Caveats on Data Cloning

Brian Zaharatos* William Navidi†

The authors gratefully acknowledge Mark Campanelli and Luis Tenorio for their important comments on early stages of this work.

Abstract

For the maximum likelihood estimator (MLE) to be unique, the parameter must be both identifiable and estimable. A parameter is identifiable if there is a one-to-one correspondence between parameter values and density functions. A parameter is estimable if the likelihood function has a unique mode. The method of data cloning has been proposed as a way to diagnose structural deficiencies—such as non-identifiability and inestimability—in a model. In this paper, we discuss cases in which the number of clones required to detect model deficiencies may be impractically large, and provide guidelines for avoiding such cases.

Key words: identifiability, estimability, data cloning, maximum likelihood estimation.

1. Introduction

The method of maximum likelihood is the most commonly used method for estimating parameters in statistical models. For the maximum likelihood estimator (MLE) to be unique, the parameter must be both identifiable and estimable. There is some debate about how one ought to define parameter identifiability in different contexts (for example, see **?**, **?**, **?**). However, the most widely used definition is given in the statement of the classical regularity conditions (for a statement of these regularity conditions, see **?** or **?**). Given data $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ with a likelihood $L(\boldsymbol{\theta}|\boldsymbol{x})$, the model parameter $\boldsymbol{\theta}$ is said to be *identifiable* if, for all $\boldsymbol{\theta}_2, \boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$, $L(\boldsymbol{\theta}_2|\boldsymbol{x}) = L(\boldsymbol{\theta}_1|\boldsymbol{x})$ for almost all $\boldsymbol{x}$ implies that $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_1$. Otherwise, $\boldsymbol{\theta}$ is *non-identifiable*. In the context of regularity conditions, the term "estimability" used in **?** is not standard. However, the concept being referred to is that the likelihood has a unique global maximum. When the likelihood does not have a unique global maximum, $\boldsymbol{\theta}$ is said to be *inestimable*. Note that estimability can depend on the data.

For most models in applications, checking whether these conditions are satisfied is difficult to the point that such checks are rarely attempted (**?**). Data cloning, a technique originally developed in **?** to compute maximum likelihood estimates for linear mixed models and generalized linear mixed models, has been proposed as a way to check whether a model has an inestimable parameter. In this paper, we seek to promote the effective use of data cloning by discussing cases in which the number of clones required to detect multiple

*Department of Applied Mathematics University of Colorado Boulder
†Department of Applied Mathematics and Statistics, Colorado School of Mines

maxima in the likelihood function may be impractically large, and by providing guidelines for avoiding such cases. In Section **??**, we describe data cloning as a diagnostic tool, provide simple examples of multimodal likelihood functions (that arise from non-identifiable or inestimable parameters), and discuss an application of data cloning to photovoltaic (PV) performance modeling. The purpose of the discussion of PV performance modeling is to provide a brief example of a robust application of data cloning. Then, in Section **??** and Section **??**, we describe situations where the data cloning procedure maps a multimodal likelihood function onto a posterior distribution that is, for practical purposes, unimodal. In such cases, inferences made about the uniqueness of the MLE that are based on properties of the posterior distribution, such as those made through data cloning, can be misleading. Our descriptions include practical guidelines for avoiding these issues.

2. Data Cloning as a Diagnostic Tool

Suppose that the observations $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ arise from a model $f(\boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ is a vector of model parameters. In **?**, data cloning was proposed as a way of finding MLEs for cases where $f(\boldsymbol{x}; \boldsymbol{\theta})$ is intractable, including cases of general hierarchical models for which maximum likelihood estimation requires high-dimensional integration. To implement data cloning for the purpose of calculating MLEs, one develops a Bayesian model for the problem at hand and uses Markov chain Monte Carlo (MCMC) to compute MLEs or diagnose model inadequacies. We note that, although data cloning uses Bayesian tools, such as Bayes' Theorem and MCMC, it is not necessarily a "Bayesian method"; rather, it uses Bayesian computational tools to provide frequentist inferences, i.e., MLEs and their standard errors. Consequently, data cloning does not require any adherence to the assumptions of Bayesian inference (e.g., a Bayesian interpretation of probability or the modeling of parameters as random variables).

We can cast data cloning as a thought experiment: suppose that, by coincidence, one observed k independent identical samples of size n:

$$\boldsymbol{x}_{n_k} = \underbrace{(x_1, ..., x_n, ..., x_1, ..., x_n)}_{k \text{ independent repeats}}.$$

The resulting likelihood function would be $L_k(\boldsymbol{\theta}|\boldsymbol{x}_{n_k}) = [L(\boldsymbol{\theta}|\boldsymbol{x})]^k$. To construct a Bayesian model, we let $\pi(\boldsymbol{\theta})$ be any proper prior density for $\boldsymbol{\theta}$. The k^{th} *posterior distribution* is defined, up to a constant of proportionality, as

$$\pi_k(\boldsymbol{\theta}|\boldsymbol{x}) \propto L_k(\boldsymbol{\theta}|\boldsymbol{x}_{n_k})\pi(\boldsymbol{\theta}).$$

Properties of the k^{th} posterior distribution are informative about MLE theory and potential model inadequacies. For sufficiently large k, $\pi_k(\boldsymbol{\theta}|\boldsymbol{x})$ is nearly degenerate around $\widehat{\boldsymbol{\theta}}_{\text{ML}}$ with covariance matrix approximately equal to $\frac{1}{k}\mathcal{I}(\widehat{\boldsymbol{\theta}}_{\text{ML}})^{-1}$, where $\mathcal{I}(\widehat{\boldsymbol{\theta}}_{\text{ML}})$ is the *Fisher information matrix* evaluated at $\widehat{\boldsymbol{\theta}}_{\text{ML}}$, and is defined as

$$\mathcal{I}(\widehat{\boldsymbol{\theta}}_{\text{ML}}) = \mathbb{E}\Big[-\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\log(L(\boldsymbol{\theta}|\boldsymbol{x}))\Big]\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{ML}}}$$

(**?**). So, to calculate MLEs for $\boldsymbol{\theta}$, one can choose a sufficiently large value for k (some guidelines are given in **?**); the value for which the posterior is degenerate around is the MLE. Further, standard errors for the MLE can be calculated using $\frac{1}{k}\mathcal{I}(\widehat{\boldsymbol{\theta}}_{\text{ML}})^{-1}$.

The data cloning procedure also provides a way to assess whether the MLE is unique (**?**). If the variance of the k^{th} marginal posterior $\pi_k(\theta_i|\boldsymbol{x})$ does not converge to zero, then θ_i is inestimable. Further as k gets large, $\pi_k(\theta_i|\boldsymbol{x})$ converges to a truncated prior distribution, truncated over the space of MLEs. These results suggest that, in addition to providing MLEs, data cloning can also be used as a diagnostic tool for assessing parameter inestimability.

2.1 Clarifications on identifiability and estimability

Here, we provide several examples to help clarify the definitions for identifiability and estimability given in Section **??**. Data cloning makes direct claims about estimability and not necessarily identifiability; however, if the goal of the practitioner is to find evidence that the likelihood in question has multiple maxima—which can occur because of non-identifiability or inestimability—then data cloning is a useful tool and the distinction is of less importance. For each of the following examples, we let $i \in \{1, 2, ..., n\}$.

1. Let $X_i \sim N(\theta, 1)$. The unknown parameter θ is both identifiable and estimable.

2. Let $X_i \sim U(\theta,\ \theta + 1)$. The unknown parameter θ is identifiable. However, θ is inestimable because any value in the interval $\big[\max\{\boldsymbol{x}\} - 1, \min\{\boldsymbol{x}\}\big]$ maximizes the model's likelihood function.

3. Let $Y_i \sim N(\theta_1\theta_2, 1)$. Neither θ_1 nor θ_2 is identifiable or estimable; however, the product $\theta_1\,\theta_2$ is both identifiable and estimable.

4. Let $X_i \sim N(|\theta|, 1)$. The unknown parameter θ is not identifiable because for $\theta_1 = -\theta_2$, $L(\theta_2, \sigma^2|\boldsymbol{x}) = L(\theta_1, \sigma^2|\boldsymbol{x})$. θ is inestimable because both $\widehat{\theta}_{\text{ML}}$ and $-\widehat{\theta}_{\text{ML}}$ maximize the likelihood function.

Models with a non-identifiable parameter admit multiple values for that parameter that explain the data equally well. Without independent a priori knowledge about the problem under investigation, there is no way of knowing which value of the parameter actually generated the data. Attempting to estimate an inestimable model parameter may yield disagreements between different optimization methods using the same data. Non-identifiability and inestimability may have important practical and philosophical implications.

2.2 Data cloning—a robust application

Data cloning has been used as a diagnostic tool in a number of robust applications. To illustrate the fact that data cloning can provide evidence of multimodal likelihoods in complex models, we present an application in modeling the performance of a photovoltaic (PV) device, which converts solar energy into electricity. For other examples of data cloning applications, see **?**, **?**, and **?**.

One method for determining the performance of a PV device is to measure current (I), voltage (V), and operating conditions temperature (T), and "effective" irradiance (E), and use these measurements in a model to estimate important performance parameters. These parameters include a device's short circuit current (I_{SC}), diode reverse saturation current (I_{S}), ideality factor (n), series resistance (R_{S}), parallel resistance (R_{P}), open circuit voltage (V_{OC}), and maximum power output (P_{max}). These parameters have physical importance. For example, I_{SC} represents the largest current that can be drawn from the PV cell. For a large number of PV devices, the *single-diode model* describes the relationship between the measured data (I, V, E) and parameters, and is thought to strike a good compromise between accuracy and simplicity for modeling PV performance (**?**). The authors of **?** and **?** show that for parameter estimation purposes, at a fixed temperature of $T = T_0 = 25°\text{C}$, and for values of E near 1000 W/m^2, the nonlinear and implicitly defined single-diode PV performance model can be expressed as $E = g_E(I, V; \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (I_{\text{SC}}, I_{\text{S}}, n, R_{\text{S}}, R_{\text{P}})$ is a vector of parameters to be estimated from measurements of I, V, and E. The parameters V_{OC} and P_{max} are functions of $\boldsymbol{\theta}$.

Under the (often reasonable) assumption that measurements of I and V contain negligible error, and that the measurement error in E is log-normal, we get that, for N measurements and $i \in \{1, ..., N\}$, the observable E_i is modeled by the random variable

$$\mathcal{E}_i \sim \text{lognormal}\Big(\log(g_E(I_i, V_i; \boldsymbol{\theta})), \sigma_E^2\Big).$$

The parameter $\boldsymbol{\theta}$ may be estimated by maximum likelihood, but given the complexity of g_E, it is unclear whether the likelihood has a unique maximum. Data cloning has been

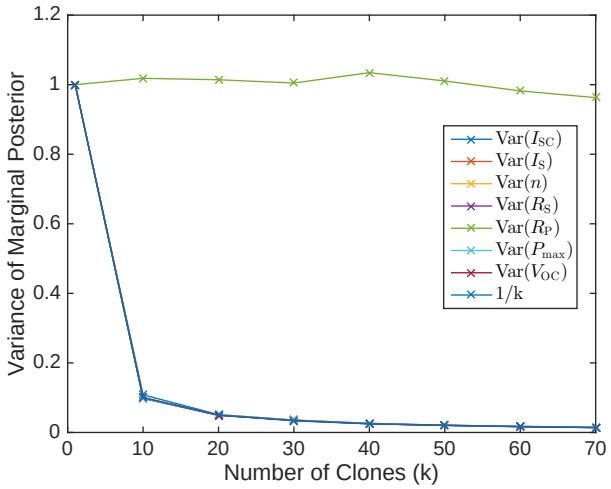

Figure 1: The variance of the marginal posterior distribution for $\boldsymbol{\theta}$ and the corresponding key performance parameters $P_{\text{MAX}} = P_{\text{MAX}}(\boldsymbol{\theta})$ and $V_{\text{OC}} = V_{\text{OC}}(\boldsymbol{\theta})$.

used to understand this problem.

For several synthetically generated datasets—i.e., data generated using pre-determined values of $\boldsymbol{\theta}$ from different regions of the parameter space—data cloning was performed for diagnostic purposes (for more details on the generation of the synthetic datasets and implementation of data cloning, see **?**). It was shown that, in some regions of the parameter space—specifically, for some values of R_{P}—there is evidence that likelihood function has multiple maxima. Figure 1 shows the standardized variance of the marginal posterior distribution for $\boldsymbol{\theta}$ and the corresponding key performance parameters $P_{\text{MAX}} = P_{\text{MAX}}(\boldsymbol{\theta})$ and $V_{\text{OC}} = V_{\text{OC}}(\boldsymbol{\theta})$. The variance of $\pi_k(R_{\text{P}}|\boldsymbol{E})$ does not decay at rate $1/k$, providing evidence that the likelihood function has multiple maxima.

3. Choosing a Prior for Detecting Inestimability

In principle, any proper prior can be used to detect multiple maxima, because the variance of the posterior distribution will not converge to 0 when the likelihood has multiple maxima. In practice, however, if the prior mass around one maximum is considerably larger than the masses around the others, the number of clones required to detect lack of convergence of the posterior variance may be impractically large.

To see this, let $X_1, ..., X_n$ be jointly distributed with likelihood function $L(\theta \,|\, \boldsymbol{x})$ of a single parameter θ. Assume there are two equal maxima, at at m_1 and m_2. Let $\pi(\theta)$ be any continuous prior. For a large number k of clones, the posterior distribution will be approximately equal to a mixture of distributions with means m_1 and m_2, and variances σ_1^2/nk

and σ_1^2/nk for constants σ_1 and σ_2. The weights on the two distributions are proportional to $\pi(m_1)$ and $\pi(m_2)$. Let $p = \pi(m_1)/[\pi(m_1) + \pi(m_2)]$. The variance of the posterior distribution is approximated by

$$V(\theta \mid \boldsymbol{x}^{(k)}) = \frac{p\sigma_1^2}{nk} + \frac{(1-p)\sigma_2^2}{nk} + p(1-p)(m_1 - m_2)^2. \tag{1}$$

By analogy with analysis of variance, we may refer to the the sum of the first two terms as the *within groups* variance and the last term as the *between groups* variance. As $k \to \infty$, the posterior variance converges to the between groups variance, rather than 0. If the ratio $\pi(m_1)/\pi(m_2)$ is very large or very small, the between groups variance may be for practical purposes indistinguishable from 0.

For a specific example, let $X_1, ..., X_n$ be i.i.d. $N(|\theta|, 1)$, and assume we observe $\bar{X} = 4$. This model is clearly nonidentifiable, because the distributions specified by θ and $-\theta$ are the same. The likelihood has equal maxima at ± 4. Figure 2 presents the likelihood function for $n = 10$. Now consider the prior $\pi(\theta) = N(5, 1)$. The prior masses at the maxima are $\pi(4) = e^{-1/2}/\sqrt{2\pi}$ and $\pi(-4) = e^{-81/2}/\sqrt{2\pi}$. Figure 3 presents the posterior distribution. The mode at -4 is undetectable.

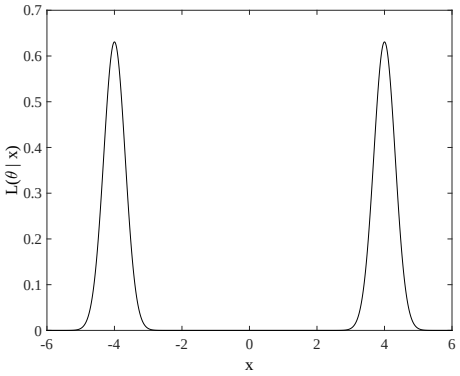

Figure 2

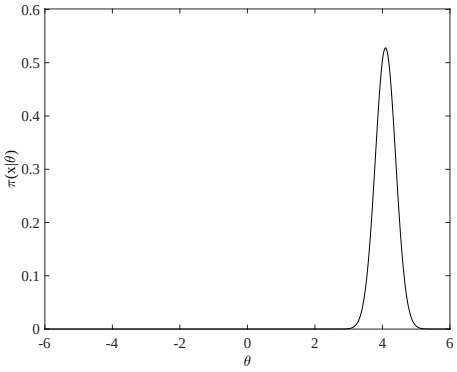

Figure 3

For a reasonably large number k of clones, the posterior variance is well approximated by

$$V(k) = \frac{1}{nk} + 2.7 \times 10^{-16}.$$

The authors of **?** suggest plotting $V(k)/V(1)$ against k and comparing it with a plot of $1/k$ to detect inestimability. Figure 4 presents such a plot. The curves are indistinguishable.

This problem can be avoided in any of several ways. One can use a uniform prior over a region that one feels is certain to contain all maxima of the likelihood function. Another possibility is to use a diffuse normal prior. Finally, as suggested by **?**, one may repeat the process with several priors concentrated on different regions of the parameter space. We present results for the previous example, with the $N(5, 1)$ prior replaced with a uniform

prior on $(-1000, 1000)$. The posterior variance is now approximately equal to $1/nk + 16$. Figure 5 presents the plot of $V(k)/V(1)$ superimposed on a plot of $1/k$, for $n = 10$. Failure of the posterior variance to converge to 0 is clear.

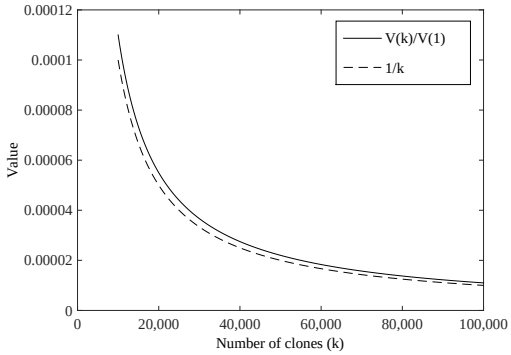

Figure 4

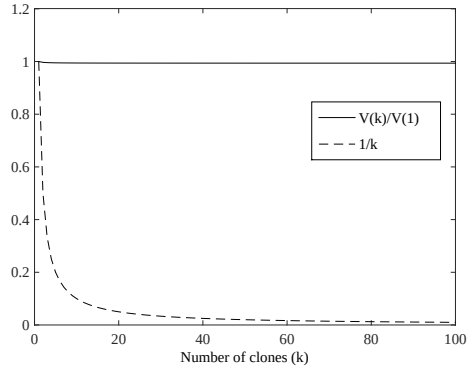

Figure 5

4. Choosing an MCMC Algorithm for Detecting Inestimability

In most applications of data cloning, the k^{th} posterior distribution is estimated using Monte Carlo Markov Chains (MCMC). In some cases, the tuning of the MCMC algorithm affects the data cloning diagnostic results. One of the most common MCMC algorithms used to estimate a posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{x})$, is the Metropolis-Hastings (MH) algorithm. MH produces a Markov chain whose stationary distribution is the desired (target) posterior distribution by the following algorithm:

1. set $t = 1$ and choose an initial state of the chain, $\boldsymbol{\theta}^{(0)}$;
2. randomly choose a new state of the chain, $\boldsymbol{\theta}^*$, from a proposal distribution, $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$;
3. calculate the acceptance probability

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{x})q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{x})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})}\right\};$$

4. draw $u \sim U(0, 1)$;
5. if $u \leq \alpha$, then set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$; otherwise, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$;
6. set $t = t + 1$ and repeat steps 2–6 until $t = T$.

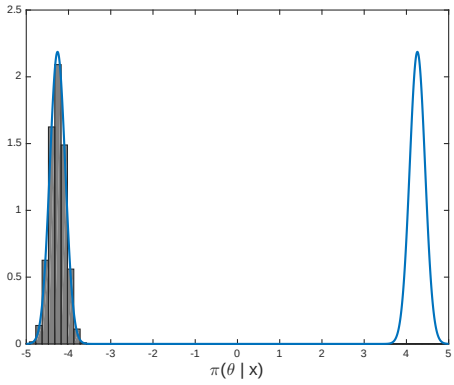

Figure 4

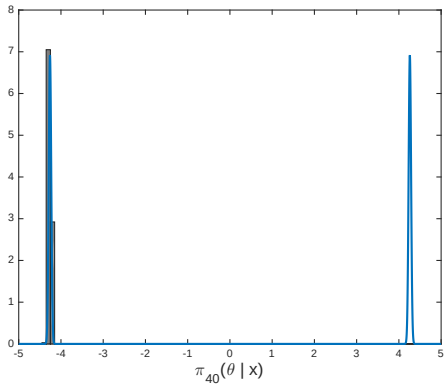

Figure 5

If T is large enough, the resulting chain, after some "burn in" period, say, $\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^{(s+1)}, ..., \boldsymbol{\theta}^{(T)}$, will be a sample from the posterior distribution.

To gain some intuition about MH, suppose that the proposal distribution is symmetric around the value of a parameter (e.g., normal). Then, $q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$ and

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{x})}{\pi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{x})}\right\}.$$

If $\pi(\boldsymbol{\theta}^*|\boldsymbol{x}) \geq \pi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{x})$, then $\alpha = 1$ and the proposed step is accepted as part of the chain. If $\pi(\boldsymbol{\theta}^*|\boldsymbol{x}) < \pi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{x})$ then $0 \leq \alpha < 1$ and the proposed step is accepted sometimes and rejected others; this ensures that the full density can be explored.

MH requires that one choose the initial state of the chain, $\boldsymbol{\theta}^{(0)}$, and that one "tune" the proposal distribution such that the chain converges to the posterior in a reasonable amount of time. If the proposal distribution is normal, then tuning amounts to choosing a reasonable proposal variance-covariance matrix. If the proposal variance is set too large or small, the Markov chain will not converge to the posterior in a reasonable amount of time.

In practice, especially in high-dimensional problems, tuning the MH algorithm can be difficult. In particular, for certain multimodal target distributions and certain tunings of MH, the number of simulations needed for MH to discover multiple modes of $\pi_k(\boldsymbol{\theta}|\boldsymbol{x})$ may be impractically large. If, for each clone k, the Markov chain produced by MH only detects one mode of $\pi_k(\boldsymbol{\theta}|\boldsymbol{x})$, the data cloning procedure would incorrectly conclude that the likelihood is unimodal.

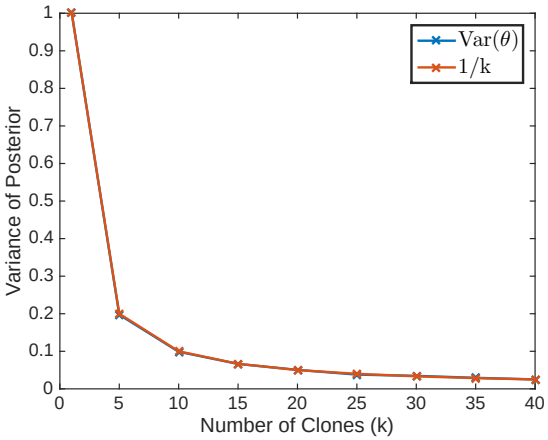

Figure 7

For a specific example, consider again $X_1, ..., X_n$ i.i.d. from $N(|\theta|, 1)$, and assume we observe $\bar{X} = 4$. Now consider the prior $\pi(\theta) = N(0, 5)$ and proposal variance 1. Figure 6 presents the true posterior distribution and histogram from the MH simulations for $k = 1$ and $k = 40$. Note that only one peak is explored by the Markov chain. Figure 7 presents the plot of the standardized variance of $\pi_k(\theta|\boldsymbol{x})$ and the curve $1/k$ against k. The curves are indistinguishable.

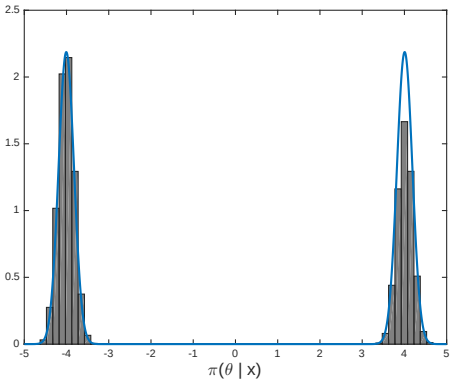

Figure 4

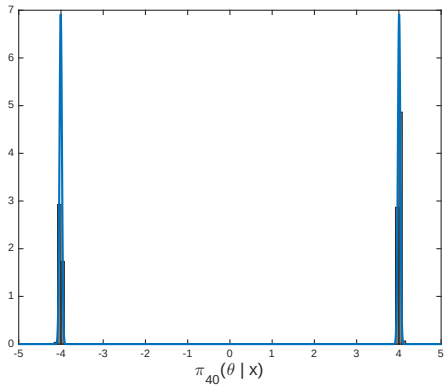

Figure 5

To fix this issue, one can choose the proposal variance in a way that sufficiently explores the target posterior distribution. To sufficiently explore a multimodal posterior distribution, such as the one from the example in the previous paragraph, the proposal variance should be large enough to make jumps between modes. When adjusting the proposal variance of 5 in this example, we see that data cloning does detect both modes, as in Figure 8. In this case, data cloning correctly diagnoses multimodality (as seen in the variance plot in Figure 9).

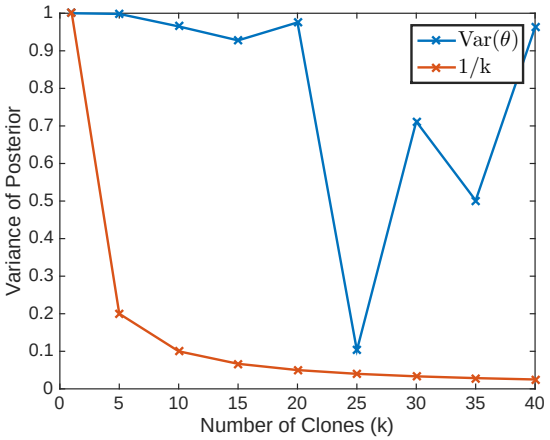

Figure 9

Another set of commonly used MCMC algorithms adapt the proposal variance-covariance matrix at certain steps in the chain. One such algorithm, adaptive metropolis (AM), is an attempt to optimize the rate of convergence of MH to the target posterior distribution by adapting the variance-covariance matrix of the proposal distribution based on the history of the chain. At step t of the chain, the normal proposal distribution is set so its mean is the current position of the chain, $\theta^{(t)}$, and its covariance is set to be

$$C^{(t)} = s_d \text{Cov}(\theta_0, \theta_1, ..., \theta_{t-1}) + s_d \varepsilon I_d,$$

where s_d is a parameter that depends on the dimension of the state space[1], I_d is the d-dimensional identity matrix, and $\varepsilon > 0$ is a constant chosen to be very small (ε ensures that C_t does not become singular). Although the AM modification of MH can be helpful in tuning MH, adaptations of MH are often not well-equipped to explore multimodal posterior distributions. The authors of **?** claim that

> the use of a multivariate normal proposal distribution with covariance matrix adaptation works well for Gaussian-shaped target distributions, but cannot sample adequately multimodal distributions with long tails...it is relatively easy for a single chain to become stuck in a local mode and common diagnostics would not detect that the chain has not explored adequately the full posterior model.

As with MH, we recommend that practitioners use a large initial proposal variance-covariance for AM or similar adaptive algorithms. Large, of course, depends on the given

[1] the authors of **?** show that $s_d = 2.42/d$ is often optimal.

application. Thus, we also recommend normalizing the parameter space before implementing data cloning—e.g., performing a transformation so that each parameter lies in the interval $[-1, 1]$. Further, we note that some MCMC algorithms are well-suited to estimating multimodal posterior distributions. For example, the Differential Evolution Markov Chain (DREAM) algorithm has been shown to perform well for Bayesian inference problems where the posterior distribution is multimodal.

5. Concluding Remarks

The method of data cloning can be a useful tool for calculating MLEs and diagnosing certain model inadequacies. In this paper, we have shown that data cloning can be a useful diagnostic tool for detecting whether a likelihood function has more than one maximum, but because of practical constraints on implementation, in some cases data cloning can lead practitioners astray. In particular, certain prior distributions and MCMC algorithms map a multimodal likelihood to a posterior distribution that is, for all practical purposes, unimodal. In such cases, inferences made about the uniqueness of the MLE that are based on properties of the posterior distribution, such as those made through data cloning, can be misleading.

Because there are cases in which data cloning may yield misleading results, we recommend that practitioners think carefully about their choice of priors and MCMC algorithm. In particular, priors with large variances centered at values near an MLE will work best. Further, with respect to the choice of MCMC algorithm, we recommend using a larger proposal variance for MH, or a larger initial proposal variance for adaptive methods. Large, of course, depends on the given application, and so we also recommend normalizing the parameter space before implementing data cloning

Finally, we also suggest being cautious with the language used in describing the results of data cloning. As our methodological investigations in Section **??** and Section **??** indicate, the conclusions based on data cloning may be mistaken. So, we suggest the following:

1. Degeneracy of the k^{th} posterior distribution in the data cloning procedure does not entail that the model in question has estimable (or identifiable) parameters. Following the language of hypothesis testing, we suggest that, in the case of degeneracy, the proper conclusion is that there is "*no evidence* of inestimability" of the model parameter in question.

2. Similarly, non-degeneracy of the k^{th} posterior distribution in the data cloning procedure does not entail that the model in question has inestimable (or non-identifiable)

parameters. It only provides *evidence* of inestimability.