred component relevances for $f^{(j)}$, $j = 1, \ldots, 5$ are $0.097, 0.098, 0.077, 0.043, 0.015$, respectively. The selected covariates are id, age, diseaseAge and sex

3.2 Longitudinal proteomics data analysis

We used *lgpr* to analyse a longitudinal dataset from a recent T1D study (Liu *et al.*, 2018), where the longitudinal profile of protein intensities from plasma samples was measured for 11 cases and 10 controls at nine time points that span the initiation of the disease pathogenesis, resulting in a total of 189 data points for most proteins. We chose to analyse 1538 proteins which were chosen by requiring that at least 50% of the measurements must have non-missing values. The exact sample sizes after discarding missing data for each protein are shown in Supplementary Table S2. Eleven children developed T1D, and for those individuals we defined the disease effect time to be the seroconversion age, which was defined as age at the first detection of one or multiple T1D autoantibodies (Liu *et al.*, 2018). We performed our modelling using five covariates: id, age, diseaseAge, sex and group (case/control). We followed the preprocessing described in (Liu *et al.*, 2018) to get normalized protein intensities. Of the categorical covariates, id and sex are modelled as age-dependent category-specific deviations from the shared age effect, and group is a constant group offset variable.

Covariate relevances and selection results for all proteins are included in Supplementary Tables S2 and S3. As an example, both models confirm the sex association of the Mullerian inhibiting factor (uniprot id P03971) (Liu *et al.*, 2018), assigning a relevance score of 0.912 for the *sex* × *age* interaction term. The homogeneous model finds 38 and the heterogeneous model finds 66 proteins associated with the disease-related age covariate, with intersection of 20 proteins. Figure 5a shows the normalized measurements for protein Q8WA1 and Figures 5c and d show the inferred covariate effects using the two different disease effect modelling approaches. The new heterogeneous modelling approach is seen to detect a stronger average disease effect, because it allows the effect sizes to vary between individuals. Moreover, the posterior distributions of individual-specific disease effect magnitude parameters (Fig. 5e), reveal four individuals ($id = 15, 16, 17, 21$) (Fig. 5b), that experience a strong disease effect near the seroconversion time.

3.3 Longitudinal RNA-seq data analysis

We analysed also read count data from CD4+ T cells of 14 children measured at 3, 6, 12, 18, 24 and 36 months age (Kallionpää *et al.*, 2019). The number of available data points was 6 (for 8 children), 5 (2 children), 4 (2 children) or 3 (2 children), resulting in a total of 72 data points. Seven children had become seropositive for T1D during the measurement interval (cases), while the other seven children were autoantibody negative (controls). We included 519 highly variable genes in our *lgpr* analysis, based on preprocessing steps explained in Supplementary Material. We included the same covariates and components in our *lgpr* model as in the proteomics data analysis, and age at the first detection of one or more T1D autoantibodies was again used to compute the disease related age.

Covariate relevances and selection results for all genes are included in Supplementary Table S4. Our analysis confirms the differential expression profile of the IL32 gene between the case and control individuals (Kallionpää *et al.*, 2019), as the group covariate is selected with relevance 0.196. The disease-related age was initially selected as relevant for a total of 73 genes. As the data is sparse and noisy, we defined a stricter rule and required that the relevance of the disease-related age component alone is larger than 0.05. This way we detected 12 interesting, potentially disease development-related genes (highlighted in blue in Supplementary Table S4). As an example, Figure 6 shows the inferred covariate effects for the SIAH3 (*Seven in absentia homolog 3*) gene.

4 Conclusions

The *lgpr* tool provides several important novel features for modelling longitudinal data and offers a good balance between flexibility and interpretability. We have shown that the interpretable kernels, heterogeneous disease modelling, uncertainty modelling of effect times and covariate selection strategy of *lgpr* significantly improve previous longitudinal modelling methods. The tool has an intuitive syntax, and thus provides an easy transition from the standard linear mixed modelling tools to Bayesian non-parametric longitudinal regression. It is widely applicable as the data can involve irregular sampling intervals, different numbers of measurement points over individuals and crossed categorical factors. Moreover, many types of response variables that are common in post-genomic studies (continuous, discrete, binary, proportion) can be modelled with the proper observation models. The comprehensive software implementation of *lgpr* enjoys state-of-the-art sampling efficiency and diagnostics (Vehtari et al., 2020) offered by Stan. The user can choose from the numerous presented modelling options and set various parameter priors (which have widely applicable defaults). Overall, *lgpr* has the potential to become a standard tool for statistical analysis of longitudinal data.

Acknowledgements

The authors acknowledge the computational resources provided by the Aalto Science-IT project.

Funding

This work was supported by the Academy of Finland [292660].

Conflict of Interest: none declared.

Data availability

No new data were generated or analysed in support of this research.

References

Bates,D. *et al.* (2015) Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, **67**, 1–48.

Betancourt,M. (2017) A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434*.

Carpenter,B. *et al.* (2017) Stan: a probabilistic programming language. *J. Stat. Softw.*, **76, 1–32**.

Cheng,L. *et al.* (2019) An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. *Nat. Commun.*, **10**, 1798.

Diggle,P. *et al.* (2002) *Analysis of Longitudinal Data*. Oxford University Press, United Kingdom.

Duvenaud,D.K. *et al.* (2011) Additive Gaussian processes. *Adv. Neur. Inf. Proc. Syst.*, **24**, 226.

Gelman,A. *et al.* (2012) Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.*, **5**, 189–211.

Gelman,A. *et al.* (2019) R-squared for Bayesian regression models. *Am. Stat.*, **73**, 307–309.

Goutis,C. and Robert,C.P. (1998) Model choice in generalised linear models: a Bayesian approach via Kullback-Leibler projections. *Biometrika*, **85**, 29–37.

Hastie,T. and Tibshirani,R. (1986) Generalized additive models. *Stat. Sci.*, **1**, 297–310.

Hoffman,M.D. and Gelman,A. (2014) The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, **15**, 1593–1623.

Kallionpää,H. *et al.* (2019) Early detection of peripheral blood cell signature in children developing β-cell autoimmunity at a young age. *Diabetes*, **68**, 2024–2034.

Kaufman,C.G. and Sain,S.R. (2010) Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Anal.*, **5**, 123–149.

Kuznetsova,A. *et al.* (2017) lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.*, **82**, 1–26.

Liu,C.-W. *et al.* (2018) Temporal expression profiling of plasma proteins reveals oxidative stress in early stages of Type 1 Diabetes progression. *J. Proteomics*, **172**, 100–110.

O'Hara,R.B. and Kotze,D.J. (2010) Do not log-transform count data. *Methods Ecol. Evol.*, **1**, 118–122.

Pietropaolo,M. *et al.* (2007) The heterogeneity of diabetes. *Diabetes*, **56**, 1189–1197.

Piironen,J. and Vehtari,A. (2017) Comparison of Bayesian predictive methods for model selection. *Stat. Comput.*, **27**, 711–735.

Plate,T. (1999) Accuracy versus interpretability in flexible modeling: implementing a tradeoff using Gaussian process models. *Behaviourmetrika*, **26**, 29–50.

Quintana,F.A. *et al.* (2016) Bayesian nonparametric longitudinal data analysis. *J. Am. Stat. Assoc.*, **111**, 1168–1181.

R Core Team (2018) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen,C.E. and Williams,C.K.I. (2006) *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts.

Roberts,S. *et al.* (2013) Gaussian processes for time-series modelling. *Phil. Trans. R. Soc. A*, **371**, 20110550.

Robin,X. *et al.* (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**,

Sharma,A. *et al.* (2018) Longitudinal single-cell RNA sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy. *Nat. Commun.*, **9**, 4931.

Snoek,J. *et al.* (2014) Input warping for Bayesian optimization of non-stationary functions. *Int. Conf. Mach. Learn.*, **31**, 1674–1682.

Stroup,W.W. (2012) *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Chapman & Hall/CRC Texts in Statistical Science*. CRC Press, Boca Raton, FL.

Vanhatalo,J. *et al.* (2013) GPstuff: Bayesian modeling with Gaussian processes. *J. Mach. Learn. Res.*, **14**, 1175–1179.

Vatanen,T. *et al.* (2016) Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell*, **165**, 842–853.

Vehtari,A. *et al.* (2020) Rank-normalization, folding, and localization: an improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Anal.* Advance publication, doi:10.1214/20-BA1221

Wood,S. (2006) *Generalized Additive Models: An Introduction with R. Texts in Statistical Science*. Chapman & Hall, United Kingdom.

Ziegler,A.G. *et al.* (2013) Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children. *J. Am. Med. Assoc.*, **309**, 2473–2479.