REVSTAT – Statistical Journal
Volume 22, Number 1, January 2024, 87–110
https://doi.org/10.57805/revstat.v22i1.457

Identifiability Analysis Using Data Cloning

Authors: José Augusto Sartori Junior ✉
– Institute of Mathematics and Statistics, University of São Paulo, Brazil
jsartori@ime.usp.br

Márcia D'Elia Branco
– Institute of Mathematics and Statistics, University of São Paulo, Brazil
mbranco@ime.usp.br

Received: March 2021 Revised: March 2022 Accepted: March 2022

Abstract:

- Lack of identifiability in statistical models may hinder unique inferential conclusions. Therefore, the search for parametric constraints that ensure identifiability is of utmost importance in statistics. However, for complex modeling strategies, even acquiring the knowledge that the model is unidentifiable may prove very difficult. In this paper, we investigate the use of Data Cloning, a modern algorithm for classical inference in latent variable models, as a tool for assessing model identifiability. We discuss its advantages and disadvantages and illustrate its use with a dynamic linear model.

Keywords:

- *identifiability; data cloning; dynamic models; MCMC algorithms.*

AMS Subject Classification:

- 62A99, 62F10.

✉ Corresponding author.

1. INTRODUCTION

The specification of identifiable statistical models is an extremely important step in statistical inference. Data-driven decisions in unidentifiable models may be non-unique, *i.e.* it may not be possible to choose a single optimal decision based solely on the data at hand (San Martín 2018 [22]). Therefore, be it in the classical or Bayesian framework, unidentifiability may lead to severely wrong answers to scientific inquiries.

In classical statistics, for instance, lack of identifiability implies there does not exist a consistent estimator for some or all of the model parameters (Paulino and Pereira 1994 [15]). In other words, no matter how large the sample size, with an unidentifiable model we will never be able to distinguish the true parameter value from at least one other alternative value.

Unidentifiability can also cause problems in the Bayesian setting. When using flat prior distributions for unidentifiable parameters, the resulting posterior can still be flat. Moreover, if the prior is improper, then the posterior may also be improper (Lindley 1971 [14]). However, even if informative priors are used in this situation, it is not very clear what inferences can be drawn *a posteriori*. San Martín (2018) [22] argues for inference on the sufficient parameter (see meaning therein) and shows how the influence of the prior distribution never vanishes for unidentifiable parameters.

Most of the models in the statisticians' basic toolkit enjoy solid theoretical foundations. However, the recent advances in computational power have led to the appearance of more complex models for which identifiability is not always guaranteed. Earlier, the theoretical development of statistics was followed by the studies of computational feasibility of the models, however, now a lot of theoretical work is to understand the properties of the newer modeling strategies.

Due to the difficulty in answering the question of whether or not a particular statistical model is identifiable, there is a sizable literature suggesting a diverse range of methods. There have been approaches using differential geometry (Villaverde *et al.* 2019 [25]), differential algebra (Bellu *et al.* 2007 [3]), measure theory (San Martín 2015 [21]) etc. However, sometimes theory alone does not end the issue and computational strategies are called upon to provide empirical evidence of model identifiability.

A closely related concept is that of estimability or practical identifiability. It may be the case that the statistical model is identifiable, but the data available is of poor quality or the model has been incorrectly specified. This may hinder the ability to estimate the model parameters and the uncertainty associated with such estimates, which can affect both inferential and predictive tasks. In other words, estimability deals with the question of whether the data at hand can reliably estimate the desired quantities. Identifiability, however, is concerned with the existence, for any two distinct parameter values, of a hypothetical data set which can differentiate between them. As such, lack of estimability does not imply lack of identifiability, although the converse is always true (Paulino and Pereira 1994 [15]).

Lack of estimability is commonly caused by low signal-to-noise ratio in the data, low sample size, or inappropriate sampling scheme (Lele *et al.* 2010 [13]). These problems often result in a likelihood function of the parameters that have many local maxima or an almost flat region.

In these scenarios, any given estimation algorithm might result in parameter estimates which yield considerably distinct inferential conclusions. Furthermore, confidence regions may present one or more coordinates assuming unreasonably large values.

Recently, Lele *et al.* (2007) [12] proposed an algorithm for maximum likelihood estimation, called Data Cloning, which is of particular relevance in latent variable models. The method is quite intuitive and draws motivation from the idea of replicability of experiments in frequentist statistics. To be more specific, the data cloning algorithm starts from a prior distribution on the parameter space and sequentially updates it using the same data set until some diagnostic measures reach specified thresholds.

As Lele *et al.* (2010) [13] show, if the model is unidentifiable then convergence issues can be easily spotted with the tools for diagnosing convergence of the algorithm. Data cloning has been used earlier for studying model estimability. Campbell and Lele (2014) [4]) proposed an ANOVA test of estimability based on data cloning and Peacock *et al.* (2017) [16] employed data cloning to assess estimability of a spatio-temporal model under distinct study designs based on an observed data set.

Our objective in this paper is to introduce data cloning as a practical tool for the assessment of identifiability of statistical models. For this, we show how to plan and perform a simulation study that can shed light on possible problems in the structure of the statistical model. Our idea is somewhat similar to that of Peacock *et al.* (2017) [16] in that we also employ simulated data. However, instead of exploring possible alternative study designs based on observed data, we advocate the exploration of a multitude of possible data based on as many as possible parameter values to study the structure of the model itself.

In Section 2 we present the data cloning algorithm and its main diagnostic measures that can be used to study model identifiability. In Section 3 we present the formal definitions of identifiability of statistical models, relate them to data cloning, and show, theoretically, how the identifiability issue reveals itself in the Gaussian dynamic linear model. Finally, in Section 4 we present a simulation study using the package *dclone* (Solymos 2010 [23]) from *R* (R Core Team 2020 [19]) and *JAGS* (Plummer 2017 [18]), and discuss the evidence it brings about identifiability in the adopted model.

2. DATA CLONING

In the subjective realm of Bayesian inference, a great deal of discomfort in the prior specification vanishes for highly informative data. An important result in Bayesian asymptotics, due to Walker (1969) [26], shows that, under some regularity conditions, for large n the posterior distribution $\pi(\boldsymbol{\theta}|y_1, ..., y_n)$ is approximately Gaussian with mean $\hat{\boldsymbol{\theta}}$, the maximum likelihood estimate of $\boldsymbol{\theta}$, and covariance matrix $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$, the inverse of the Fisher information evaluated at this maximum. For this, see also Turkman *et al.* (2019), Sec. 8.1 [24].

Suppose we performed an experiment k times independently and happened to observe the exact same realization $\mathbf{y}^{(j)} = \mathbf{y} = (y_1, ..., y_n)$ for all $j \in \{1, ..., k\}$ with probability density function $f(\mathbf{y}|\boldsymbol{\theta})$ for each experiment. Let $\pi_k(\boldsymbol{\theta}|\mathbf{y})$ denote the posterior distribution updated with samples for k such experiments. Since the k experiments are independent, Bayes theorem says this distribution is

$$\pi_k(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \prod_{j=1}^{k} f(\mathbf{y}|\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta})\{f(\mathbf{y}|\boldsymbol{\theta})\}^k \quad . \tag{2.1}$$

Let $L(\boldsymbol{\theta}; \mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)})$ and $l(\boldsymbol{\theta}; \mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)})$ denote the likelihood and log-likelihood functions of these experiments, respectively. If $\hat{\boldsymbol{\theta}}_{(n)}$ is maximum likelihood estimate under any one of the k experiments, then it follows immediately that

$$\hat{\boldsymbol{\theta}} = \arg\sup_{\theta\in\Theta} L(\boldsymbol{\theta}; \mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)}) = \arg\sup_{\theta\in\Theta} [L(\boldsymbol{\theta}; \mathbf{y})]^k = \arg\sup_{\theta\in\Theta} L(\boldsymbol{\theta}; \mathbf{y}) = \hat{\boldsymbol{\theta}}_{(n)} \tag{2.2}$$

and if we let $\mathbb{V}(\mathbf{X})$ denote the covariance matrix of a random vector $\mathbf{X}$, then

$$\mathcal{I}(\hat{\boldsymbol{\theta}}) = \mathbb{V}\left(\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y}^{(1)}, ..., \mathbf{Y}^{(k)})}{\partial \boldsymbol{\theta}}\right)\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbb{V}\left(\sum_{j=1}^{k} \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y}^{(j)})}{\partial \boldsymbol{\theta}}\right)\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = k\mathcal{I}(\hat{\boldsymbol{\theta}}_{(n)}) \quad , \tag{2.3}$$

in which the last equality follows from the independence of the experiments. Therefore, for a fixed n and k arbitrarily large, our posterior distribution would be well approximated by a Gaussian distribution with mean $\hat{\boldsymbol{\theta}}_{(n)}$ and covariance matrix $\frac{1}{k}\mathcal{I}^{-1}\left(\hat{\boldsymbol{\theta}}_{(n)}\right)$.

Note that we have not made a single comment about what prior $\pi(\boldsymbol{\theta})$ we started with. In fact, the previous results are valid as long as the prior distribution and the likelihood function satisfy some mild regularity conditions. In other words, for any two such priors π_1 and π_2 over Θ, there is a number of experiments, k, for which the posterior distributions would be arbitrarily close to each other (Lele *et al.* 2010 [13]).

Similarly, with minor modifications, the results above can be applied to latent variable models. Since the experiments are performed independently, realizations of the hidden stochastic process $\{\mathbf{X}^{(j)}\}$, $j \in \{1, ..., k\}$, are also assumed to have occurred k times independently. We begin by assigning a joint prior distribution $\pi(\boldsymbol{\theta}, \mathbf{x}) = \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ for the parameters and latent variables. The resulting posterior distribution is given by

$$\pi_k(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}) \prod_{j=1}^{k} f(\mathbf{y}^{(j)}|\mathbf{x}^{(j)}, \boldsymbol{\theta})\pi(\mathbf{x}^{(j)}|\boldsymbol{\theta})}{f(\mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)})} \quad . \tag{2.4}$$

For inference on the parameter vector, it suffices to marginalize on $\mathbf{x}$, which is made easier by the assumption of independence:

$$\begin{aligned}
\pi_k(\boldsymbol{\theta}|\mathbf{y}) &= \int_{\mathcal{X}} \pi(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}^{(j)}, ..., \mathbf{y}^{(k)}) d\mathbf{x} \\
&= \frac{\left\{\int_{\mathcal{X}} \prod_{j=1}^{k} f(\mathbf{y}|\mathbf{x}^{(j)}, \boldsymbol{\theta})\pi(\mathbf{x}^{(j)}|\boldsymbol{\theta}) d\mathbf{x}^{(j)}\right\}\pi(\boldsymbol{\theta})}{f(\mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)})} \\
&= \frac{\left\{\prod_{j=1}^{k} L(\boldsymbol{\theta}; \mathbf{y})\right\}\pi(\boldsymbol{\theta})}{f(\mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)})} \\
&= \frac{\{L(\boldsymbol{\theta}; \mathbf{y})\}^k \pi(\boldsymbol{\theta})}{\int_{\Theta} \{L(\boldsymbol{\theta}; \mathbf{y})\}^k \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \quad . \qquad (2.5)
\end{aligned}$$

In summary, if we obtained such an odd data set over a very large number of independent experiments, from (almost) arbitrary initial prior distributions, we could perform frequentist inference within the Bayesian setting. All that is required is we take samples from the posterior distribution of $\boldsymbol{\theta}$ and compute the mean vector and covariance matrix.

At first glance, it may seem that we replaced the high-dimensional integral required for maximum likelihood estimation, namely $L(\boldsymbol{\theta}; \mathbf{y})$, with a possibly much more complicated one in the denominator of (2.5). However, commonly employed Bayesian software for probabilistic sampling avoid the integration procedure altogether, which surely is an important reason why there is increasing adoption of the subjective paradigm amongst researchers dealing with complex latent variable models (Lele *et al.* 2010 [13]).

2.1. The Algorithm

Admittedly, repeating the same experiment may be just as infeasible as simply increasing the sample size. Thus, given a dataset, Lele *et al.* (2007) [12] propose that we clone it k times, with k as large as is computationally possible, and then draw samples from the k-times cloned posterior π_k. Although what we are using is in fact fake data, the machine on which the sampling algorithm will run cannot tell the difference.

As before, let $\mathbf{y} = (y_1, ..., y_n)$ denote the available realization of a measurement process generated by a hidden latent process. Suppose $\boldsymbol{\theta}$ is a continuous random vector, $p(\boldsymbol{\theta})$ is a proposal distribution, and define the burn-in length $N_{burn} < N_{sim}$, the simulation length for the Metropolis-Hastings algorithm. Algorithm 1 provides a way to sample from π_k; for the regularity conditions we direct the reader to Lele *et al.* (2010) [13].

Algorithm 1: Data Cloning Metropolis-Hastings (Lele *et al.* 2007 [12])

```
Generate $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$ and $\mathbf{x}^{*(1)}, ..., \mathbf{x}^{*(k)} \sim \pi(\mathbf{x}|\boldsymbol{\theta}^*)$;
for $l \in \{1, ..., N_{sim}\}$ do
    Compute $q^* = \prod_{j=1}^{k} f(\mathbf{y}|\mathbf{x}^{*(j)}, \boldsymbol{\theta}^*)$;
    Generate $\boldsymbol{\theta}^{\#} \sim p(\boldsymbol{\theta})$ and $\mathbf{x}^{\#(1)}, ..., \mathbf{x}^{\#(k)} \sim \pi(\mathbf{x}|\boldsymbol{\theta}^{\#})$;
    Compute $q^{\#} = \prod_{j=1}^{k} f(\mathbf{y}|\mathbf{x}^{\#(j)}, \boldsymbol{\theta}^{\#})$;
    Generate $U \sim \text{Uniform}(0, 1)$;
    if $U < \min\{1, q^{\#}/q^*\}$ then
        Set $(\boldsymbol{\theta}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(k)})_l = (\boldsymbol{\theta}^{\#}, \mathbf{x}^{\#(1)}, ..., \mathbf{x}^{\#(k)})$;
    else
        Set $(\boldsymbol{\theta}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(k)})_l = (\boldsymbol{\theta}^*, \mathbf{x}^{*(1)}, ..., \mathbf{x}^{*(k)})$;
    end
end
Discard $(\boldsymbol{\theta}, \mathbf{x}^{(j)}, ..., \mathbf{x}^{(k)})_1, ..., (\boldsymbol{\theta}, \mathbf{x}^{(j)}, ..., \mathbf{x}^{(k)})_{N_{burn}-1}$;
```

As long as the number of clones k is large enough and the regularity conditions are satisfied, the mean of the samples drawn from Algorithm 1 is a numerical approximation to the maximum likelihood estimate. Also, their covariance matrix is the inverse of the k-times scaled observed Fisher information matrix. As pointed out in Lele *et al.* (2007) [12], increasing the number of clones provides better numerical accuracy in the estimates. However, this does increase the computational cost of the algorithm considerably, since for each new clone we must simulate the latent process generating it. For models in which the number of latent variables grows exponentially with the sample size, the data cloning algorithm will surely demand an unreasonable amount of computational time. Nevertheless, it performs incredibly well for longitudinal and time series data, since the number of latent variables is usually on the order of the sample size.

2.2. Convergence Diagnostics

Another great feature of data cloning is the plethora of diagnostic measures available. On one hand, since we are using probabilistic sampling algorithms, it is mandatory to diagnose convergence of the sampling algorithm itself. For this, measures such as the potential scale reduction factor $\hat{R}$ (Gelman and Rubin 1992 [8]) and the effective sample size N_{eff}, for example, assess convergence of the Markov chains and the within-chains autocorrelations, respectively. For these and other measures and their implementation, see for instance Turkman *et al.* (2019), Ch. 9 [24].

On the other hand, for likelihood-based inference, we need to ensure the posterior distribution is well approximated by a Gaussian distribution and also that this distribution degenerates at a point. Indeed, only then we can be confident that the influence of the choice of prior distribution has vanished and the mean of the posterior samples is a good approximation of the desired maximum likelihood estimate. Specific to the data cloning algorithm, we need to check whether the posterior distribution

(i) has become nearly degenerate and

(ii) nearly Gaussian.

The number of clones required for these behaviors depends heavily on the likelihood function and prior distribution chosen. Fortunately, the diagnostic measures recommended by Lele *et al.* (2010) [13] are simple to compute and allow the selection of an adequate number of clones for the problem at hand.

If the assumptions of the data cloning algorithm are satisfied, then the Fisher information matrix is positive definite in a neighborhood around the maximum likelihood estimate. Furthermore, recall from Equation (2.3) that, in the k-times cloned posterior π_k, the estimate of the inverse Fisher information matrix from the posterior samples is scaled by the inverse of k. Therefore, as we increase the number of clones, the eigenvalues of the estimated covariance matrix from the samples should decrease at approximately a rate k^{-1}.

For a positive definite matrix, the Courant-Fischer Theorem ensures that the greatest eigenvalue provides an upper bound on the elements of the main diagonal (Horn and Johnson 2012 [10]). Hence, since the greatest eigenvalue should decrease at the rate k^{-1}, we have an upper bound on the rate at which the elements of the main diagonal of the estimated inverse Fisher information matrix must decrease. This enables us to measure the rate at which the posterior distribution is degenerating to a point mass probability measure on the maximum likelihood estimate, since the elements in the main diagonal of the said matrix represent an estimate of the posterior variances for the model parameters.

Let $\lambda_{max,k}$ denote the maximum eigenvalue of the k-times cloned posterior covariance matrix. Then, for large k and under regularity conditions,

$$\lambda^S_{max,k} = \frac{\lambda_{max,k}}{\lambda_{max,1}} \approx \frac{1}{k} \quad . \tag{2.6}$$

Lele *et al.* (2010) [13] call $\lambda^S_{max,k}$ the scaled maximum eigenvalue for k clones. The authors suggest monitoring this quantity to assess its rate of convergence to zero as we increase the number of clones. The closer this measure is to zero, the more mass the posterior distribution assigns to small neighborhoods of the (possibly) maximum likelihood estimate.

For the second item, the normality of the posterior distribution, Lele *et al.* (2010) [13] suggest computing two statistics from the posterior samples. As before, let p be the dimension of the parameter vector $\boldsymbol{\theta}$ and N denote the number of samples obtained from Algorithm 1 after discarding the ones from the burn-in period. Let E_i denote the $(i-0.5)/N$ quantile of a χ^2_p distribution, $i \in \{1, ..., N\}$. Define

$$O_i = (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^\top \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}}) \tag{2.7}$$

for $i \in \{1, ..., N\}$, with $\widehat{\boldsymbol{\Sigma}}$ the estimated posterior covariance matrix and let $O_{(i)}$ denote their ordered values. Since $O_{(i)}$ are simply estimates of E_i, the statistics

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (O_{(i)} - E_{(i)})^2 \tag{2.8}$$

and

$$r^2 = 1 - \hat{\rho}^2 \big(O_{(1)}, ..., O_{(N)}; E_1, ..., E_N\big) \quad , \tag{2.9}$$

in which $\hat{\rho}$ denotes the estimated Pearson correlation coefficient, approach zero as the number of clones k increases.

Solymos (2010) [23] provides an implementation of data cloning for *R* (R Core Team 2020 [19]). The package allows the use of common probabilistic sampling software amongst Bayesian practitioners, such as *JAGS* (Plummer 2017 [18]) and Stan (Carpenter *et al.* (2017) [5]) to perform sampling from the cloned posterior distribution and includes all diagnostic measures described above. The paper by Solymos (2010) [23] and also the original papers by Lele *et al.* (2007) [12] and Lele *et al.* (2010) [13] provide plenty of examples to get acquainted with data cloning.

3. IDENTIFIABILITY

Let $\mathcal{Y}$ denote a sample space, $\mathcal{A}$ a σ-algebra of subsets of $\mathcal{Y}$ and $\mathcal{M}(\mathcal{Y}, \mathcal{A})$ denote the set of probability measures on $(\mathcal{Y}, \mathcal{A})$. In statistical theory, the inferential procedure is enabled by equipping the measurable space $(\mathcal{Y}, \mathcal{A})$ with a family of probability measures $\mathcal{F} \subset \mathcal{M}(\mathcal{Y}, \mathcal{A})$. For practical purposes, this family is defined through a known map $\boldsymbol{\Phi} : \Theta \to \mathcal{M}(\mathcal{Y}, \mathcal{A})$, with Θ being the parameter space in the parametric scenario. Specifically, a statistical model is a triple $\mathcal{E} = (\mathcal{Y}, \mathcal{A}, \mathcal{F} = \{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\})$, in which $\mathcal{F}$ is a family of probability measures on $(\mathcal{Y}, \mathcal{A})$ indexed by the parameter space Θ.

Notice the definition of a statistical model imposes no restrictions on Θ, allowing for parametric, semiparametric and nonparametric structures (San Martín 2018 [22]). We assume in this paper that:

- **(i)** $\Theta \subset \mathbb{R}^p$ is an Euclidean space;
- **(ii)** the sample space $\mathcal{Y}$ is equipped with a topology and $\mathcal{A} = \mathcal{B}(\mathcal{Y})$ is the Borel σ-algebra obtained from the topology on $\mathcal{Y}$;
- **(iii)** the probability measures in $\mathcal{F}$ are absolutely continuous with respect to some measure on $(\mathcal{Y}, \mathcal{A})$.

The latter constraint allows for much simplification in the discussion since now we can represent $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ as a family of probability density functions. We are now ready to define identifiability of a parameter in the sampling theoretical framework.

Definition 3.1. Let $\mathcal{E}$ be a statistical model. A parameter $\boldsymbol{\theta} \in \Theta$ is said to be **identifiable** if for any $\boldsymbol{\theta}^* \in \Theta$, $f_{\boldsymbol{\theta}}(\mathbf{y}) = f_{\boldsymbol{\theta}^*}(\mathbf{y})$ occurs, for all $\mathbf{y} \in \mathcal{Y}$, if and only if $\boldsymbol{\theta}^* = \boldsymbol{\theta}$.

The concept above is referred to as global identifiability (Koopmans and Reiersol 1950 [11]). This is done to distinguish it from local identifiability, which we define below. However, in this paper, we refer to the former property as simply *identifiability.*

Definition 3.2. Let $\mathcal{E}$ be a statistical model. A parameter $\boldsymbol{\theta} \in \Theta$ is said to be **locally identifiable** if there exists $\epsilon > 0$ and a neighborhood $N_\epsilon(\boldsymbol{\theta}) \subset \Theta$ of $\boldsymbol{\theta}$ such that for any $\boldsymbol{\theta}^* \in N_\epsilon(\theta)$, $f_{\boldsymbol{\theta}}(\mathbf{y}) = f_{\boldsymbol{\theta}^*}(\mathbf{y})$ occurs, for all $\mathbf{y} \in \mathcal{Y}$, if and only if $\boldsymbol{\theta}^* = \boldsymbol{\theta}$.

We can see thus that local identifiability is weaker than (global) identifiability. In fact, a globally identifiable parameter is always locally identifiable. The converse, however, need not be true.

So far the definitions allow us to talk only about a single point in the parameter space. For the inferential procedure to be satisfactory, we would like to know whether all parameter values $\theta \in \Theta$ are identifiable. Fortunately, the definitions above are easily extended to the entire parameter space Θ.

Definition 3.3. A statistical model $\mathcal{E}$ is said to be identifiable (locally identifiable) if for all $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\theta}$ is identifiable (locally identifiable).

Parameters which yield the same likelihood function are said to be *observationally equivalent*, and it is possible to construct an equivalence relation using the concept of identifiability; see for example Picci (1977) [17] or Florens and Simoni (2011) [6] and references therein. This relation $\sim$ is such that, for any $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta$, $\boldsymbol{\theta} \sim \boldsymbol{\theta}^*$ if, and only if, $f_{\boldsymbol{\theta}}(\mathbf{y}) = f_{\boldsymbol{\theta}^*}(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{Y}$. Through the equivalence relation defined above we obtain the quotient space $\tilde{\Theta} = \Theta/\!\sim$. The elements of the quotient spaces are the equivalence classes induced by $\sim$ on Θ. Thus, there always exists a canonical statistical model $\mathcal{E}_{\tilde{\Theta}} = (\mathcal{Y}, \mathcal{A}, \mathcal{F} = \{f_{[\boldsymbol{\theta}]} : [\theta] \in \tilde{\Theta}\})$ which is set identifiable, *i.e.* the family $\mathcal{F}$ is indexed by the equivalence classes.

An easy-to-prove property of equivalence classes which makes them very convenient for studying identifiability is that they are disjoint. Thus, it is sufficient for statistical identifiability to define a function which maps each equivalence class to a single element in that class, since then the equivalence classes will be reduced to singletons. Florens and Simoni (2011) [6] call such functions *sections*. Let $[\boldsymbol{\theta}] \in \tilde{\Theta}$ denote the class of equivalence of $\boldsymbol{\theta} \in \Theta$, *i.e.* $[\boldsymbol{\theta}] = \{\boldsymbol{\theta}^* \in \Theta : \boldsymbol{\theta}^* \sim \theta\}$.

Definition 3.4. A **section** is a function $\sigma : \tilde{\Theta} \to \Theta$ such that for all $[\boldsymbol{\theta}] \in \tilde{\Theta}$, $\sigma([\boldsymbol{\theta}]) \in [\boldsymbol{\theta}]$.

As previously mentioned, the equivalence classes being disjoint leads us to an identifiable statistical model $\mathcal{E}_\sigma = (\mathcal{Y}, \mathcal{A}, \mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \sigma(\tilde{\Theta})\})$, in which σ is any one section. Moreover, the choice of section is really irrelevant, as Paulino and Pereira (1994) [15] point out, since for any two sections σ and σ^*, there exists a bijective function $h : \sigma(\tilde{\Theta}) \to \sigma^*(\tilde{\Theta})$.

3.1. Implications to Data Cloning

Another important consequence of identifiability which we exploit in this paper is that, if $\mathcal{E}$ happens to be unidentifiable, then the maximum likelihood estimate of $\boldsymbol{\theta}$ for any given sample (if it exists) is not unique. In fact, given any point estimate $\hat{\boldsymbol{\theta}}(y) \in \Theta$, there is $\hat{\boldsymbol{\theta}}^*(y) \in [\hat{\boldsymbol{\theta}}]$ such that $L(\hat{\boldsymbol{\theta}}; y) = L(\hat{\boldsymbol{\theta}}^*; y)$. It becomes clear now why consistency of the maximum likelihood procedure is no longer guaranteed. Under model unidentifiability thus there is a class of equivalence of undistinguishable candidates to the maximum likelihood estimate

$$[\hat{\boldsymbol{\theta}}] = \arg \sup_{[\boldsymbol{\theta}] \in \tilde{\Theta}} L([\boldsymbol{\theta}]; y) = \left\{ \boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} = \arg \sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; y) \right\} \quad . \tag{3.1}$$

Lele *et al.* (2010) [13] discuss the behavior of the k-times cloned posterior distribution under model unidentifiability. In this scenario, let $[\hat{\boldsymbol{\theta}}]$ be the equivalence class of the maximum likelihood estimate. The authors show that if $[\hat{\boldsymbol{\theta}}]$ is not a singleton, then

$$\pi_k(\boldsymbol{\theta}|\mathbf{y}) \xrightarrow{\mathcal{L}} \frac{\pi(\boldsymbol{\theta})}{\int_{[\hat{\boldsymbol{\theta}}]} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \quad , \quad \forall \boldsymbol{\theta} \in [\hat{\boldsymbol{\theta}}] \quad . \tag{3.2}$$

Therefore, it seems we can, in theory, use data cloning to investigate the identifiability of complex statistical models. If the posterior samples generated using data cloning, for increasing values of k, do not seem normally distributed and/or seem to degenerate at a set of values, then the model may be unidentifiable.

In reality, however, we must not hurry to conclusions. Identifiability of statistical models can only be assessed using analytical techniques and it is a mathematical question in general. It precedes statistical inference (Koopmans and Reiersol 1950 [11]). What we can study with data cloning diagnostics is model estimability under a particular data set. However, detecting estimability issues over many distinct data sets would lead us to question the very structure of the statistical model we are employing. Thus, a general guideline for a simulation study for assessing identifiability using data cloning is:

- **(i)** for various sample sizes, simulate several data sets from a postulated statistical model;
- **(ii)** for each data set, fit the model using data cloning with distinct prior distributions and with increasing values of the number of clones k;
- **(iii)** analyze the posterior samples to study the behavior of the algorithm.

Data cloning diagnostics should always reveal convergence issues when the algorithm is used to estimate the parameters of an unidentifiable model. However, under possibly very informative priors and weakly informative data, it may happen that even for large values of k the sampling algorithm will output samples from π_k which do not indicate problems. This is resonant with the arguments of Lindley (1971) [14] on the irrelevance of the question of identifiability within the subjective Bayesian paradigm.

Furthermore, even if the statistical model is identifiable, there are plenty of other ways in which the data cloning algorithm may fail to converge. It is no coincidence that advances in Bayesian computational methods have been accompanied by developments of techniques for diagnosing convergence issues. Since data cloning uses MCMC algorithms for likelihood-based inference, the same convergence issues of sampling algorithms that Bayesian analyses face must also be considered.

3.2. Unidentifiability of the Gaussian DLM

The Gaussian dynamic linear model is particularly convenient for our purposes since it illustrates many aspects of identifiability in a simple manner.

Definition 3.5. *The dynamic model with state and observation equations*

$$\begin{cases} Y_t = FX_t + \nu_t & , \quad \nu_t \overset{iid}{\sim} \mathcal{N}(0,V) \\ X_t = GX_{t-1} + \omega_t & , \quad \omega_t \overset{iid}{\sim} \mathcal{N}(0,W) \\ X_0 \sim \mathcal{N}(m_0, C_0) \end{cases} \tag{3.3}$$

is called a ***univariate Gaussian dynamic linear model*** *with parameter vector* $\boldsymbol{\theta} = (F,G,V,W) \in \mathbb{R}^2 \times \mathbb{R}^2_+$ *and initial information set* $D_0 = \{m_0, C_0\}$.

The statistical model of Definition 3.5 is not identifiable as it is. This is a well known result in the literature of dynamic models and some identifiability constraints for the multivariate scenario can be found in Harvey (1989) [9]. The usual path to a proof of the unidentifiability of the dynamic linear model employs a change of variables in its defining observation and process equations. Under Gaussian errors it is easy to see that for any real number $s \neq 0$

$$\begin{cases} Y_t = FX_t + \nu_t \\ X_t = GX_{t-1} + \omega_t \end{cases} \iff \begin{cases} Y_t = (Fs^{-1})(sX_t) + \nu_t \\ sX_t = G(sX_{t-1}) + s\omega_t \end{cases} \iff \begin{cases} Y_t = F^*X_t^* + \nu_t \\ X_t^* = GX_{t-1}^* + \omega_t^* \end{cases}, \tag{3.4}$$

in which $F^* = Fs^{-1}$, $\omega_t^* = s\omega_t$ and $X_t^* = sX_t$, for all $t \in \mathcal{T}$. Notice now the process equation random error is distributed as $\omega_t^* \overset{iid}{\sim} \mathcal{N}(0, W^*)$, with $W^* = s^2 W$.

Therefore, this change of variables implies that, if $\boldsymbol{\theta} = (F,G,V,W)$ is the original parameter vector and $\boldsymbol{\theta}^* = (F^*,G,V,W^*)$ is the parameter vector that results from the transformation proposed, then it follows that for all $(y_1,...,y_T) \in \mathcal{Y}$ we have $f_{\boldsymbol{\theta}}(y_1,...,y_T) = f_{\boldsymbol{\theta}^*}(y_1,...,y_T)$ for any $s \neq 0$.

The structural equations define the density functions uniquely and the arguments above are sufficient to prove this model is unidentifiable. The general suggestions for enforcing model identifiability in this context are to

(i) fix the parameter F to a known non-zero constant or

(ii) fix the process variance W to a known positive constant and constrain F to be either strictly positive or strictly negative (Harvey 1989 [9]).

Thus, in the identifiable statistical model, one must choose between estimating F or estimating W.

There are, however, an infinite number of other restrictions, or sections, on the parameter space which lead to the same statistical model up to a bijective function. Firstly, note that for the Gaussian dynamic linear model, the equivalence classes are readily built from the proof in (3.4). As a matter of fact, we know that $\boldsymbol{\theta} \sim \boldsymbol{\theta}^*$ if there exists $s \in \mathbb{R}$ such that

$$\boldsymbol{\theta} = (F, G, V, W) \sim (F/s, G, V, s^2 W) = (F^*, G, V, W^*) = \boldsymbol{\theta}^* \quad .$$

This, in turn, implies $\boldsymbol{\theta} \sim \theta^*$ whenever $F^2 W = (F^*)^2 W^*$. If we let $(a, b, c) \in \mathbb{R} \times \mathbb{R}^2_+$, then we can write the quotient space as

$$\tilde{\Theta} = \bigcup_{(a,b,c) \in \mathbb{R} \times \mathbb{R}^2_+} \left\{ \{(F, G, V, W) \in \Theta : G = a, V = b \text{ and } F^2 W = c\} \right\} \quad . \tag{3.5}$$

We recall once again that the equivalence classes are disjoint. Therefore, once we build them, it is sufficient for model identifiability that we find a section $\sigma : \tilde{\Theta} \to \Theta$ such that the equivalence classes of $\sigma(\tilde{\Theta}) \subset \Theta$ are singletons (Paulino and Pereira 1994 [15]). For clarity of exposition, let $[(a, b, c)] = \{(F, G, V, W) \in \Theta : G = a, V = b, F^2 W = c\} \in \tilde{\Theta}$ denote the equivalence classes on Θ for all $(a, b, c) \in \mathbb{R} \times \mathbb{R}^2_+$. The general identifiability constraints can now be stated as

Proposition 3.1. *Let $\mathcal{E}$ be the Gaussian dynamic linear model as in Definition 3.5. A sufficient condition for the function $\sigma : \tilde{\Theta} \to \Theta$ to be a section on $\tilde{\Theta}$ is that for all $(a, b, c) \in \mathbb{R} \times \mathbb{R}^2_+$, the set function $\sigma : [(a, b, c)] \mapsto (u_1(a, b, c), G, V, u_2(a, b, c))$, with $u_1^2 u_2 : (a, b, c) \mapsto c$.*

Proof: We need to show that σ is injective and $\sigma([(a, b, c)] \in [(a, b, c)]$ for all such equivalence classes. The latter follows immediatly from the fact that if $\theta = (F, G, V, W)$ is such that $G = a$, $V = b$ and $F^2 W = c$, then $\boldsymbol{\theta} \in [(a, b, c)]$. Moreover, since equivalence classes are disjoint this implies σ is injective. Therefore, taking $F = u_1(a, b, c)$ and $W = u_2(a, b, c)$, the proof is complete. □

We can now write the commonly suggested restrictions for the Gaussian DLM as sections on the parameter space. Fixing $F = s$, for some real constant $s \neq 0$, is equivalent to conducting inference over the section $\sigma_F : [(a, b, c)] \mapsto (s, G, V, c/s^2)$. Also, fixing $W = s$, for some $s \in \mathbb{R}_+$ is equivalent to adopting the section $\sigma_W : [(a, b, c)] \mapsto (\sqrt{c/s}, G, V, s)$.

Nevertheless, there is nothing wrong with using an unidentifiable statistical model as long as inference (or prediction) is conducted on identifiable quantities. An example, suggested to us by one of the reviewers, is that of a linear model with rank defficient design matrix:

even though some, or all, of the regression parameters are unidentifiable, the mean response can always be estimated uniquely from the data. We emphasize, however, that we would refer to the previous problem as an unidentifiability problem only when the design matrix is always rank defficient no matter what data we observe, such as in high-dimensional scenarios. In case some exploratory variables collected are perfectly (or highly) correlated only for a particular data set, we would refer to it as a problem of model estimability.

4. SIMULATION STUDY

In this section we present and discuss the results of several simulation studies in which the data cloning algorithm is employed to assess identifiability of the Gaussian dynamic linear model. Ideally, data cloning should not present any convergence issues when used for maximum likelihood estimation in an identifiable model. On the other hand, we would expect to see clear failures in all of the convergence measures available whenever data cloning is employed in an unidentifiable model. Before proceeding to the results there are some important points that need to be discussed so that the motivation behind the simulation study is clear.

Firstly, our main objective is to show how to use data cloning as a tool to assess identifiability statistical models. Our choice to illustrate the procedure through the dynamic linear model is justified by the fact that it is a latent variable model for which the identifying constraints are known. Hence, we can perfectly discern convergence issues due to model unidentifiability from those due to poor performance of the sampling algorithms.

Secondly, Lele *et al.* (2010) [13] advise using distinct prior distributions when performing data cloning. A simulation study for identifiability analysis should also take this into account as we need to be sure that there exists a number of clones for which the influence of any prior distribution vanishes. The more prior distributions we test the better is the study. As proposed by Lele *et al.* (2010) [13], we adopt three prior setups: uninformative, informative and disinformative. The first is simply as vague as possible, the second puts most of its probability mass around the true parameter value and the third is also an informative prior distribution, but most of its probability mass is allocated somewhat far from the true parameter values.

Lastly, we recommend employing both varying sample sizes and parameters. The former allows a view of the convergence of the maximum likelihood estimator, while the latter allows us to explore regions of the parameter space that may be of practical interest.

Data cloning is computationally demanding, although for a single data set setting the number of clones to a high value may not be a problem. For our purposes, we will be fitting the same model under multiple distinct settings and it is just not feasible to use a high number of clones. It does not matter, however, because we are not interested in finding the maximum likelihood estimate, but in gathering evidence of whether or not it can be found uniquely. Multiple starting values plus strong diagnostic measures of convergence allow us to gather solid evidence of model identifiability and issues thereof.

4.1. Simulation Parameters

We have chosen to simulate time series of sizes 25, 50, and 100. The simulated states are in Figure 1. To test distinct parameter vectors we vary the amount of noise added to the sample by the measurement process. This is done by considering the ratio between the variance of the process and measurement errors to be $W/V = 0.5$, 1, 2, and 10.

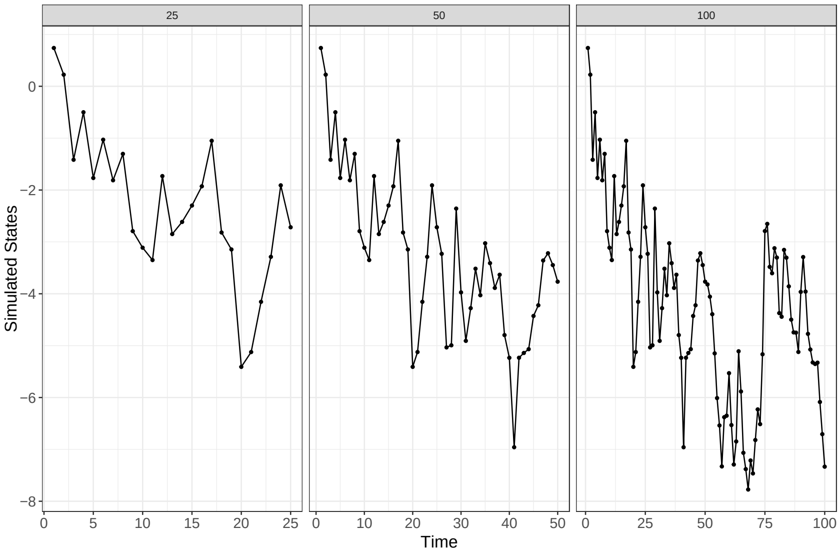

Figure 1: Simulated states for each adopted sample size.
The true parameter values used are displayed in Table 1.

We recall the parameter vector for the dynamic linear model under normality assumptions is $\theta = (F, G, V, W) \in \mathbb{R}^2 \times \mathbb{R}^2_+ = \Theta$. The four true parameter setups we use for simulating the time series are displayed in Table 1. Some illustrations of the effect of increasing the measurement error are available in Figure 2. We can see in the plots that increasing the measurement noise makes it harder to visually detect any patterns from the hidden states. We would expect that noisy and/or small datasets would be very challenging for data cloning since the likelihood function might not be well-behaved around the (possibly non-unique) maximum likelihood estimate. Nevertheless, similar situations would be challenging for most alternative estimation methods as well.

Table 1: True parameter values for simulation of the Gaussian dynamic linear model.

Setup	F	G	V	W
$W/V = 1/2$	1	1	2	1
$W/V = 1$	1	1	1	1
$W/V = 2$	1	1	0.5	1
$W/V = 10$	1	1	0.1	1

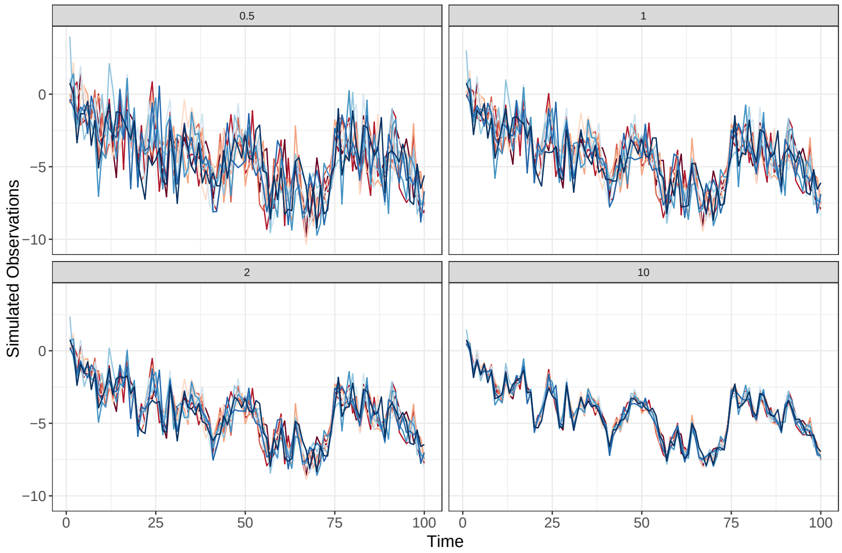

Figure 2: Plot of 10 simulated time series of length 100 arising from the hidden states in Figure 1. Each panel represents a signal-to-noise ratio W/V as presented in Table 1.

For the standard deviations $\sqrt{W}$ and $\sqrt{V}$, we adopted half-Cauchy prior distributions with scale equal to 10 as uninformative priors. This prior distribution is recommended by Gelman (2006) [7] for hierarchical models as an ideal alternative to the widely used gamma prior with small hyperparameters. Since our true parameter values are quite small compared to the tails of these prior distributions, we expect their added information to be insignificant compared to the data.

The F parameter receives a $\mathcal{N}(0, 10^4)$ in the uninformative setup. If we were to be faithful to the identifiability constraints required for this model, we would have to employ prior distributions which assign zero mass to negative values for F. However, the model is unidentifiable whether or not we restrict this parameter to the positive real line. Nonetheless, when performing some pre-tests for the simulation study, sampling F from priors on $\mathbb{R}_+$ resulted in running times up to three times longer than when using priors on $\mathbb{R}$.

The parameter G regulates the autoregressive behavior of the hidden states. The data we simulate assumes that these latent variables behave as a Gaussian random walk. We know that for values of G outside the open interval $(-1, 1)$ the latent process is non-stationary (Harvey 1989 [9]). Lele *et al.* (2007) [12] use a uniform prior distribution on the interval $(-1, 1)$ for this parameter, enforcing stationarity of the latent stochastic process. In our simulations, the data is clearly non-stationary. Therefore, we consider a $\mathcal{N}(0, 10^4)$ as the uninformative prior setup for G. This prior allows the process to present highly explosive growth behaviors if the data behaves as such. It is highly unlikely that this prior distribution would be used in a purely Bayesian framework, but data cloning allows us to use such largely uninformative prior distributions with ease.

In Table 2 we present the uninformative prior setup just discussed together with the informative and disinformative ones. The choice of the latter two, as previously discussed, simply aims to assign more probability mass closer or further (respectively) from the true parameter values. Notice that since the parameterization of the Gaussian distribution in

JAGS (Plummer 2007 [18]) is given by the mean and precision (*i.e.* the inverse of the variance), the prior distributions in Table 2 are provided in terms of the precision parameters $1/W$ and $1/V$ instead of W and V, respectively.

Table 2: Prior distributions chosen to represent the uninformative, informative and disinformative prior setups. The notation $\mathcal{HC}^{-2}$ indicates the distribution of the inverse of the square of a Half-Cauchy random variable, while λ denotes the signal-to-noise ratio W/V.

Prior Setup	F	G	$1/V$	$1/W$
Uninformative	$\mathcal{N}(0, 10^4)$	$\mathcal{N}(0, 10^4)$	$\mathcal{HC}^{-2}(0, 10)$	$\mathcal{HC}^{-2}(0, 10)$
Informative	$\mathcal{N}(1, 1)$	$\mathcal{N}(1, 1)$	$\Gamma(4^{-1}, (4\lambda)^{-1})$	$\Gamma(4^{-1}, 4^{-1})$
Disinformative	$\mathcal{N}(10, 5)$	$\mathcal{N}(-1, 1)$	$\Gamma(1, 5^{-1})$	$\Gamma(1, 5^{-1})$

4.2. Data Cloning Diagnostics

We begin our study of identifiability through the scaled maximum eigenvalue, $\lambda^S_{max,k}$, of the posterior covariance matrix, which should decay at about the theoretical rate of $1/k$, in which k denotes the number of clones used. In Figure 3 we display this measure for the case of the unidentifiable dynamic linear model. Since some of the resulting eigenvalues presented very high values, the graph with all of the observed measures is uninteresting due to the scaling of the ordinate axis.

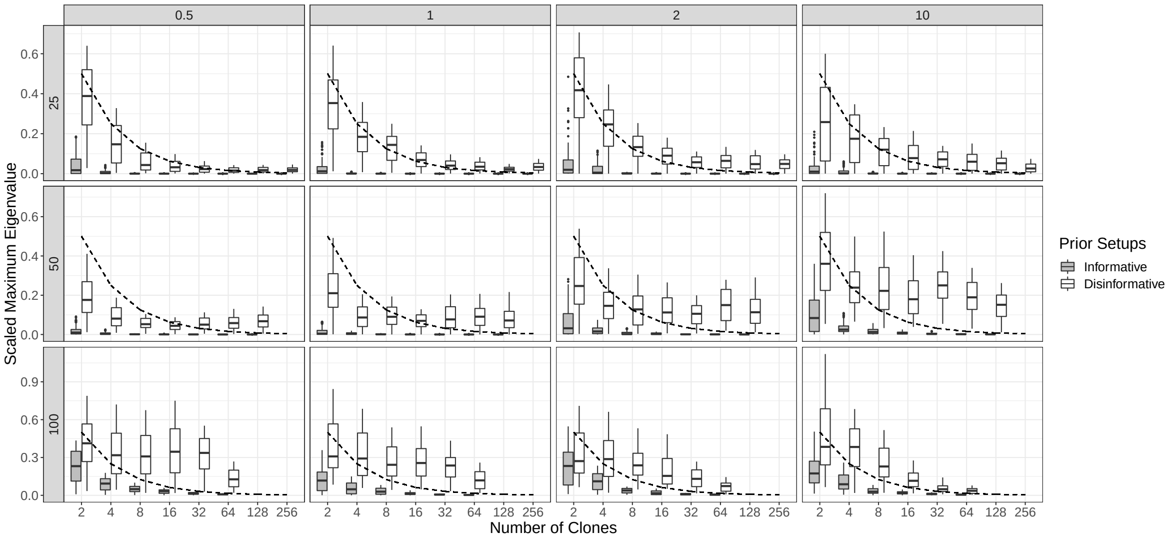

Figure 3: Box-plots of the scaled maximum eigenvalues of the posterior covariance matrix for the unidentifiable dynamic linear model. The dashed line represents the theoretical rate of convergence.

We have chosen thus to display in Figure 3 only the 50% smallest scaled eigenvalues obtained from the simulations using the informative and disinformative setup. The posteriors arising from uninformative prior distributions presented extremely high scaled eigenvalues even at the maximum number of clones for each sample size, which is a very strong evidence of identifiability issues. Nevertheless, the quantities in Figure 3 should still follow the theoretical convergence rate as long as the model is identifiable (which we know it is not).

It is easy to see in Figure 3 that $\lambda^S_{max,k}$ does not decay at the theoretical rate for the disinformative prior distributions. On the other hand, the informative prior setup seems to yield reasonable values for this diagnostic measure as the number of clones increases. However, we can see some odd behavior, particularly in the lower sample sizes with a low signal-to-noise ratio. Note that many $\lambda^S_{max,k}$ are already much below the theoretical rate of convergence in the first steps of the data cloning. This may be a sign that the variance of the unidentifiable parameters is being held low by the variance of the prior distribution.

Summing up, the scaled maximum eigenvalues observed from fitting an unidentifiable model resulted in undesired convergence properties and, in the case of the uninformative prior distributions, unreasonably high values for the eigenvalues of the posterior covariance matrix. This is in accordance with what we would expect from an unidentifiable model and indicates that data cloning is pointing towards identifiability issues when they are indeed present.

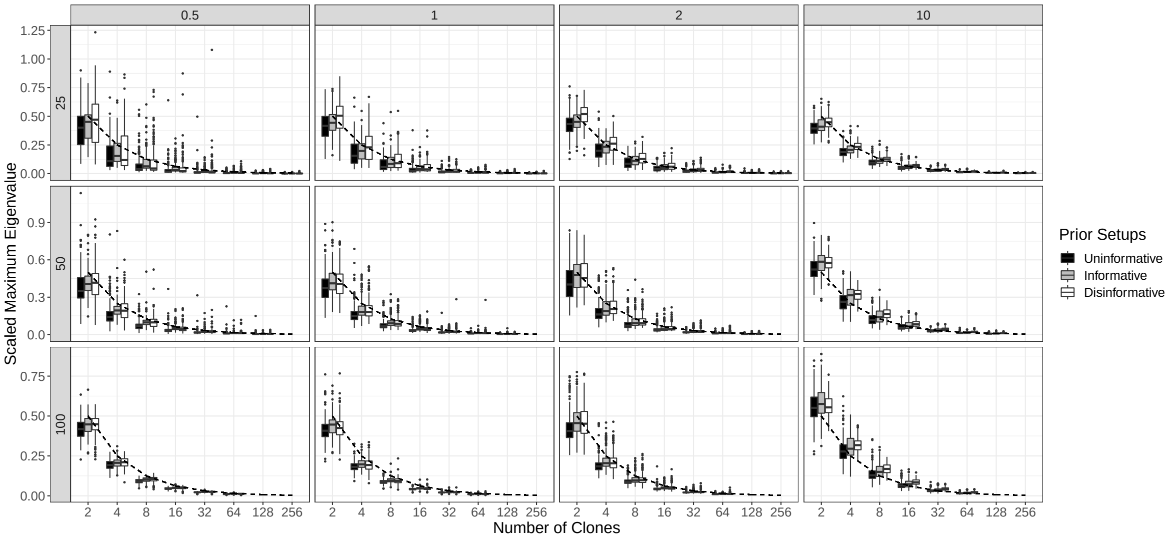

Figure 4: Box-plots of the scaled maximum eigenvalues of the posterior covariance matrix for the identifiable dynamic linear model in which we have set F to its true value. The dashed line represents the theoretical rate of convergence.

In Figure 4 we present the $\lambda^S_{max,k}$ drawn from an identifiable model in which we have fixed the parameter F to its true value. As the number of clones increases the theoretical rate of convergence is followed tightly by the quantities resulting from all three prior distributions. This is a clear indicator that the posterior distributions are becoming increasingly degenerated at the expected rate. Another important consideration is that all three prior setups differ only within small clone numbers, which indicates the influence of the prior distribution is indeed vanishing. Once again, therefore, the main diagnostic for data cloning has provided satisfactory results for the dynamic linear model. It exhibited no convergence issues when there are indeed no identifiability problems. This indicates constraining the parameter space by fixing the parameter F has led to a statistical model for which all parameters can be reliably estimated.

Table 3 presents the quartiles for $\lambda_{max,k}$ taken at the best-case scenario: the sample size is 100 and the number of clones k is 64. We can see the quartiles in the identifiable model are very close to each other, while the ones for the unidentifiable model are not. This is yet another strong evidence that there is still a considerable influence of the prior distributions on the joint posterior distribution of the parameters. There is an issue, however, because the quartiles for the maximum eigenvalues are quite small under the informative and disinformative setups even under unidentifiability. This indicates the posterior variance is decreasing as we increase the number of clones, which obviously should not happen since the model is unidentifiable. Nonetheless, for the uninformative setup, the maximum eigenvalues are of an order 10^4 higher than those from the informative setup.

Table 3: Quartiles for the maximum eigenvalues of the posterior covariance matrix of the parameters of the dynamic linear model. The values are for both the unidentifiable and identifiable models at the sample size of 100 and with number of clones equal to 64.

W/V	Prior	Identifiable			Unidentifiable		
		P_{25}	P_{50}	P_{75}	P_{25}	P_{50}	P_{75}
0.5	Uninformative	0.0035	0.0045	0.0055	32.1684	65.6898	140.4389
	Informative	0.0035	0.0044	0.0054	0.0048	0.0079	0.0115
	Disinformative	0.0034	0.0044	0.0054	0.0795	0.1493	0.3428
1	Uninformative	0.0019	0.0024	0.0032	40.4107	76.1869	132.6472
	Informative	0.0019	0.0025	0.0032	0.0024	0.0037	0.0066
	Disinformative	0.0019	0.0024	0.0032	0.1070	0.1670	0.2772
2	Uninformative	0.0015	0.0019	0.0024	38.6173	61.5385	113.3972
	Informative	0.0015	0.0019	0.0024	0.0017	0.0029	0.0048
	Disinformative	0.0015	0.0018	0.0023	0.0496	0.1045	0.2125
10	Uninformative	0.0012	0.0014	0.0016	29.9088	49.6416	72.4899
	Informative	0.0012	0.0014	0.0016	0.0014	0.0019	0.0029
	Disinformative	0.0012	0.0013	0.0016	0.0221	0.0454	0.0802

If we compare the quartiles over the three prior setups, it becomes clear there is still strong influence of the prior distribution even at 100 clones of the original dataset when the model is unidentifiable. However, under a single prior setup the conclusions related to the variance of the posterior distribution would differ considerably. In the informative prior setting, in particular, the quartiles of the maximum eigenvalues indicate no identifiability problems at all. The quartiles in this case are all small and reasonably close to each other, indicating the variance of the posterior distribution is small since the maximum eigenvalue provides an upper bound on the variances of the parameters. Therefore, for our purposes it would seem that the observed decay rate for the scaled maximum eigenvalue is the most appropriate of the two measures of degeneracy. By measuring the decay of $\lambda^S_{max,k}$, we were able to detect possible identifiability problems across all prior settings.

We can also check whether the posterior is reasonably close to a Gaussian distribution. This is done, as suggested by Lele *et al.* (2010) [13], through the measures presented in (2.8) and (2.9). In Table 4 we present the average of these diagnostics for each of the scenarios explored in the simulations. Overall, these diagnostic measures are greater, on average, in the unidentifiable than in the identifiable model. Furthermore, their averages seem to decrease in magnitude as the sample size increases, which is also to be expected.

Table 4: Diagnostic measures for normality of the samples from the posterior distribution of the parameters of the dynamic linear model.

Size	Prior Setup	Constraint	$W/V = 0.5$		$W/V = 1$		$W/V = 2$		$W/V = 10$	
			MSE	$\tilde{r}^2$	MSE	$\tilde{r}^2$	MSE	$\tilde{r}^2$	MSE	$\tilde{r}^2$
25	Uninformative	F = 1	3.173	0.088	1.718	0.068	1.590	0.065	1.891	0.078
		None	15.327	0.178	10.277	0.155	8.449	0.137	4.551	0.105
	Informative	F = 1	2.335	0.078	1.416	0.062	1.030	0.052	1.909	0.074
		None	13.662	0.175	10.891	0.159	7.148	0.130	3.779	0.091
	Disinformative	F = 1	2.430	0.083	1.398	0.062	0.989	0.052	1.194	0.054
		None	11.578	0.164	8.214	0.148	7.640	0.135	5.015	0.102
50	Uninformative	F = 1	3.699	0.092	1.600	0.057	0.624	0.039	0.393	0.030
		None	7.237	0.130	4.569	0.110	3.036	0.091	1.933	0.068
	Informative	F = 1	1.697	0.068	0.965	0.050	0.557	0.037	0.468	0.033
		None	5.467	0.112	3.947	0.101	2.704	0.086	1.301	0.055
	Disinformative	F = 1	2.556	0.083	1.229	0.055	0.678	0.040	0.409	0.031
		None	4.660	0.105	3.876	0.102	2.596	0.083	1.433	0.060
100	Uninformative	F = 1	0.591	0.037	0.364	0.028	0.253	0.021	0.224	0.020
		None	1.541	0.057	1.046	0.050	0.832	0.041	1.051	0.026
	Informative	F = 1	0.529	0.035	0.344	0.026	0.218	0.020	0.241	0.022
		None	1.425	0.059	0.915	0.047	0.578	0.034	0.345	0.024
	Disinformative	F = 1	0.623	0.038	0.394	0.029	0.255	0.022	0.177	0.017
		None	1.329	0.057	0.905	0.046	0.511	0.032	0.450	0.029

However, we would be hard-pressed to say these quantities have provided evidence of model unidentifiability (or identifiability). The values obtained under both scenarios, especially for the $\tilde{r}^2$, are satisfactory and also not very far apart from each other. For the MSE, in particular, it is to be expected that a model with one extra parameter, which is the case for the unidentifiable model, would require larger sample sizes or number of clones to achieve the same precision as a model with a lower number of parameters.

Furthermore, the quadratic form in Equation (2.7) may follow a chi-squared distribution even if the underlying probability distribution is not Gaussian. Azzalini and Valle (1996) [1], for example, show that this result holds for the quadratic form of p-variate Skew-Gaussian random variables. Therefore, these measures alone do not suffice to assess identifiability issues when using data cloning because it is possible these present reasonable values even when the posterior distribution is not Gaussian.

Therefore, although both the MSE and $\tilde{r}^2$ certainly serve their purpose when the interest is in obtaining the maximum likelihood estimates, for identifiability purposes they have not presented themselves as useful indicators of identifiability issues for this simple model and we do not advocate them to be heavily relied upon.

4.3. MCMC Diagnostics

We now move to the assessment of the quality of the posterior samples for maximum likelihood estimation. Firstly, we would like to know whether the Markov chains resulting from each of the three distinct prior setups, at the largest number of clones adopted, are targeting the same posterior distribution. For this task, we simply pretend we have run three independent chains under the same initial conditions, when in fact we used three distinct prior distributions. Within the Bayesian paradigm, individuals carrying distinct prior information about the same quantities need not arrive at the same inferential conclusions for finite samples, although asymptotic theory ensures this happens under some regularity conditions (Walker 1969 [26]). The differences in the posterior distributions for such individuals are even more pronounced whenever complex models and small and/or weakly informative data is at hand.

We emphasize yet again, however, that data cloning is a maximum likelihood estimation algorithm. Being so, it can not be affected by prior opinions. By collecting the chains from the three distinct prior setups, diagnostics such as the $\hat{R}$ can be used to check if the samples are being drawn from the same posterior distribution.

We provide in Table 5 the proportion of the simulations in which the $\hat{R}$ comparing three chains, one from each prior setup, is below the usual thresholds of 1.05 and 1.10. It is immediately clear that none of the simulations for the unidentifiable model have yielded joint posterior distributions for the parameter vector which are acceptably close enough from each other when starting from different prior distributions. For the identifiable model, the proportion starts low for the lower sample size of 25 and a low signal-to-noise ratio of 0.5 and reaches 1 for the sample size of 100. This is as to be expected, if not for the fact that many simulations do not yield close enough joint posterior distributions for some of the scenarios explored in this identifiable statistical model.

Table 5: Proportion of the simulations for which there is evidence that, starting from distinct prior distributions, the Markov chains are targeting the same posterior distribution. The values for the Gelman-Rubin diagnostic are computed at the highest number of clones for each sample size of the dynamic linear model.

Size	W/V	Identifiable		Unidentifiable	
		$\hat{R} < 1.05$	$\hat{R} < 1.10$	$\hat{R} < 1.05$	$\hat{R} < 1.10$
25	0.5	0.74	0.76	0	0
	1.0	0.90	0.91	0	0
	2.0	0.84	0.86	0	0
	10.0	0.44	0.51	0	0
50	0.5	0.51	0.55	0	0
	1.0	0.86	0.90	0	0
	2.0	0.98	0.99	0	0
	10.0	0.97	0.98	0	0
100	0.5	1.00	1.00	0	0
	1.0	1.00	1.00	0	0
	2.0	1.00	1.00	0	0
	10.0	0.94	0.98	0	0

It is possible that for some of the length 25 time series, the likelihood is very flat in a region around the maximum likelihood estimate. In this scenario, we would need a much larger number of clones to see that the Markov chains target the same joint posterior distribution. Nevertheless, the results from the Gelman-Rubin diagnostic point towards the clear failure of the data cloning when the model is unidentifiable, and present extremely promising results for the identifiable one. Were we unaware of the model's identifiability issues, these results, albeit not proof of unidentifiability, would surely lead us to reconsider the model structure.

There is one last point we need to verify when checking for model identifiability: the posterior means. The Gelman-Rubin diagnostic does not necessarily indicate that the posterior means, which are maximum likelihood estimates for a sufficiently large number of clones, are different. It can happen that the posterior means are very close to each other, while the posterior variance is not. Recall the Gelman-Rubin diagnostic indicates whether independent Markov chains happen to target the same posterior distribution. In other words, we could start from two or more distinct prior distributions and arrive at posterior distributions with the same mean but different variances. These are, thus, different posterior distributions and the Gelman-Rubin diagnostic will point towards convergence issues.

From Table 5 and Figure 3 we already know there is evidence of model unidentifiability. However, if the posterior means from distinct prior distributions were the same, we would have some evidence that we are able to reliably estimate the model parameters. Moreover, it might be the case that the unidentifiability issues found so far arise from poor tuning of the sampling algorithm.

In Table 6 we provide the average of the posterior means for the unidentifiable and identifiable Gaussian dynamic linear model with true signal-to-noise ratio $W/V = 1$ and sample size of 100. We also display the average effective sample size as a measure of the quality of the estimation of the posterior mean. Since we have drawn 1000 samples from each posterior distribution, we would want the effective sample size to be as close as possible to the total number of samples drawn. However, due to the very nature of MCMC algorithms it is expected that N_{eff} will be lower than the number of samples even if the model is identifiable. When using this measure, we are looking for parameters for which the N_{eff} is noticeably lower than both the total number of samples and the N_{eff} for other parameters.

If we focus on the parameters G and V, we can see that the averages of their posterior means are not considerably far apart from each other. However, we see that for the troublesome parameters W and F the posterior means under unidentifiability are heavily influenced by the choice of the prior distribution. Furthermore, the average effective sample size ranges from 1% to 5% of the total number of samples, indicating the chains for both parameters are highly autocorrelated. These results point towards extremely poor mixing of the Markov chains and, together with the diagnostics previously discussed, indicate clear failure of the model fitting procedure for the unidentifiable dynamic linear model.

However, the results for the identifiable functional F^2W, although not as good as that for the identifiable model, are still close to each other and also to the true value. This illustrates our previous comment that there is no harm in using unidentifiable statistical models, as long as the inferences are based on identifiable quantities. Hence, if we were interested in maximum likelihood estimation of any identifiable functional of $\boldsymbol{\theta}$, data cloning would yield good numerical approximations even if we had chosen to use the unidentifiable dynamic linear model.

Table 6: Averages of posterior means and effective sample sizes for the unidentifiable and identifiable Gaussian DLM with true signal-to-noise ratio $W/V = 1$ and sample size of 100. The true value for all parameters is 1 (see Table 1).

Parameter	Prior Setup	Identifiable		Unidentifiable	
		Mean	N_{eff}	Mean	N_{eff}
F	Uninformative	—	—	0.09	3.81
	Informative	—	—	1.89	3.69
	Disinformative	—	—	8.34	2.69
G	Uninformative	1.01	507.96	1.01	500.96
	Informative	1.01	504.54	1.01	503.79
	Disinformative	1.01	494.54	1.01	507.68
V	Uninformative	1.23	369.44	1.23	370.97
	Informative	1.23	284.53	1.23	282.88
	Disinformative	1.23	291.20	1.23	274.98
W	Uninformative	0.76	302.04	110.5	16.81
	Informative	0.76	208.40	0.22	16.87
	Disinformative	0.76	213.76	0.01	8.37
F^2W	Uninformative	0.76	302.04	0.76	311.37
	Informative	0.76	208.40	0.77	210.41
	Disinformative	0.76	213.76	0.80	218.34

Furthermore, as expected, the behavior within the identifiable model, in which we set $F = 1$, is exactly what we would want to see if we were using data cloning for maximum likelihood estimation. The posterior means, which we would like to call maximum likelihood estimates, seem to be independent of the choice of the prior distribution at the largest number of clones. The effective sample sizes are all satisfactory and indicate that the chains may be adequately exploring the posterior distribution. Given that the $\hat{R}$ diagnostics in Table 5 revealed the posterior distribution seems to be independent of the choice of prior distribution in the identifiable model at the highest number of clones adopted, we could gather the samples from all three chains to increase the effective sample size even further. Doing so would reduce the Monte Carlo variance of the numerical approximation to the maximum likelihood estimate and, consequently, improve the estimation of the Fisher information matrix.

5. FINAL COMMENTS

In this paper, we have explored the capabilities of data cloning as a tool for identifiability analysis of statistical models through a simulation study with the Gaussian dynamic linear model. Through an example, we have shown how such a simulation study can be planned and performed to gather evidence of possible model unidentifiability and how to interpret the most relevant diagnostic measures for the data cloning algorithm.

We found the bounds on the posterior covariance matrix of the parameters, its maximum eigenvalue $\lambda_{max,k}$, to be a good indicator of model identifiability. Its scaled version, $\lambda^S_{max,k}$, also yielded strong results since it exhibited convergence problems when they existed, while also indicating proper convergence of the algorithm in the identifiable model. The measures of normality did not present results as interesting as did $\lambda^S_{max,k}$. Both $\tilde{r}^2$ and MSE, suggested by Lele *et al.* (2010) [13], did not show significantly distinct values under either the identifiable or unidentifiable model. If we also consider the fact that these diagnostics can be satisfactory for quadratic forms of distributions other than the Gaussian, then our conclusion is that they are unreliable for identifiability analysis. However, for the purpose of maximum likelihood estimation using data cloning they must not be overlooked.

By exploiting distinct prior distributions, we were able to find clear parameter identifiability issues through the Gelman-Rubin diagnostic $\hat{R}$. Together with the data cloning diagnostics and the posterior means of the parameters, the evidence gathered through the diagnostics led us to the correct conclusion that the unconstrained Gaussian dynamic linear model is unidentifiable. Nonetheless, it also allowed us to conclude that fixing the parameter F to a known constant was enough to ensure statistical identifiability.

Overall, we find the results from the simulation study are very promising and indicate data cloning can (and should) be used as a tool for identifiability analysis, although some care must be taken as to how to do it properly. We emphasize here, once more, the importance of employing distinct prior distributions, parameter values and sample sizes in the simulation study to ensure that the evidence of identifiability, or lack of it, are consistent across an as wide as possible range of real possibilities.

There are also models for which MCMC algorithms either perform poorly or are simply too computationally demanding, for example those involving stochastic partial differential equations. As mentioned by one of the reviewers, the integrated nested Laplace approximation (Rue *et al.* 2009 [20]), or INLA for short, employs deterministic approximations to posterior distributions and has been paired up with data cloning for maximum likelihood estimation (see Baghishani *et al.* 2012 [2]). Although not as widely applicable as MCMC algorithms, INLA has been shown to be both extremely fast and precise when compared to the former. Furthermore, we are unaware of any studies on the usage of INLA and data cloning specifically for identifiability analysis and this may be an interesting venture within this topic.

Finally, we must also emphasize that identifiability cannot in general be proved based on simulation studies. After all, identifiability is a structural property of statistical models and it is impossible to exhaust the possible combinations of parameters and infinite sample sizes in a simulation study. Therefore, we are restricted to finite samples and a few points of interest in the parameter space. This implies that, at best, we can gather evidence of local identifiability in a region of practical interest of the postulated parameter space. The enterprise is nevertheless worth the effort since any evidence even of local unidentifiability in a statistical model can indicate undesired behavior of inferential procedures.

ACKNOWLEDGMENTS

We would like to thank the Brazilian research funding agencies CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for financial support. We also thank the valuable comments by the two reviewers and the Editor, which substantially contributed to the improvement of the paper.

REFERENCES

[1] Azzalini, A. and Valle, A.D. (1996). The multivariate skew-normal distribution, *Biometrika*, **83**(4), 715–726.

[2] Baghishani, H.; Rue, H. and Mohammadzadeh, M. (2012). On a hybrid data cloning method and its application in generalized linear mixed models, *Statistics and Computing*, **22**(2), 597–613.

[3] Bellu, G.; Saccomani, M.P.; Audoly, S. and D'Angiò, L. (2007). DAISY: A new software tool to test global identifiability of biological physiological systems, *Computer methods and Programs in Biomedicine*, **88**(1), 52–61.

[4] Campbell, D. and Lele, S. (2014). An ANOVA test for parameter estimability using data cloning with application to statistical inference for dynamic systems, *Computational Statistics and Data Analysis*, **70**, 257–267.

[5] Carpenter, B.; Gelman, A.; Hoffman, M.D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P. and Riddel, A. (2017). Stan: a probabilistic programming language, *Journal of Statistical Software*, **76**(1).

[6] Florens, J.P. and Simoni, A. (2011). Bayesian identification and partial identification (Unpublished Paper). Available at:
`https://cdn.uclouvain.be/public/Exports%20reddot/core/documents/Simoni.pdf`

[7] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, **1**(3), 515–533.

[8] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**(4), 457–472.

[9] Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

[10] Horn, R.A. and Johnson, C.R. (2012). *Matrix Analysis*, Cambridge University Press, Cambridge.

[11] Koopmans, T.C. and Reiersol, O. (1950). The identification of structural characteristics, *The Annals of Mathematical Statistics*, **21**(2), 165–181.

[12] Lele, S.R.; Dennis, B. and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Mote Carlo methods, *Ecology Letters*, **10**(7), 551–563.

[13] Lele, S.R.; Nadeem, K. and Scmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning, *Journal of the American Statistical Association*, **105**, 1617–1625.

[14] Lindley, D.V. (1971). *Bayesian Statistics: A Review*, Society for Industrial and Applied Mathematics, Philadelphia.

[15] Paulino, C.D.M. and Pereira, C.A.B. (1994). On identifiability of parametric statistical models, *Journal of the Italian Statistical Society*, **3**(1), 125–151.

[16] Peacock, S.J.; Krkošek, M.; Lewis, M.A. and Lele, S. (2017). Study design and parameter estimability for spatial and temporal ecological models, *Ecology and Evolution*, **7**(2), 762–770.

[17] Picci, G. (1977). Some connections between the theory of sufficient statistics and the identifiability problem, *SIAM Journal on Applied Mathematics*, **33**(3), 383–398.

[18] Plummer, M. (2017). *JAGS Version 4.3.0 User Manual.*

[19] R Core Team (2020). *R: A Language and Environment for Statistical Computing*, Vienna, Austria.

[20] Rue, H.; Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society, Series B*, **71**(2), 319–392.

[21] San Martín, E.; González, J. and Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model, *Psychometrika*, **80**(2), 450–467.

[22] San Martín, E. (2018). Identifiability of structural characteristics: how relevant is it for the Bayesian approach?, *Brazilian Journal of Probability and Statistics*, **32**(2), 346–373.

[23] Solymos, P. (2010). dclone: data cloning in R, *The R Journal*, **2**(2), 29–37.

[24] Turkman A.; Paulino, C.D. and Müller, P. (2019). *Computational Bayesian Statistics – An Introduction*, Cambridge University Press, Cambridge.

[25] Villaverde, A.F.; Tsiantis, N. and Banga, J.R. (2019). Full observability and estimation of unknown inputs, states and parameters of nonlinear biological models, *Journal of the Royal Society Interface*, **16**.

[26] Walker, A.M. (1969). On the asymptotic behaviour of posterior distributions, *Journal of the Royal Statistical Society, Series B*, **31**(1), 80–88.