# A simulation-based evaluation of a hidden Markov model to characterize disease transitions using frequently sampled spirometry data

Ludvig Jakobsson[1,2,3], Jacob Leander[3], Marcus Baaz[1], Philip Gerlee[2], Mats Jirstrand[1]

(1) Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Gothenburg, Sweden
(2) Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden
(3) Clinical Pharmacology and Quantitative Pharmacology, Clinical Pharmacology and Safety Sciences, R&D, AstraZeneca, Gothenburg, Sweden

## Introduction

**Hidden Markov models** (HMM) have been used to describe discrete switches in disease dynamics (*1*). By explicitly including latent states in the disease model, inference methods can be used to learn otherwise unobserved dynamics.

**In respiratory diseases** such as asthma, sudden worsening events called exacerbations are associated with a distinct drop in lung function. This drop may be modelled as a discrete state switch in an HMM, potentially enabling new ways to make inference about the risk of exacerbations.

This work explores the possibility of modelling home-measured peak expiratory flow as an HMM via simulated data and estimation method evaluation.

## Methods

### Hidden Markov Model

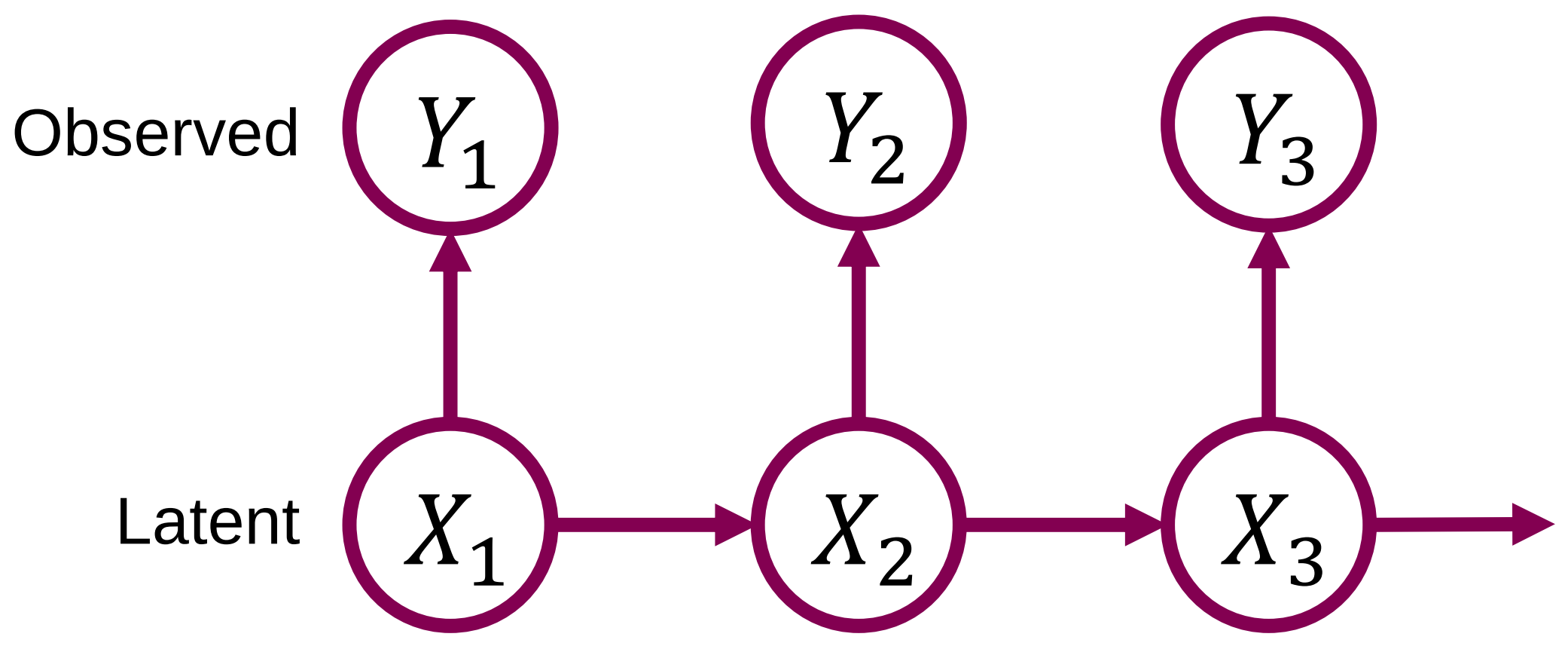
Observations were modelled using a Gaussian process with state dependent mean and variance, representing a baseline and a worsened disease state, respectively.

$$Y_t = \mu_{X_t} + \varepsilon_t,\ \ \varepsilon_t \sim N(0, \sigma^2_{X_t})$$

The latent states were modelled as a discrete-time Markov chain, taking values 0 and 1.

$$P(X_t = j \mid X_{t-1} = k) = p_{kj}$$

**Figure 1:** Schematic view of the HMM used in this work.



### Model Simulation

1000 individual data series were simulated to resemble peak expiratory flow from clinical trials in which home-measured spirometry is carried out. Model parameters for each individual $i$ were drawn from probability distributions based on previous results (*2*).

$$\log(\mu_{0,i}) = \log(\mu_0) + \eta_{1,i},\ \ \mu_{1,i} = d_i \times \mu_{0,i}$$
$$\log(\sigma^2_{0,i}) = \log(\sigma^2_0) + \eta_{2,i}$$
$$\log(\sigma^2_{1,i}) = \log(\sigma^2_1) + \eta_{3,i}$$
$$\operatorname{logit}(d_i) = \operatorname{logit}(d) + \eta_{4,i}$$
$$\operatorname{logit}(p_{01,i}) = \operatorname{logit}(p_{01}) + \eta_{5,i}$$
$$\operatorname{logit}(p_{10,i}) = \operatorname{logit}(p_{10}) + \eta_{6,i}$$

where $\eta_i = (\eta_{1,i}, \eta_{2,i}, \eta_{3,i}, \eta_{4,i}, \eta_{5,i}, \eta_{6,i}) \sim N(0, \Omega)$ with covariance matrix $\Omega$.

### Estimation Methods

The parameters of the HMM were estimated per simulated individual using the **Baum-Welch** algorithm (*3*). This was done for varying time horizons $T$ (ranging from 100 to 1000), simulating short and long clinical trials, respectively.

**Latent state estimation** was evaluated using confusion matrices, showing the accuracy in classifying state transitions. This was done twice, using true parameters and using estimated parameters, respectively.

**Model parameter estimation** was evaluated by how well the histograms of estimated individual parameters resembled the true parameter distributions. Further, the median absolute error of the parameter estimates was calculated for varying time horizons $T$.

## Results

### Latent State Estimation

**Figure 2:** Simulated PEF measurements (points) and latent state estimation (line) given true model parameters.

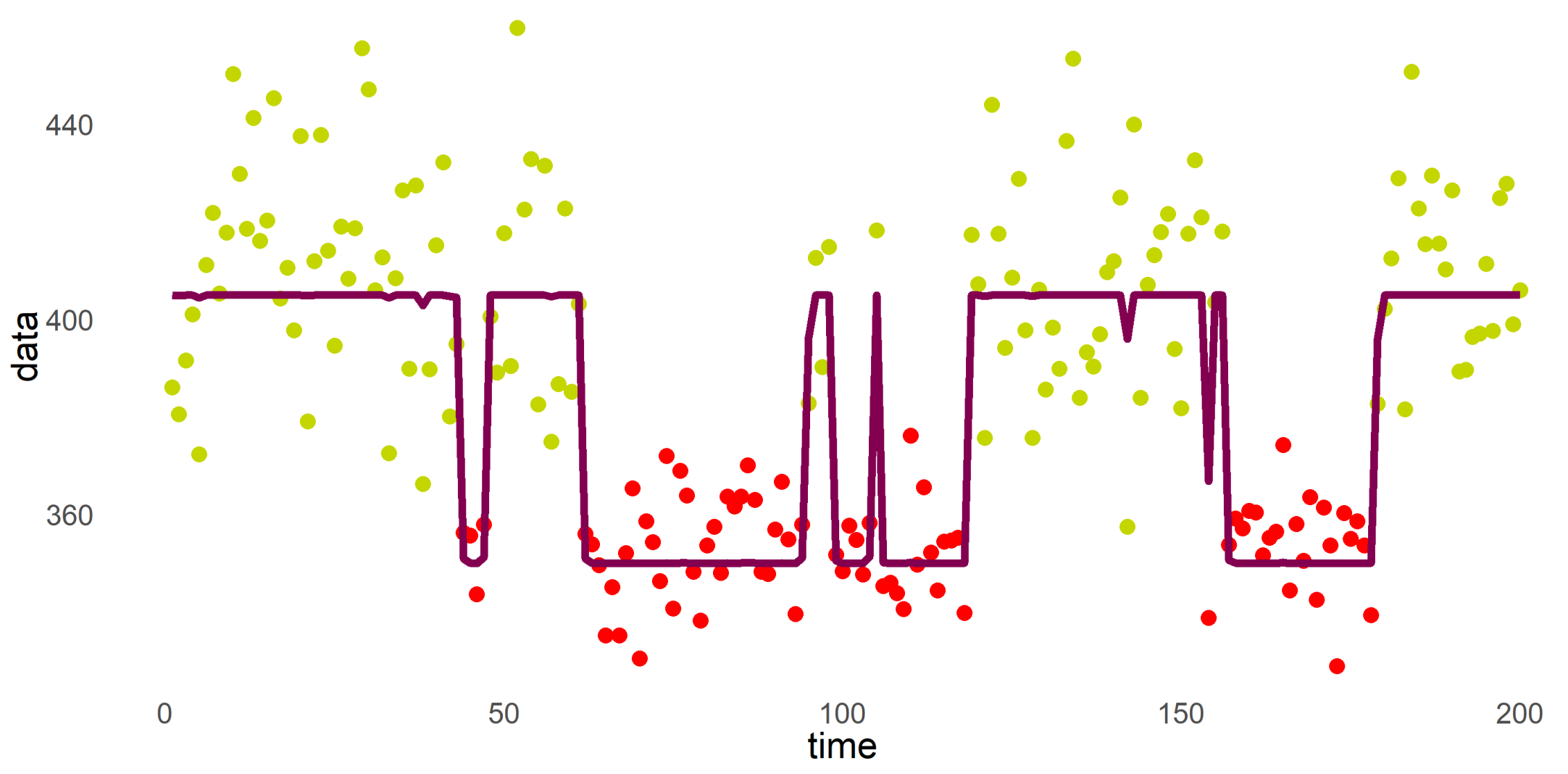


**Table 1:** Accuracy (% of total data points) of estimated latent states given true and estimated model parameters, respectively ($T = 100$).

| | Estimated state 0 | Estimated state 1 |
|---|---|---|
| True state 0 | 0.65<br>0.44 | 0.01<br>0.18 |
| True state 1 | 0.01<br>0.10 | 0.33<br>0.28 |

### Individual Parameter Estimation

Figure 3 shows the distribution of estimated model parameters against true parameter distributions. It is clear that large errors exist for parameters related to measurement noise and transition probability.

Figure 4 shows that median absolute errors of estimated parameters, relative to the true median parameter values, increase substantially as the length of the clinical trial decreases towards $T = 100$.

**Figure 3:** Parameter estimates and true parameter distributions ($T = 100$).

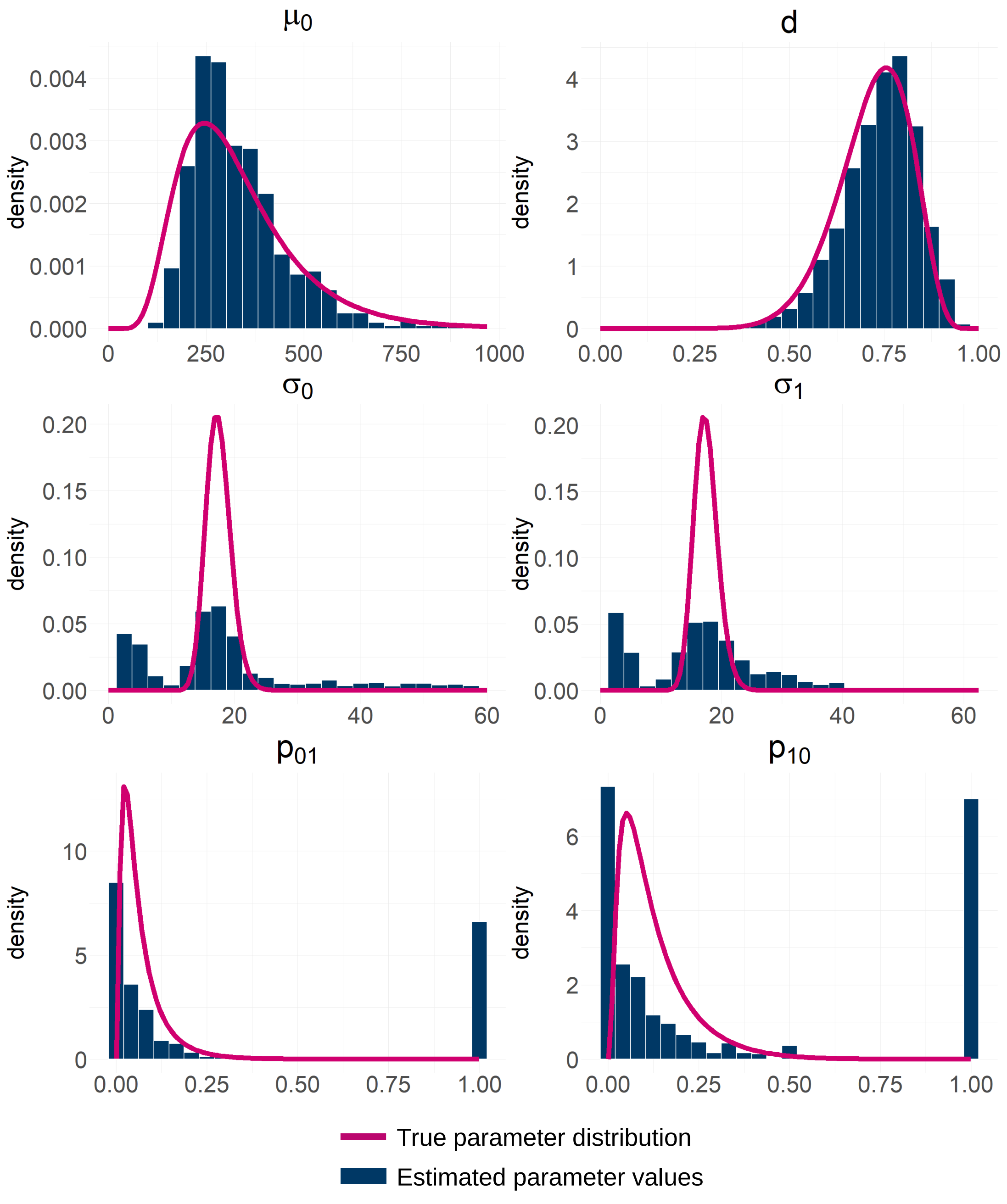
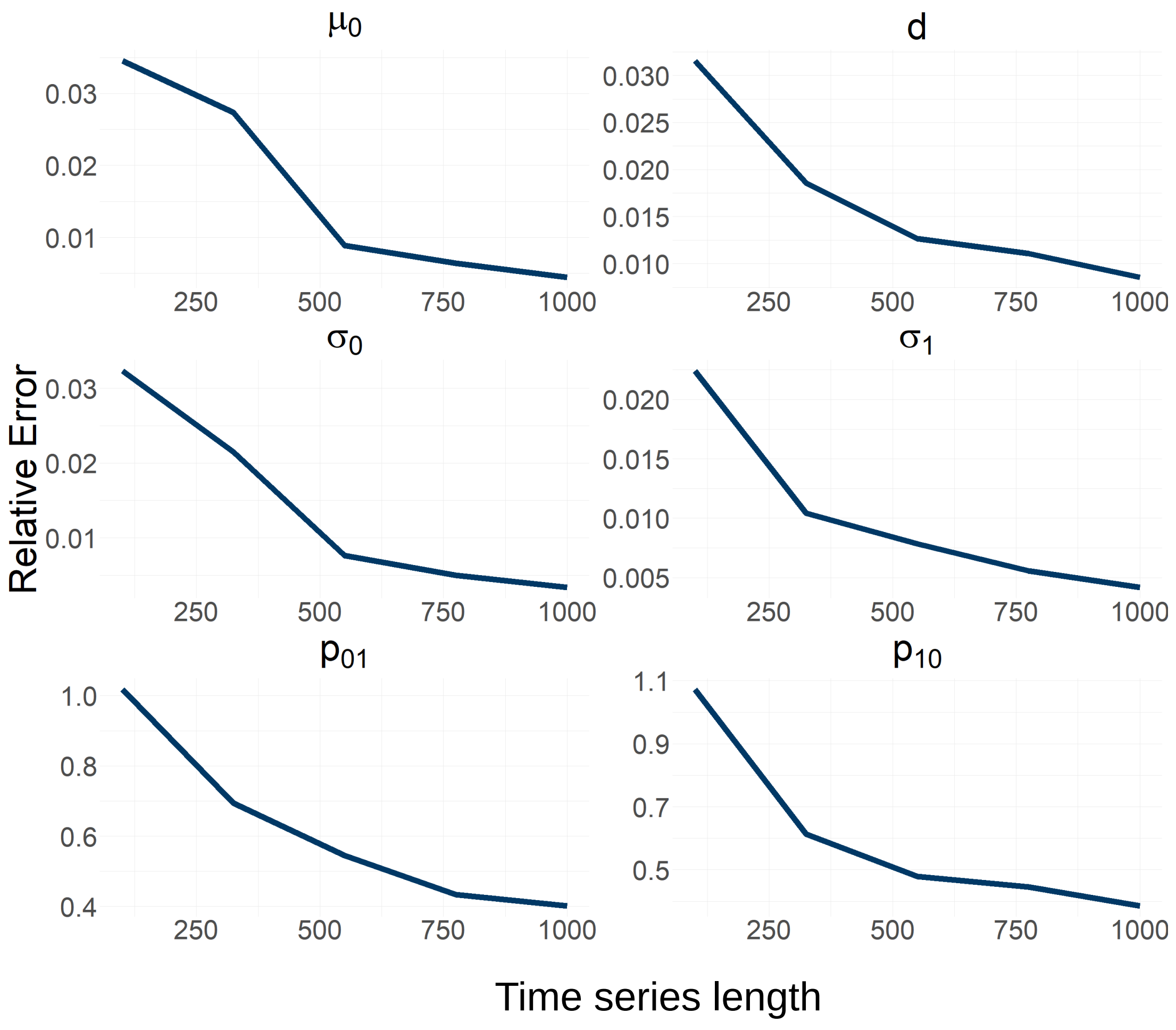


**Figure 4:** Median absolute error of estimated individual parameters relative to true median for varying data series lengths.

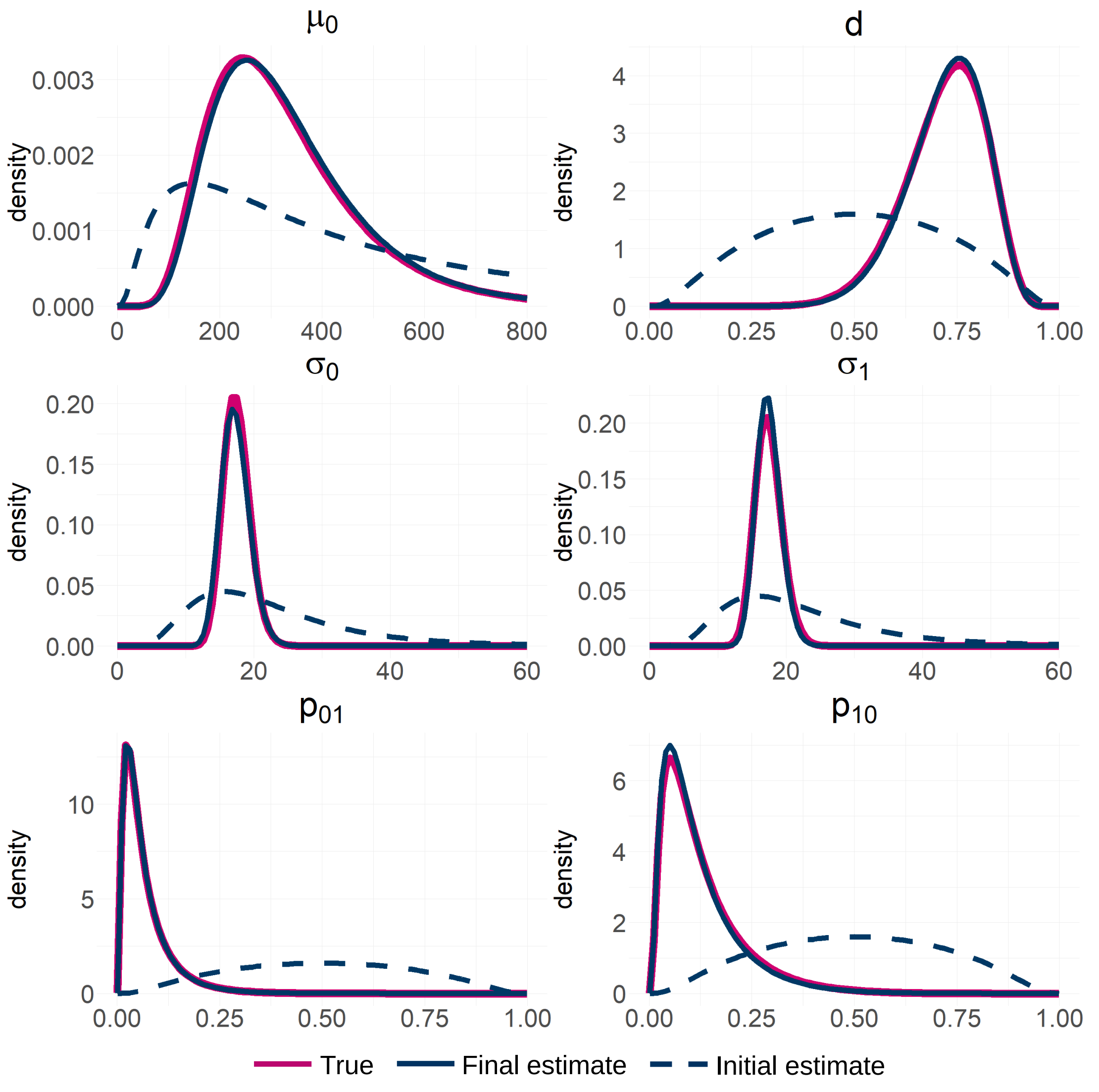


The large error in parameter estimates for short data series is believed to be driven by **individuals without any true transitions**. Inference for these cases is indeed difficult and remains an important issue to solve before applying the methods to clinical trial data.

## Future Work

The issues outlined here are being addressed in further work by using a mixed HMM framework with random effects on all model parameters and implementing the SAEM algorithm. Preliminary results indicate good alignment between estimated and true parameter distributions (Figure 5).

**Figure 5:** Estimated parameter distributons obtained with the SAEM algorithm, and true distributions ($T = 100$).



## Conclusions

The implemented algorithm performed quite well on longer time series in which the dynamics of the model were explicitly present, but failed on shorter series. Further work suggests that mixed HMMs and the SAEM algorithm may solve these issues.

**References**

1. Lavielle, M. Journal of Pharmacokinetics and Pharmacodynamics, 45(1), 91–105. (2018)
2. Jakobsson, L. et al. PAGE 32 (2024) Abstr 10868 [www.page-meeting.org/?abstract=10868]
3. Bilmes, Jeff A. Berkeley, CA: International Computer Science Institute. pp. 7–13 (1998)

*Presented at PAGE, Thessaloniki, June 2025*

JL is an employee of AstraZeneca and may hold shares or stock options