SUPPLEMENTARY MATERIALS

Marcelo Hartmann[1,*] **Georgi Agiashvili**[1] **Paul Bürkner**[2] **Arto Klami**[1]
[1] Department of Computer Science, University of Helsinki
[2] Department of Computer Science, Aalto University
∗ marcelo.hartmann@helsinki.fi

Prior predictive probability

In this section we highlight the steps to obtain the prior predictive probability, by rewriting it as a expected value w.r.t. the prior distribution as follows. Given that the probabilistic models, $\pi_{\boldsymbol{Y}\,|\,\boldsymbol{\theta}}(\mathbf{y}\,|\,\boldsymbol{\theta})$ and the prior $\pi_{\boldsymbol{\theta}}$ are positive functions, we can rearrange the order of the integration (See Folland, 2013, Fubini-Tonelli theorem). Hence we have

$$\begin{aligned}\mathbb{P}_{A|\,\boldsymbol{\lambda}} &:= \int_A \pi_{\boldsymbol{Y}}(\mathbf{y}\,|\,\boldsymbol{\lambda})\,\mathrm{d}\mathbf{y}\\ &= \int_A \int_\Theta \pi_{\boldsymbol{Y}\,|\,\boldsymbol{\theta}}(\mathbf{y}\,|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta}\,|\,\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\mathbf{y}\\ &\overset{\text{Fubini}}{=} \int_\Theta \int_A \pi_{\boldsymbol{Y}\,|\,\boldsymbol{\theta}}(\mathbf{y}\,|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta}\,|\,\boldsymbol{\lambda})\,\mathrm{d}\mathbf{y}\,\mathrm{d}\boldsymbol{\theta}\\ &= \int_\Theta \mathbb{P}_{\boldsymbol{Y}\,|\,\boldsymbol{\theta}}(\boldsymbol{Y}\in A|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta}\,|\,\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\theta}\\ &= \mathbb{E}_{\boldsymbol{\theta}}\left(\mathbb{P}_{\boldsymbol{Y}\,|\,\boldsymbol{\theta}}(\boldsymbol{Y}\in A|\,\boldsymbol{\theta})\right). \end{aligned} \tag{1}$$

Approximate role of the precision measure

Here we show the approximate behaviour of the precision parameter α for the general case when covariates are present. The simplification to other cases in straightforward. Recall the likelihood function of $\boldsymbol{\lambda}$ given expert data reads,

$$\mathcal{D}(\mathbf{p}_1,\ldots,\mathbf{p}_J\,|\alpha,\boldsymbol{\lambda}) = \frac{\Gamma(\alpha)^J}{\prod_{j=1}^{J}\prod_{i_j=1}^{n_j}\Gamma(\alpha\,\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda}})}\times \prod_{j=1}^{J}\prod_{i_j=1}^{n_j} p_{j,i_j}^{\alpha\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda}}-1}. \tag{2}$$

Consider the Stirling's approximation[1] to the $\Gamma(\cdot)$ function given by,

$$\Gamma(x) \approx \sqrt{\frac{2\pi}{x}}\left(\frac{x}{e}\right)^x. \tag{3}$$

Rewriting the likelihood function in terms of the above approximation and removing terms that does not depend on α with a simplified notation we get,

$$\begin{aligned}\mathcal{D}(\mathbf{p}\,|\alpha,\boldsymbol{\lambda}) &\approx \frac{\left(\sqrt{\frac{2\pi}{\alpha}}\left(\frac{\alpha}{e}\right)^\alpha\right)^J}{\prod_{j,i_j}\sqrt{\frac{2\pi}{\alpha\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda}}}}\left(\frac{\alpha\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda}}}{e}\right)^{\alpha\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda}}}}\\ &\quad\times\exp\left(\sum_{i,j}\alpha(\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda}}-1)\log p_{j,i_j}\right)\\ &\approx \frac{\alpha^{\sum_j n_j/2-J/2}\prod_{i,j}\mathbb{P}^{1/2}_{A_{j,i_j}|\,\boldsymbol{\lambda}}}{\exp\left(\alpha\sum_{i,j}\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda}}\log\frac{\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda}}}{p_{i,i_j}}\right)}\end{aligned} \tag{4}$$

Take the logarithm of the above function and the derivative w.r.t. α. Setting it to zero and solving for α we obtain,

$$\hat{\alpha} \approx \frac{\sum_j n_j/2 - J/2}{\sum_j KL(\mathbb{P}_j\,||\,\mathbf{p}_j)} \tag{5}$$

where the notation $\mathbb{P}_j = [\mathbb{P}_{A_{j,1}|\,\boldsymbol{\lambda}}\cdots\mathbb{P}_{A_{j,n_j}|\,\boldsymbol{\lambda}}]^\top$ and $KL(P||Q)$ denotes the Kullback-Leibler divergence in this order.

[1] This is a precise approximation.

Hyperparameters' Fisher information matrix

The Fisher information matrix for the unknown hyperparameters can be obtained in closed-form by the fact that, in the original parametrisation of the Dirichlet distribution, the Fisher information is already known. In the original parametrisation and in its basic form, the probability density function reads

$$\mathcal{D}(\mathbf{p}\,|\alpha,\mathbb{P}) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{n}\Gamma(\alpha\,\mathbb{P}_i)}\prod_{i=1}^{n} p_i^{\alpha\mathbb{P}_i-1} \tag{6}$$

where $\mathbb{P} = [\mathbb{P}_1 \cdots \mathbb{P}_n]^\top$. Also knowing that the Dirichlet distribution belongs the exponential family, the Fisher information matrix reads,

$$H_{\mathbb{P}} = \alpha^2(\operatorname{diag}(\psi'(\alpha\mathbb{P})) - \psi'(\alpha)\mathbb{1}\mathbb{1}^\top), \tag{7}$$

whose inverse is given in closed-form as

$$\begin{aligned} H_{\mathbb{P}}^{-1} =& \tfrac{1}{\alpha^2}\Bigg(\operatorname{diag}(\psi'(\alpha\mathbb{P}))^{-1} \\ &+ \frac{\operatorname{diag}(\psi'(\alpha\mathbb{P}))^{-1}\mathbb{1}\mathbb{1}^\top\operatorname{diag}(\psi'(\alpha\mathbb{P}))^{-1}}{(1/\psi'(\alpha) - \mathbb{1}^\top\operatorname{diag}(\psi'(\alpha\mathbb{P}))^{-1}\mathbb{1})}\Bigg) \end{aligned} \tag{8}$$

where $\mathbb{1}$ is $n \times 1$ vector with each component equals to unity.

In the main paper, the vector of parameters $\mathbb{P}$ of the Dirichlet distribution is written as a function of $\boldsymbol{\lambda}$. Using the change of variables for a new parametrisation (see Calderhead, 2012; Girolami and Calderhead, 2011, page 64, Section 3.2.5, equation 3.27), the Fisher information matrix with respect to $\boldsymbol{\lambda}$ can be obtained directly (by passing any need of recalculating integrals) as,

$$H_{\boldsymbol{\lambda}} = \left[\tfrac{d}{d\lambda_1}\mathbb{P}\cdots\tfrac{d}{d\lambda_M}\mathbb{P}\right]^\top H_{\mathbb{P}}\left[\tfrac{d}{d\lambda_1}\mathbb{P}\cdots\tfrac{d}{d\lambda_M}\mathbb{P}\right] \tag{9}$$

where the vector $\frac{d}{d\lambda_m}\mathbb{P} = \left[\frac{d}{d\lambda_m}\mathbb{P}_1\cdots\frac{d}{d\lambda_m}\mathbb{P}_n\right]^\top$ (the Jacobian matrix). Note that $H_{\mathbb{P}}$ is invertible and positive-definide, so as $H_{\boldsymbol{\lambda}}$. Hence $H_{\boldsymbol{\lambda}}$ is also invertible and its cholesky decomposition is stable to compute.

Presence of covariates (inputs): When set of covariates are present, we have to consider that different partitions are provided. Since the likelihood function will still factorise for distinct covariates, note equation (2), the resulting Fisher information matrix will be the sum of Fisher information matrices (Casella and Berger, 2001). Hence, we can write,

$$H_{\boldsymbol{\lambda}} = \sum_j \left[\tfrac{d}{d\lambda_1}\mathbb{P}_j\cdots\tfrac{d}{d\lambda_M}\mathbb{P}_j\right]^\top H_{\mathbb{P}_j}\left[\tfrac{d}{d\lambda_1}\mathbb{P}_j\cdots\tfrac{d}{d\lambda_M}\mathbb{P}_j\right] \tag{10}$$

Non-closed form prior predictive probabilities and hierachical structures

For the case where $\mathbb{P}_j$ does not have closed-form expression we can estimate $\mathbb{P}_j$ and its derivatives w.r.t $\boldsymbol{\lambda}$ using the *reparametrisation gradients* and automatic differentiation. The main idea is to find a pivotal function (see Casella and Berger, 2001, page 427, Section 9.2.2) and obtain Monte-Carlo estimates of $\mathbb{P}_j$ and gradients $d/d\lambda_m\mathbb{P}_j$ with low computational cost according to Figurnov et al. (2018) and Mohamed et al. (2019).

With a simplified notation, recall the prior distribution $\pi_{\boldsymbol{\theta}}$ and that the prior predictive probability can be rewritten as a expected value

$$\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \mathbb{E}_{\boldsymbol{\theta}}\left(\mathbb{P}(\boldsymbol{Y}\in A|\,\boldsymbol{\theta})\right) \tag{11}$$

which depends on $\boldsymbol{\lambda}$. Here the expression $\mathbb{P}(\boldsymbol{Y}\in A|\,\boldsymbol{\theta})$ depends only on $\boldsymbol{\theta}$. Then, find a pivotal function $X = T(\boldsymbol{\theta})$ such that the distribution of X does not depend on $\boldsymbol{\lambda}$. We then can rewrite the expectation,

$$\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \mathbb{E}_X\left(\mathbb{P}(\boldsymbol{Y}\in A|T_X^{-1}(\boldsymbol{\lambda}))\right) \tag{12}$$

The gradients can be computed by interchanging the order of integration and derivation,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda_m}\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \mathbb{E}_X\left(\frac{\mathrm{d}}{\mathrm{d}\lambda_m}\mathbb{P}(\boldsymbol{Y}\in A|T_X^{-1}(\boldsymbol{\lambda}))\right). \tag{13}$$

Where $T_X^{-1}(\cdot)$ is a inverse function of T and depends on X and $\boldsymbol{\lambda}$. The important notion here is that there is no need for resampling X since the distribution $\pi_X(\cdot)$ is free of $\boldsymbol{\lambda}$ by definition.

Hierachical structures: Assume a hierarchical probabilistic model defined in form of layers as in the representation $\boldsymbol{Y} \leftarrow \boldsymbol{\theta}_1 \leftarrow \cdots \leftarrow \boldsymbol{\theta}_L \leftarrow \boldsymbol{\lambda}$, where the letter L indicate the number of hierarchical layers. Formally one could write the hierarchical probabilistic model,

$$\begin{aligned} \boldsymbol{Y}\,|\,\boldsymbol{\theta}_1 &\sim \pi(\mathbf{y}\,|\,\boldsymbol{\theta}_1) \\ \boldsymbol{\theta}_1\,|\,\boldsymbol{\theta}_2 &\sim \pi(\boldsymbol{\theta}_1\,|\,\boldsymbol{\theta}_2) \\ &\vdots \\ \boldsymbol{\theta}_L\,|\,\boldsymbol{\lambda} &\sim \pi(\boldsymbol{\theta}_L\,|\,\boldsymbol{\lambda}) \end{aligned} \tag{14}$$

whose prior predictive probability reads,

$$\begin{aligned} \mathbb{P}_{A|\,\boldsymbol{\lambda}} &= \int_\Theta \mathbb{P}(\boldsymbol{Y}\in A|\,\boldsymbol{\theta}_1)\prod_{\ell=1}^{L-1}\pi(\boldsymbol{\theta}_\ell\,|\,\boldsymbol{\theta}_{\ell+1}) \\ &\quad\times\pi(\boldsymbol{\theta}_L\,|\,\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\theta} \\ &= \int_{\Theta_L}\pi(\boldsymbol{\theta}_L\,|\,\boldsymbol{\lambda})\int_{\Theta_{L-1}}\pi(\boldsymbol{\theta}_{L-1}\,|\,\boldsymbol{\theta}_L)\cdots \\ &\quad\int_{\Theta_1}\pi(\boldsymbol{\theta}_1\,|\,\boldsymbol{\theta}_2)\mathbb{P}(\boldsymbol{Y}\in A|\,\boldsymbol{\theta}_1)\mathrm{d}\boldsymbol{\theta}_1\ldots\mathrm{d}\boldsymbol{\theta}_L \end{aligned} \tag{15}$$

where $\Theta = \cup_{\ell=1}^{L} \Theta_j$ and $\boldsymbol{\theta}_\ell \in \Theta_\ell$. Note that the above equation can be rewritten via the *tower property* by applying it sequentially due to the model hierarchy.

$$\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \mathbb{E}_{\boldsymbol{\theta}_L}\left(\mathbb{E}_{\boldsymbol{\theta}_{L-1}} \cdots \left(\mathbb{E}_{\boldsymbol{\theta}_1}\left(\mathbb{P}_{A|\,\boldsymbol{\theta}_1}\right)\right)\right) \tag{16}$$

with shortened notation $\mathbb{P}(\boldsymbol{Y} \in A|\,\boldsymbol{\theta}_1) = \mathbb{P}_{A|\,\boldsymbol{\theta}_1}$.

In this case, to apply the reparametrisation gradients technique, first find a pivotal function $X_\ell = T_\ell(\boldsymbol{\theta}_\ell)$ for each layer ℓ whose inverse function is denoted as $T^{-1}_{X_\ell}(\boldsymbol{\theta}_{\ell+1})$.

Note the fact when we assume a pivotal quantity for every layer ℓ, by definition the distribution of $\pi_{X_\ell}(x_\ell) = \pi_{\boldsymbol{\theta}_\ell\,|\,\boldsymbol{\theta}_{\ell+1}}(T^{-1}_{x_\ell})|\det J(T^{-1}_{x_\ell})|$ does not dependent on $\boldsymbol{\theta}_{\ell+1}$ or $\boldsymbol{\lambda}$. Hence, define the composite of inverse functions for each layer as

$$\boldsymbol{\theta}_\ell = f_\ell(\boldsymbol{\lambda}) = (T^{-1}_{X_\ell} \circ T^{-1}_{X_{\ell+1}} \circ \cdots \circ T^{-1}_{X_L})(\boldsymbol{\lambda})$$

This way, the above expected value as a function of $\boldsymbol{\lambda}$ can be rewritten as,

$$\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \mathbb{E}_{X_L}\left(\mathbb{E}_{X_{L-1}} \cdots \left(\mathbb{E}_{X_1}\left(\mathbb{P}_{A|f_1(\boldsymbol{\lambda})}\right)\right)\right) \tag{17}$$

To estimate $\mathbb{P}_{A|\,\boldsymbol{\lambda}}$ via Monte Carlo first remember that $\boldsymbol{\lambda}$ is fixed. Sample from π_{X_ℓ} for each ℓ and obtain the respectively the value of the function $f_\ell(\boldsymbol{\lambda})$ for each ℓ. Calculate the sample mean of $\mathbb{P}_{A|f_1(\boldsymbol{\lambda})}$.

Gradients of $\mathbb{P}_{A|\,\boldsymbol{\lambda}}$ w.r.t. λ can be obtained similarly, the extra step needed is in the calculation of the following expression,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda_m}\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \mathbb{E}_{\boldsymbol{X}}\left(\frac{\mathrm{d}f_1}{\mathrm{d}\lambda_m}\frac{\mathrm{d}}{\mathrm{d}f_1}\mathbb{P}_{A|f_1(\boldsymbol{\lambda})}\right) \tag{18}$$

where the notation of the expectation $\mathbb{E}_{\boldsymbol{X}}(\cdot)$ is the same as in (17), but shortened. The first derivative on the right-hand side of the equation above then reads,

$$\frac{\mathrm{d}f_1}{\mathrm{d}\lambda_m} = \prod_{r=1}^{L-1}\frac{\mathrm{d}T^{-1}_{X_r}}{\mathrm{d}T^{-1}_{X_{r+1}}}\frac{\mathrm{d}T^{-1}_{X_L}}{\mathrm{d}\lambda_m}. \tag{19}$$

In cases where the derivative of the inverse function $T^{-1}_{X_\ell}$ above cannot be obtained in closed-form we proceed similar as Figurnov et al. (2018) equation (6). Knowing that T_ℓ is one-to-one function, we can write

$$X_\ell = T_\ell(T^{-1}_{X_\ell}(\boldsymbol{\theta}_{\ell+1})) \tag{20}$$

Take *implicit* and *explicit* derivatives (total derivative) with respect to $\boldsymbol{\theta}_{\ell+1}$ to get that

$$\begin{aligned} 0 &= \frac{\mathrm{d}T_\ell}{\mathrm{d}\,\boldsymbol{\theta}_{\ell+1}}\bigg|_{\text{explicit}} + \frac{\mathrm{d}T_\ell}{\mathrm{d}\,\boldsymbol{\theta}_{\ell+1}}\bigg|_{\text{implicit}} \\ &= \frac{\mathrm{d}T_\ell}{\mathrm{d}\,\boldsymbol{\theta}_{\ell+1}} + \frac{\mathrm{d}T_\ell}{\mathrm{d}\,\boldsymbol{\theta}_\ell}\frac{\mathrm{d}\,\boldsymbol{\theta}_\ell}{\mathrm{d}\,\boldsymbol{\theta}_{\ell+1}} \end{aligned} \tag{21}$$

Identifying the notation $\boldsymbol{\theta}_\ell = T^{-1}_{X_\ell}$ for all ℓ and solving for $\dfrac{\mathrm{d}\,\boldsymbol{\theta}_\ell}{\mathrm{d}\,\boldsymbol{\theta}_{\ell+1}}$ yields,

$$\frac{\mathrm{d}T^{-1}_{X_\ell}}{\mathrm{d}T^{-1}_{X_{\ell+1}}} = -\left(\frac{\mathrm{d}T_\ell}{\mathrm{d}T^{-1}_{X_\ell}}\right)^{-1}\frac{\mathrm{d}T_\ell}{\mathrm{d}T^{-1}_{X_{\ell+1}}} \tag{22}$$

We can now plug (22) into (19) to estimate (18) and in turn to have the estimate for hyperparmeters' Fisher information matrix in (9) and (10). Hence, we can proceed with *stochastic natural gradient descent* to estimate hyperparameters $\boldsymbol{\lambda}$ for general types of probabilistic models.

Predictive elicitation in practice: Example

The probabilistic model for observed data (stature of male human being) is specified as follows,

$$\begin{aligned} Y_t|\,\boldsymbol{\theta}, b &\sim \mathcal{W}(h(t;\boldsymbol{\theta}), b) \\ b &\sim \mathcal{G}(a_0, b_0) \\ \theta_d &\overset{i.i.d}{\sim} \mathcal{LN}(a_d, b_d) \end{aligned} \tag{23}$$

where Y_t is univariate $S = 1$ and denotes the stature of the human being at time t. The parameters of the growth-model $h(t;\boldsymbol{\theta})$ are denoted as $\boldsymbol{\theta} = [\theta_1\ \theta_2\ \theta_3\ \theta_4\ \theta_5] = [h_1, h_{t_*}, t_*, s_0, s_1]^\top$, where h_1 is the average height of an adult human, h_{t_*} is the average high for the event "growth-spurt" (Preece and Baines, 1978), t_* is when that event happens, s_0 and s_1 are constants from the model. The parameter b controls the variance of the variable Y_t around $h(t;\boldsymbol{\theta})$. Large the values of b less variance around the $h(t;\boldsymbol{\theta})$ and vice-versa. $\mathcal{W}$, $\mathcal{G}$ and $\mathcal{LN}$ stands for respectively, Weibull, Gamma and log-Normal distributions.

We used the Weibull distribution in the mean-variance parametrisation which means that the probability distribution of $Y_t|\,\boldsymbol{\theta}, b$ is given by,

$$\begin{aligned} \pi_{Y_t|\,\boldsymbol{\theta},b}(y) = b\,\tfrac{\Gamma(1+1/b)}{\exp(h(t;\boldsymbol{\theta}))}\left(y\,\tfrac{\Gamma(1+1/b)}{\exp(h(t;\boldsymbol{\theta}))}\right)^{b-1} \\ \times \exp\left(-\left(y\,\tfrac{\Gamma(1+1/b)}{\exp(h(t;\boldsymbol{\theta}))}\right)^{b}\right) \end{aligned} \tag{24}$$

The other distribution used for the prior are used in their standard parametrisation scale-shape for Gamma and mean-variance for log-Normal distribution. The vector of hyperparameters is $\boldsymbol{\lambda} = \{a_m, b_m, m = 0, \ldots, 5\}$. The human-growth mode was obtained by Preece and Baines (1978) and given in Section 2, Model 1 in their paper. In our notation this growth-model reads

$$h(t;\boldsymbol{\theta}) = h_1 - \frac{2(h_1 - h_{t_*})}{\exp[s_0\,(t - t_*)] + \exp[s_1\,(t - t_*)]}. \tag{25}$$

The only general background information provided to the participants was the following brief description characterizing the overall growth process and providing general numerical values as reminders:

"During the early stages of life the stature of female and male are about the same, but their stature start to clearly to differ during growth and in the later stages of life. In the early stage man and female are born roughly with the same stature, around 45*cm - 55cm. By the time they are born reaching around 2.5 years old, both male and female present the highest growth rate (centimetres pey year). It is the time they grow the fastest. During this period, man has higher growth rate compared to female. For both male and female there is a spurt growth in the pre-adulthood. For man, this phase shows fast growth rate varying in between 13-17 years old and female varying from 11-15. Also, male tend to keep growing with roughly constant rate until the age of 17-18, while female with until the age of 15-16. After this period of life they tend to stablish their statures mostly around* 162 *-* 190*cm and* 155 *-* 178*cm respectively."*

Given the background information we asked each user to provide the distribution for statures of males at given ages $t = \{t_1, t_2, t_3, t_4\} = \{0, 2.5, 10, 17.5\}$ in form of probabilistic assessments. For eliciting the probabilities we asked them to provide the thresholds y_i determining the statures that partition the sample space with the following probabilities

$$\begin{aligned}
\mathbb{P}(Y_t \leq y_1) &= 0.10 \\
\mathbb{P}(Y_t \leq y_2) &= 0.25 \\
\mathbb{P}(Y_t \leq y_3) &= 0.50 \\
\mathbb{P}(Y_t \leq y_4) &= 0.75 \\
\mathbb{P}(Y_t \leq y_5) &= 0.90
\end{aligned} \tag{26}$$

where naturally $y_1 < y_2 < \ldots < y_5$. The data used as each t_j was hence given by

$$\begin{aligned}
\mathbb{P}(Y_{t_j} \in (0, y_1)) &= p_{j,i_j} = 0.10 \\
\mathbb{P}(Y_{t_j} \in (y_1, y_2)) &= p_{j,i_j} = 0.15 \\
\mathbb{P}(Y_{t_j} \in (y_2, y_3)) &= p_{j,i_j} = 0.25 \\
\mathbb{P}(Y_{t_j} \in (y_3, y_4)) &= p_{j,i_j} = 0.25 \\
\mathbb{P}(Y_{t_j} \in (y_4, y_5)) &= p_{j,i_j} = 0.15 \\
\mathbb{P}(Y_{t_j} \in (y_5, \infty)) &= p_{j,i_j} = 0.1
\end{aligned} \tag{27}$$

Results for the prior predictive elicitation

The main manuscript provided the results for one example user. The results for other four users are provided here in Tables 1 to 4.

The general trend of prior predictive elicitation matching better the data-dependent values of Preece and Baines (1978) remains, and for some users the direct parameter elicitation approach resulted in very poor prior (e.g. h_{t_*} for User 3).

Table 1: User 2

		Predictive		**Parametric**	
Parameter	Reference	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$
h_1	174.6	191.74	4.32	172.7	101.6
h_{t_*}	162.9	153.73	1.6	129.1	31.0
s_0	0.1	0.04	< 0.01	0.51	< 0.04
s_1	1.2	2	4.3	0.5	< 0.04
t_*	14.6	15.9	0.7	12.9	0.5
b		61.4	111.4	3.1	2.6
α	–	14.0	–	1.3	–

Table 2: User 3

		Predictive		**Parametric**	
Parameter	Reference	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$
h_1	174.6	177.14	3.68	174.6	146.3
h_{t_*}	163.0	148.8	1.86	78.5	37.2
s_0	0.1	0.07	< 0.001	0.2	0.004
s_1	1.2	4.54	37.83	0.9	0.004
t_*	14.6	11.31	0.21	6.9	2.9
b	–	18.4	12.5	25.8	74.1
α	–	9.5	–	1.5	–

Table 3: User 4

		Predictive		**Parametric**	
Parameter	Reference	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$
h_1	174.6	174.5	< 0.01	50.5	64.5
h_{t_*}	162.9	162.8	0.02	129.1	31.0
s_0	0.1	0.1	< 0.01	5.1	2.7
s_1	1.2	1.6	1.7	5.1	2.7
t_*	14.60	14.7	0.9	12.9	0.6
b	–	14.5	14.3	1	< 0.02
α	–	17.1	–	1.2	–

Table 4: User 5

		Predictive		**Parametric**	
Parameter	Reference	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$
h_1	174.6	174.4	0.91	159.66	155.96
h_{t_*}	162.9	162.6	0.85	121.75	57.27
s_0	0.1	0.1	< 0.01	3.3	3.3
s_1	1.2	3.4	< 0.01	3.3	3.3
t_*	14.6	14.6	0.02	11.7	5.36
b	–	17.8	17.8	9.5	8.3
α	–	7.7	–	1.5	–

References

Calderhead, B. (2012) *Differential geometric MCMC methods and applications*. Ph.D. thesis, University of Glasgow.

Casella, G. and Berger, R. L. (2001) *Statistical Inference*. Duxbury Press, 2 edn.

Figurnov, M., Mohamed, S. and Mnih, A. (2018) Implicit reparameterization gradients. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, 439–450.

Folland, G. (2013) *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley.

Girolami, M. and Calderhead, B. (2011) Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods. *Journal of the Statistical Royal Society B*, **73**, 123–214.

Mohamed, S., Rosca, M., Figurnov, M. and Mnih, A. (2019) Monte Carlo Gradient Estimation in Machine Learning. *arXiv e-prints*.

Preece, M. A. and Baines, M. J. (1978) A new family of mathematical models describing the human growth. *Annals of Human Biology*, **5**, 1–24.