Transformed Gaussian Processes for Characterizing a Model's Discrepancy

Aurélien Nioche NIOCHE.AURELIEN@GMAIL.COM *School of Computing Science, University of Glasgow, Glasgow G12 8RZ, UK*

Ville Tanskanen VILLE.TANSKANEN@HELSINKI.FI *Department of Computer Science, University of Helsinki, 00560 Helsinki, Finland*

Marcelo Hartmann MARCELO.HARTMANN@HELSINKI.FI *Department of Computer Science, University of Helsinki, 00560 Helsinki, Finland*

Arto Klami ARTO.KLAMI@HELSINKI.FI *Department of Computer Science, University of Helsinki, 00560 Helsinki, Finland*

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Mathematical models of observational phenomena are at the core of experimental sciences. By learning the parameters of such models from typically noisy observations, we can interpret and predict the phenomena under investigation. This process, however, assumes that the model itself is correct and that we are only uncertain of its parameters. In practice, this is rarely true, but rather the model is a simplification of the actual generative process. One proposed remedy is a post hoc investigation of how the model differs from reality, by explicitly modeling the discrepancy between the two. In this paper, we use transformed Gaussian processes as flexible models for this. Our formulation relaxes the assumption on the correctness of the model by assuming it is only correct in expectation, and it directly supports both additive and multiplicative corrections, treated separately in the literature, using suitable transformations. We demonstrate the approach in two example cases: modeling human growth (relation age-height) and modeling the risk attitude (relation reward-utility). The former provides a simple example, while the second case highlights the importance of the transformations in obtaining meaningful information about the discrepancy.

Keywords: Model discrepancy; Gaussian process; Applied machine learning; cognitive modeling.

1. Introduction

An important part of scientific activity consists of building mathematical models of observational phenomena. For example, physicists build models of motion, ecologists of population evolution, and cognitive scientists of decision-making. On a more abstract level, we consider a general class of mathematical models for describing an input-output relationship such that $y \approx M(x)$, where x corresponds to a vector of covariates and y is a vector of outcomes. For instance, a simple growth model describes how the (scalar) height y depends on age x.

The model $M(x, \theta)$ usually takes the form of a parameterized function designed, e.g., from first principles or as the result of decades of research in the field, but where the param-

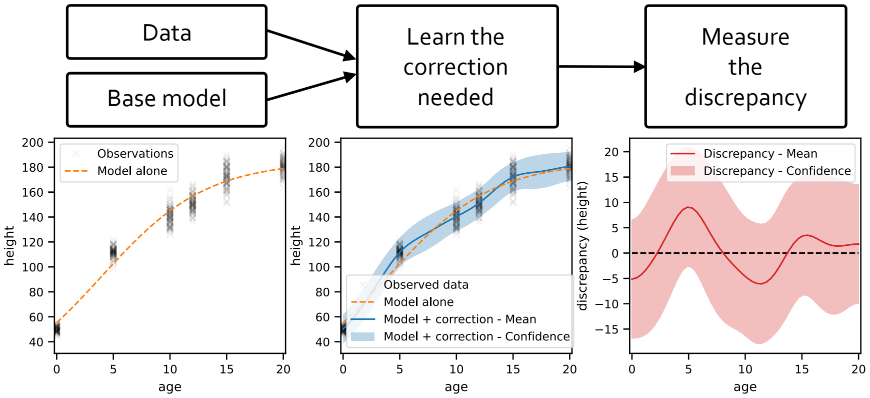

Figure 1: General workflow for a model's discrepancy modeling. We assume that we have some data and a model for that data (left panel). We propose to use a transformed Gaussian process (middle panel) to estimate the expected discrepancy between the model and the actual generative process given the data (right panel), by constraining the discrepancy model so that the underlying model is assumed to be correct in expectation.

eters θ are unknown. The parameters are determined by fitting the model to a particular data set $D = (X, Y)$, collected for the study at hand. In the context of a growth model, the data could be measurements made for a specific sample of humans. This simplified process implicitly assumes that the model accurately describes the underlying process and that the uncertainty lies only in the values of the parameters and in the measurement process.

In most cases, however, we are also unsure about the model itself, for instance, because it is an oversimplification of a complex process (e.g., models of human cognition). The classical solution is to consider multiple alternative models, using *model selection* techniques to pick the most suitable one amongst a set of alternatives, or *model averaging* techniques to weight multiple models according to their relative quality (Wasserman, 2000; Claeskens and Hjort, 2008). The former has the fundamental problem of committing to a single model and its faults when none of the alternatives is correct. The latter solution often provides good predictions but is problematic from the perspective of scientific analysis as we no longer have a single model to interpret.

An alternative strategy, first proposed by Kennedy and O'Hagan (2001) and extended in this work, is to use the assumed model but separately investigate its fit by modeling the discrepancy between the model and the data generative process using a flexible data-driven model. This allows the researcher to investigate how the model deviates from the actual generative process. This information can be used, e.g., for deciding which parameters can be safely interpreted, how to improve the model, or for instance, in the context of cognitive modeling, to identify which subjects match the assumed model. This approach has been used in a broad range of applications from experimental physics (Arendt et al., 2012; Bhat et al., 2017) to health economics (Strong et al., 2012).

Since the goal is to model any type of discrepancy between the model and the underlying generative process, the model for the discrepancy needs to be very flexible. The most common choice in the literature has been to use Gaussian processes (GP) (Brynjarsdóttir and O'Hagan, 2014; Gardner et al., 2021) that provide a highly flexible family while also characterizing the uncertainty over the estimates. There are now two sources of uncertainty that we aim to deal with: the parameters of the theory-based model and the parameters of the discrepancy model. This can cause identifiability issues by the discrepancy model correcting for bad parameter values of $M(x, \theta)$ (Brynjarsdóttir and O'Hagan, 2014; Gardner et al., 2021). Therefore, one core challenge is to keep certain flexibility to model potentially large deviations while constraining the optimization problem enough to guarantee interpretable results.

Building on the literature of GP-based discrepancy models, we propose to model the discrepancy as *suitably transformed* Gaussian process. We argue they provide the right balance between flexibility and constraints needed for interpretation. The main contributions relative to the existing literature on discrepancy models are:

- We explicitly formulate the assumption of models being correct *in expectation* as a more relaxed alternative consisting of assuming correct models. This offers a more solid basis compared to arbitrary discrepancy models.
- We show how this assumption can be ensured for different transformations, therefore providing the means of having *unbiased corrections*, a property that is only implicitly assumed by additive discrepancy models.
- We show that this approach based on transformations allows characterizing both additive and multiplicative discrepancies, while also supporting more constrained types of discrepancies, in contrast to previous methods like Gardner et al. (2021) that only support additive discrepancy.
- Finally, we consider exploring the discrepancy as part of an iterative inference workflow (Gelman et al., 2020).

We demonstrate our framework in two empirical case studies. The first one deals with human growth modeling and is designed to be maximally simple to support the understanding of the approach. The second one deals with (human) decision-making modeling, a prototypical case where the models describing the utilities and the distorted perception of probabilities are simplifications that do not necessarily match reality (see, e.g., Stott (2006)). The second example highlights the importance of the choice of transformations for meaningful information about the model discrepancy.

2. Modeling a Model's Discrepancy

2.1. Problem Formulation

We assume that the users are working with some model $y \approx M(x, \theta)$, often theory-based and interpretable, and that they have already estimated the parameters $\hat{\theta}$ of the model conditional on some available data D. Our task is then to study how well the model matches the actual generative process $f(x)$ that is unknown. We accomplish this task by modeling it as a combination of $M(x, \hat{\theta})$ — written thereafter as $M(x)$ for compactness — and a *correction term* $r(x)$. The purpose of the correction term is to model the discrepancy between the base model and the true data generative process and explain how the base model $M(x)$ could be changed to improve the match. The approach is illustrated conceptually in Figure 1, which gives an overview of the workflow.

To describe arbitrary discrepancies, we need a flexible family of functions for the correction term $r(x)$. Furthermore, to provide reliable information about the discrepancy, we want to estimate the posterior distribution $p(r(\cdot)|M(\cdot), X, Y)$ of the correction as the justified basis for verifying possible discrepancies. Gaussian processes (GP) (Rasmussen and Williams, 2006) are a natural tool for achieving this, allowing computationally efficient modeling of the posterior over a family of flexible smooth functions, and indeed GPs have been broadly used for modeling discrepancy (Kennedy and O'Hagan, 2001; Brynjarsdóttir and O'Hagan, 2014; Gardner et al., 2020, 2021). Our work deviates from the previous approaches by considering a more general but still interpretable family of discrepancies, obtained by combining the model $M(x)$ and the correction term $r(x)$ in a novel manner.

Understanding the computational details is not critical for grasping the nature of the approach: it is sufficient to think of $r(x)$ as an arbitrary smooth function. Hence we will postpone the formal treatment of GPs to Section 2.4, and start by introducing the model from a higher-level perspective.

2.2. Model

We assume the data-generative process follows

$$f(x) = h(h^{-1}(M(x)) + r(x)), \tag{1}$$

$$\text{such that: } \mathbb{E}_{r(x)}[f(x)] = M(x). \tag{2}$$

In practice, this means that the inference starts from the base model (which we argue is the best guess for the prior given data) and proceeds to infer the posterior of $r(x)$.

Note that at the core of this expression is an *additive correction* – we assume that the generative process for any x can be obtained by modulating the base model by adding a suitable correction term $r(x)$. However, this correction is done in a *transformed space* in order to better match the scope of possible deviations. This transformation is controlled by the link function $h(\cdot)$ chosen so that the additive correction is reasonable in the space of $h^{-1}(M(x))$; the inverse transformation maps the outputs to a real space where additive noise assumption is often reasonable. For example, to model discrepancy for models that output probabilities we need the sigmoid transformation so that the corrections are meaningful also for the extreme values.

The other key element is the constraint (2), which states that we assume the model to be *correct in expectation*. This is a milder assumption than directly assuming the model to be the correct description of the phenomenon but still offers a more solid theoretical basis for modeling the discrepancy. The assumption is justified in most applications since models that violate it significantly would likely be rejected in the early stage. As will be clarified in Section 2.5, it is in general not sufficient to assume the correction term $r(x)$ itself to have a mean of zero, and hence the constraint needs to be respected either by adjusting the prior mean or the process (1) itself.

Finally, the observed data D relates to the model defined by Eq. (1) and (2) via a suitable likelihood. The discrepancy model is often estimated using the same data D that was used to fit the base model $M(x)$, but if additional observations are available, we can also fit the discrepancy using another data set D'. We assume that the observations are noisy evaluations of the data generative process $f(x)$, so that $\mathbb{E}[y|f(x)] = f(x)$. This implies $\mathbb{E}\left[\mathbb{E}[y|f(x)]\right] = M(x)$, which holds in our case studies. In general, it can be satisfied by a suitable reparameterization of the likelihood, even when $\mathbb{E}[y|f(x)] \neq f(x)$ for some parameterization (e.g. switching from rate to scale parameterization in the case of exponential distribution).

2.3. Transformations

The intuition on how h controls the nature of the deviation can be obtained by inspecting specific choices. For the identity function $h(z) = z$ (where z is a generic dummy variable), Eq. (1) becomes simply $f(x) = M(x)+r(x)$ and models purely *additive* discrepancies, which has been extensively studied in the field since the work of Kennedy and O'Hagan (2001). *Multiplicative* discrepancies, studied previously by He and Xiu (2016), can be modelled using $h(z) = \exp(z)$ as the model then becomes $f(x) = M(x)\exp(r(x))$ with $\exp(r(x)) > 0$.

Our formulation covers both of these special cases but also allows for inspecting a broader family of discrepancies not addressed by previous work. For instance, for models and true processes restricted to be probabilities ($f(x) \in [0, 1]$) or otherwise constrained to a finite interval, the possible discrepancies will also have a finite range and they will be skewed away from the limits. In contrast to previous methods, our approach supports such constraints directly. For instance, for the specific case of $f(x) \in [0, 1]$, a natural choice would be the sigmoid transformation $h(z) = (1 + \exp(-z))^{-1}$. Then $h^{-1}(z)$ maps the outputs to a real space for modeling the discrepancy and $h(z)$ ensures that all corrected outputs still satisfy the constraint.

Upon first inspection, the need to choose the transformation $h(\cdot)$ may seem like a limitation that makes the process more complicated. However, since the primary goal of the approach is the exploration and inspection of possible discrepancies, it should be seen as an advantage. The set of possible transformations is typically relatively small, corresponding to the standard link functions used e.g. for generalized linear models, and the researcher can inspect the solutions for all choices for a better understanding of the kinds of discrepancies their model may suffer. Suitable functions for different types of constraints can also be automatically determined if desired, by selecting transformations that map the image of $M(x)$ to the whole real line; see, e.g., the transformations the `Stan` probabilistic programming language (Carpenter et al., 2017) uses.

2.4. Implementation using Gaussian Processes

Having completed the high-level description and motivation, we now turn our attention to implementing the discrepancy model. In principle, we could use any flexible family of functions as $r(x)$, such as deep neural networks, but the choice of Gaussian processes offers advantages in terms of easy uncertainty quantification, direct control of the functions' smoothness, and reliable learning from small data. Next, we provide sufficient background on Gaussian processes and the computational details needed in our case.

Formally, a Gaussian process is a stochastic process over a collection of some (often continuous) index set $\mathcal{X}$ such that, for every finite collection $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$, the function values associated to the set X follow a multivariate Gaussian distribution. Hence, $r(\cdot)$ can be expressed as:

$$r(X)|\tau \sim \mathcal{N}\left(\mu(X), K(X,X)\right) \tag{3}$$

where $\mu(\cdot)$ and $K(\cdot)$ are respectively the mean and covariance functions, and where τ are the hyperparameters of (here only) the kernel. The covariance function determines the smoothness of the functions. Our formulation is agnostic to the choice of the kernel, allowing the user either to specify a kernel based on domain expertise or to default to standard choices (e.g., squared exponential kernel). This choice is entirely analogous to all applications of GP in machine learning, and we refer the reader to the excellent source by Rasmussen and Williams (2006) for further information on how to choose the right kernel.

The choice of the mean function is often treated lightly in GP literature, typically setting $\mu(\cdot) = 0$. However, it plays an essential role in our approach since the mean $\mu(\cdot)$ influences the condition (2), corresponding to the assumption of the base models being on average correct. We provide more details about the choice of the mean function in Section 2.5.

Given a model $M(\cdot)$, the prior $r(\cdot) \sim GP(\mu(\cdot), K(\cdot))$, and the observed data $D = (X, Y)$, we form the posterior distribution

$$p(r(\cdot)|M(\cdot), X, Y) \propto p(Y|M(\cdot), X, r(\cdot))p(r(\cdot)|M(\cdot), X),$$

where $p(Y|M(\cdot), X, r(\cdot))$ is the application-specific likelihood. For the identity transformation $h(z) = z$ and Gaussian likelihood $p(Y|M(\cdot), X, r(\cdot))$, this can be done analytically in closed form since the prior is conjugate to the likelihood, following the standard GP regression procedure. The covariance hyperparameters τ can be selected to maximize the marginal likelihood $p(Y|M(\cdot), X, \tau)$, which also has an analytic expression.

For other choices of $h(\cdot)$ and for non-Gaussian likelihoods, we need approximate inference. We use variational inference (VI) for computational efficiency, but because our approach is indifferent to the inference method, other standard choices of MCMC and Laplace approximation would be applicable as well. We approximate the posterior $p(r(\cdot)|M(\cdot), X, Y, \tau)$ with approximate distribution $q(r(\cdot)|\lambda)$ that depends on *variational parameters* λ. A variational inference procedure is conducted through the optimization of a KL-divergence measure s.t. $\hat{\lambda}, \hat{\tau} = \arg\min_{\lambda,\tau} \mathrm{KL}(q(r(\cdot)|\lambda)||p(r(\cdot)|M(\cdot), X, Y, \tau))$; see Blei et al. (2017) for an excellent overview and detailed description of variational inference. Our implementation builds on GPyTorch (Gardner et al., 2019), uses inducing points for scalability, and optimizes both λ and τ simultaneously, but the technical details of the inference are not particularly essential for this work. For more information about VI for GPs and sparse GPs, we invite the reader to refer to Hensman et al. (2015) or Damianou (2015).

2.5. Ensuring Unbiased Corrections

The assumption according to which the model is correct in expectation is encoded by the constraint in Eq. (2). For the simplest choice of $h(z) = z$, the condition (2) is trivially achieved by setting $\mu(x) = 0$, but for other choices, we need to either use non-zero mean functions or change the model itself slightly.

To see this, let us consider the case of $h(z) = \exp(z)$. The expectation of (log-normally distributed) $exp(r(x))$ is then $\exp(\mu(x) + \sigma^2(x)/2)$ and to remove the bias we need to set

$\mu(x) = -\frac{\sigma^2(x)}{2}$ where $\sigma^2(x) = K(x,x)$ is the prior variance of the Gaussian process. For some transformations, we cannot compute the required prior mean analytically, but standard approximations are easy to derive for transformations of interest. For example, for the sigmoid transformation, the prior mean needs to be $\mu(x) = \Phi^{-1}(M(x))\sqrt{0.588^{-2} + \sigma^2(x)} - h^{-1}(M(x))$, where Φ^{-1} is the probit function (see Appendix for derivation).

An alternative, general-purpose approach is to assume an approximate process of (1) by using a second-order Taylor series approximation. We then have

$$\tilde{f}(x) = h\Big(h^{-1}(M(x)) + r(x)\Big) - \frac{\sigma^2}{2} h''\Big(h^{-1}(M(x))\Big),$$

so that $\mathbb{E}[\tilde{f}(x)] \approx M(x)$ when $\mathbb{E}[r(x)] = 0$. This provides a general-purpose solution that only requires the computation of the second derivatives $h''(z)$ that are easy for all transformations.

2.6. Measuring and Visualizing the Discrepancy

The outcome of the discrepancy model is the posterior distribution of the corrections $r(x)$, which can be interpreted in numerous ways. In practice, we recommend separately characterizing the overall discrepancy with a numeric summary and visualizing the local discrepancy as a function of x.

For summaries, one can consider any standard distance measure over x that measures the bias (how much the corrected process differs from the theoretical model M) and the variance (the uncertainty about the correction). As a concrete example, in the case studies, we use as overall *discrepancy measure* δ the expected ℓ_1-distance over the x-domain

$$\delta = \frac{\int_{x \in [x_{min}, x_{max}]} |\mathbb{E}[f(x)|M(\cdot), X, Y] - M(x)| dx}{x_{max} - x_{min}}.$$

As overall *measure of the uncertainty* u_δ, we use the expected amplitude of the credible interval at 95% over the x-domain

$$u_\delta = \frac{\int_{x \in [x_{\min}, x_{\max}]} P_{97.5\%}(f(x)) - P_{2.50\%}(f(x)) dx}{x_{max} - x_{min}},$$

with $P_{i\%}(z)$, the i-percentile of z. In practice, both measures are estimated by sampling $f(x)$ at a dense grid of x.

For visualizing the discrepancy, one possibility is to plot the posterior distribution for each dimension of x separately, as done in all figures of this paper. The mean tells about where the initial model has potential flaws, and the uncertainty tells about the confidence one can have about each of these supposed "flaws" given the current set of observations.

2.7. Workflow

To better explain how the previous section relates to the overall modeling workflow, we briefly summarize the typical steps of the process. Our solution is easy to incorporate into existing workflows since it is fully decoupled from the process of fitting the model M, similarly to Gardner et al. (2020) but unlike the majority of previous work that estimates the correction simultaneously with the model parameters.

We assume that the scientist has already collected data D about their problem of interest and has selected a particular model M to investigate. To study the model discrepancy, they would then:

1. Fit the model M, estimating $\hat{\theta}$ with any optimization routine; our approach does not depend on how the parameters were estimated.

2. Choose a covariance function K for controlling the smoothness of the discrepancy. In the absence of domain knowledge, a standard choice such as squared exponential can be used.

3. Select the function h based on prior constraints for the model outputs and the expected form of discrepancy. Alternatively, run the model with several choices.

4. Find the posterior distribution of the discrepancy as explained in Section 2.4.

5. Investigate the discrepancy, using measures of overall discrepancy (Section 2.6) and by visual inspection of the discrepancy plots (e.g. as in Fig. 2, 3 or 4).

6. Interpret and possibly modify or reject the model M according to the observed discrepancy.

As support for following this workflow for new applications, we provide an implementation building on top of *GPyTorch* (Gardner et al., 2019), at `https://github.com/AurelienNioche/TransformedGP`.

3. Related Work

In the early literature, authors studied the discrepancy by learning the model and the correction simultaneously (Kennedy and O'Hagan, 2001). Brynjarsdóttir and O'Hagan (2014) thoroughly investigated the identifiability problems of the additive correction model proposed by Kennedy and O'Hagan (2001) and proposed using stronger priors for the correction term to mitigate the issues. This strategy, however, has the fundamental problem of requiring prior information directly on the discrepancy, which is very rarely available. Recently, Gardner et al. (2020, 2021) proposed decoupling the inference process by first inferring the theoretical model's parameters θ and only estimating the discrepancy afterwards. However, they only considered additive corrections. We build on this stream of research, providing a more flexible family for discrepancies and considering the overall workflow.

Our work also relates to the broader literature on hybrid models combining theory-based and data-driven models in alternative ways, presented here from the perspective of our case example of cognitive decision-making models. Plonsky et al. (2017) combine data-driven models such as random forests and decision-making models to improve predictive accuracy, whereas Afrabandpey et al. (2020) considered learning Bayesian non-parametric models that retain interpretability of theory-based reference models also in other application domains. Bourgin et al. (2019) proposed to train neural networks with data simulated from classical decision-making models to facilitate training these flexible models with fewer observations,

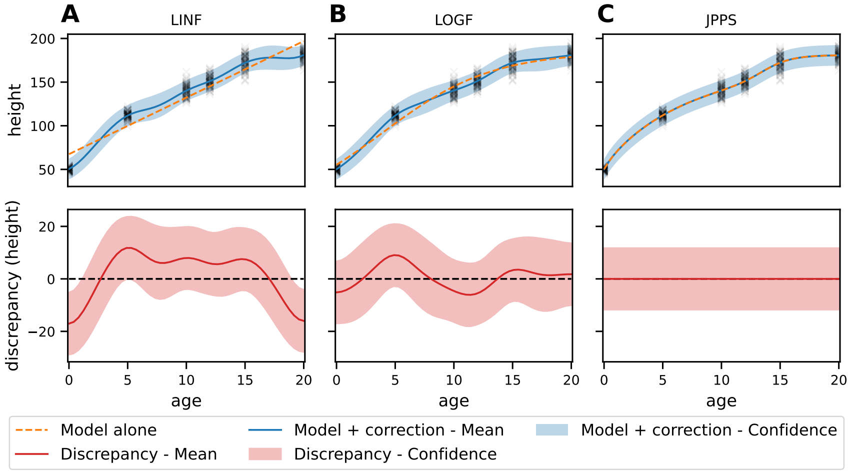

Figure 2: Human growth modeling. Each subfigure (column) presents a different base model: (A) linear, (B) logistic, (C) theory-based model. The top row presents the predictions of the model alone and with the correction. The bottom row represents the expected discrepancy.

and Rajan and Miyapuram (2020) proposed to make use of hybrid models to individualize the inference of the latent processes. Recently, Peterson et al. (2021) proposed to keep the structure of interpretable models but to replace the deterministic definitions of each component with a flexible representation learned using a neural network. Finally, recent studies on residual analysis on behavioral models (Agrawal et al., 2020) (equivalent to additive correction) have been proposing to use of flexible function approximators to make data-driven corrections for the theoretical model. Hybrid models often lose some of the interpretability of the theoretical model by either using theory-based features as features of a black-box model or by using data-driven elements inside theoretical models, whereas our focus is specifically on analyzing the behavior of the base model itself.

4. Case Study I: Human growth

The general goal of this case study is to illustrate (i) the basic idea as clearly as possible, with an elementary setting, while showing (ii) how the approach can be used to support model comparison and (iii) how an excellent model is identified as one having no or almost no discrepancy.

Context We consider a series of height measures $Y = \{y_i\}_{i \in [1,n]}$, made for humans at corresponding ages $X = \{x_i\}_{i \in [1,n]}$. The task is to model the growth by learning the interpretable model $M(x)$ that provides the height for each age, used already as a motivational example in Figure 1.

Data We use simulated data of 100 observations for six age groups ($n = 600$), generating the data according to statistical information about Finnish children provided by the Finnish Health Institution(data available at http://kasvukayrat.fi/wp-content/uploads/2018/08/Pojat-0-2v1.pdf and http://kasvukayrat.fi/wp-content/uploads/2018/08/Pojat-1-20v1.pdf). The simulation process is described in the Appendix.

Base models We consider three different models to demonstrate different types of discrepancies and their interpretations. Two are generic statistical models (linear and logistic regression), and one is a classical theory-based growth model by Jolicoeur et al. (1988).

The linear regression model is given by $M_{\text{LINF}}(x) = \beta_0 + \beta_1 x$, where $\beta_0 \in \mathbb{R}$ is the intercept, and $\beta_1 \in \mathbb{R}$ is the slope ($\boldsymbol{\theta}_{\text{LINF}} = \{\beta_0, \beta_1\}$). The logistic model (LOGF) is given by $M_{\text{LOGF}}(x) = \frac{a}{1+\exp(-k(x-x_0))}$, where $x_0 \in \mathbb{R}^+$ is the newborn age, $a \in \mathbb{R}^+$ is the height at the maturity, and $k \in \mathbb{R}^+$ is the logistic growth rate ($\boldsymbol{\theta}_{\text{LOGF}} = \{x_0, a, k\}$). Finally, the Jolicoeur et al. (1988)'s model (JPPS) is given by

$$M_{\text{JPPS}}(x) = a\left(1 - \frac{1}{1 + \sum_{i=1}^{3}(x'/b_i)^{c_i}}\right),$$

where $x' = x + 0.75$ (since the model JPPS covers the height growth from conception, we use an adjustment term assuming a constant gestation of nine months); $a \in \mathbb{R}^+$ is the height at the maturity; b_1, b_2, b_3, c_1, c_2, $c_3 \in \mathbb{R}^+$ are scale parameters ($\boldsymbol{\theta}_{\text{JPPS}} = \{a, b_1, b_2, b_3, c_1, c_2, c_3\}$).

For each model M, we obtained point estimates for its parameters θ_M using the L-BFGS-B algorithm.

Learning the correction We demonstrate a simple use case of an additive correction and hence use the identity function $h(z) = z$. We use the square exponential kernel $K(x_i, x_j) = \alpha^2 \exp\left(-\frac{(x_i - x_j)^2}{2\rho^2}\right)$, with $\alpha, \rho \in \mathbb{R}^+$ controlling the output scale and smoothness.

We assume a Gaussian likelihood. Hence, the posterior has an analytic solution. The hyperparameters of the kernel $\tau = \{\alpha, \rho\}$ are optimized for maximising the marginal likelihood (which also has analytic form) using Adam optimizer with a learning rate of 0.1 for 1000 epochs.

Results The results are depicted in Fig. 2. The main result is that the overall discrepancy is extremely small for the theory-based model (JPPS; $\delta < 0.001$), compared to the alternatives of linear regression (LINF; $\delta = 7.844$) and logistic regression (LOGF; $\delta = 3.635$). All three models show similar discrepancy uncertainty (LINF: $u_\delta = 25.028$; LOGF: $u_\delta = 24.983$; JPPS: $u_\delta = 23.119$), which stems from the generative process; the expected uncertainty for the observation error used for simulating the data for each of the 5 age groups is $2 \times 1.96 \times \text{SD} = 7.84, 15.68, 27.44, 27.44, 31.36, 23.52$. The visualizations reveal that LINF overestimates the height for young ages and young adults, while it underestimates it for intermediary ages (Fig. 2A). LOGF under- and overestimates the height over the whole domain (Fig. 2B), whereas JPSS shows an expected deviation of zero for all inputs.

Conclusion The case shows how the approach can be used for comparing alternative models. A researcher using the simpler models would learn information about their limitations, whereas one considering all three alternatives would learn that the model of Jolicoeur et al. (1988) is to be preferred and matches the true process extremely well.

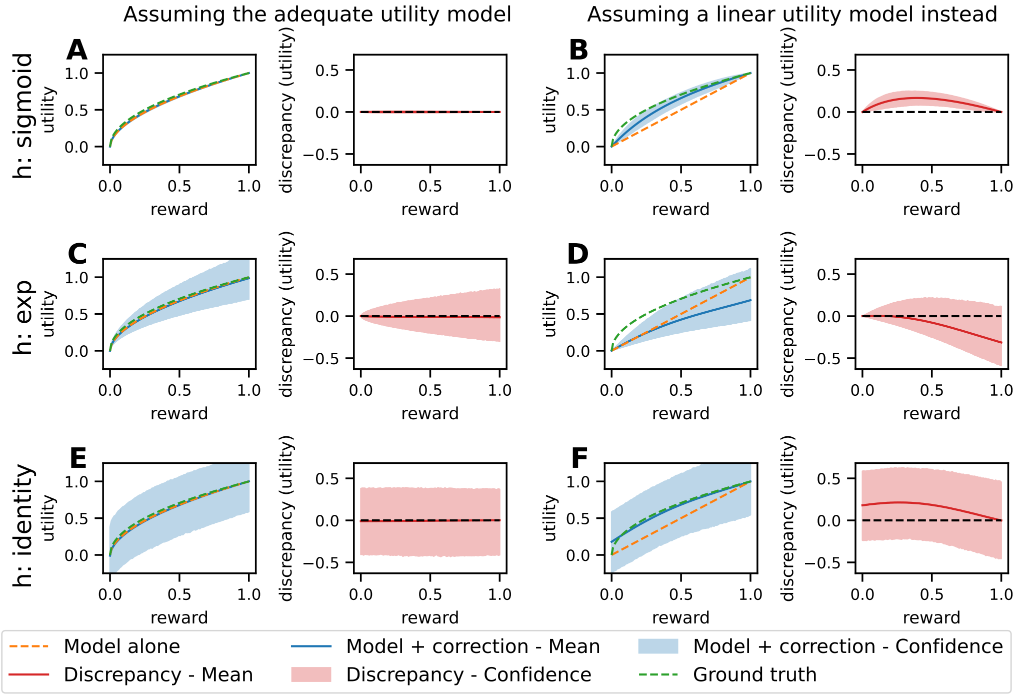

Figure 3: Risk modeling – Artificial data. (A, C, E) Using the correct model as the base model, (B, D, F) using an identity utility (i.e., “wrong” model) instead, and using respectively a sigmoidal, exponential, and identity transformation.

5. Case Study II: Risk Attitude

This case study demonstrates the approach in a prototypical use case in cognitive science and is an example that needs the full flexibility of our framework in terms of the transformations for a meaningful inference of the discrepancy.

Context We consider a series of choices $Y = \{y_i\}_{i\in[1,n]}$ made by a (human) subject. Each choice y_i is between two lotteries L_{i1} and L_{i2}. Each lottery L_{ij} gives the reward x_{ijk} with probability p_{ijk}, with $\sum_k p_{ijk} = 1$. For detailed explanations of the concepts, see e.g. Erev et al. (2017). The underlying decision-making process is latent (i.e. not observable), so we want to model it. As we expect large inter-individual differences, we learn a separate model for each user, and one of the goals is to detect for which users the model can be relied on as a basis for interpreting their risk behavior.

Data We consider two datasets: artificial data simulated from the model for technical validation and the CPC18 data set (`https://zenodo.org/record/845873#.WeDg9GhSw2x`) of real observations of human decision-makers, explained in detail in Erev et al. (2017). We used the data from 125 subjects for which we had $n = 325$ observations. All rewards are normalized between 0 and 1. For artificial data, $n = 325$ observations and features were generated using similar stimuli.

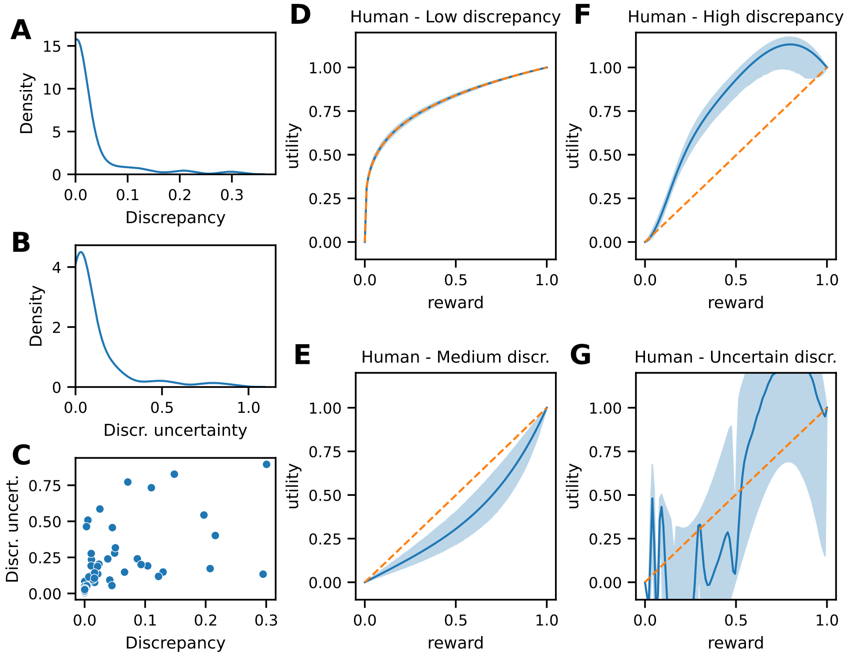

Figure 4: Risk modeling – Human data (CPC dataset). (A) Distribution of the overall discrepancy (δ). (B) Distribution of the overall discrepancy uncertainty (u_δ). (C) Overall discrepancy uncertainty over the overall discrepancy. (D) Subject with a "low" discrepancy ($\delta < 0.001$). (E) A subject with a "moderate" discrepancy ($\delta = 0.130$). (G) Subject with a "strong" discrepancy ($\delta = 0.295$). (G) Subject where no reliable discrepancy has been found ($u_\delta = 0.895$).

Base models We use probabilistic version of Expected Utility (Von Neumann and Morgenstern, 1944) model. The probability of a choice y_i is

$$p(y_i = L_{i1}) = s\left(\beta\left[\sum_k p_{i1k}u(x_{i1k}) - \sum_k p_{i2k}u(x_{i2k})\right]\right),$$

where $s(\cdot)$ is the sigmoid function, $\beta \in \mathbb{R}^+$ the inverse temperature parameter, and $u(\cdot)$ the utility function. A classic formulation of the utility function (Holt and Laury, 2002) is $u(x) = x^\alpha$, with $\alpha \in \mathbb{R}^+$ as the risk aversion parameter. $\alpha < 1$ indicates risk aversion, while $\alpha > 1$ indicates risk-seeking.

Note that for modeling the discrepancy, we consider the utility function itself as a target model for our discrepancy investigation. Therefore, contrary to the human growth modeling experiment, the goal is to measure the discrepancy of a *component* of the full model.

We again obtained a point estimate for each user's parameters $\theta = \{\alpha, \beta\}$ using the L-BFGS-B algorithm.

Learning the correction Similar to Case Study I, we used the squared exponential kernel as a covariance function. We assume here that the user does not know in advance which transformation should be used. Hence, we considered three alternatives: identity (id.), exponential (exp.), and sigmoid (sig.). The exponential transformation constrains the solution to lie in the positive domain, while the sigmoid transformation constrains it to be in the interval $[0, 1]$. We also use here a second-order Taylor series approximation for f.

Here the likelihood is non-Gaussian (a Bernoulli distribution), and hence we use VI with full-rank multivariate Gaussian distribution to approximate the posterior. We optimized the GP hyperparameters (τ: kernel's output scale and length scale) at the same optimization loop as the variational parameters (λ: mean and covariance matrix of the q distribution) using *GPyTorch* (Gardner et al., 2019) with the Adam optimizer, using a learning rate of 0.05 for 1000 epochs, 100 samples for the Monte Carlo estimation of the log probabilities, and 50 inducing points linearly spaced over the interval $[0, 1]$.

Results First, we ran experiments with artificial agents where the ground truth is known to illustrate the approach in a controlled case. When using the correct model (i.e. the one used for generating the data; see Fig 3A 3C 3E), the overall discrepancy (δ) is low for all the transformations, although higher for the identity transformation (sig.: 0.003, exp.: 0.077, id. 0.158). However, the uncertainty estimation (u_δ) is meaningful only when using the sigmoid transformation, as it is overly large otherwise (sig.: 0.017, exp.: 0.374, id.:0.778). For this simulated data, we can also compare the corrected model to the true data generative process by computing an analogous metric as δ but using the true process in place of $M(x)$. For all transformations, the corrected model is close to the true process (sig.: 0.021, exp.: 0.033, id.: 0.013).

For evaluation purposes, we also consider a case where a "wrong" model is used as the base model, i.e., an identity function $u(x) = x$, instead of the one used for generating the data $u(x) = x^\alpha$. In this case, the observed discrepancy depends significantly on the transformation (see Fig 3B 3D 3F), and using the sigmoid transformation is critical for correct analysis. In the other cases, either the mean of the corrected model is relatively far from the ground truth (0.059, 0.272, 0.020 for sig., exp., and id.), or the uncertainty is very large (u_δ of 0.119, 0.436, 0.847 for sig., exp., and id.).

Figure 4 presents the results for the CPC18 dataset, using only the sigmoid that was robust to discrepancy in the case of both correct and incorrect models. From a general perspective, we note that the overall discrepancy is relatively low — as one can expect since we are using a very classic utility function (Holt and Laury, 2002) — and that the uncertainty of the discrepancy is also low (see Fig. 4A–C).

We learn a separate model for each individual and hence can study how well each one follows the assumed model. Fig 4D–G shows the results of four subjects that are prototypical cases: a subject with an overall discrepancy close to 0 (Fig 4D; $\delta < 0.001$), a subject with a "moderate" discrepancy (Fig 4E; $\delta = 0.130$), a subject with a high discrepancy (Fig 4F; $\delta = 0.295$), and a subject with an "uncertain" discrepancy (Fig 4G; $u_\delta = 0.895$). We labeled the latter as "uncertain" as the mean-variance is so high that it does not clearly indicate how the initial model could be wrong (e.g., overestimates utility for lower values), but rather that no model could bring reliable predictions for this subject.

Conclusion The artificial data results show that the choice of the transformation (h) is crucial for a good estimate of the overall discrepancy and its uncertainty. The results in the CPC18 dataset show how a model's discrepancy can be used to distinguish, for each individual separately, whether the utility model's parameters can be safely interpreted to describe the subject's attitude towards risk.

6. Discussion and Conclusion

We proposed to use transformed Gaussian processes for estimating the discrepancy between a base model and the data generative process given some observations, using transformed Gaussian processes to model this discrepancy. Our work extends previous work on modeling discrepancy with Gaussian processes (Brynjarsdóttir and O'Hagan, 2014; Gardner et al., 2021) by offering a family of possible discrepancy models that allows a better balance between flexibility and optimization constraints. We showed how additive and multiplicative discrepancies are specific choices of transformation, how we can encode the assumption of the model being correct in expectation irrespective of the transformation, and how an adequate transformation that constraints the optimization problem can be the key to a meaningful estimation of the discrepancy. We also covered an important application case in human decision-making modeling where combining theory-based and data-driven approaches has recently been shown to bring significant progress (Plonsky et al., 2017; Agrawal et al., 2020; Peterson et al., 2021) but discrepancy models have not been considered before.

Our approach easily integrates into existing modeling workflows, covering all processes where point estimates of θ are used. We provided the fundamental elements required for an interactive tool, but further work remains in developing an easy-to-use interface, particularly from the perspective of efficiently supporting the exploration of discrepancies for models with more than a few parameters. Extending the approach for a scenario where the researcher is conducting Bayesian inference over θ would also be an interesting future direction.

A limitation is that the visualization techniques proposed here will not scale well with a high number of covariates (i.e., high dimensionality). Another limitation is that we have no guarantees of obtaining a reliable model's discrepancy in every context. Our experimental results show that we can, in practice, identify potential failures, using the uncertainty (variance) of the discrepancy as an indicator, but further work would be needed to identify potential theoretical conditions for reliable and meaningful results. Still, our approach allows the inspection of discrepancy from multiple perspectives by using different transformations and can accelerate the development of theory-based models by enhancing their diagnostic.

Acknowledgments

We thank our colleagues and the reviewers for their helpful comments. This work was funded by the Engineering and Physical Sciences Research Council (EPSRC) under grant number BC307829-06 ("Quest"), Aalto University's Department of Communications and Networking (Comnet), the Finnish Center for Artificial Intelligence (FCAI), the Foundation for Aalto University Science and Technology, and the Research Council of Finland (projects 328813, 318559, 348952 and 324852). We also acknowledge the computation resources provided by Aalto University's Science-IT project.

References

Homayun Afrabandpey, Tomi Peltola, Juho Piironen, Aki Vehtari, and Samuel Kaski. A decision-theoretic approach for model interpretability in bayesian framework. *Machine Learning*, 109(9):1855–1876, 2020.

Mayank Agrawal, Joshua C. Peterson, and Thomas L. Griffiths. Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, 117(16): 8825–8835, 2020.

Paul D. Arendt, Daniel W. Apley, and Wei Chen. Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability. *Journal of Mechanical Design*, 134 (10), 09 2012.

K Sham Bhat, David S Mebane, Priyadarshi Mahapatra, and Curtis B Storlie. Upscaling uncertainty with dynamic discrepancy for a multi-scale carbon capture system. *Journal of the American Statistical Association*, 112(520):1453–1467, 2017.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

David D Bourgin, Joshua C Peterson, Daniel Reichman, Stuart J Russell, and Thomas L Griffiths. Cognitive model priors for predicting human decisions. In *International Conference on Machine Learning*, pages 5133–5141, 2019.

Jenný Brynjarsdóttir and Anthony O'Hagan. Learning about physical parameters: the importance of model discrepancy. *Inverse Problems*, 30(11):114007, oct 2014.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017.

Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging.* Cambridge University Press, October 2008.

Andreas Damianou. *Deep Gaussian processes and variational propagation of uncertainty.* PhD thesis, University of Sheffield, 2015.

Ido Erev, Eyal Ert, Ori Plonsky, Doron Cohen, and Oded Cohen. From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, 124(4):369, 2017.

Jacob R. Gardner, Geoff Pleiss, David Bindel, Kilian Q. Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *arXiv preprint*, arXiv:1809.11165, 2019.

P. Gardner, T.J. Rogers, C. Lord, and R.J. Barthorpe. Learning model discrepancy: A gaussian process and sampling-based approach. *Mechanical Systems and Signal Processing*, 152:107381, 2021.

Paul Gardner, Charles Lord, and Robert J Barthorpe. Bayesian history matching for structural dynamics applications. *Mechanical Systems and Signal Processing*, 143:106828, 2020.

Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint*, arXiv:2011.01808, 2020.

Yanyan He and Dongbin Xiu. Numerical strategy for model correction using physical constraints. *Journal of Computational Physics*, 313:617–634, 2016.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360. PMLR, 2015.

Charles A Holt and Susan K Laury. Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655, 2002.

Pierre Jolicoeur, Jacques Pontier, Marie-Odile Pernin, and Michel Sempé. A lifetime asymptotic growth curve for human height. *Biometrics*, pages 995–1003, 1988.

Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

Joshua C. Peterson, David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.

Ori Plonsky, Ido Erev, Tamir Hazan, and Moshe Tennenholtz. Psychological forest: Predicting human behavior. In *AAAI Conference on Artificial Intelligence*, 2017.

Prakash Rajan and Krishna P Miyapuram. Psychfm: Predicting your next gamble. In *International Joint Conference on Neural Networks*, 2020.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT press Cambridge, 2006.

Henry P Stott. Cumulative prospect theory's functional menagerie. *Journal of Risk and uncertainty*, 32(2):101–130, 2006.

Mark Strong, Jeremy E Oakley, and Jim Chilcott. Managing structural uncertainty in health economic decision models: a discrepancy approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(1):25–45, 2012.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1944.

Larry Wasserman. Bayesian model selection and model averaging. *Journal of mathematical psychology*, 44(1):92–107, 2000.