Efficient Uncertainty Propagation in Bayesian Multi-Step Procedures

DAGStat 2025

Svenja Jedhoff   Paul Bürkner

TU Dortmund University

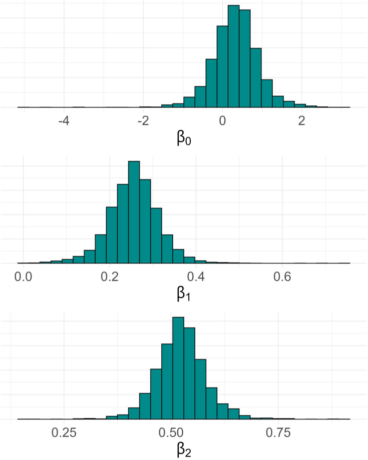
25.03.2025

Transregio 391

Dataset D

y	x_1	x_2
6	5	8
7	10	8
6	7	8
4	4	5
4	10	2
5	8	5
6	8	8
4	4	5
8	10	9
7	7	9

$$y \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma^2)$$



Dataset D

y	x_1	x_2
6	5	8
7	10	8
NA	7	NA
4	4	5
NA	NA	2
5	8	5
6	8	8
4	4	5
NA	NA	9
7	7	9

Multiple Imputation Chained Equation (MICE)

Dataset D_1

y	x_1	x_2
6	5	8
7	10	8
4	7	5
4	4	5
4	4	2
5	8	5
6	8	8
4	4	5
7	7	9
7	7	9

Dataset D_2

y	x_1	x_2
6	5	8
7	10	8
6	7	9
4	4	5
4	8	2
5	8	5
6	8	8
4	4	5
7	10	9
7	7	9

Dataset D_3

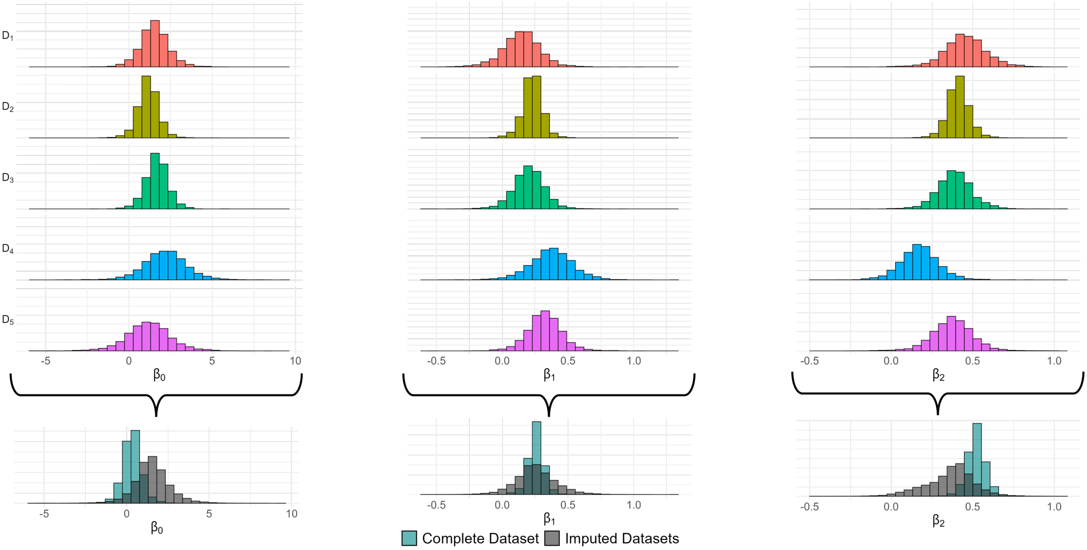
y	x_1	x_2
6	5	8
7	10	8
6	7	8
4	4	5
4	4	2
5	8	5
6	8	8
4	4	5
7	7	9
7	7	9

Dataset D_4

y	x_1	x_2
6	5	8
7	10	8
6	7	2
4	4	5
6	8	2
5	8	5
6	8	8
4	4	5
7	7	9
7	7	9

Dataset D_5

y	x_1	x_2
6	5	8
7	10	8
7	7	8
4	4	5
6	10	2
5	8	5
6	8	8
4	4	5
7	5	9
7	7	9


D1
D2
D3
D4
D5
-5
0
5
10
β0
-0.5
0.0
0.5
1.0
β1
β2
Complete Dataset
Imputed Datasets

- Dataset D with target variable y and predictors $x_1, \dots, x_p$.
- Interested in posterior distribution of parameter estimates θ of some model describing y depending on $x_1, \dots, x_p$ with

$$y \sim p(y \mid \theta)$$
$$\theta \sim p(\theta)$$
$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \propto p(y \mid \theta)p(\theta)$$

- Use MCMC to get posterior draws $\theta^{(1)}, \dots, \theta^{(S)}$.

technische universität dortmund

Goal: Propagating the uncertainty induced by the missing values into the posterior densities.

- Use MICE algorithm to get m imputed datasets $D_1, \ldots, D_m$ with underlying distribution $p_{\text{MICE}}(y \mid y^\star)$ dependent on original data $y^\star$.
- Calculate the posterior distribution given the original data $y^\star$:

$$p_{\text{MICE}}(\theta \mid y^\star) = \int p(\theta \mid y) p_{\text{MICE}}(y \mid y^\star) dy \overset{\text{Monte Carlo}}{\approx} \frac{1}{m} \sum_{i=1}^{m} p(\theta \mid y_i) \tag{1}$$

But:

- MCMC needs to be run separately for each dataset.

$\Rightarrow$ For large m and complex models this means high computational effort.

Solution:

- Posterior distributions $p(\theta \mid y_i)$ and $p(\theta \mid y_j)$ are similar to each other.
- Use importance sampling methods to approximate posterior distributions.

Goal: Approximating posterior distributions $p(\theta \mid y_i)$ for all $i = 1, \dots, m$ without running MCMC separately.

$\Rightarrow$ Use **Importance Sampling**

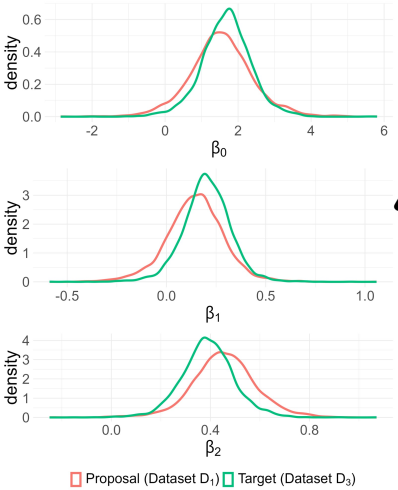
- Target distribution $p(\theta \mid y_i)$.
- Proposal distribution $q(\theta)$ with samples $\theta^{(1)}, \dots, \theta^{(S)}$.
- Importance ratios/weights:

$$w_s = w(\theta^{(s)}) = \frac{p(\theta^{(s)} \mid y_i)}{q(\theta^{(s)})} \tag{2}$$

- Use importance resampling to gain draws of target distribution:

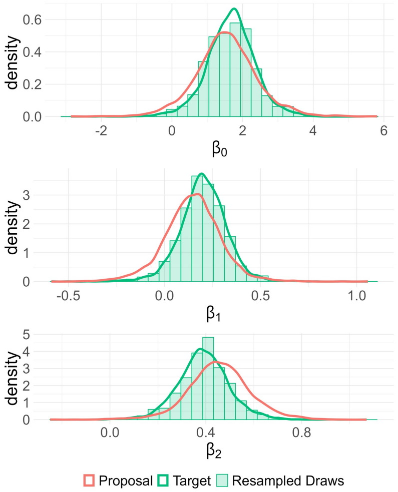
$$\theta_i^{(s)} \sim \text{multinomial}(S, (\theta^{(1)}, \dots, \theta^{(S)}), (w_1, \dots, w_S)) \tag{3}$$

$\Rightarrow$ Use **Pareto Smoothed Importance Sampling** [1] (PSIS) to stabilize importance weights and get a diagnostic tool $\hat{k}$.

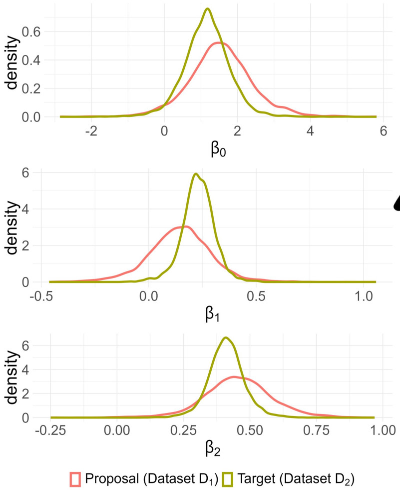


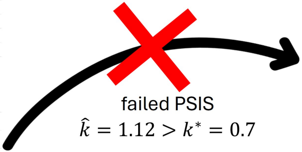
successful PSIS

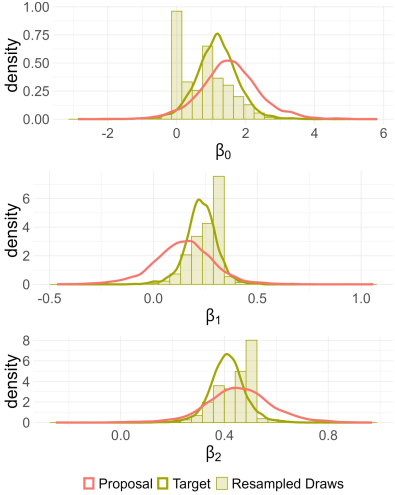
$\hat{k} = 0.46 < k^* = 0.7$



Methods Importance Sampling







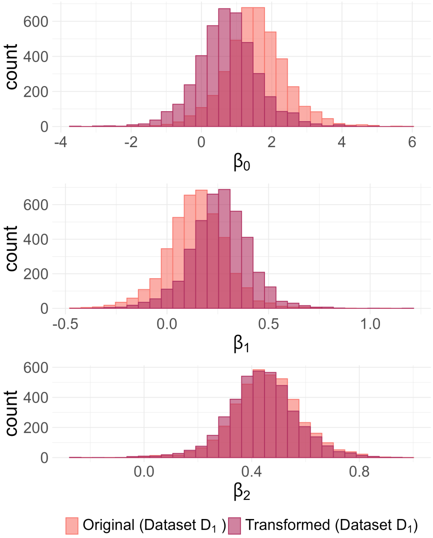
If PSIS fails a more advanced method is needed:

Importance Weighted Moment Matching (IWMM)[2]

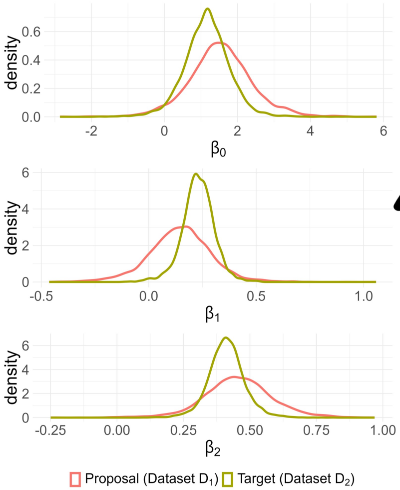
- Adaptive importance sampling method: Proposal distribution q is iteratively updated.
- Use affine transformations T to transform Monte Carlo samples:

$$T : \theta^{(s)} \mapsto \mathbf{A}\theta^{(s)} + \mathbf{b} =: \breve{\theta}^{(s)} \quad (4)$$

- Three different complexities of transformations to match the samples to the mean/marginal variance/covariance of the importance weights.

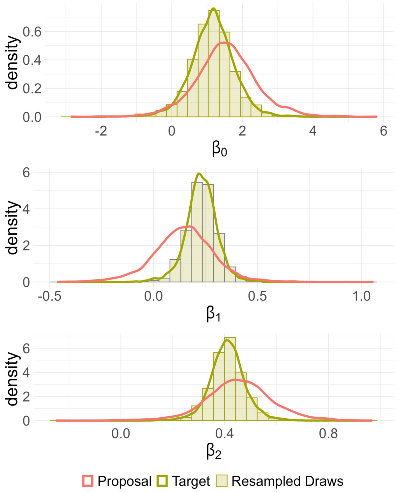


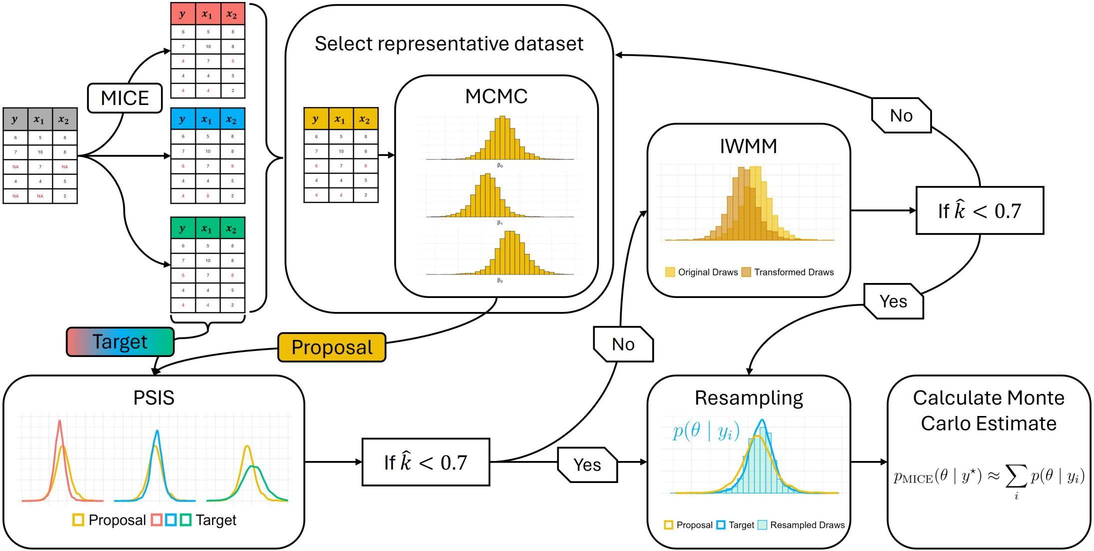
Methods IWMM



successful IWMM

$$\hat{k} = -0.01 < k^* = 0.7$$




MICE
Target
Proposal
Select representative dataset
MCMC
IWMM
Original Draws
Transformed Draws
If $\hat{k} < 0.7$
No
Yes
PSIS
Proposal
Target
If $\hat{k} < 0.7$
No
Yes
Resampling
$p(\theta \mid y_i)$
Proposal
Target
Resampled Draws
Calculate Monte Carlo Estimate
$p_{\text{MICE}}(\theta \mid y^{\star}) \approx \sum_i p(\theta \mid y_i)$

technische universität dortmund

Evaluation of Iterative Method

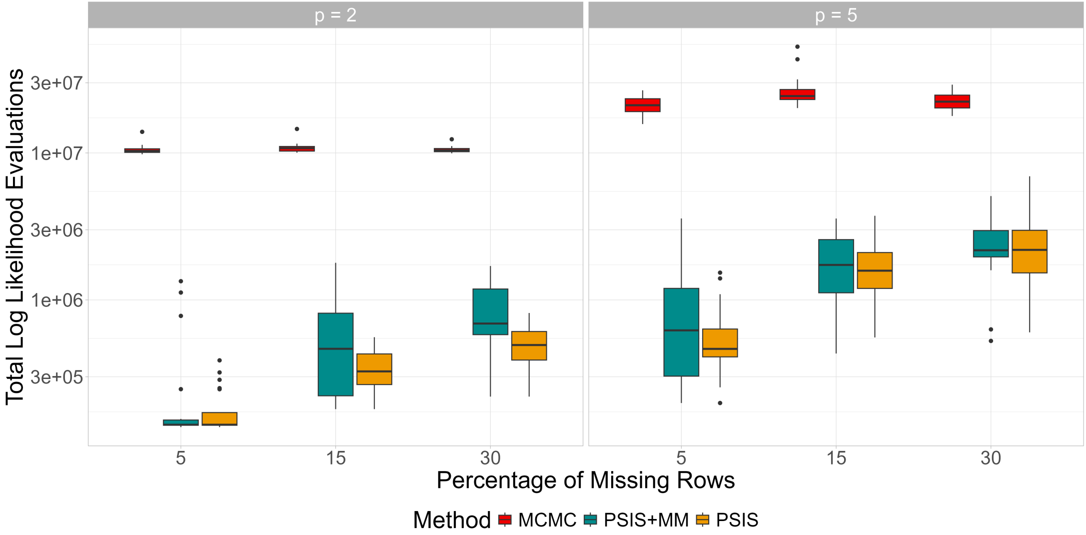
- Instead of runtime use number of log-likelihood evaluations necessary for the calculation.
- For MCMC (no-U-turn-sampler (NUTS) [3]) the log-probability and log-gradient needs to be evaluated.
- For PSIS and IWMM, log-ratios need to be calculated:

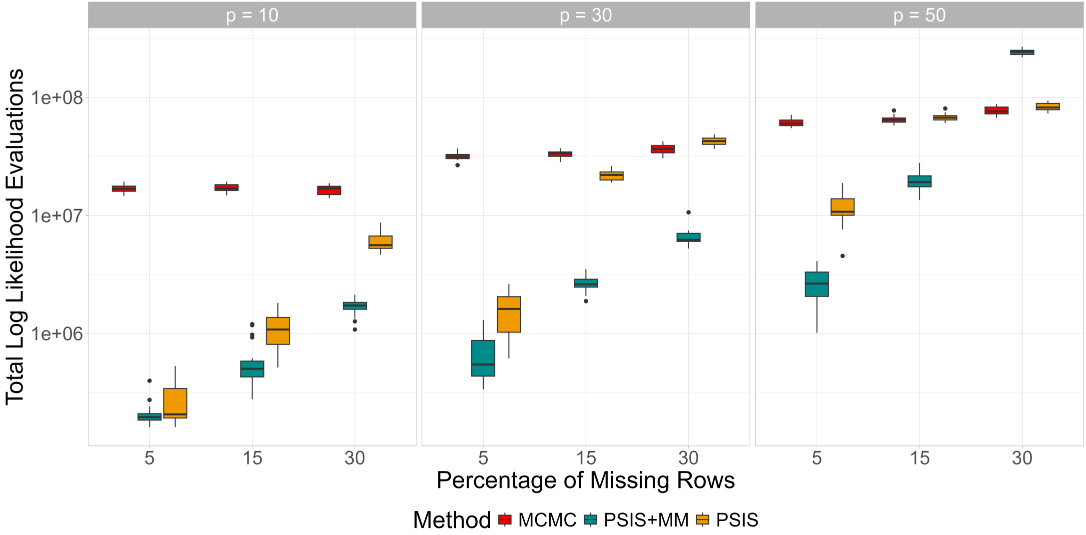
$$\log(w_s) = \log\left(\frac{\prod_{j=1}^{n} p(y_i^{(j)} \mid \theta)}{\prod_{j=1}^{n} p(y_*^{(j)} \mid \theta)}\right) = \log\left(\frac{\prod_{j\in I^*} p(y_i^{(j)} \mid \theta)}{\prod_{j\in I^*} p(y_*^{(j)} \mid \theta)}\right), \quad (5)$$

where I^* indicates the index set where datasets D_i and D_* have different rows.
$\Rightarrow$ Improves number of log-lik evaluations by factor $\mid I^* \mid /n$.

- Dataset D with
  - $n \in \{10, 100\}$ observations and
  - $p \in \{2, 5\}$ for $n = 10$ and $p \in \{10, 30, 50\}$ for $n = 100$ covariates.
- Linear Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \,, \epsilon \sim N(0, 1)$.
- Each dataset was imputed $m = 100$ times using `mice` [4] in `R` [5].
- Bayesian models were fitted using `brms` [6] in `R` and standard priors were used.

Simulation Study Results $n = 10$




p = 10
p = 30
p = 50
1e+08
1e+07
1e+06
Total Log Likelihood Evaluations
5
15
30
Percentage of Missing Rows
Method
MCMC
PSIS+MM
PSIS

Conclusion

- Big improvement in number of evaluations compared to MCMC.
- Further improvement possible by building a mixture distribution as a proposal for PSIS.
- Method can be applied on different problems.
- Simulation studies for problems with surrogate models [7] already running.

References I

[1] A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry. **"Pareto Smoothed Importance Sampling".** In: *Journal of Machine Learning Research* 25.72 (2024), pp. 1–58.

[2] T. Paananen, J. Piironen, P.-C. Bürkner, and A. Vehtari. **"Implicitly adaptive importance sampling".** In: *Statistics and Computing* 31.2 (Feb. 9, 2021), p. 16. DOI: 10.1007/s11222-020-09982-2. (Visited on 01/13/2025).

[3] M. D. Hoffman and A. Gelman. **"The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo".** In: *Journal of Machine Learning Research* 15.47 (2014), pp. 1593–1623. URL: http://jmlr.org/papers/v15/hoffman14a.html.

[4] S. van Buuren and K. Groothuis-Oudshoorn. **"mice: Multivariate Imputation by Chained Equations in R".** In: *Journal of Statistical Software* 45.3 (2011), pp. 1–67. DOI: 10.18637/jss.v045.i03.

[5] R Core Team. ***R: A Language and Environment for Statistical Computing.*** R Foundation for Statistical Computing. Vienna, Austria, 2024. URL: https://www.R-project.org/.

[6] P.-C. Bürkner. **"brms: An R Package for Bayesian Multilevel Models Using Stan".** In: *Journal of Statistical Software* 80.1 (2017), pp. 1–28. DOI: 10.18637/jss.v080.i01.

References II

[7] P. Reiser, J. E. Aguilar, A. Guthke, and P.-C. Bürkner. **"Uncertainty Quantification and Propagation in Surrogate-based Bayesian Inference".** en. In: *arXiv preprint* (2024).

[8] J. Zhang and M. A. Stephens. **"A New and Efficient Estimation Method for the Generalized Pareto Distribution".** In: *Technometrics* 51.3 (2009), pp. 316–325.

technische universität dortmund

- Potential Problems with Importance Sampling: proposal distribution q is not suitable, importance weights unstable, weights have a larger tail than suitable
- Solution: replace the M largest weights of $\{w_s\}_{s=1}^S$ with quantiles of the generalized Pareto distribution

Pareto Smoothed Importance Sampling (PSIS) [1]

- Order importance weights from lowest to highest $w_{(s)}, s = 1, \dots, S$
- Set $M = \min(0.2S, 3\sqrt{S})$ and $\mathrm{w}_s = w_{(s)}, k = 1, \dots, S - M$
- Estimate parameters of the generalized Pareto distribution: $\hat{u} = w_{(S-M)}$ and $\hat{k}$ and $\hat{\sigma}$ are estimated using the algorithm of Zhang and Stephens [8]
- Set $\mathrm{w}'_{(S-M+z)} = \min\left(F^{-1}\left(\frac{z-1/2}{M}\right), \max_s(w_s)\right)$, for each $z = 1, \dots, M$

$\Rightarrow$ we get smoothed weights $\{\mathrm{w}_s\}_{s=1}^S$ and a diagnostic tool $\hat{k}$

technische universität dortmund

Transformation T_1 is used to match the mean of the samples to the importance weighted mean:

$$\breve{\theta}_*^{(s)} = T_1(\theta_*^{(s)}) = \theta_*^{(s)} - \bar{\theta}_* + \tilde{\theta}_* \text{ with } \bar{\theta}_* = \frac{1}{S}\sum_{s=1}^{S}\theta_*^{(s)} \text{ and } \tilde{\theta}_* = \frac{\sum_{s=1}^{S} w_i(\theta_*^{(s)})\theta_*^{(s)}}{\sum_{s=1}^{S} w_i(\theta_*^{(s)})} \,. \tag{6}$$

Transformation T_2 is used to match the marginal variance in addition to matching the mean:

$$\breve{\theta}_*^{(s)} = T_2(\theta_*^{(s)}) = \tilde{\mathbf{v}}^{1/2} \circ \mathbf{v}^{-1/2} \circ (\theta_*^{(s)} - \bar{\theta}_*) + \tilde{\theta}_* \tag{7}$$

$$\text{with } \mathbf{v} = \frac{1}{S}\sum_{s=1}^{S}(\theta_*^{(s)} - \bar{\theta}_*) \circ (\theta_*^{(s)} - \bar{\theta}_*) \text{ and } \tilde{\mathbf{v}} = \frac{\sum_{s=1}^{S} w_i(\theta_*^{(s)})(\theta_*^{(s)} - \bar{\theta}_*) \circ (\theta_*^{(s)} - \bar{\theta}_*)}{\sum_{s=1}^{S} w_i(\theta_*^{(s)})} \,, \tag{8}$$

with $\circ$ indicating the pointwise product of two vectors. To also match the covariance and the mean, transformation T_3 can be applied:

$$\breve{\theta}_*^{(s)} = T_3(\theta_*^{(s)}) = \tilde{\mathbf{L}}\mathbf{L}^{-1}(\theta_*^{(s)} - \bar{\theta}_*) + \tilde{\theta}_* \tag{9}$$

$$\text{with } \mathbf{L}\mathbf{L}^T = \Sigma = \frac{1}{S}\sum_{s=1}^{S}(\theta_*^{(s)} - \bar{\theta}_*)(\theta_*^{(s)} - \bar{\theta}_*)^T \text{ and } \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T = \frac{\sum_{s=1}^{S} w_i(\theta_*^{(s)})(\theta_*^{(s)} - \tilde{\theta}_*)(\theta_*^{(s)} - \tilde{\theta}_*)^T}{\sum_{s=1}^{S} w_i(\theta_*^{(s)})} \,. \tag{10}$$

Input: PSIS threshold $k_{\text{threshold}}$, proposal density $p(\theta|\tau^*)$, draws $\{\theta_*^{(s)}\}_{s=1}^S$ from $p(\theta|\tau^*)$

Output: $\hat{k}$, updated draws $\{\breve{\theta}_*^{(s)}\}_{s=1}^S$ and weights $\{\breve{w}(\breve{\theta}_*^{(s)})\}_{s=1}^S$

Compute importance weights $\{w(\theta_*^{(s)})\}_{s=1}^S$ and diagnostic $\hat{k}$.

while $\hat{k} > k_{threshold}$ **do**

  for *j in 1:3* **do**

    Transform the draws with $T_j : \theta_*^{(s)} \mapsto \breve{\theta}_*^{(s)}$.

    Recompute weights $\{\breve{w}(\breve{\theta}_*^{(s)})\}_{s=1}^S$ and diagnostic $\hat{\hat{k}}$.

    if $\hat{\hat{k}} < \hat{k}$ **then**

      Accept the transformation and update $\{\theta_*^{(s)}\}_{s=1}^S = \{\breve{\theta}_*^{(s)}\}_{s=1}^S$, $\{w(\theta_*^{(s)})\}_{s=1}^S = \{\breve{w}(\breve{\theta}_*^{(s)})\}_{s=1}^S$ and $\hat{k} = \hat{\hat{k}}$.

      Exit for loop.

    else

      Discard the transformation.

    if *j == 3* **then**

      Moment matching failed because $\hat{k} > k_{\text{threshold}}$, end algorithm with a warning about sampling anaccuracy.

return Moment Matching successful: Return $\hat{k}$, updated draws $\{\breve{\theta}_*^{(s)}\}_{s=1}^S$ and weights $\{\breve{w}(\breve{\theta}_*^{(s)})\}_{s=1}^S$

Appendix Iterative Method Algorithm (Part I)

technische universität dortmund

Input: Variable of interest $\theta \sim p(\theta)$, underlying variable $\tau \sim p(\tau)$
Output:

Draw m samples $\tau^{(1)}, \ldots, \tau^{(m)} \sim p(\tau)$.
Set index set $I = \{1, \ldots, m\}$.
while $|I| > 0$ **do**
    Select representative values from $\{\tau^{(i)}\}_{i \in I}$ and run MCMC to access their posterior distributions through samples $\{\theta_*^{(s)}\}_{s=1}^S$. If multiple values are selected, build a mixture distribution and gain samples through resampling.
    Remove the selected values from the index set I.
    for $i \in I$ **do**
        Run PSIS with target $p(\theta|\tau^{(i)})$ and proposal $p(\theta|\tau^*)$ resp. $q(\theta)$ and gain importance weights and metric $\hat{k}_i$.
        if $\hat{k}_i < 0.7$ **then**
            Use calculated importance weights and draws $\{\theta_*^{(s)}\}_{s=1}^S \sim p(\theta|\tau^*)$ resp. $q(\theta)$ for importance resampling and save the resampled draws as posterior draws corresponding to $p(\theta|\tau^{(i)})$.
            Remove i from I: $I \leftarrow I \setminus i$.

Appendix Iterative Method Algorithm (Part II)

```
    else
        Run IWMM with with target p(θ|τ^(i)) and proposal p(θ|τ*) resp. q(θ) and gain transformed
         importance weights {θ̆_*^(s)}_{s=1}^S and new metric k̂_i.
        if k̂_i < 0.7 then
            Use calculated transformed importance weights and transformed draws {θ̆_*^(s)}_{s=1}^S for importance
             resampling and save the resampled draws as posterior draws corresponding to p(θ|τ^(i)).
            Remove i from I: I ← I \ i.
return Posterior draws corresponding to all p(θ|τ^(i)), i = 1, ..., m.
```

else

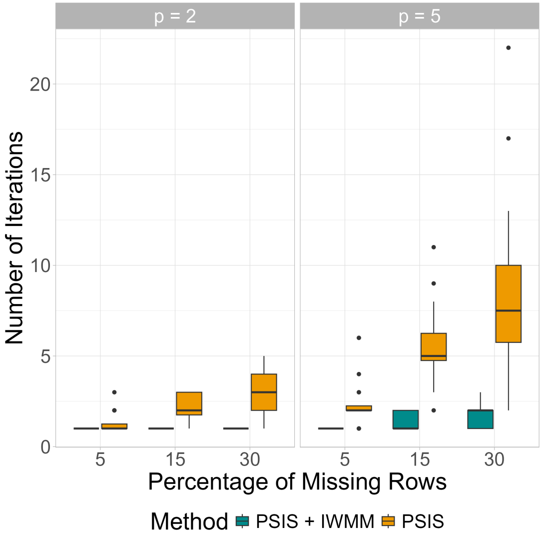
Run IWMM with with target $p(\theta|\tau^{(i)})$ and proposal $p(\theta|\tau^*)$ resp. $q(\theta)$ and gain transformed importance weights $\{\breve{\theta}_*^{(s)}\}_{s=1}^S$ and new metric $\hat{\breve{k}}_i$.

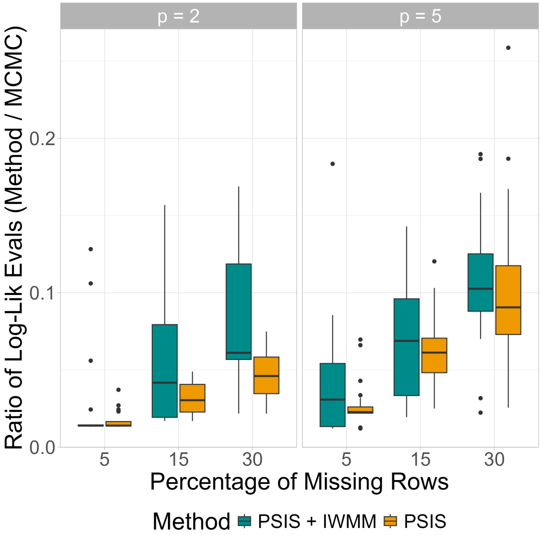
if $\hat{\breve{k}}_i < 0.7$ **then**

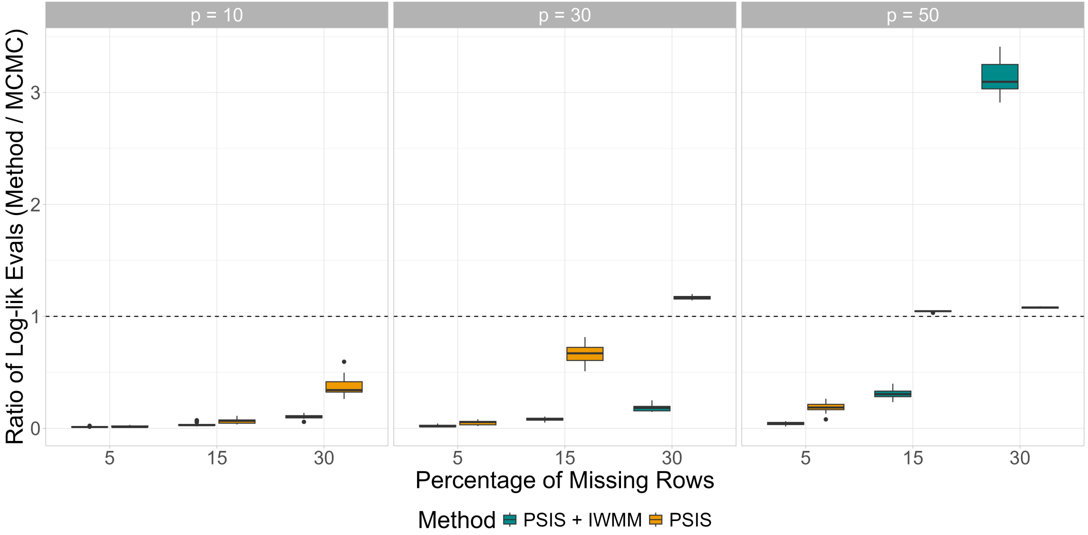
Use calculated transformed importance weights and transformed draws $\{\breve{\theta}_*^{(s)}\}_{s=1}^S$ for importance resampling and save the resampled draws as posterior draws corresponding to $p(\theta|\tau^{(i)})$.

Remove i from I: $I \leftarrow I \setminus i$.

return Posterior draws corresponding to all $p(\theta|\tau^{(i)}), i = 1, \ldots, m$.


p = 2
p = 5
Number of Iterations
0
5
10
15
20
5
15
30
Percentage of Missing Rows
Method
PSIS + IWMM
PSIS


p = 2
p = 5
Ratio of Log-Lik Evals (Method / MCMC)
0.0
0.1
0.2
5
15
30
Percentage of Missing Rows
Method
PSIS + IWMM
PSIS


p = 10
p = 30
p = 50
Ratio of Log-lik Evals (Method / MCMC)
0
1
2
3
5
15
30
Percentage of Missing Rows
Method
PSIS + IWMM
PSIS