Metropolis-Hastings via Classification
Supplemental Materials

March 25, 2022

Contents

1 Notation

The following notation has been used throughout the manuscript. We employ the operator notation for expectation, e.g., $P_0 f = \int f dP_0$ and $\mathbb{P}_m^\theta f = \frac{1}{m}\sum_{i=1}^m f(X_i^\theta)$. The *$\varepsilon$-bracketing number* $N_{[]}(\varepsilon, \mathcal{F}, d)$ of a set $\mathcal{F}$ with respect to a premetric d is the minimal number of ε-brackets in d needed to cover $\mathcal{F}$.[1] The *δ-bracketing entropy integral* of $\mathcal{F}$ with respect to d is

$$J_{[]}(\delta, \mathcal{F}, d) := \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, d)} d\varepsilon.$$

We denote the usual Hellinger semi-metric for independent observations as

$$d_n^2(\theta, \theta') = \frac{1}{n}\sum_{i=1}^n \int (\sqrt{p_{\theta,i}} - \sqrt{p_{\theta',i}})^2 \mathrm{d}\mu_i.$$

Next, $K(p_{\theta_0}^{(n)}, p_\theta^{(n)}) = \sum_{i=1}^n K(p_{\theta_0,i}, p_{\theta,i})$ denotes the Kullback-Leibler divergence between product measures and $V_2(f, g) = \int f |\log(f/g)|^2 \mathrm{d}\mu$. Define $\langle a, b\rangle = \sum_{i=1}^d a_i b_i$ for $a, b \in \mathbb{R}^d$.

2 Proof of Theorem 4.1

The following lemma bounds the Kullback-Leibler divergence and variation by possibly non-diverging multiples of the Hellinger distance.[2] This can be used to derive sharper rates of posterior contraction in models with unbounded likelihood ratios [see also 6, p. 199 and Appendix B].

Lemma 2.1. *For probability measures P and P_0 such that $P_0(p_0/p) < \infty$, let $M := \inf_{c \geq 1} cP_0(\frac{p_0}{p} \mid \frac{p_0}{p} \geq [1 + \frac{1}{2c}]^2)$ where $P_0(\cdot \mid A) = 0$ if $P_0(A) = 0$. For $k \geq 2$, the following hold.*

(i) $-P_0 \log \frac{p}{p_0} \leq (3 + M) h(p, p_0)^2$.

(ii) $P_0 |\log \frac{p}{p_0}|^k \leq 2^{k-1}\Gamma(k+1)(2 + M) h(p, p_0)^2$.

(iii) $P_0 |\log \frac{p}{p_0} - P_0 \log \frac{p}{p_0}|^k \leq 2^{2k-1}\Gamma(k+1)(2 + M) h(p, p_0)^2$.

[1] A *premetric* on $\mathcal{F}$ is a function $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ such that $d(f, f) = 0$ and $d(f, g) = d(g, f) \geq 0$.

[2] Lemma 2.1 (iv) first appeared in Kaji et al. [10, Lemma 5]. We reproduce the proof here as it is used to prove other statements.

(iv) $\|\frac{1}{2}\log\frac{p}{p_0}\|_{P_0,B}^2 \le (2+M)h(p,p_0)^2$.

(v) $\|\frac{1}{4}(\log\frac{p}{p_0} - P_0\log\frac{p}{p_0})\|_{P_0,B}^2 \le (2+M)h(p,p_0)^2$.

Here, $\|f\|_{P,B} := \sqrt{2P(e^{|f|}-1-|f|)}$ *is the Bernstein "norm".*

Proof. (iv) Using $e^{|x|}-1-|x| \le (e^x-1)^2$ for $x \ge -\frac{1}{2}$ and $e^{|x|}-1-|x| < e^x - \frac{3}{2}$ for $x > \frac{1}{2}$,

$$\left\|\log\sqrt{\tfrac{p}{p_0}}\right\|_{P_0,B}^2 \le 2P_0\Big(\sqrt{\tfrac{p}{p_0}}-1\Big)^2\mathbb{1}\Big\{\tfrac{p}{p_0}\ge\tfrac{1}{e}\Big\} + 2P_0\Big(\sqrt{\tfrac{p_0}{p}}-\tfrac{3}{2}\Big)\mathbb{1}\Big\{\tfrac{p_0}{p}>e\Big\}.$$

The first term is bounded by $2h(p,p_0)^2$. For every $c \ge 1$,

$$\begin{aligned}P_0\Big(\sqrt{\tfrac{p_0}{p}}-\tfrac{3}{2}\Big)\mathbb{1}\Big\{\tfrac{p_0}{p}>e\Big\} &\le P_0\Big(\sqrt{\tfrac{p_0}{p}}-1-\tfrac{1}{2c}\Big)\mathbb{1}\Big\{\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big\}\\ &= P_0\Big(\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)\Big[P_0\Big(\sqrt{\tfrac{p_0}{p}}-1\;\Big|\;\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)-\tfrac{1}{2c}\Big].\end{aligned}$$

Since $x - \frac{1}{2c} \le \frac{c}{2}x^2$ for every x,

$$\begin{aligned}P_0\Big(\sqrt{\tfrac{p_0}{p}}-1\;\Big|\;\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)-\tfrac{1}{2c} &\le \tfrac{c}{2}\Big[P_0\Big(\sqrt{\tfrac{p_0}{p}}-1\;\Big|\;\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)\Big]^2\\ &\le \tfrac{c}{2}P_0\Big(\tfrac{p_0}{p}\;\Big|\;\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)P_0\Big(\Big[1-\sqrt{\tfrac{p}{p_0}}\Big]^2\;\Big|\;\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)\end{aligned}$$

by the Cauchy-Schwarz inequality. Then the result follows.

(i) Write $-P_0\log\frac{p}{p_0} = P_0(\frac{p}{p_0}-1-\log\frac{p}{p_0}) + P(p_0=0)$. With $x-1-\log x \le 3(\sqrt{x}-1)^2$ for $x > \frac{1}{3}$ and $\frac{1}{x}-1-\log\frac{1}{x} < 2(\sqrt{x}-\frac{3}{2})$ for $x \ge 3$,

$$P_0\Big(\tfrac{p}{p_0}-1-\log\tfrac{p}{p_0}\Big) \le 3P_0\Big(\sqrt{\tfrac{p}{p_0}}-1\Big)^2\mathbb{1}\Big\{\tfrac{p}{p_0}>\tfrac{1}{3}\Big\} + 2P_0\Big(\sqrt{\tfrac{p_0}{p}}-\tfrac{3}{2}\Big)\mathbb{1}\Big\{\tfrac{p_0}{p}\ge 3\Big\}.$$

The second term is bounded as above. The first term and $P(p_0=0) = \int(\sqrt{p}-\sqrt{p_0})^2\mathbb{1}\{p_0=0\}$ are collectively bounded by $3h(p,p_0)^2$.

(ii) Since $e^x - 1 - x \ge x^k/\Gamma(k+1)$ for $k \ge 2$ and $x \ge 0$,[3] $P_0|\log\frac{p}{p_0}|^k \le 2^{k-1}\Gamma(k+1)\|\frac{1}{2}\log\frac{p}{p_0}\|_{P_0,B}^2$. Then, apply (iv).

(iii) By the triangle and Jensen's inequalities, $P_0|\log\frac{p}{p_0} - P_0\log\frac{p}{p_0}|^k \le [(P_0|\log\frac{p}{p_0}|^k)^{1/k} + |P_0\log\frac{p}{p_0}|]^k \le 2^k P_0|\log\frac{p}{p_0}|^k$ for $k \ge 1$. Then, use (ii).

(v) By the convexity of $e^{|x|}-1-|x|$ and Jensen's inequality, $\|\frac{1}{4}(\log\frac{p}{p_0} - P_0\log\frac{p}{p_0})\|_{P_0,B}^2 \le \frac{1}{2}\|\frac{1}{2}\log\frac{p}{p_0}\|_{P_0,B}^2 + \frac{1}{2}\|P_0\frac{1}{2}\log\frac{p}{p_0}\|_{P_0,B}^2 \le \|\frac{1}{2}\log\frac{p}{p_0}\|_{P_0,B}^2$. With (iv) follows the result. □

[3] $\Gamma(k-1) \ge \int_x^\infty y^{k-2}e^{-y}dy \ge x^{k-2}e^{-x}$ implies $\frac{d^2}{dx^2}(e^x-1-x) \ge \frac{d^2}{dx^2}x^k/\Gamma(k+1)$.

Proof of Theorem 4.1. For $D \in \mathcal{D}^{\theta}_{n,\delta_n}$, write $\mathbb{P}_n(\log \frac{1-D}{1-D_\theta} - \log \frac{D}{D_\theta})$ as

$$P_0 \log \tfrac{1-D}{1-D_\theta} - P_0 \log \tfrac{D}{D_\theta} + (\mathbb{P}_n - P_0) \log \tfrac{1-D}{1-D_\theta} - (\mathbb{P}_n - P_0) \log \tfrac{D}{D_\theta}.$$

Since $\log(x) \le 2(\sqrt{x} - 1)$ for $x > 0$, we have

$$-2P_0\Big(\sqrt{\tfrac{D_\theta}{D}} - 1\Big) \le P_0 \log \tfrac{D}{D_\theta} \le 2P_0\Big(\sqrt{\tfrac{D}{D_\theta}} - 1\Big).$$

By the Cauchy-Schwarz inequality and Assumption 2,

$$P_0\Big|\sqrt{\tfrac{D}{D_\theta}} - 1\Big| \le \sqrt{P_0\Big(\sqrt{\tfrac{D}{D_\theta}} - 1\Big)^2} = h_\theta(D, D_\theta) \le \delta_n,$$
$$P_0\Big|\sqrt{\tfrac{D_\theta}{D}} - 1\Big| \le \sqrt{P_0 \tfrac{D_\theta}{D}} \sqrt{P_0\Big(1 - \sqrt{\tfrac{D}{D_\theta}}\Big)^2} \le \sqrt{M}\delta_n.$$

Therefore, $|P_0 \log \frac{D}{D_\theta}| \le 2(1 \vee \sqrt{M})\delta_n$. Next, let $W := \sqrt{\frac{1-D}{1-D_\theta}} - 1$ and define a function R by $\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{2}x^2 R(x)$, which implies R is increasing and $R(x) < 1$ for $x > -1$, and $R(x) = O(x)$ as $x \to 0$. With this, write

$$P_0 \log \tfrac{1-D}{1-D_\theta} = 2P_0 W - P_0 W^2 + P_0 W^2 R(W).$$

By the Cauchy-Schwarz inequality,

$$P_0|W| \le \sqrt{P_0 \tfrac{p_0}{p_\theta}} \cdot h_\theta(1-D, 1-D_\theta) \le \sqrt{M}\delta_n,$$
$$P_0 W^2 \le \sqrt{(P_0 + P_\theta)\Big(\tfrac{p_0}{p_\theta}\Big)^2 (\sqrt{1-D} - \sqrt{1-D_\theta})^2} \cdot h_\theta(1-D, 1-D_\theta).$$

Since D and D_θ are bounded by 0 and 1,

$$(P_0 + P_\theta)\Big(\tfrac{p_0}{p_\theta}\Big)^2 (\sqrt{1-D} - \sqrt{1-D_\theta})^2 \le P_0\Big(\tfrac{p_0}{p_\theta}\Big)^2 + P_0 \tfrac{p_0}{p_\theta} \le 2M.$$

Therefore, $P_0 W^2 \le \sqrt{2M}\delta_n$. Next, the residual is bounded as

$$\begin{aligned} |P_0 W^2 R(W)| &\le P_0 W^2 |R(W)| \mathbb{1}\{W \le -\tfrac{1}{5}\} + P_0 W^2 |R(W)| \mathbb{1}\{W > -\tfrac{1}{5}\} \\ &\le P_0(-R(W)\mathbb{1}\{W \le -\tfrac{1}{5}\}) + P_0 W^2 |R(-\tfrac{1}{5}) \vee R(W)|, \end{aligned}$$

where the second inequality uses $W \ge -1$ and R increasing. Since $R < 1$ and $P_0 W^2 \le \sqrt{2M}\delta_n$, the second term is also bounded by $\sqrt{2M}\delta_n$. With $0 < -R(x) < -2\log(1+x)$

for $x \le -\frac{1}{5}$, the first term is bounded by

$$P_0\Big(\log \tfrac{1-D_\theta}{1-D}\mathbb{1}\{W \le -\tfrac{1}{5}\}\Big) = P_0\Big(\tfrac{1-D}{1-D_\theta}\log \tfrac{1-D_\theta}{1-D}\cdot \tfrac{1-D_\theta}{1-D}\mathbb{1}\{W \le -\tfrac{1}{5}\}\Big)$$
$$\le \sup_{\sqrt{x}-1\le -1/5} |x\log\tfrac{1}{x}| \cdot P_0\Big(\tfrac{1-D_\theta}{1-D}\mathbb{1}\{W \le -\tfrac{1}{5}\}\Big).$$

The supremum is $1/e$. The second term is bounded by $P_0(W \le -\frac{1}{5})P_0(\frac{1-D_\theta}{1-D} \mid \frac{1-D_\theta}{1-D} \ge \frac{25}{16}) \le P_0(W \le -\frac{1}{5})M$ by Assumption 2. By Markov's inequality, $P_0(W \le -\frac{1}{5}) \le 25P_0W^2 \le 25\sqrt{2M}\delta_n$. Thus, $|P_0W^2R(W)| \le (1+25M/e)\sqrt{2M}\delta_n$. Altogether, we have $|P_0\log\frac{1-D}{1-D_\theta}| \le (\sqrt{2}+2+25M/e)\sqrt{2M}\delta_n$.

Next, we bound $\mathbb{E}^*\sup_{D\in\mathcal{D}^\theta_{n,\delta_n}}|\sqrt{n}(\mathbb{P}_n-P_0)\log\frac{D}{D_\theta}|$. Under Assumption 2, an analogous argument as Lemma 2.1 (iv) yields

$$\left\|\tfrac{1}{2}\log\tfrac{D}{D_\theta}\right\|^2_{P_0,B} \le 2(1+M)h_\theta(D,D_\theta)^2 \le 2(1+M)\delta_n^2.$$

By van der Vaart and Wellner [21, Lemma 3.4.3], we have

$$\mathbb{E}^*\sup_{D\in\mathcal{D}^\theta_{n,\delta_n}}\left|\sqrt{n}(\mathbb{P}_n-P_0)\log\tfrac{D}{D_\theta}\right| \lesssim J\Big(1+\tfrac{J}{\delta_n^2\sqrt{n}}\Big)$$

for $J := J_{[]}(\delta_n, \{\log\frac{D}{D_\theta} : D\in\mathcal{D}^\theta_{n,\delta_n}\}, \|\cdot\|_{P_0,B})$. Note that a δ_n-bracket in $\mathcal{D}^\theta_{n,\delta_n}$ induces a $C\delta_n$-bracket in $\{\log\frac{D}{D_\theta}\}$ for some constant C since

$$\left\|\log\tfrac{u}{D_\theta}-\log\tfrac{\ell}{D_\theta}\right\|^2_{P_0,B} \le 4P_0\Big(\sqrt{\tfrac{u}{\ell}}-1\Big)^2 = O(d_\theta(u,\ell)^2)$$

by Assumption 2. Therefore, $J \le J_{[]}(\delta_n, \mathcal{D}^\theta_{n,\delta_n}, d_\theta)$ and hence $J(1+\frac{J}{\delta_n^2\sqrt{n}}) \lesssim \delta_n^2\sqrt{n}$ by Assumption 1.

Finally, we bound $\mathbb{E}^*\sup_{D\in\mathcal{D}^\theta_{n,\delta_n}}|\sqrt{n}(\mathbb{P}_n-P_0)\log\frac{1-D}{1-D_\theta}|$. As in Lemma 2.1 (iv), we obtain $\rho^2 := \left\|\frac{1}{2}\log\frac{1-D}{1-D_\theta}\right\|^2_{P_0,B} \le 2(1+M)P_0W^2 \le 2(1+M)\sqrt{2M}\delta_n$. Therefore, by van der Vaart and Wellner [21, Lemma 3.4.3], we have $\mathbb{E}^*\sup_{D\in\mathcal{D}^\theta_{n,\delta_n}}|\sqrt{n}(\mathbb{P}_n-P_0)\log\frac{1-D}{1-D_\theta}| \lesssim J\Big(1+\frac{J}{\delta_n^2\sqrt{n}}\Big)$ for $J = J_{[]}(\rho, \{\log\frac{1-D}{1-D_\theta} : D\in\mathcal{D}^\theta_{n,\delta_n}\}, \|\cdot\|_{P_0,B})$. With a δ_n-bracket in $\mathcal{D}^\theta_{n,\delta_n}$, Assumption 2 implies

$$\left\|\log\tfrac{1-\ell}{1-D_\theta}-\log\tfrac{1-u}{1-D_\theta}\right\|^2_{P_0,B} \le 4P_0\Big(\sqrt{\tfrac{1-\ell}{1-u}}-1\Big)^2 = O(\delta_n).$$

Therefore, the expectation of the supremum is of order $O(\delta_n\sqrt{n})$. □

3 Proof of Theorem 4.2

Let h_n be a bounded sequence and denote $\theta_n := \theta_0 + \frac{h_n}{\sqrt{n}}$ and $W_n := \sqrt{\hat{p}_{\theta_n}/\hat{p}_{\theta_0}} - 1$. Define R by $\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{2}x^2 R(x)$ for $R(x) = O(x)$. Then,

$$n\mathbb{P}_n \log \tfrac{\hat{p}_{\theta_n}}{\hat{p}_{\theta_0}} = 2n\mathbb{P}_n W_n - n\mathbb{P}_n W_n^2 + n\mathbb{P}_n W_n^2 R(W_n).$$

By Assumption 4 (ii) and $P_{\theta_0}\dot{\ell}_{\theta_0} = 0$,

$$2n\mathbb{P}_n W_n - n\mathbb{P}_n W_n^2 = 2nP_{\theta_0} W_n + \sqrt{n}\mathbb{P}_n h_n' \dot{\ell}_{\theta_0} - nP_{\theta_0} W_n^2 + o_P(1).$$

By Assumption 4 (i),

$$nP_{\theta_0} W_n^2 = \tfrac{1}{4} P_{\theta_0} h_n' \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}' h_n + o_P(1) = \tfrac{1}{4} h_n' I_{\theta_0} h_n + o_P(1).$$

Also, since $P_{\theta_0}\dot{\ell}_{\theta_0} = 0$,

$$\begin{aligned} 2nP_{\theta_0} W_n &= 2n\hat{P}_{\theta_0} W_n + 2n(P_{\theta_0} - \hat{P}_{\theta_0}) W_n \\ &= -n\int \left(\sqrt{\hat{p}_{\theta_n}} - \sqrt{\hat{p}_{\theta_0}}\right)^2 + n(c_{\theta_n} - c_{\theta_0}) - \sqrt{n}\hat{P}_{\theta_0} h_n' \dot{\ell}_{\theta_0} \\ &\quad + 2n\int \left(\sqrt{p_{\theta_0}} - \sqrt{\hat{p}_{\theta_0}}\right)\left(\sqrt{p_{\theta_0}} + \sqrt{\hat{p}_{\theta_0}}\right)\left(W_n - \tfrac{h_n'\dot{\ell}_{\theta_0}}{2\sqrt{n}}\right). \end{aligned}$$

By Assumption 4 (i), $n\int(\sqrt{\hat{p}_{\theta_n}} - \sqrt{\hat{p}_{\theta_0}})^2 = \frac{1}{4}\hat{P}_{\theta_0} h_n' \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}' h_n + o_P(1) = \frac{1}{4} h_n' I_{\theta_0} h_n + o_P(1)$. By the Cauchy-Schwarz inequality,

$$\begin{aligned} &\left|\int \left(\sqrt{p_{\theta_0}} - \sqrt{\hat{p}_{\theta_0}}\right)\left(\sqrt{p_{\theta_0}} + \sqrt{\hat{p}_{\theta_0}}\right)\left(W_n - \tfrac{h_n'\dot{\ell}_{\theta_0}}{2\sqrt{n}}\right)\right| \\ &\qquad \le \left[\int \left(\sqrt{p_{\theta_0}} - \sqrt{\hat{p}_{\theta_0}}\right)^2 \int \left(\sqrt{p_{\theta_0}} + \sqrt{\hat{p}_{\theta_0}}\right)^2 \left(W_n - \tfrac{h_n'\dot{\ell}_{\theta_0}}{2\sqrt{n}}\right)^2\right]^{1/2}, \end{aligned}$$

which is $O_P(\delta_n n^{-3/4}) = o_P(n^{-1})$ under Assumption 4 (i) and $\delta_n = o(n^{-1/4})$.

Since $|n\mathbb{P}_n W_n^2 R(W_n)| \le |n\mathbb{P}_n W_n^2| \max_{1\le i\le n} |R(W_n(X_i))|$ and $n\mathbb{P}_n W_n^2$ "converges" to $nP_{\theta_0} W_n^2 = O_P(1)$ by Assumption 4 (ii), it remains to show that the maximum is $o_P(1)$. Write $V_n := W_n - \frac{h_n'\dot{\ell}_{\theta_0}}{2\sqrt{n}}$. Then,

$$\max_i |W_n(X_i)| \le \max_i \left|\tfrac{1}{2\sqrt{n}} h_n' \dot{\ell}_{\theta_0}(X_i)\right| + \max_i |V_n(X_i)|.$$

By Markov's inequality,

$$P\Big(\max_{1\le i\le n}\Big|\tfrac{1}{\sqrt{n}}h_n'\dot\ell_{\theta_0}(X_i)\Big|>\varepsilon\Big)\le nP\Big(\Big|\tfrac{1}{\sqrt{n}}h_n'\dot\ell_{\theta_0}(X_i)\Big|>\varepsilon\Big)$$
$$\le\varepsilon^{-2}P_{\theta_0}((h_n'\dot\ell_{\theta_0})^2\mathbb{1}\{(h_n'\dot\ell_{\theta_0})^2>n\varepsilon^2\}),$$

which converges to zero as $n\to\infty$ for every $\varepsilon>0$. Thus, $\max_i\big|\frac{1}{\sqrt{n}}h_n'\dot\ell_{\theta_0}(X_i)\big|$ converges to zero in probability. Since Assumption 4 (ii) and (i) imply that $n\mathbb{P}_nV_n^2=nP_{\theta_0}V_n^2+o_P(1)=o_P(1)$, we have $\max_i V_n^2(X_i)=o_P(1)$ and hence $\max_i|V_n(X_i)|=o_P(1)$. Conclude that $\max_i|W_n(X_i)|$ converges to zero in probability and so does $\max_i|R(W_n(X_i))|$.

4 Proof of Theorem 4.4

We will prove Theorem 4.4 under weaker assumptions. In particular, we slightly relax Assumption 4.4 by considering the aggregate behavior of $u_\theta(X^{(n)})$ around θ_0 with respect to the prior $\Pi_n(\cdot)$. Instead, we assume

$$P_{\theta_0}^{(n)}\left(I_n(\Pi_n,X^{(n)},\varepsilon_n)\le \mathrm{e}^{-\widetilde{C}_n n\varepsilon_n^2}\right)=o(1)$$

where

$$I_n(\Pi_n,X^{(n)},\epsilon)=\int_{B_n(\theta_0,\epsilon)}\mathrm{e}^{u_\theta(X^{(n)})}\mathrm{d}\Pi_n(\theta)\tag{4.1}$$

and, at the same time,

$$P_{\theta_0}^{(n)}\left[\sup_{\Theta_n^c\cup d_n(\theta,\theta_0)>\epsilon}|u_\theta(X^{(n)})|>\widetilde{C}_n n\varepsilon_n^2\right]=o(1)$$

for any $\epsilon>\varepsilon_n$. Assumption (4.5) is not needed if one is only interested in the concentration inside Θ_n. Alternatively, we could also replace Assumption (4.4) with the following condition to lower-bound the denominator in (4.3)

$$\sup_{\theta\in B_n(\theta_0,\varepsilon_n)}P_{\theta_0}^{(n)}\left[\ln(p_\theta^{(n)}/p_{\theta_0}^{(n)})+u_\theta<-n\varepsilon_n^2\right]=o(n\varepsilon_n^2).$$

Instead of relying on the existence of exponential tests (through Lemma 9 in [5]), we could then directly assume that for any $\epsilon>\varepsilon_n$ and for all $\theta\in\Theta_n$ such that $d(\theta,\theta_0)>j\epsilon$ for any $j\in\mathbb{N}$ there exists a test $\phi_n(\theta)$ satisfying

$$P_{\theta_0}^{(n)}\phi_n\lesssim\mathrm{e}^{-n\epsilon^2/2}\quad\text{and}\quad\int_{\mathcal{X}}(1-\phi_n)p_\theta^{(n)}\mathrm{e}^{u_\theta}\le\mathrm{e}^{-j^2n\epsilon^2/2}.$$

We will use the following Lemma (an analogue of Lemma 10 [5]).

Lemma 4.1. *Recall the definition $I_n(\Pi_n, X^{(n)}, \epsilon)$ in (4.1) and define $q_\theta^{(n)} = p_\theta^{(n)}/p_{\theta_0}^{(n)} \mathrm{e}^{u_\theta}$. Then we have for any $C, \varepsilon > 0$*

$$P_{\theta_0}^{(n)} \left(\int_{B(\theta_0,\varepsilon)} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta) \leq \mathrm{e}^{-(1+C)n\varepsilon^2} \times I_n(\Pi_n, X^{(n)}, \epsilon) \right) \leq \frac{1}{C^2 n\varepsilon^2}.$$

Proof. Define a changed prior measure $\Pi_n^\star(\cdot)$ through $\mathrm{d}\Pi_n^\star(\theta) = \frac{\mathrm{e}^{u_\theta(X^{(n)})}}{\int \mathrm{e}^{u_\theta(X^{(n)})}\mathrm{d}\theta} \mathrm{d}\Pi_n(\theta)$. Lemma 10 of [5] then yields

$$\begin{aligned} &P_{\theta_0}^{(n)} \left(\int_{B(\theta_0,\varepsilon)} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta) \leq \mathrm{e}^{-(1+C)n\varepsilon^2} I_n(\Pi_n, X^{(n)}, \epsilon) \right) \\ &= P_{\theta_0}^{(n)} \left(\int_{B(\theta_0,\varepsilon)} p_\theta^{(n)}/p_{\theta_0}^{(n)} \mathrm{d}\Pi_n^\star(\theta) \leq \Pi_n^\star(B(\theta_0,\varepsilon)) \mathrm{e}^{-(1+C)n\varepsilon^2} \right) \leq \frac{1}{C^2 n\epsilon^2}. \end{aligned}$$

□

Recall the definition $I_n(\Pi_n, X^{(n)}, \varepsilon_n) = \int_{B(\theta_0,\varepsilon_n)} \mathrm{e}^{u_\theta(X^{(n)})} \mathrm{d}\Pi_n(\theta)$ and define an event

$$\mathcal{A}_n = \left\{ X^{(n)} : \int_{B(\theta_0,\varepsilon_n)} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta) > \mathrm{e}^{-2n\varepsilon_n^2} I_n(\Pi_n, X^{(n)}, \varepsilon_n) \right\}$$

where $q_\theta^{(n)} = p_\theta^{(n)}/p_{\theta_0}^{(n)} \mathrm{e}^{u_\theta}$. From our assumptions, there exists a sequence $\widetilde{C}_n > 0$ such that the complement of the set

$$\mathcal{B}_n = \left\{ X^{(n)} : I_n(\Pi_n, X^{(n)}, \varepsilon_n) > \mathrm{e}^{-\widetilde{C}_n n\varepsilon_n^2} \text{ and } \sup_{\Theta_n^c \cup d_n(\theta,\theta_0) > \varepsilon_n} |u_\theta(X^{(n)})| \leq \widetilde{C}_n n\varepsilon_n^2 \right\}$$

has a vanishing probability. Lemma 4.1 then yields $P_{\theta_0}^{(n)}[\mathcal{A}_n^c \cup \mathcal{B}_n^c] = o(1)$ as $n \to \infty$. The following calculations are thus conditional on the set $\mathcal{A}_n \cap \mathcal{B}_n$. On this set, we can lower-bound the denominator of (4.3) as follows

$$\int_\Theta q_\theta^{(n)} \mathrm{d}\Pi_n(\theta) > \int_{B(\theta_0,\varepsilon_n)} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta) > \mathrm{e}^{-2n\varepsilon_n^2} I_n(\Pi_n, X^{(n)}, \varepsilon_n) \geq \mathrm{e}^{-(2+\widetilde{C}_n)n\varepsilon_n^2}.$$

We first show that $P_{\theta_0}^{(n)}[\Pi_n^\star(\Theta \backslash \Theta_n \mid X^{(n)})] = o(1)$ as $n \to \infty$. On the set $\mathcal{A}_n \cap \mathcal{B}_n$ we have from (4.5) and from the Fubini's theorem

$$\begin{aligned} P_{\theta_0}^{(n)} \left[\Pi_n^\star(\Theta \backslash \Theta_n \mid X^{(n)}) \right] &= P_{\theta_0}^{(n)} \left[\frac{\int_{\Theta \backslash \Theta_n} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta)}{\int_\Theta q_\theta^{(n)} \mathrm{d}\Pi_n(\theta)} \right] \leq \mathrm{e}^{2n\varepsilon_n^2} \frac{\Pi_n^\star(\Theta \backslash \Theta_n)}{\Pi_n^\star(B_n(\theta_0, \varepsilon_n))} \\ &= \mathrm{e}^{2(1+\widetilde{C}_n)n\varepsilon_n^2} \frac{\Pi_n(\Theta \backslash \Theta_n)}{\Pi_n(B_n(\theta_0, \varepsilon_n))} = o(1). \end{aligned}$$

For some $J > 0$ (to be determined later) we define the complement of the ball around the truth as a union of shells

$$U_n = \{\theta \in \Theta_n : d_n(\theta, \theta_0) > MJ\varepsilon_n\} = \bigcup_{j \geq J} \Theta_{n,j}$$

where each shell equals

$$\Theta_{n,j} = \{\theta \in \Theta_n : Mj\varepsilon_n < d_n(\theta, \theta_0) \leq M(j+1)\varepsilon_n\}.$$

We now invoke the local entropy Assumption (3.2) in [5] which guarantees (according to Lemma 9 in [5]) that there exist tests ϕ_n (for each n) such that

$$P_{\theta_0}^{(n)}\phi_n \lesssim \mathrm{e}^{n\varepsilon_n^2 - nM^2\varepsilon_n/2} \quad \text{and} \quad P_{\theta}^{(n)}(1-\phi_n) \leq \mathrm{e}^{-nM^2\varepsilon_n^2 j^2/2} \tag{4.2}$$

for all $\theta \in \Theta_n$ such that $d_n(\theta, \theta_0) > M\varepsilon_n j$ and for every $j \in \mathbb{N}\backslash\{0\}$ and $M > 0$. One can then write

$$\begin{aligned} P_{\theta_0}^{(n)}\Pi\left(\theta \in \Theta : d(\theta, \theta_0) > MJ\varepsilon_n \,|\, X^{(n)}\right) &\leq P_{\theta_0}^{(n)}\Pi(\Theta_n^c \,|\, X^{(n)}) + P_{\theta_0}^{(n)}\phi_n + P_{\theta_0}^{(n)}(\mathcal{A}_n^c) + P_{\theta_0}^{(n)}(\mathcal{B}_n^c) \\ &\quad + \sum_{j \geq J} P_{\theta_0}^{(n)}[\Pi(\Theta_{n,j} \,|\, X^{(n)})(1-\phi_n)\mathbb{I}(\mathcal{A}_n \cap \mathcal{B}_n)] \end{aligned}$$

For the last term above, we recall that $\Pi(\Theta_{n,j} \,|\, X^{(n)}) = \frac{\int_{\Theta_{n,j}} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta)}{\int_\Theta q_\theta^{(n)} \mathrm{d}\Pi_n(\theta)}$. We bound the denominator as before. Regarding the numerator, on the event $\mathcal{B}_n$ we have from (4.2) and from the Fubini's theorem

$$P_{\theta_0}^{(n)} \int_{\Theta_{n,j}} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta)(1-\phi_n) \leq \mathrm{e}^{-nM^2\varepsilon_n^2 j^2/2 + \widetilde{C}_n n\varepsilon_n^2} \Pi_n(\Theta_{n,j}) \tag{4.3}$$

Putting the pieces together, we obtain

$$P_{\theta_0}^{(n)}[\Pi(\Theta_{n,j} \,|\, X^{(n)})(1-\phi_n)\mathrm{I}(\mathcal{A}_n \cap \mathcal{B}_n)] \leq \mathrm{e}^{-nM^2\varepsilon_n^2 j^2/2 + 2(1+\widetilde{C}_n)n\varepsilon_n^2} \frac{\Pi_n(\Theta_{n,j})}{\Pi_n[B_n(\theta_0, \varepsilon_n)]}.$$

Assumption (3.4) of [5] writes as

$$\frac{\Pi_n(\Theta_{n,j})}{\Pi_n[B_n(\theta_0, \varepsilon_n)]} \leq \mathrm{e}^{nM^2\varepsilon_n^2 j^2/4} \tag{4.4}$$

which yields

$$P_{\theta_0}^{(n)}\Pi\left(\theta \in \Theta : d(\theta, \theta_0) > MJ\varepsilon_n \,|\, X^{(n)}\right) \leq o(1) + \sum_{j \geq J} \mathrm{e}^{-n\varepsilon_n^2(M^2 j^2/4 - 2 - 2\widetilde{C}_n)}.$$

The right hand side converges to zero as long as $J = J_n \to \infty$ fast enough so that $\widetilde{C}_n = o(J_n)$ and $n\varepsilon_n^2$ is bounded away from zero.

5 Posterior Concentration Rate: Misspecification Lens

The following Theorem 5.1 quantifies concentration in terms of a KL neighborhoods around $\widetilde{P}_{\theta^*}^{(n)}$ defined as $B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)}) = \left\{\widetilde{P}_{\theta}^{(n)} \in \widetilde{\mathcal{P}}^{(n)} : K(\theta^*, \theta_0) \leq n\epsilon^2, V(\theta^*, \theta_0) \leq n\epsilon^2\right\}$,where $K(\theta^*, \theta_0) \equiv P_{\theta_0}^{(n)} \log \frac{\widetilde{p}_{\theta^*}^{(n)}}{\widetilde{p}_{\theta}^{(n)}}$ and $V(\theta^*, \theta_0) = P_{\theta_0}^{(n)} \left|\log \frac{\widetilde{p}_{\theta^*}^{(n)}}{\widetilde{p}_{\theta}^{(n)}} - K(\theta^*, \theta_0)\right|^2$.

Theorem 5.1. *Denote with $Q_{\theta}^{(n)}$ a measure defined through $\mathrm{d}Q_{\theta}^{(n)} = \frac{p_{\theta_0}^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \mathrm{d}P_{\theta}^{(n)}$ and let $d(\cdot, \cdot)$ be a semi-metric on $\mathcal{P}^{(n)}$. Suppose that there exists a sequence $\varepsilon_n > 0$ satisfying $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$ such that for every $\epsilon > \varepsilon_n$ there exists a test ϕ_n (depending on ϵ) such that for every $J \in \mathbb{N}_0$*

$$P_{\theta_0}^{(n)} \phi_n \lesssim \mathrm{e}^{-n\epsilon^2/4} \quad \textit{and} \quad \sup_{\widetilde{P}_{\theta}^{(n)} : d(\widetilde{P}_{\theta}^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) > J\epsilon} Q_{\theta}^{(n)}(1 - \phi_n) \leq \mathrm{e}^{-nJ^2\epsilon^2/4}. \tag{5.1}$$

Let $B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})$ be as before and let $\widetilde{\Pi}_n(\theta)$ be a prior distribution with a density $\widetilde{\pi}(\theta) \propto C_\theta \pi(\theta)$. Assume that there exists a constant $L > 0$ such that, for all n and $j \in \mathbb{N}$,

$$\frac{\widetilde{\Pi}_n\left(\theta \in \Theta : j\varepsilon_n < d(\widetilde{P}_{\theta}^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) \leq (j+1)\varepsilon_n\right)}{\widetilde{\Pi}_n\left(B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})\right)} \leq \mathrm{e}^{n\varepsilon_n^2 j^2/8}. \tag{5.2}$$

Then for every sufficiently large constant M, as $n \to \infty$,

$$P_{\theta_0}^{(n)} \Pi_n^\star\left(\widetilde{P}_{\theta}^{(n)} : d(\widetilde{P}_{\theta}^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) \geq M\varepsilon_n \,|\, X^{(n)}\right) \to 0. \tag{5.3}$$

Proof. We define the event

$$\mathcal{A} = \left\{X^{(n)} \in \mathcal{X} : \int \frac{\widetilde{p}_{\theta}^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \mathrm{d}\widetilde{\Pi}_n(\theta) > \mathrm{e}^{-(1+C)n\epsilon^2} \widetilde{\Pi}_n[B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})]\right\}.$$

The following lemma shows that $P_{\theta_0}^{(n)}[\mathcal{A}^c] = o(1)$ as $n \to \infty$.

Lemma 5.2. *For $k \geq 2$, every $\epsilon > 0$ and a prior measure $\widetilde{\Pi}_n(\theta)$ on Θ, we have for every $C > 0$*

$$P_{\theta_0}^{(n)}\left(\int \frac{\widetilde{p}_{\theta}^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \mathrm{d}\widetilde{\Pi}_n(\theta) \leq \mathrm{e}^{-(1+C)n\epsilon^2} \widetilde{\Pi}_n[B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})]\right) \leq \frac{1}{C^2 n\epsilon^2}.$$

Proof. This follows directly from Lemma 10 in [5].

We now define $U_n(\epsilon) = \Pi_n(\theta \in \Theta : d(\widetilde{P}_\theta^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) > \epsilon \,|\, X^{(n)})$. For every $n \geq 1$ and $J \in \mathbb{N}\backslash\{0\}$, we can decompose

$$\begin{aligned} P_{\theta_0}^{(n)} U_n(JM\varepsilon_n) =& P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)\phi_n] + P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)(1-\phi_n)\mathbb{I}(\mathcal{A}^c)] \\ &+ P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)(1-\phi_n)\mathbb{I}(\mathcal{A})]. \end{aligned}$$

The first term is bounded (from the assumption (5.1)) as

$$P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)\phi_n] \leq P_{\theta_0}^{(n)}\phi_n \lesssim \mathrm{e}^{-n\varepsilon_n^2 J^2 M^2}.$$

The second term can be bounded by $P_{\theta_0}^{(n)}[\mathbb{I}(\mathcal{A}^c)] \leq \frac{1}{C^2 J^2 M^2 n\varepsilon_n^2}$ which converges to zero as $n\varepsilon_n^2 \to \infty$. The last term satisfies

$$\begin{aligned} P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)(1-\phi_n)\mathbb{I}(\mathcal{A})] &= P_{\theta_0}^{(n)}\left[(1-\phi_n)\mathbb{I}(\mathcal{A})\frac{\int_{\theta: d(\widetilde{P}_\theta^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) > JM\varepsilon_n} \frac{\widetilde{p}_\theta^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \widetilde{\Pi}_n(\theta)\mathrm{d}\theta}{\int_\Theta \frac{\widetilde{p}_\theta^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \widetilde{\Pi}_n(\theta)\mathrm{d}\theta}\right] \\ &\leq \frac{\mathrm{e}^{(1+C)n\epsilon^2}}{\widetilde{\Pi}_n[B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})]} \int_{\theta: d(\widetilde{P}_\theta^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) > JM\varepsilon_n} \left[\int_{\mathcal{X}} (1-\phi_n) p_{\theta_0}^{(n)} \frac{\widetilde{p}_\theta^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}}\right] \widetilde{\Pi}_n(\theta)\mathrm{d}\theta \\ &\leq \frac{\mathrm{e}^{(1+C)n\epsilon^2}}{\widetilde{\Pi}_n[B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})]} \sum_{j \geq J} \int_{U_{n,j}} Q_\theta^{(n)}(1-\phi_n)\mathrm{d}\widetilde{\Pi}_n(\theta), \end{aligned}$$

where $U_{n,j} = \{\theta : jM\varepsilon_n < d(\widetilde{P}_\theta^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) \leq (j+1)M\varepsilon_n)\}$. The tests (from the assumption (5.1)) satisfy $Q_\theta^{(n)}(1-\phi_n) \leq \mathrm{e}^{-nj^2M^2\varepsilon_n^2/4}$ uniformly on $U_{n,j}$. Then we find (using the assumption (5.2))

$$P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)(1-\phi_n)\mathbb{I}(\mathcal{A})] \leq \mathrm{e}^{(1+C)n\varepsilon_n^2} \sum_{j \geq J} \mathrm{e}^{-nj^2M^2\varepsilon_n^2/4 + nj^2M^2\varepsilon_n^2/8}.$$

The sum converges to zero when $n\varepsilon_n^2$ is bounded away from zero and $J \to \infty$. □

Remark 1. *For iid data, [11] introduce a condition involving entropy numbers under misspecification which implies the existence of exponential tests for a testing problem that involves non-probability measures. Since we have a non-iid situation, we assumed the existence of tests directly.*

Remark 2. *(Friendlier Metrics) In parametric models indexed by θ in a metric space (Θ, d), it is more natural to characterize the posterior concentration in terms of $d(\cdot,\cdot)$*

rather than the Kullback-Leibler divergence[4]*. Section 5 of [11] clarifies how Theorem 5.1 can be reformulated in terms of some metric $d(\cdot,\cdot)$ on Θ.*

6 Normal Location-Scale Example

Let $X_i \sim P_0 = N(0,1)$ and $P_\theta = N(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$ are the unknown parameters and $\theta_0 = (0,1)$ are the true values. This model satisfies Assumption 3 with the score $\dot{\ell}_{\theta_0}(x) = \begin{bmatrix} x \\ (x^2-1)/2 \end{bmatrix}$ and the Fisher information matrix $I_{\theta_0} = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$. The oracle discriminator of P_0 from P_θ is $D_\theta(x) = \left[1 + \exp\left(-\frac{1}{2}\log\sigma^2 + \frac{x^2}{2} - \frac{(x-\mu)^2}{2\sigma^2}\right)\right]^{-1}$. Let us use the logistic regression using regressors $(1, x, x^2)$ to estimate D_θ, i.e.,

$$D_\theta(x) = [1 + \exp(-\beta_0 - \beta_1 x - \beta_2 x^2)]^{-1}.$$

Thus, the true parameter for the logistic regression is $\beta = (\beta_0, \beta_1, \beta_2) = \left(\frac{1}{2}\log\sigma^2 + \frac{\mu^2}{2\sigma^2}, -\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2} - \frac{1}{2}\right)$. Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ be the estimator of β. Then,

$$\hat{p}_\theta(x) = \frac{\exp\left(-\frac{x^2}{2} - \hat{\beta}_0 - \hat{\beta}_1 x - \hat{\beta}_2 x^2\right)}{\sqrt{2\pi}} \qquad \text{and} \quad c_\theta = \frac{\exp\left(-\hat{\beta}_0 + \frac{1}{2}\frac{\hat{\beta}_1^2}{1+2\hat{\beta}_2}\right)}{\sqrt{1+2\hat{\beta}_2}}.$$

Being a MLE, $\hat{\beta}$ is regular and efficient, so $\sqrt{n}(\hat{\beta} - \beta) = \Delta + o_P(1)$ for a normal vector Δ. Moreover, if we generate X_i^θ through $X_i^\theta = \mu + \sigma\widetilde{X}_i$, $\widetilde{X}_i \sim N(0,1)$, there is one-to-one correspondence between $X_i^{\theta_1}$ and $X_i^{\theta_2}$ for every θ_1 and θ_2, so the dependence of Δ on θ disappears as $n \to \infty$ for otherwise a more efficient estimator exists to contradict efficiency. Therefore, the formula for $\hat{p}_\theta$ implies that Assumption 4 (i) is satisfied with the oracle score function $\dot{\ell}_{\theta_0}$; since $\hat{p}_\theta$ is twice differentiable, it holds with a faster rate of $O_P(\|h\|^4)$. Meanwhile, if inflated with $\sqrt{n}$, the dependence of Δ on θ may not be ignorable. Simulation suggests that this dependence is linear and of order $O(n^{-1/2})$, so write $\sqrt{n}(\hat{\beta}-\beta) = \Delta + n^{-1/2}\dot{\Delta}(\theta-\theta_0) + o_P(n^{-1/2})$ for some $\dot{\Delta}$ independent of θ. Considering

[4] Hellinger neighborhoods are less appropriate for misspecified models

c_θ as a function of $\hat{\beta}$ and β as a function of θ, Taylor's theorem implies

$$\begin{aligned} n(c_\theta - c_{\theta_0}) &= \sqrt{n}\tfrac{\partial c_\theta}{\partial \beta'}\sqrt{n}(\hat{\beta}_\theta - \hat{\beta}_{\theta_0}) + \tfrac{1}{2}\sqrt{n}(\hat{\beta}_\theta - \beta_{\theta_0})'\tfrac{\partial^2 c_\theta}{\partial\beta\partial\beta'}\sqrt{n}(\hat{\beta}_\theta - \beta_{\theta_0}) \\ &\quad - \tfrac{1}{2}\sqrt{n}(\hat{\beta}_{\theta_0} - \beta_{\theta_0})'\tfrac{\partial^2 c_\theta}{\partial\beta\partial\beta'}\sqrt{n}(\hat{\beta}_{\theta_0} - \beta_{\theta_0}) + o_P(1), \\ \sqrt{n}(\hat{\beta}_\theta - \hat{\beta}_{\theta_0}) &= \tfrac{\partial\beta}{\partial\theta'}\sqrt{n}(\theta - \theta_0) + \tfrac{1}{2}(\mu - \mu_0)\tfrac{\partial^2\beta}{\partial\mu\partial\theta'}\sqrt{n}(\theta - \theta_0) \\ &\quad + \tfrac{1}{2}(\sigma^2 - \sigma_0^2)\tfrac{\partial^2\beta}{\partial\sigma^2\partial\theta'}\sqrt{n}(\theta - \theta_0) + \tfrac{\dot{\Delta}}{\sqrt{n}}(\theta - \theta_0) + o_P(n^{-1/2}). \end{aligned}$$

At $\theta = \theta_0$,

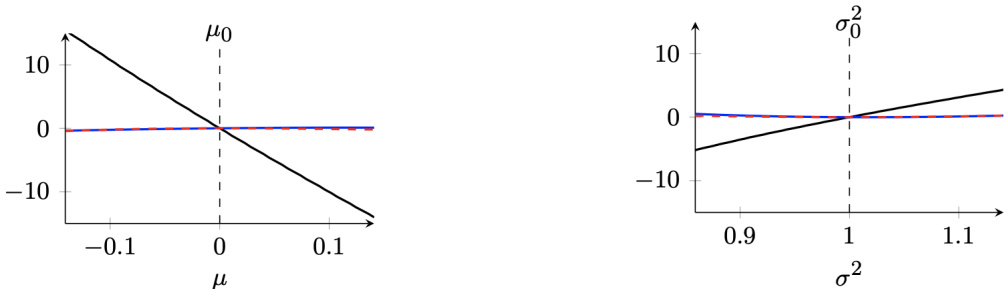
$$\tfrac{\partial c_\theta}{\partial\beta} = \begin{bmatrix} -1 \\ 0 \\ -1 \end{bmatrix}, \tfrac{\partial^2 c_\theta}{\partial\beta\partial\beta'} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{bmatrix}, \tfrac{\partial\beta}{\partial\theta'} = \begin{bmatrix} 0 & \frac{1}{2} \\ -1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}, \tfrac{\partial^2\beta}{\partial\mu\partial\theta'} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \tfrac{\partial^2\beta}{\partial\sigma^2\partial\theta'} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Substituting these, we can derive that

$$n(c_\theta - c_{\theta_0}) = \sqrt{n}(\theta - \theta_0)'\left(\begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}\Delta + \dot{\Delta}'\begin{bmatrix} -1 \\ 0 \\ -1 \end{bmatrix}\right) + o_P(1),$$

yielding Assumption 4 (iii). Finally, Figure 1 illustrates Assumption 4 (ii) and (iii). The black lines plot $n(c_\theta - c_{\theta_0})$ as we change θ; they are linear and its quadratic curvatures are ignorable. The blue lines represent $n(\mathbb{P}_n - P_{\theta_0})\left(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1 - (\theta - \theta_0)'\dot{\ell}_{\theta_0}/2\right)$ and the red lines $n(\mathbb{P}_n - P_{\theta_0})\left(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1\right)^2$; compared to the values of $n(c_\theta - c_{\theta_0})$, both are uniformly ignorable.

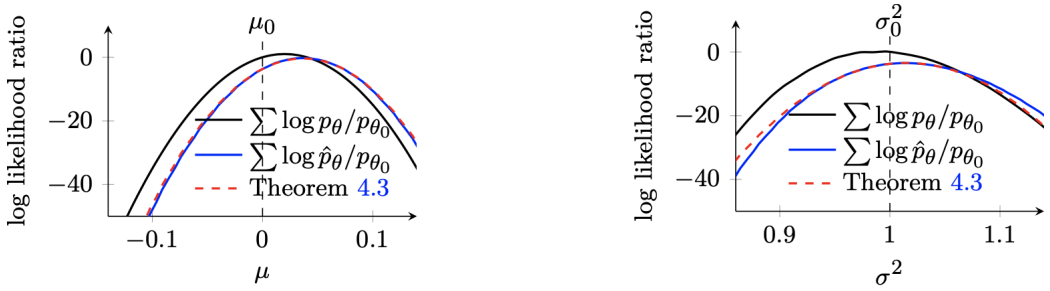
Since this model with the logistic classifier satisfies Assumptions 3 and 4, it is susceptible to Theorem 4.2. This is supported by a diagnostics plot in Figure 2 which portrays true and estimated likelihood ratios. In Figure 2a, μ is varied with σ^2 fixed at σ_0^2 while, in Figure 2b, σ^2 is varied with μ held at μ_0. The difference between the estimated log likelihood (blue) and the quadratic approximation (dashed red) is negligible, demonstrating that the validity of Theorem 4.2 is justifiable. Compared to the oracle log likelihood (black), the estimated log likelihood is shifted by the random term $\sqrt{n}(\dot{c}_{n,\theta_0} - \hat{P}_{\theta_0}\dot{\ell}_{\theta_0})$. The curvature, however, is the same as oracle since the red line curves by the Fisher information I_{θ_0}. Thus, we expect Algorithm 1 to produce a biased sample and Algorithm 2 a dispersed sample. Note that we can compute $\sqrt{n}\hat{P}_{\theta_0}\dot{\ell}_{\theta_0} = c_{\theta_0}\sqrt{n}\left[-\frac{\hat{\beta}_1}{1+2\hat{\beta}_2}, -\frac{1}{2} + \frac{1}{2(1+2\hat{\beta}_2)} + \frac{\hat{\beta}_1^2}{2(1+2\hat{\beta}_2)^2}\right]'$, which is asymptotically linear in Δ by the delta method. It is then reasonable to expect that this term has mean zero when averaged over $\widetilde{X}$ since $\hat{\beta}$ is asymptotically unbiased. If $\dot{c}_{n,\theta_0}$ also has mean zero, then Algorithm 2 is unbiased and Algorithm 3 recovers the exact normal posterior.



(a) The black line $n(c_\theta - c_{\theta_0})$; the blue line $n(\mathbb{P}_n - P_{\theta_0})(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1 - (\theta-\theta_0)'\dot{\ell}_{\theta_0}/2)$; the red line $n(\mathbb{P}_n - P_{\theta_0})(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1)^2$. σ^2 is fixed at σ_0^2.

(b) The black line $n(c_\theta - c_{\theta_0})$; the blue line $n(\mathbb{P}_n - P_{\theta_0})(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1 - (\theta-\theta_0)'\dot{\ell}_{\theta_0}/2)$; the red line $n(\mathbb{P}_n - P_{\theta_0})(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1)^2$. μ is fixed at μ_0.

Figure 1: Illustration of Assumption 4 (ii–iii) in the normal location-scale example with $n = m = 5000$.



(a) True log likelihood, estimated log likelihood, and quadratic approximation by Theorem 4.2. $\sigma^2 = \sigma_0^2$.

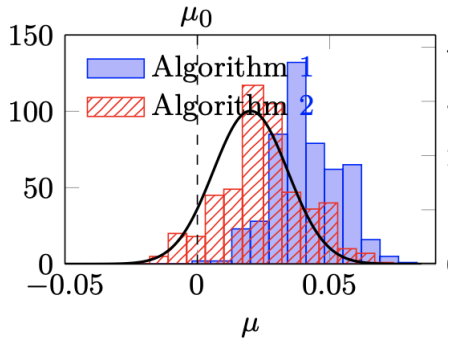
(b) True log likelihood, estimated log likelihood, and quadratic approximation by Theorem 4.2. $\mu = \mu_0$.

Figure 2: Illustration of Theorem 4.2 in the normal mean-scale example with $n = m = 5000$.

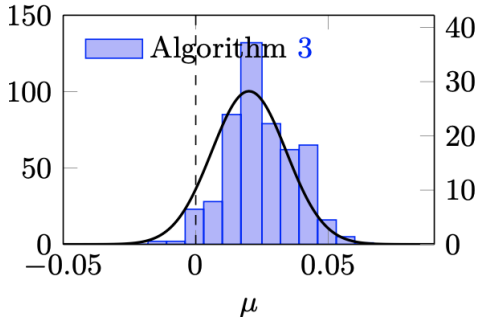
To see that this is indeed the case, we impose a conjugate normal-inverse-gamma prior, $\theta \sim N\Gamma^{-1}(\mu_0, \nu, \alpha, \beta)$, that is, the marginal prior of σ^2 is the inverse-gamma $\Gamma^{-1}(\alpha, \beta)$ and the conditional prior of μ given σ^2 is $N(\mu_0, \frac{\sigma^2}{\nu})$. The posterior is then analytically calculated as (for $\bar{X}_n = \frac{1}{n}\sum_i X_i$)

$$\theta \mid X \sim N\Gamma^{-1}\Big(\frac{\nu\mu_0 + n\bar{X}_n}{\nu+n}, \nu+n, \alpha+\frac{n}{2}, \beta + \frac{1}{2}\sum_i (X_i - \bar{X}_n)^2 + \frac{n\nu}{\nu+n}\frac{(\bar{X}_n - \mu_0)^2}{2}\Big).$$

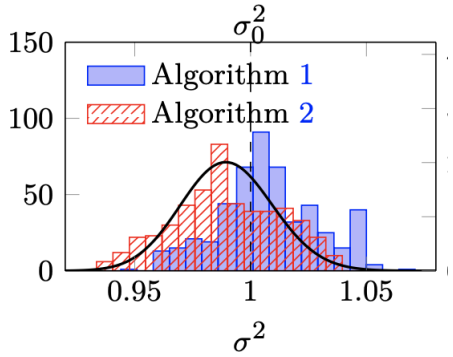
Figure 3 shows the histograms of Algorithm 1, 2 and 3 after $K = 500$ MCMC steps. Since the estimated log likelihood has a rightward bias (as seen from Figure 2), Algorithm 1



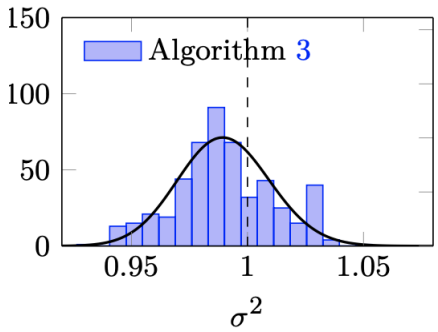
(a) Algorithm 1 and 2.



(b) Algorithm 3.



(c) Algorithm 1 and 2.



(d) Algorithm 3.

Figure 3: Histograms of the MHC samples of μ and σ^2 in the normal location-scale model. Algorithm 1 (resp. 2) yield more biased (resp. dispersed) samples compared to the true posterior (black curve). Algorithm 3 (on the right) tracks the black curve more closely.

produces a sample that is shifted to the right (Figures 3a and 3c). Algorithm 2, on the other hand, gives a sample that is more dispersed than the posterior but is correctly placed, indicating that the random bias has mean zero. Consequently, Algorithm 3 generates a sample that is placed and shaped correctly (Figures 3b and 3d).

7 Mixing Properties of MHC

A critical issue for MCMC algorithms is the determination of the number of iterations needed for the result to be approximately a sample from the distribution of interest. This section sheds light on the mixing rate of Algorithm 1. Under standard assumptions on $q(\cdot \mid \cdot)$ (such as positivity almost surely, see Corollary 4.1 in [19]), the distribution of the

MHC Markov chain after t steps will converge to $\pi_n^\star(\theta \,|\, X^{(n)})$ from any initialization in Θ in total variation as $t \to \infty$. [14] derive necessary and sufficient conditions for the Metropolis algorithms (with independent or symmetric candidate distributions) to converge at a geometric rate to a prescribed continuous distribution. [3] studied the speed of convergence of MH when both $n \to \infty$ *and* $d \to \infty$ where $\theta \in \Theta \subset \mathbb{R}^d$.

We can reformulate their sufficient conditions for showing polynomial mixing times of MHC. Recall that the stationary distribution $\pi_n^\star(\theta \,|\, X^{(n)})$ of the MHC sampler in (3.4) normalized to a compact set $K \subset \Theta$, writes as $\Pi_K^\star(B) = \int_B \pi_n^\star(\theta \,|\, X^{(n)}) / \int_K \pi_n^\star(\theta \,|\, X^{(n)})$. We are interested in bounding the number of steps needed to draw a random variable from Π_K^* with a given precision. We denote with Π_K^{*t} the distribution obtained after t steps of the MHC algorithm starting from Π_K^{*0}. It is known (see e.g. [13]) that the total variation distance between Q and Q_t can be bounded by $\|\Pi_K^* - \Pi_K^{*t}\|_{TV} \leq \sqrt{M}(1 - \phi^2/2)^t$, where M is a constant which depends on the initial distribution Π_K^{*0} and ϕ is the *conductance* of the Markov chain defined, e.g., in (3.13) in [3]. To obtain bounds on the conductance, the Markov chain needs to transition somewhat smoothly (see assumption D1 and D2 in [3]). These assumptions pertain to the continuity of the transitioning measure and are satisfied by the Gaussian random walk with a suitable choice of the proposal variance (see Section 3.2.4 in [3]) The following Lemma summarizes Theorem 2 of [3] in the context of Algorithm 1 under asymptotic normality assumptions examined in more detail in Section 8.

Lemma 7.1. *(Mixing Rate) Under conditions in equations* (8.7)-(8.8) *and a Gaussian random walk* $q(\cdot \,|\, \cdot)$ *satisfying Lemma 4 of [3], the global conductance* ϕ *of the Markov chain obtained from Algorithm 1 satisfies* $1/\phi = \mathcal{O}(d)$ *in* $P_{\theta_0}^{(n)}$*-probability. In addition, the minimal number of MCMC iterations needed to achieve* $\|\Pi_K^* - \Pi_K^{*t}\|_{TV} < \epsilon$ *is* $\mathcal{O}(d^2 \log(M/\epsilon))$ *for some suitable constant* M *depending on the initial distribution* Π_K^{*0}.

MHC thus attains bounds on the mixing rate that are *polynomial* in d (i.e. rapid mixing) under suitable Bernstein-von Mises conditions formalized later in Section 8. This section investigates how fast the Markov chain converges to its target $\pi_n^\star(\theta \,|\, X^{(n)})$ as the number of iterations t grows. In Section 4.2.1 (resp. Section 4.2.2), we investigate a fundamentally different question. We assess the speed at which the target $\pi_n^\star(\theta \,|\, X^{(n)})$ shrinks around the truth θ_0 (resp. a Kullback-Leibler projection) as n grows.

The multiplication constant M in Lemma 7.1 depends on the initial distribution. Namely, the initial distribution needs to be "M-warm" according to assumption (3.5) in [3]. Loosely speaking, M quantifies the amount of overlap between the initial and stationary distributions. Our convergence rate result thereby implicitly incorporates the properties of the initialization algorithm by regarding the constant M as dependent on the initialization routine. In our Lotka-Volterra example, we found that the mixing mixing performance of MHC depends on the classifier. With random forests, the initialization was not as important since the shape of the likelihood approximation did not have a sharp peak (compare Figure 5 and 6 in the main manuscript). On the other hand, `glmnet` yields likelihood approximations with only a very narrow area of likelihood support and the initialization needed to be close in order to avoid a very long burn-in. We have considered an ABC pilot run for initialization. Alternatively, one could try less accurate/costly classifiers in a pilot run to obtain a good initialization.

8 Bernstein-von Mises Theorem

The Bernstein-von Mises (BvM) theorem asserts that the posterior distribution of a parameter in a suitably regular finite-dimensional model is approximately normally distributed as the number of observations grows to infinity. More precisely, if p_θ is appropriately smooth and identifiable in θ and the prior $\Pi_n(\cdot)$ puts positive mass around the true parameter θ_0, then the posterior distribution of $\sqrt{n}(\theta - \widehat{\theta}_n)$ tends to $N(0, I_{\theta_0}^{-1})$ for most observations $X^{(n)}$, where $\widehat{\theta}_n$ is an efficient estimator and I_θ is the Fisher information matrix of the model at θ. In this section, we want to understand the effect of the tilting factor $\mathrm{e}^{u_\theta(X^{(n)})}$ on the limiting shape of the pseudo-posterior in (3.4) that is proportional to $\pi_n(\theta \mid X^{(n)})\mathrm{e}^{u_\theta(X^{(n)})}$. Exponential tilting is particularly intuitive for linear $u_\theta(X^{(n)})$ and for Gaussian posteriors where it implies a location shift. Example 1 below reveals how the behavior of $u^*(X^{(n)})$ affects the centering of the posterior limit (under linearity and Gaussianity)

Example 1. *(Linear u_θ) Suppose that the posterior $\pi_n(\theta \mid X^{(n)})$ is Gaussian with some mean μ and covariance Σ. This holds approximately in regular models according to the BvM theorem (Theorem 10.1 in [20]). Assume that there exists an invertible mapping $\tau : \Theta \to \Theta$*

such that $\theta = \tau(\bar{\theta})$ *where the density for* $\bar{\theta}$ *satisfies* $\pi_n(\theta \mid X^{(n)})\mathrm{e}^{u_\theta(X^{(n)})}\mathrm{d}\theta \propto \pi_n^*(\bar{\theta} \mid X^{(n)})\mathrm{d}\bar{\theta}$. *Assuming the following linear form (justified in Remark 4.3)*

$$u_\theta(X^{(n)}) = a^*(X^{(n)}) + \theta' u^*(X^{(n)}) \tag{8.1}$$

we obtain $\bar{\theta} \sim \mathcal{N}(\mu + \Sigma\, u^*(X^{(n)}), \Sigma)$. *In this case, the mapping* τ *satisfies* $\theta = \tau(\bar{\theta}) = \bar{\theta} - \Sigma u^*(X^{(n)})$, *implying a location shift. We had concluded a similar property below Theorem 4.2 at the end of Section 4.1.*

We now turn to more precise statements by recollecting the BvM phenomenon under misspecification in LAN models [12]. The centering and the asymptotic covariance matrix will be ultimately affected by θ^* in (4.8).

Lemma 8.1. *(Bernstein von-Mises) Assume that the posterior (4.7) concentrates around* θ^* *at the rate* ε_n^* *and that for every compact* $K \subset \mathbb{R}^d$

$$\sup_{h \in K} \left| \log \frac{\widetilde{p}^{(n)}_{\theta^* + \varepsilon_n^* h}(X^{(n)})}{\widetilde{p}^{(n)}_{\theta^*}(X^{(n)})} - h' \widetilde{V}_{\theta^*} \widetilde{\Delta}_{n,\theta^*} - \frac{1}{2} h' \widetilde{V}_{\theta^*} h \right| \to 0 \quad \textit{in } P^{(n)}_{\theta_0}\textit{-probability} \tag{8.2}$$

for some random vector $\widetilde{\Delta}_{n,\theta^*}$ *and a non-singular matrix* $\widetilde{V}_{\theta^*}$. *Then the pseudo-posterior converges to a sequence of normal distributions in total variation at the rate* ε_n^*, *i.e.*

$$\sup_B \left| \Pi_n^* \left(\varepsilon_n^{*-1}(\theta - \theta^*) \in B \mid X^{(n)} \right) - N_{\widetilde{\Delta}_{n,\theta^*}, \widetilde{V}_{\theta^*}}(B) \right| \to 0 \quad \textit{in } P^{(n)}_{\theta_0}\textit{-probability.}$$

Proof. Follows from Theorem 2.1 of [12].

It remains to examine the assumption (8.2). For `iid` data, [12] derived sufficient conditions (Lemma 2.1) for (8.2) to hold. Due to the non-separability of the term $u_\theta(X^{(n)})$, the mis-specified model cannot be regarded as arriving from an `iid` experiment. In Lemma 8.2 below we nevertheless provide intuition for when (8.2) is expected to hold if $u_\theta(X^{(n)})$ is linear. Recall that in Remark 4.3 we have concluded that under differentiability, the posterior residual $u_\theta(X^{(n)})$ does converge to a linear function in θ. Below, we provide sufficient conditions for the LAN assumption (8.2), relaxing slightly Lemma 2.1 in [12]. The assumptions in Lemma 8.2 are closely related to the ones in Theorem 4.2. The main difference is that Lemma 8.2 is concerned with the behavior of the (misspecified) likelihood around θ^* as opposed to θ^0.

Lemma 8.2. *Assume that* $P_{\theta_0}^{(n)} = P_{\theta_0}^n$ *with a density* $\prod_{i=1}^n p_{\theta_0}(x_i)$ *where the function* $\theta \to \log p_\theta(x)$ *is differentiable at* θ^* *with a derivative* $\dot{\ell}_\theta$. *Assume there exists an open neighborhood* U *of* θ^* *such that* $\left|\log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)}\right| \leq m_{\theta^*}\|\theta_1 - \theta_2\|$ $P_{\theta_0} - a.s. \forall \theta_1, \theta_2 \in U$ *where* m_θ *is a square integrable function. Assume that the log-likelihood has a 2nd order Taylor expansion around* θ^* *(i.e.* (8.5) *holds). Assume that* u_θ *is asymptotically linear around* θ^* *(i.e.* (8.6) *holds), then* (8.2) *holds with* $\varepsilon_n^* = 1/\sqrt{n}$ *and*

$$\widetilde{V}_\theta = V_\theta \quad \text{and} \quad \widetilde{\Delta}_{n,\theta} = V_\theta^{-1}\left[\frac{\dot{C}_\theta}{\sqrt{n}} + \sqrt{n}\mathbb{P}_n\dot{\ell}_\theta + \frac{u^*(X^{(n)})}{\sqrt{n}}\right] \tag{8.3}$$

Proof. We can write

$$\log \frac{\widetilde{p}_{\theta^*+\varepsilon_n h}^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} = \log \frac{C_{\theta^*+\varepsilon_n h}}{C_{\theta^*}} + \log \frac{p_{\theta^*+\varepsilon_n h}^{(n)}}{p_{\theta^*}^{(n)}} + u_{\theta^*+\varepsilon_n h} - u_{\theta^*}. \tag{8.4}$$

This yields, from Lemma 19.31 in [20], that

$$\mathbb{G}_n\left(\sqrt{n}\log\frac{p_{\theta^*+h/\sqrt{n}}}{p_\theta^*} - h'\dot{\ell}_{\theta^*}\right) \to 0 \quad \text{in } P_0,$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_{\theta_0})$ is the empirical process. Assuming that

$$P_{\theta_0}\log\left(\frac{p_\theta}{p_{\theta^*}}\right) = P_{\theta_0}\dot{\ell}'_{\theta^*}(\theta-\theta^*) + \frac{1}{2}(\theta-\theta^*)'V_{\theta^*}(\theta-\theta^*) + o(\|\theta-\theta^*\|^2) \quad \text{as } \theta \to \theta^* \tag{8.5}$$

one obtains

$$\begin{aligned}\log\frac{p_{\theta^*+h/\sqrt{n}}^{(n)}}{p_{\theta^*}^{(n)}} = n\mathbb{P}_n\log\frac{p_{\theta^*+h/\sqrt{n}}}{p_{\theta^*}} &= o_P(1) + \mathbb{G}_n h'\dot{\ell}_{\theta^*} + nP_{\theta_0}\log\frac{p_{\theta^*+h/\sqrt{n}}}{p_{\theta^*}} \\ &= o_P(1) + \mathbb{G}_n h'\dot{\ell}_{\theta^*} + \frac{h_n' V_{\theta^*} h}{2} + \sqrt{n}P_{\theta_0}h'\dot{\ell}_\theta\end{aligned}$$

If we assume asymptotic linearity of u_θ around θ^*, i.e.

$$u_{\theta^*+h/\sqrt{n}}(X^{(n)}) - u_{\theta^*}(X^{(n)}) = \frac{1}{\sqrt{n}}h'u^\star(X^{(n)}) + o_P(1) \tag{8.6}$$

for some $u^\star(X^{(n)})$ and

$$\log\frac{C_{\theta^*+h_n/\sqrt{n}}}{C_{\theta^*}} = \frac{\dot{C}'_{\theta^*}h_n}{\sqrt{n}} + o(1)$$

then (8.2) holds with (8.3). □

Related BvM conditions have been characterized in [3]. We restate these conditions utilizing the localized re-parametrization $h = \sqrt{n}(\theta - \theta_0) - s$, where $s = \sqrt{n}(\hat{\theta} - \theta_0)$ is

a *zero-mean* vector where $\hat{\theta}$ is some suitable estimator. We first define a localized criterion function $\ell(h) \equiv \frac{\widetilde{p}_{\hat{\theta}+h/\sqrt{n}}(X^{(n)})\widetilde{\pi}(\hat{\theta}+h/\sqrt{n})}{\widetilde{p}_{\hat{\theta}}(X^{(n)})\widetilde{\pi}(\hat{\theta})}$, which corresponds to the normalized pseudo-posterior $\pi^*(\theta \,|\, X^{(n)})/\pi^*(\hat{\theta} \,|\, X^{(n)})$. [3] impose a centered variant of (8.2) requiring that $\ell(h)$ approaches a quadratic form on a closed ball K (such that[5] $\Lambda \equiv \sqrt{n}(\Theta - \theta_0) - s = K \cup K^c$) in the sense that

$$|\log \ell(h) - (-h'Jh)/2| \leq \epsilon_1 + \epsilon_2 \times h'Jh/2 \quad \forall h \in K, \tag{8.7}$$

for some matrix $J > 0$ with eigenvalues bounded away from zero. If

$$\epsilon_1 = o(1) \quad \text{and} \quad \epsilon_2 \times \lambda^2_{max}(J)(\sup_{h \in K} \|h\|)^2 = o(1) \quad \text{in } P^{(n)}_{\theta_0}\text{-probability.} \tag{8.8}$$

Theorem 1 of [3] shows that $\ell(h)/\int_\Lambda \ell(h)\mathrm{d}h$ approaches the standard normal density in $P^{(n)}_{\theta_0}$-probability as $n, d \to \infty$. The condition (8.7) (a) allows for mild deviations from smoothness and log-concavity, (b) involves also the prior (unlike (8.2)) but, (c) requires the existence of a $\sqrt{n}-$consistent estimator $\hat{\theta}$. Lemma 8.1 is more general, where the rate ε^*_n does not need to be $1/\sqrt{n}$ and where the posterior is allowed to have a non-vanishing bias. The requirement (8.8) imposes certain restrictions on $u_\theta(X^{(n)})$. For example, in the linear case (8.1) one would need $u^\star(X^{(n)}) = o(\sqrt{n})$ in $P^{(n)}_{\theta_0}$-probability from (8.8).

9 Alternatives to MHC

A recent paper [9] suggests a related Metropolis-Hastings strategy which relies on a simulation-based likelihood ratio estimator trained separately from the Markov chain simulation. This estimator is based on contrastive learning between two fake data-parameter pairs, with parameters sampled from the prior and with fake data generated either from the marginal or the conditional likelihood evaluated at sampled prior parameters [18]. See also [17] (Chapter 12) for conditional density estimation using machine learning. Using the marginal distribution $p(\cdot)$ as a reference and denoting with $D^m_\theta(X) = \frac{p(X)\pi(\theta)}{p(X)\pi(\theta)+p_\theta(X)\pi(\theta)}$ we can re-write (2.4) as $p^{(n)}_\theta = p^{(n)}(X^{(n)}) \exp\left(\sum_{i=1}^n \log \frac{1-D^m_\theta(X_i)}{D^m_\theta(X_i)}\right)$, where $p^{(n)}(X^{(n)})$ is the marginal likelihood. Similarly as in (2.5), a likelihood estimator can be then obtained by replacing

[5] $\int_K \ell(h)\mathrm{d}h/\int_\Lambda \ell(h)\mathrm{d}h \geq 1 - o_{P_{\theta_0}}(1)$ and $\int_K \phi(h)\mathrm{d}h$ for $\phi(\cdot)$ standard Gaussian density

D_θ^m with $\hat{D}_\theta^m$, which is now trained solely on simulated data. The expression (2.5) then still holds with u_θ now defined using D_θ^m and $\hat{D}_\theta^m$. We implement this approach in Section 5.2 (main document) and discuss its theoretical properties in Remark 3. This approach will be advantageous when the cost of learning the likelihood ratio simulator prior to MCMC simulation outweighs the costs of performing classification at each MH step. Another related strategy was proposed in [15], where no reference is used and the likelihood ratio inside Metropolis-Hastings is estimated by contrastive learning between two fake data generated from conditional likelihoods evaluated at new versus current parameter values. This approach also requires classification at each step but is not limited by the sample size n when choosing the fake data sample size m for classification. We also implement this approach later in Section 5.2 and make comparisons with our approach in Section 12 in the Supplement. The choice of the contrasting density in the context of parameter estimation in unnormalized models is discussed in [8].

10 Ricker Model

The Ricker model is a classic discrete model that describes partially observed population dynamics of fish and animals in ecology. The latent population $N_{i,t}$ follows

$$\log N_{i,t+1} = \log r + \log N_{i,t} - N_{i,t} + \sigma\varepsilon_{i,t}, \qquad \varepsilon_{i,t} \sim N(0,1),$$

where r denotes the intrinsic growth rate and σ is the dispersion of innovations. The index t represents time and runs through 1 to $T = 20$. The index i represents independent observations and runs through 1 to $n = 300$. The initial population $N_{i,0}$ may be set as 1 or set randomly after some burn-in period. We observe $X_{i,t}$ such that

$$X_{i,t} \mid N_{i,t} \sim \text{Poisson}(\varphi N_{i,t}),$$

where φ is a scale parameter. The objective is to make inference on $\theta := (\log r, \sigma^2, \varphi)$. Each time sequence $X_i := (X_{i,1}, \ldots, X_{i,T})$ constitutes an observation, where i runs through n. In our notation, we can define the underlying data-generating process as $\widetilde{X}_{i,t} := (U_{i,t}, \varepsilon_{i,t})$ for $U_{i,t} \sim U[0,1]$ and set the function T_θ to map ε_i to N_i and then (U_i, N_i) to X_i through the Poisson inverse transform sampling of $U_{i,t}$ into $X_{i,t}$. We set the true parameter as

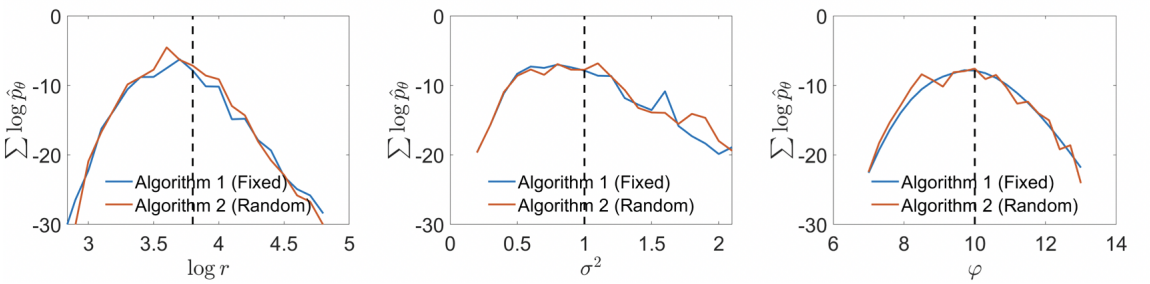


Figure 4: Estimated log likelihood ratio for the Ricker model: (Left) function of $\log r$ fixing $\sigma = \sigma_0$ and $\varphi = \varphi_0$, (Middle) function of σ^2 fixing $r = r_0$ and $\varphi = \varphi_0$, (Right) function of φ fixing $\sigma = \sigma_0$ and $r = r_0$.

$(\log r_0, \sigma_0^2, \varphi_0) = (3.8, 1, 10)$ and employ an improper, flat prior. Note that our method can accommodate an improper prior, unlike ABC.

There is no obvious sufficient statistic for this model, and the likelihood is intractable due to the nontrivial time dependence of $N_{i,t}$. We use an average of neural network discriminators to adapt to the unknown likelihood ratio. First, we estimate D_θ by a neural network with one hidden layer with 50 nodes, each of which is equipped with the hyperbolic tangent sigmoid activation function. Then, we compute the log likelihood of the data $\sum_i \log \frac{1-\hat{D}_\theta}{\hat{D}_\theta}$. We repeat this for 20 times with independently drawn $\widetilde{X}$ and take the average of the log likelihood. This specification produces approximately quadratic likelihood-ratio curves (Figure 4). Unlike the location-scale normal model, the fixed design does not produce entirely smooth curves due to the averaging aspect over many discriminators. The quadratic shape is nevertheless recovered here, implying that the differentiability assumptions from Section 4.1 are not entirely objectionable.

Figure 5 shows the marginal histograms of the MHC samples (500 MCMC iterations). The proposal distribution is independent across parameters; $\log r$ uses the normal distribution, σ^2 the inverse-gamma distribution, and φ the gamma distribution; each of them has the mean equal to the previous draw and variance $1/n$. The vertical dashed lines indicate the true parameter θ_0. Note that the posterior is asymptotically centered at the MLE, not θ_0. However, the blue histograms on the left (Algorithm 1) seem too far away from θ_0 relative to the widths of the histograms. On the other hand, the red histograms

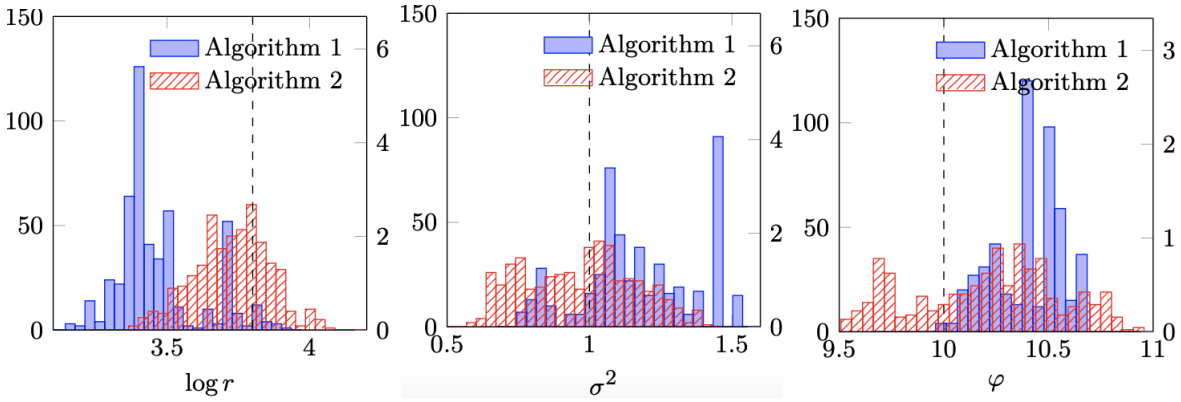


Figure 5: MHC samples for the Ricker model.

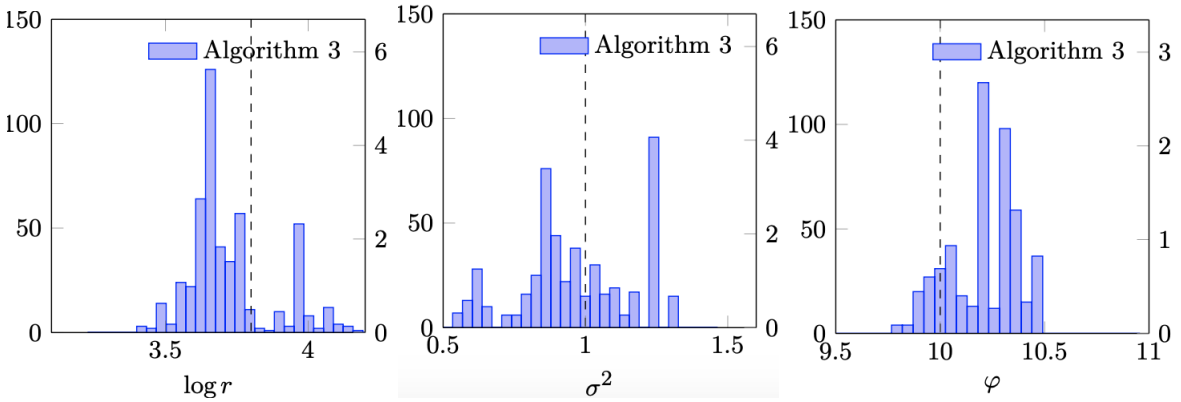


Figure 6: MHC samples for the Ricker model (Algorithm 3)

(Algorithm 1) are more dispersed but located closer to θ_0. These observations confirm our theoretical findings. Histograms of Algorithm 3 (Figure 6) look reasonable as a posterior sample, center around the true values.

Figure 6 and 7 compare our method with the MCWM pseudo-marginal Metropolis-Hastings algorithm [1]. We have implemented the default pseudo-marginal method which deploys an average of conditional likelihoods for X_i, given N_i,

$$\hat{p}(X_i) = \frac{1}{K}\sum_{k=1}^{K}\prod_{t=1}^{T} p(X_{i,t} \mid N_{i,t,k}) = \frac{1}{K}\sum_{k=1}^{K}\prod_{t=1}^{T} \frac{(\varphi N_{i,t,k})^{X_{i,t}} e^{-\varphi N_{i,t,k}}}{X_{i,t}!}$$

as the likelihood approximation, where K is some positive integer and where $N_{i,t,k}$ are independently drawn across $k = 1, \ldots, K$. In our comparisons, we let $K = 20n$. Figure 6 shows that the two methods produce posterior draws that are located at similar places, and the widths of the histograms are also comparable. We would like to point out, again,

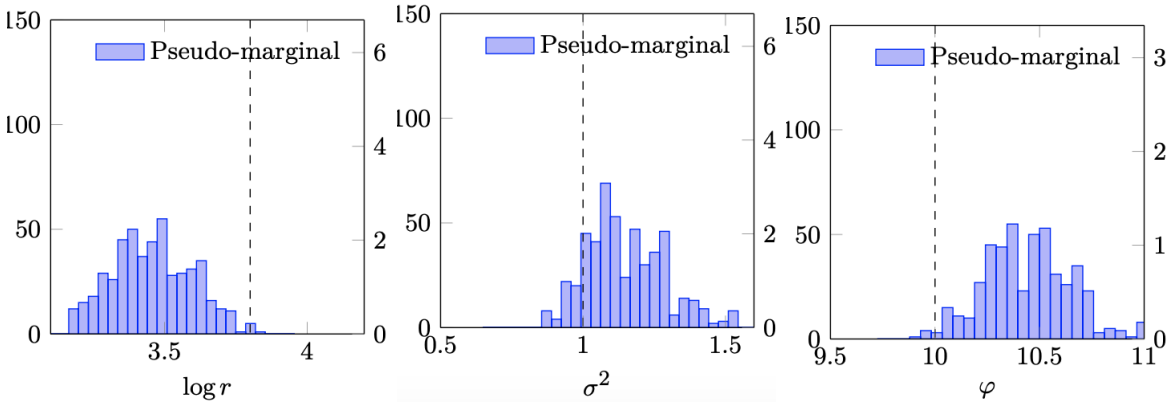


Figure 7: Posterior samples for the Ricker model using the pseudo-marginal MCWM method

that our method does not require that a tractable conditional likelihood is available nor that a user-specified summary statistic is supplied.

11 Bayesian Model Selection

The performance of summary statistic-based methods is ultimately sensitive to the quality of summary statistics whose selection can be a delicate matter. One such instance is model selection, where it is known that when ABC may fail even when the summary statistic is sufficient for *each* of the models considered [16]. Our method *does not* require a summary statistic but a sieve of discriminators that can adapt to the oracle discriminator in the limit. This creates hope that our method can tackle model selection problems. To illustrate this point we consider a toy model choice problem considered in [16]. The actual data follows $X_i \sim N(0,1)$ for $i = 1, \ldots, n = 500$. We have two candidate models $P_{1,\mu} = N(\mu, 1)$ and $P_{2,\mu} = N(\mu, 1 + 3/\sqrt{n})$ to choose from. We let the parameters be $\theta := (m, \mu)$, where $m \in \{1, 2\}$ is the model indicator and μ is unknown mean with a prior $N(0, 1)$. The model is assigned a uniform prior, i.e. $P(m = 1) = P(m = 2) = 0.5$. Following the traditional Bayesian model selection formalism, we collect evidence for model $m = 1$ with a Bayes factor

$$B_{12} := \frac{\pi_n(m = 1 \mid X)}{\pi_n(m = 2 \mid X)}.$$

The Bayes factor is the ratio of the marginal likelihoods (or posterior probabilities) of $m = 1$ over $m = 2$. The actual Bayes factor value is $B_{12} = 9$, indicating strong evidence

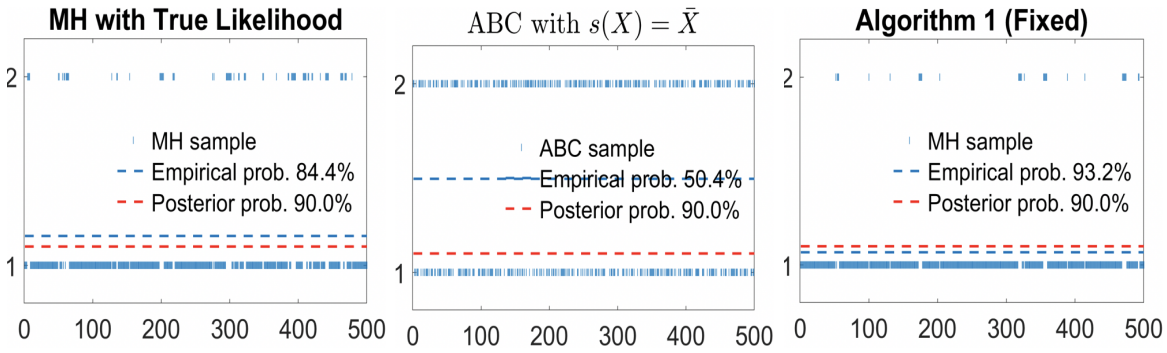


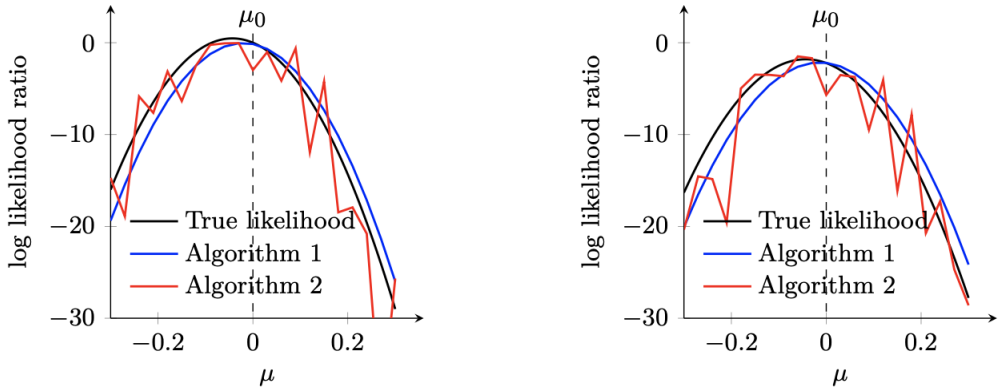
Figure 8: Trace plots of sampled models using: (Left) MH with the true likelihood ratio, (Middle) ABC with $s(X^{(n)}) = \bar{X}_n$ and (Right) fixed generator MHC.

in favor of $m = 1$. The Bayes factor will be estimated by the ratio of the frequencies of the posterior samples given by ABC or our method. Since our parameter of interest m is discrete, there is no de-biasing for this example. [16] in their Lemma 2 show that when the summary statistic is $\sum_i X_i$, the Bayes factor estimated by ABC asymptotes to 1. This is equivalent to choosing the model with a coin toss. For our method, we use the logistic regression on regressors $(1, X_i, X_i^2)$, which can mimic the oracle discriminator. The trace plots of sampled models for exact MH, MHC and ABC are provided in Figure 8. Table 1 summarizes the posterior model frequencies. The true posterior probabilities are $\pi_n(m = 1 \mid X) \approx 0.9$ and $\pi_n(m = 2 \mid X) \approx 0.1$, so the Bayes factor is 9. The "Oracle MH" is the Metropolis-Hastings with the true likelihood, in which 84.4% of the posterior draws choose model 1. Algorithms 1 and 2 choose model 1 respectively 93.2% and 70% of the times. ABC based on the sum, on the other hand, chooses the model randomly. Finally, Figure 9 in Appendix gives the estimated log-likelihood ratio for each model. In terms of μ, we again see that Algorithm 1 is slightly biased with the correct shape and Algorithm 2 is less biased but more dispersed on average.

Figure 9 shows true likelihood ratio and and classification-based estimates for fixed and random designs for the Bayesian model selection example from Section 11. Under the fixed design, the curve is smooth and slightly biased with a similar shape to the true log-likelihood. For the random design, there is no smoothness (due to the fake data refreshing aspect).

	Posterior	Oracle MH	Algorithm 1	Algorithm 2	ABC
Model 1	90%	422	466	350	252
Model 2	10%	78	34	150	248
Bayes factor	9.00	5.41	13.71	2.33	1.02

Table 1: "Posterior" column gives the posterior probability of each model, $\pi_n(m = j \mid X)$. Other columns give the frequencies of the corresponding sample of size 500. "Oracle MH" refers to the Metropolis-Hastings algorithm with the true likelihood. "ABC" is based on the summary statistics $s(X) = \bar{X}_n$.



(a) Estimated log likelihood for $m = 1$. The vertical dashed line indicates $\mu_0 = 0$.

(b) Estimated log likelihood for $m = 2$. The vertical dashed line indicates $\mu_0 = 0$.

Figure 9: Estimated log likelihood for models 1 and 2. The figures indicate that it is smooth in μ and have the same curvature as the true log likelihood.

12 The CIR Model: Further Details

This section presents additional plots for the CIR analysis from Section 5.1. Figure 10 shows smoothed posterior samples for MHC (fixed generator) and $nrep \in \{1, 5\}$. These plots look qualitatively similar to the random generator results presented in Figure 2 in the main manuscript. Next, Figure 11 and 12 show trace-plots of the MHC samples. We can see that (1) using larger $nrep$ reduces variance, (2) random generators have smaller acceptance rates for the same proposal distribution. Trace-plots for the MCWM method (Figure 13) show bias in estimation of σ. Table 2 shows posterior summaries for the various algorithms

we tried, including acceptance rates and effective sample size (computed using the `coda` R package). Since MHC (random generator) resembles GIMH [2] in that it recycles the fake data, one would expect the effective sample of MHC to be smaller than for MCWM. However, making the MCWM likelihood estimator more accurate (increasing M and N) made the effective sample size (ESS) smaller even though the acceptance rate was still around 10%. Interestingly, the random generator MHC also showed a decreased ESS (as well as the acceptance rate) once we used "better" log-likelihood estimator (i.e. averaging over $nrep = 5$ estimators using different fake data). Lastly, histograms of the posterior samples together with demarkations of the 95% credible intervals are in Figure 14 and 15.

Method	α			β			σ			AR	Time	ESS
	$\bar{\alpha}$	l	u	$\bar{\beta}$	l	u	$\bar{\sigma}$	l	u			
MH Exact	0.0693	0.683	0.703	0.1558	0.1507	0.1608	0.07	0.696	0.704	9.1	3.3	255
Alg1 ($nrep = 1$)	0.0691	0.0644	0.0735	0.1505	0.1374	0.1636	0.0703	0.0669	0.0734	16.8	4.6	191
Alg2 ($nrep = 1$)	0.0691	0.0644	0.0741	0.1476	0.1353	0.1632	0.693	0.0667	0.0725	10.7	4.9	155
Alg1 ($nrep = 5$)	0.0698	0.0667	0.0725	0.1468	0.1377	0.1574	0.0699	0.676	0.725	7.8	13.9	104
Alg2 ($nrep = 5$)	0.0691	0.0665	0.0715	0.1468	0.1366	0.1571	0.0691	0.0674	0.0714	5.6	13.9	112
MCWM ($M = 2$)	0.0693	0.0658	0.0733	0.1469	0.1287	0.1632	0.067	0.0657	0.0684	13.1	15.9	316
MCWM ($M = 5$)	0.0694	0.0662	0.723	0.1538	0.1423	0.1634	0.0689	0.0676	0.0698	10.1	238.6	63

Table 2: Posterior means and 95% credible interval boundaries (lower (l) and upper (u)). AR is the acceptance rate and `Time` is computing time (in hours) for 10 000 iterations. `ESS` is the average effective sample size for the three chains computed using the R package `coda`.

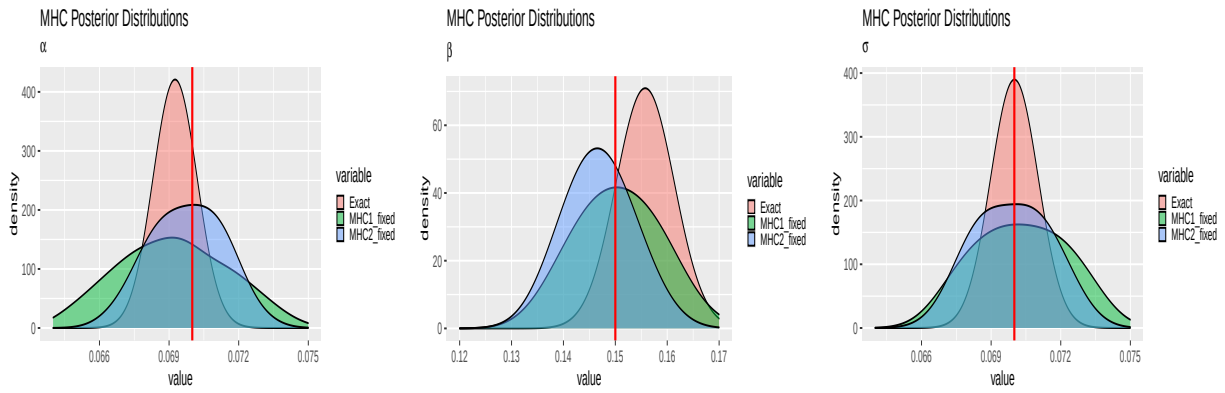


Figure 10: Smoothed posterior densities obtained by simulation using the exact MH and MHC fixed generator using $nrep \in \{1, 5\}$

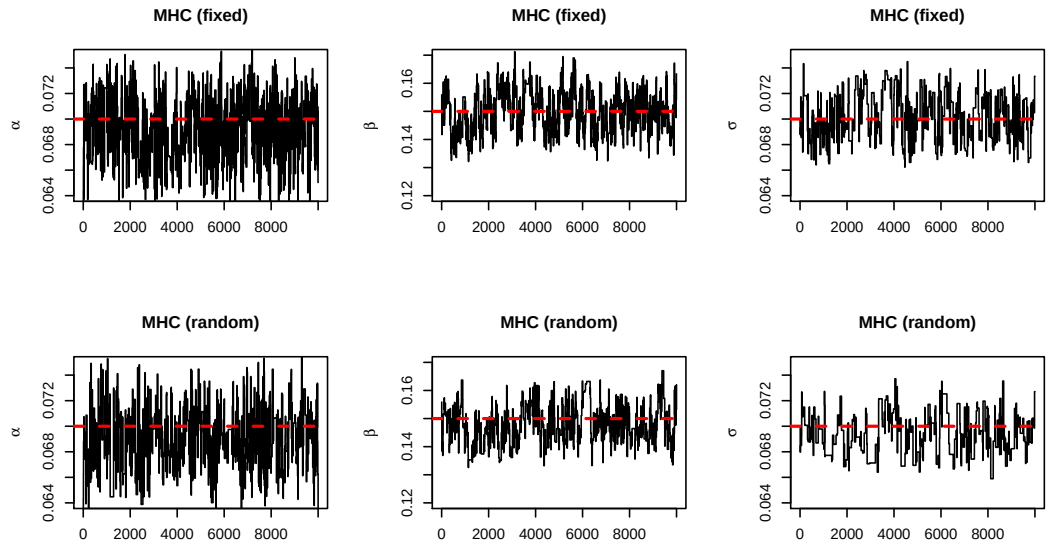


Figure 11: Trace-plots of 10 000 MHC iterations with $nrep = 1$

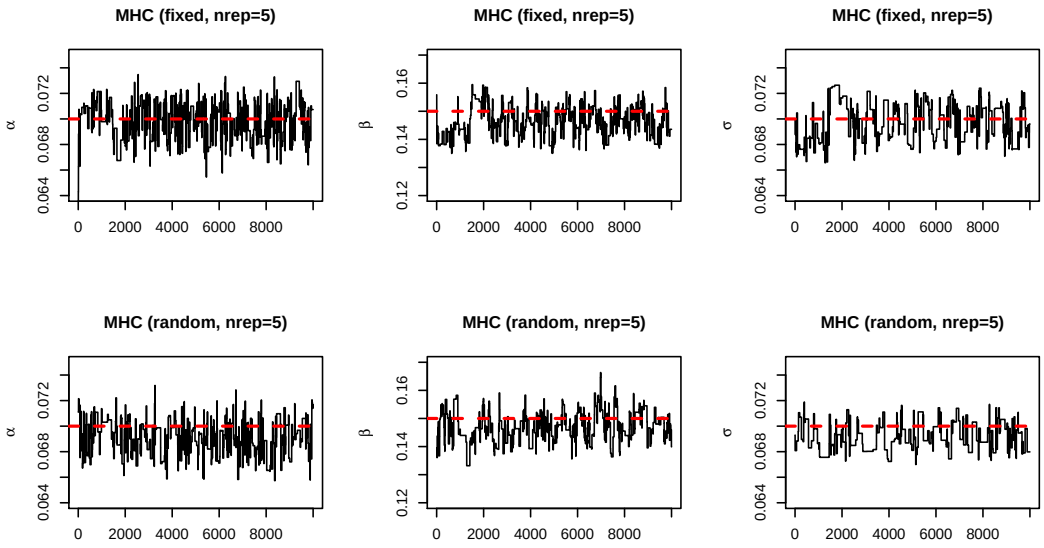


Figure 12: Trace-plots of 10 000 MHC iterations with $nrep = 5$

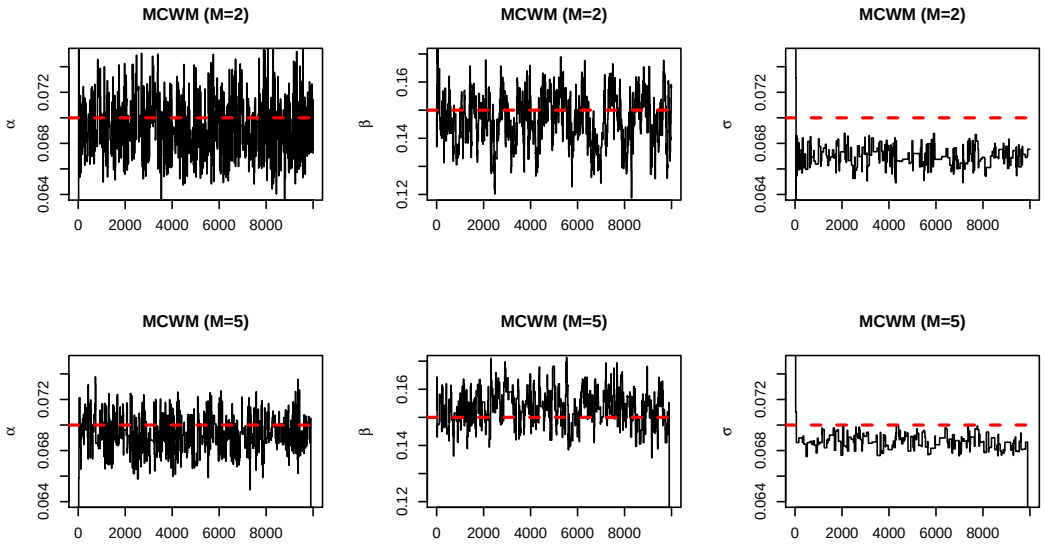


Figure 13: Trace-plots of 10 000 MCWM iterations with $M \in \{2, 5\}$

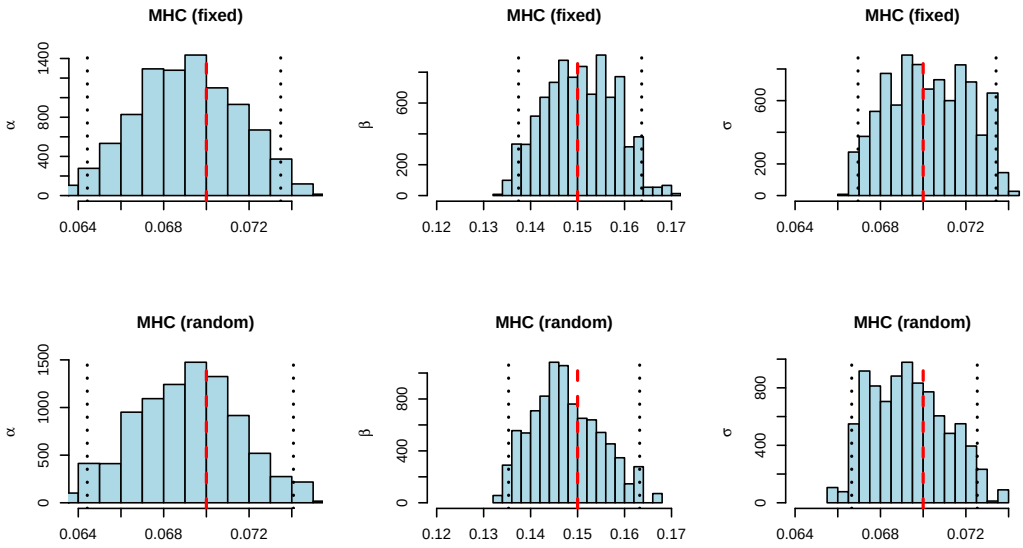


Figure 14: Histogram of 9 000 MHC iterations (after 1 000 burnin) with $nrep = 1$

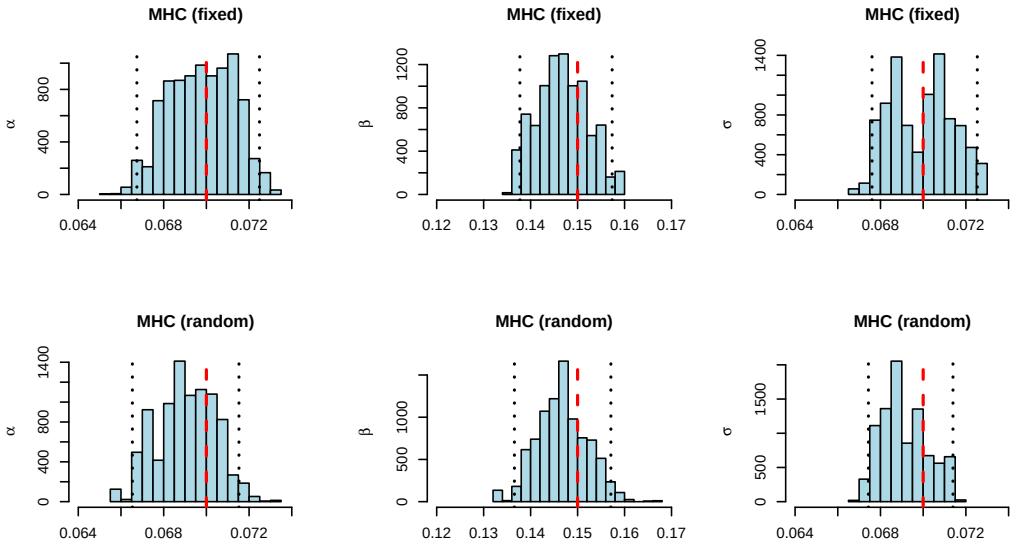


Figure 15: Histogram of 9 000 MHC iterations (after 1 000 burnin) with $nrep = 5$

13 The Lotka-Volterra Model: Further Details

13.1 Timing Comparisons

The complexity of MHC depends on the complexity of the classifier as well as on how costly it is to simulate fake data. This will be problem-specific. For example, for the Lotka-Volterra model, we have used the Gillespie algorithm [7] which can be quite costly. This will have some implication for the algorithm of [15] which generates two (not just one) fake data sets at each step. This will be slower than our approach (which generates just one fake dataset and uses observed data for contrasting) even when $m = n$. In order to get a more concrete idea about the dependence on n, m and p (which depends on the length of the time series), we have measured the cost of a single iteration of MHC for various m, n and p for the default implementation of `cv.glmnet` (10 fold cross-validation) and `randomForests` (500 trees). The computing times are in Figure 16. The default implementation of `glmnet` appears to scale less favorably with n compared to random forests and the complexity, of course, increases with m. The method of [15] requires simulating two (as opposed to one) fake dataset at each step and is, thereby, slower. This seemingly minor timing gap can aggregate in long Monte Carlo simulations. For example, 10 000 iterations of MHC with default random forests took 2.5 hours for $n = m = 20$, where [15] takes more than 6 hours with the same classifier and $n = m = 20$. This gap is particularly prominent when p (i.e. the length of the time series) is large. Random forests scale less favorably with p, compared to `glmnet` logistic regression.

Our LASSO implementation uses `glmnet` [4] where the complexity depends on the number of penalty parameters and the number of iterations of the inner coordinate ascent algorithm. As shown in Section 3 of [4], the `glmnet` algorithm for logistic regression has three nested loops. For each penalty parameter, one performs a penalized variant of iterated reweighted least squares. Because the weights are changing throughout the iterations, one cannot use faster covariance updates (Section 2.2 in [4]) and each inner cycle thereby costs $O(np)$. The complexity (without cross-validation) thus depends on the number of re-weighting steps, the number of inner iteration cycles and the length of the regularization path.

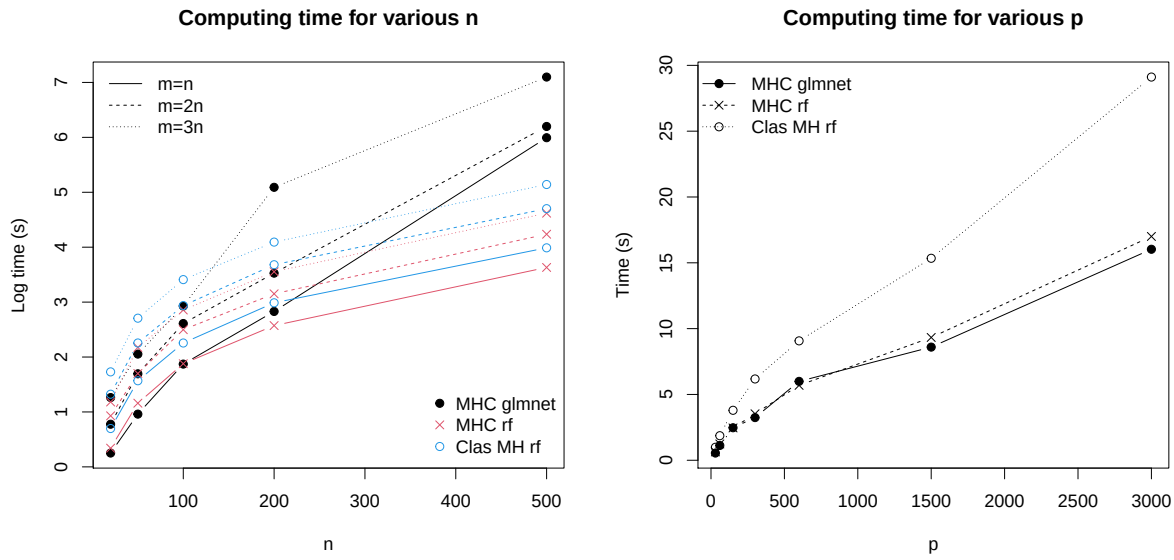


Figure 16: Computing times for one iteration of MHC and Clas MH (Classification MH of [15]) in the Lotka-Volterra example. (Left) Fixed $p = 603$ and various n and m (fake data sample size). (Right) Fixed $n = 20$ and various p (depending on the length of the time series).

13.2 The effect of m and $nrep$

We found that computing the classification estimator separately for $nrep$ many fake data (using observed data as a reference) and averaging them out stabilizes estimation. Since our MHC approach uses real observed data for contrasting, it will have the limitation that the choice of m cannot be much larger than n in order for the classification to yield good results. Indeed, we found that for small n, increasing m does help as long as it is not overly large to make the classification problem too imbalanced. This can be seen from Figure 17 below where using $n = 20$ and $m = 1\,000$ yielded unstable classification (using cross-validation and the `glmnet` classifier). Averaging over $nrep$ log-likelihood estimators is a heuristic for stabilizing estimation when n is small and, thereby, m cannot be chosen overly large. In addition, while increasing m may result in estimators which concentrate more sharply around the truth, averaging out $nrep$ estimators will result in a smoother final estimator.

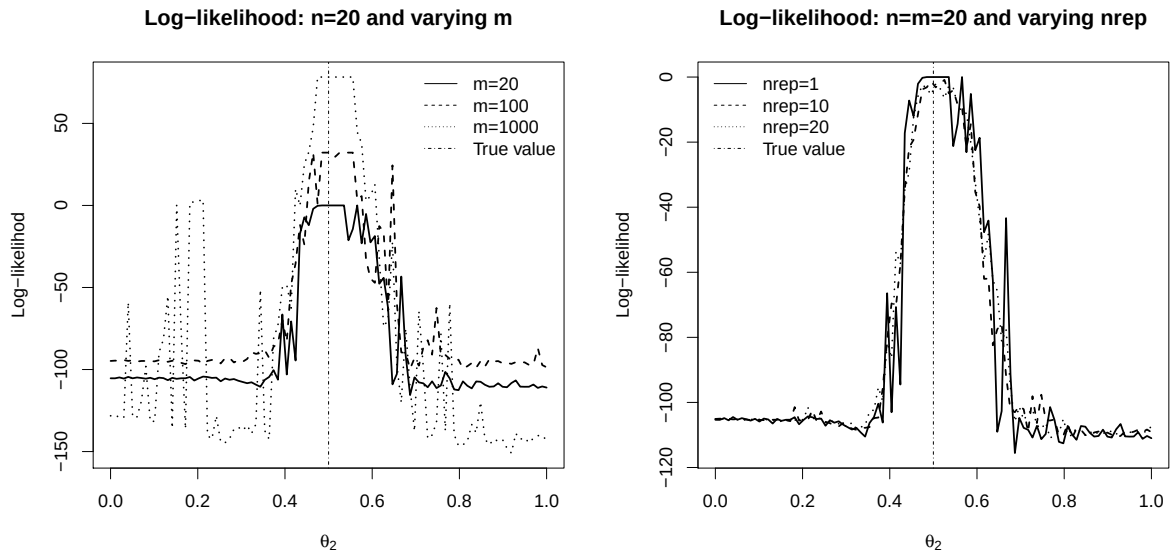


Figure 17: Log-likelihood estimators for varying m and $nrep$ and fixed $n = 20$.

13.3 Comparisons

Referees suggested comparisons with other classification MCMC approaches which use the conditional likelihood with fake data as a reference [15] or the marginal likelihood as a reference [9]. We explore the extent to which using the conditional fixed reference (i.e. the observed data) in our MHC approach is beneficial. [9] point out that using a fixed reference point might be problematic if there is not enough overlap between the conditional densities. MHC uses the truth (i.e. the real data) as the fixed reference, tacitly assuming that if the Markov chain is initialized in the vicinity of the truth, the lack of overlap between the two likelihood densities would not be a practical concern. We anticipated that using other fixed reference point (i.e. not contrasting agains observed data) might increase variance in the random generator design since the fake reference data would introduce extra randomness. This is indeed the case when looking at the width of the 95% credible interval in Table 3 (comparing MHC with random forests and Classif MH of [15] with $n = m = 20$). The only difference between these two methods is that [15] generates another set of fake data as a reference.

In particular, the method of [15] directly computes the likelihood ratio of the new

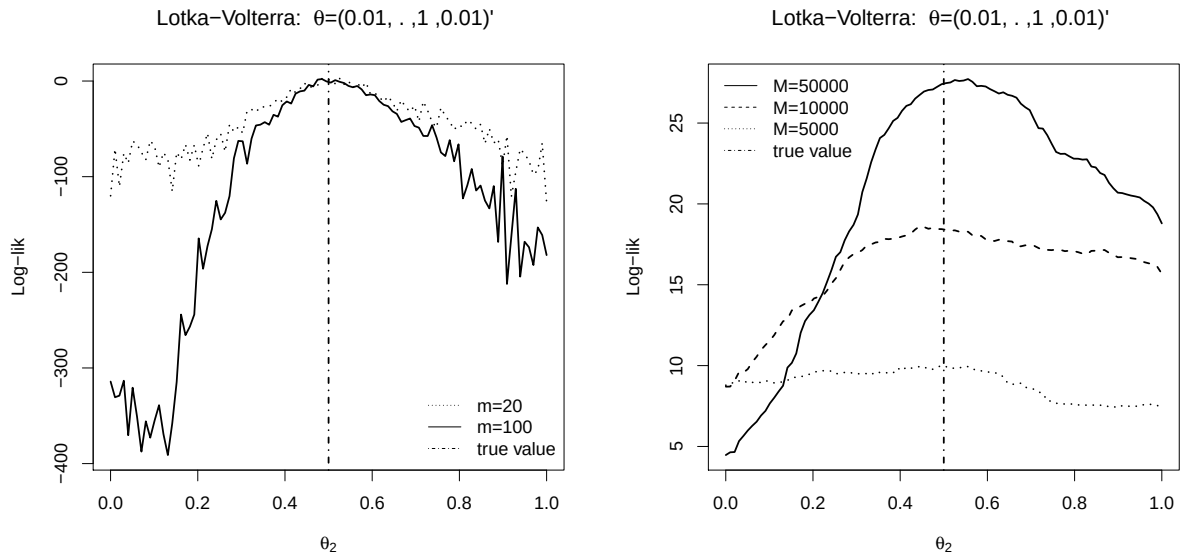


Figure 18: (**Conditional versus Marginal Reference**) Plot of estimated log-likelihood as a function of θ_2, keeping all the other parameters at the truth. (Left) The conditional approach of [15] using various m and using the default random forest classifier (R package `randomForest`), (Right) the marginal approach of [9] using various m and the random forest classifier.

versus old proposed value by contrasting two fake datasets without any fixed reference. We have implemented their approach which uses a random forest discriminator (the default `randomForest` setting in R). The plot of the estimated log-likelihod (a variant of Figure 17 on the right) is depicted in the left panel of Figure 18. We tried fake datasets of size $m = n = 20$ and $m = 100$. Learning, of course, improves with increased m but at much increased computational cost (see Section 13.1).

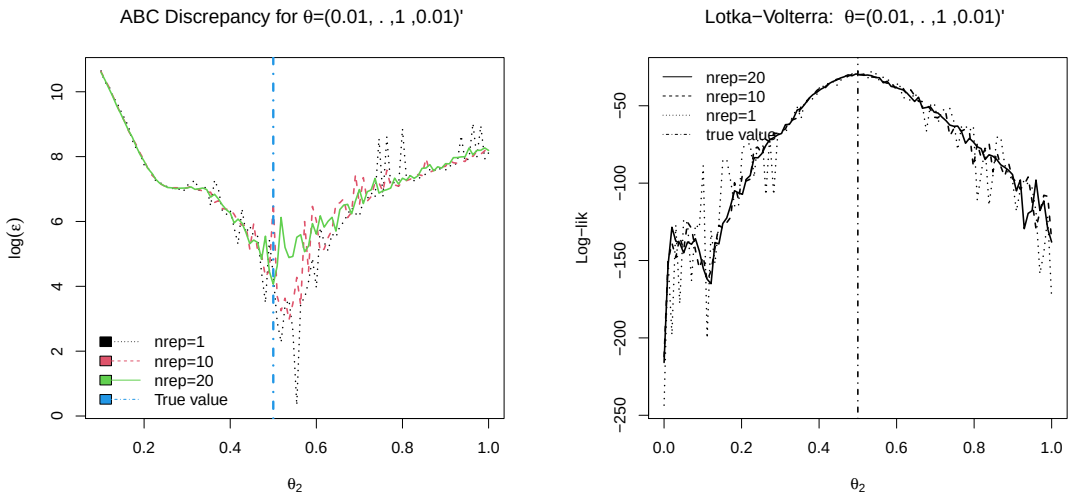
[9] suggest a marginal model trained ahead of the Monte Carlo simulation which compares dependent and independent data-parameter pairs. A related marginal technique is in [18]. We applied the technique of [9] using, again, the default random forest classifier. Due to the compact support of the parameters (a rather small subset of the cube $[0, 1]^4$), we can learn the likelihood surface quite well. If the parameters had an unbounded support, very many observations-parameter pairs would need to be generated and this would drastically increase the learning time. For example, [9] use 1 million training samples. However,

performing random forests on such a large dataset would not be practical. For our Lotka-Volterra model, we trained the classifier using $m = 10\,000$ and $m = 50\,000$ (which took roughly 2.7 hours). Additional time is needed for the actual MCMC sampling.

To see the effect of the fake data-set size m on the estimator of the (log)-likelihood, we plot the estimator obtained using the marginal reference [9] in Figure 18 in the right panel. We found that for the marginal approach, the interaction terms between parameters and data are essential for obtaining good prediction. This is why we did not choose the LASSO but a non-linear random forest classifier. For the marginal approach, we try $m \in \{5\,000, 10\,000, 50\,000\}$ using the default implementation of random forests (R package `randomForests`). With enough training samples, the estimator is quite smooth. However, as will be seen from histograms and traceplots (Figure 24 and Figure 20 below) there is certain bias in the posterior reconstruction. Choosing $m = 10\,000$, the estimator still peaks around the truth but is wigglier. The conditional approach of [15] also yields estimators peaked around the truth. The shape is similar to our fixed reference approach using random forests (Figure 17 on the right). However, both of these plots yield curves that are not nearly as peaked as with the `glmnet` classifier. This has at least two implications: (1) the Metropolis-Hastings with the `glmnet` classifier will be far more sensitive to initializations where we need to perhaps run ABC or other pilot run to obtain a satisfactory guess (see Figure 22), (2) if initialized properly and if the chain mixes well, the `glmnet` classifier might provide tighter credible intervals. The choice of the proposal distribution will be also important and it should reflect the curvature of these likelihood shapes.

To see whether our ABC summary statistics are able to capture the oscillatory behavior (at different frequencies) and distinguish it from exploding population growth, we have plotted the squared $\|\cdot\|_2$ distance of the summary statistics[6] (i.e. the ABC tolerance threshold ϵ) relative to the real data for a grid of values θ_2, fixing the rest at the true values $\theta_1^0 = 0.01, \theta_3^0 = 1, \theta_4^0 = 0.01$ (see Figure 19a). We can see a V-shaped evolution of ϵ reaching a minimum near the true value $\theta_2^0 = 0.5$, especially for $nrep = 20$. This creates hope that ABC based on these summary statistics has the capacity to provide a reliable posterior reconstruction. Contrastingly, we have plotted the estimated log-likelihood $\eta \equiv$

[6] Out of curiosity, we have considered a single fake dataset as well as the average tolerance over $nrep$ fake data replications.



(a) ABC tolerance ε (b) Log-lik Estimator (`rf`)

Figure 19: Lotka-Volterra model. ABC discrepancy ϵ and the classification-based log-likelihood 'estimator' η using observed data as a reference with the `randomForest` classifier.

$\sum_{i=1}^{n} \log[(1-\hat{D}(\boldsymbol{x}_i))/\hat{D}(\boldsymbol{x}_i)]$ (as a function of θ_2) where $\boldsymbol{x}_i = (X_1^i, \ldots, X_T^i, Y_1^i, \ldots, Y_T^i)'$ after training the LASSO-penalized logistic regression classifier (Figure 17 on the right) on $m = n$ fake data observations $\widetilde{\boldsymbol{x}}_i = (\widetilde{X}_1^i, \ldots, \widetilde{X}_T^i, \widetilde{Y}_1^i, \ldots, \widetilde{Y}_T^i)'$ for $1 \leq i \leq m$ using the cross-validated penalty λ (using the R package `glmnet`). We also use the default implementation of random forests using the R package `randomForest` (Figure 19 on the right). We can see that random forests provide estimators which are not as sharply peaked, suggesting less sensitivity to Markov chain initialization.

The trace-plots of MHC and the approaches of [15] and [9] are in Figures 20, 21 and 22). Figure 23 portrays histograms of ABC samples (top $r = 100$ out of $M = 10\,000$ in the upper panel and top $r = 1\,000$ out of $M = 100\,000$ in the lower panel). Finally, Figure 24 shows histograms of MH samples (MHC, Classification MCMC of [15] and ALR MH approach of [9]).

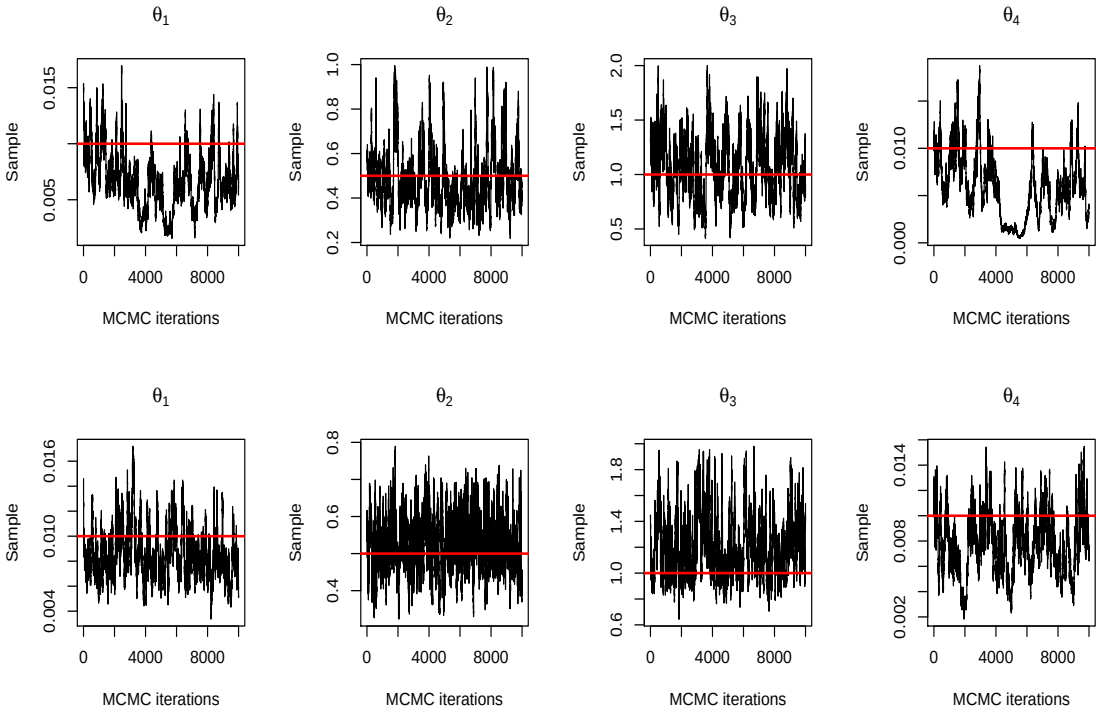


Figure 20: Traceplots of ALR MH of [9] with $m = 10\,000$ (top) and $m = 50\,000$ (bottom)

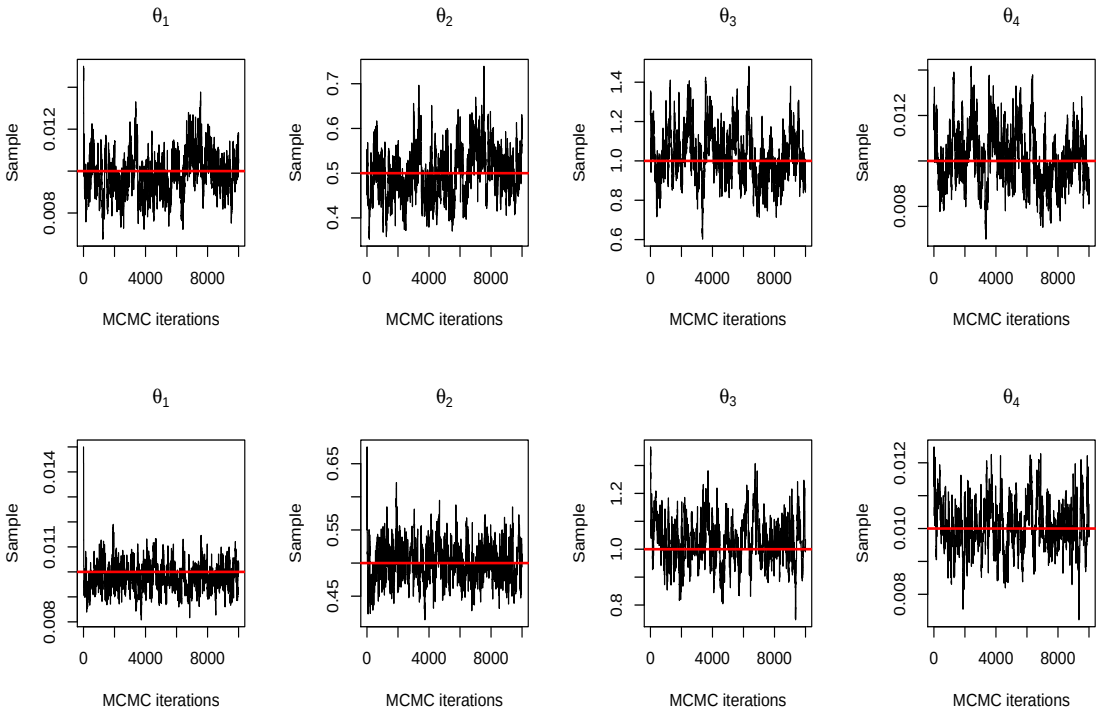


Figure 21: Traceplots of Classif MH of [15] with $m = 20$ (top) and $m = 100$ (bottom)

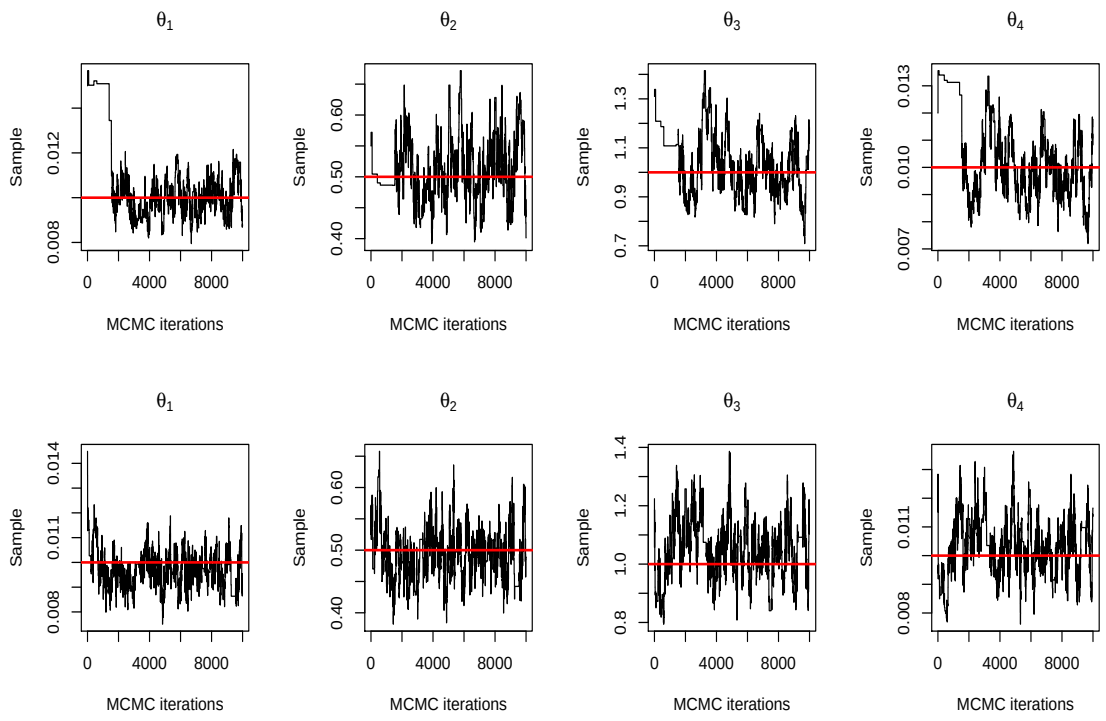


Figure 22: Traceplots of MHC with `glmnet` (top) and random forests (bottom)

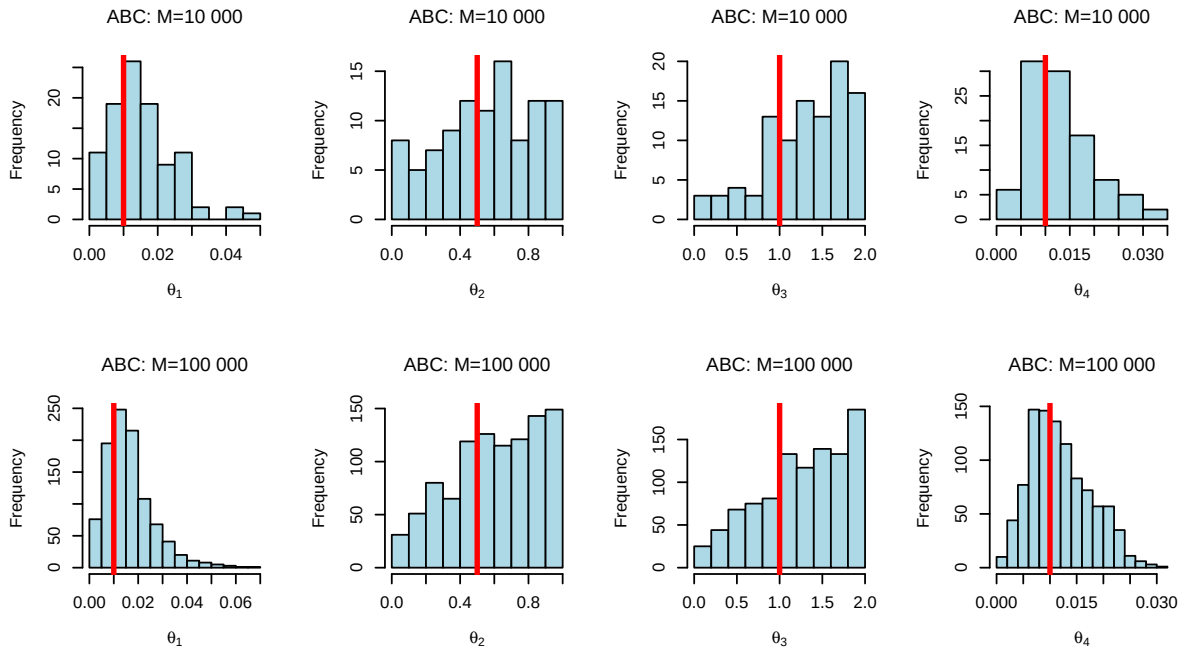


Figure 23: ABC analysis of the Lotka-Volterra model. Upper panel uses $M = 10\,000$ and $r = 100$ whereas the lower panel uses $M = 100\,000$ and $r = 1\,000$. Vertical red lines mark the true values.

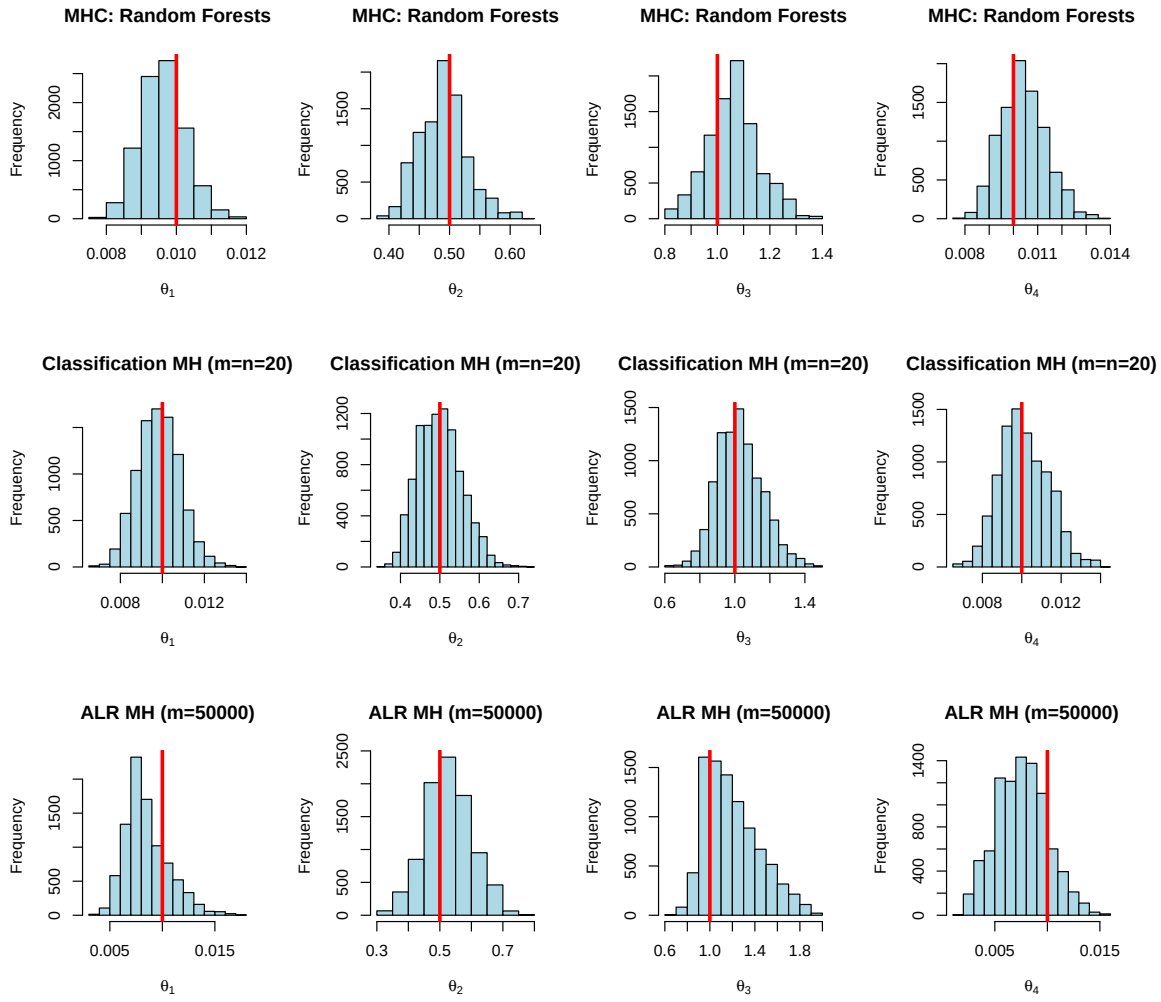


Figure 24: MH analysis of the Lotka-Volterra model (9 000 MCMC iterations after 1 000 burnin). Upper panel shows results for MHC with random forests, the middle panel uses the classification MCMC approach of [15] (using $m = n = 20$) and the lower panel is the ALR MH approach of [9]. Vertical red lines mark the true values.

References

[1] Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

[2] Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.

[3] Belloni, A. and Chernozhukov, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *The Annals of Statistics*, 37(4):2011–2055.

[4] Friedman, T., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):407–499.

[5] Ghosal, S. and van der Vaart, A. W. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223.

[6] Ghosal, S. and van der Vaart, A. W. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge.

[7] Gillespie, D. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

[8] Gutmann, M. and Hyvarinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361.

[9] Heermans, J., Begy, V., and Louppe, G. (2020). Likelihood-free MCMC with amortized approximate ratio estimators. *Proceedings of the* 37^{th} *International Conference on Machine Learning*, 119(37):15112–15117.

[10] Kaji, T., Manresa, E., and Pouliot, G. (2020). An adversarial approach to structural estimation. *arXiv*.

[11] Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34:837–877.

[12] Kleijn, B. and van der Vaart, A. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354 – 381.

[13] Lovasz, L. and Simonovits, M. (1993). Random walks in a convex body and an improved volume algorithm. *Random Structures and Algorithms*, 4:359–412.

[14] Mengersen, K. and Tweedie, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121.

[15] Pham, K., Nott, D., and Chaudhuri, S. (2014). A note on approximating ABC-MCMC using flexible classifiers. *The ISI's Journal for the Rapid Dissemination of Statistics Research*, 3:218–227.

[16] Robert, C. P., Cornuet, J.-M., Marin, J.-M., and Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117.

[17] Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press, Cambridge.

[18] Thomas, O., Dutta, R., Corander, J., Kaski, S., and Gutmann, M. (2021). Likelihood-free inference by ratio estimation. *Bayesian Analysis*, 1:1–31.

[19] Tsvetkov, D., Hristov, L., and Angelova-Slavova, R. (2017). On the convergence of the Metropolis-Hastings Markov chains. *Serdica Math. J.*, 43(2):93–110.

[20] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[21] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.