Transformed Gaussian processes for characterizing a model's discrepancy: SUPPLEMENTARY MATERIAL

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

Introduction

This document provides supplementary material complementing the main paper. For the convenience of reading, the numbers for Sections, Figures and Equations are prefixed by the letter S to avoid confusion with the main paper.

Sections S1 and S2 complement Section 2.5 of the main paper by providing the derivations for the proposed prior means for specific choices of the transformation $h(z)$ and for the Taylor series approximation of the generative process, respectively. The sections S3, S4, and S5 provide additional information about the case studies (Sections 4 and 5 of the main paper), including also comprehensive visualizations of the discrepancies for all of the subjects of the CPC18 data set provided in Figure S1. Finally, Section S6 analyses the discrepancy uncertainty for the special case of identity transformation in Case Study II, providing an explanation for the large credible intervals.

S1. Selecting Prior Mean to Achieve $\mathbb{E}[f(x)] = M(x)$

For convenience, we repeat the main defined by Equations (1) and (2),

$$f(x) = h(h^{-1}(M(x)) + r(x)), \tag{S1}$$

$$\text{such that: } \mathbb{E}[f(x)] = M(x), \tag{S2}$$

where $r(x) \sim GP(\mu(x), K)$. As explained in Section 2.5, we can satisfy (S2) by selecting the prior mean $\mu(x)$ suitably but the choice depends on the transformation $h(z)$. We provide the derivations for the three examples considered in this paper below, but similar results could easily be derived also for other transformations.

© 2023 A. Nioche, V. Tanskanen, M. Hartmann & A. Klami.

Identity: For the identity function $h(z) = z$ it is sufficient to use $\mu(x) = 0$ since

$$\begin{aligned}\mathbb{E}[f(x)] &= \mathbb{E}[M(x) + r(x)] \\ &= M(x) + \underbrace{\mathbb{E}[r(x)]}_{=0}.\end{aligned}$$

Exponential: For the multiplicative correction, we use $h(z) = \exp(z)$ and have

$$\begin{aligned}\mathbb{E}[f(x)] &= \mathbb{E}[M(x)\exp(r(x))] \\ &= M(x)\,\mathbb{E}[\exp(r(x))] \\ &\neq M(x)(, \text{ necessarily}).\end{aligned}$$

As $r(x) \sim N(\mu(x), \sigma^2)$ for each x, we know that $\exp(r(x)) \sim \text{Log-Normal}(\mu(x), \sigma^2)$, which has the expectation $\mathbb{E}[\exp(r(x))] = \exp\left(\mu(x) + \frac{\sigma^2}{2}\right)$. To obey Condition (S2), we hence need

$$\begin{aligned}&\mathbb{E}[\exp(r(x))] = 1 \\ \iff\ &\exp(\mu(x) + \frac{\sigma^2}{2}) = 1 \\ \iff\ &\mu(x) = -\frac{\sigma^2}{2}.\end{aligned}$$

Sigmoid: For the sigmoid function $h(z) = \mathbf{s}(z)$, the equation

$$\mathbb{E}[f(x)] = \mathbb{E}[s(s^{-1}(M(x)) + r(x))] = M(x) \tag{S3}$$

has no analytic solution. However, we can recognize that the sigmoid function can be approximated by the inverse probit function, and arrive at an approximation following Demidenko (2004), pp 336–337:

$$\begin{aligned}\mathbb{E}[s(\overbrace{s^{-1}(M(x)) + r(x)}^{Y\sim N(\mu_Y,\sigma^2)})] &= \int s(y)p(y|\mu_Y,\sigma^2)dy \\ &\approx \int \Phi(\lambda y)p(y|\mu_Y,\sigma^2)dy \\ &= \int P(Z < \lambda y)p(y|\mu_Y,\sigma^2)dy \\ &= \int P(Z/\lambda < Y|Y = y)p(y|\mu_Y,\sigma^2)dy \\ &= P(Z/\lambda < Y) \\ &= P(\overbrace{Z/\lambda - Y < 0}^{\sim N(-\mu_Y,\sigma^2+\lambda^{-2})}) \\ &= \Phi\left(\frac{\mu_Y}{\sqrt{\lambda^{-2}+\sigma^2}}\right) \\ &= \Phi\left(\frac{s^{-1}(M(x)) + \mu(x)}{\sqrt{0.588^{-2}+\sigma^2}}\right),\end{aligned}$$

where σ^2 is the prior variance of the Gaussian process and $\lambda \approx 0.588$ can be achieved by any optimization method. To respect Condition (S2), we solve for $\mu(x)$

$$\Phi\left(\frac{s^{-1}(M(x)) + \mu(x)}{\sqrt{0.588^{-2} + \sigma^2}}\right) = M(x)$$
$$\iff \mu(x) = \Phi^{-1}(M(x))\sqrt{0.588^{-2} + \sigma^2} - s^{-1}(M(x)).$$

Finally, we remark that, instead of $s(z)$, we could also use the inverse probit function as transformation. We would then have an exact analytical expression for the mean.

S2. Adjusting the process to respect $\mathbb{E}[f(x)] = M(x)$

Instead of the case-specific derivations for the required mean, we can alternatively satisfy Condition (S2) by slightly changing the assumed generative model, as explained in Section 2.5. This has the advantage of providing a general solution for sufficiently smooth $h(z)$, namely for all $h(z)$ that are twice differentiable.

We approximate the process using second-order Taylor series as

$$\begin{aligned}
\mathbb{E}[f(x)] &= \mathbb{E}[h(\overbrace{h^{-1}(M(x)) + r(x)}^{:=\omega \sim N(\mu_\omega, \sigma^2)})] \\
&\approx \mathbb{E}[h(\mu_\omega) + (\omega - \mu_\omega)^T h'(\mu_\omega) + \frac{1}{2}(\omega - \mu_\omega)^2 h''(\mu_\omega)] \\
&= h(\mu_\omega) + \frac{1}{2}h''(\mu_\omega)\sigma^2, \qquad \text{(S4)}
\end{aligned}$$

where $\mu_\omega = h^{-1}(M(x)) + \mu(x)$. If we now redefine Equation (S1) using $\tilde{f}(x) := h(h^{-1}(M(x)) + r(x)) - \frac{\sigma^2}{2}h''(\mu_\omega)$, it is easy to see that

$$\begin{aligned}
\mathbb{E}\left[\tilde{f}(x)\right] &= \mathbb{E}[h(h^{-1}(M(x)) + r(x)) - \frac{\sigma^2}{2}h''(\mu_\omega)] \\
&= \mathbb{E}[h(h^{-1}(M(x)) + r(x))] - \frac{\sigma^2}{2}h''(\mu_\omega) \\
&\approx h(\mu_\omega) + \frac{\sigma^2}{2}h''(\mu_\omega) - \frac{\sigma^2}{2}h''(\mu_\omega) \\
&= h(h^{-1}(M(x)) + \mu(x)),
\end{aligned}$$

for all x. If we now assume the prior mean $\mathbb{E}[r(x)] = \mu(x) = 0$, we obtain $h(h^{-1}(M(x)) = M(x)$ and hence satisfy Condition (S2).

S3. Case Study I: Data simulation

Even though growth models could be trained on real observations of individuals, we decided to use simulated data to avoid difficulties in interpretation caused by potential deviations from the assumed likelihood. Consequently, we estimated the model from artificial data. We simulated heights for hypothetical individuals from a normal distribution with mean

Table S1: Statistics used to generate the data for Case Study I.

Age	μ	σ
0	50.0	2.0
5	112.0	4.0
10	141.0	7.0
12	152.0	7.0
15	172.0	8.0
20	180.0	6.00

and standard deviation matching the observed statistics of Finnish children provided by the Finnish Health Institution[1]. We selected 6 different age groups and simulated 100 individuals for each, corresponding to a simulated data of 600 individuals in total. The ages and the corresponding statistics are provided in Table S1. From the perspective of this experiment, the data is equally valid as a real sample would be.

S4. Case Study II: CPC18 dataset

The human data for Case Study II is a subset of the CPC18 dataset (https://zenodo.org/record/845873#.WeDg9GhSw2x). The CPC18 dataset consists of 510,750 entries, where each entry is a choice made by one of the 686 subjects between two risky and/or ambiguous options. The task follows the same paradigm as the CPC15 dataset and is explained in detail by Erev et al. (2017) (Section *Space of Choice Problems*).

For the purpose of this current work, we selected entries such that the following criteria are met:

- The choice options are without ambiguity (i.e., probabilities and rewards are known to the subject). In the original CPC18 data file, it corresponds to selecting the entries where the flag *Amb* is set to 0.
- The lotteries of both choice options are simple lotteries such that each lottery has the form "gives x with probability p and x' with probability $1-p$" (i.e., we excluded the compounded lotteries). In the CPC18 original data file, it corresponds to selecting entries where the flags *LotNumA* and *LotNumB* are set to 1.
- Lotteries' rewards are positive (0 included). In the CPC18 original data file, it corresponds to selecting entries where *Ha*, *Hb*, *La*, *Lb* ≥ 0.
- Once every criterion above is satisfied, we only keep the entries such that the subject to whom this entry corresponds also has the maximum possible number of entries in total, that is 325. In other words, we discard subjects for which we have fewer observations.

In total, 40,625 entries met these criteria and were used in Case Study II. This corresponds to studying 125 subjects for which we have always observed the decisions for $n = 325$ choices for simple lotteries. We train a separate model for each subject.

1. 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

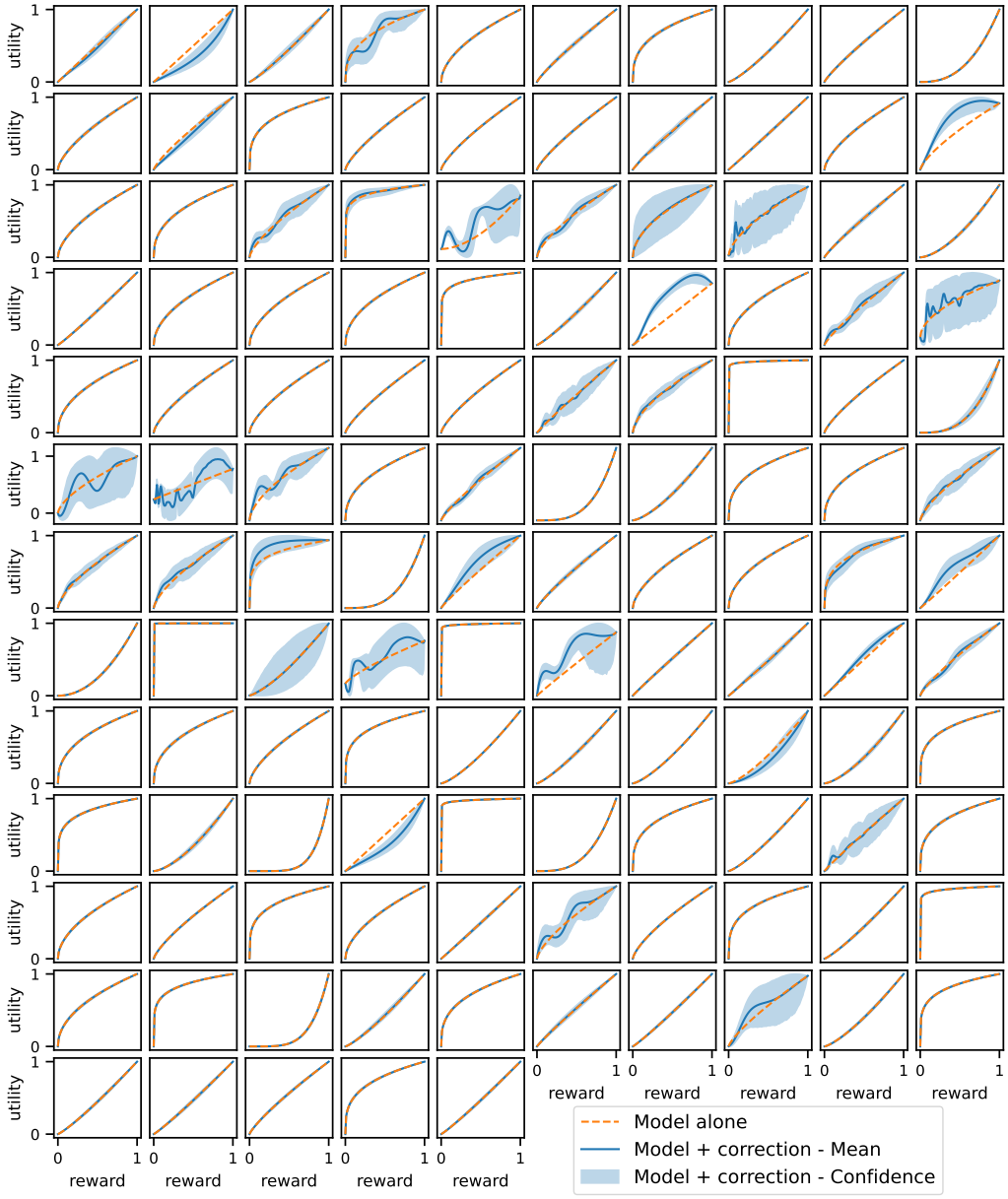

Figure S1: Individual results for the Case Study II.

Figure 4 in the main paper presented the results for four example subjects. For comprehensiveness, Figure S1 presents the results for all 125 subjects. This kind of illustration of the discrepancies could be used to quickly screen subjects that may not follow the assumed model.

S5. Case Study II: Artificial data

The artificial data used for studying the technical validity of the approach was simulated to mimic the real CPC18 data. That is, the task was structurally identical to the subset of the CPC18 dataset used here: each choice was among a set of two lotteries; each lottery gives x with probability p and x' with probability $1-p$. For each trial, the p-values and x-values for both lotteries were drawn from a uniform distribution between 0 and 1. The trials where both the amount and the probability of the greatest amount of one lottery was greater than the amount and the probability of the greatest amount of the other lottery were discarded, and the procedure was repeated (i.e., we excluded the trials with first-order stochastic dominance).

We generated $n = 325$ choices using parameters $\theta = 0.5$ and $\beta = 100$, corresponding to an example subject that is risk-averse.

S6. Large uncertainty in the risk model's discrepancy when using h : linear

As shown in Figure 3 and discussed in Section 5, the uncertainty of the discrepancy for Case II is reasonable when using the sigmoid transformation but not credible in the other cases. In particular, the uncertainty remains large for the identity transformation, even when expecting it to be small as the true generating process is given as prior mean. Here, we explain how this is a property of the underlying model, not our approach. The reason is that the likelihood model is non-identifiable with respect to additive constants. The log-likelihood is given by

$$\log P(Y = y|L_1, L_2) = \mathbb{1}\{y = 1\} \log p + \mathbb{1}\{y = 0\} \log(1 - p),$$

where $p := s(EU(L_1) - EU(L_2)) = \frac{1}{1+\exp(-EU(L_1)+EU(L_2))}$. The following shows how addition of a constant c to the utility (which is what we model the discrepancy for) does not change the likelihood

$$\begin{aligned}
&s\left(\sum_{x\in L_1} p_x(U(x,\theta)+c) - \sum_{x'\in L_2} p_{x'}(U(x',\theta)+c)\right)\\
&= s\left(\sum_{x\in L_1} p_x U(x,\theta) + c\underbrace{\sum_{x\in L_1} p_x}_{=1} - \sum_{x'\in L_2} p_{x'}U(x',\theta) - c\underbrace{\sum_{x'\in L_2} p_{x'}}_{=1}\right)\\
&= s\left(\sum_{x\in L_1} p_x U(x,\theta) - \sum_{x'\in L_2} p_{x'}U(x',\theta)\right)\\
&= s(EU(L_1) - EU(L_2)),
\end{aligned}$$

where p_x is the probability of an outcome in a lottery (Lottery L is a discrete random variable, where $P(L = x) = p_x$). A possible remedy is to fix a point e.g. $f(0, \theta) = 0$. This

can be done by not placing a uniform GP as a prior, but a conditioned one: $f(x)|\{f(0) = 0\} \sim GP(M(x),\, K_{xx} - K_{x0}^T K_{00}^{-1} K_{x0})$, where K_{xx} is the covariance matrix for x and K_{x0} is the covariance between x and $[0]$ and $K_{00} = K([0], [0])$.

References

E Demidenko. *Mixed Models Theory and Application.* John Wiley & Sons, 2004.

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.