Bending the Line: Bayesian Inference in the Face of Misspecified Models

24223499
Department of Statistical Science
University College London

Abstract

We conduct a Bayesian analysis of regression under controlled misspecification with skew-normal errors. Hamiltonian Monte Carlo (HMC) methods with gradient-based dynamics are benchmarked against BayesBag for linear and quadratic models. On synthetic data (n = 100) with bimodal covariates and skew-normal errors ($\alpha = 3$), HMC demonstrates superior mixing properties, under model misspecification. While the correctly specified quadratic model yields superior recovery (MSE = 0.97), BayesBag offers massively reduced variance, though little bias correction (MSE = 3.21 vs. 3.22) over a misspecified linear model at a steep $24\times$ computational cost. These results underscore that algorithmic improvements mitigate, but do not resolve, the limitations of structural misspecification.

1 Motivation

Modern statistical practice increasingly grapples with model misspecification in complex datasets, making robust inference methods essential. While Bayesian approaches theoretically accommodate uncertainty through posterior distributions, their practical performance depends critically on both algorithmic efficiency and model adequacy. This investigation evaluates parallel tempering Hamiltonian Monte Carlo (PT-HMC) under controlled misspecification, building on foundational work by Gelman et al. [2020] and Huggins and Miller [2022, 2024].

Misspecified models present a dual challenge in Bayesian workflows: computationally, they often induce complex posterior geometries with poor mixing properties; inferentially, they yield systematically biased estimates regardless of sample size. The literature has approached these challenges from two directions, improving sampling efficiency through geometric and tempering methods [Neal et al., 2011, Betancourt et al., 2017, Geyer et al., 1991] or enhancing robustness through posterior averaging [Bühlmann, 2014, Huggins and Miller, 2022]. However, the interplay between these approaches remains underexplored, particularly in settings where structural misspecification dominates parameter uncertainty.

Our contribution lies in rigorously comparing these approaches under a controlled experimental design. Specifically, we investigate whether tempering strategies in PT-HMC can mitigate multimodality while simultaneously exposing the limitations of posterior averaging under severe model mismatch. The bimodal covariate design with skew-normal errors creates an ideal testbed that mirrors real-world scenarios where practitioners must choose between more complex models and more robust algorithms

Preprint. Under review.

2 Methodology

2.1 Data Generating Process

The synthetic dataset was generated following a quadratic regression model with skew-normal errors. We consider $n = 100$ observations and $p = 10$ parameters from the quadratic model:

$$Y_i = \gamma_0 x_i^2 + \mathbf{Z}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \text{SN}(0, 1, 3) \tag{1}$$

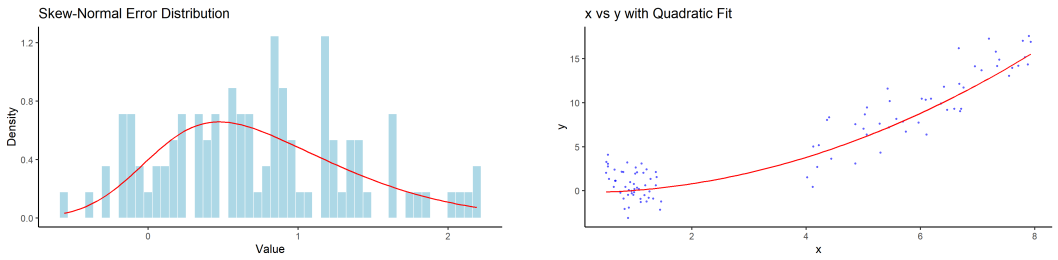
Where $SN(0, 1, 3)$ denotes the skew-normal distribution with location 0, scale 1, and shape parameter 3. Here, x_i is the predictor variable, $\mathbf{Z}_i$ is a vector of covariates, and $(\gamma_0, \boldsymbol{\beta}_0)$ are the true parameters. The design matrix $\mathbf{Z} \in \mathbb{R}^{100\times 10}$ incorporates controlled multicollinearity through engineered relationships between covariates. Specific columns were constructed as linear combinations of base normal variables to induce moderate correlations while maintaining full rank. This structure emulates realistic scenarios where covariates exhibit inherent dependencies. Parameters were set with $\gamma_0 = 0.25$ and $\boldsymbol{\beta}_0$ forming a linear sequence from -0.2 to 0.2 to ensure identifiable effects.

The critical predictor x follows a bimodal mixture distribution to test model robustness under covariate shift:

$$x_i \sim \mathcal{U}(a_1, b_1) \cdot \mathbb{I}(i \leq n/2) + \mathcal{U}(a_2, b_2) \cdot \mathbb{I}(i > n/2) \tag{2}$$

Where (a_1, b_1) and (a_2, b_2) define disjoint intervals, creating distinct data regimes. This design tests model flexibility in capturing non-linear relationships across heterogeneous subpopulations.

Skew-normal errors were generated via latent thresholding, leveraging the constructive definition. This algorithm is denoted in Appendix A.



(a) Distribution of Simulated Skew-Normal Errors (b) Relationship between predictor x and response y

Figure 1: Data Generating Process

Figure 1 illustrates the data generation process underlying our simulation study. Panel 1a shows the Skew-Normal error distribution (SN(0,1,3)) used to introduce realistic noise into the model. Panel 1b demonstrates how these errors manifest in the relationship between predictor x and response y. The true quadratic relationship is perturbed by the skew-normal errors, resulting in the observed data points. The bimodal distribution of x values reveals how the error structure affects different regions of the predictor space, due to the engineered correlation structure.

2.2 Model Specification

We compare two Bayesian formulations under $N(0, I)$ priors:

$$\text{(\textit{Misspecified}) Linear Model:} \quad Y_i = \gamma x_i + \mathbf{Z}_i^\top \boldsymbol{\beta} + \varepsilon_i \tag{3}$$

$$\text{(\textit{True}) Quadratic Model:} \quad Y_i = \gamma x_i^2 + \mathbf{Z}_i^\top \boldsymbol{\beta} + \varepsilon_i \tag{4}$$

where Y_i represents the response variable, x_i is the predictor of interest, $\mathbf{Z}_i$ is a vector of covariates with corresponding coefficient vector $\boldsymbol{\beta}$, γ is the coefficient for the predictor of interest (either linear or quadratic), and ε_i represents the error term.

2.3 MCMC Framework

We estimated both the misspecified linear and the true quadratic models using a multivariate standard normal prior. In order to better understand the models, we used 20,000 iterations, with a warmup of 8,000, $L = 15$, $\epsilon_0 = 0.0015$. The full algorithm framework is given in Appendix B.

2.3.1 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) offers a substantial advancement over traditional Metropolis-Hastings methods by leveraging gradient information to generate proposal states that are simultaneously distant from the current state yet maintain high acceptance probabilities Neal et al. [2011], Duane et al. [1987]. This approach transforms the sampling problem into a physical system by introducing an auxiliary momentum variable, enabling efficient exploration of complex posteriors.

The central insight of HMC is to augment the parameter space or position $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$ with a momentum variable $\mathbf{r} \in \mathbb{R}^{d+1}$, constructing a joint distribution:

$$\pi(\boldsymbol{\theta}, \mathbf{r}) \propto \exp(-H(\boldsymbol{\theta}, \mathbf{r})) \tag{5}$$

where $H(\boldsymbol{\theta}, \mathbf{r})$ denotes the Hamiltonian function:

$$H(\boldsymbol{\theta}, \mathbf{r}) = -\log \pi(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{r}^T\mathbf{r} \tag{6}$$

This formulation represents a physical interpretation where $-\log \pi(\boldsymbol{\theta})$ corresponds to potential energy and $\frac{1}{2}\mathbf{r}^T\mathbf{r}$ to kinetic energy. The momentum components are typically drawn from a standard multivariate normal distribution, independent of $\boldsymbol{\theta}$.

The evolution of the system follows Hamilton's equations:

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\partial H}{\partial \mathbf{r}} = \mathbf{r} \tag{7}$$

$$\frac{d\mathbf{r}}{dt} = -\frac{\partial H}{\partial \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}) \tag{8}$$

These dynamics preserve the Hamiltonian value and generate trajectories that maintain constant energy[1]. Consequently, proposals generated along these trajectories exhibit theoretical acceptance probabilities of one, enabling efficient exploration of the target distribution's high-density regions.

2.3.2 Stormer-Verlet (Leapfrog) Integration

Analytical solutions to Hamiltonian dynamics are generally unavailable for complex posteriors, necessitating numerical integration. The leapfrog integrator is ideal due to its symplectic properties and time-reversibility, which maintain detailed balance.

For step size ϵ and current state $(\boldsymbol{\theta}, \mathbf{r})$, each leapfrog step proceeds as:

1. Half-step momentum update: $\mathbf{r}(t + \epsilon/2) = \mathbf{r}(t) + (\epsilon/2)\nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}(t))$
2. Full-step position update: $\boldsymbol{\theta}(t + \epsilon) = \boldsymbol{\theta}(t) + \epsilon\mathbf{r}(t + \epsilon/2)$
3. Half-step momentum update: $\mathbf{r}(t + \epsilon) = \mathbf{r}(t + \epsilon/2) + (\epsilon/2)\nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}(t + \epsilon))$

This symmetric update enhances numerical stability. After L leapfrog steps, the proposal $(\boldsymbol{\theta}^*, \mathbf{r}^*)$ is accepted with probability:

$$\alpha(\theta^*, r^*) = \min\left[1, \frac{\exp\left(H(\boldsymbol{\theta}^*, \mathbf{r}^*)\right)}{\exp(H(\boldsymbol{\theta}, \mathbf{r}))}\right] \tag{9}$$

If accepted, $\boldsymbol{\theta}^*$ enters the Markov chain while momentum is discarded and resampled in the next iteration. This ensures convergence to the target distribution despite integration errors.

2.3.3 Computing Gradients for Posterior Dynamics

Key to efficient HMC implementation is accurate gradient evaluation of the log-posterior density. For parameter vector $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta})^T$, the gradient decomposes as:

$$\nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta} \mid \mathbf{y}) = \nabla_{\boldsymbol{\theta}} \log p(\mathbf{y} \mid \boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) \tag{10}$$

[1] HMC borrows from classical physics - imagine rolling a ball across a hilly landscape, where the shape of the terrain (posterior) guides its motion. The potential energy is the elevation (likelihood), and kinetic energy is how fast it's moving (momentum).

With skew-normal errors ($\alpha = 3$), the score function for each observation incorporates the Mills ratio:

$$s_i = \varepsilon_i - \alpha \frac{\phi(\alpha\varepsilon_i)}{\Phi(\alpha\varepsilon_i) + \delta} \tag{11}$$

where ϕ and Φ denote the standard normal PDF and CDF, and $\delta \approx 10^{-10}$ maintains numerical stability. For residuals $\varepsilon_i = y_i - \mu_i$, the gradients differ by model specification:

$$\frac{\partial \log p(\mathbf{y} \mid \boldsymbol{\theta})}{\partial \gamma} = \begin{cases} \sum_{i=1}^{n} s_i x_i & \text{(linear)} \\ \sum_{i=1}^{n} s_i x_i^2 & \text{(quadratic)} \end{cases} \quad \frac{\partial \log p(\mathbf{y} \mid \boldsymbol{\theta})}{\partial \beta_j} = \sum_{i=1}^{n} s_i Z_{ij} \tag{12}$$

With standard normal priors $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, the prior gradient simplifies to $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) = -\boldsymbol{\theta}$.

2.3.4 Parallel Tempering

While HMC is powerful, it often struggles with complex posteriors which exhibit multiple modes or regions of high curvature. To address this, we implement Parallel Tempering for our HMC (PT-HMC henceforth), which runs multiple Markov chains at different "temperatures" to enhance exploration. We define a temperature ladder $T = T_1, T_2, \ldots, T_n$ with $T_1 = 1 < T_2 < \ldots < T_n$, where each chain samples from:

$$\pi_i(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta})^{1/T_i} \tag{13}$$

Higher temperatures flatten (temper) the posterior, allowing chains to traverse energy barriers more easily. The tempered log-posterior is:

$$\log \pi_i(\boldsymbol{\theta}) = \frac{1}{T_i} \log p(\mathbf{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \tag{14}$$

We use a geometric temperature ladder, $T_i = \exp(\log(T_{\max}) \cdot (i-1)/(n-1))$, with $T_{\max} = 5$ and $n = 5$ chains.

A key feature of PT-HMC is periodic swapping of states between adjacent temperature chains. After every 10 HMC iterations, swaps are proposed with acceptance probability:

$$\alpha_{i,j} = \min\left(1, \exp\left(\left(\frac{1}{T_i} - \frac{1}{T_j}\right)(\log p(\mathbf{y}|\boldsymbol{\theta}_j) - \log p(\mathbf{y}|\boldsymbol{\theta}_i))\right)\right) \tag{15}$$

These swaps allow high-temperature chains to facilitate exploration for the target chain ($T_1 = 1$), from which samples are retained for inference[2].

The step size, ϵ, directly controls HMC's exploration-rejection balance. We implement a geometric adaptation scheme during warmup: $\epsilon_{k+1} = \epsilon_k \cdot \exp(2(\alpha_k - 0.8))$ where α_k is the acceptance rate in segment k. This adaptation targets 80% as an empirically effective acceptance rate for moderate-dimensional targets. After warmup, ϵ is fixed to preserve detailed balance and ensure valid convergence to the posterior. While this approach promotes efficient sampling, it does not guarantee geometric ergodicity, particularly in the presence of multimodality or skewed posteriors.

2.4 BayesBag Extension

The BayesBag implementation leverages non-parametric bootstrap aggregation to enhance robustness under model misspecification [Bühlmann, 2014, Huggins and Miller, 2022]. By generating 50 parallel HMC chains on bootstrap resamples, this approach mitigates sensitivity to structural model errors through distributional averaging. The aggregated posterior:

$$\pi^*(\boldsymbol{\theta} \mid \mathbf{y}) := \frac{1}{B} \sum_{b=1}^{B} \pi(\boldsymbol{\theta} \mid \mathbf{y}^{(b)}) \tag{16}$$

[2]Tempered chains "heat up" the posterior landscape - imagine softening a rugged mountain range so the steep peaks melt into gentler hills. This makes it easier for the ball (from HMC dynamics) to roll freely across regions that would otherwise trap it.

where $\mathbf{y}^{(b)}$ represents the b-th bootstrapped dataset. This averaging procedure reduces variance in parameter estimates and helps partially counteract bias induced by model misspecification, while maintaining consistency. By marginalising over data perturbations, BayesBag improves the bias-variance trade-off, offering more robust predictions compared to standard bagging. Unlike model-based corrections, BayesBag leverages empirical variability in the data itself, yielding robustness through perturbation rather than regularisation.

3 Results

Table 1: Effective sample sizes (ESS) for each covariate of linear and quadratic PT-HMC models

Model	γ	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
Linear	1210	886	268	280	796	232	154	152	679	239	384
Quadratic	1419	1259	416	422	905	296	232	224	647	307	589

Table 1 reveals important insights about posterior geometry under model misspecification. The quadratic model achieves higher ESS for the key γ parameter (1419 vs. 1210), confirming improved mixing when the functional form is correctly specified. This pattern extends to most coefficients, with the quadratic model generally yielding 15-40% higher ESS values.

Parameter-specific mixing challenges persist in both models, with substantial ESS heterogeneity across coefficients. The lowest ESS values (β_5 & β_6) directly correspond to variables with engineered multicollinearity in the design matrix Z, creating ridge-like posterior geometries that challenge even gradient-informed HMC proposals. These geometric issues are visible in trace plots (Appendix C).

While the correctly specified quadratic model generally improves mixing for structural parameters by capturing the true data-generating process, it also restructures the posterior geometry. This redistribution of correlation patterns throughout parameter space highlights how model specification fundamentally alters the sampling efficiency landscape beyond simple improvements in fit metrics.

3.1 Model Comparison

Table 2 compares parameter estimates across all three approaches against true values.

Table 2: Posterior means and 95% credible intervals for parameters under different models

X	**True**	**Linear HMC**	**Quadratic HMC**	**BayesBag**
γ	0.25	1.84 (1.78, 1.89)	0.25 (0.25, 0.26)	1.84 (1.83, 1.84)
β_0	-0.20	-2.23 (-2.46, -1.98)	-0.27 (-0.43, -0.11)	-2.22 (-2.22, -2.21)
β_1	-0.16	-0.40 (-0.93, 0.17)	0.05 (-0.57, 0.56)	-0.37 (-0.42, -0.34)
β_2	-0.11	0.19 (-0.35, 0.73)	-0.22 (-0.71, 0.35)	0.16 (0.12, 0.20)
β_3	-0.07	-0.27 (-0.54, -0.00)	-0.22 (-0.48, 0.04)	-0.28 (-0.30, -0.27)
β_4	-0.02	0.37 (-0.75, 1.77)	-0.89 (-2.13, 0.29)	0.32 (0.16, 0.49)
β_5	0.02	0.09 (0.01, 0.17)	0.08 (0.00, 0.16)	0.12 (0.10, 0.13)
β_6	0.07	-1.16 (-2.66, 0.49)	-1.00 (-2.56, 0.50)	-1.60 (-1.95, -1.35)
β_7	0.11	0.34 (0.08, 0.59)	0.00 (-0.23, 0.25)	0.34 (0.31, 0.36)
β_8	0.16	0.11 (-0.04, 0.23)	0.25 (0.12, 0.38)	0.11 (0.10, 0.13)
β_9	0.20	0.20 (0.04, 0.36)	0.13 (-0.03, 0.29)	0.16 (0.13, 0.18)
MSE	—	3.22	0.97	3.21

Note: Values represent posterior means with 95% credible intervals in parentheses.

As expected, Table 2 shows the quadratic model's estimates are closest to the true values, particularly for the critical γ parameter. HMC sampling allows precise recovery of this parameter (0.25 vs. true 0.25) due to efficient exploration of the posterior. In contrast, the misspecified linear model and even the more robust BayesBag show substantial bias in the γ estimate (both 1.84), indicating improper functional form rather than sampling inadequacies. This bias illustrates a core insight of Bayesian misspecification theory, posteriors concentrate around pseudo-true parameters that minimise Kullback–Leibler divergence under the wrong model, not the data-generating truth [Berk, 1966].

More interestingly, despite its correct functional form, the quadratic model still shows mild bias in certain parameters failing to recover the parameters (e.g., β_6 = -1.00 vs. true 0.07), likely a result of the manufactured correlations and the skew-normal errors. This illustrates the layered nature of misspecification in practical modelling scenarios. However, MSE values decisively rank performance: quadratic (0.97), BayesBag (3.21), linear (3.22) - compelling evidence that proper model specification delivers substantially greater benefits than algorithmic refinements to misspecified models.

4 Discussion

While HMC is well known for its superior computational efficiency per effective sample compared to traditional MCMC methods, its behaviour under model misspecification merits careful examination. Table 3 summarises performance across experimental conditions.

Table 3: Computational Performance of MCMC Algorithms

Algorithm	**Runtime (s)**	**Memory (MB)**	**Med ESS**	**Min ESS**	**ESS/s**	**Accept %**
PT-HMC (Q)	101.81	1.69	421.7	223.6	2.20	76.2%
PT-HMC (L)	101.24	1.68	280.3	152.3	1.50	85.1%
BayesBag (L)	2449.06	84.00	310.5	152.7	0.06	80.5%
RWM (L)	4.30	1.68	8.7	3.7	0.86	37.4%
HMC (L)	48.74	1.67	73.6	13.4	0.28	86.7%

Note: Q = true quadratic; L = misspecified linear. ESS/s = minimum ESS per second. All algorithms used 20k iterations (8k warmup) on identical hardware (Intel i7-10700, 16GB RAM). BayesBag and PT-HMC used 15 cores. Standard HMC and adaptive RWM are included for comparison under misspecification.

The quadratic model achieves the highest efficiency (2.2 ESS/s), outperforming the linear model (1.5 ESS/s). Although the linear model has a higher acceptance rate (85.1%), its more correlated samples indicate that misspecification leads to unfavourable posterior geometry for gradient-based sampling, akin to the cold posterior effect in Bayesian deep learning. While tempering doubled runtime, its ability to overcome the complex geometry from multi-collinearity and bimodality greatly improved sampling performance over adaptive RWM and single-chain HMC.

BayesBag yields tighter credible bands, consistent with Huggins and Miller [2022], but at a steep computational cost (0.06 ESS/s). Its distributional averaging cannot address structural misspecification (omitted x_i^2 terms) so it fails to recover the true γ (1.84 vs. 0.25), highlighting that robustness to distributional uncertainty cannot compensate for incorrect model structure.

5 Conclusion

Cutting-edge bayesian computational advances like HMC and PT-HMC substantially improve sampling efficiency, but our results show they cannot overcome the fundamental limitations of structural model misspecification. Only the correctly specified quadratic model achieved accurate parameter recovery (MSE 0.97 vs. 3.22), underscoring that model adequacy is more critical than algorithmic sophistication. While PT-HMC enhanced mixing and BayesBag reduced variance, neither method corrected bias from an incorrect functional form. These findings highlight that robust inference ultimately depends on domain knowledge and thoughtful model specification, not just computational innovation. We acknowledge the sensitivity of HMC to step-size (ϵ) and number of steps (L) tuning, future work should explore even more adaptive and flexible algorithms, specifically NUTS, to further address the challenges posed by real-world data complexities.

References

Robert H Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

Michael Betancourt, Sam Livingstone, Simon Byrne, , and Mark Girolami. The geometric foundations of hamiltonian monte carlo. *Bernoulli*, pages 2257–2298, 2017.

Peter Bühlmann. Discussion of big bayes stories and bayesbag. *Statistical science*, 29(1):91–94, 2014.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

Charles J Geyer et al. Markov chain monte carlo maximum likelihood. In *Computing science and statistics: Proceedings of the 23rd Symposium on the Interface*, volume 156163. New York, 1991.

Jonathan H Huggins and Jeffrey W Miller. Reproducible model selection using bagged posteriors. *Bayesian analysis*, 18(1):79, 2022.

Jonathan H Huggins and Jeffrey W Miller. Reproducible parameter inference using bagged posteriors. *Electronic Journal of Statistics*, 18(1):1549–1585, 2024.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Appendix

A Skew-Normal Distribution Sampling Algorithm

Algorithm 1 Sampling from a Skew-Normal Distribution SN(m, σ, α)

for $i = 1$ to n **do**
  Draw $Z_i \sim \mathcal{N}(m, \sigma^2)$
  Set $b_i = 1$ with probability $\Phi\left(\alpha \frac{Z_i}{\sigma}\right)$, otherwise set $b_i = -1$
  Set $\varepsilon_i = b_i Z_i$
end for

B HMC with Parallel Tempering and Adaptive Step Size

Algorithm 2 Parallel Tempering HMC with Adaptive Step Size

Given θ_i^0 for chains $i = 1, \ldots, n$, temperatures T_i, initial ϵ, L:
Warmup Phase: For chunks $k = 1, \ldots, 10$:
for steps in chunk k **do**
  for each chain i **do**
    Sample $r^0 \sim \mathcal{N}(0, I)$.
    $H_{\text{current}} \leftarrow -\mathcal{L}(\theta_i | T_i) + \frac{1}{2} r^{0,T} r^0$.
    $\tilde{\theta}, \tilde{r} \leftarrow \text{Leapfrog}(\theta_i, r^0, \epsilon, L, T_i)$.
    $H_{\text{proposed}} \leftarrow -\mathcal{L}(\tilde{\theta} | T_i) + \frac{1}{2} \tilde{r}^T \tilde{r}$.
    Accept $\tilde{\theta}$ with probability $\min\{1, \exp(H_{\text{current}} - H_{\text{proposed}})\}$.
  end for
  Attempt swaps between adjacent chains with probabilities α_{ij}.
end for
$\epsilon \leftarrow \epsilon \cdot \exp(2(\text{accept\_rate} - 0.65))$.
Sampling Phase: Repeat above without adaptation.
return samples from coldest chain ($T_1 = 1$).

function LEAPFROG($\theta, r, \epsilon, L, T$)
  for $j = 1$ to L **do**
    Set $r \leftarrow r + (\epsilon/2) \nabla_\theta \mathcal{L}(\theta | T)$.
    Set $\theta \leftarrow \theta + \epsilon r$.
    Set $r \leftarrow r + (\epsilon/2) \nabla_\theta \mathcal{L}(\theta | T)$.
  end for
  return θ, r.
end function

C Trace Plots and Diagnostics

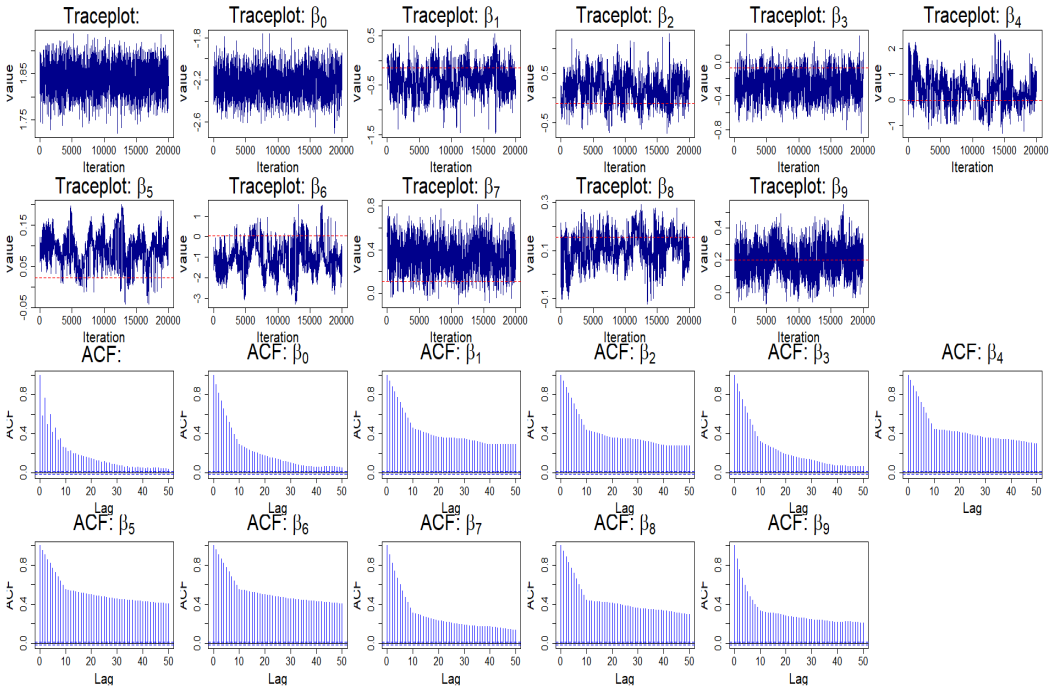


Figure 2: Trace Plots and Autocorrelation Function for *Misspecified* Linear Model (γ, β_0:β_9)

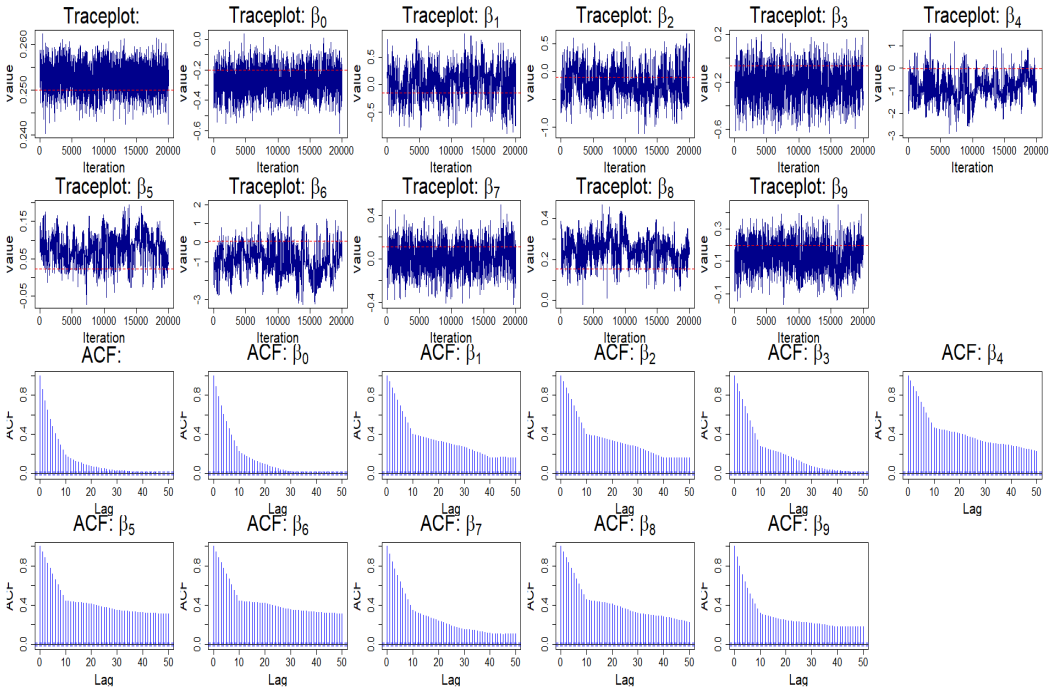


Figure 3: Trace Plots and Autocorrelation Function for *True* Quadratic Model (γ, β_0:β_9)