GAUSSIAN PROCESS PANEL MODELING — MACHINE LEARNING INSPIRED ANALYSIS OF LONGITUDINAL PANEL DATA

Julian D. Karch

leiden university, the netherlands
and
max planck institute for human development, germany

Andreas M. Brandmaier

max planck institute for human development, germany
and
max planck ucl centre for computational psychiatry and ageing research, germany

Manuel C. Voelkle

humboldt university berlin, germany
and
max planck institute for human development, germany

Abstract

In this article, we extend the Bayesian nonparametric regression method *Gaussian Process Regression* to the analysis of longitudinal panel data. We call this new approach *Gaussian Process Panel Modeling (GPPM)*. GPPM provides great flexibility because of the large number of models it can represent. It allows classical statistical inference as well as machine learning inspired predictive modeling. GPPM offers frequentist and Bayesian inference without the need for commonly used Markov chain Monte Carlo-based approximations, which makes the approach exact and fast. GPPMs are defined using the kernel-language, which can express many traditional modeling approaches for longitudinal data, such as linear structural equation models, multilevel models, or state-space models but also machine learning approaches. As a result, GPPM is uniquely able to represent hybrid models combining traditional parametric longitudinal models and nonparametric machine learning models. In the present paper, we introduce GPPM and illustrate its utility through theoretical arguments as well as simulated and empirical data.

Key words: longitudinal data analysis; machine learning; statistical learning; Bayesian statistics; continuous-time modeling; prediction

1. Introduction

Longitudinal data are crucial for addressing a host of psychological research questions, including questions related to child development, aging, and intervention research. In this paper, we focus on the analysis of (longitudinal) panel data, which we define as data that contain measurements of one or more variables from multiple individuals, each measured at multiple time points. Based on this rather broad definition, panel data encompass intensive longitudinal data, which are characterized by a relatively large number of measurements from few individuals (e.g., Walls & Schafer, 2006), as well as traditional panel data sets, which are characterized by few measurements from a relatively large number of individuals (e.g., Hsiao, 2014).

In psychological research, panel data are commonly analyzed using the general linear model (Cohen, 1968), multilevel modeling (Raudenbush & Bryk, 2001), or structural equation modeling (Bollen, 1989). Such approaches have the advantage that specification, inference, and interpretation are straightforward and well understood. However, these benefits come at the price that only relatively simple models can be expressed. Often, for example, the assumption of linear relationships between all variables is central. In addition, traditional modeling approaches are almost focused on explanatory data analysis (Shmueli, 2010; Yarkoni & Westfall, 2017). The main goal of explanatory data analysis is to estimate (parameters of) the probability distribution that generated the data, and thus, to recover relationships that hold in the population, although these relationships do not necessarily have to be causal. To this end, a model is formulated and assumed to be correctly specified, that is, to contain the population distribution. Consequently, the statistical conclusions drawn from an explanatory analysis (e.g., standard errors, p-values, confidence intervals) are only guaranteed to be valid if the chosen model is correct, which arguably is often not the case when analyzing panel data (e.g., Ghisletta et al., 2019).

In contrast, machine learning, with its underlying inference framework of statistical learning, changes the goal of the analysis to quantifying how well a certain model predicts. Prediction may be a valuable goal in itself (e.g., prediction of treatment success or risk of developing a disease) but, also, prediction may help to generate or improve explanatory models, e.g., by providing a reference model that a purely theory-driven model has to compete with (in terms of predictive accuracy) or by providing information as to where in the input space a theory-driven model is making unsatisfying predictions (Brandmaier, Prindle, McArdle, & Lindenberger, 2016). This shift towards predictive modeling enables relatively complex models, and inferences based on the statistical learning (SL) framework do not require correctness of the model (Breiman, 2001). For example, the standard method for classification in psychology is linear logistic regression whereas in statistical learning support vector machines with a Gaussian radial basis function kernel (Vapnik, 1998) are often used, which, in contrast to linear logistic regression, allows for nonparametric models including interactions and higher-order relationships of outcomes and predictors. Inferences about the generalization performance based on statistical learning from these relatively complex models are also valid when the model is not correctly specified. As a matter of fact, in machine learning, models are often misspecified on purpose to obtain better predictions. This idea becomes particularly evident in regularization, in which parameter estimates are biased (shrunken towards zero away from their unbiased estimates) to decrease the variance of the estimates, which ultimately can lead to improved predictive accuracy (cf. the

bias-variance tradeoff, see Yarkoni et al., 2017 for a detailed description).

One statistical learning method that recently has been promoted as a useful analysis tool in psychological research is Gaussian process regression (GPR) (Schulz, Speekenbrink, & Krause, 2018). Additionally, many publications (e.g., Brahim-Belhouari & Bermak, 2004; Roberts et al., 2013; Saatçi, Turner, & Rasmussen, 2010; Turner, 2012) demonstrate the utility of GPR for analyzing time-series data. However, GPR cannot be easily used for the analysis of panel data. The reason is that there are currently no means to accommodate the nested nature of the data (typically, time points within persons). In this article, we extend GPR to allow for the analysis of panel data and call the resulting method Gaussian process panel modeling (GPPM). To this end, we extend GPR such that both a within-person model and a between-person model can be specified. We adapt the statistical learning inference methods used in GPR for the resulting class of GPPM models, which provides us with methods for model selection, methods to obtain person-specific predictive distributions, and methods for model validation. We provide an implementation of GPPM in form of the R (R Core Team, 2018) package 'gppm' (REMOVED TO MAINTAIN REVIEW INTEGRITY).

Although we strongly believe psychological research can profit from incorporating ideas from statistical learning (see also, Brandmaier et al., 2016; Yarkoni & Westfall, 2017), the GPPM approach proposed in the present paper can also be used for explanatory data analysis. By expanding the class of possible models, GPPM may be equally beneficial for explanatory data analysis, because the ability of GPPM to specify a broad set of models might increase chances to specify a correct model. However, given its roots in statistical learning, frequentist inference procedures for GPPM – most notably hypothesis testing and confidence interval estimators for model parameters but also methods for model selection – have not yet been developed. To close this gap, we develop the standard frequentist inference procedures for GPPM in the present paper. As a result, GPPM may be conceived as a hybrid of a statistical learning and an explanatory approach that allows inference using both frameworks. Importantly, in contrast to the statistical learning conclusions, explanatory conclusions drawn from GPPM are not robust to misspecification.

GPPM is based on the so-called kernel-language for model specification. Kernels are functions that generate model-based covariances of pairs of measurements in continuous time and will be explained in more detail later on. The kernel-language builds on the concepts of the Mercer kernel (Rasmussen & Williams, 2006), which is used by many statistical learning methods such as GPR or support vector machines (Vapnik, 1998). From the perspective of longitudinal modeling, the kernel-language represents a new approach for specifying a within-person model, and thus complements the two existing approach (see, Ram & Grimm, 2015, for an overview): mathematical functions, as used in multilevel models and structural equation models, and differential equations, as used in state-space models. Importantly, the kernel-language can represent models that are not representable by either of the two existing approaches; most notably flexible nonparametric models as typically used in machine learning. However, the kernel-language is also able to represent traditional model classes such as (linear Gaussian) structural equation models or (linear) multilevel models. Additionally, a specific strength of the kernel-language is the ability to combine models by standard combination operators easily. Consequently, GPPM enables the researcher to employ models typically used in statistical learning, models that are commonly

used in psychological research, as well as a combination of the former two. We will place a particular emphasis on the utility of these combined, hybrid models, which GPPM is uniquely able to represent.

There have been previous efforts to extend GPR for $N > 1$ data. In this regard, the work by Cox, Kachergis, and Shiffrin (2012) is closest to our approach. They have adapted GPR for the analysis of computer mouse trajectories, which are nested within participants, which again are nested within conditions. Thus, these data can be considered an example of nested longitudinal data and consequently, their work as an example of using GPR for analyzing longitudinal data. However, Cox et al. (2012) tailored GPR to their specific analysis problem, whereas we aim at giving a broader perspective on GPR as a general method for panel data analysis. In addition, GPPM and the approach proposed by Cox differ with regard to model specification, estimation, and model selection. While Hall, Müller, and Yao (2008) did not discuss how to extend GPR for $N > 1$, they used Gaussian processes as a mathematical tool to implement a functional analysis method for panel data. Given this entirely different focus, their method is very different from GPPM as introduced in this paper.

The remainder of this paper is structured as follows. In the next section, Section 2, we recapitulate the statistical learning method GPR as Bayesian nonparametric regression approach. In Section 3, we introduce GPPM, our extension of GPR models for the analysis of panel data. In this section, we also discuss the relationship of the GPPM model class with other modeling classes; specifically, we show that both linear Gaussian structural equation modeling and linear state-space models are subsets of the GPPM model class. In Section 4, we develop frequentist inference procedures, such as hypothesis testing and confidence interval estimators, for the GPPM model class. In Section 5 we adapt the statistical learning inference procedures from GPR to GPPMs. In Section 6, we illustrate the use of GPPM based on both simulated and a real panel data set in which participants' stance towards authoritarianism was modeled. In our demonstration, we focus on the utility of hybrid models of parametric and nonparametric kernels that GPPM is uniquely able to represent. We close with a discussion and conclusion section.

2. Gaussian Process Regression

2.1. Introduction

In this section, we briefly review GPR, which is an established statistical learning method. For an in-depth treatment, see, Rasmussen and Williams (2006) and for a tutorial introduction aimed at psychologists see, Schulz et al. (2018). GPR is based on multiple linear regression. In multiple regression, the goal is to find a regression function of the form

$$f : \mathcal{X} \to \mathbb{R}, \quad f(x) = x^\top b, \quad Y(x) = f(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2).$$

The input vector $x \in \mathcal{X} \subseteq \mathbb{R}^p$ contains p predictors and the parameter vector $b \in \mathbb{R}^p$ p corresponding parameters. We assume that the input vector always contains a constant predictor such that an explicit intercept is not needed. The outcome variable $Y(x)$ represents the to-be-predicted quantity, which is assumed to vary across the predictions $f(x)$ according to a Gaussian random variable ϵ with error variance σ_ϵ^2. The distribution for the outcome variable

$Y(x)$ implied by its input vector x is thus $Y(x) \sim \mathcal{N}(f(x), \sigma_\epsilon^2)$; and for any two input vectors x, x' with $x \neq x'$, $\mathrm{Cov}\left(Y(x), Y(x')\right) = 0$.

The first step towards GPR is to extend the linear regression model such that it allows for nonlinear relationships between the input vector x and the outcome variable (OV). This is achieved by the introduction of a function $\phi(x)$ that maps the input vector x into a new space, which changes the regression function to

$$f(x) = \phi(x)^\top b.$$

The second step towards GPR is to employ Bayesian inference. A prior distribution is introduced for the parameters. A prior is only imposed on the coefficient vector b and is assumed to be Gaussian: $b \sim \mathcal{N}(\mu_b, \Sigma_b)$. The error variance σ_ϵ^2 is assumed to be a fixed quantity, which is considered to be part of the model and thus estimated as part of the model selection procedure (see Section 2.3).

The third step towards GPR is to describe the prior directly at the level of the regression function. Since every value v of the coefficient vector b translates to one particular regression function via the equation $f(x|b = v) = \phi(x)^\top v$, imposing a prior on the coefficient vector b implies a prior over regression functions. Specifically, for a matrix $X = [x_1, \ldots, x_N]$, containing input vectors as columns, the prior implied for the corresponding values of the regression functions $f(X) = [f(x_1), \ldots, f(x_n)]^\top$ can be compactly described using the matrix of transformed input vectors $\phi(X) = [\phi(x_1), \ldots, \phi(x_N)]$ as follows

$$f(X) = \phi(X)^\top b \sim \mathcal{N}\left(\phi(X)^\top \mu_b, \phi(X)^\top \Sigma_b \phi(X)\right).$$

Thus, the prior implied for the predictions of the regression function $f(x)$ at a finite set of input vectors X can be described directly using a Gaussian distribution.

However, typically the set of possible input vectors $\mathcal{X} \ni x$ is of infinite size (e.g., time is generally considered infinite). To fully describe the prior on the level of the regression functions, the distribution of the infinite set $\{f(x) : x \in \mathcal{X}\}$ has to be appropriately represented. This set is not a random vector because it is of infinite size and, consequently, its prior distribution cannot be described using a Gaussian distribution. Thus, we need to operate with an infinite-sized generalization of a random vector, which is called a stochastic process.

Definition 1 (Gaussian Process)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* (S, Σ) *a measurable space, a* stochastic process *is a set of* S*-valued random variables on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*. It can be written as* $\{f(x) : x \in \mathcal{X}\}$ *using an index set* $\mathcal{X}$*.*

A Gaussian process *is a stochastic process for which any finite subset of* $\{f(x) : x \in \mathcal{X}\}$ *(which is a random vector) is distributed according to a Gaussian distribution.*

Thus, to completely describe the prior over regression function, the distribution for the Gaussian process $\{f(x) : x \in \mathcal{X}\}$ needs to be specified. Just like for Gaussian random vectors, the distribution of a Gaussian process can be completely described by its first and second-order statistics. While for Gaussian random vectors a mean vector and a covariance matrix are used, for Gaussian processes their infinitely sized equivalents are employed; the mean function and the (Mercer) kernel.

Definition 2 (Mean Function and Kernel)**.** *Let* $\{f(x) : x \in \mathcal{X}\}$ *be a stochastic process, and* $x, x' \in \mathcal{X}$*, then* $m(x) := \mathbb{E}(f(x))$ *is called the mean function and* $k(x, x') := \mathrm{Cov}(f(x), f(x'))$ *the kernel of the stochastic process.*

The implied mean function and kernel for the Gaussian process representing the prior over regression functions are $m(x) = \phi(x)\mu_b$ and $k(x, x') = \phi(x)^\top \Sigma_b \phi(x')$. Thus, the choice of the transformation function $\phi(x)$ and the prior for the coefficient vector b determine the mean function and the kernel. For example, using the identity as transformation function and the regularizing prior $b \sim \mathcal{N}(0, I\sigma_b^2)$, where I is the appropriately sized identity matrix, results in mean function $m(x) = 0$ and kernel $k(x, x') = x^\top \sigma_b^2 x'$. This is the Bayesian equivalent of ridge regression (Hastie, Tibshirani, & Friedman, 2009, Chapter 3.4.1).

The Gaussian process prior on the regression function $f(x)$ also implies a prior on the outcome variable $Y(x)$. Since the regression function $f(x)$ is related to the outcome variable Y by the measurement equation $Y(x) = f(x) + \epsilon$, the prior implied for the outcome variable $Y(x)$ is a Gaussian process with mean function $m(x) = \phi(x)^\top \mu_b$ and kernel $k_y(x, x') = \phi(x)^\top \Sigma_b \phi(x') + \delta(x - x')\sigma_\epsilon^2$, where $\delta(\cdot)$ is the Dirac delta function, that is, it is 0 everywhere, except at 0. We will abbreviate this as

$$Y(x) \sim \mathcal{GP}\left(\phi(x)^\top \mu_b, \phi(x)^\top \Sigma_b \phi(x) + \delta(x - x')\sigma_\epsilon^2\right). \tag{1}$$

Thus, the GPR model for the outcome variable $Y(x)$ can be fully described by a mean function and a kernel. We will refer to kernels including the measurement error as k_y and kernels without the measurement error as k in the remainder of the manuscript. Model specification in GPR thus consists of choosing a mean function and a kernel. For example, the mean function and the kernel representing Bayesian linear regression with a regularizing prior and Gaussian measurement error are $m(x) = 0, k_y(x, x') = x^\top \sigma_b^2 x' + \delta(x - x')\sigma_\epsilon^2$.

2.2. Inference

Inference in GPR is traditionally focused on optimal predictions. To obtain unbiased estimates of the predictive accuracy, data is split into a training and a test set (or repeatedly so as in cross-validation). The training set $D = \{(x_i, y_i) : i \in 1, \ldots, N_1\} = (X, y)$ is used to fit the model, and the goal is to obtain optimal predictions when using input vectors x^* that have not been in the training set. We denote such inputs vectors in a test set as the matrix $X^* = [x_1^*, \ldots, x_{N_2}^*]$. Bayesian statistical learning procedures typically first obtain parameter estimates in the form of the posterior distribution and then link the posterior distribution with the likelihood to obtain predictions in the form of the predictive distribution (e.g., Bishop, 2006, Section 3.3). In contrast, in GPR, the predictive distribution is directly obtained in a single step.

The predictive distribution is the distribution given the model and the training set for the

test set predictions, that is, $f(X^*)|D$. For notational convenience, we introduce the following:

$$M(X) = \begin{bmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_{N_1}) \end{bmatrix}, \quad K(X, X^*) = \begin{bmatrix} k(x_1, x_1^*) & k(x_1, x_2^*) & \dots & k(x_1 x_{N_2}^*) \\ k(x_2, x_1^*) & k(x_2, x_2^*) & & \vdots \\ \vdots & & \ddots & \\ k(x_{N_1}, x_1^*) & \dots & & k(x_{N_1}, x_{N_2}^*). \end{bmatrix}$$

This allows expressing the joint distribution of observations $Y(X)$ and predictions $f(X^*)$ as follows:

$$\begin{bmatrix} Y(X) \\ f(X^*) \end{bmatrix} \sim \mathcal{N}\left(M\left(\begin{bmatrix} X \\ X^* \end{bmatrix} \right) \begin{bmatrix} K(X,X) + I\sigma_\epsilon^2 & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right).$$

The predictive distribution is obtained by conditioning on the observations y. It has an analytical solution, which is:

$$f(X^*)|D \sim \mathcal{N}(\mathbb{E}(f(X^*)|D), \mathrm{Cov}(f(X^*)|D)), \text{ with} \tag{2}$$

$$\mathbb{E}(f(X^*)|D) = M(X^*) + K(X^*, X)[K(X, X) + \sigma_\epsilon^2 I]^{-1}(y - M(X)) \tag{3}$$

$$\mathrm{Cov}(f(X^*)|D) = K(X^*, X^*) - K(X^*, X)[K(X, X) + \sigma_\epsilon^2 I]^{-1} K(X, X^*). \tag{4}$$

2.3. Model Selection

Before obtaining predictions based on a model, the model is chosen from a set of candidate models. The model in GPR is represented by a prior over regression functions, specified by a mean function and a kernel. Thus, model selection in GPR is formally equivalent to choosing a prior.

In GPR, the prior is typically obtained using the empirical Bayes approach (Rasmussen & Williams, 2006, Chapter 5). This means the prior is chosen based on the training data D. For GPR, this is typically done in two ways. One approach is to choose the prior that optimizes the model evidence. This approach is well in line with GPR being a Bayesian method. The other approach is to choose the prior that optimizes the predictive accuracy as measured by cross-validation. This is well in line with GPR being a statistical learning method.

Both approaches start with a set of models as represented by a parameterized mean function $m(x; \theta)$ and a parameterized kernel $k_y(x, x'; \theta)$. The parameters θ are so-called hyper-parameters, because every parameter value corresponds to a model. In both approaches, the hyper-parameters $\hat{\theta}$ are chosen that optimize an objective function given the training data $D = \{(x_i, y_i) : i \in 1, \dots, N_1\} = (X, y)$ with respect to the hyper-parameters θ.

When using the model evidence approach, the objective function is the model evidence, that is, the likelihood of the training data given the model, which we denote as $p(y|X, \theta)$. While the model evidence can only by approximated for many models (Bishop, 2006; Kruschke, 2014), it can be computed analytically for GPR models. It corresponds to a simple evaluation of the Gaussian likelihood:

$$p(y|X, \theta) = \mathcal{N}\left(y; M(X; \theta), K_y(X, X; \theta)\right).$$

When using the cross-validation approach for choosing a model, the objective function is the k-fold cross-validated prediction performance estimate (Kohavi, 1995). For a cross-validation procedure, a loss function has to be chosen, which quantifies the loss of predicting a value $\hat{y}_i$ when the true value is y_i. For GPR models, which return a predictive distribution, the negative log predictive probability loss is a natural loss function (Rasmussen & Williams, 2006, p. 112) as it also takes the uncertainty of the predictions into account. The negative log predictive probability of a vector of true values $y^* = [y_1, \ldots, y_{N_2}]$ under the predictive distribution is

$$-\log\left(\mathcal{N}(y^*; \mathbb{E}[f(X^*)|D], \mathrm{Cov}[f(X^*)|D])\right),$$

with the predictive mean $E[f(X^*)|D]$ and the predictive covariance matrix $\mathrm{Cov}[f(X^*)|D])$ being as defined in Equations 3 and 4 respectively. Consequently, the higher the likelihood of the true values y^* under the model, the lower its negative log predictive probability.

Usually, model selection approaches in both statistical learning and inferential analyses are used to select between a finite set of candidate models. In contrast, in GPR they are used to select between a typically infinite number of candidate models as represented by hyper-parameters from an uncountable hyper-parameter space $\Theta \ni \theta$. To this end, iterative optimization algorithms based on the gradient of the objective function are employed. They find the best model by optimizing the chosen objective function. This model is then used to obtain predictions as described in the previous section.

3. Gaussian Process Panel Models

In this section we generalize GPR to GPPM by introducing a between-person model, before introducing frequentist and statistical learning inference procedures in Section 4 and Section 5. To facilitate the discussion in these sections, we offer a short reinterpretation of a GPR model in the following.

3.1. Reinterpreting a Set of Priors as a Statistical Model

Understanding how a GPR model can be extended by a between-person model as well as complementing GPR with frequentist inference procedures is facilitated by reinterpreting the set of priors represented by a parameterized mean function $m(x;\theta)$ and kernel $k_y(x, x';\theta)$ pair as a statistical model.

Each hyper-parameter value θ describes the distribution of the Gaussian process $Y(x)$. In GPR, this distribution is interpreted to represent a prior and thus one model. Consequently, the set of distributions implied by the hyperparameter space Θ is interpreted as a set of models. An alternative interpretation is that each of these distributions is a candidate distribution for the Gaussian process $Y(x)$. It follows that now the set of distributions implied by the parameter space Θ is one model, which in the previous interpretation was a set of models. In notation, we write

$$Y(x) \sim \{\mathcal{GP}\left(m(x;\theta), k_y(x, x';\theta)\right) : \theta \in \Theta\},$$

which reads: It is assumed that there exists one parameter value $\theta^* \in \Theta$ such that the true mean function and kernel for the Gaussian process $Y(x)$ are $m(x;\theta^*)$ and $k(x, x';\theta^*)$.

3.2. Introducing Between-person Models

Let us assume that in a longitudinal data set, $N \in \mathbb{N}$ time series are observed. Each time series $y_i \in \mathbb{R}^{J_i}$ originates from one person i and contains J_i observations. With $y_{ij} \in \mathbb{R}$, we describe the j-th observation of person i. In analogy to GPR, we assume that each observation is accompanied by a corresponding input vector $x_{ij} \in \mathcal{X} \subseteq \mathbb{R}^p$, which is also observed. As for GPR, $x, x' \in \mathcal{X}$ describe two arbitrary input vectors. In the simplest case, the input vector can just contain the time point of the observation (e.g., days elapsed since inception of the study), but in principle, any variable is allowed to be a member of the input vector.

For modeling, we assume that for each person their time series y_i is a realization[1] of a stochastic process. Note that this assumption is reasonably general and encompasses virtually all available probabilistic methods for analyzing longitudinal data, including advanced methods such as nonlinear state-space models (Chow & Zhang, 2013).

The reinterpretation of GPR models is in line with this formalism. Essentially, with GPR, a model for the stochastic process of a single person can be defined assuming the stochastic process is a Gaussian process. Then, each person's time series y_i is considered a realization of a Gaussian process $Y_i(x)$ with true distribution as follows

$$Y_i(x) \sim \mathcal{GP}(m_i^*(x'), k_i^*(x, x')).$$

The mean function $m_i^*(x)$ and the kernel $k_i^*(x, x')$ represent the true distribution of the person-specific Gaussian process. Thus, for each person, a model can be formalized using a parameterized mean function and kernel.

However, so far it is not possible to specify relationships between persons, that is, a between-person model. The most straightforward approach to specify a between-person model is to assume that each person's time series is a realization of the same Gaussian process and that the person-specific Gaussian processes are mutually independent. The statistical model implied by this approach is

$$Y_i(x) \sim \{\mathcal{GP}(m(x;\theta), k_y(x, x';\theta)) : \theta \in \Theta\} \tag{5}$$

for every person. We call such a model a Gaussian process panel model (GPPM). Equivalently to GPR models, GPPMs are specified by choosing the predictors, the mean function, and the kernel.

Although the mean function and the kernel are assumed to be identical for each person, this does not imply that there are no between-person differences. Many forms of between-person difference can be specified using this formalism. More specifically, the model displayed in Equation 5 does not necessarily assume that the *true* distribution of the person-specific Gaussian processes is the same for each person. We will discuss this in more detail in the following section.

[1]Technically, a realization of a stochastic process is of infinite size as the stochastic process contains infinitely many random variables. So, technically y_i is a realization of a finite subset of the random variables contained in the stochastic process.

3.3. Supported Between-person Models

Let θ_{ip} be the person-specific variant of parameter θ_p. If the person-specific parameter θ_{ip} is considered a realization of a between-person probability distribution $\mathbb{P}(\theta)$, then we speak of probabilistic between-person model. If, in contrast, the person-specific parameter θ_{ip} is determined by a function $f(\tau_i, \theta_\tau) = \theta_{ip}$ with a trait (we define traits to be stable within a person) vector t_i and a parameter vector θ_τ as input then we speak of a deterministic between-person model. Note that this formalism also covers conditional distributions $\mathbb{P}(\theta|\tau_i)$, as this can be achieved by combining deterministic and probabilistic models.

In GPPM, a Gaussian between-person distributions for linear mean parameters can be implemented by simply modifying the mean function and kernel. Let the mean function be of the form

$$m(x;\theta) = f(x;\theta_1)^\top \theta_2 + h(x;\theta_3), \tag{6}$$

where the parameter vector $\theta = [\theta_1, \theta_2, \theta_3]$ is partitioned into parameter vectors $\theta_1, \theta_2, \theta_3$. $f(\theta_1)$ is a vector-valued, and $h(\theta_3)$ a scalar-valued function. The parameters in the vector θ_2 are what refer to as linear mean parameters. A probabilistic between-person model is introduced by individualizing the linear parameters in the vector θ_2 and assuming that the corresponding individualized parameter has the between-person distribution $\theta_{i2} \sim \mathcal{N}(\mu_{\theta_2}, \Sigma_{\theta_2})$. As a result, the mean function itself becomes a Gaussian process with mean function and kernel

$$\begin{aligned}
\mathbb{E}[m(x;\theta)] &= \mathbb{E}[f(x;\theta_1)^\top \theta_{i2} + h(x;\theta_3)] = f(x;\theta_1)^\top \mu_{\theta_2} + h(x;\theta_3) \\
\mathrm{Cov}(m(x;\theta), m(x';\theta)) &= f(x;\theta_1)^\top \Sigma_{\theta_2} f(x';\theta_1)
\end{aligned}$$

Consequently, the resulting GPPM is

$$\begin{aligned}
\tilde{m}(x;\theta) &= f(x;\theta_1)\mu_{\theta_2} + h(x;\theta_3) \\
\tilde{k}(x,x';\theta) &= k(x,x';\theta) + f(x;\theta_1)^\top \Sigma_{\theta_2} f(x';\theta_1)
\end{aligned} \tag{7}$$

To investigate whether other types of between-person models besides a Gaussian between-person model on mean parameters can be defined, we use the mathematical equivalence of between-person models and priors. Introducing a between-person distribution for a given parameter is equivalent to introducing a prior distribution over that parameter, even though either approach may have quite different implications in practice. With this equivalence in mind, we may then regard the original mean function and kernel as the likelihood. To express the new statistical model, the marginal likelihood has to be obtained, that is, a weighted average of the likelihood using the prior as weighing function. We showed that for a Gaussian likelihood and a Gaussian prior on linear parameters of the mean the marginal likelihood is again Gaussian (also see Bishop, 2006, Chapter 2.3.3). Between-person distributions other than the Gaussian will typically not lead to the marginal likelihood being Gaussian. The Gaussian prior is the only commonly used prior that has this property. The same is true for between-person distributions on the kernel parameters, as for most commonly used priors on variance parameters the resulting marginal likelihood is not a Gaussian. To implement non-Gaussian between-person distributions and

between-person distributions of variance parameters in the GPPM framework, an extension of the basic GPPM formalism is needed, which remains subject to future research.

Deterministic between-person models, that is, parameter heterogeneity that is governed by a deterministic function $f(\tau_i, \theta_\tau) = \theta_{ip}$, can be implemented by changing the mean function, the kernel, and the predictors in the input vector x. For this, we use a method that is able to implement parameter heterogeneity for every single observation: Let θ_p be a parameter of the mean function for which a person- and potentially observation-specific variant θ_{ijp}, determined by a function $f(\tau_{ij}, \theta_\tau) = \theta_{ijp}$, is desired. We now assume that the vector τ also contains states, which in contrast to traits are assumed to vary within person. To implement this, one simply needs to replace the parameter θ_p by $f(t_{ij}, \theta_\tau)$ in the mean function, add the trait and state variables in the vector τ_{ij} to the input vector x, and add the parameter vector θ_τ to the parameter vector θ.

The same concept can be used to implement person- and observation specific heterogeneity for parameters in the kernel. When implementing observation-specific heterogeneity, it is important to note that a kernel describes the model for the pairwise covariance between all observations. Thus, observation-specific heterogeneity needs to take the state values of two observations into account. This however, can be easily accommodated for by letting the function governing the parameter heterogeneity depend on a pair of trait-state vectors τ_{ij}, τ_{ik}, that is, it changes to $f(\tau_{ij}, \tau_{ik}, \theta_\tau)$. Besides that small change, everything works analogously for mean functions.

3.4. Model Specification & Relation to Other Model Families

Given a panel data set, a GPPM for an outcome variable Y is defined by identifying predictor variables in the input vector, which at least includes time of measurement, and by specifying a mean function and a kernel. Model specification is facilitated by the fact that models can be specified by simple combination rules. For example, a valid mean function can be created by summation of, multiplication of, and scaling of base mean functions. The same operators can be used to create a new kernel based on combinations of base mean functions and kernels (see Duvenaud, Lloyd, Grosse, Tenenbaum, & Ghahramani, 2013, for details). Some of these operators have straightforward interpretations. For example, the sum of two Gaussian processes with mean functions m_1, m_2 and kernels k_1, k_2 has mean function $m = m_1 + m_2$ and kernel $k = k_1 + k_2$.

With GPPM being a hybrid of a statistical learning and an explanatory method, a model can be specified using different strategies.

One approach to model specification is to translate a substantive theory into its GPPM representation. For example, a very simple theory could posit that there is no change over time and no between-person differences; that is, all observed differences across persons and measurement occasions are solely attributed to measurement error. This would translate into the follwing GPPM

$$m(x; \theta) = \mu_I, \quad k_y(x, x'; \theta) = \delta(x - x')\sigma_\epsilon^2,$$

with $\mu_I \in \mathbb{R}$.

Another approach for specifying a model is to translate a traditional model into a GPPM. For example, we now translate the linear latent growth curve model (Preacher, Wichman, Briggs, & MacCallum, 2008), which assumes that each person's trajectory over time follows a linear trend while allowing individual differences in intercept and linear slope. The only predictor needed is

time. We, thus, denote the input vector with t instead of x. The GPPM representation of the univariate linear latent growth curve model is

$$m(t;\theta) = \underbrace{\mu_I}_{\text{constant}} + \underbrace{\mu_S t}_{\text{linear}}, \quad k_y(t,t';\theta) = \underbrace{\sigma_I^2}_{\text{constant}} + \underbrace{t\sigma_S^2 t'}_{\text{linear}} + \underbrace{\sigma_{IS}(t+t')}_{\text{covariance}} + \underbrace{\delta(t-t')\sigma_\epsilon^2}_{\text{noise}}, \tag{8}$$

with μ_I being the mean of the intercept, μ_S the mean of the slope, σ_I^2 the variance of the intercept, σ_S^2 the variance of the slope, σ_{IS} the covariance between intercept and slope, and σ_ϵ^2 the variance of the measurement error, which is assumed to be constant across time. The latent growth curve model illustrates the combination rules that are foundational to the construction of GPPM. We annotated Equation 8 with the commonly used names of the base mean functions and kernels that form the LGCM in our representation. An important advantage of translating a traditional model into a GPPM is that it can be used with a statistical learning inference approach, as discussed in more detail further below. Another advantage is that the GPPM representation is inherently a continuous-time model that allows to consider person-specific time points of measurement, irregular intervals between measurements, and naturally allows us to interpolate and extrapolate unobserved time points.

Another approach for model specification is to adapt a model typically used for GPR. Many GPR practitioners rely on a default model, which is flexible enough to approximate most (smooth) functions, known as the universal approximating property (Micchelli, Xu, & Zhang, 2006). This model is known as the squared exponential model and is defined as follows:

$$m(x;\theta) = 0, \quad k(x,x';\theta) = \sigma_{se}^2 \exp\left(-\frac{||x-x'||^2}{l}\right).$$

The parameter σ_{se}^2 governs the variance of the process and the strictly positive parameter l governs how fast the correlation drops between two variables $Y(x)$, $Y(x')$ as a function of the squared euclidean distance of their input vectors x and x'. We will adapt and explore the utility of the squared exponential model for longitudinal data in the illustrations section.

Another valuable property for model specification is that the family of GPPMs subsumes many other model families such as longitudinal (linear Gaussian) structural equation models, and (linear Gaussian continuous-time) state-space models (REMOVED TO MAINTAIN REVIEW INTEGRITY). Essentially, a linear Gaussian continuous-time state space model describes a model for a Gaussian process via stochastic differential equations, whereas in GPPM a model for a Gaussian process is described using the mean function and the kernel. By definition, any distribution of a Gaussian process can be specified via a mean function and a kernel, whereas only a subset of distributions can be represented by stochastic differential equations. For example, the squared exponential model cannot be represented as a state-space model (Särkkä & Hartikainen, 2012). Similarly, in structural equation modeling, a model for a Gaussian random vector is specified by restricting its mean vector and covariances matrix whereas in GPPM a model is described on a Gaussian process (the generalization of a Gaussian random vector) by restricting its mean function (the generalization of a mean vector) and kernel (the generalization of a covariance matrix).

The rules for combining models along with the fact that GPPM can represent a wide range of different models can also be used to mix models from different traditions. In the illustration section, we will explore this idea by combining the squared exponential model typically used in statistical learning with a growth curve model.

4. Frequentist Inference for Gaussian Process Panel Models

4.1. Implied Statistical Model

Frequentist inference theory requires a statistical model, which is a set of candidate distributions for a random vector. A GPPM, as defined in Equation 5, is a set of candidate distributions for a stochastic process and thus not a proper statistical model. However, while a stochastic process is of infinite size, the observations drawn from it, in our case a panel data, set are necessarily finite. Thus, the data set can be seen as a realization of a finite dimensional subset of the stochastic process, which is a random vector.

The statistical model implied by a GPPM is as follows. Let $X_i \in \mathbb{R}^{p \times J_i}$ be a matrix where each column $x_{ij} \in \mathcal{X} \subseteq \mathbb{R}^p$ contains the input vector for the j-th observation of person i, that is, for the observation $y_{ij} \in \mathbb{R}$. The statistical model for all observations $y_i = [y_1, \ldots, y_{J_i}]$ of person i implied by a GPPM with mean function m and kernel k_y is

$$p(y_i|X_i) \in \{\mathcal{N}(y_i; M(X_i;\theta), K_y(X_i, X_i;\theta)) : \theta \in \Theta\}.$$

The statistical model implied for a longitudinal data set $D = (X, y)$, with $X = (X_1, \ldots, X_N)$ and $y = (y_1, \ldots, y_N)$, follows from the mutual independence assumption and is

$$p(y|X) \in \left\{ \prod_{i=1}^{N} \mathcal{N}(y_i; M(X_i;\theta), K_y(X_i, X_i;\theta)) : \theta \in \Theta \right\}.$$

This is a regular statistical model for which regular inference procedures can be derived as we will show in the following.

4.2. Point Estimation

Here, we show how to obtain maximum likelihood estimates for a GPPM and investigate their frequentist properties. To this end, the parameters $\hat{\theta}$ need to be found that maximize the likelihood of the data, that is,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} p_\theta(y|X)$$

with likelihood function

$$p_\theta(y|X) = \prod_{i=1}^{N} \mathcal{N}(y_i; M(X_i;\theta), K_y(X_i, X_i;\theta)).$$

Equivalently, the log likelihood

$$\log(p_\theta(y|X)) = \sum_{i=1}^{N} \log\left(\mathcal{N}(y_i; M(X_i;\theta), K_y(X_i, X_i;\theta))\right)$$

can be maximized.

The maximum likelihood estimates for a GPPM can typically not be derived analytically. For this reason, we employ gradient descent algorithms as they are commonly used in, for example, structural equation modeling. The required gradient of the log likelihood function $\log(p(y|X,\theta))$ can be calculated analytically:

$$\frac{\partial \log(p_\theta(y|D))}{\partial \theta_p} = \sum_{i=1}^{N} \frac{1}{2} \tilde{y_i}^\top \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \theta_p} \Sigma_i^{-1} \tilde{y_i} - \frac{1}{2} \operatorname{tr}\left(\Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \theta_p}\right) + \frac{\partial \mu_i}{\partial \theta_p} \Sigma_i^{-1} \tilde{y_i}$$

where $\mu_i(\theta) = M(X_i;\theta)$ is the model-implied mean vector for person i, $\Sigma_i(\theta) = K_y(X_i, X_i;\theta)$ the model implied covariance matrix, and $\tilde{y_i}(\theta) = y_i - M(X_i;\theta)$ is the derivation of the observations from the model implied mean. For notational simplicity, we have decided to not explicity state the dependence on θ for these terms.

Under certain regularity conditions (Taboga, 2012b) maximum likelihood estimates are consistent, efficient, and have asymptotically a Gaussian sampling distribution with the Fisher information matrix as covariance matrix (Taboga, 2012b). A comprehensive discussion of all regularity conditions is beyond the scope of this text. However, it is important to note that some conditions, such as the integratability of the log-likelihood function, are always met if a Gaussian statistical model is assumed. Others depend on the specific choice of mean and kernel function, and may be violated. For example, the model $m(x) = 0; k_y(x, x'; [\sigma_1, \sigma_2]) = \sigma_1^2 + \sigma_2^2$ is not identified and thus its maximum likelihood estimate does not have the favorable properties. However, since the possible violations are largely shared among modeling approaches, we refer the reader to the discussion in the context of structural equation modeling (e.g., Stoel, Garre, Dolan, & van den Wittenboer, 2006). We thus expect the maximum likelihood estimates for GPPMs to be consistent, efficient, and to have an asymptotically Gaussian sampling distribution for all well-behaved models.

4.3. Hypothesis Tests, Confidence Intervals, Model Selection & Validation

For hypothesis tests, the likelihood ratio test with an asymptotic Chi-squared sampling distribution of the test statistic can be used (Taboga, 2012a). This follows directly from the maximum likelihood estimators being asymptotically normal.

As for structural equation modeling (Pek & Wu, 2015), two main approaches for computing confidence intervals can be used: Wald-type and likelihood-based methods. Essentially, Wald-type confidence intervals invert the Wald-test whereas likelihood-based methods invert the likelihood-ratio test. Thus, the validity of these methods relies on the validity of their corresponding tests, which in turn follows from the maximum likelihood estimator to have an approximate Gaussian sampling distribution.

For model selection, the prototypical frequentist approach to test between two competing nested models using a hypothesis like the likelihood-ratio test can be used. Alternatively, when two models are not nested, many different approaches for selecting between two models exist. As a start, we adapt two popular, general, and simple approaches; namely, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Both measures only depend on the log-likelihood of a model at the maximum likelihood estimate $\hat{\theta}$ and the number of parameters and thus can also be used to select between two GPPMs.

For model validation, the prototypical frequentist approach is to test the model itself using a hypothesis test. In structural equation modeling, for example, this is done by testing the model against the saturated model, that is, the model "any Gaussian distribution." This essentially results in a comparison of the covariance matrix estimate under the model with the sample covariance matrix. When using a GPPM, the equivalent would be to test the obtained kernel under a GPPM, with a kernel estimate under the model "any kernel function." Intuitively, estimating this kernel seems impossible without making additional assumptions, since with every new data point, also a new parameter is introduced. Whether and how this kernel can be estimated remains to be investigated.

5. Statistical Learning Inference for Gaussian Process Panel Models

5.1. Prerequisites

To develop statistical learning inference methods for GPPMs, we interpret a GPPM as a prior over all potential observations. That is, we interpret a GPPM as represented by a mean function and a kernel to describe a prior for the set of potentially infinitely many Gaussian processs (GPs) $\{Y_i(x) : i \in I, x \in \mathcal{X}\}$ with I containing the indices for all, potentially infinitely many, persons and $\mathcal{X}$ denoting the set of possible input vectors. Since in a GPPM independence and identical distribution is assumed between the person-specific GPs, it is sufficient to specify one shared mean function and kernel. The interpretation of a GPPM as representing a set of priors is in line with the interpretations used in GPR. Consequently, the statistical learning procedures used for GPR can be applied to GPPMs with only slight adaptations.

Statistical learning is primarily concerned with estimating generalization error to make decisions about which model is the best-fitting model. To obtain unbiased estimates of this generalization error, one way is to use independent training and test sets. The training set is used for model fitting and the test set for model evaluation or selection. The standard approach to obtain a training and a test set is to split the data set. A common approach is random splitting of $(x_i, y_i) \in D$ assuming independence of all observed cases. In the context of digit classification, for example, x_i is a matrix of brightness values for an image and y_i the corresponding digit for that image. Neither the brightness values x_i nor the digit label y_i of one particular image typically contains information about any other image.

For longitudinal data, the situation is more complex. First, the data set is inherently nested. There are different persons, and each person has been observed at multiple measurement occasions. Mathematically, we denote the j-th observation of person i as (x_{ij}, y_{ij}). Thus, when splitting a longitudinal data set, a first crucial question is whether to split based on persons, measurement occasions, or both. As we will show in the next section, GPPM can obtain

predictions for all these scenarios. When splitting based on persons, all observations (x_{ij}, y_{ij}) for which the index (e.g., representing time in study) i is greater than some threshold could be put in the test set and all others in the training set. Importantly, the data of each person is either entirely in the training or in the test set. In a similar fashion, when splitting a longitudinal data set based on measurement occasions, one could distribute all observations earlier than a given time point to the training set and all remaining measurements to the test set.

How to split the data is guided by which generalization performance of the model is of core interest. When the ability of the model to predict observations for persons who are not in the training set is of interest, the appropriate split is by persons. In contrast, if the ability of the model to predict observations in the data set for future measurement occasions that are not in the training set is of interest, the appropriate split is by measurement occasions.

When splitting by measurement occasions, special care has to be taken, because the common assumption that observations in the training and test set are independent, can be easily violated. This problem is extensively discussed in the time series literature, and we refer the interested reader to Bergmeir and Benítez (2012).

5.2. Prediction

The procedure of how to obtain predictions for data not in the training set follows closely the idea underlying GPR. The joint distribution of the training data and the test data is implied by the model and then conditioned on the training observations y. This process can be simplified by the following observation: As a result of the independence between persons assumption, the predictions for a particular person are only influenced by observations from the same person, and the predictive distributions for different persons are independent of each other. Thus, the predictive distribution can be calculated independently for each person i in the test set.

Two scenarios to obtain a predictive-distribution for a person i must be distinguished. First, if there are no observations from person i in the training data. In this case, the predictive distribution is independent of the training data. The predictive distribution for predictions of interest $Y_i(X_i^*) = [Y_i(x_{i1}^*), \ldots, Y_i(x_{iJ_i^*}^*)]$ is simply:

$$Y_i(X_i^*)|D \sim \mathcal{N}(M(X_i^*; \hat{\theta}), K_y(X_i^*, X_i^*; \hat{\theta})).$$

Second, if there are observations from person i in the training data, the joint distribution of observations $Y_i(X_i) = [Y_i(x_{i1}), \ldots, Y_i(x_{iJ_i})]$ and predictions of interest $Y_i(X_i^*)$ has to be formulated:

$$\begin{bmatrix} Y_i(X_i) \\ Y_i(X_i^*) \end{bmatrix} \sim \mathcal{N}\left(M\left(\begin{bmatrix} X_i \\ X_i^* \end{bmatrix}; \hat{\theta} \right), \begin{bmatrix} K(X_i, X_i; \hat{\theta}) & K(X_i, X_i^*; \hat{\theta}) \\ K(X_i^*, X_i; \hat{\theta}) & K(X_i^*, X_i^*; \hat{\theta}) \end{bmatrix} \right) \tag{9}$$

In complete equivalence to the predictive distribution in GPR, the predictive distribution for $Y_i(X_i^*)$ is obtained by conditioning on the training data D. As discussed before, only the observations y_i from person i are needed because $Y_i(X_i^*)|D = Y_i(X_i^*)|X_i, y_i$:

$$Y_i(X_i^*)|X_i, y_i \sim \mathcal{N}(\mathbb{E}(Y_i(X_i^*)|X_i, X_i^*, y_i), \mathrm{Cov}(Y_i(X_i^*)|X_i, X_i^*, y_i)), \text{ with} \tag{10}$$

$$\mathbb{E}(Y_i(X_i^*)|X_i, y_i) = M(X_i^*) + K(X_i^*, X_i)[K(X_i, X_i)]^{-1}(y_i - M(X_i)) \tag{11}$$

$$\mathrm{Cov}(Y_i(X_i^*)|X_i, y_i) = K(X_i^*, X_i^*) - K(X_i^*, X_i)[K(X_i, X_i)]^{-1}K(X_i, X_i^*). \tag{12}$$

where we dropped the dependence on $\hat{\theta}$ for notational convenience.

The predictive distribution can be reduced to both an interval or a point estimate.

For point estimation, in principle, any Bayesian technique to reduce a posterior distribution to a parameter estimate can be used. However, since the predictive distribution of a GPPM is Gaussian, the two most common techniques, using the mode (maximum a posterior estimation) or expectation (minimum mean square error estimation) of the posterior, both correspond to the mean of the predictive distribution. That is, the recommended point estimate for the prediction implied by a input vector x_i^* is $\mathbb{E}(Y_i(x_i^*)|X_i, y_i)$, which a special case of Equation 11.

To obtain an interval estimate for the predictions, credible intervals can be constructed from the Gaussian predictive distribution. Since the predictive distribution is Gaussian, a credible interval can be obtained by using $\mathbb{E}(Y_i(x_i^*)|X_i, y_i) \pm c_\alpha \sqrt{\mathrm{Var}(Y_i(x_i^*)|X_i, y_i)}$. Because the variance of a variable is equivalent to the covariance with itself, $\mathrm{Var}(Y_i(x_i^*)|X_i, y_i)$ is a special case of Equation 12. The critical value c_α has to be chosen based on the cumulative density function of the Gaussian distribution to obtain the desired credibility $1 - \alpha$.

Predictions can also be obtained for latent variables using the same framework. All that is needed is a model for the joint distribution of latent variables and observations. In the illustration section, we will demonstrate this idea.

5.3. Model Selection and Validation

The statistical learning approaches used in GPR for model selection and validation can be readily adapted to GPPM. Remember that for the statistical learning perspective on GPPM each hyper-parameter vector value θ of the mean function and kernel represents one prior and consequently one model.

For the model evidence maximization approach to select a model and thus a hyperparameter vector value θ, the hyperparameter vector value θ that maximizes

$$p(y|X, \theta) = \prod_{i=1}^{N} \mathcal{N}(y_i; M(X_i; \theta), K_y(X_i, X_i; \theta)).$$

is selected. Note that this expression is identical to the likelihood function used for maximum likelihood estimation. Thus, the best model $\hat{\theta}$ from the statistical learning perspective and the maximum likelihood parameter $\hat{\theta}$ from the explanatory perspective are the same, only their interpretation differs. Also, the gradient-descent algorithm developed for maximum likelihood estimation can be re-used for model selection.

Since GPPM comes with mechanisms to obtain predictions, model selection procedures that estimate the predictive performance most notably cross-validation can also be employed. Because

cross-validation is essentially a repeated splitting in training and test sets, the same complications discussed earlier, apply. Another issue of using cross-validation for model selection is that cross-validation estimates tend to have a high variance when using small data sets (Piironen & Vehtari, 2017). This issue can be partly resolved by repeated cross-validation, which decreases the variance but increases the computational demands. Another approach is to use the model evidence maximization approach instead, which we will consequently focus on in this paper.

To validate a selected GPPM also its predictive performance is estimated. However, the data set for performance estimation must be independent of the data set used for model selection to avoid overly optimistic estimates. If cross-validation is used for model selection and performance estimation this leads to a process called nested cross-validation, which is described in detail in Karch, Sander, von Oertzen, Brandmaier, and Werkle-Bergner (2015).

6. Illustrations

6.1. Simulated Data: Utility of the Statistical Learning Perspective

Based on simulated data, we will first demonstrate how the statistical learning inference methods for GPPMs enable valid estimation of the predictive accuracy from standard restrictive longitudinal models such as the latent growth curve model (LGCM) even if the assumptions are violated. Second, we will demonstrate the utility of the more flexible models representable in GPPM. Third, and most importantly, we will showcase the ability of GPPM to express hybrid models that consist of a combination of standard restrictive models as well as flexible statistical learning models and the utility of these combinations. GPPM is uniquely equipped to express such models as it can represent the majority of restrictive longitudinal models typically used in psychology as well as a large class of statistical learning models and contains a set of easy rules to combine models.

To begin with, we start with the linear LGCM, which is one of the most frequently used model for analyzing longitudinal panel data in psychological research. It is a prototypical example of a restrictive model as it assumes that within-person change is linear. Additionally, inference is typically performed using classical frequentist inference methods, which crucially assume the correctness of the model. It is well known that this can lead to dramatically wrong conclusions if the assumptions are not met (e.g., Ghisletta et al., 2019).

To demonstrate this, we assume that an "exponential rise to the limit" is the true data generating model. This model can be interpreted to represent the typical skill development observed in training studies. The within-person model is

$$Y_i(t) = b_i + d_i \exp(-ts) + \epsilon_i(t), \text{ with } \epsilon_i(t) \sim \mathcal{GP}(0, \sigma_\epsilon^2)$$

At time $t = 0$, the model implies $\mathbb{E}(Y_i(0)) = b_i + d_i$. Thus, the parameter combination $b_i + d_i$ serves as intercept in this model. For, $t \to \infty$, the model implies $\lim_{t\to\infty} \mathbb{E}(Y_i(t)) = b_i$. Consequently, the model implies that each persons skill level saturates at some point. The strictly positive parameter s_i represents how fast person i reaches their natural limit. For the between-person model, we assume $b_i \sim \mathcal{N}(\mu_b, \sigma_b^2), d_i \sim \mathcal{N}(\mu_d, \sigma_d^2)$, $s_i \sim \mathcal{N}(\mu_s, \sigma_s^2)$, and b_i, d_i, s_i being mutually independent. For s_i, we use a truncated normal to avoid negative values.

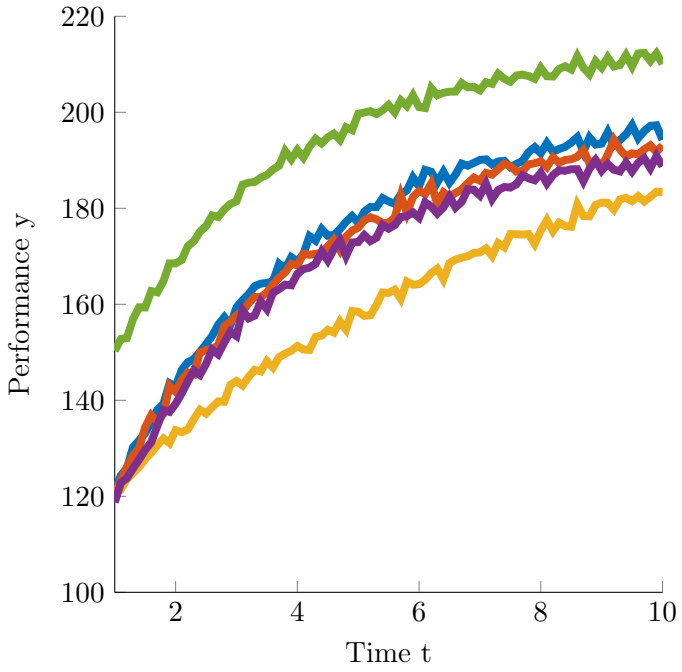


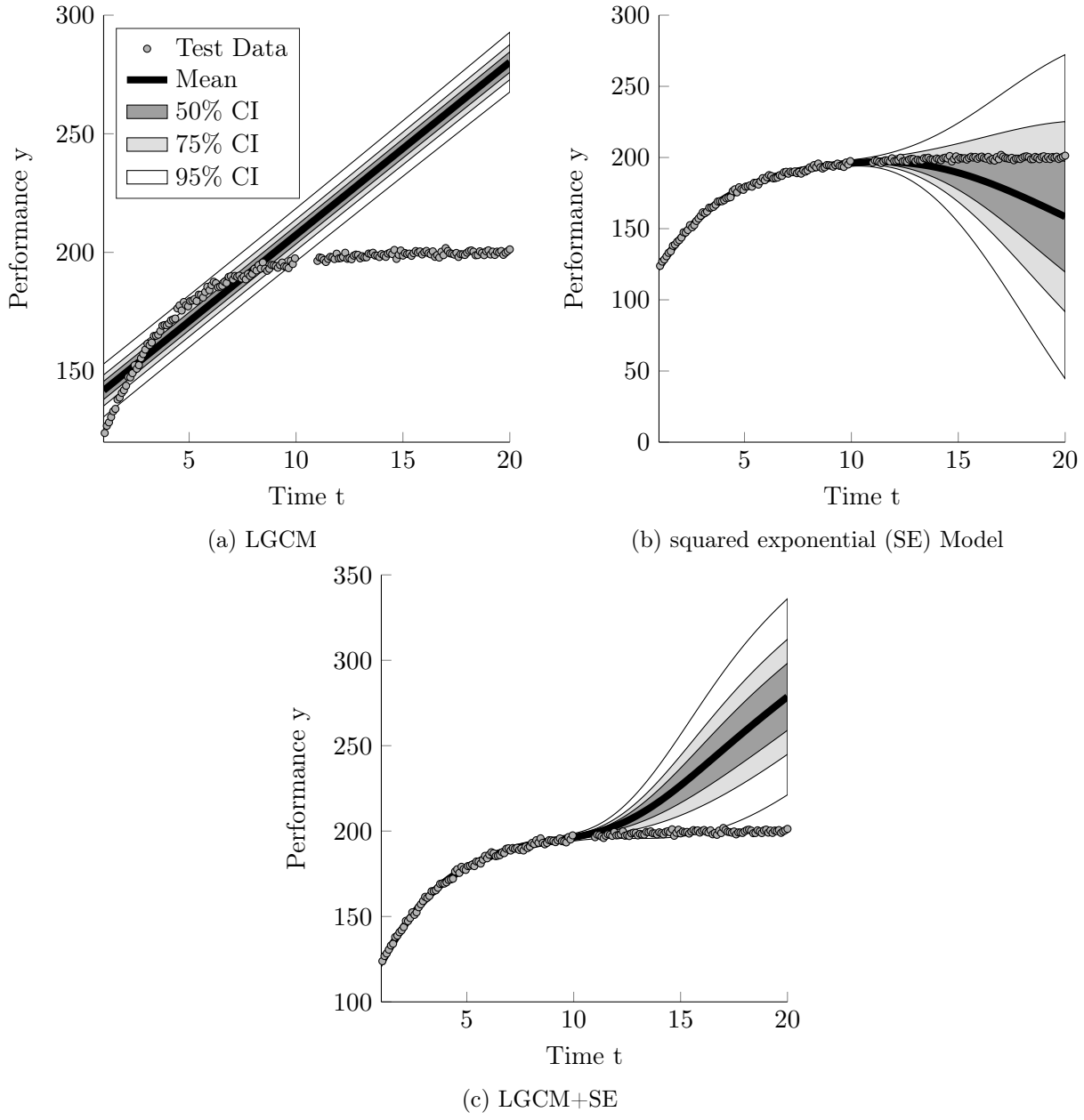
Figure 1: Five model-implied trajectories simulated from the "exponential rise to limit model". Each line represents the skill trajectory of one person.

For generating data from this model, we used the following parameter values $\mu_b = 200$, $\sigma_b^2 = 50$, $\mu_d = -100$, $\sigma_d^2 = 50$, $\mu_s = .25$, $\sigma_s^2 = 0.01, \sigma_\epsilon^2 = 1$. We generated data for 200 persons with 91 measurements each; all taken at the same time points $\{1, 1.1, 1.2, \ldots, 10\}$. Example trajectories of this model are displayed in Figure 1.

The linear LGCM or any of the typical extensions to polynomials of a higher order, such as quadratic or cubic, do not contain the data generating model and are thus misspecified.

In contrast, representing the LGCM as a GPPM allows performing valid inference using the LGCM on data simulated from the "exponential rise to the limit" model. The statistical learning framework applied to the LGCM first results in parameter estimates, which are equivalent to the maximum likelihood estimates. However, importantly, they are virtually ignored and only used to obtain the inferential object of interest, the predictive distribution. Using the maximum likelihood parameters $\hat{\theta}$, the predictive distribution for each person can be obtained according to Equations 10-12. We display the predictive distribution for one selected person in Figure 2a.

Whereas the predictive distributions crucially depend on the model, their performance is evaluated in a model-free fashion, which leads to their evaluation being independent of the validity of the model assumptions. For choosing the appropriate test set, it is crucial to distinguish which kind of predictive performance we want to assess. Here, we focus on assessing how well the model can make predictions for unobserved time points for persons within the sample. A second decision has to be made regarding the assessment of inter- or extrapolation capabilities. For assessing how well the model interpolates, we created a test set that contains the measurements for all persons in



(a) LGCM

(b) squared exponential (SE) Model

(c) LGCM+SE

Figure 2: Visualization of the predictive distributions on the "exponential rise to the limit" data for the three considered models

the original sample at time points $\{1.05, 1.15, \ldots, 9.95\}$. For assessing how well the model extrapolates, we used time points $\{11, 11.1, \ldots, 20\}$.

As loss function, we use the negative log predictive probability (NLPP). We report the average NLPP across persons. To make the average NLPP loss more interpretable we normalized it using the best possible model, the Bayes-optimal model, as reference. Because we have generated data with an error variance of 1 (also referred to as "irreducible error"), the Bayes-optimal model has an expected NLPP on the test set of the size of the interpolation set of $90 \cdot \log(\sqrt{2\pi e}) = 127.70$, where 90 the number of measurements per person and $\log(\sqrt{2\pi e})$ the entropy of the standard normal distribution. We subtracted this number from the estimated average NLPPs to obtain normed NLPPs.

Not surprisingly, the results reveal that the LGCM interpolates and extrapolates rather poorly. The interpolation normed NLPP was 152.1. The extrapolation NLPP was even higher at 317.77. The reason for this difference can be understood by looking at Figure 2a. In the interpolation range, the LGCM still provides a decent approximation of the nonlinear trend. However, in the extrapolation range, the LGCM confidently makes wrong predictions, which is caused by the predictions being based on incorrect, strict assumptions.

Model	Interpolation	Extrapolation	Combined
LGCM	152.1	317.77	600.2
SE Model	2.33	11.89	13.862
LGCM+SE Model	2.21	16.86	18.853

Table 1: Negative log predictive probabilities on the “exponential rise to the limit data” for the compared models as estimated by the different test sets.

The flexible statistical learning models representable in GPPM address this issue of the LGCM since they have specifically been designed to be able to interpolate a large set of functions well. Consequently, given enough data, they will reach an almost perfect interpolation performance for a large set of developmental trajectories. Thus, while those models are misspecified, given enough data, they predict essentially equally well as the true model. One such model is represented by the squared exponential (SE) kernel, which we have introduced earlier and repeat here:

$$m(t;\theta) = 0, \quad k_{se}(t,' t;\theta) = \sigma_{se}^2 \exp\left(-\frac{t-t'}{l}\right).$$

The SE model represents the family of smooth predictive functions. Importantly, in regions where no data has been observed the SE model falls back to predicting 0. Thus, it can be interpreted as regularizing towards zero mean predictions.

Before applying the SE model to longitudinal panel data it needs to be adapted slightly. Instead of regularizing towards zero, we regularize towards the person-specific mean. This is easily achieved using the established combination rules for GPPMs. One simply adds the GPPM representation of the random intercept model to the SE model. This results in the following

random intercept SE model

$$Y_i(t) \sim \mathcal{GP}\left(\mu_I, \sigma_I^2 + k_{se}(t,t') + \delta(t-t')\sigma_\epsilon^2\right).$$

The random intercept SE model, which we will abbreviate to SE model in the remainder, achieves, as expected, a substantively better interpolation performance (NLPP= 2.33). This almost perfect interpolation performance is also apparent in the visualization of the predictive distribution for one person in Figure 2b.

The SE model also extrapolates better than the LGCM (NLPP= 11.89). This seems to be caused by the SE model increasing the variance of the predictive distribution for data points far away from the training data whereas the variance of predictive distribution from the LGCM remains almost constant (compare Figures 2a and 2b). As a consequence, the LGCM makes wrong predictions with high confidence for the data points far away from the training data. However, in contrast to the extrapolation performance of the SE model, the interpolation performance can still be considerably improved, as is visible in Figure 2b. Essentially, the SE model falls back to a constant predictive distribution centered around the person-specific mean with a large variance.

This observation motivates the development of a class of hybrid models that consist of a combination of a parametric model, such as the LGCM, and a flexible nonparametric statistical learning models, such as the SE. Such models can also be motivated using more theoretical arguments. Within-person trajectories are often conceptualized as consisting of a combination of intraindividual change and intraindividual variability (Nesselroade, 1991; Ram & Grimm, 2015). Intra-individual change is believed to reflect the true change and is characterized by a relatively slow, well-behaved trajectories; whereas intraindividual variability is believed to occur at a much smaller time scale and is believed to reflect more chaotic, short-lived fluctuations around the intraindividual change. The hybrid of a parametric model and a flexible statistical learning model seems perfectly suited for this situation. The parametric part captures the long-term intraindividual changes whereas the flexible nonparametric part captures the intraindividual variability. The random intercept SE is also a hybrid model as it combines the parametric random intercept with the nonparametric SE model.

We demonstrate the utility of such models using the LGCM+SE model as an example. Importantly, GPPM would also allow the parametric model to be a more complex model such as the "exponential rise to the limit model". Mixing those two models, leads to the following model

$$Y_i(t) \sim \mathcal{GP}\left(\mu_I + \mu_S t, \sigma_I^2 + t\sigma_S^2 t' + \sigma_{IS}(t+t') + \delta(t-t')\sigma_\epsilon^2 k_{se}(t,t') + \delta(t-t')\sigma_\epsilon^2\right).$$

Effectively, this model regularizes the SE model using the LGCM. Thus, it falls back to a LGCM in regions with few data and to a SE model in regions with much data. As a result, it essentially behaves like the flexible SE in regions with many training samples and is thus able to fit a large class of functions, in these regions. In regions with no training samples it behaves like the LGCM (with larger predictive variance to reflect for the presence of intraindividual variability), and thus might be better at extrapolation.

The example data are generated as a combination of a LGCM and an unknown deviation

from the LGCM.

$$Y_i(t) = \underbrace{I_i + S_i t + \epsilon_i(t)}_{\text{LGCM}} + \underbrace{f(t;\theta_{i2})}_{\text{deviation}}$$

$$\begin{bmatrix} I_i \\ S_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_I \\ \mu_S \end{bmatrix}, \begin{bmatrix} \sigma_I^2 & \sigma_{IS} \\ \sigma_{IS} & \sigma_S^2 \end{bmatrix} \right), \sigma_{i2} \sim \mathcal{N}(\mu_{\theta_2}, \Sigma_{\theta_2}), \epsilon_i(t) \sim \mathcal{GP}(0, \delta(t-t')\sigma_\epsilon^2)$$

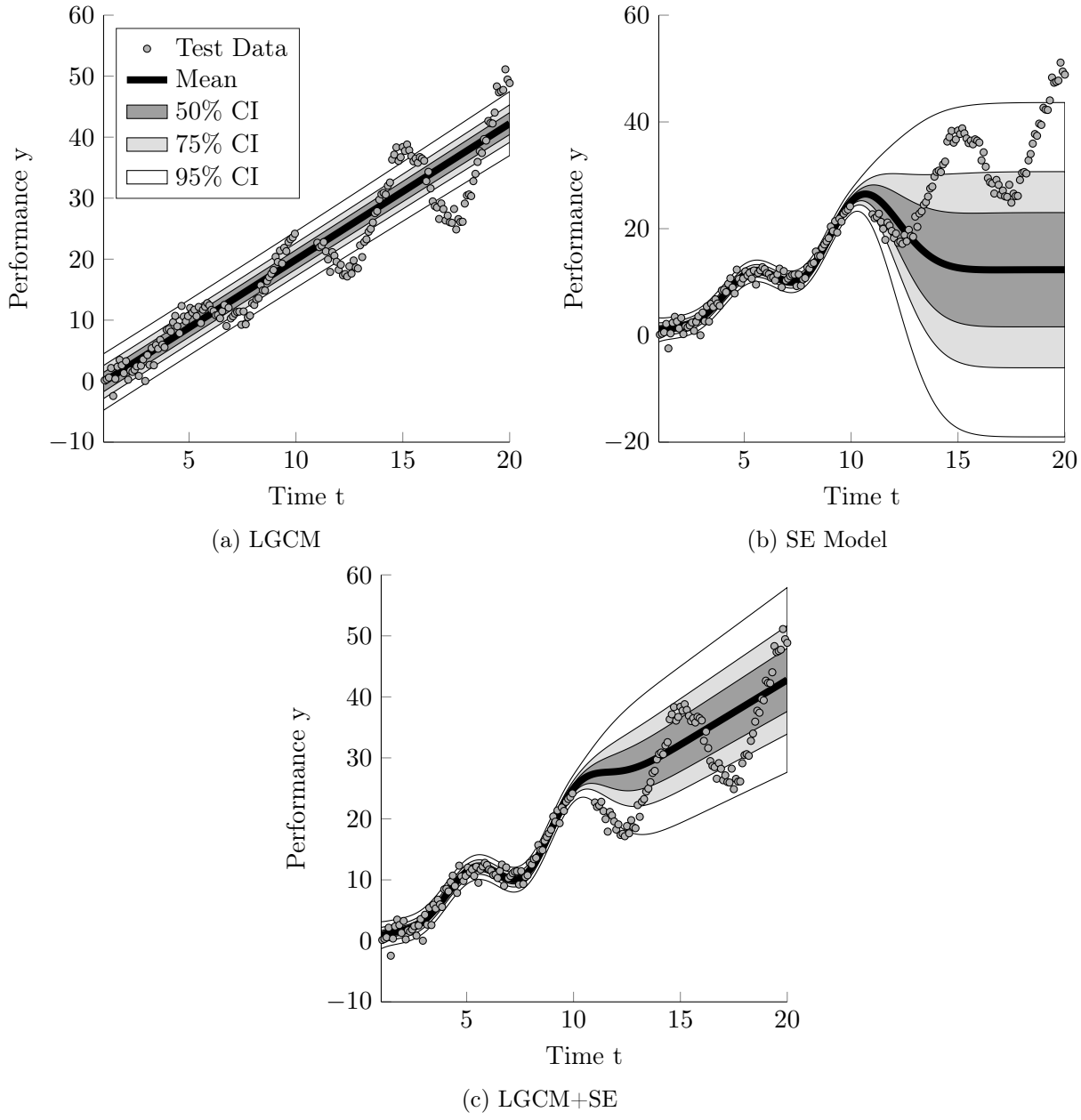
This can be interpreted as the intraindividual long-term change being appropriately represented by a LGCM but no or little knowledge is present about the short-term intraindividual variability. As parameter value for the LGCM, we used $\mu_I = 0, \mu_S = 3, \sigma_I^2 = 20, \sigma_S = 5, \sigma_{IS} = 2, \sigma_\epsilon^2 = 1$. For the deviation term, we used $f(t;\varphi_i) = \frac{1}{2} t \cos(2\pi \frac{1}{10} t - \varphi_i)$ with $\varphi_i \sim \mathcal{N}(0, 4\pi^2)$, which corresponds to an oscillation with person-varying phase and time-varying (increasing) amplitude. For creating the training, the test, and the interpolation sets, we used the same time points and numbers of persons as before.

Model	Interpolation	Extrapolation	Combined
LGCM	74.97	299.14	375.37
SE Model	2.85	31.641	34.278
LGCM+SE Model	2.87	30.275	33.038

Table 2: Negative log predictive probabilities on the data generated from the LGCM + unknown deviation distribution for the compared models as estimated by the different test sets.

We compared the performance of the LGCM, the SE and the LGCM+SE model. Overall, the LGCM+SE model performs best. With regard to the interpolation performance (NLPP= 2.87) it performs relatively close to the expected optimal performance, whereas the extrapolation performance is far from optimal (NLPP = 30.275). The SE is, as expected, less accurate than the LGCM+SE model in extrapolation (NLPP difference is 1.36) and has only a slight advantage over the LGCM+SE model in terms of the interpolation performance (NLPP difference is 0.02). The difference for the improved interpolation performance is caused by the LGCM+SE model regularizing towards the LGCM, so a person-specific linear trajectory instead of a person specific mean (compare Figures 3b and 3c). The LGCM, as expected, performs much worse than the former two. The interpolation performance is reduced by the lacking flexibility of the LGCM (NLPP= 74.968), whereas the extrapolation performance (NLPP= 299.14) is diminished by the LGCM adapting its uncertainty, as expressed by the predictive variance, too slow.

While we expect hybrid models such as LGCM+SE to perform best in situations where intraindividual change and intraindividual variability are present, and the parametric model for the intraindividual change is correctly specified, we also expect the hybrid models to perform almost as well as the flexible statistical learning models even if the parametric model is completely misspecified. The reason for this is that they essentially inherit the ability of the flexible statistical learning model to fit most functional forms and thus to achieve near-optimal interpolation performance. We demonstrate this by applying the LGCM+SE model to the data from the "exponential rise to the limit model". The LGCM+SE model achieves near-optimal interpolation



(a) LGCM

(b) SE Model

(c) LGCM+SE

Figure 3: Visualization of the predictive distributions on the data generated from the LGCM + unknown deviation distribution for the three considered models

performance (see, also Figure 2c). The interpolation performance of the LGCM+SE model was even slightly better than that of the SE model (NLPP difference 0.12).

6.2. Real Data: Smooth Models

We demonstrate that the hybrid random intercept SE model, which is one of the models uniquely representable by GPPM, is a suitable alternative to models routinely used in psychological research. This is the case considering the statistical learning as well as the explanatory perspective. We demonstrate this by showing that the random intercept model leads to more accurate predictions (statistical learning perspective) as well as a higher model probability (explanatory perspective) of the model compared to the continuous-time random intercept autoregressive model of order 1, which was previously used to analyze the example data set.

We start with the observation that the random intercept SE is very similar to a popular model used in psychological research, the continuous-time random intercept autoregressive model of order 1. The $n = 1$ continuous-time autoregressive model of order 1 is the Ornstein-Uhlenbeck process, which is one particular Gaussian process. The stationary Ornstein-Uhlenbeck process has mean function and kernel as follows

$$m(t;\theta) = \mu_I, \quad k(t,t';\theta) = \sigma_{se}^2 \exp\left(-\frac{|t-t'|}{l}\right).$$

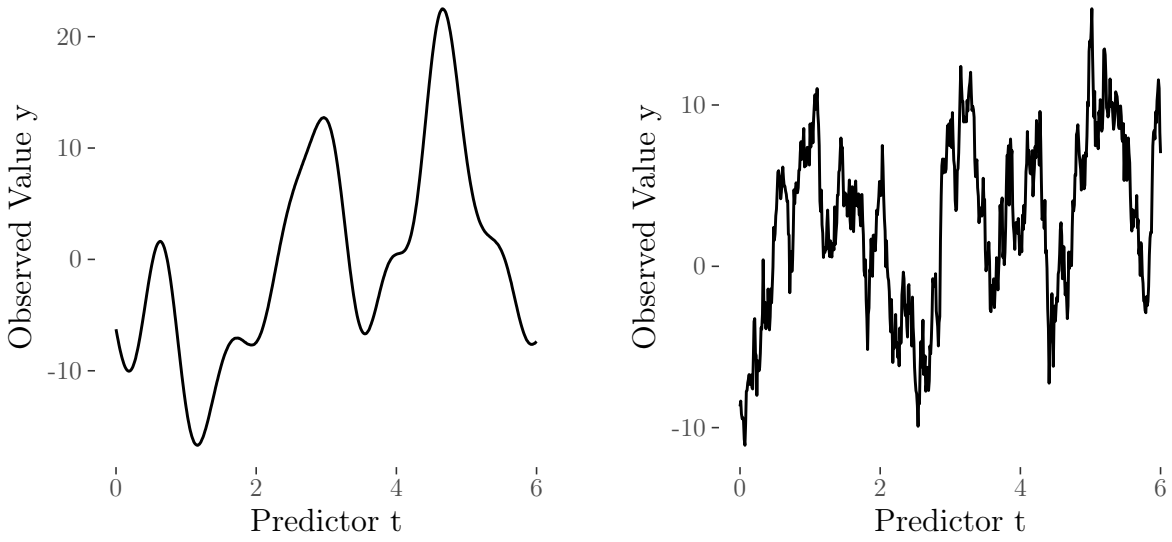
To use this model for $n > 1$ data, it is typically extended with a random intercept, which leads to the continuous-time version of the random intercept autoregressive model of order 1, which has the following GPPM representation

$$m(t;\theta) = \mu_I, \quad k(t,t';\theta) = \sigma_I^2 + \sigma_{se}^2 \exp\left(-\frac{|t-t'|}{l}\right).$$

Comparing the kernel functions of the random intercept SE model and the random intercept continuous-time autoregressive model reveals that in the former the within-person autocorrelation is assumed to decline according to a squared exponential and for the latter according to an exponential.

Despite their mathematical similarity, there is a substantial difference between the exponential and the squared exponential kernel. Both kernels are special cases of the so-called Matérn kernel (Schulz et al., 2018). From a Matérn kernel perspective, they represent two endpoints on a continuum (Schulz et al., 2018): The squared exponential kernel implies very smooth (that is infinitely differentiable) trajectories, whereas the exponential kernel implies rather unsmooth, rough, trajectories. In Figure 4, we visualize this difference by contrasting a trajectory generated from a squared exponential kernel with a trajectory generated from an exponential kernel.

On a more conceptual level, smoothness can be regarded as the mathematical implementation of the "nature does not jump" assumption, which implies that changes in nature typically do not occur abruptly, and this has already been proposed as fundamental principle in nature by, for example, Darwin (1859) and Leibniz (1704). The rough trajectories implied by the exponential model, on the other hand, are not in line with this assumption. Thus, if the "nature does not



(a) Squared exponential (b) Exponential

Figure 4: Graphical illustration of the differences between the squared exponential and the exponential kernel. Example trajectories implied by each kernel are shown. To generate the data, the variance parameter was set to $\sigma^2_{se} = 2$ and the length scale parameter to $l = 1$ for the exponential and to $l = 0.25$ for the squared exponential kernel.

jump" assumption is fulfilled using a model that implements it should lead to better predictions (statistical learning perspective) and better parameter estimates (explanatory perspective).

To investigate the usefulness of the random intercept squared exponential (SE) model (from now on simply called SE model) for psychological data analysis, we reanalyzed data that have previously been analyzed using a continuous-time autoregressive model (Voelkle, Oud, Davidov, & Schmidt, 2012). The data originate from a German panel study (Heitmeyer, 2004), measuring people aged 16 years and older who do not have an immigration background using computer-assisted interviews. Measurements were performed in 2002, 2003, 2004, 2006 and 2008, but not in 2005 and 2007.

Among other variables, authoritarianism was measured, which reflects a person's preference to submit under authorities, to orient along with the perceived conventions of the in-group, and to aggressive stances toward outgroups. For illustrative purposes, we will focus on this measure in the following. A total of $n = 2,722$ people took part in the first wave of the study, with considerable drop out over time (see Voelkle et al., 2012, for details).

To investigate whether the SE model should be preferred over the exponential model, we used an explanatory as well as statistical learning model selection procedure. For the explanatory procedure, we compared the models based on the Bayesian information criterion. We did not use the likelihood-ratio test, because the two models are not nested. As statistical learning procedure, we assessed the predictive performance of the models using cross-validation. As splitting strategy,

Measure	Exponential Model	Squared Exponential Model
BIC	10876.14	**10850.39**
NLPP	10846.54	**10821.99**

Table 3: Bayesian information criterion (BIC), and negative log predictive probability (NLPP) for the exponential and the SE model. Bold face marks the model selected on the basis of the corresponding measure. The smaller value of a measure indicates which model to select.

we split by persons. More specifically, we used leave-one-person out cross-validation. This estimates the ability of the models to predict trajectories of previously unseed persons. As before, we used the negative log predictive probability as loss functions.

As can be seen in Table 3, the prediction inaccuracy as measured by the negative log predictive probability as well as the BIC were both lower for the SE model. The Bayesian information criterion values can be translated into model posterior probabilities (Wagenmakers & Farrell, 2004). The obtained values translate into a probability of $> .99$ that the SE model is the better model for this dataset. Note, however, that this is merely a measure of relative model fit and cannot be interpreted as measure of absolute fit.

After having established that the SE model should be preferred, we investigate the impact of using the traditional exponential model instead on both explanatory and statistical learning results.

We start with the explanatory perspective. The parameter estimates and their corresponding 95%-confidence intervals are displayed in Table 4. We focus an all parameter but the length scale parameter l and the variance parameter σ^2_{se} as they implement different concepts across the two models. The estimated mean function as represented by the intercept parameter μ_I is identical across both models. In contrast, all remaining parameters are different. For example, in the exponential model, the confidence interval for the intercept variance contains 0 whereas it does not for the SE model. Thus, using a classical hypothesis testing approach, one could only conclude that there are no significant differences with regard to the starting level of authoritarianism across persons, that is, the null hypothesis of no differences in starting level cannot be rejected. However, the preferred SE model indicates significant differences in the starting levels across persons.

For the statistical learning perspective, we investigate the impact of the model choice on the predictive distribution. We have already seen that the predictive distribution of the SE model is more accurate as quantified by the lower cross-validated negative log predictive probability. We now also compare the two predictive distributions visually. In Figure 5, we show the predictive distribution obtained for one exemplary person. We plot the predictive distribution only for latent authoritarianism, that is, the authoritarianism score without being contaminated by measurement error. The most notable difference between the two predictive distributions is that the predictive mean, as well as the predictive variance, is smooth for the SE model, whereas it is not for the exponential model. This again implements the "nature does not jump" assumption.

Parameter	Lower Bound	Estimate	Upper Bound
μ_I	2.82	2.85	2.87
σ_I^2	0.00	0.00	0.11
σ_{se}^2	0.37	0.47	0.50
l	13.24	13.42	15.26
σ_ϵ^2	0.04	0.05	0.06

(a) Exponential Model (Auto-Regressive Model)

Parameter	Lower Bound	Estimate	Upper Bound
μ_I	2.82	2.85	2.87
σ_I^2	0.21	0.26	0.30
σ_{se}^2	0.16	0.19	0.23
l	20.95	21.39	30.54
σ_ϵ^2	0.07	0.08	0.08

(b) Squared Exponential Model

Table 4: 95% confidence intervals as well as maximum likelihood estimates for the parameters from exponential and the squared exponential model.

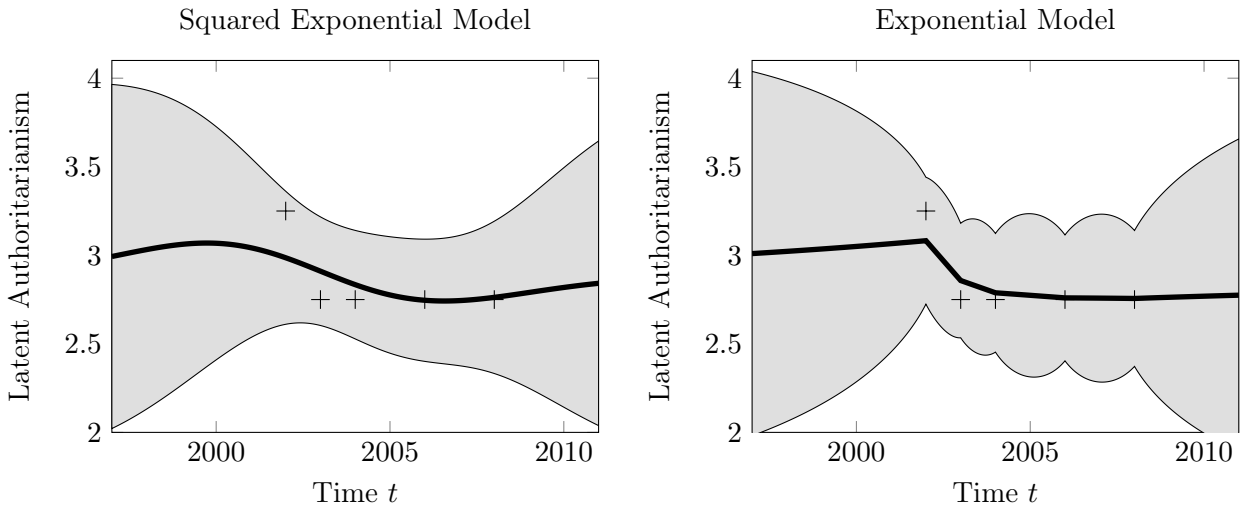


Figure 5: Person-specific predictions of the squared exponential and the exponential model for one randomly selected person. The bold line indicates the mean of the predictive distribution for every time point. The grey area displays the 95% credible region. Crosses depict observed training data.

7. Summary and Discussion

In the present paper, we have introduced Gaussian process panel modeling (GPPM), an extension of Gaussian process regression (GPR) for the analysis of panel data. GPPM provides great flexibility in specifying parametric models, nonparametric models, or combinations of both. It offers a choice of two inference frameworks focusing on either explanation or prediction. It subsumes many standard modeling approaches for longitudinal data such as linear structural equation models and state-space models as special cases but also extends the space of expressible models beyond those common approaches.

GPPMs are specified by a kernel language consisting of a mean function and a kernel. Throughout this manuscript, we have demonstrated how the flexibility of the kernel-language and its combination rules can be used to specify novel panel data models. Specifically, we used GPPM to express hybrid models such as the random intercept squared exponential (SE) model and the LGCM+ SE model. In a simulation study, we showed that the LGCM+SE model combines the advantages of the parametric LGCM and the nonparametric statistical learning SE model. The random intercept SE model was also featured in the empirical illustration, in which it was shown to be a viable alternative to the popular random intercept autoregressive model (Hamaker, Kuiper, & Grasman, 2015) when smoothness of the process is a reasonable prior belief.

Regarding inference procedures, the frequentist inference procedures for GPPM enable explanatory data analysis, which aims to recover the population distribution. The Bayesian statistical learning inference procedures provide a predictive modeling perspective that is relatively uncommon in the analysis of panel data. As demonstrated in this article, one advantage of the statistical learning inference framework is that its conclusions about the predictive accuracy of a model are also valid when the model is not correct. It is often unrealistic to assume that a chosen model is correctly specified, and adopting a predictive modeling perspective that does not rely on this assumption may be beneficial. Additionally, GPPM allows operating with hybrid models that are partly informed by theory and partly informed by data with a specific focus on maintaining generalization performance and avoiding overfitting (see,also , Brandmaier et al., 2016). Thus, we believe that the statistical learning inference perspective provides a viable addition to the methodological toolbox for analyzing panel data. Because all Bayesian quantities can be obtained analytically, statistical learning inference in GPPM is exact and faster than in commonly used Markov chain Monte Carlo-based approaches.

Within the psychometric modeling community, there have been many previous efforts to provide robust inference on covariance models both within the frequentist (e.g., Bollen, Kirby, Curran, Paxton, & Chen, 2007; Satorra, 1990) and the Bayesian inference framework (e.g., S.-Y. Lee & Xia, 2008). In contrast to our work, these approaches retain the focus on explanatory modeling. Most approaches (e.g., S.-Y. Lee & Xia, 2008; Satorra, 1990) focus on robustness concerning outliers or distributional assumptions. Other approaches go beyond this and consider more serious misspecification, which is known as structural misspecification (Satorra, 1990). However, the structural misspecification considered is often relatively mild. For example, Bollen et al. (2007) investigate structural misspecification in the sense of a few paths missing from a factor model, while the majority of the model is correctly specified. In contrast to this, the statistical learning inference approach, which estimates how well a certain model predicts, is valid

under all forms of misspecification (Breiman, 2001).

7.1. Extensions, Limitations & Future Research Directions

In the present paper, we have introduced Gaussian process panel modeling (GPPM), an extension of Gaussian process regression (GPR) for the analysis of panel data. GPPM provides great flexibility in specifying parametric models, nonparametric models, or combinations of both. It offers a choice of two inference frameworks focusing on either explanation or prediction. It subsumes many standard modeling approaches for longitudinal data such as linear structural equation models and state-space models as special cases but also extends the space of expressible models beyond those common approaches.

GPPMs are specified by a kernel language consisting of a mean function and a kernel. Throughout this manuscript, we have demonstrated how the flexibility of the kernel-language and its combination rules can be used to specify novel panel data models. Specifically, we used GPPM to express hybrid models such as the random intercept squared exponential (SE) model and the LGCM+ SE model. In a simulation study, we showed that the LGCM+SE model combines the advantages of the parametric LGCM and the nonparametric statistical learning SE model. The random intercept SE model was also featured in the empirical illustration, in which it was shown to be a viable alternative to the popular random intercept autoregressive model (Hamaker et al., 2015) when smoothness of the process is a reasonable prior belief.

Regarding inference procedures, the frequentist inference procedures for GPPM enable explanatory data analysis, which aims to recover the population distribution. The Bayesian statistical learning inference procedures provide a predictive modeling perspective that is relatively uncommon in the analysis of panel data. As demonstrated in this article, one advantage of the statistical learning inference framework is that its conclusions about the predictive accuracy of a model are also valid when the model is not correct. It is often unrealistic to assume that a chosen model is correctly specified, and adopting a predictive modeling perspective that does not rely on this assumption may be beneficial. Additionally, GPPM allows operating with hybrid models that are partly informed by theory and partly informed by data with a specific focus on maintaining generalization performance and avoiding overfitting (see, also, Brandmaier et al., 2016). Thus, we believe that the statistical learning inference perspective provides a viable addition to the methodological toolbox for analyzing panel data. Because all Bayesian quantities can be obtained analytically, statistical learning inference in GPPM is exact and faster than in commonly used Markov chain Monte Carlo-based approaches.

Within the psychometric modeling community, there have been many previous efforts to provide robust inference on covariance models both within the frequentist (e.g., Bollen et al., 2007; Satorra, 1990) and the Bayesian inference framework (e.g., S.-Y. Lee & Xia, 2008). In contrast to our work, these approaches retain the focus on explanatory modeling. Most approaches (e.g., S.-Y. Lee & Xia, 2008; Satorra, 1990) focus on robustness concerning outliers or distributional assumptions. Other approaches go beyond this and consider more serious misspecification, which is known as structural misspecification (Satorra, 1990). However, the structural misspecification considered is often relatively mild. For example, Bollen et al. (2007) investigate structural misspecification in the sense of a few paths missing from a factor model, while the majority of the

model is correctly specified. In contrast to this, the statistical learning inference approach, which estimates how well a certain model predicts, is valid under all forms of misspecification (Breiman, 2001).

7.2. Extensions, Limitations & Future Research Directions

For lack of space, we only briefly hint at some further opportunities for modeling with GPPM that may be useful in practice. Correlated error structures can be implemented by using the appropriate kernel instead of the white noise kernel $\delta(t-t')\sigma_\epsilon^2$. The autoregressive error structure, for example, is represented by the autoregressive kernel displayed in Equation 6.2. Time-varying errors can be implemented using the same approach. For example, a linear increase in measurement error is implemented by $\delta(t-t')(\sigma_1^2+\sigma_2^2 t)$. Representing more complex hierarchies beyond the simple two-level model with observations nested in persons is also possible in GPPM. We demonstrate this in Appendix B.

One current limitation of GPPM is that random effects can only be specified for linear parameters of the mean function. Consequently, multiplicative random effects (Ram & Grimm, 2015) or random effects on kernel parameters, needed to implement probabilistic person-varying measurement error, can currently not be implemented. Deterministic person-varying measurement error can already be implemented. GPPM can be extended to allow for random effects for all parameters. However, we expect that with this extension exact inference is not possible anymore, and one has to fall back to approximate inference; similar to other approaches allowing for random effects on all parameters (c.f. Asparouhov, Hamaker, & Muthén, 2017; Driver & Voelkle, 2018).

Specifying a multivariate GPPM is possible given our current framework but it may appear more intricate than in standard state-space modeling and structural equation modeling specification. Beyond the kernel for the auto-covariance of each variable, we also need cross-covariance kernels for each pair of variables (Alvarez, Luengo, Titsias, & Lawrence, 2010).

GPPM, as introduced in the present paper, is limited to continuous data. To extend GPPM to nominal or ordinal data, one can build on a rich library of methods developed for extending GPR. Just like in generalized linear models, so-called link functions (Rasmussen & Williams, 2006, Chapters 3 and 9.3) are used to accommodate non-Gaussian data. Using the same approach, non-normal measurement error for continuous data, for example, Laplace errors as commonly used in robust methods, can also be implemented. As in other extensions of linear models to accommodate non-Gaussian observations, these generalizations complicate inference. However, the appropriate algorithms have already been developed (Rasmussen & Williams, 2006) and await to be adapted to GPPM.

GPPM generalizes all methods that are restricted to Gaussian processes and use either frequentist or statistical learning inference. While this subsumes many methods, this excludes methods that imply non-Gaussian stochastic processes at the latent level such as nonlinear structural equation modeling (Jöreskog & Yang, 1996) or nonlinear state-space modeling (Chow & Zhang, 2013).

A main contribution of this work is explicitly drawing the connection of the field of kernel methods to the analysis of longitudinal data in psychological research. This opens multiple opportunities for future research: Besides the squared exponential model we have emphasized

here, many other GPR models can be readily applied to panel data (Duvenaud, 2014; Roberts et al., 2013). Among the most promising candidates are periodic models (Rasmussen & Williams, 2006, Chapter 4), and change point detection models (Turner, 2012), which could be viable alternatives to their existing state-space equivalents (Chow et al., 2018; Chow, Ram, Boker, Fujita, & Clore, 2005).

When appropriately safeguarding against overfitting, exploratory analysis has many opportunities for the analysis of panel data and can profit from research in kernel methods. One generic approach is to define a model that is flexible enough to fit most functions given enough training data. The squared exponential kernel we introduced is a prototypical example. However, exploratory analysis has been taken one step further by an algorithm that automatically learns the kernel from data and then describes the model in natural language (Lloyd, Duvenaud, Grosse, Tenenbaum, & Ghahramani, 2014). This algorithm also exploits the fact that complex models can be specified by combining a small set of base models, as we also discussed in this paper. Extending this algorithm for use in GPPM would result in a method that learns the between- and the within-person model from empirical data. This approach has the potential to find better models than the current practice of searching for a model by heuristics or merely relying on default models. Future research will have to address the right trade-offs between bias (over/underfitting) and variance (model selection uncertainty) in applying such automated model searches and, how and to what extent prior knowledge can be incorporated in this model search.

Speeding up model-fitting algorithms for panel data models is becoming increasingly important as technological progress enables us to obtain unprecedented amounts of data at little cost. Especially, fitting structural equation models on intensive longitudinal data will become a problem as the time required for parameter estimation grows cubically with the number of time points due to the necessary inversion of the model-implied covariance matrix. GPR researchers faced this problem much earlier and thus have already developed approximation algorithms, which are fast, specialized alternatives to the general, and commonly used Markov chain Monte Carlo samplers (for example, Hartikainen & Särkkä, 2010; Lawrence, Seeger, & Herbrich, 2003; Leithead & Zhang, 2007). Adopting those for GPPM holds the potential to speed up inference for longitudinal structural equation models substantially. Whether the resulting approximation error is in an acceptable range needs to be investigated.

GPPM promises to deepen our understanding of the close connections between different families of models and modeling approaches. Specifically, while we have demonstrated that GPPM generalizes linear structural equation modeling, and linear state-space modeling, it has also been shown that GPR subsumes smoothing splines, (kernel) ridge regression, Bayesian (kernel) regression, and it is closely related to other Machine Learning methods such as support vector machines and (deep) neural networks (J. Lee et al., 2017; Rasmussen & Williams, 2006). One interesting result that follows from the identification of structural equation modeling as a special case of GPPM, and GPPM's close relation to Bayesian kernel regression, is that every conventional structural equation model is equivalent to Bayesian linear regression in some high-dimensional space. We believe that making these connections explicit has the potential to foster innovations from seemingly distant research areas, such as kernel learning or deep learning, for the analysis of psychological data. In this regard, we share the hope of Yarkoni and Westfall (2017) that the predictive modeling approach is regarded as an opportunity, not a threat, and

psychological researchers equipped with a mix of classical and new methods will have a higher likelihood of finding the appropriate modeling and inference framework for their research question.

References

Alvarez, M. A., Luengo, D., Titsias, M. K., & Lawrence, N. D. (2010). Efficient multioutput Gaussian processes through variational inducing kernels. In *International Conference on Artificial Intelligence and Statistics* (pp. 25–32).

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2017). Dynamic latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 257–269. doi:10.1080/10705511.2016.1253479

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*. Data Mining for Software Trustworthiness, *191*, 192–213. doi:10.1016/j.ins.2011.12.028

Bishop, C. M. (2006). *Pattern recognition and machine learning* (1st ed.). New York, NY: Springer.

Bollen, K. A. (1989). *Structural equations with latent variables* (1st ed.). New York, NY: Wiley.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent Variable Models Under Misspecification: Two-Stage Least Squares (2SLS) and Maximum Likelihood (ML) Estimators. *Sociological Methods & Research*, *36*(1), 48–86. doi:10.1177/0049124107301947

Brahim-Belhouari, S., & Bermak, A. (2004). Gaussian process for nonstationary time series prediction. *Computational Statistics & Data Analysis*, *47*(4), 705–712. doi:10.1016/j.csda.2004.02.006

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological methods*, *21*(4), 566.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231. doi:10.1214/ss/1009213726

Chow, S.-M., Ou, L., Ciptadi, A., Prince, E. B., You, D., Hunter, M. D., ... Messinger, D. S. (2018). Representing sudden shifts in intensive dyadic interaction data using differential equation models with regime switching. *Psychometrika*, 1–35. doi:10.1007/s11336-018-9605-1

Chow, S.-M., Ram, N., Boker, S. M., Fujita, F., & Clore, G. (2005). Emotion as a thermostat: Representing emotion regulation using a damped oscillator model. *Emotion*, *5*(2), 208–225. doi:10.1037/1528-3542.5.2.208

Chow, S.-M., & Zhang, G. (2013). Nonlinear regime-switching state-space (RSSS) models. *Psychometrika*, *78*(4), 740–768. doi:10.1007/s11336-013-9330-8

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*(6, Pt.1), 426–443. doi:10.1037/h0026714

Cox, G. E., Kachergis, G., & Shiffrin, R. M. (2012). Gaussian process regression for trajectory analysis. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1440–1445).

Darwin, C. (1859). *On the origin of species*. London, England: John Murray.

Driver, C. C., & Voelkle, M. C. (2018). Hierarchical Bayesian continuous time dynamic modeling. *Psychological Methods*. doi:10.1037/met0000168

Duvenaud, D. K. (2014). *Automatic model construction with Gaussian processes* (Doctoral Dissertation, University of Cambridge).

Duvenaud, D. K., Lloyd, J. R., Grosse, R., Tenenbaum, J. B., & Ghahramani, Z. (2013). Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922*.

Duvenaud, D. K., Nickisch, H., & Rasmussen, C. E. (2011). *Additive Gaussian processes* (J. Shawe-Taylor, R. S. Zemel, J. C. Bartlett, F. Pereira, & K. Weinberger, Eds.). Red Hook, NY: Curran Associates.

Ghisletta, P., Mason, F., von Oertzen, T., Hertzog, C., Nilsson, L.-G., & Lindenberger, U. (2019). On the Use of Growth Models to Study Normal Cognitive Aging. *Lifebrain*.

Hall, P., Müller, H.-G., & Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent Gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(4), 703–723. doi:10.1111/j.1467-9868.2008.00656.x

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*. Longitudinal Topics, *20*(1), 102–116. doi:10.1037/a0038889

Hartikainen, J., & Särkkä, S. (2010). Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 379–384). doi:10.1109/MLSP.2010.5589113

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction, Springer Series in Statistics*. Springer New York.

Heitmeyer, W. (Ed.). (2004). *Deutsche Zustände. Folge 3 [Current state in Germany. Series 3]*. Frankfurt am Main: Suhrkamp.

Hsiao, C. (2014). *Analysis of panel data*. Cambridge University Press.

Jöreskog, K. G., & Yang, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. *Advanced structural equation modeling: Issues and techniques*, (3), 57–88.

Karch, J. D., Sander, M. C., von Oertzen, T., Brandmaier, A. M., & Werkle-Bergner, M. (2015). Using within-subject pattern classification to understand lifespan age differences in oscillatory mechanisms of working memory selection and maintenance. *NeuroImage*, *118*, 538–552. doi:10.1016/j.neuroimage.2015.04.038

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on artificial intelligence*.

Kruschke, J. (2014). *Doing bayesian data analysis* (2nd ed.). San Diego, CA: Academic Press.

Lawrence, N. D., Seeger, M., & Herbrich, R. (2003). Fast sparse Gaussian process methods: The informative vector machine. In *Advances in neural information processing Systems* (Vol. 15, pp. 609–616). Cambridge, MA: MIT Press.

Leckie, G. (2013). Module 12: Cross-Classified Multilevel Models. https://www.cmm.bris.ac.uk/lemma/. LEMMA VLE, University of Bristol, Centre for Multilevel Modelling.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., & Sohl-Dickstein, J. (2017). Deep Neural Networks as Gaussian Processes. *arXiv:1711.00165 [cs, stat]*. arXiv: 1711.00165 `[cs, stat]`

Lee, S.-Y., & Xia, Y.-M. (2008). A Robust Bayesian Approach for Structural Equation Models with Missing Data. *Psychometrika*, *73*(3), 343. doi:10.1007/s11336-008-9060-5

Leibniz, G. W. (1704). *Nouveaux essais sur l'entendement humain [New essays on human understanding]*. Paris, France: Hachette.

Leithead, W. E., & Zhang, Y. (2007). $O(N^2)$-operation approximation of covariance matrix inverse in Gaussian process regression based on quasi-Newton BFGS method. *Communications in Statistics – Simulation and Computation*, *36*(2), 367–380. doi:10.1080/03610910601161298

Lloyd, J. R., Duvenaud, D. K., Grosse, R., Tenenbaum, J. B., & Ghahramani, Z. (2014). Automatic construction and natural-language description of nonparametric regression models. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.

Micchelli, C. A., Xu, Y., & Zhang, H. (2006). Universal kernels. *Journal of Machine Learning Research*, *7*(Dec), 2651–2667.

Nesselroade, J. R. (1991). The warp and the woof of the developmental fabric. In *Visions of aesthetics, the environment & development: The legacy of Joachim F. Wohlwill* (pp. 213–240). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Pek, J., & Wu, H. (2015). Profile likelihood-based confidence intervals and regions for structural equation models. *Psychometrika*, *80*(4), 1123–1145. doi:10.1007/s11336-015-9461-1

Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. doi:10.1007/s11222-016-9649-y

Preacher, K. J., Wichman, A. L., Briggs, N. E., & MacCallum, R. C. (2008). *Latent growth curve modeling*. Sage.

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

Ram, N., & Grimm, K. J. (2015). Growth curve modeling and longitudinal factor analysis. *Handbook of child psychology and developmental science*, 1–31.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., & Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A*, *371*(1984), : 20110550. doi:10.1098/rsta.2011.0550

Saatçi, Y., Turner, R. D., & Rasmussen, C. E. (2010). Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning* (pp. 927–934).

Särkkä, S., & Hartikainen, J. (2012). Infinite-dimensional Kalman filtering approach to spatio-temporal Gaussian process regression. In *International Conference on Artificial Intelligence and Statistics* (pp. 993–1001).

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality and Quantity*, *24*(4), 367–386. doi:10.1007/BF00152011

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16.

Seeger, M. (2002). *Relationships between Gaussian processes, support vector machines and smoothing splines*. Institute for Adaptive and Neural Computation (ANC), University of Edinburgh, UK.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. doi:10.1214/10-STS330. arXiv: 1101.0891

Stoel, R. D., Garre, F. G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*(4), 439–455. doi:10.1037/1082-989X.11.4.439

Taboga, M. (2012a). Likelihood ratio test. In *Lectures on Probability Theory and Mathematical Statistics* (2nd ed.). Lexington, KY: CreateSpace Independent Publishing Platform.

Taboga, M. (2012b). Maximum likelihood. In *Lectures on Probability Theory and Mathematical Statistics* (2nd ed.). Lexington, KY: CreateSpace Independent Publishing Platform.

Turner, R. D. (2012). *Gaussian processes for state space models and change point detection* (Doctoral Dissertation, University of Cambridge).

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Voelkle, M. C., Oud, J. H., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods*, *17*(2), 176–192. doi:10.1037/a0027543

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. doi:10.3758/BF03206482

Walls, T. A., & Schafer, J. L. (Eds.). (2006). *Models for intensive longitudinal data*. Oxford; New York: Oxford University Press.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 1100–1122. doi:10.1177/1745691617693393

A. Exemplary Gaussian Process Panel Models

In Table 5, we present a selection of models expressible in GPPM with corresponding mean function and kernel. For brevity, all models are shown without measurement error, which crucially needs to be added when applying these models. For white noise Gaussian error, add $\delta(t-t')\sigma_\epsilon^2$ to the kernel. For simplicity, we assumed no covariance between the intercept and the slope terms for the quadratic LGCM. The quadratic LGCM can be further extended to polynomials of arbitrary degree p. The mean function and kernel representing a polynomial of degree p are $m(t;\theta) = \sum_{i=0}^{p} \mu_i t^i$ and $k(t,t';\theta) = \sum_{i=0}^{p} t^i \sigma_i^2 t'^i$.

For the statistical learning models, the number of predictors is typically quite high, which is why we use x, which denotes a vector of predictors, instead of t, which denotes one single predictor, namely time. There are many more statistical learning methods that can be expressed as a GPR and thus as a GPPM. Presenting their mean and kernel function goes beyond the scope of this text. We refer the interested reader to Seeger (2002) and Rasmussen and Williams (2006, Chapter 6.3) for smoothing splines; Duvenaud, Nickisch, and Rasmussen (2011) for generalized additive models; and J. Lee et al. (2017) for (deep) neural networks.

Mixing the statistical learning models and the longitudinal psychological models lead to the two hybrid models that we discuss in detail in the Illustration section. Here, we also present the

Name	Mean Function	Kernel
	Psychometric Longitudinal Models	
Random Intercept	μ_I	σ_I^2
Random Intercept, Fixed Slope	$\mu_I + \mu_S t$	σ_I^2
Linear LGCM	$\mu_I + \mu_S t$	$\sigma_I^2 + t\sigma_S^2 t' + \sigma_{IS}(t + t')$
Quadratic LGCM	$\mu_I + \mu_S + \mu_{S2} t^2$	$\sigma_I^2 + t\sigma_S^2 t' + t^2\sigma_{S2}^2 t'^2$
Random Intercept AR(1)	μ_I	$\sigma_I^2 + \sigma_e^2 \exp\left(-\frac{\lvert t-t'\rvert}{l}\right)$
	Statistical Learning Models	
Ridge Regression	0	$x^\top \sigma_b^2 x'$
SE Model	0	$\sigma_{se}^2 \exp\left(-\frac{\lvert\lvert x-x'\rvert\rvert^2}{l}\right)$
	Hybrid Models	
Random Intercept + SE	μ_I	$\sigma_I^2 + \sigma_{se}^2 \exp\left(-\frac{(t-t')^2}{l}\right)$
LGCM + SE	$\mu_I + \mu_S t$	$\sigma_I^2 + t\sigma_S^2 t' + \sigma_{IS}(t + t') + \sigma_{se}^2 \exp\left(-\frac{(t-t')^2}{l}\right)$
Exponential Decay + SE	$\mu_b + \mu_d \exp(-ts)$	$\sigma_b^2 + \sigma_d^2 \exp\left(-(t + t')s\right) + \sigma_{se}^2 \exp\left(-\frac{(t-t')^2}{l}\right)$

Table 5: Mean functions and kernels for exemplary Gaussian process panel models.

exponential decay + SE model, a hybrid model whose parametric part implies the nonlinear "exponential rise to the limit" trajectories.

B. More Complex Hierarchies

To showcase the ability of GPPM to represent more complex hierarchies, we demonstrate how to specify the longitudinal version of the two-way cross-classified model introduced in Leckie (2013). The model expressed using the multilevel notation is as follows:

$$Y_{ijk}(t) = \beta_0 + v_k + u_j + \epsilon_{ijk}(t), \text{ with } v_k \sim \mathcal{N}(0, \sigma_v^2), u_j \sim \mathcal{N}(0, \sigma_u^2), \epsilon_{ijk}(t) \sim \mathcal{GP}(0, \sigma_\epsilon^2)$$

where $Y_{ijk}(t)$ is the age t score of student i who attended primary school j and secondary school k, β_0 is the mean score across all schools, v_k is the effect of secondary school k, u_j is the effect of primary school j, and $e_{ijk}(t)$ is the student-time level residual error term. The random effects and residual errors are assumed independent of one another. Scores are thus nested within students. Students are nested within primary and secondary schools but primary schools are not nested in secondary schools or vice versa.

If we assume that we observe predictors time t, primary school index j, and secondary school index k, the resulting GPPM is

$$Y_{ijk}(t) \sim \mathcal{GP}(\beta_0, \delta(k - k')\sigma_v^2 + \delta(j - j')\sigma_u^2 + \delta(t - t')\sigma_\epsilon^2).$$