Bachelor Thesis Marginal-Sampling

Michael Fedders

January 21, 2021

Table of Contents

Introduction

Model

We assume that we can describe a (biological) process through a function $x(t, \theta)$ with time t and unknown model parameters θ. Through (un)voluntary limitations the measured data is not x but

$$\bar{y} = c + h\big(x(t, \theta)\big) + \varepsilon$$

where

- $\bar{y}$ is the measured data
- c is an offset parameter
- h is the observation function
- ε is a noise - for now $\epsilon \sim \mathcal{N}(0, 1/\lambda)$

Standard approach

The standard approach is to use a data set D to determine the model parameters θ and the offset c and noise parameter λ with Bayes theorem:

$$p(\theta, c, \lambda \mid D) = \frac{p(D \mid \theta, c, \lambda) \cdot p(\theta, c, \lambda)}{p(D)}.$$

We can then use Markov chain Monte Carlo (MCMC) methods to proportionally sample the posterior distribution with the product of **likelihood** $p(D \mid \theta, c, \lambda)$ and **prior** $p(\theta, c, \lambda)$. We will call this way **FP-approach** from now on

Hierarchical approach

For Maximum Likelihood methods it was shown that it can be faster to first derive the model parameter θ and then in a second step the noise and transformation (e.g. offset, scaling) parameter [Loos, Krause, and Hasenauer 2018] and [Schmiester et al. 2019].

$\Longrightarrow$ We would like to apply the concept to posterior sampling.

Hierarchical approach

For this **MP-approach** we use again Bayes theorem

$$p(\theta \mid D) = \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)}.$$

If we can calculate the marginalized likelihood we can again use MCMC methods to proportionally sample from the marginalized posterior $p(\theta \mid D)$.

Hierarchical approach

We calculate the required marginalized likelihood with

$$p(D \mid \theta) = \iint p(D \mid \theta, c, \lambda) \cdot p(\theta, c, \lambda) \, \mathrm{d}c \, \mathrm{d}\lambda.$$

Hierarchical approach

We calculate the required marginalized likelihood with

$$
\begin{aligned}
p(D \mid \theta) &= \iint p(D \mid \theta, c, \lambda) \cdot p(\theta, c, \lambda) \, \mathrm{d}c \, \mathrm{d}\lambda \\
&= \iint p(D \mid \theta, c, \lambda) \cdot p(c, \lambda) \, \mathrm{d}c \, \mathrm{d}\lambda \cdot p(\theta)
\end{aligned}
$$

MCMC

In our tests we used 2 different algorithms for the sampling, an adaptive Metropolos sampler and a Parallel Tempering Sampler which is a multi chain method. Every chain gets tempered with a temperature $\beta \in (0, 1]$ i.e. instead of $p(D \mid \theta) \cdot p(\theta)$ we sample

$$p(D \mid \theta)^{\beta} \cdot p(\theta).$$

The first chain always has $\beta = 1$ such that it sample the correct distribution. Then the different chains can exchange values - as the chains with small β sample a more uniform distribution we expect more jumps which then lead to better mixing in the first chain.

Gaussian noise

Normal-Gamma prior

The normal-gamma prior depends on 4 shape parameters, $\mu \in \mathbb{R}, \kappa, \alpha, \beta \in \mathbb{R}_+$ and has the following structure:

$$\begin{aligned} p(c, \lambda) &= f(c, \lambda \mid \mu, \kappa, \alpha, \beta) \\ &= \mathcal{N}(c \mid \mu, \hat{\lambda} = \kappa\lambda) \cdot \Gamma(\lambda \mid \alpha, \beta). \end{aligned}$$

For the prior $p(\theta)$ we will always consider a uniform distribution for a reasonable large interval.

Likelihood

We recall that the measurements are defined as $\overline{y} = c + h + \epsilon$. We assume independent noise for different points in time $t_1, \ldots, t_N,\ N \in \mathbb{N}$. Therefore the likelihood is

$$p(D \mid \theta, c, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{N/2} \cdot \exp\left(-\frac{\lambda}{2}\sum_{k=1}^{N}(\overline{y}_k - (c + h_k))^2\right)$$

Marginal Likelihood

The integral which we have to solve is defined as

$$p(D \mid \theta) = \int_{\mathbb{R}\times\mathbb{R}_+} p(D \mid \theta, c, \lambda) p(c, \lambda) \, d(c, \lambda)$$

Marginal Likelihood

The integral which we have to solve is defined as

$$
\begin{aligned}
p(D \mid \theta) &= \int_{\mathbb{R}\times\mathbb{R}_+} p(D \mid \theta, c, \lambda) p(c, \lambda)\, d(c, \lambda) \\
&= \int_{\mathbb{R}_+} \int_{\mathbb{R}} \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left(-\frac{\lambda}{2}\sum_{k=1}^{N}(\overline{y}_k - (c + h_k))^2\right) \\
&\quad \cdot \frac{\beta^\alpha \sqrt{\kappa}}{\Gamma(\alpha)\sqrt{2\pi}} \lambda^{\alpha - 1/2} \exp\left(-\frac{\lambda}{2}\left(\kappa(c-\mu)^2 + 2\beta\right)\right)\, dc\, d\lambda
\end{aligned}
$$

Marginal Likelihood

We recieve the following form:

$$p(D \mid \theta) = \frac{(\beta/C)^{\alpha}}{\Gamma(\alpha)(2\pi C)^{\frac{N}{2}}} \cdot \sqrt{\frac{\kappa}{N+\kappa}} \cdot \Gamma\left(\frac{N}{2} + \alpha\right)$$

with

$$\begin{aligned} C \equiv \frac{1}{2}\left(\left(\sum_{k=1}^{N}(\overline{y_k} - h_k)^2\right) + \kappa\mu^2 + 2\beta\right) \\ - \frac{1}{2(N+\kappa)}\left(\left(\sum_{k=1}^{N}\overline{y_k} - h_k\right) + \kappa\mu\right)^2. \end{aligned}$$

Distribution of c and λ

To sample c and λ in a second step we need to derive their distribution from the integrand of the marginalized likelihood. We have

$$\lambda \propto \mathrm{Gamma}(\alpha' = \alpha + N/2, \beta' = C)$$

and

$$c \propto \mathcal{N}\left(\mu' = \frac{\left(\sum_{k=1}^{N} \overline{y_k} - h_k\right) + \kappa\mu}{N + \kappa}, \hat{\lambda} = \lambda(N + \kappa)\right)$$

mRNA-transfection model

For $\delta, \xi, k_{TL} \in \mathbb{R}$ we consider the mRNA-transfection model which defines x through

$$\frac{d}{dt}X_1 = k_{TL} \cdot X_2 - \xi \cdot X_1 \qquad \text{and} \qquad \frac{d}{dt}X_2 = -\delta \cdot X_2.$$

In this model we will observe the value of X_2.

Solutions

For an initial condition $X_1(t_1) = 0,\ X_2(t_1) = m_1$ we have the analytical solution

$$X_2(t, \theta = \{\delta, k_{TL} \cdot m_1, \xi, t_1\}) = \frac{k_{TL} \cdot m_1}{\xi - \delta}\left(e^{-\xi(t-t_1)} - e^{-\delta(t-t_1)}\right)$$

Especially

$$X_2(t, \delta = \delta_1, \xi = \xi_1) = X_2(t, \delta = \xi_1, \xi = \delta_1)$$

Sampling

As we are observing more variables and want to sample bimodal parameters we increased the aomount of steps per run to 1.000.000 steps. As underlying data we used experimental data and added an offset of 0,2. For the sampling we used a Parallel Tempering algorithm from pyPESTO with 4 chains.

FP-approach

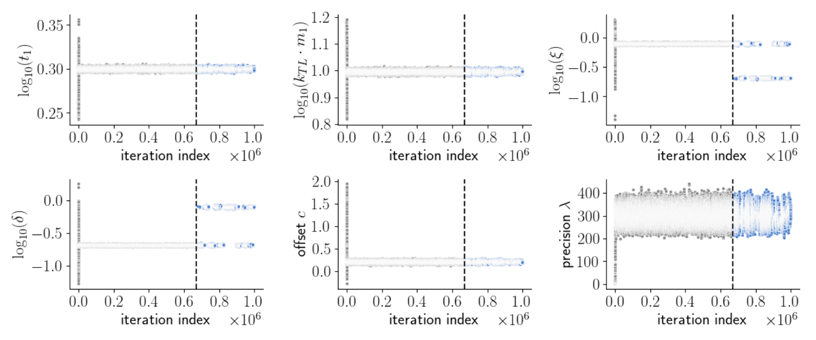

Figure 1: A converging run for the FP-approach.

FP-approach

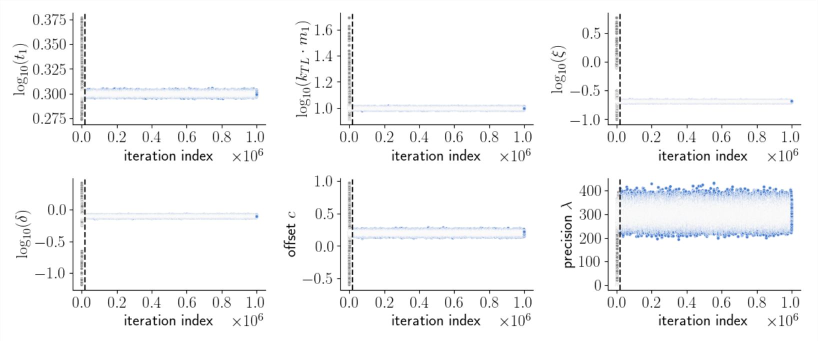

Figure 2: A run which only samples in one mode.

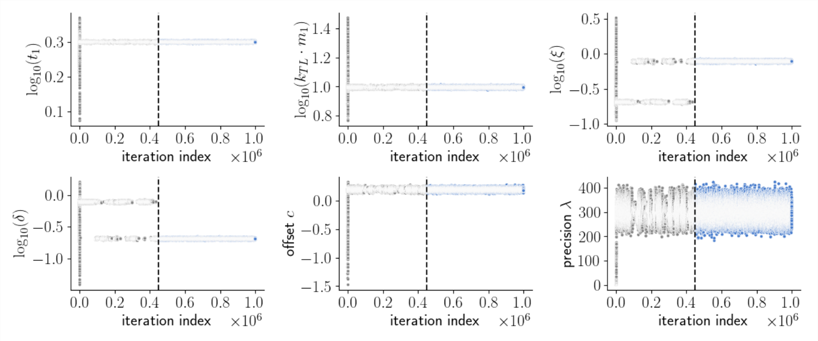

Figure 3: A run which stops to sample in both modes.

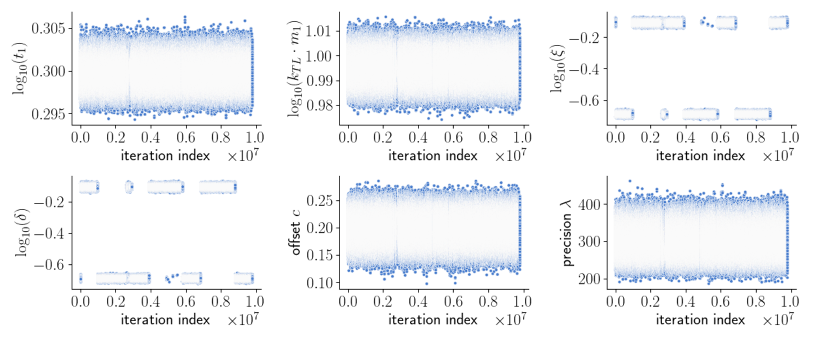

Figure 4: 10 independent runs merged together.

FP-approach

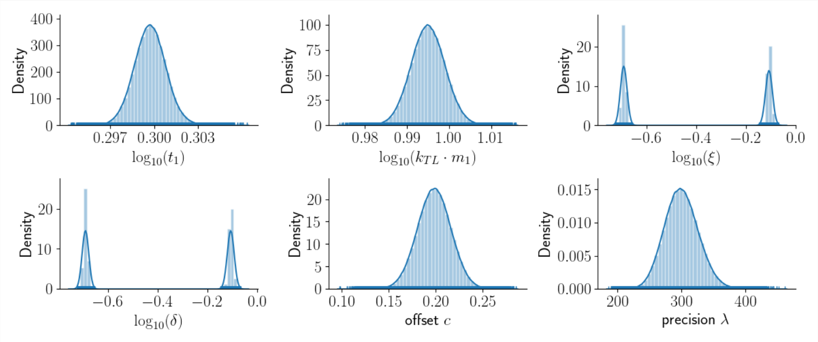

Figure 5: The marginal densities for the 10 runs.

We can use the median for the parameters with one mode and take the highest value separately for each mode for ξ and δ.

FP-approach

We receive

$$\log_{10}(t_1) = 0.2997 \qquad c = 0.9948$$
$$\log_{10}(k_{TL} \cdot m_1) = 0.1986 \qquad \lambda = 299.80$$

and two combiantions for ξ and δ:

$$\log_{10}(\xi) = -0.1076 \text{ and } \log_{10}(\delta) = -0.6909$$

or

$$\log_{10}(\xi) = -0.6909 \text{ and } \log_{10}(\delta) = -0.1077.$$

FP-approach

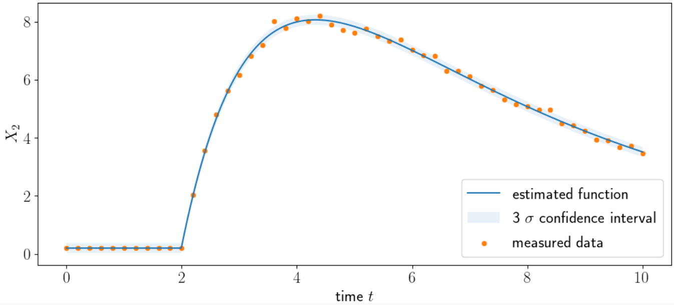

Figure 6: Estimated value for X_2 and the easured data.

MP-approach

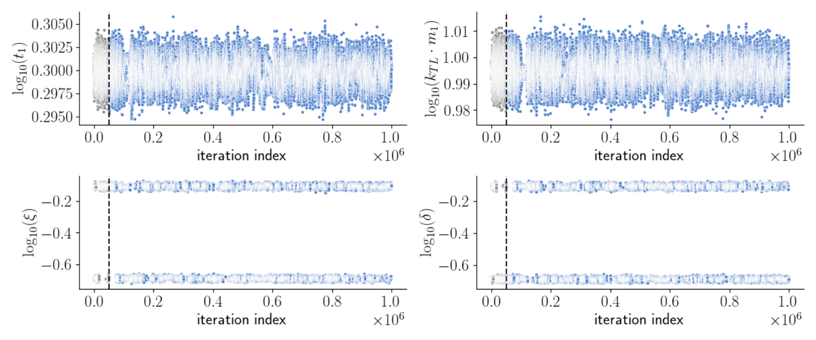

Figure 7: One run with the MP-approach.

Comparison of both approaches

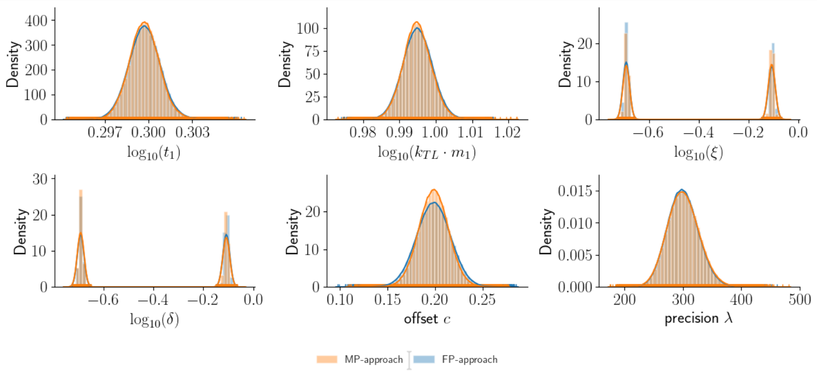

Figure 8: Marginal densities for 10 independent runs each for both approaches.

Comparison of both approaches

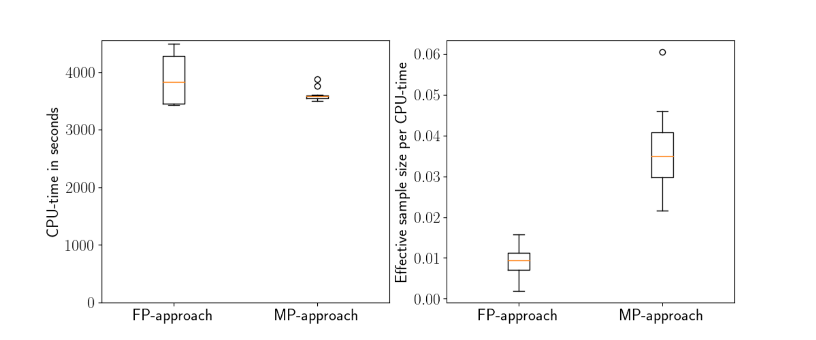

Figure 9: Performance only for θ parameter and **converged** runs.

Comparison of both approaches

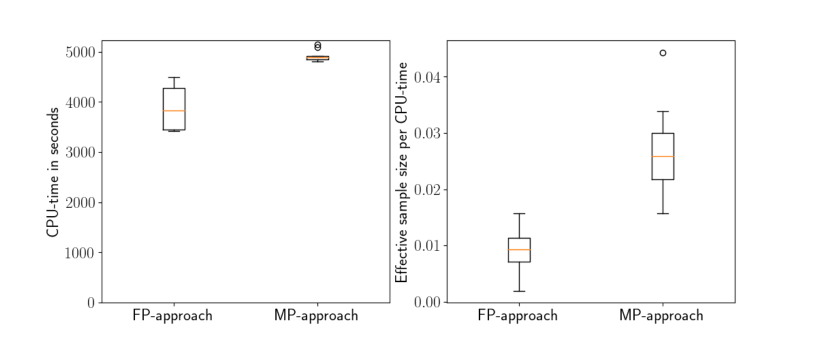

Figure 10: Performance for all parameters with **converged** runs.

Laplacian noise

Likelihood

Now we make the assumption that

$$\epsilon_k \sim \mathit{Laplace}(0, \sigma),\ \sigma \in (0, \infty)$$

i.e. it has a Laplace distribution. The new likelihood has the following form:

$$\begin{aligned} p(D \mid \theta, c, \sigma) &= \prod_{k=1}^{N} \text{Laplace}\left(\overline{y}_k \mid c + h_k, \sigma\right) \\ &= \prod_{k=1}^{N} \frac{1}{2\sigma} \cdot \exp\left\{-\frac{|\overline{y}_k - c - h_k|}{\sigma}\right\} \end{aligned}$$

Marginalised likelihood

The integral we receive is

$$\iint p(D \mid \theta, c, \sigma) p(c) p(\sigma) \, \mathrm{d}c \, \mathrm{d}\sigma$$
$$= \int_0^\infty \int_{-\infty}^\infty \left(\prod_{k=1}^N \frac{1}{2\sigma} \cdot \exp\left\{ -\frac{|c - (\overline{y}_k - h_k)|}{\sigma} \right\} \right) p(c) p(\sigma) \, \mathrm{d}c \, \mathrm{d}\sigma$$

Marginalised likelihood

For $b_0 = -\infty,\ b_i = \overline{y}_i - h_i\ (i = 1, \ldots, N),\ b_{N+1} = \infty$ we can split up the integral in the following parts:

$$\int_0^\infty \frac{p(\sigma)}{2\sigma} \sum_{i=0}^{N} \int_{b_i}^{b_{i+1}} \exp\left\{-\frac{\sum_{k=1}^{N} |c - (\overline{y}_k - h_k)|}{\sigma}\right\} p(c)\, \mathrm{d}c\, \mathrm{d}\sigma.$$

Marginalised likelihood

We finally receive

$$\int_0^\infty \frac{p(\sigma)}{2\sigma}\left(\sum_{i=0}^{N} \exp\left\{\frac{\left(\sum_{k=1}^{i} \overline{y}_k - h_k\right) - \left(\sum_{k=i+1}^{N} \overline{y}_k - h_k\right)}{\sigma}\right\} \cdot \int_{b_i}^{b_{i+1}} e^{c\cdot(N-2i)} p(c)\,\mathrm{d}c\right)\mathrm{d}\sigma$$

Marginalised likelihood

With

$$l_i \equiv \left(\sum_{k=1}^{i} \overline{y}_k - h_k \right) - \left(\sum_{k=i+1}^{N} \overline{y}_k - h_k \right) \quad \text{for } i = 0, \ldots, N.$$

we can write the integral as

$$\sum_{i=0}^{N} \int_0^{\infty} \frac{p(\sigma)}{2\sigma} \exp\left\{ \frac{l_i}{\sigma} \right\} \int_{b_i}^{b_{i+1}} e^{c\cdot(N-2i)} p(c) \, \mathrm{d}c$$

Exponential c prior

We were not aware of any standard choice for the priors so we tried out different possibilities. We will start with an exponential distribution, i.e.

$$p(c) = \lambda e^{-\lambda \cdot c} \text{ with } \lambda > 0.$$

Exponential c prior

We were not aware of any standard choice for the priors so we tried out different possibilities. We will start with an exponential distribution, i.e.

$$p(c) = \lambda e^{-\lambda \cdot c} \text{ with } \lambda > 0.$$

The integral becomes

$$\int_0^\infty \frac{\lambda \cdot p(\sigma)}{2\sigma} \sum_{i=0}^{N} e^{l_i/\sigma} \int_{b_i}^{b_{i+1}} \mathbb{1}_{[0,\infty)}(c) \cdot e^{-c(2i-N+\lambda)} \,\mathrm{d}c \,\mathrm{d}\sigma.$$

Exponential c prior

Let $r = \min\{i = 0, \dots, N \mid b_i \geq 0\}$. We also introduce the notation

$$\begin{aligned} b_{0,\dots,r-1} &\equiv 0 \\ b_r &\equiv \frac{\lambda}{2(N-2r-\lambda)} \cdot \left(e^{b_{r+1}(N-2r-\lambda)} - 1\right) \\ b_{i=r+1,\dots,N-1} &\equiv e^{l_i/\sigma} \frac{\lambda}{2(N-2i-\lambda)} \left(e^{b_{i+1}(N-2(i+1)-\lambda)} - e^{b_i(N-2i-\lambda)}\right) \\ b_N &\equiv e^{l_N/\sigma} \frac{\lambda}{2(N+\lambda)} e^{-b_N(N+\lambda)}. \end{aligned}$$

c prior

We finally have

$$p(D \mid \theta) = \sum_{i=0}^{N} b_i \int_0^\infty \frac{p(\sigma)}{\sigma} \exp\left\{\frac{l_i}{\sigma}\right\} \mathrm{d}\sigma.$$

In general also for a Gaussian or Laplacian c prior we arrive at such a form just with different constants (and possibly different support).

Amoroso σ prior

For $a, b \neq 0, c \in \mathbb{R}$ and $d \in \mathbb{R}_+$ we have

$$\text{Amoroso}(\sigma \mid a, b, c, d) = \frac{1}{\Gamma(d)}\left|\frac{c}{b}\right|\left(\frac{\sigma - a}{b}\right)^{d \cdot c - 1} \exp\left\{-\left(\frac{\sigma - a}{b}\right)^{c}\right\}$$

with $\text{supp}(\sigma) = [a, \infty)$ if $b > 0$ and $\text{supp}(\sigma) = (-\infty, a]$ if $b < 0$.

$\implies$ We will use $d = 1, c = -2, b = 1, a = 0$.

Amoroso σ prior

We need $a = 0$ for the correct support and $c < -1$ so that our integral converges, the other values are chose to simplify the calculation and can be generalized. The prior has the form

$$p(\sigma) = 2 \cdot \sigma^{-3} \exp\left\{ -\frac{1}{\sigma^2} \right\}$$

and therefore the integral has the form

$$\sum_{i=0}^{N} \int_0^{\infty} \frac{2b_i}{\sigma^4} \exp\left\{ \frac{l_i}{\sigma} - \frac{1}{\sigma^2} \right\} \mathrm{d}\sigma.$$

Outlook

Outlook

There are several parts which can be extended.

Outlook

There are several parts which can be extended.

- In the same setting we can test the efficiency for the Gaussian noise in more complex models.

Outlook

There are several parts which can be extended.

- In the same setting we can test the efficiency for the Gaussian noise in more complex models.
- For the Laplacian noise we can finish the derivation and start the tests with models as well.

Outlook

There are several parts which can be extended.

- In the same setting we can test the efficiency for the Gaussian noise in more complex models.
- For the Laplacian noise we can finish the derivation and start the tests with models as well.
- Also we can extend the setting to also include scaling parameters.

End

Thank you for your attention!