UNIVERSITÄT ZU LÜBECK
INSTITUTE FOR ELECTRICAL
ENGINEERING IN MEDICINE

Benchmarking MCMC Samplers on Challenging Synthetic Posteriors

MC-FiT: A Synthetic Benchmarking Framework

Fabian Kohrs

September 25, 2025

Master's Thesis - University of Lübeck

Summary

Problem: Markov Chain Monte Carlo (MCMC) performance depends strongly on posterior geometry (multimodality, correlation, dimensionality, tail weight). Guidance is fragmented and often heuristic.

Approach: MC-FiT: define synthetic *posteriors directly*, vary attributes systematically, and evaluate samplers against IID reference samples using distributional distances + diagnostics.

Contributions:

- A reusable, controlled benchmark framework for posterior geometries.
- Empirical mapping of attribute effects and break points for multiple samplers.
- Practical guidelines for sampler choice conditioned on anticipated geometry.

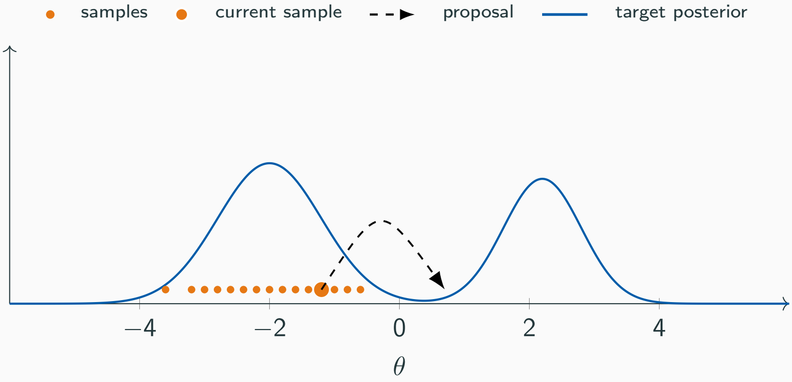
Roadmap

Bayesian Inference & the Challenge

- Goal: characterize the posterior $p(\theta \mid D) \propto p(D \mid \theta)p(\theta)$.
- Intractable evidence $\Rightarrow$ approximate inference; MCMC widely used.
- Real constraint: finite compute budgets $\Rightarrow$ need to know when we get accurate samples.
- Poor approximation $\Rightarrow$ biased estimates, misleading uncertainty.
- **Key insight:** posterior *geometry* drives sampler efficiency/accuracy.

Geometry attributes studied: multimodality, dimensionality, correlation, tail weight.

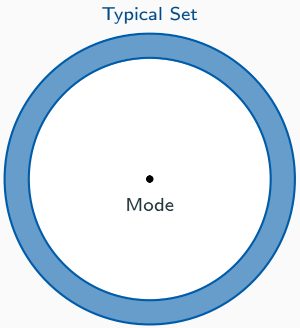
Multimodality: Chains Get Stuck



Problem: Low-density valleys block transitions.

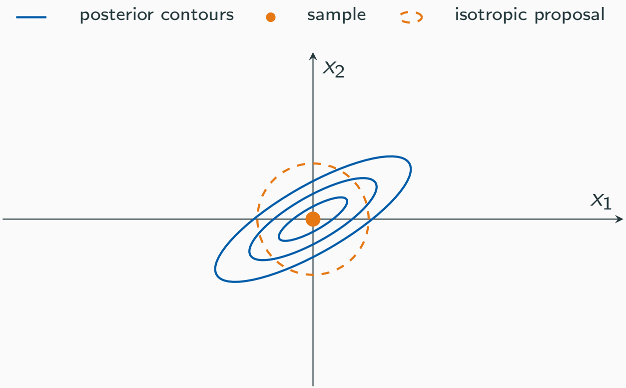
Consequence: Chains remain stuck in one mode $\Rightarrow$ biased samples.

High Dimensionality & the Typical Set



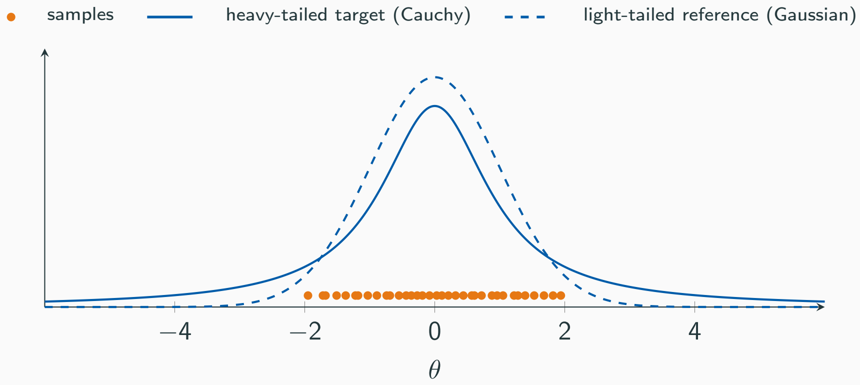
Problem: In high dimensions, most mass lies in the thin typical set rather than at the mode.
Consequence: Proposals must be tuned to this scale, otherwise acceptance decays and chains mix poorly.

Correlation / Curvature: Narrow Ridges



Problem: Posterior mass lies along narrow ridges.
Consequence: Isotropic proposals waste moves orthogonal to the ridge $\Rightarrow$ slow exploration.

Heavy Tails: Slow Convergence



Problem: Proposals struggle to balance center and heavy tails.
Consequence: Chains under-sample tails ⇒ unstable, slow convergence.

Samplers (Quick Intro)

- **Metropolis–Hastings (MH)[1]:**
  random-walk proposals + accept/reject.
- **Hamiltonian Monte Carlo (HMC)[2]:**
  gradient-informed proposals + accept/reject.
- **Differential Evolution Metropolis (DEM)[3]:**
  adaptive proposals from differences of two past samples (scaled).
- **Sequential Monte Carlo (SMC) [4]:**
  sequence of tempered distributions + resampling.

[1] Metropolis et al. (1953); Hastings (1970)
[2] Duane et al. (1987); Neal et al. (2011)
[3] Braak et al.(2006)
[4] Doucet et al. (2001)

Existing Benchmarking Frameworks

PosteriorDB: realistic models + some reference posteriors; limited control over geometry.[5]

MCBench: synthetic targets + IID distances; limited set of fixed distributions.[6]

Gap: Need *systematic*, multi-attribute control (dim, correlation, tails, modes) with IID references for accuracy *and* efficiency comparisons.

[5] Magnusson et al. (2024)
[6] Ding et al. (2025)

Roadmap

MC-FiT: Concept

Idea: Define target posteriors directly (single or mixture of Normal / Student-t), then **vary attributes parametrically**.

- Supports **single** and **mixture** posteriors.
- **Initialization:** uniform over IID-derived bounding box.

Evaluation: Metrics & Rationale

Diagnostics ($\hat{R}$, ESS) and **efficiency** (runtime, ESS/s).

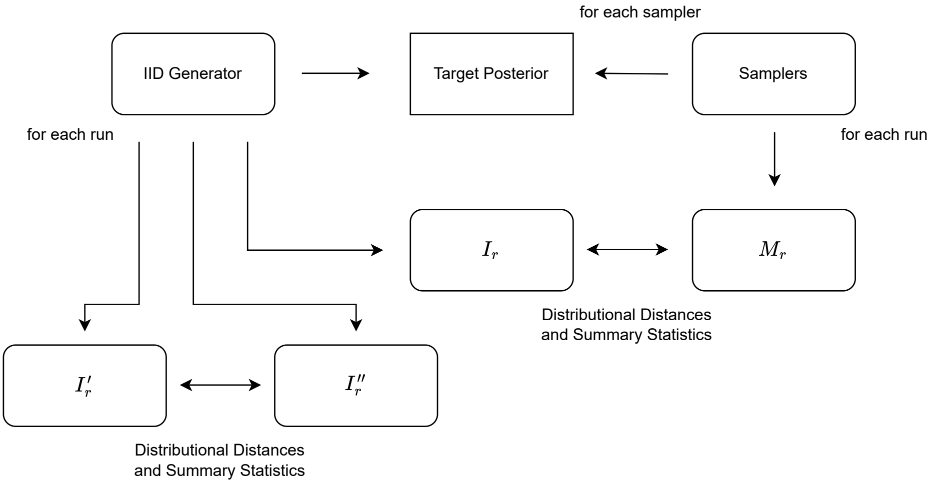
Summary discrepancies: RMSE of per-dimension mean/variance vs. IID.

Distributional distances: Sliced Wasserstein Distance (SWD) (many 1D projections) and Maximum Mean Discrepancy (MMD).

Why baselines?

- Even perfect samplers show non-zero finite-sample distance.
- Enables normalization (Glass's Δ).

Workflow per Posterior



A schematic view of one full posterior evaluation in MC-FiT.

Glass's Δ (Effect Size Normalization)

Definition

$$\Delta = \frac{\bar{x}_{\text{MCMC}} - \bar{x}_{\text{IID}}}{s_{\text{IID}}}$$

where $\bar{x}_{\text{MCMC}}$ is the metric from sampler output, $\bar{x}_{\text{IID}}$ and s_{IID} are mean and std. from IID baselines.

Intuition

- Accounts for finite-sample variability in baselines.
- $\Delta \approx 0$: sampler indistinguishable from IID baseline.
- Larger Δ: stronger deviation .

Roadmap

Design Overview

Experiment stages from single-attribute to multi-attribute combinations.

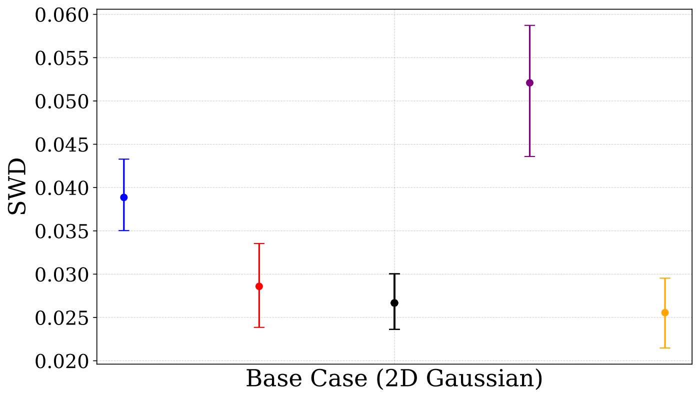
Value grids per attribute (dimension, correlation strength, tail weight, mode distance).

Protocol with fixed defaults (samples, chains, repetitions), identical random seeds *Goal:* reveal *thresholds / break points* where performance changes sharply.

Roadmap

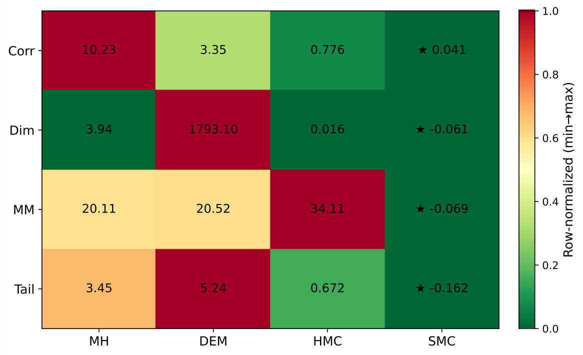
Baseline: 2D Gaussian

- All samplers near IID baseline.
- HMC and SMC are the best



● MH ● HMC ● IID baseline ● DEM ● SMC

Single-Attribute Effects

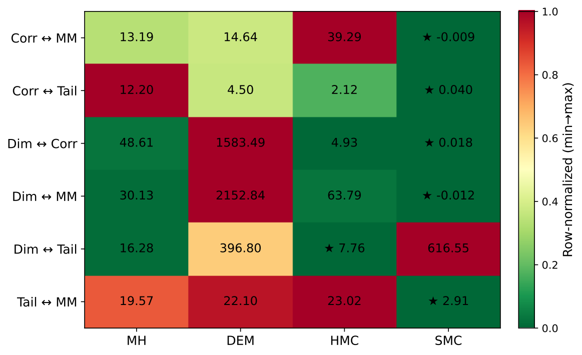


Dark green = closer to IID (better), red = worse.

Most important findings:

- **SMC** consistently best across all attributes
- **HMC** strong overall, but struggles with **multimodality**
- **DEM** fails badly with increasing **dimension**
- **MH** weak under strong **correlation**

Two-Attribute Interactions

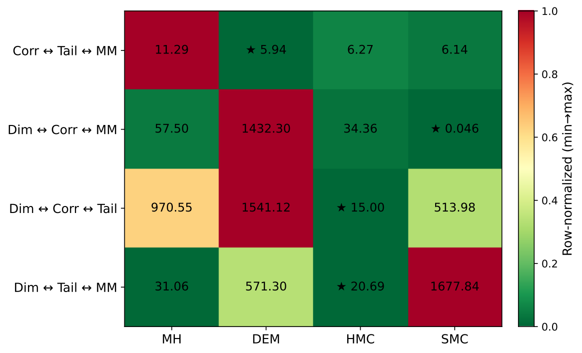


Dark green = closer to IID (better), red = worse.

Most important findings:

- **SMC** strong overall, but **collapses for Dim × Tail**
- **HMC** robust to dimensions/tails, but **fails under multimodality**
- **DEM** consistently poor whenever **dimension** is involved
- **MH** intermediate, handles correlation × tails reasonably

Three-Attribute Interactions

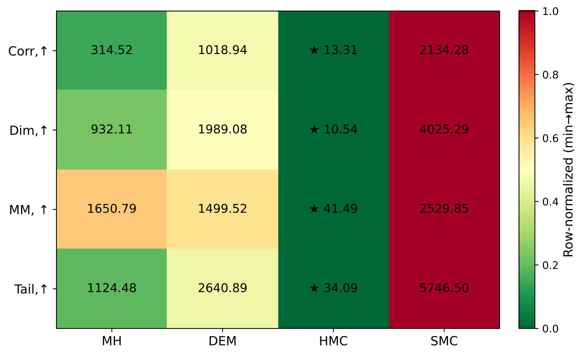


Dark green = closer to IID (better), red = worse.

Most important findings:

- **HMC** most stable across triplets
- **SMC** loses dominance - struggles with heavy tails
- **DEM** collapses, with one rare success (Corr–Tail–MM)

Four-Attribute Interactions



Dark green = closer to IID (better), red = worse.

Most important findings:

- Fully stressed scenario: three attributes fixed high, vary the fourth
- **Only HMC remains usable** ($\Delta \approx$ 10–40)
- **MH** better than DEM/SMC, but still highly inaccurate
- **DEM & SMC** collapse (huge Δ, often in the thousands)

Guidelines derived from observations

If you expect strong correlation/curvature
Use gradient-informed samplers like HMC; avoid isotropic MH.

If you expect multimodality
Consider tempered methods like SMC; MH/HMC risk mode trapping.

If you expect high dimension
HMC scales better than MH. If also heavy tails do not use SMC.

If you expect extreme stresses
Only HMC remains usable (though accuracy degrades).

Roadmap

Conclusion

- MC-FiT enables **controlled**, **reproducible** benchmarking across geometries.
- Distributional distances + IID baselines reveal failures missed by basic diagnostics.
- Clear empirical guidance emerges for sampler choice under geometry assumptions.

Outlook

- Extend posterior families (e.g., skewness).
- Extend samplers in framework
- Integrate option to include own MCMC samples.

Questions

Thank you!

Questions welcome.