Bayesian model comparison and asymptotics for state-space models

Bayesian model comparison and asymptotics for state-space models

A dissertation presented

by

Stephane Shao

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

April 2019

Dissertation advisor: Pierre Jacob Stephane Shao

Bayesian model comparison and asymptotics for state-space models

Abstract

This thesis studies the implementation and properties of a novel criterion for model comparison, with a keen interest in the task of selecting Bayesian state-space models. This criterion, based on the Hyvärinen score and termed the H-factor, was recently advocated in the decision-theoretic literature as an appealing alternative to the ubiquitous Bayes factor, particularly in settings where the presence of vague prior distributions renders the latter unreliable. The practical use of H-factors requires them to be numerically estimated, which we propose to consistently achieve by using sequential Monte Carlo methods. The uncertainty of the model choice, resulting from this estimation, is quantified by using new advances in unbiased Markov chain Monte Carlo methods to construct confidence intervals for the exact H-factors. Proving theoretical guarantees for this new criterion in large samples will bring us to the realm of Bayesian asymptotics. It will require us to look into the consistency and asymptotic Normality of posterior distributions, à la Bernstein-von Mises, in general and possibly misspecified state-space models.

Contents

Acknowledgments

I would like to extend a million thanks to my advisor Pierre Jacob, who has been a role model to me for the past couple of years, in so many ways. I look up to your incredible work ethics and integrity. Words can barely describe how grateful I am to you, for your infinite patience during those long hours spent debugging my code, for your contagious passion that has kept me motivated throughout my many projects, for the phenomenal attention you put into your teaching and your students, for the memorable conferences and workshops you have encouraged me to attend, for being approachable and always keen to share your knowledge, for being compassionate and understanding when I needed your help. I have become a better researcher and person under your guidance, and I cannot thank you enough for that.

My sincere thanks to Neil Shephard and Randal Douc for agreeing to be part of my thesis committee. As a great enthusiast of your works, I feel extremely honored to have benefited from your expertises as you reviewed my dissertation. I thank you deeply for your time.

I would like to thank Judith Rousseau and Arnaud Doucet for their warm welcome during my visit at Oxford University. You are both giants in your respective areas, yet you explain and communicate ideas very humbly, which I find admirable and truly inspiring.

My heartfelt thanks to Tirthankar Dasgupta and Marie-Abèle Bind. Your kindness and perpetual smiles have eased my difficult start, and brought me much positivity and hope.

Many thanks to the faculty and administrative staff members of Harvard University's Department of Statistics, the former for their valuable feedbacks on my work, the latter for all the organizational efforts seamlessly invested behind the scenes.

A special thought to the devoted students whom I had the pleasure to teach in my first few years. Your eagerness to learn has made every minute of extra preparation worthwhile.

My deepest gratitude to my parents, for supporting my education from the get-go, for enabling opportunities that have led me here, for all the love you give me in your own way. A ton of thanks to my sister for always having my back and helping me feel grounded.

Finally, to all my friends here and across the world, my warmest thanks to you for sticking by my side and supporting me in my hour of need. I have grown and learned immensely from you. I will dearly treasure the memories of our impromptu travels, our late-night conversations, our spicy hotpot and dumpling-wrapping sessions, our intense pool games and clumsy English lessons, our mind-boggling blocus at the library, our carefree fruit breaks on the terrace, the Thanksgiving potlucks, the secret Santas, the surprise mooncakes, the Canadian chocolates, the colorful temaris, the Lego Eiffel tower, the mood-lifting pep talks, the invisible muffins, the paper-quilling owls, and so many more.

Preface

Stemming from the use of a novel criterion to perform Bayesian model comparison in contexts where vague prior distributions are involved, this thesis explores the estimation of Bayesian model comparison criteria, the quantification of their consequent uncertainty, and the derivation of their asymptotic properties. A brief summary of each chapter is provided below.

Chapter 1 [Joint work with Pierre E. Jacob, Jie Ding, and Vahid Tarokh]

We consider the H-factor, based on the Hyvärinen score, as an alternative to traditional Bayes factor in settings where the models at hand may involve vague prior distributions on their parameters. We propose a method to consistently estimate H-factors for parametric models, using sequential Monte Carlo algorithms. We prove the consistency of H-factors under strong regularity assumptions, which notably include the yet unproved consistency of posterior distributions for general state-space models.

Chapter 2 [Joint work with Pierre E. Jacob]

Practical uses of Bayes or H-factors to perform model selection typically involve estimating these intractable quantities, and using estimated factors as proxies to guide the decision process. This raises the question of quantifying the uncertainty induced by this estimation on the final choice. We address this concern by constructing confidence intervals for the intractable factors. Our proposed construction uses recent advances in unbiased Markov chain Monte Carlo methods and is conveniently parallel by design.

Chapter 3 [Joint work with Judith Rousseau, Arnaud Doucet, and Pierre E. Jacob]

The consistency of posterior distributions and their asymptotic Normality are often required to prove theoretical guarantees of Bayesian procedures, and are implicitly assumed in the implementation of many numerical methods. Yet, general proofs of these results in the context of state-space models on non-compact spaces are still elusive. We investigate the ingredients needed for such results to hold, and we look for technical conditions that would allow for possibly misspecified state-space models.

Notations

$\mathbb{K}_+, \mathbb{K}_-, \mathbb{K}^*$ . . . Subsets of non-negative, non-positive, non-zero elements of $\mathbb{K} \subseteq \mathbb{R}$.

$[\![m, n]\!]$ . . . . . Subset of integers $\{i \in \mathbb{Z} : m \leq i \leq n\}$.

$Y_{m:n}$ . . . . . Subsequence $(Y_i)_{m \leq i \leq n}$ for $m \leq n$, and $\emptyset$ if $m > n$ by convention.

$\lambda_k(A)$ . . . . . k-th smallest eigenvalue of a symmetric $A \in \mathbb{R}^{n \times n}$ for $k \in [\![1, n]\!]$.

$\|A\|$ . . . . . Spectral norm of $A \in \mathbb{R}^{m \times n}$ given by $\sqrt{\lambda_n(A^\top A)}$.

$|A|$ . . . . . Determinant of $A \in \mathbb{R}^{m \times n}$.

I_n . . . . . Identity matrix of size $n \times n$.

$\mathbb{1}_B$ . . . . . Indicator function of a set B, i.e. $\mathbb{1}_B(x) = 1$ if $x \in B$ and 0 otherwise.

δ_x . . . . . Dirac measure at x, i.e. $\delta_x(A) = \mathbb{1}_A(x)$ for all events A.

δ_{ij} . . . . . Kronecker delta of (i, j), i.e. $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

$\partial f(y)/\partial y_{(k)}$ . . . k-th partial derivative of $f : \mathbb{R}^d \to \mathbb{R}$ at $y \in \mathbb{R}^d$ for $k \in [\![1, d]\!]$.

$\mathrm{NB}(m, v)$ . . . . Negative binomial distribution with $0 < m < v$ parametrized by its mean and variance. With $p = (v - m)/m$ and $r = m^2/(v - m)$, its probability mass function is given by $k \mapsto \binom{k+r-1}{k}(1-p)^r p^k$ for $k \in \mathbb{N}$.

$\mathrm{Gamma}(\alpha, \beta)$ . . Gamma distribution with shape $\alpha > 0$ and rate $\beta > 0$. Its probability density function is given by $x \mapsto \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}_{(0,+\infty)}(x)$.

$\text{Inv-}\chi^2(\nu, s^2)$ . . . Scaled inverse chi square distribution with degrees of freedom $\nu > 0$ and scale $s > 0$, corresponding to the distribution of the inverse of a $\mathrm{Gamma}(\nu/2, s^2\nu/2)$ random variable. Its probability density function is given by $x \mapsto \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} s^\nu x^{-(\nu/2+1)} e^{-\nu s^2/(2x)} \mathbb{1}_{(0,+\infty)}(x)$.

$t_\nu(\mu, s^2)$ . . . . . Scaled Student's t-distribution with degrees of freedom $\nu > 0$, mean $\mu \in \mathbb{R}$, and scale $s > 0$. Its probability density function is given by $x \to \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi s^2}} \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{s}\right)^2\right)^{-(\nu+1)/2}$.

$d_{TV}(p, q)$ . . . . . Total variation distance between the probability distributions p and q.

“ Il n’y a pas de modèle pour qui cherche ce qu’il n’a jamais vu. ”

— Paul Éluard, *Donner à voir*

1

Bayesian model comparison with the Hyvärinen score

Contents

The work in this chapter has been done in collaboration with Pierre E. Jacob (*Department of Statistics*, Harvard University), Jie Ding (*School of Statistics*, University of Minnesota), and Vahid Tarokh (*Department of Electrical and Computer Engineering*, Duke University). It corresponds to the paper entitled *Bayesian model comparison with the Hyvärinen score: computation and consistency*, recently published in the *Journal of the American Statistical Association*.

1.1 Introduction

1.1.1 Bayesian model comparison

Bayesian model comparison is challenging in situations where the candidate models involve either vague or improper prior distributions on some of their parameters. The Bayes factor (Jeffreys, 1961) between two models — defined as the ratio of their marginal likelihoods — is a widely used approach to model comparison. If one of the candidate models includes the data-generating process, that model is termed well-specified or correct, and the Bayes factor can be interpreted as a ratio of odds, which updates the relative probabilities of the models being correct. In the misspecified or M-open setting (Bernardo and Smith, 2000), the marginal log-likelihood can be interpreted as a measure of out-of-sample predictive performance assessed with the logarithmic scoring rule (e.g. Kass and Raftery, 1995; Key, Pericchi and Smith, 1999; Bernardo and Smith, 2000). Scoring rules are loss functions for the task of predicting an observation y with a probability distribution p, and the logarithmic scoring rule quantifies predictive performance with $-\log p(y)$. Under regularity conditions, the Bayes factor leads to consistent model selection as the number of observations goes to infinity (e.g. Dawid, 2011; Lee and MacEachern, 2011; Walker, 2013; Chib and Kuffner, 2016).

However, if any of the models involves either vague or improper prior distributions on their parameters, the Bayes factor can take arbitrary values and becomes unreliable for any fixed sample size. This is problematic as vague priors are extensively used in practice,

for instance when uniform distributions are specified on intervals of plausible values (e.g. Knape and de Valpine, 2012, see Section 1.4.2). Improper priors also arise from theoretical considerations, for instance as Jeffreys priors (e.g. Chapter 3 of Robert, 2007). Our paper takes the use of such priors by practitioners as a starting point, and addresses the question of model comparison in this context where one cannot rely on the Bayes factor. This limitation of the Bayes factor, sometimes referred to as Bartlett's paradox (Bartlett, 1957; Kass and Raftery, 1995), is a long-lasting challenge in Bayesian model comparison (Chapter 7 of Robert, 2007), as it seems to suggest that prior specification should take into account the potential use (or misuse) of Bayes factors. Many approaches have been proposed to tackle this issue, either by modifying the Bayes factor (e.g. O'Hagan, 1995; Berger and Pericchi, 1996; Berger, Pericchi and Varshavsky, 1998; Berger and Pericchi, 2001) or bypassing it altogether (e.g. Kamary, Mengersen, Robert and Rousseau, 2014, and references therein). In this paper, we investigate an alternative criterion that is 1) principled for any sample size, thanks to an interpretation in terms of predictive performance and scoring rules, 2) enjoys asymptotic consistency properties, and 3) is robust to the arbitrary vagueness of prior distributions.

Since the Bayes factor is associated with predictive performance under the logarithmic scoring rule, natural alternatives arise by considering other scoring rules (Dawid and Musio, 2015; Dawid, Musio and Ventura, 2016). We consider the *Hyvärinen score* (Hyvärinen, 2005), which is *proper*, *local*, and *homogeneous* (Dawid and Lauritzen, 2005; Parry, Dawid and Lauritzen, 2012; Ehm and Gneiting, 2012). Given T observations $(y_1, ..., y_T) \in \mathbb{Y}^T$ and a finite set $\mathcal{M}$ of candidate models, each inducing a joint marginal density of $(Y_1, ..., Y_T)$ denoted by p_M for $M \in \mathcal{M}$, we can regard the log-Bayes factor as a comparison of predictive sequential (or *prequential*, Dawid, 1984) log-score $-\log p_M(y_{1:T}) = \sum_{t=1}^{T} -\log p_M(y_t|y_{1:t-1})$, where by convention $p_M(y_1|y_{1:0})$ denotes the prior predictive distribution of Y_1 under model M. By contrast, for any d_y-dimensional observation $y \in \mathbb{R}^{d_y}$ and twice differentiable density

p on $\mathbb{R}^{d_y}$, the Hyvärinen score is defined as

$$\mathcal{H}(y,p) = 2\,\Delta \log p(y) + \|\nabla \log p(y)\|^2, \tag{1.1}$$

where ∇ and Δ respectively denote the gradient and Laplacian operators with respect to the variable y. We would then select the model with the smallest prequential Hyvärinen score

$$\mathcal{H}_T(M) = \sum_{t=1}^{T} \mathcal{H}\left(y_t, p_M(dy_t|y_{1:t-1})\right). \tag{1.2}$$

We will refer to this prequential Hyvärinen score as the *H-score*. Homogeneity is the key property of the Hyvärinen score which is not shared by the logarithmic scoring rule. It ensures that the score does not depend on normalizing constants of candidate densities, hence offering robustness to vague priors and allowing for improper priors. For example, if M denotes the toy model $Y_1, ..., Y_T \,|\, \mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ with prior $\mu \sim \mathcal{N}(0, \sigma_0^2)$ and known hyperparameter $\sigma_0 > 0$, then $Y_t \,|\, Y_{1:t-1} \sim \mathcal{N}\left(\mu_{t-1}, \sigma_{t-1}^2 + 1\right)$ for all $t \in [\![0, T]\!]$ by conjugacy, where $\sigma_t^2 = (t + \sigma_0^{-2})^{-1}$ and $\mu_t = \sigma_t^2 \sum_{i=1}^{t} Y_i$ for all $t \in [\![1, T]\!]$. The log-score $-\log p_M(y_{1:T})$ becomes equivalent to $\log \sigma_0$ when $\sigma_0 \to +\infty$, and thus diverges to $+\infty$ as σ_0 increases. In other words, one could obtain Bayes factors that prefer virtually any other model over this one, by simply increasing σ_0 thus making the prior on μ arbitrarily vague, for any fixed number of observations T. On the other hand, the prequential Hyvärinen score, computed from (1.1) and (1.2) using conjugacy, converges to a finite limit as $\sigma_0 \to +\infty$, so that increasing σ_0 can only influence the prequential Hyvärinen score to a limited extent. Throughout the article, the notion of robustness to arbitrary vagueness of priors is to be understood in that sense. Such a robustness is desirable when models are misspecified or when the specification of vague priors is dictated by practical considerations rather than a genuine reflection of one's prior knowledge, as is sometimes the case for parameters of complex state-space models (e.g. see Section 1.4.2). The limit of $\mathcal{H}_T(M)$ as $\sigma_0 \to +\infty$ also unambiguously defines the value of the score for a flat improper prior $p(\mu) \propto 1$.

1. Bayesian model comparison with the Hyvärinen score

Without conjugacy, the calculation of the Hyvärinen score involves typically intractable integrals with respect to the sequence of partial posteriors. In this chapter, we show how to use sequential Monte Carlo (SMC) methods to consistently estimate H-scores, thereby enabling their use in Bayesian model comparison for general parametric models. More specifically, we show that this estimation can be achieved for models with tractable likelihoods via SMC samplers (Chopin, 2002; Del Moral, Doucet and Jasra, 2006; Zhou, Johansen and Aston, 2016). Furthermore, the case of generic state-space models can be covered by using SMC^2 (Fulop and Li, 2013; Chopin, Jacob and Papaspiliopoulos, 2013) under the mild requirement that we can simulate the latent state process and evaluate the measurement density (Bretó, He, Ionides and King, 2009; Andrieu, Doucet and Holenstein, 2010), plus some integrability conditions. Our second contribution is to prove that, under regularity conditions allowing for misspecified settings, the H-score is consistent for model selection. Finally, motivated by an application to count-valued data in a population dynamics context, we propose a modified score for discrete observations that builds on recent complete characterizations of proper scoring rules on discrete spaces (McCarthy, 1956; Hendrickson and Buehler, 1971; Dawid, Lauritzen and Parry, 2012; Dawid, Musio and Columbu, 2017).

This chapter is organized as follows. In Section 1.2, we consider parametric models with tractable likelihoods. We present how the H-score can be estimated via SMC samplers, and show that it leads to consistent model selection, under regularity assumptions. In Section 1.3, we generalize the approach to nonlinear non-Gaussian state-space models, using SMC^2, and we present a simulation study with Lévy-driven stochastic volatility models. In Section 1.4, we extend the proposed criterion to discrete observations and compare diffusion models for population dynamics. Possible limitations and directions for future research are discussed in Section 1.5. Implementation details are presented in Appendix A. Formal proofs and additional simulations are provided in Appendix B. The R code producing the figures is available at github.com/pierrejacob/bayeshscore.

1.1.2 Terminology and additional notations

Given two models M_1 and M_2, the difference of their H-scores $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)$ will be termed the *H-factor of* M_1 *against* M_2. Each observation $y = (y_{(1)}, ..., y_{(d_y)})^\top$ is a vector of dimension $d_y \in \mathbb{N}^*$ and takes values in $\mathbb{Y} \subseteq \mathbb{R}^{d_y}$. Aside from Section 1.4, the observations are assumed to be continuous variables. Continuous probability distributions are assumed to admit densities with respect to the Lebesgue measure. We let $\mathbb{P}_\star$ (resp. $\mathbb{E}_\star$) denote the probability (resp. expectation) induced by the data-generating mechanism of the stochastic process $(Y_t)_{t\in\mathbb{N}^*}$. We use the abbreviation $\mathbb{P}_\star$-a.s. for $\mathbb{P}_\star$-*almost surely*. Assuming its existence, we let $p_\star$ denote the probability density or mass function associated with $\mathbb{P}_\star$. When dealing concurrently with several models from a set $\mathcal{M} = \{M_j : j \in [\![1, m]\!]\}$, we use the subscript $j \in [\![1, m]\!]$ to condition on a particular model. Each candidate model M_j is parametrized by a parameter θ_j in a space $\mathbb{T}_j \subseteq \mathbb{R}^{d_{\theta_j}}$ of dimension $d_{\theta_j} \in \mathbb{N}^*$. Explicit dependence on models is dropped from the notation whenever possible. The H-score $\mathcal{H}_T(M)$ of a model M is a random variable that depends on the observations, although the notation makes this dependence implicit. When the context is clear, we use the notation $\mathcal{H}_T(M)$ to denote both the random score and its realization.

1.2 H-score for models with tractable likelihoods

We first describe how the H-score can be estimated with SMC samplers, before turning to asymptotic properties and numerical investigations. The H-score of a model M, defined in (1.2), can be rewritten as

$$\mathcal{H}_T(M) = \sum_{t=1}^{T} \sum_{k=1}^{d_y} \left(2 \frac{\partial^2 \log p(y_t|y_{1:t-1})}{\partial y_{t(k)}^2} + \left(\frac{\partial \log p(y_t|y_{1:t-1})}{\partial y_{t(k)}} \right)^2 \right). \tag{1.3}$$

The marginal predictive densities appearing in (1.3) correspond to integrals with respect to posterior distributions, as $p(y_t|y_{1:t-1}) = \int p(y_t|\theta, y_{1:t-1})\, p(\theta|y_{1:t-1})\, d\theta$.

1.2.1 Computation of the H-score using SMC

As noted in Dawid and Musio (2015), an interchange of differentiation and integration under appropriate regularity conditions (see Appendix B.4) shows that $\mathcal{H}_T(M)$ is equal to

$$\sum_{t=1}^{T}\sum_{k=1}^{d_y}\left(2\,\mathbb{E}_t\left[\frac{\partial^2 \log p(y_t|y_{1:t-1},\Theta)}{\partial y_{t(k)}^2}+\left(\frac{\partial \log p(y_t|y_{1:t-1},\Theta)}{\partial y_{t(k)}}\right)^2\right]-\left(\mathbb{E}_t\left[\frac{\partial \log p(y_t|y_{1:t-1},\Theta)}{\partial y_{t(k)}}\right]\right)^2\right) \quad (1.4)$$

where the conditional expectations $\mathbb{E}_t$ are taken with respect to the posterior distributions $\Theta \sim p(d\theta|y_{1:t})$. The terms of the sum in (1.4) might not be well-defined when improper posterior distributions arise from improper priors. If τ denotes the first index such that the posterior $p(d\theta|y_{1:\tau})$ is proper, then we would redefine the H-score as $\sum_{t=\tau}^{T}\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1})\right)$. This issue is not specific to the H-score, and for simplicity of exposition, we will thereafter assume that posterior distributions are proper after assimilating one observation.

In general, expectations with respect to $p(d\theta|y_{1:t})$ for all successive $t \in [\![1, T]\!]$ can be consistently estimated using sequential or annealed importance sampling (Neal, 2001) and SMC samplers (Chopin, 2002; Del Moral et al., 2006). An SMC sampler starts by sampling a set of N_θ particles $\theta^{(1:N_\theta)} = (\theta^{(1)}, \ldots, \theta^{(N_\theta)})$ independently from an initial distribution $q(d\theta)$, e.g. a uniform distribution on a set (e.g. Fearnhead and Taylor, 2013), the prior distribution $p(d\theta)$ when it is proper, or more generally an approximation of the first proper posterior distribution. The algorithm then assigns weights, resamples, and moves these particles in order to approximate $p(d\theta|y_{1:t})$ for each $t \in [\![1, T]\!]$. We can move samples from a posterior distribution to the next by successively targeting intermediate distributions whose densities are proportional to $p(\theta|y_{1:t})p(y_{t+1}|y_{1:t},\theta)^{\gamma_{t,j}}$, where $0 = \gamma_{t,0} < \gamma_{t,1} < \ldots < \gamma_{t,J_t} = 1$ with $J_t \in \mathbb{N}^*$. The temperatures $\gamma_{t,j}$ can be determined adaptively to maintain a chosen level of non-degeneracy in the importance weights of the particles, e.g. by forcing the effective sample size to stay above a desired threshold or by imposing a minimum number of unique particles. The resampling steps can be performed with various schemes (see Douc and

Cappé, 2005; Murray, Lee and Jacob, 2016; Gerber, Chopin and Whiteley, 2017), and the move steps with any Markov chain Monte Carlo (MCMC) method. Further details about the implementation of SMC samplers are provided in Appendix A.1. In the numerical experiments of this chapter and its supplementary material in Appendix B, the resampling is done with the Srinivasan sampling process (SSP) described in Gerber et al. (2017), and the moves consist of independent Metropolis–Hastings steps with proposals obtained as mixtures of Normal distributions fitted on the current weighted particles. We will use prior distributions as a choice of initial distribution $q(d\theta)$.

Sequential estimation of the H-score can thus be achieved at a cost comparable to that of estimating the log-evidence. Indeed, both can be obtained from the same SMC runs. However, numerical experiments suggest that the estimator of the H-score tends to have a larger relative variance than the estimator of the log-evidence, for a given number of particles. This can be explained informally as follows. For the evidence, the Monte Carlo approaches approximate expectations of the form $\mathbb{E}[p(y_t|y_{1:t-1}, \Theta)]$ with respect to the posterior $p(d\theta|y_{1:t-1})$. On the other hand, the H-score involves expectations such as $\mathbb{E}[\nabla_y \log p(y_t|y_{1:t-1}, \Theta)]$ with respect to $p(d\theta|y_{1:t})$. When t is large, the distributions $p(d\theta|y_{1:t-1})$ and $p(d\theta|y_{1:t})$ are similar, whereas the integrands $\theta \mapsto p(y_t|y_{1:t-1}, \theta)$ and $\theta \mapsto \nabla_y \log p(y_t|y_{1:t-1}, \theta)$ are different. In some generality, the first type of integrands will be easier to integrate than the second one, e.g. when the former is bounded in θ while the latter is polynomial in θ, as in Normal location models (see Section 1.2.3).

1.2.2 Consistency of the H-score for i.i.d. settings

Irrespective of model misspecification, the H-score can be justified for finite samples since it results from assessing predictions with a scoring rule that satisfies desirable properties such as propriety, locality, and homogeneity (Parry et al., 2012; Ehm and Gneiting, 2012). Moreover, under regularity conditions, we can show that the H-score also satisfies sensible

asymptotic properties: as the number of observations grows, choosing the model with the smallest H-score eventually leads to selecting the model closest to the data-generating process in a certain sense, as made precise below. Some general perspective on consistency of prequential scores can be found in Dawid and Musio (2015).

Here we consider i.i.d. models and assume that $(Y_t)_{t\in\mathbb{N}^*}$ is a sequence of i.i.d. observations drawn from $p_\star$. State-space models and more general data-generating processes will be covered in Section 1.3.2. For simplicity, we focus on continuous univariate ($d_y = 1$) observations. Our results will only be meaningful for models that are either non-nested, or nested with at most one model being well-specified. The case of well-specified nested models is discussed at the end of this section, with more details in Appendix B.5.4. Our consistency result rely on the expression

$$\mathcal{H}_T(M) = \left(\sum_{t=1}^{T} \mathbb{E}_t\left[\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \Theta)\right)\right]\right) + \left(\sum_{t=1}^{T} \mathbb{V}_t\left[\frac{\partial \log p(y_t|y_{1:t-1}, \Theta)}{\partial y_t}\right]\right) \tag{1.5}$$

which follows directly from rearranging the terms in (1.4), where $\mathbb{E}_t$ and $\mathbb{V}_t$ respectively denote conditional expectations and variances with respect to $\Theta \sim p(d\theta|y_{1:t})$. The key insight is that, in non-nested settings, as the number of observations grows and the posterior distribution $p(d\theta|y_{1:T})$ concentrates to a point mass, the sum of the conditional expectations in (1.5) will eventually dominate and drive the behavior of the H-score, while the sum of the conditional variances acts as a penalty term that becomes negligible. This penalty term only becomes crucial when comparing well-specified nested models, as discussed at the end of this section.

The result below considers model selection consistency for two i.i.d. models M_1 and M_2, each describing the data respectively as $Y_1, ..., Y_T \,|\, \theta_j \overset{\text{i.i.d.}}{\sim} p_j(dy|\theta_j)$, with parameter $\theta_j \in \mathbb{T}_j$ and prior density $p_j(\theta_j)$, for $j \in \{1, 2\}$.

Theorem 1. *Assume $(Y_t)_{t\in\mathbb{N}^*}$ is a sequence of i.i.d. draws from $p_\star$. Assume M_1 and M_2 both satisfy the following conditions, where models are omitted from the notation and probabilistic statements are $\mathbb{P}_\star$-almost sure:*

1. Bayesian model comparison with the Hyvärinen score

(a) For all $t \in \mathbb{N}^*$ *and* $y_{1:t} \in \mathbb{Y}^t$, $\theta \mapsto p(y_t|\theta)\, p(\theta|y_{1:t-1})$ *is integrable on* $\mathbb{T}$.

(b) For all $t \in \mathbb{N}^*$ *and* $\theta \in \mathbb{T}$, $y_t \mapsto p(y_t|\theta)$ *is twice differentiable on* $\mathbb{Y}$.

(c) For all $t \in \mathbb{N}^*$, *there exist integrable functions* $h_{1,t}$ *and* $h_{2,t}$ *such that, for all* $(y_{1:t}, \theta) \in \mathbb{Y}^t \times \mathbb{T}$, $|p(\theta|y_{1:t-1})\, \partial p(y_t|\theta)/\partial y_t| \leq h_{1,t}(\theta)$ *and* $|p(\theta|y_{1:t-1})\, \partial^2 p(y_t|\theta)/\partial {y_t}^2| \leq h_{2,t}(\theta)$.

(d) There exists $\theta^\star \in \mathbb{T}$ *such that, if* $\Theta_t \sim p(d\theta|Y_{1:t})$ *for all* $t \in \mathbb{N}^*$, *then* $\Theta_t \xrightarrow[t\to+\infty]{\mathcal{D}} \theta^\star$.

(e) There exist a constant $L > 0$ *and a neighborhood* $\mathcal{U}_{\theta^\star}$ *of* $\theta^\star$ *such that, for all* $t \in \mathbb{N}^*$, $\theta \mapsto \mathcal{H}\left(Y_t, p(dy_t|\theta)\right)$ *and* $\theta \mapsto \partial \log p(Y_t|\theta)/\partial y_t$ *are* L*-Lipschitz functions.*

(f) There exist $\alpha_1 > 1$ *and* $\alpha_2 > 1$ *such that* $\sup_{t\in\mathbb{N}^*} \mathbb{E}\left[\, |\mathcal{H}\left(Y_t, p(dy_t|\Theta_t)\right)|^{\alpha_1} \,|\, Y_{1:t}\,\right] < +\infty$ *and* $\sup_{t\in\mathbb{N}^*} \mathbb{E}\left[(\partial \log p(Y_t|\Theta_t)/\partial y_t)^{2\,\alpha_2} \,|\, Y_{1:t}\right] < +\infty$, *where the conditional expectations are with respect to the posterior distribution* $\Theta_t \sim p(d\theta|Y_{1:t})$.

(g) $\mathbb{E}_\star\left[|\mathcal{H}\left(Y, p(dy|\theta^\star)\right)|\right] < +\infty$ *and* $p_\star(y)\, \partial \log p(y|\theta^\star)/\partial y \xrightarrow[|y|\to+\infty]{} 0$.

We also assume that the data-generating density $p_\star$ *is such that* $y \mapsto p_\star(y)$ *is twice differentiable and* $\mathbb{E}_\star[|\mathcal{H}\left(Y, p_\star(dy)\right)|] < +\infty$. *If all the conditions are met, then we have*

$$\frac{1}{T}\left(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \mathrm{D}_{\mathcal{H}}(p_\star, M_2) - \mathrm{D}_{\mathcal{H}}(p_\star, M_1) \tag{1.6}$$

where, for each $j \in \{1, 2\}$, *the quantity*

$$\mathrm{D}_{\mathcal{H}}(p_\star, M_j) = \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_j(dy|\theta_j^\star)\right)\right] - \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_\star(dy)\right)\right] \tag{1.7}$$

satisfies $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) \geq 0$, *with* $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) = 0$ *if and only if* $p_j(y|\theta_j^\star) = p_\star(y)$ *for all* $y \in \mathbb{Y}$.

The assumptions listed in Theorem 1 are strong, which allows for more intuitive proofs. Our numerical experiments suggest that (1.6) can hold when these conditions are not met (e.g. see Section 1.2.3). Conditions (a) to (c) ensure the validity of (1.5); (d) assumes the concentration of the posterior to a point mass; (e) to (f) ensure suitable convergence of

posterior moments; and (g) ensures the strict propriety of the H-score and its definiteness for $p_\star$. Further discussion on these conditions and detailed proofs are provided in Appendix B.5.

Theorem 1 provides insights into the asymptotic behavior of the H-score. Using integration by parts, we have

$$\mathrm{D}_{\mathcal{H}}(p_\star, M_j) = \int \left(\frac{\partial \log p_\star(y)}{\partial y} - \frac{\partial \log p_j(y|\theta_j^\star)}{\partial y} \right)^2 p_\star(y) dy \tag{1.8}$$

so that $\mathrm{D}_{\mathcal{H}}(p_\star, M_j)$ can be interpreted as a divergence between the data-generating distribution $p_\star$ and model M_j. As long as $\mathbb{E}_\star\left[\mathcal{H}\left(Y, p_1(dy|\theta_1^\star)\right)\right] \neq \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_2(dy|\theta_2^\star)\right)\right]$, the H-score asymptotically chooses the model closest to the data-generating distribution $p_\star$ with respect to the divergence $\mathrm{D}_{\mathcal{H}}$. In particular, if M_1 is well-specified and M_2 is misspecified, then $\mathrm{D}_{\mathcal{H}}(p_\star, M_1) = 0 < \mathrm{D}_{\mathcal{H}}(p_\star, M_2)$, which leads to $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) > 0$ for all sufficiently large T, $\mathbb{P}_\star$-almost surely. In other words, the H-score eventually chooses a well-specified model M_1 over a misspecified model M_2.

The divergence $\mathrm{D}_{\mathcal{H}}(p_\star, M_j)$ appearing in (1.8) is sometimes referred to as the *relative Fisher information divergence* between $p_\star$ and $p_j(dy|\theta_j^\star)$ (e.g. Walker, 2016; Holmes and Walker, 2017). It should be contrasted to the divergence associated with the log-score: under similar assumptions, one can prove (e.g. Dawid, 2011) that

$$\frac{1}{T}\left(\left(-\log p_2(Y_{1:T})\right) - \left(-\log p_1(Y_{1:T})\right)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \mathrm{KL}(p_\star, M_2) - \mathrm{KL}(p_\star, M_1)$$

where $\mathrm{KL}(p_\star, M_j) = \mathbb{E}_\star\left[-\log p_j(Y|\theta_j^\star)\right] - \mathbb{E}_\star\left[-\log p_\star(Y)\right]$ denotes the Kullback-Leibler divergence between $p_\star$ and $p_j(dy|\theta_j^\star)$. In other words, the log-score $-\log p_j(Y_{1:T})$ asymptotically favors the model that is the closest to $p_\star$ with respect to the Kullback-Leibler divergence $\mathrm{KL}(p_\star, M_j)$, whereas the H-score $\mathcal{H}_T(M_j)$ asymptotically favors the model that is the closest to $p_\star$ with respect to the divergence $\mathrm{D}_{\mathcal{H}}(p_\star, M_j)$.

When only one of the candidate models is well-specified, the log-Bayes factor and the H-factor both agree on consistently selecting it. When both M_1 and M_2 are misspecified,

each criterion selects a model according to its associated divergence. Despite being related (e.g. Bobkov, Gozlan, Roberto and Samson, 2014, and references therein), the geometries induced by these divergences differ, leading the log-Bayes factor and the H-factor to select possibly different models (see case 3 in Section 1.2.3). In the presence of informative priors, deciding which score to use in such misspecified settings is then a matter of preferences and further practical considerations; in this article we focus on the case of vague priors for which Bayes factors are not recommendable, as discussed earlier.

If $\mathbb{E}_\star\left[\mathcal{H}\left(Y, p_1(dy|\theta_1^\star)\right)\right] = \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_2(dy|\theta_2^\star)\right)\right]$, the limit in (1.6) becomes 0 and calls for a more careful look at the higher order penalty term formed by the conditional variances in (1.5). Such a refinement is needed if M_1 is nested in M_2, in the sense of Eq. (9) in Berger and Pericchi (1996), and both models are well-specified. In other words, we have $\mathbb{T}_2 = \{(\theta_1, \eta) \in \Xi_1 \times \Xi_2\} \subseteq \mathbb{R}^{k_1} \times \mathbb{R}^{k_2 - k_1}$ and $\mathbb{T}_1 \subseteq \Xi_1$ for some $k_1, k_2 \in \mathbb{N}$ with $k_2 > k_1 > 0$, and there exists $\eta_1^\star \in \Xi_2$ such that $p_1(y|\theta_1) = p_2(y|\theta_1, \eta_1^\star)$ for all $(y, \theta_1) \in \mathbb{Y} \times \mathbb{T}_1$. There also exists $\theta_1^\star \in \mathbb{T}_1$ such that $p_\star(y) = p_1(y|\theta_1^\star) = p_2(y|\theta_2^\star)$ for all $y \in \mathbb{Y}$, where $\theta_2^\star = (\theta_1^\star, \eta_1^\star)$. The particular case of nested Normal linear models is discussed in Sections 8 and 9 of Dawid and Musio (2015). Under regularity conditions, and if the parameters are orthogonal such that $\mathbb{E}_\star[\nabla_\eta \nabla_{\theta_1} \log p_2(Y|\theta_1^\star, \eta_1^\star)] = 0$, we conjecture that

$$\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) = \delta_{21} \log T + o(\log T)$$

as $T \to \infty$, in $\mathbb{P}_\star$-probability, where the difference δ_{21} in model dimensions appears as

$$\delta_{21} = \mathbb{E}_\star\left[\left(\nabla_\eta \frac{\partial \log p_2(Y|\theta_2^\star)}{\partial y}\right)^\top \mathbb{E}_\star[-\nabla_\eta^2 \log p_2(Y|\theta_2^\star)]^{-1} \left(\nabla_\eta \frac{\partial \log p_2(Y|\theta_2^\star)}{\partial y}\right)\right] > 0.$$

This would imply that $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) \to +\infty$ as $T \to +\infty$, in $\mathbb{P}_\star$-probability, so that the H-score asymptotically chooses the model M_1 of smaller dimension, similarly to the log-Bayes factor for which $\log p_1(Y_{1:T}) - \log p_2(Y_{1:T}) = (1/2)(k_2 - k_1) \log T + o(\log T)$ under suitable assumptions (e.g. Moreno, Girón and Casella, 2010; Rousseau and Taeryon, 2012;

Chib and Kuffner, 2016). Heuristic justification and numerical illustration of this postulate are provided in Appendices B.5.4 and B.5.5. We leave more formal studies of the H-score in nested well-specified settings for future research.

As an aside, we need to contrast the prequential approach described in (1.2) with a batch approach, where one would assess the predictive performance of model M at once via $\mathcal{H}_T^{batch}(M) = \mathcal{H}\left(y_{1:T}, p_M(dy_{1:T})\right)$. This batch approach would allow approximations using standard Markov chain Monte Carlo methods. However, the batch approach is generally not consistent for model selection (see Section 8.1 in Dawid and Musio, 2015). Therefore, the prequential framework not only has a natural interpretation that relates to sequential probability forecasts (Dawid, 1984), but is also necessary for consistency. This leads to the task of approximating all the successive predictive distributions $p(dy_t|y_{1:t-1})$, as described in Section 1.2.1. This distinction does not arise for the log-score, for which we always have $-\log p(y_{1:T}) = -\sum_{t=1}^{T} \log p(y_t|y_{1:t-1})$. One consequence of the sequential approach is that different orderings of the observations lead to different sequences of predictive distributions, hence yielding different values of the H-score. This might be undesirable in settings where the observations are not naturally ordered (e.g. i.i.d. or spatial data). For large samples, this issue is mitigated by the convergence of rescaled H-scores to limits that do not depend on the ordering of the observations (cf. Theorem 1). For small samples, one could average the H-score over different permutations of the data, or use a random ordering of the data within each SMC run (see Section 1.2.3), at the cost of extra computations.

1.2.3 Numerical illustration with Normal models

Inspired by Section 3.2. of O'Hagan (1995), we consider the two Normal models

$$M_1: \quad Y_1, ..., Y_T \,|\, \theta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta_1, 1\right), \quad \theta_1 \sim \mathcal{N}\left(0, \sigma_0^2\right).$$
$$M_2: \quad Y_1, ..., Y_T \,|\, \theta_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \theta_2\right), \quad \theta_2 \sim \text{Inv-}\chi^2\left(\nu_0, s_0^2\right).$$

1. Bayesian model comparison with the Hyvärinen score

The positive hyperparameters are chosen as $\sigma_0^2 = 10$, $\nu_0 = 0.1$, and $s_0^2 = 1$. We compare M_1 and M_2, using data generated as $Y_1, ..., Y_T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_\star, \sigma_\star^2)$, in the following four settings: (1) $(\mu_\star, \sigma_\star^2) = (1, 1)$, i.e. M_1 is well-specified while M_2 is not; (2) $(\mu_\star, \sigma_\star^2) = (0, 5)$, i.e. M_2 is well-specified while M_1 is not; (3) $(\mu_\star, \sigma_\star^2) = (4, 3)$, i.e. both M_1 and M_2 are misspecified; (4) $(\mu_\star, \sigma_\star^2) = (0, 1)$, i.e. both M_1 and M_2 are well-specified.

Conjugacy allows all the posterior distributions, scores, and divergences to be computed in closed form. The posteriors under M_1 and M_2 concentrate respectively around $\theta_1^\star = \mu_\star$ and $\theta_2^\star = \sigma_\star^2 + \mu_\star^2$. We compute $\mathrm{D}_\mathcal{H}$ and the Kullback-Leibler divergence for Normal densities analytically (see Section 6.1 of Dawid and Musio, 2015) and get the theoretical limits

$$\mathrm{D}_\mathcal{H}(p_\star, M_2) - \mathrm{D}_\mathcal{H}(p_\star, M_1) = \frac{\mu_\star^2}{\sigma_\star^2 (\mu_\star^2 + \sigma_\star^2)} - \frac{(\sigma_\star^2 - 1)^2}{\sigma_\star^2}, \tag{1.9}$$

$$\mathrm{KL}(p_\star, M_2) - \mathrm{KL}(p_\star, M_1) = \frac{1}{2} \log \left(\frac{\mu_\star^2 + \sigma_\star^2}{\sigma_\star^2} \right) - \frac{(\sigma_\star^2 - 1) - \log(\sigma_\star^2)}{2}, \tag{1.10}$$

which depend on the values of $|\mu_\star|$ and $\sigma_\star^2$. For each of the four cases, we generate $T = 1000$ observations and perform 5 runs of SMC with $N_\theta = 1024$ particles to estimate the log-Bayes factors and H-factors of M_1 against M_2. Each run uses a different ordering of the data, sampled uniformly from all the possible permutations. The results are shown in Figure 1.1. H-factors and log-Bayes factors are overlaid on the same plots in order to track their evolution jointly, but their values should not be directly compared. As expected in cases 1 and 2, the H-factor selects the well-specified model and diverges to infinity at a linear rate, with respective slopes matching the theoretical limits 0.5 and -3.2 from (1.9). Similar behavior is obtained for the log-Bayes factor, which correctly diverges to infinity at the same linear rate, with theoretical slopes given by (1.10). In case 3, both models are misspecified, and (1.9)-(1.10) with $(\mu_\star, \sigma_\star^2) = (4, 3)$ yield $\mathrm{D}_\mathcal{H}(p_\star, M_2) - \mathrm{D}_\mathcal{H}(p_\star, M_1) \approx -1.05 < 0$ and $\mathrm{KL}(p_\star, M_2) - \mathrm{KL}(p_\star, M_1) \approx 0.47 > 0$. This leads the Bayes factor and the H-factor to favor different misspecified models. In fact, when both M_1 and M_2 are misspecified, there are

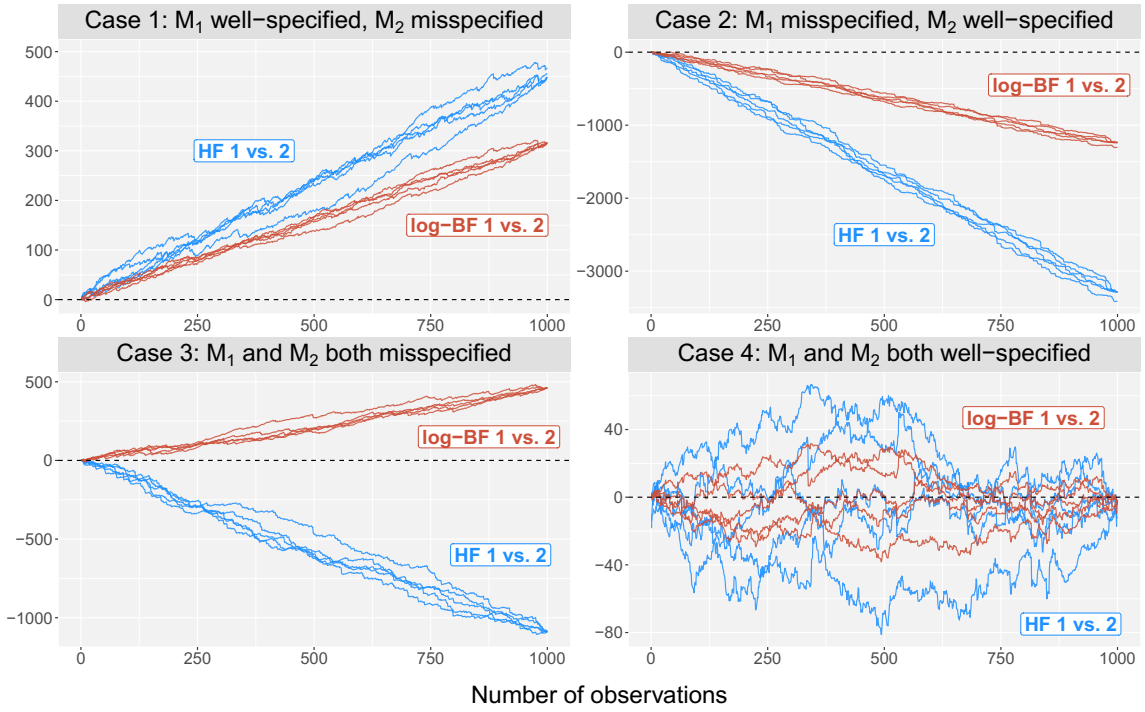

Figure 1.1. *Estimated log-Bayes factors (log-BF) and H-factors (HF) of M_1 against M_2, computed for 5 replications (thin solid lines), under four i.i.d. data-generating processes: $\mathcal{N}(1,1)$ (Case 1), $\mathcal{N}(0,5)$ (Case 2), $\mathcal{N}(4,3)$ (Case 3), and $\mathcal{N}(0,1)$ (Case 4). In each plot, the observations are fixed but randomly ordered, so that the variability within each factor is due to Monte Carlo error and random permutation of the data. See Section 1.2.3.*

infinitely many combinations of $(|\mu_\star|, \sigma_\star^2) \in \mathbb{R}_+^2$ for which $\mathrm{D}_{\mathcal{H}}(p_\star, M_2) < \mathrm{D}_{\mathcal{H}}(p_\star, M_1)$ whereas $\mathrm{KL}(p_\star, M_2) > \mathrm{KL}(p_\star, M_1)$. Indeed, if we define the boundary $\mathcal{B}_{\mathcal{H}}(\sigma_\star^2) = |\sigma_\star^2 - 1|(2 - \sigma_\star^2)^{-1/2}$ for $\sigma_\star^2 \in (0, 2)$ and $\mathcal{B}_{\mathcal{H}}(\sigma_\star^2) = +\infty$ for $\sigma_\star^2 \in [2, +\infty)$, then we have $\mathrm{D}_{\mathcal{H}}(p_\star, M_2) = \mathrm{D}_{\mathcal{H}}(p_\star, M_1)$ (resp. $>$ and $<$) for $|\mu_\star| = \mathcal{B}_{\mathcal{H}}(\sigma_\star^2)$ (resp. $>$ and $<$). By contrast, $\mathrm{KL}(p_\star, M_2) = \mathrm{KL}(p_\star, M_1)$ if and only if $|\mu_\star| = \mathcal{B}_{\mathrm{KL}}(\sigma_\star^2)$, where $\mathcal{B}_{\mathrm{KL}}(\sigma_\star^2) = (\exp(\sigma_\star^2 - 1) - \sigma_\star^2)^{1/2}$ for all $\sigma_\star^2 > 0$. Thus, whenever $\mathcal{B}_{\mathrm{KL}}(\sigma_\star^2) < |\mu_\star| < \mathcal{B}_{\mathcal{H}}(\sigma_\star^2)$, the divergences $\mathrm{D}_{\mathcal{H}}$ and KL disagree on which model is closer to $p_\star$. This is illustrated in Figure 1.2. When both divergences are sensible, deciding which one to use would require further considerations (e.g. see Jewson, Smith and Holmes, 2018). As explained in Section 1.1, the log-Bayes factor might be inappropriate in the presence of vague priors. Looking back at case 1 for example, since $\log p_{M_1}(y_{1:T}) \to -\infty$ when

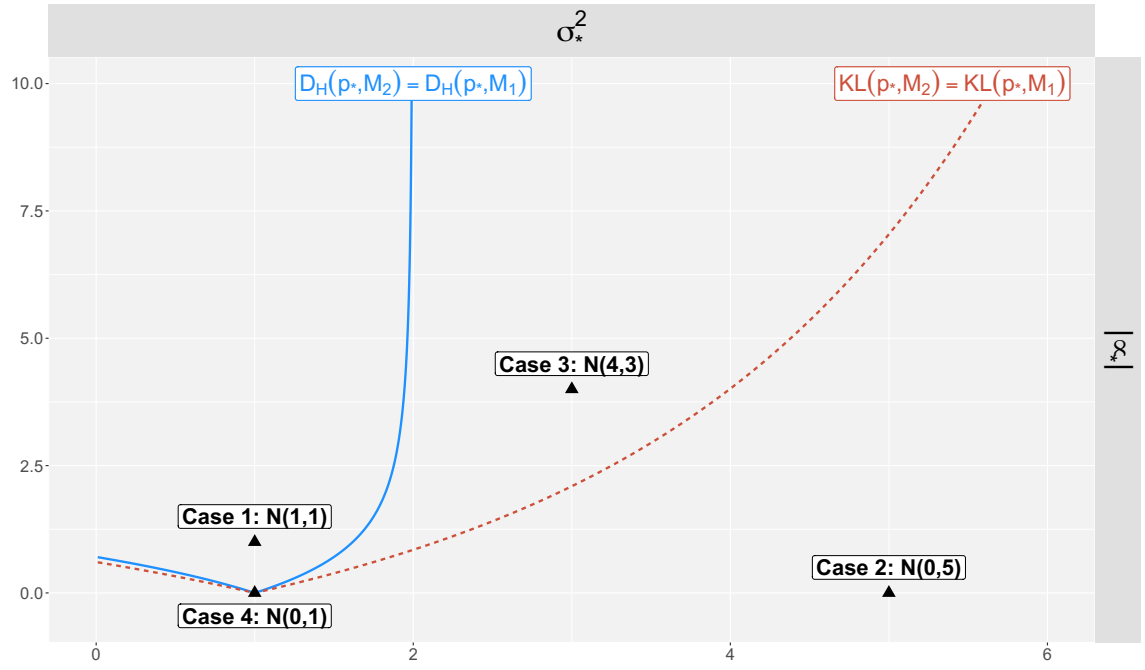

Figure 1.2. *Phase plane of* $d(p_\star, M_2) - d(p_\star, M_1)$ *as a function of* $(|\mu_\star|, \sigma_\star^2)$*, where* $d \in \{D_\mathcal{H}, \mathrm{KL}\}$*. The four cases from Section 1.2.3 are indicated as triangles. The lines (solid for* $D_\mathcal{H}$*, dashed for* KL*) are the sets of* $(|\mu_\star|, \sigma_\star^2)$ *such that* $d(p_\star, M_2) = d(p_\star, M_1)$*. The regions above (resp. below) the lines satisfy* $d(p_\star, M_2) > d(p_\star, M_1)$ *(resp.* $<$*), i.e.* M_1 *(resp.* M_2*) is closer to* $p_\star$*.*

$\sigma_0 \to +\infty$, one could always specify a σ_0 large enough such that the log-Bayes factor would wrongly pick M_2. On the other hand, the choice of M_1 by the H-factor remains unchanged when σ_0 increases. This robustness is further illustrated in Appendix B.1.

Finally, in case 4, the theoretical slopes are exactly 0, while the models are of equal dimensions, hence no model prevails.

1.3 H-score for state-space models

The H-score raises additional computational challenges in the case of state-space models. State-space models, also known as hidden Markov models, are a flexible and widely used class of time series models (Cappé, Moulines and Rydén, 2005; Douc, Moulines and Stoffer, 2014), which describe the observations $(Y_t)_{t \in \mathbb{N}^*}$ as conditionally independent given a latent

Markov chain $(X_t)_{t\in\mathbb{N}^*}$ living in $\mathbb{X}\subseteq\mathbb{R}^{d_x}$. A state-space model with parameter $\theta\in\mathbb{T}\subseteq\mathbb{R}^{d_\theta}$ specifies an initial distribution $\mu_\theta(dx_1)$ of the first state X_1, a Markov kernel $f_\theta(dx_{t+1}|x_t)$ for the transition of the latent process, a measurement distribution $g_\theta(dy_t|x_t)$, and a prior distribution $p(d\theta)$ on the parameter.

1.3.1 Computation of the H-score using SMC^2

The conditional predictive distributions $p(y_t|y_{1:t-1},\theta)$ appearing in (1.4) correspond to integrals over the latent states, i.e. $p(y_t|y_{1:t-1},\theta)=\int p(x_t|y_{1:t-1},\theta)\,g_\theta(y_t|x_t)\,dx_t$, which are in general intractable. Interchanging differentiation and integration under suitable regularity conditions yields the following results, which are similar to Fisher's and Louis' identities (Proposition 10.1.6 in Cappé et al., 2005), except that differentiation here is with respect to the observation instead of the parameter. We obtain for all $\theta\in\mathbb{T}$, all observed $y_{1:T}\in\mathbb{Y}^T$, all $k\in[\![1,d_y]\!]$, and all $t\in\in[\![1,T]\!]$,

$$\frac{\partial\log p(y_t|y_{1:t-1},\theta)}{\partial y_{t(k)}}=\mathbb{E}_t\left[\frac{\partial\log g_\theta(y_t|X_t)}{\partial y_{t(k)}}\middle|\theta\right], \tag{1.11}$$

$$\frac{\partial^2\log p(y_t|y_{1:t-1},\theta)}{\partial y_{t(k)}^2}+\left(\frac{\partial\log p(y_t|y_{1:t-1},\theta)}{\partial y_{t(k)}}\right)^2=\mathbb{E}_t\left[\frac{\partial^2\log g_\theta(y_t|X_t)}{\partial y_{t(k)}^2}+\left(\frac{\partial\log g_\theta(y_t|X_t)}{\partial y_{t(k)}}\right)^2\middle|\theta\right], \tag{1.12}$$

where the conditional expectations $\mathbb{E}_t$ are with respect to $X_t\sim p(dx_t|y_{1:t},\theta)$. Proofs of (1.11) and (1.12) under regularity assumptions are presented in Appendix B.4.2. By applying (1.11) and (1.12) to each term in (1.4) and using the tower property of conditional expectations, the expression of $\mathcal{H}_T(M)$ further simplifies to

$$\sum_{t=1}^{T}\sum_{k=1}^{d_y}\left(2\,\mathbb{E}_t\left[\frac{\partial^2\log g_\Theta(y_t|X_t)}{\partial y_{t(k)}^2}+\left(\frac{\partial\log g_\Theta(y_t|X_t)}{\partial y_{t(k)}}\right)^2\right]-\left(\mathbb{E}_t\left[\frac{\partial\log g_\Theta(y_t|X_t)}{\partial y_{t(k)}}\right]\right)^2\right) \tag{1.13}$$

where the expectations $\mathbb{E}_t$ are with respect to the joint posterior distributions of (Θ,X_t) given the observations $y_{1:t}$, whose densities are given by $p(\theta,x_t|y_{1:t})=p(\theta|y_{1:t})p(x_t|y_{1:t},\theta)$.

For many state-space models, the log-derivatives of the measurement density $g_\theta(y|x)$ can be evaluated at any point $(\theta,y,x)\in\mathbb{T}\times\mathbb{Y}\times\mathbb{X}$. Assuming that we can simulate

the transition kernel of the latent process, we can use SMC^2 (Fulop and Li, 2013; Chopin et al., 2013) to consistently estimate all the conditional expectations appearing in (1.13). At each time $t \in\in [\![1, T]\!]$, SMC^2 produces a set of weighted particles targeting the joint density $p(\theta, x_t|y_{1:t})$, which can be used to update the H-score. Further implementation guidelines are provided in Appendix A.2.

1.3.2 Consistency of the H-score for state-space models

We revisit the asymptotic consistency results of the H-score in the case of state-space models. The observations are no longer assumed to be i.i.d. and we consider two candidate models, M_1 and M_2. An additional difficulty in proving consistency of the H-score with dependent observations lies in the approximation of $\mathcal{H}_T(M_j)$ by a stationary analog, to which ergodic theorems will apply. As in the i.i.d. setting, we only give results for univariate continuous observations.

Theorem 2. *Assume $(Y_t)_{t\in\mathbb{N}^*}$ is ergodic and strongly stationary, so that we can artificially extend its set of indices to negative integers and consider the two-sided process $(Y_t)_{t\in\mathbb{Z}}$. Assume M_1 and M_2 both satisfy the following conditions, where models are omitted from the notation and probabilistic statements are $\mathbb{P}_\star$-almost sure:*

(a) For all $t \in \mathbb{N}^*$ and $y_{1:t} \in \mathbb{Y}^t$, $\theta \mapsto p(y_t|\theta)\, p(\theta|y_{1:t-1})$ is integrable on $\mathbb{T}$.

(b) For all $t \in \mathbb{N}^*$ and $\theta \in \mathbb{T}$, $y_t \mapsto p(y_t|\theta)$ is twice differentiable on $\mathbb{Y}$.

(c) For all $t \in \mathbb{N}^*$, there exist integrable functions $h_{1,t}$ and $h_{2,t}$ such that, for all $(y_{1:t}, \theta) \in \mathbb{Y}^t \times \mathbb{T}$, $|p(\theta|y_{1:t-1})\, \partial p(y_t|\theta)/\partial y_t| \leq h_{1,t}(\theta)$ and $|p(\theta|y_{1:t-1})\, \partial^2 p(y_t|\theta)/\partial {y_t}^2| \leq h_{2,t}(\theta)$.

(d) For all $t \in \mathbb{N}^*$ and $(y_{1:t}, \theta) \in \mathbb{Y}^t \times \mathbb{T}$, $x_t \mapsto p(x_t|y_{1:t-1}, \theta)\, g_\theta(y_t|x_t)$ is integrable on $\mathbb{X}$.

(e) For all $t \in \mathbb{N}^*$ and $(\theta, x_t) \in \mathbb{T} \times \mathbb{X}$, $y_t \mapsto g_\theta(y_t|x_t)$ is twice differentiable on $\mathbb{Y}$.

1. Bayesian model comparison with the Hyvärinen score

(f) There exist integrable functions $h_{3,t}$ and $h_{4,t}$ such that, for all $(y_{1:t}, \theta, x_t) \in \mathbb{Y}^t \times \mathbb{T} \times \mathbb{X}$, $|p(x_t|y_{1:t-1}, \theta)\partial g_\theta(y_t|x_t)/\partial y_t| \leq h_{3,t}(x_t)$ and $|p(x_t|y_{1:t-1}, \theta)\partial^2 g_\theta(y_t|x_t)/\partial y_t^2| \leq h_{4,t}(x_t)$.

(g) For all $t \in \mathbb{N}^*$, there exists $\theta^\star \in \mathbb{T}$ such that, if $\Theta_t \sim p(d\theta|Y_{1:t})$ for all $t \in \mathbb{N}^*$, then $\Theta_t \xrightarrow[t\to+\infty]{\mathcal{D}} \theta^\star$.

(h) There exist a constant $L > 0$ and a neighborhood $\mathcal{U}_{\theta^\star}$ of $\theta^\star$ such that, for all $t \in \mathbb{N}^*$, $\theta \mapsto \mathcal{H}(Y_t, p(dy_t|Y_{1:t-1}, \theta))$ and $\theta \mapsto \partial \log p(Y_t|Y_{1:t-1}, \theta)/\partial y_t$ are L-Lipschitz functions.

(i) There exist $\alpha_1 > 1$ and $\alpha_2 > 1$ such that $\sup_{t\in\mathbb{N}^*} \mathbb{E}\left[\,|\mathcal{H}(Y_t, p(dy_t|Y_{1:t-1}, \Theta_t))|^{\alpha_1} \,|\, Y_{1:t}\right] < +\infty$ and $\sup_{t\in\mathbb{N}^*} \mathbb{E}\left[(\partial \log p(Y_t|Y_{1:t-1}, \Theta_t)/\partial y_t)^{2\alpha_2} \,|\, Y_{1:t}\right] < +\infty$, where the conditional expectations are with respect to the posterior distribution $\Theta_t \sim p(d\theta|Y_{1:t})$.

(j) There exists a dominating probability measure η on $\mathbb{X}$ such that the transition kernel $f_{\theta^\star}(dx_{t+1}|x_t)$ has density $\nu_{\theta^\star}(x_{t+1}|x_t) = (df_{\theta^\star}(\cdot|x_t)/d\eta)(x_{t+1})$ with respect to η.

(k) There exist positive constants σ^- and σ^+ such that, for all $(x_t, x_{t+1}) \in \mathbb{X} \times \mathbb{X}$, the transition density $\nu_{\theta^\star}(x_{t+1}|x_t)$ satisfies $0 < \sigma^- < \nu_{\theta^\star}(x_{t+1}|x_t) < \sigma^+ < +\infty$.

(l) For all $y_t \in \mathbb{Y}$, the integral $\int_{\mathbb{X}} g_{\theta^\star}(y_t, x_t)\,\eta(dx_t)$ is bounded away from 0 and $+\infty$.

(m) $b = \sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} \left|\frac{\partial^2 \log g_{\theta^\star}(y|x)}{\partial y^2} + \left(\frac{\partial \log g_{\theta^\star}(y|x)}{\partial y}\right)^2\right| < +\infty$ *and* $c = \sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} \left|\frac{\partial \log g_{\theta^\star}(y|x)}{\partial y}\right| < +\infty$.

(n) $\sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} g_{\theta^\star}(y|x) < +\infty$ *and* $\mathbb{E}_\star\left[\left|\log\left(\int_{\mathbb{X}} g_{\theta^\star}(Y_1|x)\nu_{\theta^\star}(dx)\right)\right|\right] < +\infty$.

(o) The conditional density $y_1 \mapsto p_\star(y_1|Y_{-\infty:0})$ of Y_1 given $(Y_t)_{t\leq 0}$ is well-defined and twice differentiable, and $\mathbb{E}_\star\left[\left|\mathcal{H}(Y_1, p_\star(dy_1|Y_{-\infty:0}))\right|\right] < +\infty$.

If these conditions are met, we may define, for each $j \in \{1, 2\}$, the quantity

$$\mathrm{D}_{\mathcal{H}}(p_\star, M_j) = \mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p_j(dy_1|Y_{-\infty:0}, \theta_j^\star)\right)\right] - \mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p_\star(dy_1|Y_{-\infty:0})\right)\right] \tag{1.14}$$

where $p_j(y_1|Y_{-\infty:0}, \theta_j^\star)$ *is the provably well-defined conditional density of* Y_1 *given* $(Y_t)_{t\leq 0}$ *under* M_j *and* $\theta_j^\star$*. Under these conditions, we have*

$$\frac{1}{T}\Big(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\Big) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \mathrm{D}_{\mathcal{H}}(p_\star, M_2) - \mathrm{D}_{\mathcal{H}}(p_\star, M_1)\,. \tag{1.15}$$

If $p_\star(y_1|Y_{-\infty:0})\,\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)/\partial y_1 \xrightarrow[|y_1|\to+\infty]{\mathbb{P}_\star - a.s.} 0$*, then we have* $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) \geq 0$*, with* $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) = 0$ *if and only if* $p_j(y_1|Y_{-\infty:0}, \theta_j^\star) = p_\star(y_1|Y_{-\infty:0})$*,* $\mathbb{P}_\star$*-almost surely.*

Conditions (a) to (c) ensure the validity of (1.5); (d) to (f) ensure the validity of (1.11) and (1.12); (g) assumes the concentration of the posterior to a point mass; (h) to (i) yield suitable convergence of posterior moments; (j) to (l) ensure the *forgetting propriety* of the latent Markov chain and the H-score; (m) to (n) relate to the well-definiteness of the conditional density $p_j(y_1|Y_{-\infty:0}, \theta_j^\star)$; finally, (o) and the last boundary condition ensure that the H-score is strictly proper and well-defined for $p_\star$. Further discussion on these conditions and detailed proofs are provided in Appendix B.5.

For state-space models, posterior concentration results have been derived in specific cases (e.g. Lijoi, Prünster and Walker, 2007; De Gunst and Shcherbakova, 2008; Shalizi, 2009; Gassiat and Rousseau, 2014; Douc et al., 2014; Douc, Olsson and Roueff, 2019, and references therein). However, to the best of our knowledge, general results on posterior concentration for misspecified state-space models have yet to be established. As a consequence, our proof of Theorem 2 uses posterior concentration as a working assumption. Our numerical examples suggest that concentration of posterior distributions can be observed in practice, even for complex state-space models, as hinted by the posterior density plots shown in Section 3.7 of Chapter 3. Further research on Bayesian asymptotics in state-space models might provide more theoretical understanding of such phenomena.

1.3.3 Illustration with Lévy-driven stochastic volatility models

In this simulation study we illustrate the consistency of the H-score in nonlinear, non-Gaussian state-space models with continuous observations. A simpler example with linear Gaussian state-space and ARMA models can be found in Appendix B.2. Here we consider Lévy-driven stochastic volatility models (Barndorff-Nielsen and Shephard, 2001, 2002). These models feature intractable transition kernels that can only be simulated, and describe the joint evolution of the log-returns Y_t and the instantaneous volatility V_t of a financial asset. The former is modeled as a continuous time process driven by a Brownian motion, while the latter is modeled as a Lévy process. Given a triplet of parameters (λ, ξ, ω), we can generate random variables $(V_t, Z_t)_{t\geq 1}$ recursively as:

$$\left.\begin{array}{l} k \sim \text{Poisson}\left(\lambda\xi^2/\omega^2\right); \quad C_{1:k} \overset{\text{i.i.d.}}{\sim} \text{Unif}(t-1, t); \quad E_{1:k} \overset{\text{i.i.d.}}{\sim} \text{Exp}\left(\xi/\omega^2\right); \\ Z_0 \sim \text{Gamma}\left(\xi^2/\omega^2, \xi/\omega^2\right); \quad Z_t = e^{-\lambda} Z_{t-1} + \sum_{j=1}^{k} e^{-\lambda(t-C_j)} E_j; \\ V_t = \frac{1}{\lambda}\left(Z_{t-1} - Z_t + \sum_{j=1}^{k} E_j\right). \end{array}\right\} \tag{1.16}$$

The first model (M_1) describes the volatility as driven by a *single factor*, expressed in terms of a finite rate Poisson process.

M_1: (V_t, Z_t) *from* (1.16) *given* (λ, ξ, ω); $\quad X_t = (V_t, Z_t)$; $\quad Y_t \,|\, X_t \sim \mathcal{N}\left(\mu + \beta V_t, V_t\right)$; $\quad$ *with independent priors* $\lambda \sim \text{Exp}(1)$; $\quad \xi,\, \omega^2 \overset{\text{i.i.d.}}{\sim} \text{Exp}\left(1/5\right)$; $\quad \mu,\, \beta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 10)$.

The second model (M_2) introduces an additional independent component to drive the behavior of the volatility, leading to the *multi-factor* model below.

M_2: $(V_{i,t}, Z_{i,t})$ *from* (1.16) *independently for* $i \in \{1, 2\}$ *given* $(\lambda_i, \xi \mathrm{w}_i, \omega \mathrm{w}_i)$, *with* $(\mathrm{w}_1, \mathrm{w}_2) = (\mathrm{w}, 1-\mathrm{w})$; $\quad X_t = (V_{1,t}, V_{2,t}, Z_{1,t}, Z_{2,t})$; $\quad Y_t \,|\, X_t \sim \mathcal{N}\left(\mu + \beta V_t, V_t\right)$ *where* $V_t = V_{1,t} + V_{2,t}$; $\quad$ *with independent priors* $\quad \lambda_1 \sim \text{Exp}(1)$; $\quad \lambda_2 - \lambda_1 \sim \text{Exp}\left(1/2\right)$; $\quad \mathrm{w} \sim \text{Unif}(0, 1)$; $\xi,\, \omega^2 \overset{\text{i.i.d.}}{\sim} \text{Exp}\left(1/5\right)$; $\quad \mu,\, \beta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 10)$.

1. Bayesian model comparison with the Hyvärinen score

For model M_1, we can prove that there exist values of the parameter $\theta = (\lambda, \xi, \omega, \mu, \beta)$ such that $\mathbb{E}[\,|\partial \log g_\theta(y_1|X_1)/\partial y_1|\,] = +\infty$, which prevents the use of (1.11)-(1.12) to estimate the H-score of model M_1. When (1.11)-(1.12) do not hold, we can directly estimate the partial derivatives of $\tilde{y}_t \mapsto p(\tilde{y}_t|y_{1:t-1}, \theta)$ at the observed y_t, by using approximate draws from the conditional predictive distribution $p(dy_t|y_{1:t-1}, \theta)$. Approximate draws from $p(dy_t|y_{1:t-1}, \theta)$ can be obtained from a run of SMC^2, as long as one can sample from the measurement distribution $g_\theta(dy_t|x_t)$. For a chosen bandwidth $h > 0$ (e.g. Hardle, Marron and Wand, 1990, and references in Section 1.11 of Tsybakov, 2009) and a twice continuously differentiable kernel K integrating to 1, e.g. a standard Gaussian kernel $K(u) = (2\pi)^{-1/2} \exp(-u^2/2)$, we can use n draws $\tilde{y}_t^{(1)}, \ldots, \tilde{y}_t^{(n)}$ from $p(dy_t|y_{1:t-1}, \theta)$ to consistently estimate $p(y_t|y_{1:t-1}, \theta)$ by the kernel density estimator $\widehat{p}(y_t|y_{1:t-1}, \theta) = (nh)^{-1} \sum_{i=1}^n K((y_t - \tilde{y}_t^{(i)})/h)$. This kernel density estimator is twice differentiable with respect to y_t, hence we can respectively use $\partial\, \widehat{p}(y_t|y_{1:t-1}, \theta)/\partial y_{t(k)}$ and $\partial^2\, \widehat{p}(y_t|y_{1:t-1}, \theta)/\partial y_{t(k)}^2$ as consistent estimators of the partial derivatives $\partial p(y_t|y_{1:t-1}, \theta)/\partial y_{t(k)}$ and $\partial^2 p(y_t|y_{1:t-1}, \theta)/\partial y_{t(k)}^2$, as $n \to +\infty$ and $h \to 0$ at an appropriate rate (e.g. Bhattacharya, 1967).

We simulate $T = 1000$ observations from a single-factor Lévy-driven stochastic volatility model with parameters $\lambda = 0.01$, $\xi = 0.5$, $\omega^2 = 0.0625$, $\mu = 0$, and $\beta = 0$, following the simulations of Barndorff-Nielsen and Shephard (2002). The H-factor of M_1 against M_2 is computed for 15 replications of SMC^2, using $N_\theta = 1024$ particles in θ, and an adaptive number of particles in x starting at $N_x = 128$. The kernel density estimation is performed with a Gaussian kernel, using $n = 1024$ predictive draws and $h = 0.1$. The estimated log-Bayes factor and H-factor of M_1 against M_2 are plotted in Figure 1.3. Here the models are nested and well-specified, but their dimensions differ. We see that both criteria correctly select the smaller model M_1. As mentioned in Section 1.2.1, the estimated H-factor tends to have a larger relative variance than the estimated log-Bayes factor, especially in the presence of extreme observations (e.g. at times 454 and 656), and might thus call for a larger number of particles.

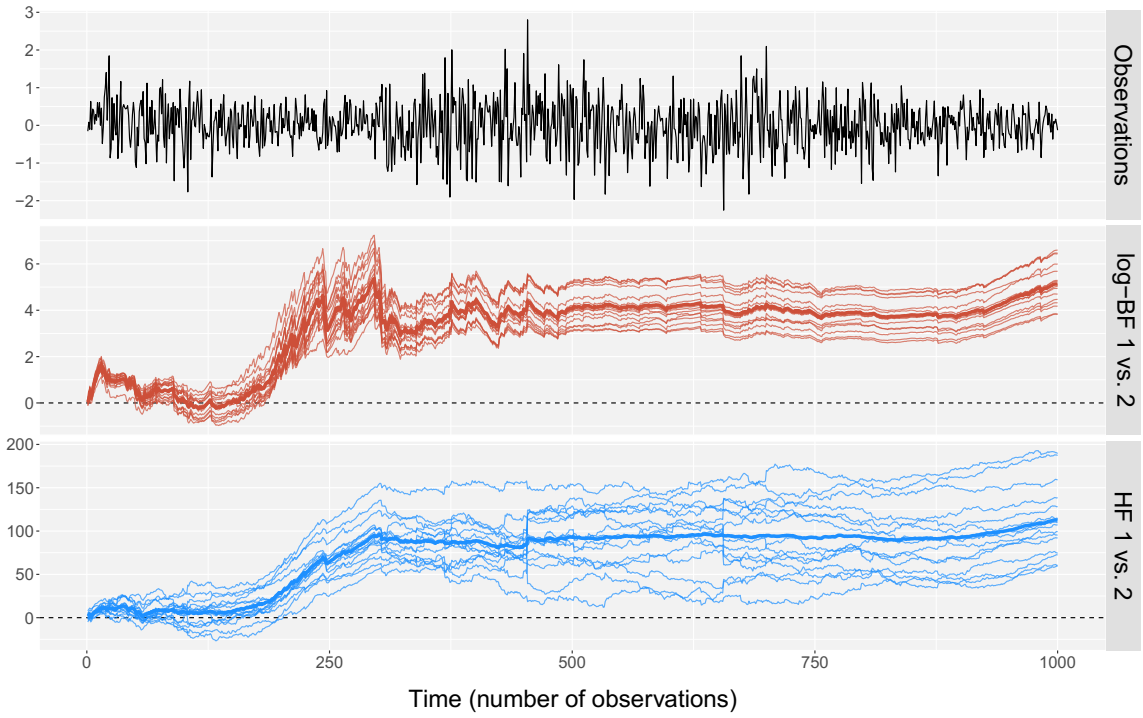

Figure 1.3. *Top panel: log-returns simulated from model M_1 with parameters $\lambda = 0.01$, $\xi = 0.5$, $\omega^2 = 0.0625$, $\mu = 0$, and $\beta = 0$. Middle and bottom panels: estimated log-Bayes factor (log-BF) and H-factor (HF) of M_1 against M_2, computed for 15 replications (thin solid lines), along with the average scores across replications (thick solid lines). In each plot, the variability within each factor is due to Monte Carlo error. See Section 1.3.3.*

1.4 H-score for discrete observations

Motivated by an application in population dynamics (Section 1.4.2), we propose an extension of the H-score to discrete observations. We assume that each observation $y = (y_{(1)}, ..., y_{(d_y)})^\top$ takes finite values (i.e. $\|y\| < +\infty$) in some discrete space $\mathbb{Y} = [\![a_1, b_1]\!] \times ... \times [\![a_{d_y}, b_{d_y}]\!]$, where $[\![a_k, b_k]\!] = [a_k, b_k] \cap \mathbb{Z}$ and $a_k, b_k \in \mathbb{Z} \cup \{-\infty, +\infty\}$, with $a_k < b_k$ for all $k \in [\![1, d_y]\!]$. For ease of exposition, assume for now that $b_k - a_k \geq 3$ for all $k \in [\![1, d_y]\!]$.

1.4.1 Extension of the H-score to discrete observations

Let e_k denote the canonical vector of $\mathbb{Z}^{d_y}$ that has all coordinates equal to 0 except for its k-th coordinate that equals 1. For all $y \in \mathbb{Y}$, all non-negative functions p on $\mathbb{Y}$, and all

$k \in [\![1, d_y]\!]$, we define $\partial_k p(y) = (p(y + e_k) - p(y - e_k))/2$ and $\partial_k \log p(y) = \partial_k p(y)/p(y)$. We define the score function

$$\mathcal{H}^D(y, p) = \sum_{k=1}^{d_y} \mathcal{H}_k^D(y, p) \tag{1.17}$$

where $\mathcal{H}_k^D(y, p) = 2\,\partial_k\left(\partial_k \log p(y)\right) + \left(\partial_k \log p(y)\right)^2$ if $a_k + 2 \leq y_{(k)} \leq b_k - 2$. If a_k or b_k is finite, we may define $\mathcal{H}_k^D(y, p)$ at the relevant boundaries respectively as $\partial_k \log p(y + e_k)$, $\partial_k \log p(y + e_k) + (\partial_k \log p(y))^2$, $-\partial_k \log p(y - e_k) + (\partial_k \log p(y))^2$, and $-\partial_k \log p(y - e_k)$ for $y \in \{a_k, a_k+1, b_k-1, b_k\}$. When the context is clear, we extend the definition of the H-score to discrete observations by overloading its notation with $\mathcal{H}_T(M) = \sum_{t=1}^T \mathcal{H}^D(y_t, p_M(dy_t|y_{1:t-1}))$.

The expression of $\mathcal{H}^D$ can be regarded as a discrete analog of the Hyvärinen score $\mathcal{H}$, where the partial derivatives are replaced by central finite differences. The seemingly arbitrary discretization turns out to yield a scoring rule that is actually proper, which can be proved by using a characterization of proper scores for discrete observations as super-gradients of concave entropy functions (McCarthy, 1956; Hendrickson and Buehler, 1971; Dawid et al., 2012). The construction of $\mathcal{H}^D$ and the proof of its propriety are detailed in Appendix B.3.

If $b_k = a_k + 1$ (e.g. for binary data) or $b_k = a_k + 2$, we could still define $\mathcal{H}_k^D$ by ignoring the cases $y_{(k)} = a_k + 1$, or $y_{(k)} = b_k - 1$, or both. Alternatively, we could use forward differences. All these definitions lead to scores that meet the requirements of being insensitive to prior vagueness, while being proper and local. Deciding which one to use is then a matter of further considerations, left for future research.

1.4.2 Diffusion models for population dynamics of red kangaroos

We illustrate the H-score for discrete observations by comparing three nonlinear non-Gaussian state-space models, describing the dynamics of a population of red kangaroos (*Macropus rufus*) in New South Wales, Australia. These models were compared in Knape and de Valpine (2012) using Bayes factors, although the authors acknowledged the undesirable sensitivity of

their results to their choice of prior distributions. The data (Caughley, Shepherd and Short, 1987) is a time series of 41 bi-variate observations $(Y_{1,t}, Y_{2,t})$, formed by double transect counts of red kangaroos, measured between 1973 and 1984 (see Figure 1.4). The small number of observations calls for a criterion that is principled for finite samples, contrarily to e.g. the Bayesian Information Criterion. The models are nested and will be referred to as M_1, M_2, and M_3, by decreasing order of complexity. The largest model (M_1) is a logistic diffusion model. Simpler versions include an exponential growth model (M_2) and a random-walk model (M_3). In these models a latent population size (X_t) follows a stochastic differential equation (see further motivation in Dennis and Costantino, 1988; Knape and de Valpine, 2012). Each model is specified below, where $(W_t)_{t\geq 0}$ denotes a standard Brownian motion.

$$M_1\text{: } X_1 \sim \mathrm{LN}(0,5)\,; \quad dX_t/X_t = (\sigma^2/2 + r - bX_t)\,dt + \sigma dW_t\,;$$
$$Y_{1,t}, Y_{2,t} \,|\, X_t, \tau \overset{\text{i.i.d.}}{\sim} \mathrm{NB}(X_t,\, X_t + \tau X_t^2)\,;$$
$$\textit{with independent priors; } \sigma,\, \tau,\, b \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,10),\; r \sim \mathrm{Unif}(-10,10).$$

$$M_2\text{: } \textit{same as } M_1 \textit{ with } b = 0\textit{; with independent priors } \sigma,\, \tau \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,10),\; r \sim \mathrm{Unif}(-10,10).$$

$$M_3\text{: } \textit{same as } M_1 \textit{ with } b = 0 \textit{ and } r = 0\textit{; with independent priors } \sigma,\, \tau \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,10).$$

We perform 5 runs of SMC^2 to estimate the log-score and H-score of each model, with an adaptive number N_x of latent particles. We use $N_\theta = 16384$ particles in θ, and $N_x = 32$ initial particles in x. For model M_1, we simulate the latent process using the Euler-Maruyama method with discretization step $\Delta_t = 0.001$. The estimated log-scores and H-scores are shown in Figure 1.4. For better readability, the log-score is rescaled by the number of observations. Using the H-scores would lead to selecting model M_3, similarly to Knape and de Valpine (2012) who use log-scores. Their conclusion was mitigated by the sensitivity of the evidence to the choice of vague priors: for instance, changing the prior on r in model M_2 to $\mathrm{Unif}(-100, 100)$ effectively divides the evidence of M_2 by a factor 10. On the other hand, we have found the

impact of that change of prior on the H-score to be indistinguishable from the Monte Carlo variation across runs. Using our R implementation and choices of algorithmic parameters, one SMC or SMC^2 replication took about a few minutes for each i.i.d. Normal models with 1000 observations (Section 1.2.3), about an hour for each kangaroo population model with 41 observations (Section 1.4.2), and about five hours for each stochastic volatility model with 1000 observations (Section 1.3.3). In all cases, the Monte Carlo error can be reduced by increasing the number of particles, albeit at the expense of an increased computational cost.

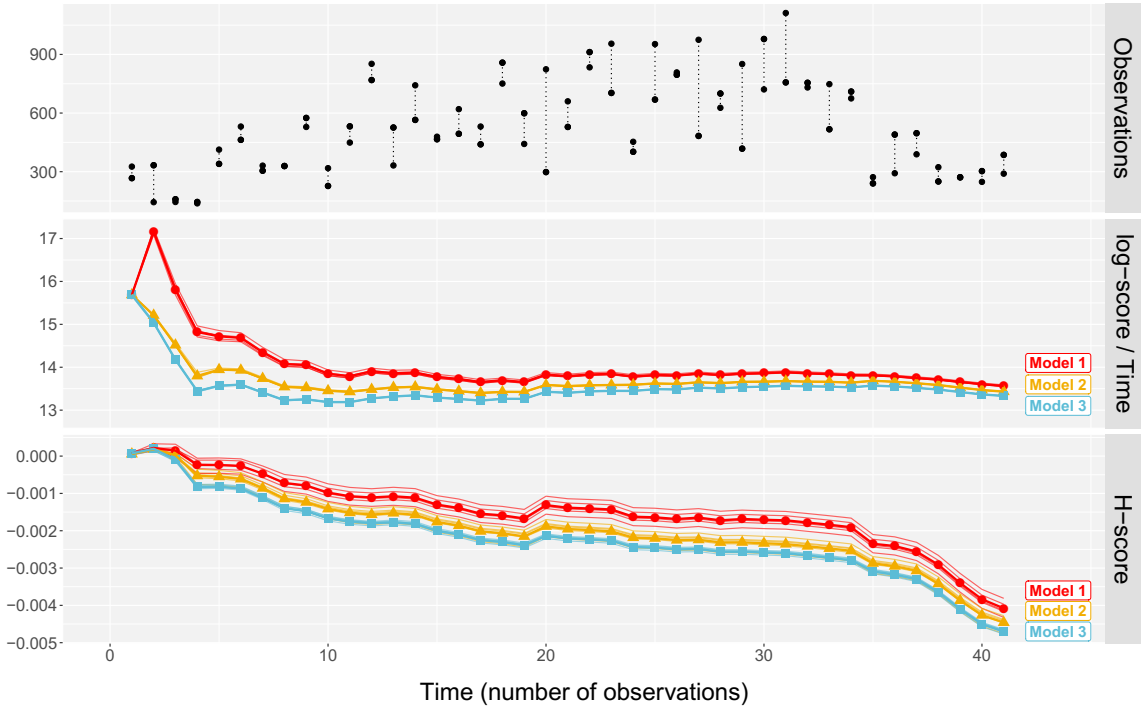

Figure 1.4. *Top panel: double transect counts of red kangaroos. Middle and bottom panels: estimated log-scores and H-scores of M_1 (circles), M_2 (triangles), and M_3 (squares), for 5 replications (thin solid lines), along with the average scores across replications (thick lines with shapes). The log-scores are rescaled by the number of observations for better readability. The variability within each model is due to Monte Carlo error. See Section 1.4.2.*

1.5 Discussion

The H-factor constitutes a competitive alternative to the Bayes factor. It is justified non-asymptotically since it relies on assessing predictive performances using a proper local scoring rule, and it is robust to the arbitrary vagueness of prior distributions. It can be applied to a large variety of models — including nonlinear non-Gaussian state-space models — and it can be estimated sequentially with SMC or SMC^2, at a cost comparable to that of the Bayes factor. However, the H-score puts additional smoothness restrictions on the models, e.g. the twice differentiability of their predictive distributions with respect to the observations (see Dawid and Musio, 2015, and its rejoinder). Thus there are models for which the Bayes factor is applicable but not the H-factor. We have also discussed in Section 1.2.3 a case where the two criteria disagree, even asymptotically, contrarily to e.g. partial and intrinsic Bayes factors (Santis and Spezzaferri, 1999) that asymptotically agree with the Bayes factor. To deal with vague or improper priors, other alternatives to the log-evidence include Bayesian cross-validation criteria, e.g. $\sum_{t=1}^{T} \log p(y_t|y_{-t})$, where $y_{-t} = \{y_s : 1 \leq s \leq T \text{ and } s \neq t\}$. Such criteria would be applicable under weaker smoothness assumptions on the predictive densities, while still being robust to arbitrary vagueness of prior distributions. Efficient computation of these criteria is challenging, and can be envisioned for i.i.d. models using MCMC (Alqallaf and Gustafson, 2001), SMC (Bornn, Doucet and Gottardo, 2010), or more recent work on unbiased MCMC (Jacob, O'Leary and Atchadé, 2017) applied to path sampling (Rischard, Jacob and Pillai, 2018); the case of state-space models would be more challenging, due to standard difficulties arising when splitting time series. Another approach suggested in Kamary et al. (2014) is to cast model selection as a mixture estimation problem, which also raises questions in the case of time series.

When using either the log-Bayes factor or the H-factor for model selection, decision theory stipulates that the choice should be dictated by the sign of these differences of scores. In

many practical settings, one only has access to SMC estimators of these differences, and a crucial question is to determine whether the estimated sign is correct or wrongly flipped due to the inherent Monte Carlo variation of the estimation method. Such uncertainty quantification relates to the task of building confidence intervals for the differences of scores and will be investigated in Chapter 2.

The sequential form of the H-score is problematic when observations are not naturally ordered, leading to different values of the H-score for different orderings. This issue is mitigated by the following facts: if the sample is large enough, any ordering of the data would yield similar H-scores. For smaller samples, one could average the H-score over random permutations of the data. In that case, quantifying and controlling the extra variability induced by these permutations would deserve further investigation.

For continuous observations and non-nested parametric models satisfying strong regularity assumptions, we have proved that the H-score leads to consistent model selection. The asymptotic behavior of the H-factor is determined by how close the candidate models are from the data-generating process, where closeness is quantified by the relative Fisher information divergence associated with the H-score, in contrast to the Kullback–Leibler divergence associated with the Bayes factor. Although our numerical experiments indicate that the results might hold in more generality, our proofs rely on strong assumptions. For non-nested models, these assumptions include suitable concentration of the posterior distribution as the number of observations grow. When dealing with well-specified nested models, the limits in (1.6) and (1.15) become meaningless, and higher order results are needed, e.g. the asymptotic Normality of the posterior. To the best of our knowledge, such results on Bayesian asymptotics are still elusive for general state-space models and will be the object of Chapter 3.

“ Rien n’est si fou que de mettre son salut dans l’incertitude; mais rien n’est si naturel [...] ”

— Madame de Sévigné, *Lettres à sa fille*

2

Uncertainty quantification for model comparison

Contents

The work in this chapter has been done in collaboration with Pierre E. Jacob (*Department of Statistics*, Harvard University). It corresponds to an upcoming paper at the time of writing.

2.1 Introduction

Given observations $y_{1:T} \in \mathbb{Y}^T$, the task of choosing between two models M_1 and M_2 can be regarded as a decision theoretic problem, which consists in selecting the model minimizing an appropriately chosen loss function (e.g. Bernardo and Smith, 2000; Robert, 2007). In Chapter 1, we have discussed the relative benefits of using the log-evidence or the H-score when choosing between Bayesian models, and we will keep focusing on these two criteria in this chapter. For either criterion, deciding which model to select would be dictated by the sign of the difference $\mathrm{LBF}_T^{1|2} = (-\log p_2(y_{1:T})) - (-\log p_1(y_{1:T}))$ or $\mathrm{HF}_T^{1|2} = \mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)$ between the models' respective scores. However, this idealized decision rule cannot be used in practice: as pointed out in the previous chapter, the exact scores are generally not available and can only be estimated, e.g. via SMC methods, so that practical decisions are actually based on estimated values $\hat{\mathrm{LBF}}_T^{1|2}$ and $\hat{\mathrm{HF}}_T^{1|2}$ in place of the exact unknown differences. The use of numerical approximations raises legitimate concerns about whether the estimated signs of those differences are reliable or wrongly misled by the Monte Carlo error of SMC samplers. This motivates the important task of quantifying the uncertainty pertaining to the estimation of such model selection criteria. In the numerical illustrations from Sections 1.3.3 and 1.4.2 of Chapter 1, the uncertainty of the model selection procedure was heuristically assessed by performing several independent replications of SMC samplers. A more rigorous approach to quantify this uncertainty would consist in constructing confidence intervals for the exact unknown differences in scores $\mathrm{LBF}_T^{1|2}$ and $\mathrm{HF}_T^{1|2}$. The construction of such valid confidence intervals will be at the heart of this chapter. Unless stated otherwise, all the probabilistic statements in this chapter will be with respect to the auxiliary random

numbers generated during the course of SMC or MCMC algorithms, conditional on some realized observations $y_{1:T} \in \mathbb{Y}^T$ treated as fixed.

In Section 2.2, we start by proposing different methods to construct approximate confidence intervals for log-evidences using SMC samplers. The first approach (Section 2.2.1) uses the work of Lee and Whiteley (2018) to construct confidence intervals that are asymptotically valid as the number of particles grows to infinity. Despite the existence of previous works on parallel implementations of SMC (e.g. Vergé, Dubarry, Del Moral and Moulines, 2015; Murray et al., 2016; Whiteley, Lee and Heine, 2016), such an asymptotic regime in the number of particles is not appropriate when computational time is budgeted, nor does it allow to fully take advantage of modern computing architectures often made of many independent processors. This motivates another approach (Section 2.2.2) based on independent runs of SMC to construct confidence intervals that are asymptotically valid as the number of processors grows and therefore much more amenable to parallel computing. However, that approach crucially relies on the non-trivial property that SMC estimators of evidences are unbiased, making them not directly applicable to H-scores, whose SMC estimators are generally biased. The more challenging construction of confidence intervals for H-scores will be studied in Section 2.3. Using the recent work of Jacob et al. (2017) on debiasing MCMC estimators, we present one practical construction (Section 2.3.2) that stems from the use of SMC samplers as proposal distributions within coupled particle independent Metropolis-Hastings algorithms (Middleton, Deligiannidis, Doucet and Jacob, 2019). The proposed method is pleasingly parallel, and produces confidence intervals whose validity is asymptotic in the number of independent processors. We conclude by discussing possible avenues of future research in Section 2.4. Throughout this chapter, we illustrate the different constructions by using the diffusion models M_2 and M_3 from Section 1.4.2 of Chapter 1 as a recurring testbed.

2.2 Confidence intervals for the log-evidence

The work of Lee and Whiteley (2018) allows us to estimate the variance of SMC-based estimators. These approximate variances can be obtained from the same run of SMC used to produce the estimators in the first place, and they allow the construction of approximate confidence intervals that are asymptotically valid as the number of particles grows to infinity, as explored in Section 2.2.1. However, such an asymptotic regime in the number of particles is not appropriate when computational time is budgeted, and is restricted by the limited amount of computational resources on any given processor. In Section 2.2.2, we present an alternative method to construct approximate confidence intervals for the log-evidence by performing independent runs of SMC on different processors. This alternative method yields confidence intervals that are asymptotically valid as the number of processors increases, and is therefore more amenable to parallel computing and budget constraints.

2.2.1 Asymptotic validity in the number of particles

Consider a generic sequence of T probability distributions of interest, identified with their densities $(\eta_t)_{t\in[\![1,T]\!]}$ on some space $\Xi \subseteq \mathbb{R}^d$. Let $(\gamma_t)_{t\in[\![1,T]\!]}$ denote the corresponding unnormalized densities, so that for all $(t,\xi) \in [\![1,T]\!] \times \Xi$, we have $\eta_t(\xi) = \gamma_t(\xi)/Z_t$ with $Z_t = \int_\Xi \gamma_t(\xi)d\xi$ known as the *normalizing constant* of γ_t. It is customary to introduce an initial distribution η_0 that is easy to sample from, with $\gamma_0 = \eta_0$ and $Z_0 = 1$. By alternating between resampling, moving, and re-weighting operations, an SMC algorithm (see Appendix A.1 and references therein) sequentially produces T systems of N weighted particles $(w_t^{(i)}, \xi_t^{(i)})_{(t,i)\in[\![1,T]\!]\times[\![1,N]\!]}$ such that, for every horizon $t \in [\![1,T]\!]$, the corresponding empirical distributions

$$\hat{\gamma}_{t,N} = \sum_{i=1}^{N} w_t^{(i)} \delta_{\xi_t^{(i)}} \qquad \text{and} \qquad \hat{\eta}_{t,N} = \sum_{i=1}^{N} \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}} \delta_{\xi_t^{(i)}}$$

respectively approximate the target distributions γ_t and η_t. One can then naturally use

$$\hat{\gamma}_{t,N}(\varphi) = \sum_{i=1}^{N} w_t^{(i)} \varphi(\xi_t^{(i)}) \qquad \text{and} \qquad \hat{\eta}_{t,N}(\varphi) = \sum_{i=1}^{N} \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}} \varphi(\xi_t^{(i)}) \tag{2.1}$$

as respective estimators of the expectations

$$\gamma_t(\varphi) = \int_{\Xi} \varphi(\xi)\gamma_t(\xi)d\xi \qquad \text{and} \qquad \eta_t(\varphi) = \int_{\Xi} \varphi(\xi)\eta_t(\xi)d\xi \tag{2.2}$$

for any suitably integrable function φ on Ξ. It follows that the normalizing constants $Z_t = \gamma_t(1)$ can be estimated with $\hat{Z}_{t,N} = \hat{\gamma}_{t,N}(1)$ for all horizons $t \in [\![1, T]\!]$. Of particular interest to us is the Bayesian setting where $\gamma_t(\theta) = p(\theta)p(y_{1:t}|\theta)$, for a given prior density $\theta \mapsto p(\theta)$, a sequence of observations $y_{1:T} \in \mathbb{Y}^T \subseteq (\mathbb{R}^{d_y})^T$, and a likelihood function $\theta \mapsto p(y_{1:t}|\theta)$, in which case $(\eta_t)_{t\in[\![1,T]\!]}$ corresponds to the sequence of successive posterior densities $\eta_t(\theta) = p(\theta|y_{1:t})$ for all $t \in [\![1, T]\!]$, and the normalizing constant $Z_t = p(y_{1:t})$ corresponds to the sought-after evidence.

The behavior as $N \to +\infty$ of the SMC estimators $\hat{\gamma}_{t,N}(\varphi)$ and $\hat{\eta}_{t,N}(\varphi)$ in (2.1) has been extensively studied, notably by Del Moral (2004, Section 9.4), Chopin et al. (2004, Theorem 1), and Del Moral et al. (2006, Proposition 2). Although the latter is generally biased for $\eta_t(\varphi)$ while the former unbiasedly estimates $\gamma_t(\varphi)$, both estimators $\hat{\gamma}_{t,N}(\varphi)$ and $\hat{\eta}_{t,N}(\varphi)$ are typically consistent and satisfy a version of the central limit theorem centered at their exact counterparts $\gamma_t(\varphi)$ and $\eta_t(\varphi)$. In particular, under appropriate regularity assumptions, the SMC estimator of the evidence satisfies

$$\sqrt{N}\left(\frac{\hat{Z}_{t,N}}{Z_t} - 1\right) \xrightarrow[N\to+\infty]{\mathcal{D}} \mathcal{N}\left(0, \frac{\sigma^2_{\mathrm{SMC},\hat{Z}_t}}{Z_t^2}\right) \tag{2.3}$$

for some finite relative variance $\sigma^2_{\mathrm{SMC},\hat{Z}_t} \in (0, +\infty)$. Since $(\partial \log(u)/\partial u)^2 = (1/u)^2 = 1$ at $u = 1$, applying the delta method to (2.3) leads to

$$\sqrt{N}\left(\log \hat{Z}_{t,N} - \log Z_t\right) \xrightarrow[N\to+\infty]{\mathcal{D}} \mathcal{N}\left(0, \frac{\sigma^2_{\mathrm{SMC},\hat{Z}_t}}{Z_t^2}\right). \tag{2.4}$$

2. Uncertainty quantification for model comparison

The recent work of Lee and Whiteley (2018) provides a novel way to consistently estimate the asymptotic relative variance $\sigma^2_{\mathrm{SMC},\hat{Z}_t}/Z_t^2$ in (2.4). Although their initial motivation was to estimate the variance of $\hat{\gamma}_{t,N}(\varphi)$ and $\hat{\eta}_{t,N}(\varphi)$ in the context of particle filters (see Appendix A.2.1 and references therein), their method readily applies to more general SMC samplers. From a non-adaptive SMC run with deliberate rejuvenation (i.e. step (iv) of Algorithm 1 in Appendix A.1.1) after each re-weighting step, we can construct the estimator

$$\widehat{V}_{t,N} \;=\; 1 - \left(\frac{N}{N-1}\right)^{t+1} + \left(\frac{N}{N-1}\right)^{t+1} \frac{1}{N^2} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \mathbb{1}_{\{E_t^j = i\}} \right)^2 \tag{2.5}$$

where $E_t^j \in [\![1, N]\!]$ denotes the index of the ancestral particle from horizon 0 whose descendants at horizon t includes particle ξ_t^j. In other words, starting from the leaf node ξ_t^j at horizon t and tracing back the respective ancestors all the way up the lineage would lead to particle $\xi_0^{E_t^j}$ as the root. Under mild assumptions (Lee and Whiteley, 2018, Theorem 1), the estimator $\widehat{V}_{t,N}$ satisfies $N\widehat{V}_{t,N} \xrightarrow[N\to+\infty]{} \sigma^2_{\mathrm{SMC},\hat{Z}_t}/Z_t^2$ in probability, so that

$$\widehat{V}_{t,N}^{-1/2} \left(\log \hat{Z}_{t,N} - \log Z_t \right) \quad \xrightarrow[N\to+\infty]{\mathcal{D}} \quad \mathcal{N}(0,1) \tag{2.6}$$

and a possible $100(1-\alpha)\%$-confidence interval for $\log Z_t$ is given by

$$\left[\; \log \hat{Z}_{t,N} \;\; \pm \;\; q_{1-\alpha/2}^{\mathcal{N}(0,1)} \sqrt{\widehat{V}_{t,N}} \; \right] \tag{2.7}$$

where $q_{1-\alpha/2}^{\mathcal{N}(0,1)}$ denotes the $(1-\alpha/2)$-quantile of a standard Normal distribution. Thanks to (2.6), the confidence interval in (2.7) is asymptotically valid as $N \to +\infty$. When comparing two models M_2 and M_3, one would run two independent SMC samplers — one for each model — to obtain log-evidence estimators $\log \hat{Z}_{t,N}^{M_2}$ and $\log \hat{Z}_{t,N}^{M_3}$ along with their variance estimators $\widehat{V}_{t,N}^{M_2}$ and $\widehat{V}_{t,N}^{M_3}$. By independence, variances can be added up and we may combine the respective confidence intervals from (2.7) to obtain

$$\left[\; \left(\log \hat{Z}_{t,N}^{M_2} - \log \hat{Z}_{t,N}^{M_3} \right) \;\; \pm \;\; q_{1-\alpha/2}^{\mathcal{N}(0,1)} \sqrt{\widehat{V}_{t,N}^{M_2} + \widehat{V}_{t,N}^{M_3}} \; \right] \tag{2.8}$$

as an asymptotically valid $100(1-\alpha)\%$-confidence interval for the difference of log-evidences i.e. the log-Bayes factor $\mathrm{LBF}_t^{2|3} = \log p(y_{1:t}|M_2) - \log p(y_{1:t}|M_3)$ of model M_2 against model M_3, for all $t \in [\![1, T]\!]$. For log-evidences of state-space models, the above method can be directly applied — provided the transition kernel can be sampled from and the observation densities can be evaluated — by using SMC^2 samplers (Chopin et al., 2013) and regarding them as particular instances of exact SMC methods on extended spaces, as reviewed in Appendix A.2. For SMC^2 samplers producing N_θ particles $(\theta_t^{(i)})_{(t,i)\in[\![1,T]\!]\times[\![1,N_\theta]\!]}$, each carrying a particle filter made of N_x particles, the asymptotic validity of (2.7) holds as $N_\theta \to +\infty$ for any fixed $N_x \in \mathbb{N}^*$, albeit with different efficiency considerations as N_x varies. The construction of confidence intervals of the form (2.7) or (2.8) is illustrated on diffusion models in Section 2.2.3.

A few caveats need to be issued here: although the estimator $\widehat{V}_{t,N}$ in (2.5) can be computed for any implementation of SMC, its consistency in (2.6) has only been proved for SMC algorithms with non-random weights and a deterministic rejuvenation schedule based on multinomial resampling (Lee and Whiteley, 2018, Theorem 1). The extension of $\widehat{V}_{t,N}$'s theoretical guarantees to SMC algorithms with adaptive resampling schedules (e.g. the iterated batch importance sampling of Chopin, 2002, and Algorithm 2 in Appendix A.1.2), more advanced resampling schemes (e.g. the SSP resampling of Gerber et al., 2017), or random weights (e.g. produced by particle filters in the case of the SMC^2 of Chopin et al., 2013) would be an interesting avenue for future research.

The main advantage of the variance estimator $\widehat{V}_{t,N}$ is that it can be computed on the fly from the same run of SMC used to produce the log-evidence estimator $\log \hat{Z}_t$, at virtually no additional computational or memory cost (e.g. no need to store any lineages), as explained in Algorithm 2 and Lemma 6 from the supplementary material of Lee and Whiteley (2018). We only use a particular version of their variance estimator, whose original formulation applies to SMC estimators of the form $\hat{\gamma}_{t,N}(\varphi)$ and $\hat{\eta}_{t,N}(\varphi)$ for general test functions φ, while also allowing for time-varying number of particles $(N_t)_{t\in[\![0,T]\!]} \in (\mathbb{N}^*)^{T+1}$. Despite

the generality of their formulation, extending the present method to the construction of confidence intervals for H-scores is not straightforward: the main difficulty comes from the fact that H-scores are not posterior expectations of the form $\gamma_t(\varphi)$ nor $\eta_t(\varphi)$, and can only be expressed as $\sum_{s=1}^{t} h(\eta_s(\varphi_1), ..., \eta_s(\varphi_J))$ for some non-linear function $h : \mathbb{R}^J \to \mathbb{R}$ of J different test functions $(\varphi_j)_{j\in[\![1,J]\!]}$. This will call for a radically different approach, which will be investigated in Section 2.3.

Another limitation of a confidence interval based on (2.7) is that its validity is asymptotic in the number of particles N. Getting more accurate estimates can only be achieved by increasing the number of particles, and along with it the computational and memory costs of the underlying SMC algorithm. To the best of our knowledge, increasing N to $N' > N$ would require re-running an SMC with N' particles from scratch, with no direct way of re-using the former run with N particles to refine the results. Besides, the extent to which one can increase N on a given processor is limited by the amount of available memory, and vast increases in computation time are not sustainable in contexts where one needs to refine estimators within a budgeted time window. Parallelizing the SMC algorithm (e.g. Vergé et al., 2015; Murray et al., 2016; Whiteley et al., 2016) would alleviate the computation time, but it would not solve the fundamental issues of having undesirable asymptotics in N. This motivates the need for a different asymptotic regime, based on the number of independent processors, which would fully take advantage of modern and increasingly parallel computing architectures.

2.2.2 Asymptotic validity in the number of processors

The SMC estimator $\hat{Z}_{t,N}$ of the evidence has the key property of being unbiased, i.e. it satisfies $\mathbb{E}[\hat{Z}_{t,N} \,|\, y_{1:t}] = Z_t = p(y_{1:t})$ for all $t \in [\![1, T]\!]$ and — more remarkably — all finite $N \in \mathbb{N}^*$, where the expectation is with respect to the random variables generated in the course of running the SMC algorithm. As a consequence, we could perform R independent runs of SMC with N particles on R processors, entirely in parallel, and produce R independent and

identically distributed (i.i.d.) estimators $(\hat{Z}_{t,N}^{(r)})_{r\in[\![1,R]\!]}$ of the evidence. From these estimators, we can form the sample average $\bar{Z}_{t,N,R} = R^{-1}\sum_{r=1}^{R}\hat{Z}_{t,N}^{(r)}$. Under the mild assumption (see e.g. Del Moral, 2004; Del Moral et al., 2006) that the estimator $\hat{Z}_{t,N}$ has a finite variance $\sigma^2_{\hat{Z}_{t,N}} \in (0,+\infty)$, the standard central limit theorem for i.i.d. variables ensures that

$$\sqrt{\frac{R}{\sigma^2_{\hat{Z}_{t,N}}}}\left(\bar{Z}_{t,N,R} - Z_t\right) \quad \xrightarrow[R\to+\infty]{\mathcal{D}} \quad \mathcal{N}(0,1) \tag{2.9}$$

for all $t \in [\![1,T]\!]$ and all finite $N \in \mathbb{N}^*$, where the centering at Z_t crucially relies on the unbiasedness of $\hat{Z}_{t,N}$. Applying logarithms and using the delta method on (2.9) yields

$$\sqrt{R\frac{Z_t^2}{\sigma^2_{\hat{Z}_{t,N}}}}\left(\log\bar{Z}_{t,N,R} - \log Z_t\right) \quad \xrightarrow[R\to+\infty]{\mathcal{D}} \quad \mathcal{N}(0,1) \tag{2.10}$$

for all $t \in [\![1,T]\!]$ and all finite $N \in \mathbb{N}^*$. By respectively using the sample mean $\bar{Z}_{t,N,R}$ and the sample variance $\hat{\sigma}^2_{\hat{Z}_{t,N},R} = R^{-1}\sum_{r=1}^{R}(\hat{Z}_{t,N}^{(r)} - \bar{Z}_{t,N,R})^2$ to consistently estimate the unknown mean Z_t and the unknown variance $\sigma^2_{\hat{Z}_{t,N}}$ as $R \to +\infty$, we can turn (2.10) into a confidence interval for $\log Z_t$ defined as

$$\left[\ \log\bar{Z}_{t,N,R} \ \pm\ q^{\mathcal{N}(0,1)}_{1-\alpha/2}\sqrt{\frac{1}{R}\frac{\hat{\sigma}^2_{\hat{Z}_{t,N},R}}{\bar{Z}^2_{t,N,R}}}\ \right]$$

which further simplifies to

$$\left[\ \log\bar{Z}_{t,N,R} \ \pm\ q^{\mathcal{N}(0,1)}_{1-\alpha/2}\sqrt{\frac{1}{R^2}\sum_{r=1}^{R}\left(\frac{\hat{Z}_{t,N}^{(r)}}{\bar{Z}_{t,N,R}} - 1\right)^2}\ \right] \tag{2.11}$$

whose validity is asymptotic as $R \to +\infty$, for any finite value of $N \in \mathbb{N}^*$. Possible numerical underflow due to the $\hat{Z}_{t,N}^{(r)}$'s can be avoided by using a traditional log-sum-exp trick with

$$\begin{cases} \log\bar{Z}_{t,N,R} &= \log\left(\dfrac{1}{R}\displaystyle\sum_{r=1}^{R}\exp\left(\log\hat{Z}_{t,N}^{(r)} - \log\hat{Z}_{t,N}^{(*)}\right)\right) + \log\hat{Z}_{t,N}^{(*)} \\ \dfrac{\hat{Z}_{t,N}^{(r)}}{\bar{Z}_{t,N,R}} &= \dfrac{\exp\left(\log\hat{Z}_{t,N}^{(r)} - \log\hat{Z}_{t,N}^{(*)}\right)}{R^{-1}\sum_{j=1}^{R}\exp\left(\log\hat{Z}_{t,N}^{(j)} - \log\hat{Z}_{t,N}^{(*)}\right)} \end{cases}$$

where $\log \hat{Z}_{t,N}^{(*)} = \max\{\log \hat{Z}_{t,N}^{(r)} : r \in [\![1, R]\!]\}$.

Going from (2.11) to asymptotic confidence intervals for the difference of log-evidences between two models can be done similarly to (2.8), i.e. by running independent estimations for each model and leveraging the independence to add up the variances, leading to

$$\left[\left(\log \bar{Z}_{t,N,R}^{M_2} - \log \bar{Z}_{t,N,R}^{M_3} \right) \pm q_{1-\alpha/2}^{\mathcal{N}(0,1)} \sqrt{\frac{1}{R^2} \sum_{r=1}^{R} \left(\left(\frac{\hat{Z}_{t,N}^{M_2\,(r)}}{\bar{Z}_{t,N,R}^{M_2}} - 1 \right)^2 + \left(\frac{\hat{Z}_{t,N}^{M_3\,(r)}}{\bar{Z}_{t,N,R}^{M_3}} - 1 \right)^2 \right)} \right] \tag{2.12}$$

as an approximate $100(1-\alpha)\%$-confidence interval for $\mathrm{LBF}_t^{2|3}$, that is valid at all horizons $t \in [\![1, T]\!]$, for any finite value of $N \in \mathbb{N}^*$ and asymptotically as $R \to +\infty$. Numerical illustrations of confidence intervals based on (2.11) and (2.12) are provided in Section 2.2.3.

It is worth mentioning that unlike (2.7), whose theoretical understanding is currently limited to SMC samplers that use multinomial resampling, the construction in (2.11) allows for the use of other resampling schemes (e.g. Liu and Chen, 1998; Gerber et al., 2017) within the underlying SMC samplers, as long as the resulting estimators of the evidence remain unbiased. This can be ensured by imposing sufficient conditions on the resampling schemes, such as Assumption 2 in Andrieu et al. (2010, equations (23) and (24)). We emphasize once more how much the validity of this approach relies on the unbiased property of $\hat{Z}_{t,N}$ to justify averaging over independent replications in (2.9). As a word of caution, this precludes the use of adaptive versions of SMC, whose evidence estimators are generally no longer unbiased, although this can be mitigated by performing a preliminary run of adaptive SMC — to tune the algorithmic parameters. The reliance on unbiasedness of SMC estimators also prevents any generalization of the current approach to the construction of confidence intervals for H-scores : although SMC samplers can be used to produce consistent estimators of H-scores, as explained in Sections 1.2.1 and 1.3.1 of Chapter 1, these estimators are typically biased. The more involved construction of confidence intervals for H-scores will be investigated in Section 2.3.

2.2.3 Numerical illustration with diffusion models

We will use the population dynamics models introduced in Section 1.4.2 of Chapter 1 as a recurring example to illustrate our different methods of constructing confidence intervals. Unless stated otherwise, we will use 95% as a confidence level by default. We restrict our attention to the models M_2 and M_3 that appeared to be the most similar, and thus the hardest to tell apart. Recall that the data consists of $T = 41$ bi-variate vectors of counts $(Y_{1,t}\,, Y_{2,t})_{t\in[\![1,T]\!]} \in (\mathbb{N}\times\mathbb{N})^T$ (Caughley et al., 1987; Knape and de Valpine, 2012), as shown in Figure 1.4 of Chapter 1, and the models are defined as

$$
\begin{aligned}
M_2\colon\ & X_1 \sim \mathrm{LN}(0,5)\,; \quad dX_t/X_t = (\sigma^2/2 + r)\,dt + \sigma dW_t\,; \\
& Y_{1,t}\,, Y_{2,t}\,|\,X_t\,, \tau \overset{\text{i.i.d.}}{\sim} \mathrm{NB}(X_t,\, X_t + \tau X_t^2)\,; \\
& \textit{with independent priors; } \sigma,\, \tau \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,10),\ r \sim \mathrm{Unif}(-10,10).
\end{aligned}
$$

$$
\begin{aligned}
M_3\colon\ & X_1 \sim \mathrm{LN}(0,5)\,; \quad dX_t/X_t = (\sigma^2/2)\,dt + \sigma dW_t\,; \\
& Y_{1,t}\,, Y_{2,t}\,|\,X_t\,, \tau \overset{\text{i.i.d.}}{\sim} \mathrm{NB}(X_t,\, X_t + \tau X_t^2)\,; \\
& \textit{with independent priors; } \sigma,\, \tau \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,10).
\end{aligned}
$$

In Figure 2.1, we demonstrate the construction of approximate confidence intervals for the log-scores (i.e. negatives of log-evidences) of models M_2 and M_3, using the methods (2.7) and (2.11) respectively presented in Sections 2.2.1 and 2.2.2. The intervals based on (2.7) are obtained from one run of non-adaptive SMC^2 for each model, using $N_\theta = 1024$ and $N_x = 32$ particles, with forced rejuvenation at every step and a multinomial resampling scheme. The intervals based on (2.11) are obtained from $R = 100$ independent runs of non-adaptive SMC^2 for each model, using $N_\theta = 1024$ and $N_x = 32$ particles, with multinomial resampling and a rejuvenation schedule learned beforehand from a preliminary adaptive run. The intervals (2.7) and (2.11) are plotted for each model and each horizon $t \in [\![1, 41]\!]$. For better readability, the log-scores and their confidence intervals are rescaled by the number of observations.

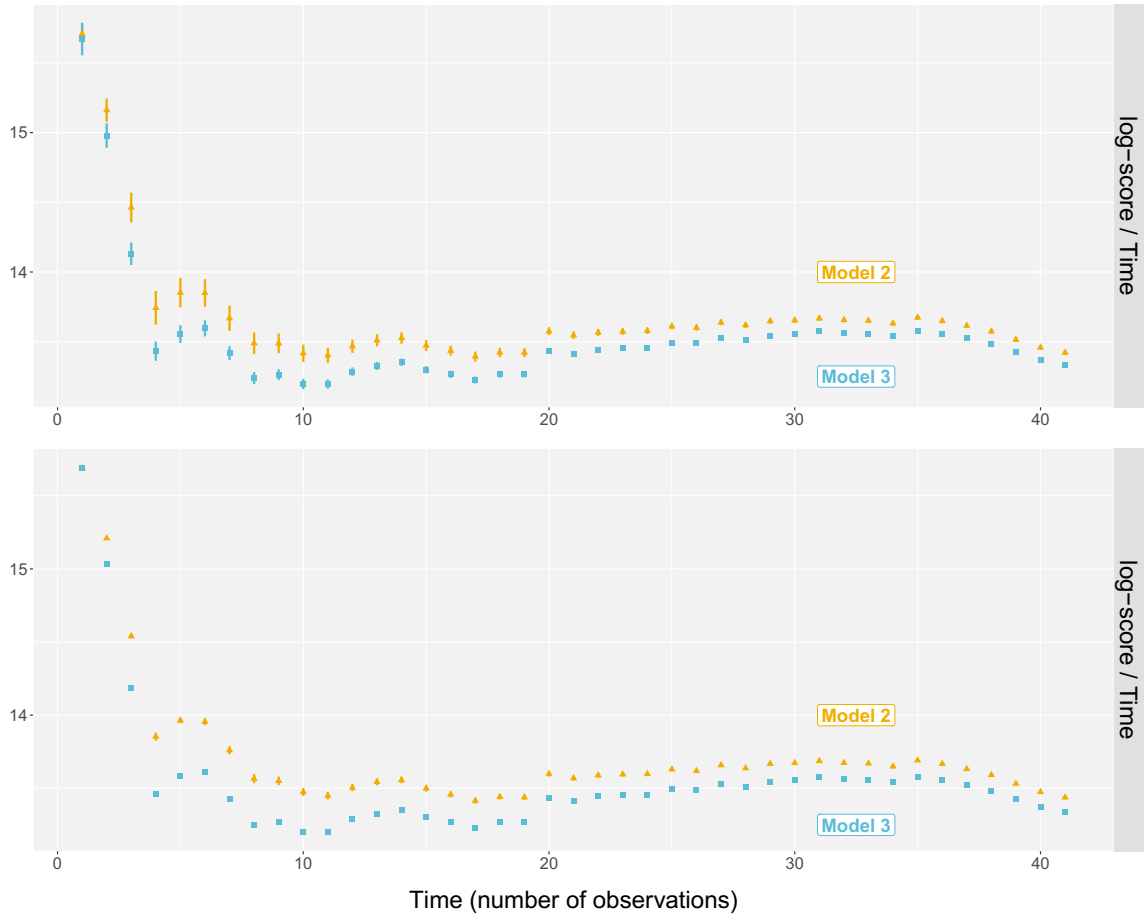

Figure 2.1. *Approximate* 95%*-confidence intervals (vertical segments) for the log-scores of models* M_2 *(yellow triangles) and* M_3 *(blue squares) scaled by the number of observations. Top panel: using equation* (2.7) *with variance estimators from single runs of SMC*^{2} *with* $N_\theta = 1024$ *and* $N_x = 32$*. Bottom panel: using equation* (2.11) *with* $R = 100$ *independent runs of SMC*^{2} *with* $N_\theta = 1024$ *and* $N_x = 32$*. See Section 2.2.3.*

The resulting approximate confidence intervals for the log-Bayes factor of model M_2 against model M_3 are shown in Figure 2.2. We see that they stop containing 0 after a few observations, and quickly drift further away to the negative side. Conditional on the observed data, Figure 2.2 would suggest a preference for model M_3, thus corroborating the initial hunch ventured in Section 1.4.2, but with stronger theoretical grounds this time around. As with classical hypothesis testing, care is advised when interpreting the results of these confidence intervals. The confidence intervals (2.8) and (2.12) asymptotically have the correct

coverage $(1-\alpha)$, either when $N \to +\infty$ or $R \to +\infty$ respectively, in the following sense: conditional on the realized observations $y_{1:T}$, these random intervals at horizon $t \in [\![1, T]\!]$ contain the exact log-Bayes factor $\mathrm{LBF}_t^{2|3}$ with probability $(1-\alpha)$, where the probabilities are taken with respect to the joint sampling distribution of all the random variables generated during the run of the corresponding SMC algorithms.

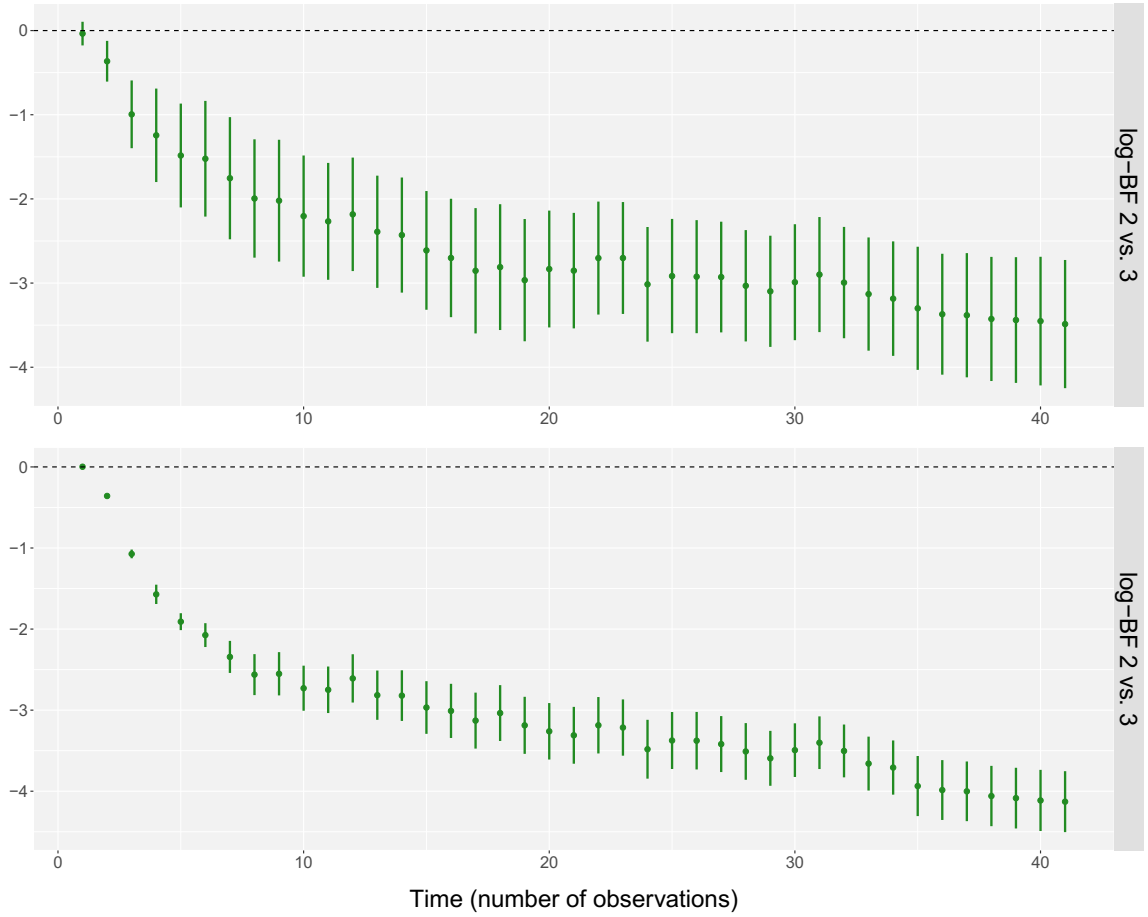

Figure 2.2. *Approximate 95%-confidence intervals for the log-Bayes factor of models M_2 against M_3 (vertical segments centered at the dots). Top panel: using equation (2.8) with variance estimators from single runs of SMC^2 with $N_\theta = 1024$ and $N_x = 32$. Bottom panel: using equation (2.12) with $R = 100$ independent runs of SMC^2 with $N_\theta = 1024$ and $N_x = 32$. See Section 2.2.3.*

2.3 Confidence intervals for the H-score

As previously mentioned, none of the methods presented so far is directly applicable to the construction of confidence intervals for H-scores. We further investigate the matter here, starting in Section 2.3.1 by expressing H-scores as explicit functions of posterior expectations. The work of Jacob et al. (2017) on unbiased MCMC methods provides a way of using R independent processors to produce estimators that satisfy a central limit theorem centered at those exact posterior expectations, with asymptotics in the number of processors. By a suitable application of the delta method, these estimators can be further combined into estimated H-scores, whose asymptotic distribution as $R \to +\infty$ will also be Normal, centered precisely at the exact H-scores. This procedure enables the construction of approximate confidence intervals for H-scores, whose validity is asymptotic in the number of independent processors, as desired. Different variants of unbiased MCMC methods will then lead to possibly different constructions. In Section 2.3.2, we propose a construction based on coupled particle independent Metropolis-Hastings (PIMH) algorithms (Middleton et al., 2019) with SMC samplers as proposal distributions. In Section 2.3.3, we provide some numerical illustrations that will raise legitimate concerns about the tuning of the aforementioned algorithm. We discuss these concerns more thoroughly in Section 2.3.4.

2.3.1 Combining unbiased MCMC and delta methods

After reviewing how unbiased MCMC can be used to derive central limit theorems for estimators of posterior expectations (Section 2.3.1.1), we explain how the delta method can enable the construction of confidence intervals for H-scores. This will rely on our ability to express the H-score as a function of posterior expectations, both in settings where the observations are continuous (Section 2.3.1.2) and discrete (Section 2.3.1.3).

2.3.1.1 Unbiased MCMC for posterior expectations

Given a target probability distribution π on some space $\mathbb{V} \subseteq \mathbb{R}^{d_\mathcal{V}}$ and a test function $\varphi : \mathbb{V} \to \mathbb{R}$ of interest, unbiased MCMC methods (Jacob et al., 2017) are a particular class of MCMC algorithms (e.g. Robert and Casella, 1999; Liu, 2008; Brooks, Gelman, Jones and Meng, 2011; Green, Łatuszyński, Pereyra and Robert, 2015) that produce a pair of Markov chains $\mathcal{V} = (\mathcal{V}^{(n)})_{n\in\mathbb{N}}$ and $\mathcal{W} = (\mathcal{W}^{(n)})_{n\in\mathbb{N}}$ in $\mathbb{V}^\mathbb{N}$ having the following properties:

- The two processes $(\mathcal{V}^{(n)})_{n\in\mathbb{N}}$ and $(\mathcal{W}^{(n)})_{n\in\mathbb{N}}$ have the same law: each Markov chain marginally starts from an initial distribution π_0, evolves according to a transition kernel P, and converges to π as its invariant distribution. By letting $\pi(\varphi)$ denote the expectation $\mathbb{E}[\varphi(\tilde{\mathcal{V}})]$ under $\tilde{\mathcal{V}} \sim \pi$, we have $\mathbb{E}[\varphi(\mathcal{V}^{(n)})] \xrightarrow[n\to+\infty]{} \pi(\varphi)$.
- The joint transition kernel of $(\mathcal{V}^{(n)}, \mathcal{W}^{(n)})_{n\in\mathbb{N}}$ uses a coupling (i.e. a constrained joint distribution) that is designed to make the two chains almost surely meet, i.e. the stopping time defined by $\tau = \min\{n \in \mathbb{N}^* : \mathcal{V}^{(n)} = \mathcal{W}^{(n-1)}\}$ is almost surely finite and satisfies $\mathbb{E}[\tau] < +\infty$. We refer to τ as the *meeting time* of the two chains.
- The two chains stay together after meeting, so that $\mathcal{V}^{(n)} = \mathcal{W}^{(n-1)}$ for all $n \geq \tau$.

Different settings, couplings, and technical conditions can be considered, leading to different variants and theoretical guarantees (Jacob et al., 2017; Middleton, Deligiannidis, Doucet and Jacob, 2018; Heng and Jacob, 2019; Middleton et al., 2019). We focus on the setting where π is the posterior distribution of a parameter, conditional on some fixed observations, and our main interest in these unbiased MCMC algorithms lies in their unique ability to produce a bias-corrected estimator

$$\Phi^{\varphi}_{k:m}(\mathcal{V}, \mathcal{W}) = \left[\frac{1}{m-k+1}\sum_{\ell=k}^{m}\varphi(\mathcal{V}^{(\ell)})\right] + \left[\sum_{\ell=k}^{\tau-1}\min\left(1, \frac{\ell-k+1}{m-k+1}\right)\left(\varphi(\mathcal{V}^{(\ell+1)}) - \varphi(\mathcal{W}^{(\ell)})\right)\right] \quad (2.13)$$

which, for all $(k,m) \in \mathbb{N}^2$ with $k \leq m$, has the particularity — under the regularity assumptions of Proposition 3.1 in Jacob et al. (2017) or Theorem 1 in Middleton et al. (2018) — of being an unbiased estimator of $\pi(\varphi)$, with finite expected computation time, and finite variance denoted by $\sigma^2_{\Phi^{\varphi}_{k:m}}$. These remarkable properties imply that, by performing R independent runs of unbiased MCMC on R processors in parallel to produce i.i.d. estimators $(\Phi^{\varphi,(r)}_{k:m})_{r\in[\![1,R]\!]}$ from (2.13), we can form the average $\bar{\Phi}^{\varphi}_{k:m,R} = R^{-1}\sum_{r=1}^{R}\Phi^{\varphi,(r)}_{k:m}$ and invoke the standard law of large numbers along with the central limit theorem for i.i.d. variables to claim

$$\bar{\Phi}^{\varphi}_{k:m,R} \quad \xrightarrow[R\to+\infty]{\mathbb{P}-a.s.} \quad \pi(\varphi) \tag{2.14}$$

$$\sqrt{R}\left(\bar{\Phi}^{\varphi}_{k:m,R} - \pi(\varphi)\right) \quad \xrightarrow[R\to+\infty]{\mathcal{D}} \quad \mathcal{N}\left(0, \sigma^2_{\Phi^{\varphi}_{k:m}}\right) \tag{2.15}$$

so that for any continuously differentiable function $h : \mathbb{V} \to \mathbb{R}$ with derivative h' satisfying $h'(\pi(\varphi)) \neq 0$, applying the delta method to (2.15) leads to

$$\sqrt{R}\left(h(\bar{\Phi}^{\varphi}_{k:m,R}) - h(\pi(\varphi))\right) \quad \xrightarrow[R\to+\infty]{\mathcal{D}} \quad \mathcal{N}\left(0, \left[h'\left(\pi(\varphi)\right)\right]^2 \sigma^2_{\Phi^{\varphi}_{k:m}}\right). \tag{2.16}$$

By respectively using the sample variance $\hat{\sigma}^2_{\Phi^{\varphi}_{k:m},R} = R^{-1}\sum_{r=1}^{R}(\Phi^{\varphi,(r)}_{k:m} - \bar{\Phi}^{\varphi}_{k:m,R})^2$ and the plug-in estimate $h'(\bar{\Phi}^{\varphi}_{k:m,R})$ as consistent estimators of the unknown $\sigma^2_{\Phi^{\varphi}_{k:m}}$ and $h'(\pi(\varphi))$ — the latter being a consequence of (2.14) and the continuous mapping theorem — we can construct

$$\left[\; h\left(\bar{\Phi}^{\varphi}_{k:m,R}\right) \;\pm\; q^{\mathcal{N}(0,1)}_{1-\alpha/2}\sqrt{\frac{1}{R}\left(h'(\bar{\Phi}_{k:m,R})\right)^2 \hat{\sigma}^2_{\Phi^{\varphi}_{k:m},R}}\;\right]$$

as an approximate $100(1-\alpha)\%$-confidence interval for $h(\pi(\varphi))$ that is asymptotically valid as $R \to +\infty$, thanks to (2.16) and Slutsky's theorem. Although valid for all $(k,m) \in \mathbb{N}^2$ with $k \leq m$, practical and efficiency considerations should guide the choice of (k,m), as will be discussed in Section 2.3.4.3.

The previous approach directly generalizes to a collection of J test functions $\varphi = (\varphi_j)_{j\in[\![1,J]\!]}$ and a multivariate transformation $h : \mathbb{V}^J \to \mathbb{R}$ as follows. Each processor $r \in [\![1, R]\!]$ can

perform one single run of unbiased MCMC to obtain a pair of coupled chains $(\mathcal{V}_n^{(r)}, \mathcal{W}_n^{(r)})_{n\in\mathbb{N}}$, independently of all the other processors. This pair of chains is then used to compute the bias-corrected estimators $\Phi_{k:m}^{\varphi_j,(r)}$ from (2.13) for each test function φ_j, thus forming a random vector $\Phi_{k:m}^{\varphi,(r)} = (\Phi_{k:m}^{\varphi_j,(r)})_{j\in[\![1,J]\!]} \in \mathbb{R}^J$. Under similar regularity assumptions as before, the component-wise expectation of $\Phi_{k:m}^{\varphi,(r)}$ is equal to the vector $\pi(\varphi) = (\pi(\varphi_j))_{j\in[\![1,J]\!]}$, and its covariance matrix $\Sigma_{k:m}^{\varphi} = (\mathrm{Cov}(\Phi_{k:m}^{\varphi_i,(r)}, \Phi_{k:m}^{\varphi_j,(r)}))_{(i,j)\in[\![1,J]\!]^2} \in \mathbb{R}^{J\times J}$ is well-defined. By forming the sample average vector across R independent processors $\bar{\Phi}_{k:m,R}^{\varphi} = R^{-1}\sum_{r=1}^{R}\Phi_{k:m}^{\varphi,(r)}$ and invoking the multivariate versions of the law of large numbers, central limit theorem, and delta method assuming $\nabla h(\pi(\varphi)) \neq 0$, we obtain multivariate analogs of (2.14) to (2.16) given by

$$\bar{\Phi}_{k:m,R}^{\varphi} \quad \xrightarrow[R\to+\infty]{\mathbb{P}-a.s.} \quad \pi(\varphi) \tag{2.17}$$

$$\sqrt{R}\left(\bar{\Phi}_{k:m,R}^{\varphi} - \pi(\varphi)\right) \quad \xrightarrow[R\to+\infty]{\mathcal{D}} \quad \mathcal{N}\left(0, \Sigma_{k:m}^{\varphi}\right) \tag{2.18}$$

$$\sqrt{R}\left(h(\bar{\Phi}_{k:m,R}^{\varphi}) - h(\pi(\varphi))\right) \quad \xrightarrow[R\to+\infty]{\mathcal{D}} \quad \mathcal{N}\left(0, \nabla h\left(\pi(\varphi)\right)^{\top}\Sigma_{k:m}^{\varphi}\nabla h\left(\pi(\varphi)\right)\right) \tag{2.19}$$

which finally leads to the construction of

$$\left[\; h\left(\bar{\Phi}_{k:m,R}^{\varphi}\right) \;\pm\; q_{1-\alpha/2}^{\mathcal{N}(0,1)}\sqrt{\frac{1}{R}\nabla h\left(\bar{\Phi}_{k:m,R}^{\varphi}\right)^{\top}\hat{\Sigma}_{k:m,R}^{\varphi}\nabla h\left(\bar{\Phi}_{k:m,R}^{\varphi}\right)}\;\right] \tag{2.20}$$

where $\hat{\Sigma}_{k:m,R}^{\varphi} = R^{-1}\sum_{r=1}^{R}(\Phi_{k:m}^{\varphi,(r)} - \bar{\Phi}_{k:m,R}^{\varphi})(\Phi_{k:m}^{\varphi,(r)} - \bar{\Phi}_{k:m,R}^{\varphi})^{\top}$ is the sample variance of the realized vectors $(\Phi_{k:m}^{\varphi,(r)})_{r\in[\![1,R]\!]}$ across the R independent processors. The interval in (2.20) is an approximate $100(1-\alpha)\%$-confidence interval for $h(\pi(\varphi))$, whose asymptotic validity is in the number of independent processors $R \to +\infty$.

By thinking of π as a posterior distribution, the derivation of (2.20) offers a convenient and inherently parallelizable way to construct confidence intervals for virtually any smooth functions of posterior expectations. This will be particularly relevant to us since, as explained in the next sections, H-scores can be explicitly expressed as continuously differentiable functions of vectors of posterior expectations.

2.3.1.2 H-scores as functions of posterior expectations: continuous case

For continuous observations $y_{1:T} \in \mathbb{Y}^T \subseteq (\mathbb{R}^{d_y})^T$, equations (1.4) and (1.13) from Chapter 1 allow us to express the H-score of a model M at any intermediate horizon $s \in [\![1, T]\!]$ as

$$\mathcal{H}_s(M) \;=\; \sum_{t=1}^{s} \sum_{d=1}^{d_y} \left(2\, E_{t,d,1} - (E_{t,d,2})^2 \right) \tag{2.21}$$

where $E_{t,d,1}$ and $E_{t,d,2}$ are posterior expectations with respect to $\Theta \sim p(d\theta|y_{1:t})$ defined as

$$\begin{aligned} E_{t,d,1} \;&=\; \mathbb{E}\left[\frac{\partial^2 \log p(y_t|y_{1:t-1}, \Theta)}{\partial y_{t(d)}^2} + \left(\frac{\partial \log p(y_t|y_{1:t-1}, \Theta)}{\partial y_{t(d)}} \right)^2 \,\middle|\, y_{1:t} \right], \\ E_{t,d,2} \;&=\; \mathbb{E}\left[\frac{\partial \log p(y_t|y_{1:t-1}, \Theta)}{\partial y_{t(d)}} \,\middle|\, y_{1:t} \right], \end{aligned}$$

which, in the context of state-space models, further simplify to posterior expectations with respect to $(\Theta, X_t) \sim p(d\theta|y_{1:t})p(dx_t|y_{1:t}, \theta)$ given by

$$\begin{aligned} E_{t,d,1} \;&=\; \mathbb{E}\left[\frac{\partial^2 \log g_\Theta(y_t|X_t)}{\partial y_{t(d)}^2} + \left(\frac{\partial \log g_\Theta(y_t|X_t), \Theta)}{\partial y_{t(k)}} \right)^2 \,\middle|\, y_{1:t} \right], \\ E_{t,d,2} \;&=\; \mathbb{E}\left[\frac{\partial \log g_\Theta(y_t|X_t), \Theta)}{\partial y_{t(d)}} \,\middle|\, y_{1:t} \right], \end{aligned}$$

under mild regularity conditions (see Assumption A11 in Appendix B.4). In other words, the H-score at horizon $s \in [\![1, T]\!]$ can be written as

$$\mathcal{H}_s(M) \;=\; h_s\left((E_{t,d,j})_{(t,d,j) \in [\![1,s]\!] \times [\![1,d_y]\!] \times [\![1,2]\!]} \right) \tag{2.22}$$

where $(E_{t,d,j})_{(t,d,j) \in [\![1,s]\!] \times [\![1,d_y]\!] \times [\![1,2]\!]}$ is the vector enumerating the $E_{t,d,j}$'s in the lexicographic order of $[\![1, s]\!] \times [\![1, d_y]\!] \times [\![1, 2]\!]$ induced by the natural order on $\mathbb{N}$, and $h_s : \mathbb{R}^{s \times d_y \times 2} \to \mathbb{R}$ is a continuously differentiable function whose gradient ∇h_s can be analytically computed. The functional forms of h_s and ∇h_s are detailed in Appendix C.1.1. Thanks to (2.22), the construction of confidence intervals from (2.20) will directly apply as soon as we can produce R unbiased estimators $((\hat{E}_{t,d,j}^{(r)})_{(t,d,j) \in [\![1,s]\!] \times [\![1,d_y]\!] \times [\![1,2]\!]})_{r \in [\![1,R]\!]}$ for $(E_{t,d,j})_{(t,d,j) \in [\![1,s]\!] \times [\![1,d_y]\!] \times [\![1,2]\!]}$ from (2.13).

2.3.1.3 H-scores as functions of posterior expectations: discrete case

Let e_d denote the canonical vector of $\mathbb{Z}^{d_y}$ that has all coordinates equal to 0 except for its d-th coordinate that equals 1. From Section 1.4.1 of Chapter 1, the general form of the H-score at horizon $s \in [\![1, T]\!]$ in the case of discrete observations is given — except at possible boundaries if $\mathbb{Y}$ is bounded, in which case some of the undefined terms are dropped and the formula simplifies even further — by the expression

$$\begin{aligned}\mathcal{H}_T(M) &= \sum_{t=1}^{T}\sum_{d=1}^{d_y}\Big(2\,\partial_d\Big(\partial_d \log p(y_t \,|\, y_{1:t-1})\Big) + \big(\partial_d \log p(y_t \,|\, y_{1:t-1})\big)^2\Big) \\ &= \sum_{t=1}^{T}\sum_{d=1}^{d_y}\left(\frac{p_{t,d,2} - p_{t,d,0}}{2\,p_{t,d,1}} - \frac{p_{t,d,0} - p_{t,d,-2}}{2\,p_{t,d,-1}} + \left(\frac{p_{t,d,1} - p_{t,d,-1}}{2\,p_{t,d,0}}\right)^2\right)\end{aligned} \tag{2.23}$$

where the $p_{t,d,j}$'s are posterior expectations with respect to $\Theta \sim p(d\theta|y_{1:t-1})$ defined as

$$p_{t,d,j} \;=\; p\left(y_t + je_d \,|\, y_{1:t-1}\right) \;=\; \mathbb{E}\Big[p(y_t + je_d \,|\, y_{1:t-1}, \Theta)\;\Big|\; y_{1:t-1}\Big]$$

for all $(t, d, j) \in [\![1, T]\!] \times [\![1, d_y]\!] \times [\![-2, 2]\!]$. For state-space models, this simplifies to

$$p_{t,d,j} \;=\; \mathbb{E}\Big[g_\Theta(y_t + je_d \,|\, X_t)\;\Big|\; y_{1:t-1}\Big]$$

where the posterior expectation is now with respect to $(\Theta, X_t) \sim p(d\theta|y_{1:t-1})p(dx_t|y_{1:t-1}, \theta)$. The H-score at horizon $s \in [\![1, T]\!]$ can thus be written as

$$\mathcal{H}_s(M) \;=\; h_s\Big((p_{t,d,j})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![-2,2]\!]}\Big) \tag{2.24}$$

where $(p_{t,d,j})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![-2,2]\!]}$ is the vector enumerating the $p_{t,d,j}$'s in the lexicographic order of $[\![1, s]\!] \times [\![1, d_y]\!] \times [\![-2, 2]\!]$ induced by the natural order on $\mathbb{Z}$, and $h_s : \mathbb{R}^{s\times d_y\times 5} \to \mathbb{R}$ is a continuously differentiable function whose gradient ∇h_s can be analytically computed. The details of their functional forms are provided in Appendix C.1.2.

2.3.2 Using coupled particle independent Metropolis–Hastings

As explained in Section 2.3.1.1, the construction of confidence intervals for H-scores using (2.20) boils down to producing i.i.d. estimators that are unbiased either for $(E_{t,d,j})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![1,2]\!]}$ from (2.22) if the observations are continuous, or for $(p_{t,d,j})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![-2,2]\!]}$ from (2.24) when the observations are discrete.

The coupled PIMH (Middleton et al., 2019), reviewed in Appendix A.3.2, is a an algorithm that can concurrently run T pairs of Markov chains such that, for each horizon $t \in [\![1, T]\!]$, the t-th pair of chains targets the intermediate posterior distribution $p(d\theta|y_{1:t})$ at horizon t, while also satisfying all the requirements presented in Section 2.3.1.1 under mild assumptions (Middleton et al., 2019, Proposition 3). Each iteration of the coupled PIMH invokes pseudo-marginal Metropolis-Hastings kernels across different horizons and thus requires T proposals, one for each horizon $t \in [\![1, T]\!]$. These proposals can be generated from a single run of SMC (or SMC^2). Indeed, running an SMC (resp. SMC^2) sampler with N_θ (resp. N_θ and N_x) particles over all T observations sequentially approximates the successive posterior distributions $p(d\theta|y_{1:t})$ at every $t \in [\![1, T]\!]$ by empirical distributions $\sum_{n=1}^{N_\theta} W_{\theta,t}^{(m)} \delta_{\theta_t^{(m)}}$, from which we can draw proposals $\tilde{\theta}_t$, along with non-negative unbiased estimators $\hat{Z}_t$ of the evidence $p(y_{1:t})$ needed to compute the acceptance ratios in the underlying pseudo-marginal Metropolis-Hastings kernel. The unbiasedness of the $\hat{Z}_t$'s is vital for the validity of the coupled PIMH, and thus limits the use of adaptive variants of SMC when using them as proposal distributions within PIMH. Further details are provided in Algorithms 4 and 5 of Appendix A.3. If one already has a working implementation of SMC, the coupled PIMH has the advantage of being naturally modularized and simply adds an external layer of Metropolis-Hastings iterations, with no alteration to the original SMC implementation. The coupled PIMH also allows for a convenient analysis and enjoys strong theoretical guarantees, especially on the distribution of its meeting times (Middleton et al., 2018, Proposition 8).

2. Uncertainty quantification for model comparison

For a single processor indexed by r, performing one run of coupled PIMH using SMC (or SMC^2) proposals produces T pairs of coupled chains $\{(\mathcal{V}_t^{(n)}, \mathcal{W}_t^{(n)})_{n\in\mathbb{N}} : t \in [\![1, T]\!]\}$, one for each horizon $t \in [\![1, T]\!]$, and each pair of chains $(\mathcal{V}_t^{(n)}, \mathcal{W}_t^{(n)})_{n\in\mathbb{N}}$ can be used with (2.13) to produce unbiased estimators $(\hat{E}_{t,d,j}^{(r)})_{(d,j)\in[\![1,d_y]\!]\times[\![1,2]\!]}$ or $(\hat{p}_{t,d,j}^{(r)})_{(d,j)\in[\![1,d_y]\!]\times[\![-2,2]\!]}$ depending on the context. By gathering these estimators across different horizons, processor r can produce vectors $(\hat{E}_{t,d,j}^{(r)})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![1,2]\!]}$ or $(\hat{p}_{t,d,j}^{(r)})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![-2,2]\!]}$ of unbiased estimators, respectively for $(E_{t,d,j})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![1,2]\!]}$ or $(p_{t,d,j})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![-2,2]\!]}$, as desired. Although the components of these random vectors generally share some dependence, the key point is that these vectors of unbiased estimators can be generated independently across R different processors, so that averaging i.i.d. replications of them will still yield a valid central limit theorem of the form (2.18), thus enabling the construction of confidence intervals for H-scores given by (2.20). We illustrate the construction of such intervals in Section 2.3.3.

Remark 3. *In the case of continuous observations, the expression of the H-score in* (2.21) *as a polynomial of conditional expectations allows for a notable alternative to the reliance on the delta method. If processor r performs 2 independent runs of coupled PIMH with SMC (or SMC^2) proposals to obtain the unbiased estimators $(\hat{E}_{t,d,j}^{(r,1)})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![1,2]\!]}$ and $(\hat{E}_{t,d,j}^{(r,2)})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![1,2]\!]}$, then $(\hat{E}_{t,d,1}^{(r,1)} + \hat{E}_{t,d,1}^{(r,2)})$ unbiasedly estimates $2E_{t,d,1}$ by linearity, and $\hat{E}_{t,d,2}^{(r,1)}\hat{E}_{t,d,2}^{(r,2)}$ unbiasedly estimates $(E_{t,d,2})^2$ by independence, so that*

$$\hat{\mathcal{H}}_s^{(r)} = \sum_{t=1}^{s}\sum_{d=1}^{d_y}\left(\left(\hat{E}_{t,1,d}^{(r,1)} + \hat{E}_{t,1,d}^{(r,2)}\right) - \hat{E}_{t,2,d}^{(r,1)}\hat{E}_{t,2,d}^{(r,2)}\right)$$

is an unbiased estimator of the H-score $\mathcal{H}_s(M)$ of model M at any horizon $s \in [\![1, T]\!]$. By using R independent processors to form the sample average $\bar{\mathcal{H}}_{s,R} = R^{-1}\sum_{r=1}^{R}\hat{\mathcal{H}}_s^{(r)}$ and the sample variance $\hat{\sigma}^2_{\hat{\mathcal{H}}_s} = R^{-1}\sum_{r=1}^{R}(\hat{\mathcal{H}}_s^{(r)} - \bar{\mathcal{H}}_{s,R})^2$, we obtain the confidence interval

$$\left[\ \bar{\mathcal{H}}_{s,R} \ \pm \ q_{1-\alpha/2}^{\mathcal{N}(0,1)}\sqrt{\frac{1}{R}\hat{\sigma}^2_{\hat{\mathcal{H}}_s}}\ \right] \tag{2.25}$$

for $\mathcal{H}_s(M)$ at horizon $s \in [\![1, T]\!]$. This approximate $100(1-\alpha)\%$-confidence interval retains the desirable asymptotic regime in the number R of independent processors, and offers an arguably simpler alternative to the interval from (2.20).

2.3.3 Numerical illustration with diffusion models

As an illustration, we use once more the population dynamics models M_2 and M_3 from Chapter 1, restated in Section 2.2.3. In Figure 2.3, we construct approximate 95%-confidence intervals for the H-scores of each model, at each horizon $t \in [\![1, T]\!]$. These intervals are based on the construction (2.20) from Section 2.3, using a coupled PIMH algorithm with SMC^2 proposals. The SMC^2 samplers use $N_\theta = 8192$ and $N_x = 128$ particles, with rejuvenation kernels and a temperature schedule learned from a preliminary adaptive run. The present attempt uses $k = 5$ and $m = 10$ to form the estimators in (2.13). The computations were performed by $R = 1000$ processors using AMD Opteron 6300-series CPU cores from Harvard University's Odyssey cluster.

Figure 2.3 also shows, for every horizon $s \in [\![1, T]\!]$, approximate 95%-confidence intervals for the H-factor $\mathrm{HF}_s^{2|3} = \mathcal{H}_s(M_3) - \mathcal{H}_s(M_2)$ of M_2 against M_3, formed by adding the estimated variance to combine the independent intervals for $\mathcal{H}_s(M_2)$ and $\mathcal{H}_s(M_3)$ in a similar fashion as in (2.8). Although these intervals are centered at negative values, which would suggest a preference for model M_3, their range is too wide to rule out $(0, +\infty)$. In other words, the Monte Carlo error in estimating the H-factor appears to be of the same magnitude as the H-factor itself, preventing any reliable conclusion in deciding which model to select. A particular advantage of the interval in (2.20) is that it can be conveniently refined by simply increasing the number R of independent processors. In the asymptotic regime, the length of the interval would approximately scale as $1/\sqrt{R}$ when R increases. Concurrently, more fundamental variance reductions and efficiency gains can be obtained by choosing better

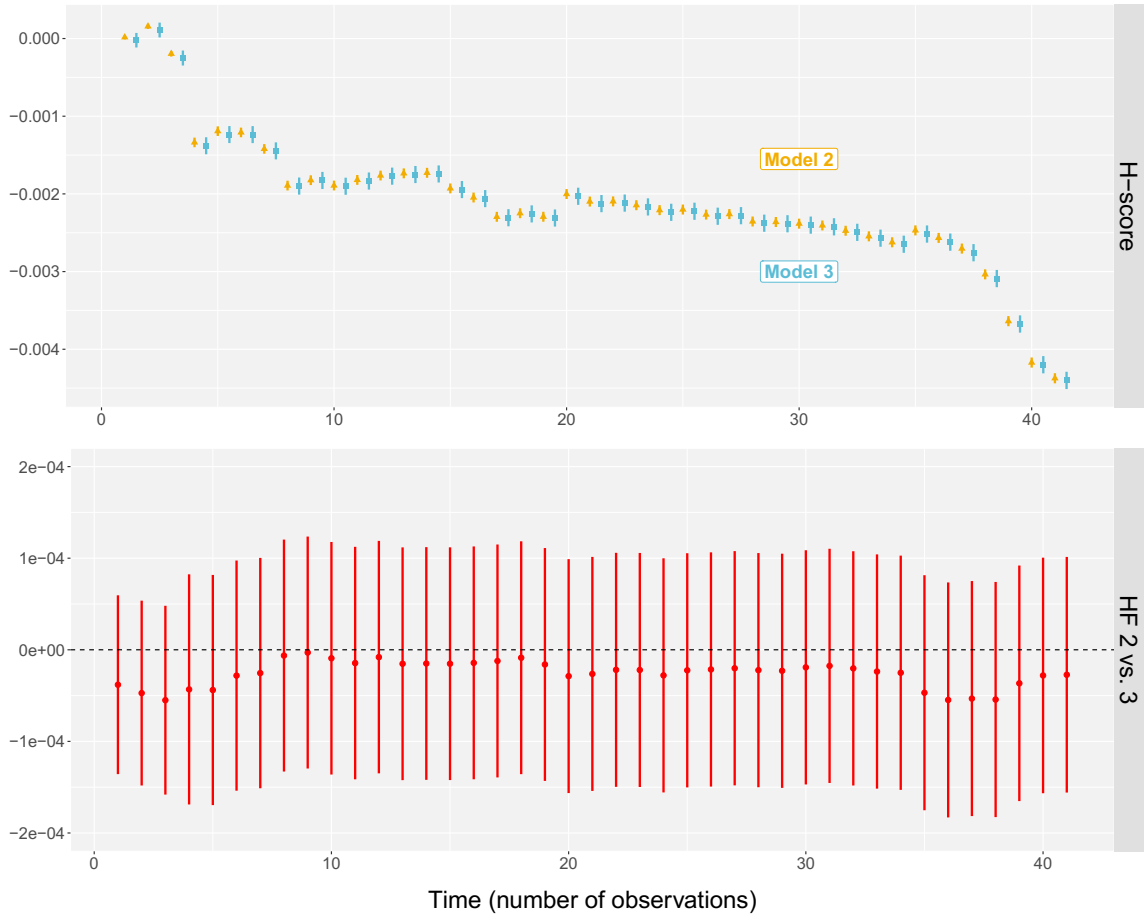

Figure 2.3. *Approximate 95%-confidence intervals (vertical segments) obtained from $R = 1000$ independent coupled PIMH using SMC^2 proposals with $N_\theta = 8192$, $N_x = 128$, $k = 5$, and $m = 10$. Top panel: H-scores of models M_2 (yellow triangles) and M_3 (blue squares). Bottom panel: H-factor of models M_2 against M_3. See Section 2.3.3.*

values of N_θ, N_x, k and m for the coupled PIMH algorithm. We investigate the tuning and trade-offs of such algorithmic parameters in Section 2.3.4.

One possible diagnostic for the coupled PIMH is to look at how fast the pairs of Markov chains meet. Figure 2.4 shows boxplots of the realized meeting times $(\tau_{M,t}^{(r)})_{r\in[\![1,R]\!]}$ across the $R = 1000$ processors, for each model $M \in \{M_2, M_3\}$ and each horizon $t \in [\![1, T]\!]$. One particular feature in the design of coupled PIMH algorithms is that pairs of chains can meet as soon as the first iteration, with probability at least $1/2$, which explains why all the boxplots are squashed at the value $\tau_{M,t}^{(r)} = 1$, as predicted by Proposition 8 in Middleton et al.

(2018). Figures 2.3 and 2.4 suggest that increasing N_θ in the SMC^2 proposals could benefit the estimation of the H-score for model M_2 while decreasing its associated meeting times.

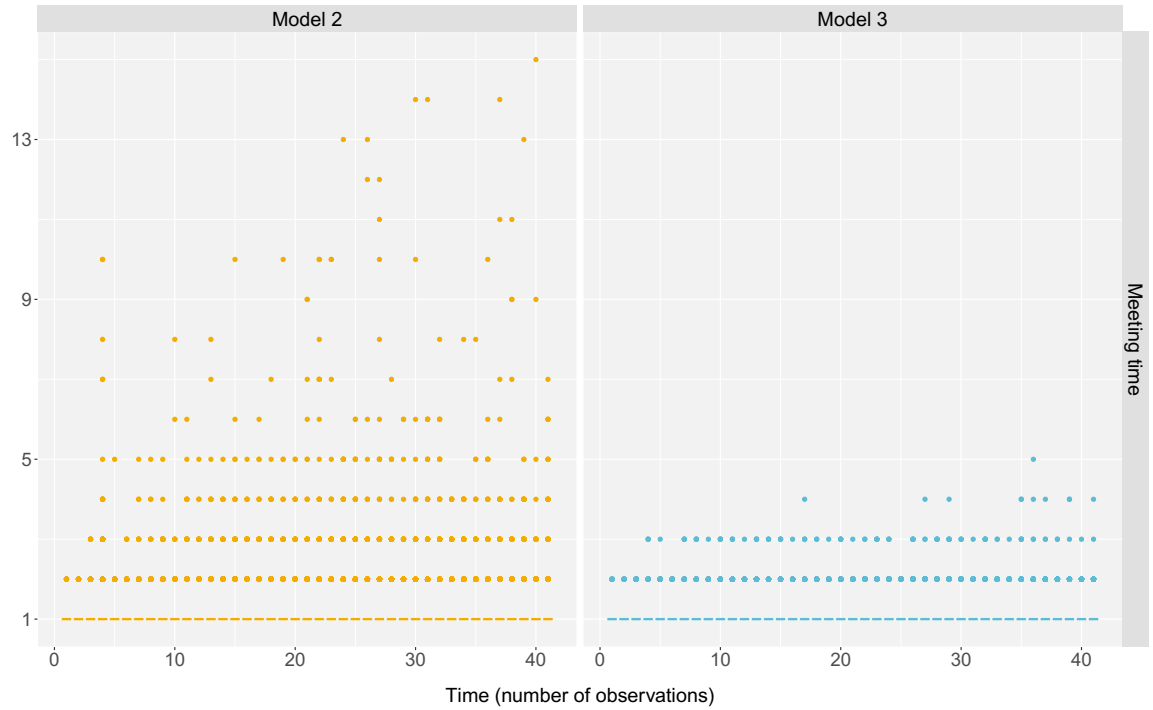

Figure 2.4. *Empirical distribution (boxplots) of the meeting times* $(\tau_{M,t}^{(r)})_{r\in[\![1,R]\!]}$ *for each model* $M \in \{M_2, M_3\}$ *(left and right panels), across* $R = 1000$ *independent coupled PIMH runs using SMC*^{2} *proposals with* $N_\theta = 8192$ *and* $N_x = 128$*, for each time horizon* $t \in [\![1, 41]\!]$*. All the boxes appear vertically squashed at* $\tau = 1$*, and the dots represent outliers whose values exceed the upper quartiles by more than 1.5 times the interquartile ranges. See Section 2.3.3.*

2.3.4 Tuning the coupled PIMH with SMC^2 proposals

As forewarned by Jacob et al. (2017, Section 3.1), the performance of unbiased MCMC methods can vary wildly depending on the choice of algorithmic parameters. This section provides some thoughts and guidelines to help tuning such parameters in the context of coupled PIMH with SMC^2 proposals.

2.3.4.1 Choice of N_θ

The choice of N_θ directly influences the quality of the particle approximations of the posteriors $p(d\theta|y_{1:t})$ from which the proposals are drawn. It also affects the variance of the SMC estimator of the evidences $p(y_{1:t})$, which directly impacts the distribution of the meeting times τ_t. For coupled PIMH algorithms, the distribution of these meeting times is surprisingly tractable (Middleton et al., 2019, Proposition 8), and can be explicitly described as a mixture consisting of

$$\begin{cases} \tau_t \;\Big|\; \hat{Z}^{(0)}_{t,N_\theta,N_x} \;\sim\; \text{Geom}\left(\mathbb{E}\left[\min\left(1, \dfrac{\hat{Z}^{(1)}_{t,N_\theta,N_x}}{\hat{Z}^{(0)}_{t,N_\theta,N_x}}\right) \;\middle|\; \hat{Z}^{(1)}_{t,N_\theta,N_x} \sim \text{SMC}^2_{N_\theta,N_x}\right]\right) \\ \hat{Z}^{(0)}_{t,N_\theta,N_x} \;\sim\; \text{SMC}^2_{N_\theta,N_x} \end{cases} \tag{2.26}$$

where $\text{Geom}(\alpha)$ denotes a geometric distribution — supported on $\mathbb{N}^*$ by convention — with probability α of success, whereas the unbiased evidence estimators $\hat{Z}^{(0)}_{t,N_\theta,N_x}$ and $\hat{Z}^{(1)}_{t,N_\theta,N_x}$ are generated independently from SMC^2 runs with (N_θ, N_x) particles. The other algorithmic choices in the implementation of SMC^2 are made implicit in the notation, and we focus mostly on the influence of N_θ and N_x on the performance of coupled PIMH algorithms.

As indicated by (2.3) in Section 2.2.1, the SMC estimators $\hat{Z}^{(0)}_{t,N_\theta,N_x}$ and $\hat{Z}^{(1)}_{t,N_\theta,N_x}$ are consistent and asymptotically Normal around the exact evidence Z_t as $N_\theta \to +\infty$, for any fixed N_x. Increasing N_θ can thus bring the acceptance probability in (2.26) arbitrarily close to 1, in the sense that $\mathbb{P}(\tau_t = 1) \xrightarrow{N_\theta \to +\infty} 1$ as stated in Proposition 8 of Middleton et al. (2019). Therefore, controlling the distribution of the meeting times τ_t can be achieved by monitoring the variance of the estimators $\hat{Z}_{t,N_\theta,N_x}$ relative to Z_t, or equivalently the variance of $\log \hat{Z}_{t,N_\theta,N_x}$ when N_θ is large enough. In practice, each of the R processor can perform one preliminary run of $\text{SMC}^2_{N_\theta,N_x}$ to produce the estimators $(\log \hat{Z}^{(r)}_{t,N_\theta,N_x})_{r\in[\![1,R]\!]}$ at each time horizon $t \in [\![1, T]\!]$. Looking at the empirical distributions of the $(\log \hat{Z}^{(r)}_{t,N_\theta,N_x})_{r\in[\![1,R]\!]}$ provides

a way to assess the relative variance of $\hat{Z}_{t,N_\theta,N_x}$ and its closeness to Normality at each time horizon $t \in [\![1, T]\!]$, as illustrated in Figure 2.5. The histograms of Figure 2.5 can be complemented by computing the sample variances across R replications at each horizon, as shown in Figure C.1 of Appendix C.2. Once these log-evidence estimators have been generated, we can preemptively assess the distribution of the meeting times τ_t at each horizon $t \in [\![1, T]\!]$ by emulating the distribution in (2.26) as follows:

- Choose a desired number M of draws, then for all $\ell \in [\![1, M]\!]$:
  - Draw r_* uniformly from $[\![1, R]\!]$.
  - Draw $\hat{\tau}_t^{(\ell)} \sim \text{Geom}\left(\frac{1}{R-1}\sum_{r \neq r_*} \min\left(1, \exp\left(\log \hat{Z}^{(r)}_{t,N_\theta,N_x} - \log \hat{Z}^{(r_*)}_{t,N_\theta,N_x}\right)\right)\right)$.
- Output $(\hat{\tau}_t^{(\ell)})_{\ell \in [\![1,M]\!]}$ as an approximate sample from the distribution of τ_t.

When the empirical distributions of $(\log \hat{Z}^{(r)}_{t,N_\theta,N_x})_{r \in [\![1,R]\!]}$ appear close enough to Normality (e.g. as in Figure 2.5), we can alternatively rely on the asymptotic Normality of log-evidence estimators to simulate approximate meeting times for each time horizon $t \in [\![1, T]\!]$ as:

- Compute the sample mean $\hat{\mu}_{t,R}$ and variance $\hat{\sigma}^2_{t,R}$ of $(\log \hat{Z}^{(r)}_{t,N_\theta,N_x})_{r \in [\![1,R]\!]}$.
- Choose desired numbers M and n of draws, then for all $\ell \in [\![1, M]\!]$:
  - Generate $(n+1)$ independent draws $(\log \tilde{Z}_t^{(i)})_{i \in [\![0,n]\!]} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\hat{\mu}_{t,R}, \hat{\sigma}^2_{t,R})$.
  - Draw $\hat{\tau}_t^{(\ell)} \sim \text{Geom}\left(\frac{1}{n}\sum_{i=1}^{n} \min\left(1, \exp\left(\log \tilde{Z}_t^{(i)} - \log \tilde{Z}_t^{(0)}\right)\right)\right)$.
- Output $(\hat{\tau}_t^{(\ell)})_{\ell \in [\![1,M]\!]}$ as an approximate sample from the distribution of τ_t.

The Normal approximations could be replaced by heavier-tailed distributions in order to obtain more conservative approximations of the tail probabilities of the meeting times. Simulated meeting times for each model and time horizon are shown in Figure 2.6. They appear in

line with the empirical meeting times observed in Figure 2.4, and suggest an exponential decay for the marginal tail probabilities of the mixture distributions from (2.26).

Solely focusing on reducing the meeting times would encourage choosing larger values of N_θ. However, doing so would quickly be prohibitive and hindered by the limited computational resources available on any given processor. Increasing N_θ would also increase the cost and computational time of each SMC^2 runs, thus allowing for fewer PIMH iterations within a budgeted time window and possibly hurting the resulting estimators. Such trade-offs have been investigated and quantified by Doucet, Pitt, Deligiannidis and Kohn (2015) in the context of pseudo-marginal algorithms, where they recommend aiming for the standard deviation of $\log \hat{Z}_{t,N_\theta,N_x}$ to be around 1 when the underlying idealized PIMH is efficient and around 1.7 otherwise. With the ability to make up for relatively inefficient PIMH estimators by using more independent processors, these guidelines do not directly apply here, and further investigation would be needed to fully understand how to choose N_θ optimally as to minimize the length of the confidence intervals for a fixed computational time budget. It is also worth mentioning that for model selection purposes, the task at hand is not so much to accurately estimate the individual H-scores of each model, but rather to correctly estimate the sign of their difference. The former task would likely require N_θ to be scaled linearly as the number T of observations increases (Del Moral et al., 2006; Chopin et al., 2013) in order to stabilize the variance of the estimated log-evidence and maintain the quality of the consequent coupled PIMH estimators, whereas the latter task — assuming consistency of H-scores as in Theorems 1 and 2 of Chapter 1 — might allow for a more lenient scaling since the magnitude of the difference of H-scores also tends to grow linearly with the number of observations, making the estimation of its sign presumably easier. Quantifying these interplays more precisely could be an interesting direction to explore.

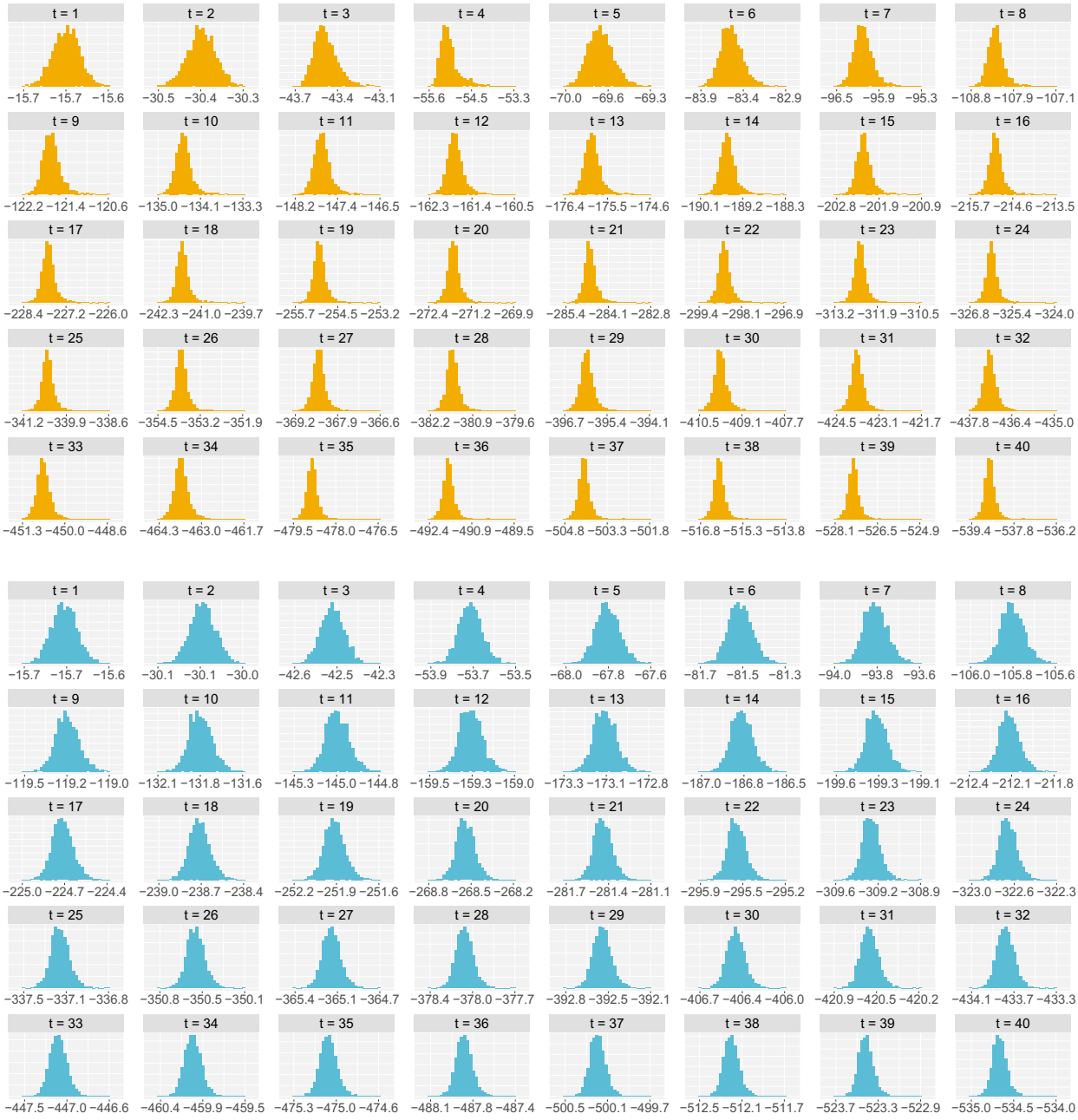

Figure 2.5. *Histograms of the estimated log-evidences* $\log \hat{Z}_t$ *from* $R = 1000$ *independent runs of SMC*^{2} *with* $N_\theta = 8192$ *and* $N_x = 128$. *Each subplot corresponds to a particular time horizon* $t \in [\![1, 40]\!]$. *The histogram for horizon* $t = 41$ *is similar to the one for* $t = 40$ *and is not shown here so as to make the layout more convenient. The y-axis representing counts are irrelevant and omitted for better readability. The top five rows (yellow) correspond to model* M_2, *and the bottom five rows (blue) to model* M_3. *See Section 2.3.4.1.*

2.3.4.2 Choice of N_x

Although SMC^2 algorithms are valid SMC samplers for any fixed N_x, the choice of N_x presents similar variance-cost trade-offs as the choice of N_θ. The necessity of unbiasedness for the evidence estimators $\hat{Z}_{t,N_\theta,N_x}$ prevents the direct use of automatic calibration procedures such as the ones described in Section 3.6 of Chopin et al. (2013), although one could envision using a preliminary run with adaptive N_x in order to heuristically guide the final choice of fixed schedule for N_x. Fewer theoretical results exist on the impact of N_x on the estimators $\hat{Z}_{t,N_\theta,N_x}$, and an interesting avenue of research would be to derive approximate finite sample bounds for the variance of $\hat{Z}_{t,N_\theta,N_x}$ as a function of both N_θ and N_x, in the spirit of Cérou, Del Moral and Guyader (2011). Such bounds would then need to be further combined with the additional computational considerations raised by coupled PIMH algorithms.

2.3.4.3 Choice of k and m

As extensively discussed by Jacob et al. (2017), a well-thought choice of k and m can lead to drastic efficiency gains for the estimator in (2.13). The general guidelines in Section 3.1 of Jacob et al. (2017) propose to choose k as a large quantile of the distribution of the meeting time τ, so that $\mathbb{P}(\tau > k)$ is small. The value of m is then set to a large multiple of k, typically $m = 10k$. Under appropriate regularity conditions, these heuristic rules ensure that the efficiency of the coupled PIMH is similar to its non-coupled counterpart, so that debiasing PIMH estimators can be performed at virtually no loss of efficiency. The tuning of (k, m) for general unbiased MCMC methods traditionally requires one to perform many replications of the full coupled MCMC algorithm in order to obtain reliable empirical distributions of the meeting times, which are then used to guide the final choice of (k, m) for the subsequent coupled MCMC replications. Interestingly, the coupled PIMH algorithm enjoys the remarkable property of having meeting times whose distributions can be conveniently emulated. As explained in Section 2.3.4.1 and shown in Figure 2.6, the distribution of the

meeting times can be approximately simulated by only performing one preliminary run of SMC^2 on each processor, instead of full runs of coupled PIMH with SMC^2 proposals. The simulated distributions of the meeting times can then inform the choice of k and m.

2.3.4.4 Budget allocation

Many practical settings consist in having access to a large number R of independent processors, each with its own relatively limited computational resources, and a global time budget to comply with. This requires us to consider a notion of efficiency that balances the variances of estimators with their expected costs, as in the framework of Glynn and Whitt (1992*a*). Such considerations are non-trivial: smaller values of N_θ and N_x would lead to estimators with larger variances, but these estimators would be cheaper and faster to compute by independent processors, thus allowing for more independent replications. This could lead to averaged estimators with smaller variances than their hypothetical counterparts formed by averaging fewer more accurate estimators, within the same fixed computational time window.

In practice, the computational time of coupled PIMH can vary significantly across different processors, due to the randomness of the algorithm and possible hardware disparities. This is conspicuous from Figure 2.7, which shows the total wall-clock times used by each of the $R = 1000$ processors to complete a full run of coupled PIMH with SMC^2 proposals using $N_\theta = 8192$, $N_x = 128$, $k = 5$, and $m = 10$. Given a fixed time budget, a better use of resources would consist in allowing each processor to keep starting new chains and producing new estimators within the allocated time window. Some precautions are in order when aggregating estimators from algorithms with random computation times within a fixed time frame, as highlighted in Glynn and Heidelberger (1990), whose Corollary 7 offers a practical solution to perform a statistically valid aggregation. Computations of unbiased MCMC estimators under such time budget constraints are demonstrated in Section 3.2.2 and Figure 9 of Middleton et al. (2018). Similar considerations can be applied at the level of each

processor, since the distributions of meeting times in Figure 2.4 appear to vary greatly across different time horizons. All things considered, the task of tuning (N_θ, N_x, k, m) to optimize the efficiency of the coupled PIMH in a budgeted time setting is challenging but of great practical importance, and would deserve further dedicated research.

2.4 Discussion

The construction of confidence intervals using (2.11) and (2.20), respectively for the log-evidence and the H-score, only requires modest implementation efforts. The former can be constructed as soon as one has access to a working SMC sampler, which generally accommodates a wide range of models whose likelihoods can be either evaluated or unbiasedly estimated. For the latter, the incremental effort needed to turn an SMC sampler into the proposal distribution of a coupled PIMH algorithm is minimal, as it only requires the addition of an external layer of coupled Metropolis-Hastings acceptance step, without any alterations to the existing SMC implementation. Although the proposed algorithms are arguably easy to implement, the question of using them to the best of their abilities remains open, and further investigation would be needed to fully understand how to optimally tune their algorithmic parameters. An alternative to the coupled PIMH used to form the interval in (2.20) would be to use coupled particle marginal Metropolis-Hastings algorithms (Middleton et al., 2018), whose implementation would require a more involved coupling procedure but whose performance would be worth exploring.

Both the proposed confidence intervals from (2.11) and (2.20) enjoy an asymptotic validity in the number R of independent replications. Their construction is thus particularly suited for parallel computing on modern architectures made of many processors with limited individual resources. The ability to refine these intervals by generating more independent replicates is well-adapted to settings with budgeted computational time windows. It also opens the door

to sequential model selection procedures: starting from a computationally cheap confidence interval for the difference in scores, further independent replications would be provided on demand until the updated confidence interval is refined enough to distinguish between the two models of interest. Such stopping rules would need to be carefully designed so as to avoid the common pitfalls of sequential hypothesis testing (e.g. Glynn and Whitt, 1992*b*, and references therein). Generalizing our approach from 2 to a large number K of candidate models would require further research, since naive comparisons of the pairwise differences in scores would suffer from multiple testing issues. By regarding the negative of the models' scores as expected rewards for which we can draw unbiased i.i.d. estimators, we could envision recasting the task sequential model selection as a best-arm identification problem in a multi-arm bandit setting with K arms (e.g. Kaufmann, Cappé and Garivier, 2016, and references therein).

There are fundamentally two sources of uncertainty when using the log-evidence or the H-score to compare models M_1 and M_2 given realized observations $y_{1:T}$ from some unknown data generating process $p_\star$. The first source of uncertainty lies in the Monte Carlo error induced by using estimated log-Bayes factors $\hat{\mathrm{LBF}}_T^{1|2}$ or H-factors $\hat{\mathrm{H}}_T^{1|2}$ as proxies for the exact but unavailable log-Bayes factors $\mathrm{LBF}_T^{1|2}$ or H-factors $\mathrm{H}_T^{1|2}$. Quantifying this uncertainty was at the heart of this chapter, and we addressed this issue by constructing confidence intervals for the exact $\mathrm{LBF}_T^{1|2}$ and $\mathrm{H}_T^{1|2}$. In the idealized case where $\mathrm{LBF}_T^{1|2}$ or $\mathrm{H}_T^{1|2}$ can be exactly computed, choosing the model with the smallest realized score constitutes a sensible selection rule, as supported by decision theory.

However, if the interest lies in future predictions instead, one would then have to look at expected scores with respect to the unknown data generating distribution $p_\star$. In such a setting, the second source of uncertainty concerns the sampling error induced by using the realized differences in scores $\mathrm{LBF}_T^{1|2}$ or $\mathrm{H}_T^{1|2}$ in place of the exact but unavailable expected differences, typically given by the differences in divergences $\Delta_{\mathrm{KL}}(M_1, M_2) = \mathrm{KL}(p_\star, M_2) - \mathrm{KL}(p_\star, M_1)$ and $\Delta_{\mathrm{D}_\mathcal{H}}(M_1, M_2) = \mathrm{D}_\mathcal{H}(p_\star, M_2) - \mathrm{D}_\mathcal{H}(p_\star, M_1)$. Assuming perfect access to $\mathrm{LBF}_T^{1|2}$ and

$\mathrm{H}_T^{1|2}$, this second kind of uncertainty would pertain to questions about how large $|\mathrm{LBF}_T^{1|2}|$ or $|\mathrm{H}_T^{1|2}|$ need to be in order to be regarded as statistically significant with respect to the sampling distribution of the observations $y_{1:T}$. Such concerns touch upon the need of assessing frequentist properties of procedures based on $\mathrm{LBF}_T^{1|2}$ and $\mathrm{H}_T^{1|2}$. Specific scales to interpret Bayes factors have been previously provided in the literature, including an arguably questionable one by Jeffreys (1961, Appendix B) that was later revisited by Kass and Raftery (1995, Section 3.2). More recent considerations rely on studying the asymptotic distribution of log-Bayes factors as the sample size increases (e.g. Vuong, 1989; Gelfand and Dey, 1994; Walker, 2004; Zhou and Guan, 2018). The lack of corresponding results for the H-score leaves room for interesting topics of research. Quantifying that second kind of uncertainty requires a good understanding of the sampling distribution of $\mathrm{LBF}_T^{1|2}$ and $\mathrm{H}_T^{1|2}$, which is still elusive even in large sample regimes. Consistency results of the form $T^{-1}\,\mathrm{LBF}_T^{1|2} \longrightarrow \Delta_{\mathrm{KL}}(M_1, M_2)$ and $T^{-1}\,\mathrm{H}_T^{1|2} \longrightarrow \Delta_{\mathrm{D}_{\mathcal{H}}}(M_1, M_2)$ only provide first-order asymptotics, without any information about the sampling fluctuations around the respective limits. Testing the significance of the difference in scores would rather require second-order asymptotics to provide more precise distributional statements, presumably for quantities of the form $\sqrt{T}\,\left(T^{-1}\mathrm{LBF}_T^{1|2} - \Delta_{\mathrm{KL}}(M_1, M_2)\right)$ and $\sqrt{T}\,\left(T^{-1}\mathrm{H}_T^{1|2} - \Delta_{\mathrm{D}_{\mathcal{H}}}(M_1, M_2)\right)$ by analogy with standard central limit theorems. As suggested by our proofs of the consistency of H-scores in Appendix B.5.1, first-order type of results typically rely on the consistency of the posterior distribution, whereas second-order asymptotics tend to require its asymptotic Normality, none of which has been established yet in the context of general state-space models, to the best of our knowledge. Studying the consistency and asymptotic Normality of posterior distributions for possibly misspecified state-space models will be the focus of Chapter 3.

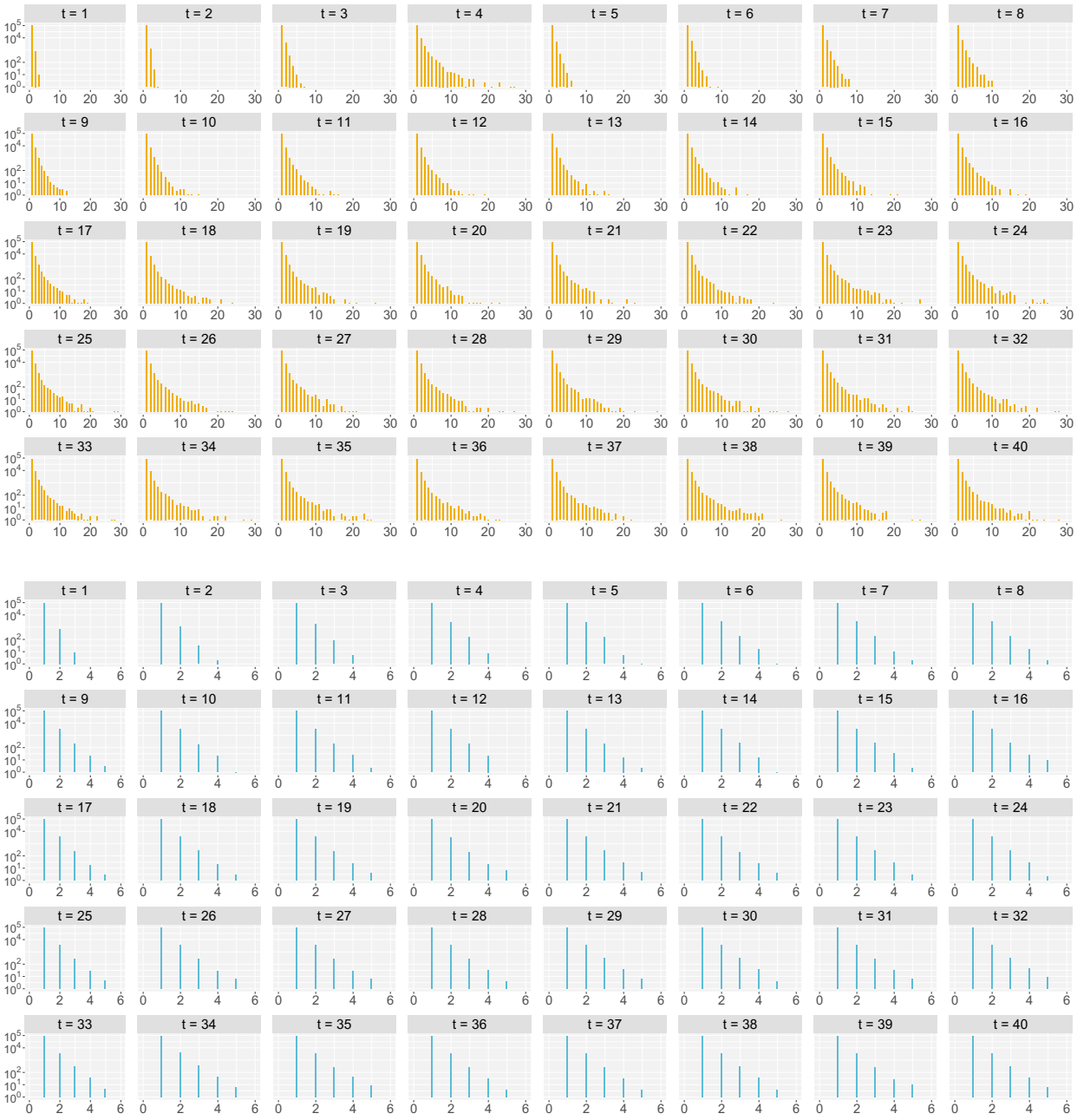

Figure 2.6. *Histograms of $M = 10^5$ simulated meeting times $(\hat{\tau}_t^{(\ell)})_{\ell \in [\![1,M]\!]}$ of coupled PIMH chains using SMC^2 proposals with $N_\theta = 8192$ and $N_x = 128$, for each horizon $t \in [\![1, 40]\!]$. The histogram for horizon $t = 41$ is similar to the one for $t = 40$ and is not shown here so as to make the layout more convenient. The top five rows (yellow) correspond to model M_2, and the bottom five rows (blue) to model M_3. A logarithmic scale in base 10 is used on the y-axis for better readability, and the displayed counts are only relevant to the extent that they suggest an exponential decay of the meeting times' tail probabilities. See Section 2.3.4.1.*

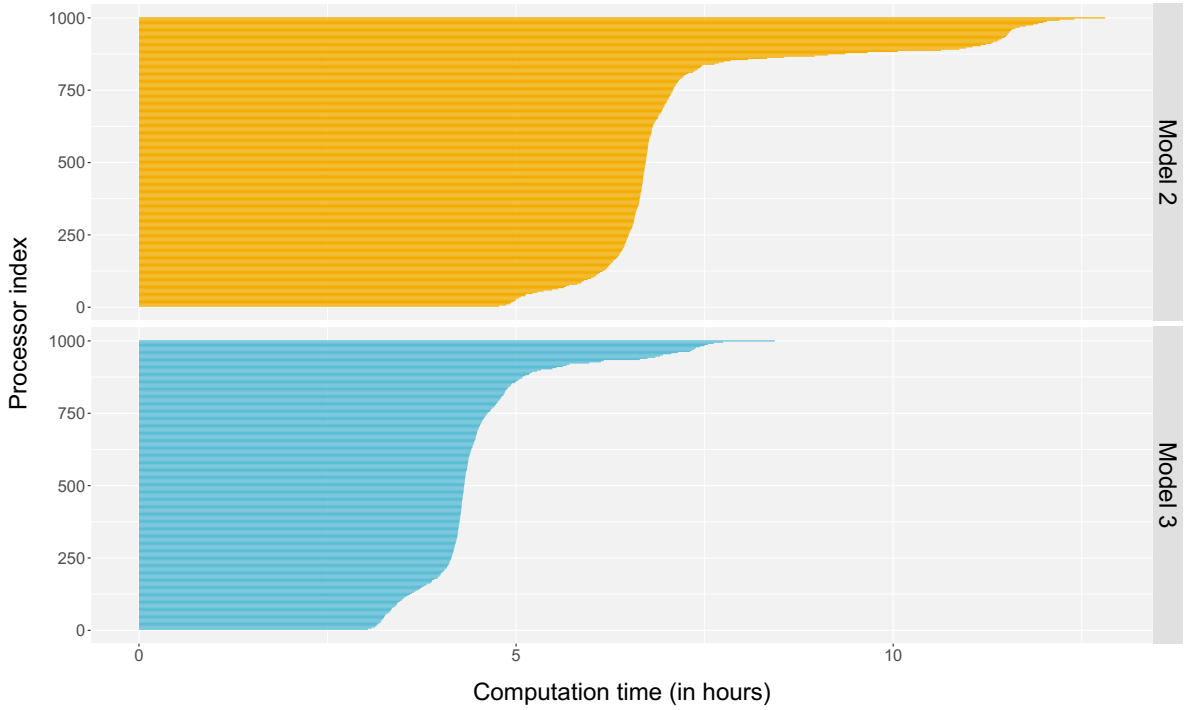

Figure 2.7. *Total wall-clock computation times across* $R = 1000$ *processors (artificially sorted in increasing order for better readability) to perform a coupled PIMH using SMC*^{2} *proposals with* $N_\theta = 8192$, $N_x = 128$, $k = 5$, *and* $m = 10$ *for models* M_2 *(top panel) and* M_3 *(bottom panel).*

“ *On devrait vivre a posteriori.*
On décide tout trop tôt. ”

— Daniel Pennac, *Aux fruits de la passion*

3

Posterior consistency and asymptotic Normality in state-space models

Contents

The work in this chapter has been done in collaboration with Judith Rousseau (*Department of Statistics*, Oxford University), Arnaud Doucet (*Department of Statistics*, Oxford University), and Pierre E. Jacob (*Department of Statistics*, Harvard University). It corresponds to an upcoming paper at the time of writing.

3.1 Introduction

Understanding the asymptotic behavior of standard statistical methods in the context of state-space models has proved to be a challenging problem, mostly due to the dependence between observations and the intractability of the likelihood function. Some major advances have been obtained on the asymptotic behavior of the maximum likelihood estimation methods. This line of research was initiated by Baum and Petrie (1966) in the case of well-specified models with a finite space $\mathbb{X}$ of latent states and a finite space $\mathbb{Y}$ of observations. The consistency of the maximum likelihood estimator (MLE) was later extended by Leroux (1992) to the case of finite state spaces $\mathbb{X}$ and general observation spaces $\mathbb{Y}$, while the asymptotic Normality of the MLE in that setting was derived in Bickel, Ritov and Ryden (1998). A series of related works under increasingly weaker assumptions followed, notably by Douc and Matias (2001) and Douc, Moulines and Ryden (2004) who proved the consistency and asymptotic Normality of the MLE under conditions that essentially amount to having a compact latent space $\mathbb{X}$. Generalizations of these results to non-compact latent spaces and misspecified models were ultimately derived by Douc, Moulines, Olsson and Van Handel (2011) and Douc and Moulines (2012). A comprehensive review on the subject can be found in Douc et al. (2014).

On the Bayesian side of things, the more challenging analysis of the asymptotic properties of posterior distributions in state-space models has been much less studied to this day, despite its importance. Understanding the asymptotic behavior of posterior distributions is paramount to establishing desirable properties of other Bayesian methods, such as the

consistency of model selection criteria like the BIC (Schwartz, 1978; Yonekura, Beskos and Singh, 2018) or the H-score (Section 1.3.2 of Chapter 1). The consistency of the posterior distribution and its asymptotic Normality — via results analogous to the Bernstein-von-Mises theorem — also play a key role in justifying practical implementation choices (e.g. designing proposal distributions), while ensuring the stability and validity of various numerical methods including SMC (Chopin, 2002; Whiteley, 2013; Chopin et al., 2013) and variational inference (Wang and Blei, 2018, and references therein). Such traditional results — which have been well-established in a variety of settings, including parametric models with i.i.d. observations (e.g. Van der Vaart, 2000; Ghosh and Ramamoorthi, 2003; Lehmann and Casella, 2006; Le Cam and Yang, 2012), as well as semi-parametric (e.g. Kim, 2006; De Blasi and Hjort, 2009; Rivoirard and Rousseau, 2012; Bickel and Kleijn, 2012; Castillo, 2012; Castillo and Nickl, 2014; Castillo and Rousseau, 2015) or nonparametric (Freedman, 1999; Kim and Lee, 2004; Boucheron and Gassiat, 2009; Johnstone, 2010; Bontemps, 2011; Castillo and Nickl, 2013; Ghosal and Van der Vaart, 2017) models — are still lacking in the context of general state-space models. On the one hand, a version of the Bernstein-von-Mises theorem for state-space models was derived by De Gunst and Shcherbakova (2008), albeit under strong mixing conditions that essentially assume the compactness of the latent space $\mathbb{X}$, and that are thus hardly met in practice even for simple linear Gaussian state-space models. On the other hand, posterior consistency in general spaces was obtained by Douc et al. (2019) for the larger class of partially observed Markov models, but only in a well-specified setting.

Our goal is to establish the consistency and asymptotic Normality of posterior distributions for state-space models, under assumptions that are as weak as the ones from Douc and Moulines (2012), while also allowing for model misspecification. Although this chapter falls short of presenting a complete proof at the time of writing, it provides some directions to explore, and highlights the main obstacles to overcome when studying Bayesian asymptotics in such a setting. In Section 3.2, we present key ingredients that ensure the consistency and

asymptotic Normality of posterior distributions. These ingredients are stated as high-level assumptions, under which posterior consistency and asymptotic Normality can be readily proved, as respectively shown in Sections 3.3 and 3.4. The main difficulty is to determine reasonable sufficient conditions on the state-space model for these high-level assumptions to hold, as discussed in Section 3.5. By relying on existing results from the previously cited works, the only problematic assumption is the uniform convergence of the observed Fisher information around the limit of the MLE, which we investigate in Section 3.6. The consistency and asymptotic Normality of posterior distributions are numerically illustrated in Section 3.7, using the population dynamics and stochastic volatility models introduced in Chapter 1. Shortcomings and directions for future research are discussed in Section 3.8.

3.2 Fundamental ingredients

We borrow some of the notations introduced in the previous chapters, letting $\mathbb{P}_\star$ (resp. $\mathbb{E}_\star$) denote the probability (resp. expectation) induced by the data-generating mechanism of the observed stochastic process $(Y_n)_{t\in\mathbb{N}}$. When the context is clear, we will respectively denote the log-likelihood function and posterior density based on n observations by $\ell_n(\theta)$ and $\pi(\cdot|Y_{0:n-1})$, without making the dependence on the initial distribution of the latent states explicit. The asymptotic irrelevance of that initial distribution will be justified in Section 3.6, by invoking suitable forgetting properties of the latent Markov chain. We now present a few high-level assumptions (A1 to A6) from which the consistency and asymptotic Normality of the posterior distribution naturally follow (Sections 3.3 and 3.4). The main difficulty will be to ensure that such assumptions hold, by imposing only weak conditions on the underlying state-space model (Section 3.5).

Assumption A1. *The parameter space $\mathbb{T} \subseteq \mathbb{R}^d$ is a compact metric space, and the prior distribution Π has a density π with respect to the Lebesgue measure.*

Assumption A2 (Convergence of the log-likelihood)**.** *For all $\theta \in \mathbb{T}$, we have*

$$\frac{1}{n}\ell_n(\theta) \xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.} \ell(\theta)$$

where the limit function $\theta \mapsto \ell(\theta) = \mathbb{E}_\star\left[\log p_\theta(Y_0|Y_{-\infty:-1})\right]$ is assumed to be well-defined and upper semi-continuous on $\mathbb{T}$.

Assumption A3 (Existence and uniqueness of $\theta^\star$)**.** *The limit function $\theta \mapsto \ell(\theta)$ from Assumption A2 is uniquely maximized on $\mathbb{T}$, i.e. there exists $\theta^\star \in \mathbb{T}$ such that*

$$\arg\max_{\theta\in\mathbb{T}} \ell(\theta) \;=\; \{\theta^\star\}.$$

Assumption A4 (Consistency of the MLE)**.** *$\mathbb{P}_\star$-almost surely, the maximum likelihood estimator $\hat{\theta}_n$ is well-defined for all $n \in \mathbb{N}^*$ and is strongly consistent, i.e.*

$$\|\hat{\theta}_n - \theta^\star\| \xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.} 0.$$

Assumption A5 (Well-separation of $\theta^\star$)**.** *$\mathbb{P}_\star$-almost surely, for all $\delta > 0$, we have*

$$\limsup_{n\to+\infty} \left(\sup_{\|\theta-\theta^\star\|\geq\delta} \frac{1}{n}\ell_n(\theta) \right) \;<\; \ell(\theta^\star).$$

Assumption A6 (Uniform convergence of the observed Fisher information around $\theta^\star$)**.** *For all $n \in \mathbb{N}^*$, the log-likelihood function $\theta \mapsto \ell_n(\theta)$ is twice continuously differentiable in the neighborhood of $\theta^\star$, and there exists a symmetric matrix $J^\star > 0$ such that $\mathbb{P}_\star$-almost surely, for all $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ and $N_\varepsilon \in \mathbb{N}$ for which*

$$\sup_{\substack{\|\theta-\theta^\star\|<\delta_\varepsilon \\ n>N_\varepsilon}} \left\| \frac{1}{n}\nabla^2\ell_n(\theta) + J^\star \right\| < \varepsilon.$$

Remark 4. *Most existing proofs of Bernstein-von Mises type of results in the context of state-space models (e.g. De Gunst and Shcherbakova, 2008) tend to use a stronger version of Assumption A2, requiring the convergence of the log-likelihood $n^{-1}\ell_n$ to be uniform over*

all compact subsets of $\mathbb{T}$, thus inducing the continuity of the limit function ℓ. Our proof gets away with weaker pointwise convergence of $n^{-1}\ell_n$ and upper semi-continuity of ℓ by exploiting instead a control of the Hessian $n^{-1}\nabla^2\ell_n$ in the neighborhood of $\theta^\star$, given by Assumption A6, which we aim to ensure under conditions similar to the ones in Douc and Moulines (2012).

Remark 5. *Assumption A6 is likely to be unnecessarily strong for our purposes. In the spirit of Theorem 3 from Douc et al. (2004), we might only ask for the less demanding assumption that for any sequence $(\tilde{\theta}_n)_{n\in\mathbb{N}^*} \in \mathbb{T}^{\mathbb{N}^*}$, if $\tilde{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}_\star-a.s.} \theta^\star$, then $-n^{-1}\nabla^2\ell_n(\tilde{\theta}_n) \xrightarrow[n\to+\infty]{\mathbb{P}_\star-a.s.} J^\star$. We will refer to this weaker version of A6 as Assumption $\tilde{A}$6.*

Before proving the consistency of the posterior distribution (Theorem 11) and its asymptotic Normality (Theorem 12), we start by proving some technical consequences of Assumptions A1 to A6, including the $\mathbb{P}_\star$-almost sure asymptotic equicontinuity of $\{\theta \mapsto n^{-1}\ell_n(\theta)\}_{n\in\mathbb{N}^*}$ at $\theta^\star$ (Lemma 6), the convergence of $n^{-1}\ell_n(\hat{\theta}_n)$ to $\ell(\theta^\star)$ (Lemma 7), the well-separation of the MLE (Lemma 8), and the convergence of $-n^{-1}\nabla^2\ell_n(\hat{\theta}_n)$ to $J^\star$ (Lemma 10). The proofs of these results are presented in Appendix D.

Lemma 6 (Asymptotic equicontinuity of $n^{-1}\ell_n$ at $\theta^\star$)**.** *If Assumptions A2 to A6 hold, then $\mathbb{P}_\star$-almost surely, for all $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ and $N_\varepsilon \in \mathbb{N}$ such that*

$$\sup_{\substack{\|\theta-\theta^\star\|<\delta_\varepsilon \\ n>N_\varepsilon}} \left|\frac{1}{n}\ell_n(\theta) - \frac{1}{n}\ell_n(\theta^\star)\right| < \varepsilon.$$

Lemma 6 combined with Assumptions A2 and A4 directly implies the $\mathbb{P}_\star$-almost sure convergence of $n^{-1}\ell_n(\hat{\theta}_n)$ to $\ell(\theta^\star)$, as stated by the following lemma.

Lemma 7 (Convergence of $n^{-1}\ell_n(\hat{\theta}_n)$ to $\ell(\theta^\star)$)**.** *If Assumptions A2 to A6 hold, then $\mathbb{P}_\star$-almost surely, for all $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that*

$$\sup_{n>N_\varepsilon} \left|\frac{1}{n}\ell_n(\hat{\theta}_n) - \ell(\theta^\star)\right| < \varepsilon.$$

The following lemma will come in handy when proving of the asymptotic Normality of the posterior distribution (Theorem 12), as it allows for a control over the behavior of the log-likelihood function far from the MLE.

Lemma 8 (Well-separation of the MLE)**.** *If Assumptions A2 to A6 hold, then $\mathbb{P}_\star$-almost surely, for all $\delta > 0$, there exists $\eta_\delta > 0$ and $N_\delta \in \mathbb{N}$ such that, for all $n > N_\delta$, we have*

$$\sup_{\|\theta - \hat{\theta}_n\| \geq \delta} \frac{1}{n}\ell_n(\theta) \; < \; \frac{1}{n}\ell_n(\hat{\theta}_n) - \eta_\delta.$$

Remark 9. *Lemma 8 is a $\mathbb{P}_\star$-almost sure version of Lemma 3.1 from De Gunst and Shcherbakova (2008). However, the proof from De Gunst and Shcherbakova (2008) cannot directly be used since it assumes that the convergence in Assumption A2 is uniform on $\mathbb{T}$, which consequently ensures the continuity of the limit function $\theta \mapsto \ell(\theta)$. Nevertheless, the result can still be proved in our setting by relying on Assumption A6 instead.*

Finally, the next lemma will play an important role as it establishes the convergence of $-n^{-1}\nabla^2\ell_n(\hat{\theta}_n)$ to $J^\star$, by exploiting the uniform control of the observed Fisher information around $\theta^\star$ provided by Assumption A6.

Lemma 10 (Convergence of $-n^{-1}\nabla^2\ell_n(\hat{\theta}_n)$)**.** *If Assumptions A4 and A6 hold, then $\mathbb{P}_\star$-almost surely, for all $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that*

$$\sup_{n > N_\varepsilon} \left| \frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n) + J^\star \right| < \varepsilon.$$

where $J^\star$ is the positive definite symmetric matrix introduced in Assumption A6.

3.3 Consistency of the posterior

Using the previous assumptions and Lemma 6, we can now prove the $\mathbb{P}_\star$-almost sure consistency of the posterior distribution, as stated by the following theorem.

Theorem 11 (Posterior consistency). *Under Assumptions A1 to A6, if the prior distribution* Π *satisfies* $\Pi(\|\theta-\theta^\star\|<\varepsilon)>0$ *for all* $\varepsilon>0$*, then for all neighborhoods* $\mathcal{U}\subseteq\mathbb{T}$ *of* $\theta^\star$*, we have*

$$\Pi(\mathcal{U}\,|\,Y_{0:n-1}) \xrightarrow[n\to+\infty]{\mathbb{P}_\star-a.s.} 1.$$

Proof of Theorem 11. Let $\mathcal{U}\subseteq\mathbb{T}$ be an arbitrary neighborhood of $\theta^\star$. By our assumption on the prior Π with density π, we have $\int_{\mathcal{U}}\exp\left(\ell_n(\theta)\right)\pi(\theta)d\theta>0$, so that we may write

$$\Pi(\mathcal{U}\,|\,Y_{0:n-1}) \;=\; \frac{\int_{\mathcal{U}}\exp\left(n\left(\frac{1}{n}\ell_n(\theta)\right)\right)\pi(\theta)d\theta}{\int_{\mathbb{T}}\exp\left(n\left(\frac{1}{n}\ell_n(\theta)\right)\right)\pi(\theta)d\theta} \;=\; \frac{1}{1+\frac{\int_{\mathbb{T}\cap\mathcal{U}^c}\exp\left(n\left(\frac{1}{n}\ell_n(\theta)\right)\right)\pi(\theta)d\theta}{\int_{\mathcal{U}}\exp\left(n\left(\frac{1}{n}\ell_n(\theta)\right)\right)\pi(\theta)d\theta}}\,.$$

For any $\delta>0$, define $\mathcal{U}_\delta=\{\theta\in\mathbb{T}:\|\theta-\theta^\star\|<\delta\}$.

Since $\mathcal{U}$ is a neighborhood of $\theta^\star$, there exists $\delta^\star>0$ such that $\mathcal{U}_\delta\subseteq\mathcal{U}$ for all $\delta\in(0,\delta^\star]$. The non-negativity of $\theta\mapsto\exp\left(\ell_n(\theta)\right)\pi(\theta)$ implies that

$$\Pi(\mathcal{U}\,|\,Y_{0:n-1}) \;\geq\; \frac{1}{1+\frac{\int_{\mathbb{T}\cap\mathcal{U}^c}\exp\left(n\left(\frac{1}{n}\ell_n(\theta)\right)\right)\pi(\theta)d\theta}{\int_{\mathcal{U}_\delta}\exp\left(n\left(\frac{1}{n}\ell_n(\theta)\right)\right)\pi(\theta)d\theta}} \tag{3.1}$$

for all $0<\delta\leq\delta^\star$. As a probability, we also naturally have $\Pi(\mathcal{U}\,|\,Y_{0:n-1})\leq 1$.

All we need now is to choose a particular $\delta\in(0,\delta^\star]$, for which we will prove that

$$\frac{\int_{\mathbb{T}\cap\mathcal{U}^c}\exp\left(n\left(\frac{1}{n}\ell_n(\theta)\right)\right)\pi(\theta)d\theta}{\int_{\mathcal{U}_\delta}\exp\left(n\left(\frac{1}{n}\ell_n(\theta)\right)\right)\pi(\theta)d\theta} \xrightarrow[n\to+\infty]{\mathbb{P}_\star-a.s.} 0. \tag{3.2}$$

This will be achieved by suitably upper-bounding the numerator and lower-bounding the denominator. We place ourselves on an intersection of events of $\mathbb{P}_\star$-probability equal to 1, on which Assumptions A2 to A6 and Lemma 6 simultaneously hold.

For the numerator, we have $\mathbb{T}\cap\mathcal{U}^c\subseteq\{\|\theta-\theta^\star\|\geq\delta^\star\}$, so by Assumption A5, there exists $\eta^\star>0$ and $N^\star\in\mathbb{N}$ such that we have, for all $n>N^\star$,

$$\sup_{\theta\in\mathbb{T}\cap\mathcal{U}^c}\frac{1}{n}\ell_n(\theta) \;\leq\; \sup_{\|\theta-\theta^\star\|\geq\delta^\star}\frac{1}{n}\ell_n(\theta) \;<\; \ell(\theta^\star)-2\eta^\star. \tag{3.3}$$

For the denominator, Assumption A2 and Lemma 6 guarantee the existence of some $\delta_{\eta^\star} > 0$ and $N_{\eta^\star} \in \mathbb{N}$ such that, for all positive $\delta < \delta_{\eta^\star}$ and all $n > N_{\eta^\star}$, we have both

$$\sup_{\theta \in \mathcal{U}_\delta} \left| \frac{1}{n}\ell_n(\theta) - \frac{1}{n}\ell_n(\theta^\star) \right| \leq \sup_{\|\theta - \theta^\star\| < \delta_{\eta^\star}} \left| \frac{1}{n}\ell_n(\theta) - \frac{1}{n}\ell_n(\theta^\star) \right| < \frac{\eta^\star}{2} \tag{3.4}$$

and

$$\left| \frac{1}{n}\ell_n(\theta^\star) - \ell(\theta^\star) \right| < \frac{\eta^\star}{2}. \tag{3.5}$$

Let $\delta = \min(\delta_{\eta^\star}, \delta^\star) \in (0, \delta^\star]$ and $N = \max(N_{\eta^\star}, N^\star)$. Then, the triangle inequality combined with (3.4) and (3.5) gives us, for all $\theta \in \mathcal{U}_\delta$ and all $n > N \geq N_{\eta^\star}$,

$$\begin{aligned}
\left| \frac{1}{n}\ell_n(\theta) - \ell(\theta^\star) \right| &\leq \left| \frac{1}{n}\ell_n(\theta) - \frac{1}{n}\ell_n(\theta^\star) \right| + \left| \frac{1}{n}\ell_n(\theta^\star) - \ell(\theta^\star) \right| \\
&\leq \sup_{\theta \in \mathcal{U}_\delta} \left| \frac{1}{n}\ell_n(\theta) - \frac{1}{n}\ell_n(\theta^\star) \right| + \left| \frac{1}{n}\ell_n(\theta^\star) - \ell(\theta^\star) \right| \\
&< \eta^\star
\end{aligned}$$

so that, for all $n > N$, we have

$$\inf_{\theta \in \mathcal{U}_\delta} \frac{1}{n}\ell_n(\theta) > \ell(\theta^\star) - \eta^\star. \tag{3.6}$$

To summarize, equations (3.3) and (3.6) ensure that, for all $n > N$, we have

$$\sup_{\theta \in \mathbb{T} \cap \mathcal{U}^c} \frac{1}{n}\ell_n(\theta) < \ell(\theta^\star) - 2\eta^\star < \ell(\theta^\star) - \eta^\star < \inf_{\theta \in \mathcal{U}_\delta} \frac{1}{n}\ell_n(\theta). \tag{3.7}$$

Using (3.7), we may now prove (3.2). For all $n > N$, we have

$$0 \leq \frac{\int_{\mathbb{T} \cap \mathcal{U}^c} \exp\left(n\left(\frac{1}{n}\ell_n(\theta)\right)\right)\pi(\theta)d\theta}{\int_{\mathcal{U}_\delta} \exp\left(n\left(\frac{1}{n}\ell_n(\theta)\right)\right)\pi(\theta)d\theta} \leq \frac{e^{(n(\ell(\theta^\star) - 2\eta^\star))}}{e^{(n(\ell(\theta^\star) - \eta^\star))}} \frac{\Pi(\mathbb{T} \cap \mathcal{U}^c)}{\Pi(\mathcal{U}_\delta)} \leq e^{-n\eta^\star} \frac{\Pi(\mathbb{T} \cap \mathcal{U}^c)}{\Pi(\mathcal{U}_\delta)} \xrightarrow[n \to +\infty]{} 0.$$

Plugging this back into (3.1) finally yields

$$\Pi(\mathcal{U} \,|\, Y_{0:n-1}) \xrightarrow[n \to +\infty]{\mathbb{P}_\star - a.s.} 1$$

which concludes the proof of Theorem 11. □

3.4 Asymptotic Normality of the posterior

In order to prove the asymptotic Normality of the posterior (Theorem 12), we make an additional assumption about the continuity of the prior density at $\theta^\star$ (Assumption A7).

Assumption A7 (Continuity and positivity of the prior density)**.** *The prior density π is continuous and strictly positive in the neighborhood of $\theta^\star$.*

Theorem 12 (Asymptotic Normality of the posterior distribution)**.** *Let $\hat{\theta}_n$ and θ_n respectively denote the MLE and a random variable following the posterior distribution with n observations $Y_{0:n-1} \in \mathbb{Y}^n$. Let $h \mapsto \pi(h \,|\, Y_{0:n-1})$ conveniently denote the posterior density on $\mathbb{R}^d$ of the random variable $h = \sqrt{n}(\theta_n - \hat{\theta}_n)$. If Assumptions A1 to A7 hold, then*

$$\int_{\mathbb{R}^d} \left| \pi(h \mid Y_{0:n-1}) - \frac{1}{\sqrt{(2\pi)^d |V^*|}} \exp\left(-\frac{h^T (V^*)^{-1} h}{2} \right) \right| dh \xrightarrow[n \to +\infty]{\mathbb{P}_\star - a.s.} 0$$

where $V^* = (J^\star)^{-1}$ and $J^\star$ is the positive definite symmetric matrix introduced in A6.

Proof of Theorem 12. We place ourselves on an intersection of events of $\mathbb{P}_\star$-probability equal to 1, on which Assumptions A2 to A6, Lemma 6, and Lemma 8 simultaneously hold. In this proof, we will say that a predicate $P(n)$ holds *for all n large enough* if there exists $N \in \mathbb{N}$ such that the predicate $P(n)$ holds for all $n > N$.

Recall that for all positive semi-definite symmetric matrices $A \in \mathbb{R}^{d \times d}$ and all vectors $x \in \mathbb{R}^d$, we have the inequalities $\lambda_1(A)\|x\|^2 \leq x^T A x \leq \lambda_d(A)\|x\|^2$ and $|x^T A x| \leq \|A\| \, \|x\|^2$, where $\|A\|$ denotes the spectral norm of A. The matrix $J^\star$ given by Assumption A6 is positive definite and thus invertible, so we may define $V^* = (J^\star)^{-1}$ which is a fortiori positive definite.

We introduce the quantities

$$\begin{aligned} D_n &= \int_{\mathbb{R}^d} \pi\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) \exp\left(\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n\left(\hat{\theta}_n\right)\right) dh\,, \\ \pi(h \mid Y_{0:n-1}) &= \frac{\pi\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) \exp\left(\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n\left(\hat{\theta}_n\right)\right)}{D_n}, \\ I_n &= \int_{\mathbb{R}^d} \left| \pi(h \mid Y_{0:n-1}) - \frac{1}{\sqrt{(2\pi)^d |V^*|}} \exp\left(-\frac{h^T (V^*)^{-1} h}{2}\right) \right| dh. \end{aligned}$$

Our goal is to prove

$$I_n \xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.} 0. \tag{3.8}$$

Let us define

$$I_n^{(1)} = \int_{\mathbb{R}^d} \left| \pi\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) \exp\left(\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n(\hat{\theta}_n)\right) - \pi\left(\theta^\star\right) \exp\left(-\frac{h^T (V^*)^{-1} h}{2}\right) \right| dh\,,$$

$$I_n^{(2)} = \int_{\mathbb{R}^d} \left| \pi\left(\theta^\star\right) \exp\left(-\frac{h^T (V^*)^{-1} h}{2}\right) - \frac{D_n}{\sqrt{(2\pi)^d |V^*|}} \exp\left(-\frac{h^T (V^*)^{-1} h}{2}\right) \right| dh.$$

In order to obtain (3.8), it is sufficient to prove

$$I_n^{(1)} \xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.} 0. \tag{3.9}$$

Indeed, if (3.9) holds, then combining it with the triangle inequality for integrals leads to

$$D_n \xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.} \int_{\mathbb{R}^d} \pi\left(\theta^\star\right) \exp\left(-\frac{h^T (V^*)^{-1} h}{2}\right) dh = \pi\left(\theta^\star\right) \sqrt{(2\pi)^d |V^*|}. \tag{3.10}$$

From (3.10), it follows that the integrand in $I_n^{(2)}$ converges to 0 as $n \to +\infty$ for all $h \in \mathbb{R}^d$, and is dominated by the integrable function $h \mapsto \exp(-h^T (V^*)^{-1} h/2)$ for all n large enough. By the dominated convergence, we get

$$I_n^{(2)} \xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.} 0. \tag{3.11}$$

Using (3.9) and (3.11) with the triangle inequality, we obtain

$$0 \;\leq\; D_n I_n \;\leq\; I_n^{(1)} + I_n^{(2)} \;\xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.}\; 0 \tag{3.12}$$

Since D_n converges to a positive finite limit via (3.10), the limit in (3.12) implies (3.8) as desired, and the asymptotic Normality of the posterior distribution is then proved.

Therefore, all we need now is to prove (3.9). Let us define

$$I_n^{(1,1)} = \int_{\mathbb{R}^d} \left| \pi\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) \exp\left(\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n(\hat{\theta}_n)\right) - \pi(\hat{\theta}_n) \exp\left(-\frac{h^T[-\frac{1}{n}\nabla^2 \ell_n(\hat{\theta}_n)]h}{2}\right) \right| dh,$$

$$I_n^{(1,2)} = \int_{\mathbb{R}^d} \left| \pi\left(\hat{\theta}_n\right) \exp\left(-\frac{h^T[-\frac{1}{n}\nabla^2 \ell_n(\hat{\theta}_n)]h}{2}\right) - \pi(\theta^\star) \exp\left(-\frac{h^T (V^*)^{-1} h}{2}\right) \right| dh.$$

By the triangle inequality, we have

$$0 \;\leq\; I_n^{(1)} \;\leq\; I_n^{(1,1)} + I_n^{(1,2)}. \tag{3.13}$$

Proving (3.9) thus boils down to proving that both $I_n^{(1,1)}$ and $I_n^{(1,2)}$ converge to 0.

By combining the convergence of $\hat{\theta}_n$ to $\theta^\star$ (Assumption A4), the continuity of the prior density π at $\theta^\star$ (Assumption A7), and the convergence of $-\,n^{-1}\nabla^2 \ell_n(\hat{\theta}_n)$ to $J^\star = (V^*)^{-1}$ (Lemma 10), we get the convergence of the integrand in $I_n^{(1,2)}$ to 0 as $n \to +\infty$, for all $h \in \mathbb{R}^d$. Moreover, the convergence of $\pi(\hat{\theta}_n)$ and $-\,n^{-1}\nabla^2 \ell_n(\hat{\theta}_n)$ to their respective limits guarantees that, for all n large enough, we have $\pi(\hat{\theta}_n) \leq \pi(\theta^\star) + 1$ and $\lambda_1(-\,n^{-1}\nabla^2 \ell_n(\hat{\theta}_n)) \geq \lambda_1((V^*)^{-1})/2 > 0$. Using the triangle inequality, this implies that, for all n large enough, the integrand in $I_n^{(1,2)}$ is dominated by the integrable function $h \mapsto 2(\pi(\theta^\star) + 1)\exp(-\lambda_1((V^*)^{-1})\|h\|^2/4)$. Therefore, the dominated convergence theorem implies that

$$I_n^{(1,2)} \;\xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.}\; 0. \tag{3.14}$$

With (3.13) and (3.14), all that remains to be done in order to prove (3.9) is to prove

$$I_n^{(1,1)} \;\xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.}\; 0. \tag{3.15}$$

Thus, the rest of the proof will focus on obtaining (3.15).

For all $\delta > 0$, we can partition $I_n^{(1,1)}$ as

$$I_n^{(1,1)} = I_{n,\delta}^{(1,1,1)} + I_{n,\delta}^{(1,1,2)} \tag{3.16}$$

where

$$I_{n,\delta}^{(1,1,1)} = \int_{\|h\|>\delta\sqrt{n}} \left| \pi\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) \exp\left(\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n(\hat{\theta}_n)\right) - \pi(\hat{\theta}_n) \exp\left(-\frac{h^T[-\frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n)]h}{2}\right)\right| dh\,,$$

$$I_{n,\delta}^{(1,1,2)} = \int_{\|h\|\leq\delta\sqrt{n}} \left| \pi\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) \exp\left(\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n(\hat{\theta}_n)\right) - \pi(\hat{\theta}_n) \exp\left(-\frac{h^T[-\frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n)]h}{2}\right)\right| dh\,.$$

We now prove that for a suitable choice of δ small enough, both $I_{n,\delta}^{(1,1,1)}$ and $I_{n,\delta}^{(1,1,2)}$ converge to 0 as $n \to +\infty$, which will imply (3.15) thanks to the decomposition in (3.16).

Starting with $I_{n,\delta}^{(1,1,1)}$, for all $\delta > 0$, we introduce

$$I_{n,\delta}^{(1,1,1,1)} = \int_{\|h\|>\delta\sqrt{n}} \left| \pi\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) \exp\left(\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n(\hat{\theta}_n)\right)\right| dh\,,$$

$$I_{n,\delta}^{(1,1,1,2)} = \int_{\|h\|>\delta\sqrt{n}} \left| \pi(\hat{\theta}_n) \exp\left(-\frac{h^T[-\frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n)]h}{2}\right)\right| dh\,,$$

so that by the triangle inequality we obtain

$$I_{n,\delta}^{(1,1,1)} \leq I_{n,\delta}^{(1,1,1,1)} + I_{n,\delta}^{(1,1,1,2)}. \tag{3.17}$$

The first term $I_{n,\delta}^{(1,1,1,1)}$ can be controlled with the well-separation of the MLE from Lemma 8. Indeed, by Lemma 8, there exists $\eta_\delta > 0$ such that

$$\sup_{\|\theta-\hat{\theta}_n\|\geq\delta} \left(\ell_n(\theta) - \ell_n(\hat{\theta}_n)\right) < -n\,\eta_\delta \tag{3.18}$$

for all n large enough. By combining (3.18) with the fact that the integral of a probability density is upper-bounded by 1, we obtain

$$0 \leq I_{n,\delta}^{(1,1,1,1)} \leq \exp\left(\sup_{\|\theta-\hat{\theta}_n\|\geq\delta} \left(\ell_n(\theta) - \ell_n(\hat{\theta}_n)\right)\right) \int_{\|h\|>\delta\sqrt{n}} \pi\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) dh \leq \exp\left(-n\,\eta_\delta\right)$$

which leads to

$$I_{n,\delta}^{(1,1,1,1)} \xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.} 0 \tag{3.19}$$

since $\eta_\delta > 0$. This result holds for all $\delta > 0$.

The second term $I_{n,\delta}^{(1,1,1,2)}$ can be controlled by exploiting the convergence of $-n^{-1}\nabla^2\ell_n(\hat{\theta}_n)$ to $J^\star$ given by Lemma 10, and the convergence of $\pi(\hat{\theta}_n)$ to $\pi(\theta^\star)$ following from the consistency of the MLE (Assumption A4) and the continuity of the prior density (Assumption A7). These results ensure that, for all n large enough, we have $\lambda_1(-n^{-1}\nabla^2\ell_n(\hat{\theta}_n)) \geq \lambda_1(J^\star)/2 > 0$ and $\pi(\hat{\theta}_n) \leq \pi(\theta^\star) + 1$, which leads to

$$\begin{aligned} 0 \;\leq\; I_{n,\delta}^{(1,1,1,2)} \;&\leq \int\limits_{\|h\|>\delta\sqrt{n}} (\pi(\theta^\star)+1) \exp\left(-\frac{\lambda_1(J^\star)\|h\|^2}{4}\right) dh \\ &\leq \int\limits_{\mathbb{R}^d} (\pi(\theta^\star)+1) \exp\left(-\frac{\lambda_1(J^\star)\|h\|^2}{4}\right) \mathbb{1}_{(\delta\sqrt{n},+\infty)}(h)\, dh \end{aligned} \tag{3.20}$$

for all n large enough. The integrand in (3.20) converges to 0 and is dominated by the integrable function $h \mapsto (\pi(\theta^\star)+1)\exp(-\lambda_1(J^\star)\|h\|^2/4)$, so that using the dominated convergence theorem leads to

$$I_{n,\delta}^{(1,1,1,2)} \xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.} 0 \tag{3.21}$$

for all $\delta > 0$. By combining (3.19) with (3.21), we finally get

$$I_{n,\delta}^{(1,1,1)} \xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.} 0 \tag{3.22}$$

for all $\delta > 0$, which takes care of the first term in (3.16).

Let's now look at the second term $I_{n,\delta}^{(1,1,2)}$ in (3.16). By the triangle inequality, we have

$$I_{n,\delta}^{(1,1,2)} \;\leq\; I_{n,\delta}^{(1,1,2,1)} + I_{n,\delta}^{(1,1,2,2)} \tag{3.23}$$

where

$$I_{n,\delta}^{(1,1,2,1)} = \int_{\|h\|\leq\delta\sqrt{n}} \left|\pi\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \pi\left(\hat{\theta}_n\right)\right| \exp\left(\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n(\hat{\theta}_n)\right) dh,$$

$$I_{n,\delta}^{(1,1,2,2)} = \int_{\|h\|\leq\delta\sqrt{n}} \pi(\hat{\theta}_n) \left|\exp\left(\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n(\hat{\theta}_n)\right) - \exp\left(-\frac{h^T[-\frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n)]h}{2}\right)\right| dh.$$

By Assumption A7, there exists a neighborhood of $\theta^\star$ on which $\theta \mapsto \pi(\theta)$ is continuous. By definition, this neighborhood contains a closed ball $\bar{\mathcal{B}}(\theta^\star, R)$ of radius R around $\theta^\star$, on which π is continuous, so a fortiori uniformly continuous by the Heine-Cantor theorem.

Let's fix any arbitrary $\varepsilon \in (0,1)$.

By uniform continuity, there exists $\delta_\varepsilon^\star \in (0, R)$ such that

$$\sup_{\|\theta-\theta^\star\|\leq\delta_\varepsilon^\star} |\pi(\theta) - \pi(\theta^\star)| \;<\; \varepsilon. \tag{3.24}$$

By Assumption A4, we have $\|\hat{\theta}_n - \theta^\star\| < \delta_\varepsilon^\star/2$ for all n large enough. For all $\delta_\varepsilon \in (0, \delta_\varepsilon^\star/2]$, the triangle inequality guarantees that $\{\|\theta - \hat{\theta}_n\| < \delta_\varepsilon\} \subseteq \{\|\theta - \theta^\star\| < \delta_\varepsilon^\star\}$ for all n large enough. By using the triangle inequality again, we get

$$\begin{aligned}
\sup_{\|h\|\leq\delta_\varepsilon\sqrt{n}} \left|\pi\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \pi\left(\hat{\theta}_n\right)\right| &\leq 2 \sup_{\|\theta-\hat{\theta}_n\|\leq\delta_\varepsilon} |\pi(\theta) - \pi(\theta^\star)| \\
&\leq 2 \sup_{\|\theta-\theta^\star\|\leq\delta_\varepsilon^\star} |\pi(\theta) - \pi(\theta^\star)| \\
&< 2\,\varepsilon
\end{aligned} \tag{3.25}$$

for all n large enough. This gives us

$$0 \;\leq\; I_{n,\delta_\varepsilon}^{(1,1,2,1)} \;\leq\; 2\,\varepsilon \int_{\|h\|\leq\delta_\varepsilon\sqrt{n}} \exp\left(\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n(\hat{\theta}_n)\right) dh \tag{3.26}$$

for all $\delta_\varepsilon \in (0, \delta_\varepsilon^\star/2]$ and all n large enough.

Recall that $J^\star > 0$, i.e. $\lambda_1(J^\star) > 0$, so from here onward we may restrict ourselves to all ε satisfying $0 < \varepsilon < \lambda_1(J^\star)/2$. By Assumption A6, there exists $\delta_\varepsilon^{J^\star} > 0$ such that

$$\sup_{\|\theta-\theta^\star\|\leq\delta_\varepsilon^{J^\star}} \left\|\frac{1}{n}\nabla^2\ell_n(\theta) + J^\star\right\| < \varepsilon \tag{3.27}$$

for all n large enough. Assumption A4 implies that $\|\hat{\theta}_n - \theta^\star\| < \delta_\varepsilon^{J^\star}/2$ for all n large enough. For all $\delta_\varepsilon \in (0, \delta_\varepsilon^{J^\star}/2]$, we have $\{\|\theta - \hat{\theta}_n\| < \delta_\varepsilon\} \subseteq \{\|\theta - \theta^\star\| < \delta_\varepsilon^{J^\star}\}$ for all n large enough, by the triangle inequality. By proceeding in the same way as in (3.25), we get

$$\sup_{\|\theta-\hat{\theta}_n\|\leq\delta_\varepsilon} \left\| \frac{1}{n}\nabla^2\ell_n(\theta) - \frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n) \right\| \;\leq\; 2 \sup_{\|\theta-\theta^\star\|\leq\delta_\varepsilon^{J^\star}} \left\| \frac{1}{n}\nabla^2\ell_n(\theta) + J^\star \right\| \;<\; 2\,\varepsilon \tag{3.28}$$

for all $\delta_\varepsilon \in (0, \delta_\varepsilon^{J^\star}/2]$ and all n large enough. By using again Assumption A6, we may choose $\delta_\varepsilon \in (0, \min(\delta_\varepsilon^\star/2, \delta_\varepsilon^{J^\star}/2))$ small enough so that $\theta \mapsto \ell_n(\theta)$ is twice continuously differentiable on $\{\|\theta - \hat{\theta}_n\| < \delta_\varepsilon\}$. For all $h \in \{\|h\| \leq \delta_\varepsilon\sqrt{n}\}$, Taylor's theorem with $\nabla\ell_n(\hat{\theta}_n) = 0$ yields

$$\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n(\hat{\theta}_n) = -\frac{h^T[-\frac{1}{n}\nabla^2\ell_n(\tilde{\theta}_n)]h}{2} \tag{3.29}$$

where $\tilde{\theta}_n = \hat{\theta}_n + \gamma\, h/\sqrt{n}$ for some $\gamma \in [0,1]$, so that $\|\tilde{\theta}_n - \hat{\theta}_n\| < \delta_\varepsilon$. This also implies that $\|\tilde{\theta}_n - \theta^\star\| < \|\tilde{\theta}_n - \hat{\theta}_n\| + \|\hat{\theta}_n - \theta^\star\| < \delta_\varepsilon^{J^\star}$ for all n large enough, so by (3.27) we get

$$\lambda_1\left(-\frac{1}{n}\nabla^2\ell_n(\tilde{\theta}_n)\right) \;>\; \frac{\lambda_1(J^\star)}{2} \tag{3.30}$$

for all n large enough. By plugging (3.29) and (3.30) back into (3.26), we get

$$\begin{aligned} I_{n,\delta_\varepsilon}^{(1,1,2,1)} &\leq 2\,\varepsilon \int\limits_{\|h\|\leq\delta_\varepsilon\sqrt{n}} \exp\left(-\frac{\lambda_1(J^\star)\|h\|^2}{4}\right) dh \\ &\leq 2\,\varepsilon \left(\frac{2}{\lambda_1(J^\star)}\right)^{d/2} \int\limits_{\mathbb{R}^d} \exp\left(-\frac{\|z\|^2}{2}\right) dz \\ &\leq 2\left(\frac{4\pi}{\lambda_1(J^\star)}\right)^{d/2} \varepsilon \end{aligned} \tag{3.31}$$

for all n large enough. Similarly, equation (3.29) can be rewritten as

$$\ell_n\left(\hat{\theta}_n + \frac{h}{\sqrt{n}}\right) - \ell_n(\hat{\theta}_n) \;=\; -\frac{h^T[-\frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n)]h}{2} + R_n(h) \tag{3.32}$$

with

$$R_n(h) \;=\; \frac{h^T[\frac{1}{n}\nabla^2\ell_n(\tilde{\theta}_n) - \frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n)]h}{2}.$$

Since $\|\tilde{\theta}_n - \hat{\theta}_n\| \leq \|h\|/\sqrt{n} \leq \delta_\varepsilon$, equation (3.28) implies that

$$\sup_{\|h\| \leq \delta_\varepsilon \sqrt{n}} |R_n(h)| \leq \|h\|^2 \, \varepsilon \tag{3.33}$$

for all n large enough. By combining (3.24), (3.27), (3.32), (3.33), and by using the inequality $|\exp(x) - 1| \leq |x| \exp(|x|)$ for all $x \in \mathbb{R}$, we obtain

$$\begin{aligned}
I_{n,\delta_\varepsilon}^{(1,1,2,2)} &\leq (\pi(\theta^\star) + 1) \int\limits_{\|h\| \leq \delta_\varepsilon \sqrt{n}} \left| \exp\left(-\frac{h^T[-\frac{1}{n}\nabla^2 \ell_n(\hat{\theta}_n)]h}{2} \right) \left(\exp\left(R_n(h)\right) - 1 \right) \right| dh \\
&\leq (\pi(\theta^\star) + 1) \, \varepsilon \int\limits_{\|h\| \leq \delta_\varepsilon \sqrt{n}} \exp\left(-\frac{h^T[-\frac{1}{n}\nabla^2 \ell_n(\hat{\theta}_n) - \varepsilon I_d]h}{2} \right) \|h\|^2 \, dh \\
&\leq (\pi(\theta^\star) + 1) \, \varepsilon \int\limits_{\|h\| \leq \delta_\varepsilon \sqrt{n}} \exp\left(-\frac{\lambda_1(J^\star)\|h\|^2}{4} \right) \|h\|^2 \, dh \\
&\leq (\pi(\theta^\star) + 1) \, \varepsilon \int\limits_{\mathbb{R}^d} \exp\left(-\frac{\lambda_1(J^\star)\|h\|^2}{4} \right) \|h\|^2 \, dh \\
&\leq (\pi(\theta^\star) + 1) \, \varepsilon \left(\frac{2}{\lambda_1(J^\star)} \right)^{1+d/2} \int\limits_{\mathbb{R}^d} \exp\left(-\frac{\|z\|^2}{2} \right) \|z\|^2 \, dz \\
&\leq (\pi(\theta^\star) + 1) \left(\frac{2}{\lambda_1(J^\star)} \right)^{1+d/2} d \, (2\pi)^{d/2} \, \varepsilon
\end{aligned} \tag{3.34}$$

for all n large enough.

Let us introduce the strictly positive constant

$$c_d^{J^\star} = \min\left(\frac{\lambda_1(J^\star)}{2} \, , \, \left[2 \left(\frac{4\pi}{\lambda_1(J^\star)} \right)^{d/2} + (\pi(\theta^\star) + 1) \left(\frac{2}{\lambda_1(J^\star)} \right)^{1+d/2} d \, (2\pi)^{d/2} \right]^{-1} \right)$$

which does not depend on ε. By plugging (3.31) and (3.34) back into (3.23), we conclude that for all $\varepsilon \in (0, c_d^{J^\star})$, there exists $\delta_\varepsilon > 0$ such that

$$0 \leq I_{n,\delta_\varepsilon}^{(1,1,2)} < \varepsilon \tag{3.35}$$

for all n large enough. Since $\delta_\varepsilon > 0$, equation (3.22) also readily yields

$$0 \leq I_{n,\delta_\varepsilon}^{(1,1,1)} < \varepsilon \tag{3.36}$$

for all n large enough. By plugging (3.35) and (3.36) back into (3.16), we can conclude that for all $\varepsilon \in (0, c_d^{J^\star})$ and its associated $\delta_\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that we have

$$0 \;\leq\; I_n^{(1,1)} \;<\; I_{n,\delta_\varepsilon}^{(1,1,1)} + I_{n,\delta_\varepsilon}^{(1,1,2)} \;<\; 2\,\varepsilon$$

for all $n > N_\varepsilon$. Since this holds for every arbitrary $\varepsilon \in (0, c_d^{J^\star})$, we obtain

$$I_n^{(1,1)} \xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.} 0.$$

which establishes (3.15), and concludes the proof of Theorem 12. □

3.5 Sufficient conditions in state-space models

Assumptions A1 to A6 describe general phenomena about the log-likelihood function and the observed Fisher information. These phenomena are not particular to state-space models and provably occur in a wide variety of settings. Our main challenge is to provide a set of verifiable and desirably weak conditions on the state-space model — in the spirit of Douc and Moulines (2012) — under which Assumptions A1 to A6 can be proved to hold.

Before stating the conditions, we introduce some additional objects specific to state-space models. For ease of analogy in the later discussions, we deliberately depart from the notations used in the previous chapters to adopt the same conventions as the ones in Douc and Moulines (2012). We start the indexing at time 0 for the processes $(X_i)_{i\in\mathbb{N}}$ and $(Y_i)_{i\in\mathbb{N}}$ of latent states living in $\mathbb{X} \subseteq \mathbb{R}^{d_x}$, equipped with its Borel σ-algebra $\mathcal{X}$, and observations living in $\mathbb{Y} \subseteq \mathbb{R}^{d_y}$, equipped with its Borel σ-algebra $\mathcal{Y}$. We denote the initial distribution of the latent Markov chain by μ, its transition kernel by $Q_\theta : \mathbb{X} \times \mathcal{X} \to [0,1]$ with density q_θ, and the observation distribution by $G_\theta(\cdot|x) : \mathcal{Y} \to [0,1]$ with density $y \mapsto g_\theta(y|x)$. These densities are with respect to some reference measure, typically the Lebesgue measure. When the context is clear, we use the shorthand notation $g_{\theta,i}(x) = g_\theta(y_i|x)$. All the previous model components are parametrized by some $\theta \in \mathbb{T} \subseteq \mathbb{R}^{d_\theta}$. For $n \in \mathbb{N}^*$ and $i \in [\![0, n-1]\!]$,

we use $\mathbb{E}_{\theta,\mu}\left[f(X_{0:n-1}) \,|\, y_{0:n-1}\right]$ to denote the induced expectation of $f(X_{0:n-1})$ conditional on some θ and realized observations $y_{0:n-1} \in \mathbb{Y}^n$, given an initial distribution μ for X_0. Under stationarity assumptions, we will artificially extend the processes $(X_i)_{i\in\mathbb{N}}$ and $(Y_i)_{i\in\mathbb{N}}$ to allow for negative indexes $i \in \mathbb{Z}$. We use the convention that sums and products over empty sets of indexes are respectively equal to 0 and 1.

Recall that any measure ν on $\mathbb{X}$ induces a real-valued map on the space of suitably integrable functions f via $\nu(f) = \int_{\mathbb{X}} f(x)\nu(dx)$. On the other hand, an integral kernel $K : \mathbb{X} \times \mathcal{X} \to \mathbb{R}_+$ induces an operator on the space of measures, as well as an operator on the space of suitably integrable functions, i.e. for all measures μ on $\mathbb{X}$ and all integrable functions $f : \mathbb{X} \to \mathbb{R}$, we can define the measure μK for all events $A \in \mathcal{X}$ by

$$(\mu K)(A) = \int_{\mathbb{X}} \mu(dx) K(x, A) \tag{3.37}$$

and the function Kf for all $x \in \mathbb{X}$ by

$$(Kf)(x) = \int_{\mathbb{X}} K(x, dx') f(x') = (\delta_x K)(f).$$

For all $(\theta, y_0) \in \mathbb{T} \times \mathbb{Y}$ and suitably integrable non-negative functions $\varphi : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$, we define the unnormalized kernel $L_\theta^\varphi\langle y_0\rangle : \mathbb{X} \times \mathcal{X} \to [0, 1]$ by

$$L_\theta^\varphi\langle y_0\rangle(x_0, A) = \int \mathbb{1}_A(x_1)\varphi(x_0, x_1)\, g_\theta(y_0|x_0) Q_\theta(x_0, dx_1). \tag{3.38}$$

We will use the shorthand notation $L_\theta\langle y_0\rangle = L_\theta^{\mathbb{1}_{\mathbb{X}\times\mathbb{X}}}\langle y_0\rangle$, so that for all $(x_0, A) \in \mathbb{X}\times\mathcal{X}$, we have

$$L_\theta\langle y_0\rangle(x_0, A) = \int \mathbb{1}_A(x_1)\, g_\theta(y_0|x_0) Q_\theta(x_0, dx_1). \tag{3.39}$$

For all $(m, n) \in \mathbb{N} \times \mathbb{N}^*$ and $(\theta, y_{-m:n-1}) \in \mathbb{T} \times \mathbb{Y}^{m+n}$, we define the unnormalized kernel $L_\theta\langle y_{-m:n-1}\rangle : \mathbb{X} \times \mathcal{X} \to [0, 1]$ as the composition of the kernels $L_\theta\langle y_l\rangle$ for $l \in [\![-m, n-1]\!]$ i.e.

$$L_\theta\langle y_{-m:n-1}\rangle(x_{-m}, A) = \int \mathbb{1}_A(x_n) \prod_{l=-m}^{n-1} \left[g_\theta(y_l|x_l) Q_\theta(x_l, dx_{l+1})\right] \tag{3.40}$$

for all $(x_{-m}, A) \in \mathbb{X}\times\mathcal{X}$, with the convention $L_\theta\langle y_{-m:n-1}\rangle(x_{-m}, A) = \delta_{x_{-m}}(A)$ if $n - 1 < -m$.

3.5.1 Convergence of the log-likelihood

The convergence of the log-likelihood function to a suitable limit (Assumption A2) has been established for general and possibly misspecified state-space models by Douc and Moulines (2012), using the following conditions.

Condition C1. *The process $(Y_t)_{t\in\mathbb{N}}$ is strongly stationary and ergodic.*

Condition C2. *The parameter space $\mathbb{T}$ is a compact metric space.*

Condition C3. *There exists $r \in \mathbb{N}^*$ and a measurable set $K \subseteq \mathbb{Y}^r$ such that*

(i) $\mathbb{P}_\star(Y_{0:r-1} \in K) > 2/3$.

(ii) For all $\eta > 0$, there exists an r-local Doeblin set $C \in \mathcal{X}$ such that, for all $\theta \in \mathbb{T}$ and all $y_{0:r-1} \in K$, we have

$$\sup_{x\in C^c} L_\theta\langle y_{0:r-1}\rangle(x,\mathbb{X}) \;\leq\; \eta \sup_{x\in\mathbb{X}} L_\theta\langle y_{0:r-1}\rangle(x,\mathbb{X}) \;<\; +\infty$$

and

$$\inf_{y_{0:r-1}\in K} \frac{\epsilon_C^-(y_{0:r-1})}{\epsilon_C^+(y_{0:r-1})} > 0$$

where $\epsilon_C^+ : \mathbb{Y}^r \to \mathbb{R}_+$ and $\epsilon_C^- : \mathbb{Y}^r \to \mathbb{R}_+$ are positive functions coming from the definition of the r-local Doeblin set C. In other words, there exists a family of probability measures $(\lambda_{\theta,y}^C)$ and positive functions $(\varphi_{\theta,y}^C)$, indexed by $\theta \in \mathbb{T}$ and $y = y_{0:r-1} \in \mathbb{Y}^r$, such that $\lambda_{\theta,y}^C(C) = 1$ for all $(\theta, y) \in \mathbb{T} \times \mathbb{Y}^r$, and such that we have

$$\epsilon_C^-(y)\varphi_{\theta,y}^C(x)\lambda_{\theta,y}^C(A) \;\leq\; L_\theta\langle y_{0:r-1}\rangle(x, A\cap C) \;\leq\; \epsilon_C^+(y)\varphi_{\theta,y}^C(x)\lambda_{\theta,y}^C(A)$$

for all $x \in C$, all $A \in \mathcal{X}$, all $\theta \in \mathbb{T}$, and all $y = y_{0:r-1} \in \mathbb{Y}^r$.

(iii) There exists a set $D \in \mathcal{X}$ such that

$$\mathbb{E}_\star\left[\log^- \inf_{\theta\in\mathbb{T}} \inf_{x\in D} L_\theta\langle Y_{0:r-1}\rangle(x, D)\right] < +\infty.$$

Condition C4. *For all $\theta \in \mathbb{T}$ and all $(x, y) \in \mathbb{X} \times \mathbb{Y}$, we have $g_\theta(y|x) > 0$ and*

$$\mathbb{E}_\star \left[\log^+ \sup_{\theta \in \mathbb{T}} \sup_{x \in \mathbb{X}} g_\theta(Y_0|x) \right] < +\infty.$$

Condition C5. *There exists $N \in \mathbb{N}^*$ such that for all $x \in \mathbb{X}$ and all $n \geq N$, the function $\theta \mapsto p(Y_{0:n-1}|X_0 = x, \theta)$ is continuous on $\mathbb{T}$, $\mathbb{P}_\star$-almost surely.*

Conditions C3, C4, and C5 respectively correspond to Assumptions (A1), (A2), and (A3) in Douc and Moulines (2012). Intuitions and discussions about these conditions can be found in their Remarks 1 to 6. In particular, these conditions are weaker than the strong mixing assumptions that are traditionally used in such settings and that we used to prove the consistency of the H-score in Chapter 1 (see Condition C9 in Appendix B.5.2.3). The latter are too restrictive to be satisfied even by simple linear Gaussian state-space models, whereas the former have been proved to hold for a reasonably wide class of models (Section 4 in Douc and Moulines, 2012) including finite states or linear Gaussian state-space models, as well as particular instances of stochastic volatility models.

For $D \in \mathcal{X}$ and $r \in \mathbb{N}^*$ as described in Condition C3, we may consider a subset $\mathcal{M}(D, r)$ from the space $\mathcal{P}(\mathbb{X}, \mathcal{X})$ of probability measures on $(\mathbb{X}, \mathcal{X})$, defined as

$$\mathcal{M}(D, r) \;=\; \left\{ \mu \in \mathcal{P}(\mathbb{X}, \mathcal{X}) : \forall u \in [\![1, r]\!],\, \mathbb{E}_\star \left[\log^- \inf_{\theta \in \mathbb{T}} \mu\, L_\theta \langle Y_{0:u-1} \rangle \mathbb{1}_D \right] < +\infty \right\}. \tag{3.41}$$

Most of the subsequent results will be derived for initial distributions of the latent Markov chain belonging to such a subset. For a given initial distribution μ, explicit sufficient conditions can be derived to ensure $\mu \in \mathcal{M}(D, r)$ (Proposition 3 and Remark 8 in Douc and Moulines, 2012).

Remark 13. *If Condition C3 holds with $r = 1$, then for all probability measures $\mu \in \mathcal{P}(\mathbb{X}, \mathcal{X})$, having $\mu(D) > 0$ is sufficient to guarantee $\mu \in \mathcal{M}(D, r)$. Indeed, the non-negativity of the integrand ensures that, for all $\mu \in \mathcal{P}(\mathbb{X}, \mathcal{X})$, we have*

$$\mu\, L_\theta \langle Y_0 \rangle \mathbb{1}_D \;=\; \int_{\mathbb{X}} \mu(dx) L_\theta \langle Y_0 \rangle (x, D) \;\geq\; \mu(D) \inf_{x \in D} L_\theta \langle Y_0 \rangle (x, D).$$

Condition C3 (iii) implies that $\inf_{x\in D} L_\theta\langle Y_0\rangle(x,D) > \inf_{\theta\in\mathbb{T}}\inf_{x\in D} L_\theta\langle Y_0\rangle(x,D) > 0$ *(as explained in Remark 14), so that if* $\mu(D)>0$*, then we have*

$$\inf_{\theta\in\mathbb{T}} \mu\, L_\theta\langle Y_0\rangle \mathbb{1}_D \;\geq\; \mu(D) \inf_{\theta\in\mathbb{T}}\inf_{x\in D} L_\theta\langle Y_0\rangle(x,D) \;>\; 0. \tag{3.42}$$

The function $u\mapsto \log^-(u) = (|\log(u)| - \log(u))/2$ *is non-increasing on* $\mathbb{R}_+^*$ *and satisfies* $\log^-(uv)\leq \log^-(u)+\log^-(v)$ *for all* $(u,v)\in(\mathbb{R}_+^*)^2$*, hence equation (3.42) implies*

$$\mathbb{E}_\star\left[\log^- \inf_{\theta\in\mathbb{T}} \mu\, L_\theta\langle Y_0\rangle \mathbb{1}_D\right] \;\leq\; \log^-\Big(\mu(D)\Big) + \mathbb{E}_\star\left[\log^- \inf_{\theta\in\mathbb{T}}\inf_{x\in D} L_\theta\langle Y_0\rangle(x,D)\right].$$

Since $\log^-(\mu(D))<+\infty$ *and* $\mathbb{E}_\star[\log^-\inf_{\theta\in\mathbb{T}}\inf_{x\in D} L_\theta\langle Y_0\rangle(x,D)]<+\infty$ *thanks to* $\mu(D)>0$ *and Condition C3 (iii), we finally get* $\mathbb{E}_\star\left[\log^-\inf_{\theta\in\mathbb{T}}\mu\, L_\theta\langle Y_0\rangle\mathbb{1}_D\right]<+\infty$ *i.e.* $\mu\in\mathcal{M}(D,r)$*. Therefore, we indeed have* $\{\mu\in\mathcal{P}(\mathbb{X},\mathcal{X}) : \mu(D)>0\}\subseteq\mathcal{M}(D,r)$ *if* $r=1$ *in Condition C3.*

3.5.2 Existence and uniqueness of $\theta^\star$

Under Conditions C1 to C5, the limit ℓ of the log-likelihood function $n^{-1}\ell_n$ is upper semi-continuous on $\mathbb{T}$ (Theorem 2 in Douc and Moulines, 2012), which guarantees the existence and non-emptiness of the set $\Theta^\star = \arg\max_{\theta\in\mathbb{T}}\ell(\theta)$, thanks to the compactness of $\mathbb{T}$. Guaranteeing the uniqueness of the maximizer i.e. $\Theta^\star=\{\theta^\star\}$ requires more subtle considerations. This has been achieved by adding further identifiability restrictions — e.g. Section 5 of Douc et al. (2004), or Lemma 13.10 and Proposition 13.13 of Douc et al. (2014) — albeit for well-specified state-space models satisfying strong mixing conditions. As far as we know, the uniqueness of $\theta^\star$ for misspecified models with less restrictive conditions remains to be investigated. However, we might not necessarily require the uniqueness of the maximizer of ℓ: by having only $\Theta^\star\neq\emptyset$ instead of $\Theta^\star=\{\theta^\star\}$ in Assumption A3, and by replacing $\|\hat\theta_n-\theta^\star\|$ with $d(\hat\theta_n,\Theta^\star)=\inf_{\theta^\star\in\Theta^\star}\|\hat\theta_n-\theta^\star\|$ in Assumption A4, we could still hope to prove a notion of posterior consistency redefined as

$$\Pi\left(\bigcup_{\theta^\star\in\Theta^\star}\mathcal{U}_{\theta^\star}\,\middle|\, Y_{0:n-1}\right) \xrightarrow[n\to+\infty]{\mathbb{P}_\star-a.s.} 1$$

for all unions of neighborhoods $\mathcal{U}_{\theta^\star}$ of $\theta^\star \in \Theta^\star$. In the rest of this chapter, we will assume that $\Theta^\star = \{\theta^\star\}$ for simplicity.

3.5.3 Consistency of the MLE

Under conditions C1 to C5, Theorem 2 in Douc and Moulines (2012) proves that

$$d(\hat{\theta}_n, \Theta^\star) \xrightarrow[n\to+\infty]{\mathbb{P}_\star - a.s.} 0$$

for all initial distribution $\mu \in \mathcal{M}(D, r)$ of the latent Markov chain, where the set of distributions $\mathcal{M}(D, r)$ is defined as in (3.41). When Assumption A3 holds, we have $d(\hat{\theta}_n, \Theta^\star) = \|\hat{\theta}_n - \theta^\star\|$ so that the previous statement reduces to Assumption A4.

3.5.4 Well-separation of $\theta^\star$

Under Conditions C1 to C5, the proof of Theorem 12 in Douc and Moulines (2012, p.2725) proceeds with batches $Z_k = Y_{kr:(k+1)r-1} \in \mathbb{Y}^r$ of r successive observations to establish

$$\lim_{n\to+\infty} \left(\sup_{\theta\in\mathbb{T}_\delta} \frac{1}{n} \log p_\mu^\theta(Z_{0:n-1}) \right) < \mathbb{E}_\star \left[\log p^{\theta^\star}(Z_0 | Z_{-\infty:-1}) \right] \tag{3.43}$$

for all $\mu \in \mathcal{M}(D, r)$. Most terms in (3.43) are the fruits of non-trivial derivations: $p_\mu^\theta(Z_{0:n-1})$ corresponds to $p(Y_{0:(n-1)r} | \theta)$ with μ as initial distribution for the latent states, $p^\theta(Z_0 | Z_{-\infty:-1})$ denotes the $\mathbb{P}_\star$-almost sure limit — which can be proved to exist with a value not depending on the initial distribution μ (Douc and Moulines, 2012, Corollary 9) — of the sequence $(p_\mu^\theta(Z_0 \,|\, Z_{-m:-1}))_{m\in\mathbb{N}}$ as $m \to +\infty$, and $\mathbb{T}_\delta$ is defined as $\{\theta \in \mathbb{T} : d(\theta, \Theta_r^\star) \geq \delta\}$ for all $\delta > 0$ with $\Theta_r^\star = \arg\max_{\theta\in\mathbb{T}} \mathbb{E}_\star \left[\log p^\theta(Z_0 | Z_{-\infty:-1}) \right]$. In the particular case where C3 holds with $r = 1$, if Assumption A3 holds, then $\Theta_r^\star = \{\theta^\star\}$ and equation (3.43) simplifies to

$$\lim_{n\to+\infty} \left(\sup_{\|\theta-\theta^\star\|\geq\delta} \frac{1}{n} \log p_\mu^\theta(Y_{0:n-1}) \right) < \mathbb{E}_\star \left[\log p^{\theta^\star}(Y_0 | Y_{-\infty:-1}) \right]$$

for all $\delta > 0$ and all initial distributions $\mu \in \mathcal{M}(D, r)$. This inequality precisely implies the well-separation of the mode $\theta^\star$ (Assumption A5). For a generic $r \in \mathbb{N}^*$, the analysis

might require more technical precautions but should not be fundamentally different. Without much loss of generality, we will always assume that $r = 1$ when referring to Condition C3 in the following sections, unless required otherwise.

3.5.5 Uniform convergence of the observed Fisher information

Out of all the assumptions listed in Section 3.2, the uniform convergence of the observed Fisher information around $\theta^\star$ (Assumption A6 or its weaker form $\tilde{A}6$) is arguably the most elusive. Providing additional conditions that provably ensure such a control over the log-likelihood's Hessian for possibly misspecified state-space models, while maintaining the practical applicability of Conditions C3 to C5, constitutes the main obstacle to overcome and would be the major part of our contribution. This will be the focus of the next section.

3.6 Controlling the Hessian of the log-likelihood

In this section, we propose further technical conditions and combine them with C1 to C5 in an attempt to guarantee Assumptions A6 or $\tilde{A}6$. Our current work falls short of a complete proof at the time of writing, but it points out relevant intricacies and suggests possible research directions to explore.

3.6.1 Additional notations and conditions

Using the objects introduced in Section 3.5, we have

$$\mu L_\theta \langle y_{-m:n-1} \rangle(h) = \int h(x_n)\mu(dx_{-m}) \prod_{l=-m}^{n-1} [g_\theta(y_l|x_l)Q_\theta(x_l, dx_{l+1})] \tag{3.44}$$

for all probability measures μ and integrable functions $h : \mathbb{X} \to \mathbb{R}$. We denote by $\mu_{\theta,-m:n-1}$ the normalized version of the measure $\mu L_\theta \langle y_{-m:n-1} \rangle$, defined for all $A \in \mathcal{X}$ as

$$\mu_{\theta,-m:n-1}(A) = \frac{\mu L_\theta \langle y_{-m:n-1} \rangle(A)}{\mu L_\theta \langle y_{-m:n-1} \rangle(\mathbb{X})} \tag{3.45}$$

with the shorthand notation $\mu_{\theta,-m} = \mu_{\theta,-m:-m}$. For two functions $f : \mathbb{X} \to \mathbb{R}$ and $h : \mathbb{Y} \to \mathbb{R}$, we define the function $f \otimes h : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ as $(f \otimes h)(x, y) = f(x)h(y)$ for all $(x, y) \in \mathbb{X} \times \mathbb{Y}$. For two measures η and μ, let $\eta \otimes \mu$ denote their product measure. We define the joint kernel $\bar{L}^\theta\langle y_{-m:n-1}\rangle$ from the product space $\mathbb{X} \times \mathbb{X}$ to $\mathcal{X} \times \mathcal{X}$ by

$$\bar{L}_\theta\langle y_{-m:n-1}\rangle(\bar{x}, \bar{A}) = L_\theta\langle y_{-m:n-1}\rangle(x^{(1)}, A^{(1)}) \,.\, L_\theta\langle y_{-m:n-1}\rangle(x^{(2)}, A^{(2)}) \tag{3.46}$$

for all $\bar{x} = (x^{(1)}, x^{(2)}) \in \mathbb{X} \times \mathbb{X}$ and $\bar{A} = (A^{(1)} \times A^{(2)}) \in \mathcal{X} \times \mathcal{X}$. For all $y_{-m:n-1} \in \mathbb{Y}^{m+n}$ and $i \in [\![-m, n]\!]$, we define the function $h_{\theta,i:n} : \mathbb{X} \to \mathbb{R}_+$ as

$$h_{\theta,i:n}(x) \;=\; L_\theta\langle y_{i:n-1}\rangle(x, \mathbb{X}). \tag{3.47}$$

In other words, the value of $h_{\theta,i:n}(x)$ is equal to the normalizing constant of the predictive density $p(dx_n|y_{i:n-1}, \theta)$ conditional on $X_i = x$, so that $h_{\theta,i:n}(x) = p(y_{i:n-1}|X_i = x, \theta)$, which also equals the normalizing constant of the filtering density $p(dx_{n-1}|y_{i:n-1}, \theta)$ at time $n - 1$. From this definition, it follows that

$$h_{\theta,i:n}(x) \;=\; \delta_x L_\theta\langle y_{i:n-1}\rangle(\mathbb{1}_{\mathbb{X}}) \qquad \text{and} \qquad h_{\theta,i:n+1}(x) \;=\; \delta_x L_\theta\langle y_{i:n-1}\rangle(g_{\theta,n}) \tag{3.48}$$

which also implies the recursive relationship

$$h_{\theta,i:n} \;=\; L_\theta\langle y_i\rangle \, h_{\theta,i+1:n}. \tag{3.49}$$

The dependence of $h_{\theta,i:n}$ on the observations is kept implicit in the notation for better readability. Using the previous notations, one can notice that, for $i < n$, we have

$$\begin{aligned}
h_{\theta,i:n}(x) \;&=\; \int_{\mathbb{X}} \delta_x(dx_i) \prod_{l=i}^{n-1} \Big[g_\theta(y_l|x_l) Q_\theta(x_l, dx_{l+1}) \Big] \\
&=\; g_\theta(y_i|x) \int_{\mathbb{X}} Q_\theta(x, dx_{i+1}) \prod_{l=i+1}^{n-1} \Big[g_\theta(y_l|x_l) Q_\theta(x_l, dx_{l+1}) \Big] \\
&=\; g_\theta(y_i|x) \int_{\mathbb{X}} Q_\theta(x, dx_{i+1}) \int_{\mathbb{X}} \delta_{x_{i+1}}(d\tilde{x}_{i+1}) \prod_{l=i+1}^{n-1} \Big[g_\theta(y_l|\tilde{x}_l) Q_\theta(\tilde{x}_l, d\tilde{x}_{l+1}) \Big] \\
&=\; g_\theta(y_i|x) \int_{\mathbb{X}} Q_\theta(x, dx_{i+1}) L^\theta\langle y_{i+1:n-1}\rangle(x_{i+1}, \mathbb{X}) \\
&=\; g_{\theta,i}(x) \,.\, (Q_\theta \, h_{\theta,i+1:n})(x)
\end{aligned}$$

which can be written more concisely as

$$h_{\theta,i:n} \;=\; (g_{\theta,i})(Q_\theta\, h_{\theta,i+1:n}). \tag{3.50}$$

Using the previous notations, we have

$$\begin{aligned}
&\mathbb{E}_{\theta,\mu}\left[\varphi(X_{i-1},X_i)\,|\,y_{-m:n-1}\right]\\
&\quad= \frac{\int_{\mathbb{X}}\varphi(x_{i-1},x_i)\mu(dx_{-m})\prod_{l=-m}^{n-1}\left[g_\theta(y_l|x_l)Q_\theta(x_l,dx_{l+1})\right]}{\int_{\mathbb{X}}\mu(dx_{-m})\prod_{l=-m}^{n-1}\left[g_\theta(y_l|x_l)Q_\theta(x_l,dx_{l+1})\right]}\\
&\quad= \frac{\int_{\mathbb{X}}(\mu L_\theta\langle y_{-m:i-2}\rangle)(dx_{i-1})\varphi(x_{i-1},x_i)g_\theta(y_{i-1}|x_{i-1})Q_\theta(x_{i-1},dx_i)h_{\theta,i:n}(x_i)}{\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n})}\\
&\quad= \frac{\int_{\mathbb{X}}[\int_{\mathbb{X}}L_\theta^\varphi\langle y_{i-1}\rangle(x_{i-1},dx_i)h_{\theta,i:n}(x_i)](\mu L_\theta\langle y_{-m:i-2}\rangle)(dx_{i-1})}{\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n})}\\
&\quad= \frac{\int_{\mathbb{X}}[(L_\theta^\varphi\langle y_{i-1}\rangle h_{\theta,i:n})(x_{i-1})](\mu L_\theta\langle y_{-m:i-2}\rangle)(dx_{i-1})}{\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n})}
\end{aligned}$$

which leads to

$$\mathbb{E}_{\theta,\mu}\left[\varphi(X_{i-1},X_i)\,|\,y_{-m:n-1}\right] \;=\; \frac{\mu L_\theta\langle y_{-m:i-2}\rangle(L_\theta^\varphi\langle y_{i-1}\rangle h_{\theta,i:n})}{\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n})} \tag{3.51}$$

for all $i\in[\![-m,n]\!]$ and all suitably integrable non-negative functions $\varphi:\mathbb{X}\times\mathbb{X}\to\mathbb{R}_+$.

In the following sections, we place ourselves by default under Conditions C1 to C5, unless specified otherwise. We will also assume sufficient regularity conditions to swap derivation with integration signs, thus allowing the use of Fisher and Louis identities. Such conditions can be found in Proposition 10.1.6 of Cappé et al. (2005) and are also discussed in Section 6.1 of Douc et al. (2004).

Remark 14. *For any random variable Z, if $\mathbb{E}_\star[|Z|]<+\infty$ then $|Z|$ is $\mathbb{P}_\star$-almost surely finite. Thus, Assumptions C3 (iii) and C4 respectively imply that, $\mathbb{P}_\star$-almost surely, we have*

$$\inf_{\theta\in\mathbb{T}}\inf_{x\in D} L_\theta\langle Y_{0:r-1}\rangle(x,D) \;>\; 0 \qquad \textit{and} \qquad \sup_{\theta\in\mathbb{T}}\sup_{x\in\mathbb{X}} g_\theta(Y_0|x) \;<\; +\infty.$$

Under stationarity, the latter implies $\sup_{\theta\in\mathbb{T}}\sup_{x\in\mathbb{X}} g_\theta(Y_l|x) < +\infty$ *for all* $l \in \mathbb{Z}$, $\mathbb{P}_\star$*-almost surely. As a consequence we have* $\sup_{\theta\in\mathbb{T}} \|L_\theta\langle Y_l\rangle(\cdot,\mathbb{X})\|_\infty \leq \sup_{\theta\in\mathbb{T}}\sup_{x\in\mathbb{X}} g_\theta(Y_l|x) < +\infty$ *for all* $l \in \mathbb{Z}$, $\mathbb{P}_\star$*-almost surely. For any function* $\varphi_\theta : \mathbb{X}^2 \to \mathbb{R}_+$ *satisfying the condition* $\mathbb{E}_\star[\log^+(\sup_{\theta\in\mathbb{T}}\sup_{(x_0,x_1)\in\mathbb{X}^2}\varphi_\theta(x_0,x_1)g_\theta(Y_1|x_1))] < +\infty$*, we may define*

$$G_l = \sup_{\theta\in\mathbb{T}}\sup_{x\in\mathbb{X}} g_\theta(Y_l|x), \quad G_l^\varphi = \sup_{\theta\in\mathbb{T}}\sup_{(x_0,x_1)\in\mathbb{X}^2}\varphi_\theta(x_0,x_1)g_\theta(Y_l|x_1), \quad F_l = \inf_{\theta\in\mathbb{T}}\inf_{x\in D} L_\theta\langle Y_l\rangle(x,D),$$

and

$$V_l = \frac{G_l}{F_l}, \qquad V_l^\varphi = \frac{G_l^\varphi}{F_l}, \qquad \widetilde{V}_l = \max(V_l, V_l^\varphi),$$

which are $\mathbb{P}_\star$*-almost surely positive and finite. The variable* V_l *essentially plays the role of* D_z *in equation (46) of Douc and Moulines (2012).*

3.6.2 Main steps and results

For all $i \in \mathbb{Z}$, we define the functions $\phi_{\theta,i}^{(1)} : \mathbb{X}^2 \to \mathbb{R}^{d_\theta}$ and $\phi_{\theta,i}^{(2)} : \mathbb{X}^2 \to \mathbb{R}^{d_\theta\times d_\theta}$ as

$$\phi_{\theta,i}^{(1)}(X_{i-1},X_i) = \nabla_\theta \log\Big(q_\theta(X_{i-1},X_i)g_\theta(y_i|X_i)\Big)$$
$$\phi_{\theta,i}^{(2)}(X_{i-1},X_i) = \nabla_\theta^2 \log\Big(q_\theta(X_{i-1},X_i)g_\theta(y_i|X_i)\Big)$$

with $\phi_{\theta,0}(X_{-1},X_0) = \nabla_\theta\log(g_\theta(y_0|X_0))$ and $\varphi_{\theta,0}(X_{-1},X_0) = \nabla_\theta^2\log(g_\theta(y_0|X_0))$ by convention. Using Louis identity, as stated e.g. in Proposition 10.1.6 from Cappé et al. (2005) and equation (22) from Douc et al. (2004), we may write

$$\frac{1}{n}\nabla^2\ell_n(\theta) = \frac{1}{n}\mathbb{E}_{\theta,\mu}\left[\sum_{i=0}^{n-1}\phi_{\theta,i}^{(2)}(X_{i-1},X_i)\,\middle|\, y_{0:n-1}\right] + \frac{1}{n}\mathrm{Var}_{\theta,\mu}\left[\sum_{i=0}^{n-1}\phi_{\theta,i}^{(1)}(X_{i-1},X_i)\,\middle|\, y_{0:n-1}\right] \quad (3.52)$$

Following equations (22) to (24) from Douc et al. (2004), we define, for any collection $\varphi = (\varphi_i)_{i\in\mathbb{Z}}$ of suitably integrable functions on $\mathbb{X}^2$, the differences

$$\Delta_{k,m,\mu}^{\mathbb{E},\theta}(\varphi) = \mathbb{E}_{\theta,\mu}\left[\sum_{i=-m}^{k}\varphi_i(X_{i-1},X_i)\,\middle|\, y_{-m:k}\right] - \mathbb{E}_{\theta,\mu}\left[\sum_{i=-m}^{k-1}\varphi_i(X_{i-1},X_i)\,\middle|\, y_{-m:k-1}\right]$$

and

$$\Delta^{\mathbb{V},\theta}_{k,m,\mu}(\varphi) \;=\; \mathrm{Var}_{\theta,\mu}\left[\sum_{i=-m}^{k}\varphi_i(X_{i-1},X_i)\,\middle|\, y_{-m:k}\right] - \mathrm{Var}_{\theta,\mu}\left[\sum_{i=-m}^{k-1}\varphi_i(X_{i-1},X_i)\,\middle|\, y_{-m:k-1}\right]$$

so that we can use telescopic sums to rewrite (3.52) as

$$\frac{1}{n}\nabla^2\ell_n(\theta) \;=\; \frac{1}{n}\sum_{k=0}^{n-1}\Delta^{\mathbb{E},\theta}_{k,0,\mu}\left(\phi^{(2)}_\theta\right) \;+\; \frac{1}{n}\sum_{k=0}^{n-1}\Delta^{\mathbb{V},\theta}_{k,0,\mu}\left(\phi^{(1)}_\theta\right). \tag{3.53}$$

Following the line of reasoning in Douc et al. (2004, cf. equations (22) to (24), p.2276), we aim to prove the existence of $\mathbb{P}_\star$-almost sure limits for the sequences $(\Delta^{\mathbb{E},\theta}_{k,m,\mu}(\varphi_\theta))_{m\in\mathbb{N}}$ and $(\Delta^{\mathbb{V},\theta}_{k,m,\mu}(\varphi_\theta))_{m\in\mathbb{N}}$ as $m\to+\infty$, respectively denoted by $\Delta^{\mathbb{E},\theta}_{k,\infty}(\phi^{(2)}_\theta)$ and $\Delta^{\mathbb{V},\theta}_{k,\infty}(\phi^{(1)}_\theta)$, and provably not depending on the initial distribution μ. We would then respectively approximate $n^{-1}\sum_{k=0}^{n-1}\Delta^{\mathbb{E},\theta}_{k,0,\mu}(\phi^{(2)}_\theta)$ and $n^{-1}\sum_{k=0}^{n-1}\Delta^{\mathbb{V},\theta}_{k,0,\mu}(\phi^{(1)}_\theta)$ by the stationary analogs $n^{-1}\sum_{k=0}^{n-1}\Delta^{\mathbb{E},\theta}_{k,\infty}(\phi^{(2)}_\theta)$ and $n^{-1}\sum_{k=0}^{n-1}\Delta^{\mathbb{V},\theta}_{k,\infty}(\phi^{(1)}_\theta)$, to which ergodic theorems would be applied to obtain the limits $\mathbb{E}_\star[\Delta^{\mathbb{E},\theta}_{0,\infty}(\phi^{(2)}_\theta)]$ and $\mathbb{E}_\star[\Delta^{\mathbb{V},\theta}_{0,\infty}(\phi^{(1)}_\theta)]$. Provided that these convergences are uniform in a neighborhood of $\theta^\star$, these results would readily translate into a uniform approximation of the Hessian $-n^{-1}\nabla^2\ell_n(\theta)$ by the positive definite limit $J^\star=\mathbb{E}_\star[\Delta^{\mathbb{E},\theta^\star}_{0,\infty}(\phi^{(2)}_{\theta^\star})]+\mathbb{E}_\star[\Delta^{\mathbb{V},\theta^\star}_{0,\infty}(\phi^{(1)}_{\theta^\star})]$ in a suitable neighborhood of $\theta^\star$, and Assumption A6 would thus be ensured.

3.6.2.1 First term: sum of smoothing expectations

The goal here is to prove that the first term in (3.53) given by

$$\frac{1}{n}\sum_{k=0}^{n-1}\Delta^{\mathbb{E},\theta}_{k,0,\mu}\left(\phi^{(2)}_\theta\right) \tag{3.54}$$

converges $\mathbb{P}_\star$-almost surely to some limit $\mathbb{E}_\star[\Delta^{\mathbb{E},\theta}_{0,\infty}(\phi^{(2)}_\theta)]$ not depending on the initial distribution μ, with a uniform control in the neighborhood of $\theta^\star$. We essentially strive for a generalization of Proposition 4 in Douc et al. (2004) to misspecified models, without strong mixing assumptions nor the boundedness of $\phi^{(2)}_\theta$. This can be envisioned by:

1. Translating the forgetting properties of the latent Markov chain (Lemma 17) into forgetting properties for sequences $(\Delta^{\mathbb{E},\theta}_{k,m,\mu}(\varphi_\theta))_{m\in\mathbb{N}}$ of conditional expectations (Lemma 20).

2. Using these forgetting properties to argue that $(\Delta^{\mathbb{E},\theta}_{k,m,\mu}(\varphi_\theta))_{m\in\mathbb{N}}$ is a Cauchy sequence converging to some limit $\Delta^{\mathbb{E},\theta}_{k,\infty}(\varphi_\theta)$ not depending on the initial distribution μ.

3. Approximating $n^{-1}\sum_{k=0}^{n-1}\Delta^{\mathbb{E},\theta}_{k,0,\mu}(\varphi_\theta)$ by $n^{-1}\sum_{k=0}^{n-1}\Delta^{\mathbb{E},\theta}_{k,\infty}(\varphi_\theta)$, and proving uniform continuity results to approximate $n^{-1}\sum_{k=0}^{n-1}\Delta^{\mathbb{E},\theta}_{k,\infty}(\varphi_\theta)$ uniformly by $n^{-1}\sum_{k=0}^{n-1}\Delta^{\mathbb{E},\theta^\star}_{k,\infty}(\varphi_{\theta^\star})$ in the neighborhood of $\theta^\star$.

4. Using ergodic theorems to justify the convergence of $n^{-1}\sum_{k=0}^{n-1}\Delta^{\mathbb{E},\theta^\star}_{k,\infty}(\varphi_{\theta^\star})$ to $\mathbb{E}_\star[\Delta^{\mathbb{E},\theta^\star}_{0,\infty}(\varphi_{\theta^\star})]$.

5. Piecing all the previous results together to approximate $n^{-1}\sum_{k=0}^{n-1}\Delta^{\mathbb{E},\theta}_{k,0,\mu}(\varphi_\theta)$ uniformly by $\mathbb{E}_\star[\Delta^{\mathbb{E},\theta^\star}_{0,\infty}(\varphi_{\theta^\star})]$ in the neighborhood of $\theta^\star$, and apply the result to $\varphi_\theta = \phi^{(2)}_\theta$.

Remark 15. *The vector-valued $\phi^{(2)}_{\theta,i}:\mathbb{X}^2\to\mathbb{R}^{d_\theta}$ can be broken down into its real-valued component functions, which can in turn be decomposed into their positive and negative parts, i.e. $\phi^{(2)}_{\theta,i}=(\phi^{(2)}_{\theta,i,1},...,\phi^{(2)}_{\theta,i,d})^\top$ with $\phi^{(2)}_{\theta,i,j}=\phi^{(2)+}_{\theta,i,j}-\phi^{(2)-}_{\theta,i,j}$ for $j\in[\![1,d]\!]$, so that it will be enough to focus on controlling quantities of the forms*

$$\left|\mathbb{E}_{\theta,\mu}[\varphi_\theta(X_{i-1},X_i)\,|\,y_{-m:k}]-\mathbb{E}_{\theta,\mu'}[\varphi_\theta(X_{i-1},X_i)\,|\,y_{-m:k}]\right|$$

and

$$\left|\mathbb{E}_{\theta,\mu}[\varphi_\theta(X_{i-1},X_i)\,|\,y_{-m:k}]-\mathbb{E}_{\theta,\mu}[\varphi_\theta(X_{i-1},X_i)\,|\,y_{-m:k-1}]\right|$$

for initial distributions μ,μ', and suitably integrable non-negative functions $\varphi_\theta:\mathbb{X}^2\to\mathbb{R}_+$.

We now present several intermediary results that will bring us closer to obtaining the uniform convergence of $n^{-1}\sum_{k=0}^{n-1}\Delta^{\mathbb{E},\theta}_{k,0,\mu}(\varphi_\theta)$ to $\mathbb{E}_\star[\Delta^{\mathbb{E},\theta^\star}_{0,\infty}(\varphi_{\theta^\star})]$ in the neighborhood of $\theta^\star$, for suitable test functions $\varphi:\mathbb{X}^2\to\mathbb{R}_+$. The proofs of these results are postponed to Section 3.6.3.

Lemma 16. *Assume that Conditions C1 to C4 hold. Then, for all probability measures μ on $\mathbb{X}$ satisfying $\mu(D) > 0$, all $(\gamma^-, \gamma^+) \in (0,1)^2$ with $\gamma^- < \gamma^+$, and all $l_0 \in \mathbb{N}$ with $l_0 \geq 1$, there exists a $\mathbb{P}_\star$-almost surely finite random variable $Z > 0$ such that, for all $\theta \in \mathbb{T}$, if*

$$\frac{1}{k_1 + k_2 + 1} \sum_{l=-k_1}^{k_2} \mathbb{1}_K(y_l) \;\geq\; \max\left(1 - \gamma^-, \frac{1+\gamma^+}{2}\right) \tag{3.55}$$

for all $(k_1, k_2) \in \mathbb{N}^2$ with $k_1 \geq l_0$ or $k_2 \geq l_0$, then we have

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \mu(D)\, Z \tag{3.56}$$

for all $(m, n) \in \mathbb{N}^2$ with $n \geq 1$, where D is the set described in Assumption C3 (iii).

Lemma 16 provides a uniform lower bound on $\mu_{\theta,-m:n-1}(D)$ for any $(m, n) \in \mathbb{N}^2$. Although the random variable Z technically depends on l_0, this will be mitigated by the almost sure existence of some large enough L_0, independent of any index, for which the condition in (3.55) will hold $\mathbb{P}^*$-almost surely. The next lemma is a restatement of Proposition 5 from Douc and Moulines (2012) for possibly unbounded integrands.

Lemma 17. *Assume that Conditions C1 to C4 hold. Then, for all probability measures μ and μ' on $\mathbb{X}$, all non-negative functions $f : \mathbb{X} \to \mathbb{R}_+$ and $h : \mathbb{X} \to \mathbb{R}_+$, and all $(\gamma^-, \gamma^+) \in (0,1)^2$ with $\gamma^- < \gamma^+$, there exists $\rho \in (0,1)$ such that, for all $\eta \in (0,1)$, we may find $\tilde{\eta} \in (0,1)$ such that, for all $\theta \in \mathbb{T}$, for all $(m, i) \in \mathbb{N}^2$ with $i \geq 1$, all $y_{-m:i-1} \in \mathbb{Y}^{m+i}$, if*

$$\frac{1}{m+i} \sum_{l=-m}^{i-1} \mathbb{1}_K(y_l) \;\geq\; \max\left(1 - \gamma^-, \frac{1+\gamma^+}{2}\right), \tag{3.57}$$

then we have

$$\begin{aligned}
&\left|(\mu \otimes \mu')\bar{L}_\theta\langle y_{-m:i-1}\rangle (f \otimes h - h \otimes f)\right| \\
&\quad\leq\quad \rho^{m+i}(\mu \otimes \mu')\bar{L}_\theta\langle y_{-m:i-1}\rangle\,(f \otimes h + h \otimes f) \\
&\qquad+\; \eta^{m+i} \int (\mu \otimes \mu')(d\bar{x}_{-m}) \left(\prod_{l=-m}^{i-1} \frac{\bar{L}_\theta\langle y_l\rangle(\bar{x}_l, d\bar{x}_{l+1})}{\tilde{\eta}^{\mathbb{1}_{\bar{C}^c}(\bar{x}_l)\mathbb{1}_K(y_l)}}\right)(f \otimes h + h \otimes f)\,(\bar{x}_i).
\end{aligned} \tag{3.58}$$

In particular, if f and h are bounded, we have

$$\begin{aligned}
&\left|(\mu \otimes \mu')\, \bar{L}_\theta\langle y_{-m:i-1}\rangle (f \otimes h - h \otimes f)\right| \\
&\quad \leq \quad \rho^{m+i} (\mu \otimes \mu') \bar{L}_\theta\langle y_{-m:i-1}\rangle \, (f \otimes h + h \otimes f) \\
&\qquad\qquad + \; \eta^{m+i} \left(\prod_{l=-m}^{i-1} \|L_\theta\langle y_l\rangle(\bullet, \mathbb{X})\|_\infty^2 \right) 2 \, \|f\|_\infty \, \|h\|_\infty \, .
\end{aligned} \tag{3.59}$$

Lemma 17 can be interpreted as a standard forgetting property. It is a consequence of having the latent Markov chain forget its initial distribution exponentially fast as the process evolves. The following lemma provides an upper bound on conditional expectations of the form $\mathbb{E}_{\theta,\mu}\left[\varphi_\theta(X_{i-1}, X_i) \,|\, y_{-m:n-1}\right]$, uniformly in (m, n) for any fixed $i \in [\![-m+1, n]\!]$.

Lemma 18. *Assume that Conditions C1 to C4 hold. For all probability measures μ on $\mathbb{X}$ such that $\mu(D) > 0$, all $(i, l_0) \in (\mathbb{N}^*)^2$ and $(\gamma^-, \gamma^+) \in (0,1)^2$ with $\gamma^- < \gamma^+$, there exists a $\mathbb{P}_\star$-almost surely finite random variable $Z_{i,l_0} > 0$ such that, for all $\theta \in \mathbb{T}$ and all non-negative functions $\varphi_\theta : \mathbb{X}^2 \to \mathbb{R}_+$, if*

$$\frac{1}{k_1 + k_2 + 1} \sum_{l=-k_1}^{k_2} \mathbb{1}_K(y_l) \;\geq\; \max\left(1 - \gamma^-, \, \frac{1+\gamma^+}{2}\right), \tag{3.60}$$

for all $(k_1, k_2) \in \mathbb{N}^2$ with $k_1 \geq l_0$ or $k_2 \geq l_0$, if

$$\min\left(\frac{1}{k}\sum_{l=i+1}^{i+k} \mathbb{1}_K(y_l), \; \frac{1}{k}\sum_{l=i-1-k}^{i-2} \mathbb{1}_K(y_l)\right) \;\geq\; \max\left(1 - \gamma^-, \, \frac{1+\gamma^+}{2}\right), \tag{3.61}$$

for all $k \in \mathbb{N}$ with $k \geq l_0$, and if

$$\mathbb{E}_\star\left[\log^+\left(\sup_{\theta\in\mathbb{T}} \sup_{(x_0,x_1)\in\mathbb{X}^2} \varphi_\theta(x_0, x_1) g_\theta(Y_1|x_1)\right)\right] < +\infty \, , \tag{3.62}$$

then we have

$$\mathbb{E}_{\theta,\mu}\left[\varphi_\theta(X_{i-1}, X_i) \,|\, y_{-m:n-1}\right] \leq Z_{i,l_0} \tag{3.63}$$

for all $(m, n) \in \mathbb{N}^2$ with $-m < i \leq n$.

Lemma 18 will turn out to be of limited use unless we can guarantee that conditions (3.60) and (3.61) hold for every i as soon as there are enough terms in the sums. For all $i_0 \in \mathbb{Z}$, strong stationarity and ergodicity under Conditions C1 and C3(i) guarantee the existence of a $\mathbb{P}_\star$-almost surely finite random integer L_{i_0} such that

$$\frac{1}{k}\sum_{l=i_0+1}^{i_0+k} \mathbb{1}_K(y_l) \;\geq\; \max\left(1-\gamma^-,\, \frac{1+\gamma^+}{2}\right)$$

for all $(\gamma^-,\gamma^+) \in (0,1)^2$ satisfying $1-\mathbb{P}_\star(Y_0 \in K) < \gamma^- < \gamma^+ < 2\mathbb{P}_\star(Y_0 \in K) - 1$ and all $k \geq L_{i_0}$ (Douc and Moulines, 2012, p.2719-2720). Although trivial for a finite number of indexes, the existence of a critical size L_0 that would be common to infinitely many $i_0 \in \mathbb{Z}$ is not guaranteed. This motivates the following assumption.

Assumption A8. *For all $(\gamma^-,\gamma^+) \in (0,1)^2$ satisfying $1-\mathbb{P}_\star(Y_0 \in K) < \gamma^- < \gamma^+ < 2\mathbb{P}_\star(Y_0 \in K) - 1$, there exists a $\mathbb{P}_\star$-almost surely finite random integer L_0 such that*

$$\frac{1}{k}\sum_{l=i+1}^{i+k} \mathbb{1}_K(y_l) \;\geq\; \max\left(1-\gamma^-,\, \frac{1+\gamma^+}{2}\right) \tag{3.64}$$

for all $k \geq L_0$ and all $i \in \mathbb{Z}$.

Assumption A8 is extremely restrictive and constitutes a major limitation of the current derivation. It points at the need for additional conditions on the mixing of the observed process $(Y_t)_{t\in\mathbb{N}}$. A weaker condition could tentatively be the existence of some $\tau^\star > 0$ such that for all $k \in \mathbb{N}^*$ and $t > 0$, $\mathbb{P}_\star(|\frac{1}{k}\sum_{i=1}^k \mathbb{1}_K(Y_i) - \mathbb{P}_\star(Y_0 \in K)| > t) \leq e^{\tau^\star t^2 k}$. These considerations still require further investigation, and the rest of this section should be read as a proof of concept.

Lemma 19. *Assume that Conditions C1 to C4 and Assumption A8 hold. For all probability measures μ and μ' on $\mathbb{X}$ satisfying $\mu(D) > 0$ and $\mu'(D) > 0$, there exists $\rho \in (0,1)$ and a $\mathbb{P}_\star$-almost surely finite random variable $N \in \mathbb{N}$ with $N \geq 2$, such that for all $n \in \mathbb{N}$ with $n \geq N$, all $i \in [\![0, n-1]\!]$, all parameters $\theta \in \mathbb{T}$, all non-negative functions $\varphi_\theta : \mathbb{X}^2 \to \mathbb{R}_+$*

satisfying $\mathbb{E}_\star[\log^+(\sup_{\theta\in\mathbb{T}}\sup_{(x_0,x_1)\in\mathbb{X}^2}\varphi_\theta(x_0,x_1)g_\theta(Y_1|x_1))] < +\infty$, *there exists a* $\mathbb{P}_\star$*-almost surely finite random variable* $Z_i \in \mathbb{R}_+$ *such that for all* $m \in \mathbb{N}$, *we have*

$$\left|\mathbb{E}_{\theta,\mu}\left[\varphi_\theta(X_{i-1},X_i)\,|\,y_{-m:n-1}\right] - \mathbb{E}_{\theta,\mu'}\left[\varphi_\theta(X_{i-1},X_i)\,|\,y_{-m:n-1}\right]\right| \;\leq\; \rho^{m+i}\,Z_i \tag{3.65}$$

if $\lceil (n-m)/2\rceil \leq i \leq n$, *and*

$$\left|\mathbb{E}_{\theta,\mu}\left[\varphi_\theta(X_{i-1},X_i)\,|\,y_{-m:n-1}\right] - \mathbb{E}_{\theta,\mu}\left[\varphi_\theta(X_{i-1},X_i)\,|\,y_{-m:n-2}\right]\right| \;\leq\; \rho^{n-i}\,Z_i \tag{3.66}$$

if $-m \leq i < \lceil (n-m)/2\rceil$, $\mathbb{P}_\star$*-almost surely.*

The majorant Z_i in Lemma 19 essentially plays the role of the term "$2\,\|\varphi_{\theta,i}\|_\infty$" in the proof of Lemma 13 from Douc et al. (2004, p.2294). Its dependence on the index i complicates the analysis and prevents careless uses of telescopic sums over i. We make the following simplifying assumption about the integrability of Z_i.

Assumption A9. *The sequence* $(Z_i)_{i\in\mathbb{N}}$ *is stationary and* $\mathbb{E}_\star[Z_i] < +\infty$ *for all* $i \in \mathbb{N}$.

The stationarity of $(Z_i)_{i\in\mathbb{N}}$ could be justified using Condition C1. On the other hand, the integrability of Z_i will be solely used for ease of exposition: the construction of Z_i involves non-analytical expressions similar to the ones in (3.76) (Section 3.6.3), for which integrability can likely not realistically be checked.

Lemma 20. *For all parameters* $\theta \in \mathbb{T}$, *all collections* $\varphi_\theta = (\varphi_{\theta,i})_{i\in\mathbb{N}}$ *of non-negative functions* $\varphi_{\theta,i} : \mathbb{X}^2 \to \mathbb{R}_+$ *with* $\mathbb{E}_\star[\log^+(\sup_{\theta\in\mathbb{T}}\sup_{(x_0,x_1)\in\mathbb{X}^2}\varphi_{\theta,i}(x_0,x_1)g_\theta(Y_i|x_1))] < +\infty$, *and all probability measures* μ *on* $\mathbb{X}$, *let*

$$\Delta^{\mathbb{E},\theta}_{k,m,\mu}(\varphi_\theta) = \mathbb{E}_{\theta,\mu}\left[\sum_{i=-m}^{k}\varphi_{\theta,i}(X_{i-1},X_i)\,\middle|\,y_{-m:k}\right] - \mathbb{E}_{\theta,\mu}\left[\sum_{i=-m}^{k-1}\varphi_{\theta,i}(X_{i-1},X_i)\,\middle|\,y_{-m:k-1}\right].$$

If Lemma 19 and Assumption A9 hold, then there exist $\mathbb{P}_\star$*-almost surely finite random variables* $K \in \mathbb{N}$ *and* $Z \in \mathbb{R}_+$ *with* $K \geq 2$, *and* $\rho \in (0,1)$, *such that,* $\mathbb{P}_\star$*-almost surely, we have*

$$\sup_{\theta\in\mathbb{T}}\sup_{k>K}\sup_{m\in\mathbb{N}}\left|\Delta^{\mathbb{E},\theta}_{k,m,\mu}(\varphi_\theta) - \Delta^{\mathbb{E},\theta}_{k,m,\mu'}(\varphi_\theta)\right| \;\leq\; Z\max(k^2,m^2)\,\rho^{k+m} \tag{3.67}$$

and

$$\sup_{\theta\in\mathbb{T}}\sup_{k>K}\sup_{\substack{(m',m)\in\mathbb{N}^2\\ m'>m}}\left|\Delta^{\mathbb{E},\theta}_{k,m',\mu}(\varphi_\theta)-\Delta^{\mathbb{E},\theta}_{k,m,\mu}(\varphi_\theta)\right| \leq Z\max(k^2,m^2)\,\rho^{k+m} \tag{3.68}$$

for all probability measures μ *and* μ' *on* $\mathbb{X}$ *satisfying* $\mu(D)>0$ *and* $\mu'(D)>0$.

Equation (3.68) from Lemma 20 implies that $(\Delta^{\mathbb{E},\theta}_{k,m,\mu}(\varphi_\theta))_{m\in\mathbb{N}}$ is a real-valued Cauchy sequence, hence it converges $\mathbb{P}_\star$-almost surely to a limit $\Delta^{\mathbb{E},\theta}_{k,\infty}(\varphi_\theta)$. This limit does not depend on μ thanks to (3.67). Our Lemma 20 serves the same purpose as Lemma 13 in Douc et al. (2004). Obtaining an analog of their Lemma 14 would immediately lead us to a version of their Lemma 15, which in turn would imply our desired convergence result by following the derivation of their Proposition 4 (Douc et al., 2004, p.2295-2297).

3.6.2.2 Second term: sum of smoothing variances

The goal here is to prove that the second term in (3.53) given by

$$\frac{1}{n}\sum_{k=0}^{n-1}\Delta^{\mathbb{V},\theta}_{k,0,\mu}\left(\phi^{(1)}_\theta\right) \tag{3.69}$$

converges $\mathbb{P}_\star$-a.s. to some limit $\mathbb{E}_\star[\Delta^{\mathbb{V},\theta}_{0,\infty}(\phi^{(1)}_\theta)]$ not depending on the initial distribution μ, with a uniform control in the neighborhood of $\theta^\star$. We may follow the same steps as described in Section 3.6.2.1 and essentially aim for a generalization of Proposition 5 in Douc et al. (2004), which we may hope to achieve by deriving an analog of their Lemma 16 by using the same decompositions as in Douc et al. (2004, p.2298-2302). We do not anticipate any major additional difficulties in comparison to the analysis of the first term in Section 3.6.2.1, whose complete treatment would then naturally guide the analysis of the second term.

3.6.3 Proofs of intermediate results

Before proving the results presented in Section 3.6.2, we start by introducing a couple of inequalities (Lemma 21) that will be repeatedly used in the subsequent proofs.

Lemma 21. *For all $\rho \in (0,1)$ and $(k_1,k_2) \in \mathbb{N}^2$ with $1 \leq k_1 \leq k_2$, we have*

$$1 \quad \leq \quad \prod_{l=k_1}^{k_2} \frac{1+\rho^l}{1-\rho^l} \quad \leq \quad \exp\left(\frac{\rho^{k_1}(2-\rho^{k_1})}{(1-\rho)(1-\rho^{k_1})}\right) \quad \leq \quad \exp\left(\frac{2}{(1-\rho)^2}\right). \tag{3.70}$$

We can prove Lemma 21 by using the inequalities $x(1+x)^{-1} \leq \log(1+x) \leq x$, which hold for all $x > -1$, so that for all $(k_1,k_2) \in \mathbb{N}^2$ with $1 \leq k_1 \leq k_2$, we have

$$0 \;\leq\; \log\left(\prod_{l=k_1}^{k_2} \frac{1+\rho^l}{1-\rho^l}\right) \;=\; \sum_{l=k_1}^{k_2} \log\left(1+\rho^l\right) - \sum_{l=k_1}^{k_2} \log\left(1-\rho^l\right) \;\leq\; \sum_{l=k_1}^{k_2} \rho^l + \sum_{l=k_1}^{k_2} \frac{\rho^l}{1-\rho^l}\,.$$

The right-hand side can be further simplified via

$$\begin{aligned}\sum_{l=k_1}^{k_2} \rho^l + \sum_{l=k_1}^{k_2} \frac{\rho^l}{1-\rho^l} \;\leq\; \left(1+\frac{1}{1-\rho^{k_1}}\right) \sum_{l=k_1}^{k_2} \rho^l \;\leq\; \frac{2-\rho^{k_1}}{1-\rho^{k_1}} \sum_{l=k_1}^{+\infty} \rho^l \;&\leq\; \frac{\rho^{k_1}(2-\rho^{k_1})}{(1-\rho)(1-\rho^{k_1})} \\ &\leq\; \frac{2}{(1-\rho)^2}\end{aligned}$$

where the geometric series converges since $\rho \in (0,1)$. By taking the exponential, we get

$$1 \quad \leq \quad \prod_{l=k_1}^{k_2} \frac{1+\rho^l}{1-\rho^l} \quad \leq \quad \exp\left(\frac{\rho^{k_1}(2-\rho^{k_1})}{(1-\rho)(1-\rho^{k_1})}\right) \quad \leq \quad \exp\left(\frac{2}{(1-\rho)^2}\right)$$

which concludes the proof of Lemma 21.

We are now ready to present the proofs of Lemmas 16 to 20.

Proof of Lemma 16. We fix any arbitrary $\theta \in \mathbb{T}$. Under Conditions C3 and C4, we may use the objects defined in Remark 14. By definition of $\mu_{\theta,-m:n-1}$ from (3.45), we have

$$\mu_{\theta,-m:n-1}(D) \;=\; \frac{\mu L_\theta\langle y_{-m:n-1}\rangle(\mathbb{1}_D)}{\mu L_\theta\langle y_{-m:n-1}\rangle(\mathbb{1}_{\mathbb{X}})} \;\geq\; \mu(D)\left(\prod_{l=-m}^{n-1} \frac{1}{V_l}\right)$$

which implies

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \mu(D)\, Z_0 \tag{3.71}$$

for all $(m,n) \in \mathbb{N}^2$ with $m \leq l_0$ and $n \leq l_0$, by defining the $\mathbb{P}_\star$-almost surely positive and finite random variable

$$Z_0 \;=\; \prod_{l=-l_0}^{l_0} \frac{1}{\max(1,V_l)}. \tag{3.72}$$

The difficulty lies in lower-bounding $\mu_{\theta,-m:n-1}$ uniformly in (m,n) if $m > l_0$ or $n > l_0$.

First, let's assume that $m > l_0$. By noticing that

$$\frac{\mu L_\theta\langle y_{-m:n-1}\rangle(\mathbb{1}_D)}{\mu L_\theta\langle y_{-m:n-1}\rangle(\mathbb{1}_\mathbb{X})} = \frac{(\mu L_\theta\langle y_{-m}\rangle)L_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_D)}{(\mu L_\theta\langle y_{-m}\rangle)L_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_\mathbb{X})} = \frac{\mu_{\theta,-m}L_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_D)}{\mu_{\theta,-m}L_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_\mathbb{X})},$$

we can write

$$\begin{aligned}|\mu_{\theta,-m+1:n-1}(D) - \mu_{\theta,-m:n-1}(D)| &= \left|\frac{\mu L_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_D)}{\mu L_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_\mathbb{X})} - \frac{\mu_{\theta,-m}L_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_D)}{\mu_{\theta,-m}L_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_\mathbb{X})}\right| \\ &= \left|\frac{(\mu\otimes\mu_{\theta,-m})\bar{L}_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_D\otimes\mathbb{1}_\mathbb{X} - \mathbb{1}_\mathbb{X}\otimes\mathbb{1}_D)}{(\mu\otimes\mu_{\theta,-m})\bar{L}_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_\mathbb{X}\otimes\mathbb{1}_\mathbb{X})}\right|. \quad (3.73)\end{aligned}$$

Let $\eta_c = \exp(-2c\,\mathbb{E}_\star[\log^+ V_0])$ for some arbitrary constant $c > 1$, so that we have $\eta_c \in (0,1)$ and $\mathbb{E}_\star\left[\log^+(V_0^2)\right] < -\log\eta_c$, with $\eta_c \to 0$ as $c \to +\infty$. Since condition (3.55) and $m > l_0$ guarantee that $(m+n-1)^{-1}\sum_{l=-m+1}^{n-1}\mathbb{1}_K(y_l) \geq \max\left(1-\gamma^-, (1+\gamma^+)/2\right)$, we may apply Lemma 17, so there exists $\rho \in (0,1)$ not depending on (m,n,η_c) such that we have

$$\begin{aligned}&\left|(\mu\otimes\mu_{\theta,-m})\bar{L}_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_D\otimes\mathbb{1}_\mathbb{X} - \mathbb{1}_\mathbb{X}\otimes\mathbb{1}_D)\right| \\ &\qquad\leq \rho^{m+n-1}(\mu\otimes\mu_{\theta,-m})\bar{L}_\theta\langle y_{-m+1:n-1}\rangle\left(\mathbb{1}_D\otimes\mathbb{1}_\mathbb{X} + \mathbb{1}_\mathbb{X}\otimes\mathbb{1}_D\right) \\ &\qquad\qquad + \eta_c^{m+n-1}\,2\prod_{l=-m+1}^{n-1}\|L_\theta\langle y_l\rangle(\cdot,\mathbb{X})\|_\infty^2.\end{aligned}$$

Plugging this back into (3.73) leads to

$$\begin{aligned}&|\mu_{\theta,-m+1:n-1}(D) - \mu_{\theta,-m:n-1}(D)| \quad (3.74)\\ &\qquad\leq \rho^{m+n-1}\left(\mu_{\theta,-m+1:n-1}(D) + \mu_{\theta,-m:n-1}(D)\right) \\ &\qquad\qquad + \eta_c^{m+n-1}\frac{2\prod_{l=-m+1}^{n-1}\|L_\theta\langle y_l\rangle(\cdot,\mathbb{X})\|_\infty^2}{(\mu\otimes\mu_{\theta,-m})\bar{L}_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_\mathbb{X}\otimes\mathbb{1}_\mathbb{X})}.\end{aligned}$$

By Remark 14, we have the following inequalities:

$$\begin{aligned}
\prod_{l=-m+1}^{n-1} \|L_\theta\langle y_l\rangle(\cdot,\mathbb{X})\|_\infty^2 &\leq \left(\prod_{l=-m+1}^{n-1} G_l^2\right), \\
\mu\, L_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_\mathbb{X}) &\geq \mu(D)\left(\prod_{l=-m+1}^{n-1} F_l\right), \\
\mu_{\theta,-m}\, L_\theta\langle y_{-m+1:n-1}\rangle(\mathbb{1}_\mathbb{X}) &\geq \frac{\mu(D)}{V_{-m}}\left(\prod_{l=-m+1}^{n-1} F_l\right).
\end{aligned}$$

Applying these inequalities to (3.74) leads to

$$\begin{aligned}
&|\mu_{\theta,-m+1:n-1}(D) - \mu_{\theta,-m:n-1}(D)| \\
&\quad\leq \rho^{m+n-1}\left(\mu_{\theta,-m+1:n-1}(D) + \mu_{\theta,-m:n-1}(D)\right) + \eta_c^{m+n-1}\,\frac{2\,V_{-m}\,\prod_{l=-m+1}^{n-1} V_l^2}{\mu(D)^2}
\end{aligned}$$

which implies

$$\begin{aligned}
\mu_{\theta,-m:n-1}(D) &\geq \left(\frac{1-\rho^{m+n-1}}{1+\rho^{m+n-1}}\right)\mu_{\theta,-m+1:n-1}(D) - \left(\frac{\eta_c^{m+n-1}}{1+\rho^{m+n-1}}\right)\frac{2\,V_{-m}\,\prod_{l=-m+1}^{n-1} V_l^2}{\mu(D)^2} \\
&\geq \left(\frac{1-\rho^{m+n-1}}{1+\rho^{m+n-1}}\right)\mu_{\theta,-m+1:n-1}(D) - \eta_c^{m+n-1}\,\frac{2\,V_{-m}\,\prod_{l=-m+1}^{n-1} V_l^2}{\mu(D)^2}. \qquad (3.75)
\end{aligned}$$

Under Conditions C1 to C4, the processes $(\max(1,V_l))_{l\in\mathbb{Z}}$ and $(V_l^2)_{l\in\mathbb{Z}}$ are stationary, while also satisfying $\mathbb{E}_\star\left[\log^+(\max(1,V_0))\right] < +\infty$ and $\mathbb{E}_\star\left[\log^+ V_0^2\right] < +\infty$. By using Lemma 6 from Douc and Moulines (2012), we may define $\xi_c = \eta_c/(\alpha\tilde{\alpha}) \in (0,1)$ for any $\eta_c < \alpha < \exp(-2\,\mathbb{E}_\star[\log^+ V_0]) < 1$ and $(\eta_c/\alpha) < \tilde{\alpha} < 1$. In particular, by choosing $\alpha = \eta_c^{1/4}$ and $\tilde{\alpha} = \eta_c^{1/4}$ with $c \geq 5$, we get $\xi_c = \eta_c^{1/2}$ and we may define

$$\tilde{Z}_c = 1 + \left(\sup_{k_1\geq 0} \eta_c^{k_1/4}\bar{V}_{-k_1}\right)\left(\sup_{k_1\geq 1}\prod_{i=-k_1+1}^{0}(V_i^2\eta_c^{1/4})\right)\left(\sup_{k_2\geq 1}\prod_{i=1}^{k_2-1}(V_i^2\eta_c^{1/4})\right)\left(\sup_{k_2\geq 1}\eta_c^{k_2/4}\bar{V}_{k_2}\right) \qquad (3.76)$$

where $\bar{V}_k = \max(1,V_k)$, so that, $\mathbb{P}_\star$-almost surely, we have $0 < \tilde{Z}_c < +\infty$, and

$$\eta_c^{k_1+k_2}\,\max(1,V_{-k_1})\left(\prod_{l=-k_1+1}^{k_2} V_l^2\right)\max(1,V_{k_2}) \leq \xi_c^{k_1+k_2}\tilde{Z}_c \leq \xi_c^{k_1+k_2}\tilde{Z}_5$$

for all $(k_1, k_2) \in \mathbb{N}^2$ with $k_1 + k_2 \geq 1$ and all $c \geq 5$. This implies both

$$\eta_c^{m+n-1} V_{-m} \prod_{l=-m+1}^{n-1} V_l^2 \quad \leq \quad \xi_c^{m+n-1} \tilde{Z}_5 \tag{3.77}$$

and

$$\eta_c^{m+n-1} V_{n-1} \prod_{l=-m}^{n-2} V_l^2 \quad \leq \quad \xi_c^{m+n-1} \tilde{Z}_5 \tag{3.78}$$

for all $c \geq 5$. Plugging (3.77) into (3.75) yields

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \left(\frac{1-\rho^{m+n-1}}{1+\rho^{m+n-1}}\right) \mu_{\theta,-m+1:n-1}(D) \;-\; \xi_c^{m+n-1} \frac{2\,\tilde{Z}_5}{\mu(D)^2}.$$

Under condition (3.55), we may apply this reasoning recursively on $\mu_{\theta,k:n-1}(D)$ in a forward fashion for $k \in [\![-m, -l_0]\!]$. This leads to

$$\begin{aligned}
\mu_{\theta,-m:n-1}(D) \;&\geq\; \left(\prod_{l=l_0+n}^{m+n-1} \frac{1-\rho^l}{1+\rho^l}\right) \mu_{\theta,-l_0:n-1}(D) \;-\; \left(\sum_{l=l_0+n}^{m+n-1} \xi_c^l \prod_{j=l+1}^{m+n-1} \frac{1-\rho^j}{1+\rho^j}\right) \frac{2\,\tilde{Z}_5}{\mu(D)^2} \\
&\geq\; \left(\prod_{l=l_0+n}^{m+n-1} \frac{1-\rho^l}{1+\rho^l}\right) \mu_{\theta,-l_0:n-1}(D) \;-\; \left(\sum_{l=l_0}^{+\infty} \xi_c^l\right) \frac{2\,\tilde{Z}_5}{\mu(D)^2}.
\end{aligned} \tag{3.79}$$

Using Lemma 21, we get

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \exp\left(-\frac{2}{(1-\rho)^2}\right) \mu_{\theta,-l_0:n-1}(D) \;-\; \left(\frac{\xi_c^{l_0}}{1-\xi_c}\right) \frac{2\,\tilde{Z}_5}{\mu(D)^2}. \tag{3.80}$$

If $n \leq l_0$, then we have

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \exp\left(-\frac{2}{(1-\rho)^2}\right) \mu(D)\, Z_0 \;-\; \left(\frac{\xi_c^{l_0}}{1-\xi_c}\right) \frac{2\,\tilde{Z}_5}{\mu(D)^2}. \tag{3.81}$$

One key point is that we may choose ξ_c independently from Z_0 and $\tilde{Z}_5$. Since $\xi_c^{l_0}(1-\xi_c)^{-1} \to 0$ as $c \to +\infty$ (i.e. $\xi_c \to 0$), then for every finite realization $(z_0, \tilde{z}_5)$ of $(Z_0, \tilde{Z}_5)$, there exists some $c_0 \geq 5$ large enough such that

$$\exp\left(-\frac{2}{(1-\rho)^2}\right) z_0 \;-\; \left(\frac{\xi_{c_0}^{l_0}}{1-\xi_{c_0}}\right) \frac{2\,\tilde{z}_5}{\mu(D)^3} \;>\; 0. \tag{3.82}$$

By choosing such a $c_0 \geq 5$ for each realized value of $(Z_0, \tilde{Z}_5)$, the left-hand side of (3.82) defines the realization of some $\mathbb{P}_\star$-almost surely finite and strictly positive random variable Z_1. Note that Z_1 does not depend on (m, n), and we have

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \mu(D)\, Z_1 \tag{3.83}$$

by construction, for all $(m, n) \in \mathbb{N}^2$ with $m > l_0$ and $n \leq l_0$.

If $n > l_0$, we proceed similarly by deriving a backward recursion. We have

$$\begin{aligned}
|\mu_{\theta,-m:n-1}(D) - \mu_{\theta,-m:n-2}(D)| \;&=\; \left| \frac{\mu L_\theta\langle y_{-m:n-1}\rangle(\mathbb{1}_D)}{\mu L_\theta\langle y_{-m:n-1}\rangle(\mathbb{1}_\mathbb{X})} - \frac{\mu L_\theta\langle y_{-m:n-2}\rangle(\mathbb{1}_D)}{\mu L_\theta\langle y_{-m:n-2}\rangle(\mathbb{1}_\mathbb{X})} \right| \\
&=\; \left| \frac{(\mu\otimes\mu)\bar{L}_\theta\langle y_{-m:n-2}\rangle(L_\theta\langle y_{n-1}\rangle\mathbb{1}_D\otimes\mathbb{1}_\mathbb{X} - \mathbb{1}_\mathbb{X}\otimes L_\theta\langle y_{n-1}\rangle\mathbb{1}_D)}{(\mu\otimes\mu)\bar{L}_\theta\langle y_{-m:n-2}\rangle(L_\theta\langle y_{n-1}\rangle\mathbb{1}_\mathbb{X}\otimes\mathbb{1}_\mathbb{X})} \right|.
\end{aligned}$$

Since $n > l_0$, we have $(m+n-1)^{-1}\sum_{l=-m}^{n-2}\mathbb{1}_K(y_l) \geq \max\left(1-\gamma^-, (1+\gamma^+)/2\right)$ thanks to Condition (3.55), so we may apply Lemma 17 again and obtain

$$\begin{aligned}
&\left|(\mu\otimes\mu)\;\bar{L}_\theta\langle y_{-m:n-2}\rangle(L_\theta\langle y_{n-1}\rangle\mathbb{1}_D\otimes\mathbb{1}_\mathbb{X} - \mathbb{1}_\mathbb{X}\otimes L_\theta\langle y_{n-1}\rangle\mathbb{1}_D)\right| \\
&\qquad\leq\; \rho^{m+n-1}(\mu\otimes\mu)\bar{L}_\theta\langle y_{-m:n-2}\rangle(L_\theta\langle y_{n-1}\rangle\mathbb{1}_D\otimes\mathbb{1}_\mathbb{X} + \mathbb{1}_\mathbb{X}\otimes L_\theta\langle y_{n-1}\rangle\mathbb{1}_D) \\
&\qquad\qquad+\; \eta_c^{m+n-1}\; 2\,G_{n-1}\prod_{l=-m}^{n-2}\|L_\theta\langle y_l\rangle(\cdot,\mathbb{X})\|_\infty^2 .
\end{aligned}$$

This leads to

$$\begin{aligned}
&|\mu_{\theta,-m:n-1}\;(D) - \;\mu_{\theta,-m:n-2}(D)| \\
&\qquad\leq\; \rho^{m+n-1}\left(\mu_{\theta,-m:n-1}(D) + \mu_{\theta,-m:n-2}(D)\right) \\
&\qquad\qquad+\; \eta_c^{m+n-1}\,\frac{2\,G_{n-1}\prod_{l=-m}^{n-2}\|L_\theta\langle y_l\rangle(\cdot,\mathbb{X})\|_\infty^2}{(\mu\otimes\mu)\bar{L}_\theta\langle y_{-m:n-2}\rangle(L_\theta\langle y_{n-1}\rangle\mathbb{1}_\mathbb{X}\otimes\mathbb{1}_\mathbb{X})}.
\end{aligned}$$

which further yields

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \left(\frac{1-\rho^{m+n-1}}{1+\rho^{m+n-1}}\right)\mu_{\theta,-m:n-2}(D) \;-\; \eta_c^{m+n-1}\,\frac{2\,V_{n-1}\prod_{l=-m}^{n-2}V_l^2}{\mu(D)^2}.$$

By using (3.78), this implies

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \left(\frac{1-\rho^{m+n-1}}{1+\rho^{m+n-1}}\right)\mu_{\theta,-m:n-2}(D) \;-\; \xi_c^{m+n-1}\,\frac{2\,\tilde{Z}_5}{\mu(D)^2}.$$

Under condition (3.55), we may apply this reasoning recursively on $\mu_{\theta,-m:k}(D)$ in a backward fashion for $k \in [\![l_0, n-1]\!]$, which leads to

$$\begin{aligned}\mu_{\theta,-m:n-1}(D) \;&\geq\; \left(\prod_{l=m+l_0}^{m+n-1}\frac{1-\rho^l}{1+\rho^l}\right)\mu_{\theta,-m:l_0-1}(D) \;-\; \left(\sum_{l=m+l_0}^{m+n-1}\xi_c^l\prod_{j=l+1}^{m+n-1}\frac{1-\rho^j}{1+\rho^j}\right)\frac{2\,\tilde{Z}_5}{\mu(D)^2}\\ &\geq\; \left(\prod_{l=m+l_0}^{m+n-1}\frac{1-\rho^l}{1+\rho^l}\right)\mu_{\theta,-m:l_0-1}(D) \;-\; \left(\sum_{l=l_0}^{+\infty}\xi_c^l\right)\frac{2\,\tilde{Z}_5}{\mu(D)^2}\end{aligned}$$

Using again Lemma 21, we get

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \exp\left(-\frac{2}{(1-\rho)^2}\right)\mu_{\theta,-m:l_0-1}(D) \;-\; \left(\frac{\xi_c^{l_0}}{1-\xi_c}\right)\frac{2\,\tilde{Z}_5}{\mu(D)^2}. \tag{3.84}$$

If $m \leq l_0$, then we have $\mu_{\theta,-m:l_0-1}(D) > \mu(D)\,Z_0$ and we retrieve (3.81), which implies

$$\mu_{\theta,-m:n-1}(D) > \mu(D)\,Z_1 \tag{3.85}$$

for all $(m,n) \in \mathbb{N}^2$ with $m \leq l_0$ and $n > l_0$, similarly to (3.83). For the general case when both $m > l_0$ and $n > l_0$, we start by using the forward recursion in (3.80) to obtain

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \exp\left(-\frac{2}{(1-\rho)^2}\right)\mu_{\theta,-l_0:n-1}(D) \;-\; \left(\frac{\xi_c^{l_0}}{1-\xi_c}\right)\frac{2\,\tilde{Z}_5}{\mu(D)^2}. \tag{3.86}$$

We then bound from below the term $\mu_{\theta,-l_0:n-1}(D)$ by applying the backward recursion in (3.84), so that (3.86) further implies

$$\begin{aligned}\mu_{\theta,-m:n-1}(D) \;&\geq\; \exp\left(-\frac{4}{(1-\rho)^2}\right)\mu_{\theta,-l_0:l_0-1}(D) - \left[1+\exp\left(-\frac{2}{(1-\rho)^2}\right)\right]\left(\frac{\xi_c^{l_0}}{1-\xi_c}\right)\frac{2\,\tilde{Z}_5}{\mu(D)^2}.\\ &\geq\; \exp\left(-\frac{4}{(1-\rho)^2}\right)Z_0 - \left[1+\exp\left(-\frac{2}{(1-\rho)^2}\right)\right]\left(\frac{\xi_c^{l_0}}{1-\xi_c}\right)\frac{2\,\tilde{Z}_5}{\mu(D)^2}.\end{aligned}$$

Similarly to (3.82), we may choose ξ_c small enough, independently from Z_0 and $\tilde{Z}_5$, such that

$$\exp\left(-\frac{4}{(1-\rho)^2}\right)z_0 - \left[1+\exp\left(-\frac{2}{(1-\rho)^2}\right)\right]\left(\frac{\xi_{c_1}^{l_0}}{1-\xi_{c_1}}\right)\frac{2\,\tilde{z}_5}{\mu(D)^2} \;>\; 0$$

for some appropriately chosen $c_1 \geq 5$ tailored to each realized value $(z_0, \tilde{z}_5)$ of $(Z_0, \tilde{Z}_5)$. The left hand side then defines some $\mathbb{P}_\star$-almost surely finite and strictly positive random variable Z_2, which does not depend on (m, n), and we have

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \mu(D)\, Z_2 \tag{3.87}$$

for all $(m, n) \in \mathbb{N}^2$ with $m > l_0$ and $n > l_0$. By putting together (3.71), (3.83), (3.85), and (3.87), we finally get

$$\mu_{\theta,-m:n-1}(D) \;\geq\; \mu(D)\, Z_{l_0}$$

for all $(m, n) \in \mathbb{N}$ with $n \geq 1$, where $Z_{l_0} = \min(Z_0, Z_1, Z_2)$ is positive and finite $\mathbb{P}_\star$-almost surely. This concludes the proof of Lemma 16. □

Proof of Lemma 17. Under Condition C3, there exists a Doeblin set C for which

$$\rho_C = 1 - \left(\inf_{y \in K} \frac{\epsilon_C^-(y)}{\epsilon_C^+(y)} \right)^2 \qquad \text{and} \qquad M_{\bar{C}^c}(\bar{x}_{-m:i-1}) = \sum_{l=-m}^{i-1} \mathbb{1}_{\bar{C}^c}(\bar{x}_l)$$

can be defined, where $\bar{C} = C \times C$ and $\bar{C}^c$ denotes its complement. Under Condition C3 and condition (3.57), equations (31), (34), and (35) from the proof of Proposition 5 in Douc and Moulines (2012) hold for all non-negative real-valued functions f and h (even unbounded ones), so that for all $\beta \in (\gamma^-, \gamma^+)$, we have

$$\begin{aligned}
&\left| (\mu \otimes \mu') \bar{L}_\theta \langle y_{-m:i-1} \rangle (f \otimes h - h \otimes f) \right| \\
&\leq\; \rho_C^{\lfloor (i+m)(\beta - \gamma^-) \rfloor} (\mu \otimes \mu') \bar{L}_\theta \langle y_{-m:i-1} \rangle \, (f \otimes h + h \otimes f) \qquad (3.88) \\
&\quad + \int (\mu \otimes \mu')(d\bar{x}_0) \mathbb{1}_{\{M_{\bar{C}^c}(\bar{x}_{-m:i-1}) \geq \frac{(i+m) - \lfloor (i+m)\beta \rfloor}{2}\}} \left[\prod_{l=-m}^{i-1} \bar{L}_\theta \langle y_l \rangle (\bar{x}_l, d\bar{x}_{l+1}) \right] (f \otimes h + h \otimes f)\,(\bar{x}_i).
\end{aligned}$$

From (3.57), we have $\sum_{l=-m}^{i-1} \mathbb{1}_K(y_l) \geq ((i+m) + \lfloor (i+m)\gamma^+ \rfloor)/2$. Since $\mathbb{1}_K(y_l) = 1 - \mathbb{1}_{K^c}(y_l)$, this leads to $-\sum_{l=-m}^{i-1} \mathbb{1}_{K^c}(y_l) \geq (\lfloor (i+m)\gamma^+ \rfloor - (i+m))/2$. When the indicator in the above

integral is non-zero, i.e. when $M_{\bar{C}^c}(\bar{x}_{-m:i-1}) \geq ((i+m) - \lfloor (i+m)\beta \rfloor)/2$, we have

$$\sum_{l=-m}^{i-1} \mathbb{1}_{\bar{C}^c}(\bar{x}_l)\mathbb{1}_K(y_l) \geq \sum_{l=-m}^{i-1} \mathbb{1}_{\bar{C}^c}(\bar{x}_l) - \sum_{l=-m}^{i-1} \mathbb{1}_{K^c}(y_l) \geq \frac{\lfloor (i+m)\gamma^+ \rfloor - \lfloor (i+m)\beta \rfloor}{2} \geq \frac{\lfloor (i+m)(\gamma^+ - \beta) \rfloor}{2}.$$

This in turns implies that, for all $\tilde{\eta} \in (0,1)$, we have

$$\left(\tilde{\eta}^{\sum_{l=-m}^{i-1} \mathbb{1}_{\bar{C}^c}(\bar{x}_l)\mathbb{1}_K(y_l)}\right) \mathbb{1}_{\{M_{\bar{C}^c}(\bar{x}_{-m:i-1}) \geq \frac{(i+m) - \lfloor (i+m)\beta \rfloor}{2}\}} \leq \tilde{\eta}^{\frac{\lfloor (i+m)(\gamma^+ - \beta) \rfloor}{2}} \leq \left(\tilde{\eta}^{\frac{\lfloor \gamma^+ - \beta \rfloor}{2}}\right)^{i+m}$$

where the last inequality follows from $0 \leq (i+m)\lfloor \gamma^+ - \beta \rfloor \leq \lfloor (i+m)(\gamma^+ - \beta) \rfloor$. Similarly, we also have $0 \leq (i+m)\lfloor \beta - \gamma^- \rfloor \leq \lfloor (i+m)(\beta - \gamma^-) \rfloor$ and $\rho_C \in (0,1)$ by Condition C3, so that $\rho_C^{\lfloor \beta - \gamma^- \rfloor (i+m)} \leq \rho_C^{\lfloor (i+m)(\beta - \gamma^-) \rfloor}$. By choosing $\beta = (\gamma^- + \gamma^+)/2$, we may rewrite (3.88) as

$$\begin{aligned}
&\left| (\mu \otimes \mu')\bar{L}_\theta\langle y_{-m:i-1}\rangle (f \otimes h - h \otimes f) \right| \\
&\leq \left(\rho_C^{\lfloor (\gamma^+ - \gamma^-)/2 \rfloor}\right)^{i+m} (\mu \otimes \mu')\bar{L}_\theta\langle y_{-m:i-1}\rangle (f \otimes h + h \otimes f) \\
&\quad + \left(\tilde{\eta}^{\frac{1}{2}\lfloor (\gamma^+ - \gamma^-)/2 \rfloor}\right)^{i+m} \int (\mu \otimes \mu')(d\bar{x}_0) \left[\prod_{l=-m}^{i-1} \frac{\bar{L}_\theta\langle y_l\rangle(\bar{x}_l, d\bar{x}_{l+1})}{\tilde{\eta}^{\mathbb{1}_{\bar{C}^c}(\bar{x}_l)\mathbb{1}_K(y_l)}}\right] (f \otimes h + h \otimes f)(\bar{x}_i)
\end{aligned}$$

which proves (3.58) by setting $\rho = \rho_C^{\lfloor (\gamma^+ - \gamma^-)/2 \rfloor} \in (0,1)$ and $\eta = \tilde{\eta}^{\frac{1}{2}\lfloor (\gamma^+ - \gamma^-)/2 \rfloor} \in (0,1)$. The result holds for any $\eta \in (0,1)$ and its corresponding $\tilde{\eta}$, with ρ not depending on $(\eta, \tilde{\eta})$. When f and h are bounded, we have $\|f \otimes h + h \otimes f\|_\infty \leq 2\|f\|_\infty \|h\|_\infty$, so that the result follows from the proof of Proposition 5 in Douc and Moulines (2012). □

Proof of Lemma 18. Using (3.51), we introduce the notation

$$I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) = \mathbb{E}_{\theta,\mu}\left[\varphi_\theta(X_{i-1}, X_i) \,|\, y_{-m:n-1}\right] = \frac{\mu L_\theta\langle y_{-m:i-2}\rangle (L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n})}{\mu L_\theta\langle y_{-m:i-2}\rangle (h_{\theta,i-1:n})}.$$

As per Remark 14, under the given assumptions we have

$$I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) \leq \left(\prod_{\substack{l=-m \\ l \neq i-1}}^{n-1} V_l\right) \frac{V_{i-1}^\varphi}{\mu(D)} \leq \frac{1}{\mu(D)} \left(\prod_{l=-m}^{n-1} \tilde{V}_l\right) \leq +\infty \tag{3.89}$$

for all $-m < i \leq n$, all $\theta \in \mathbb{T}$, and all μ such that $\mu(D) > 0$. In particular, by defining

$$Z_0 \;=\; \frac{1}{\mu(D)} \left(\prod_{l=i-l_0-1}^{i+l_0+1} \max(1, \widetilde{V}_l) \right), \tag{3.90}$$

we then have, for all $(m, n) \in \mathbb{N}^2$ with $i - l_0 - 1 \leq -m < n \leq i + l_0 + 1$,

$$I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) \;\leq\; Z_0. \tag{3.91}$$

The challenge is to bound $I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta)$ uniformly in (m, n). First, let $i + l_0 + 1 < n$. We have

$$\begin{aligned}
&I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) - I^{\theta,\mu,i}_{-m:n-2}(\varphi_\theta) \\
&= \frac{\mu L_\theta \langle y_{-m:i-2} \rangle (L_\theta^{\varphi_\theta} \langle y_{i-1} \rangle h_{\theta,i:n})}{\mu L_\theta \langle y_{-m:i-2} \rangle \, (h_{\theta,i-1:n})} - \frac{\mu L_\theta \langle y_{-m:i-2} \rangle (L_\theta^{\varphi_\theta} \langle y_{i-1} \rangle h_{\theta,i:n-1})}{\mu L_\theta \langle y_{-m:i-2} \rangle \, (h_{\theta,i-1:n-1})} \\
&= \frac{(\mu \otimes \mu) \bar{L}_\theta \langle y_{-m:i-2} \rangle \, (L_\theta^{\varphi_\theta} \langle y_{i-1} \rangle h_{\theta,i:n} \otimes h_{\theta,i-1:n-1} - h_{\theta,i-1:n} \otimes L_\theta^{\varphi_\theta} \langle y_{i-1} \rangle h_{\theta,i:n-1})}{(\mu \otimes \mu) \bar{L}_\theta \langle y_{-m:i-2} \rangle \, (h_{\theta,i-1:n} \otimes h_{\theta,i-1:n-1})} \\
&= \frac{(\mu \otimes \mu) \bar{L}_\theta \langle y_{-m:i-2} \rangle \, (L_\theta^{\varphi_\theta} \langle y_{i-1} \rangle h_{\theta,i:n} \otimes L_\theta \langle y_{i-1} \rangle h_{\theta,i:n-1} - L_\theta \langle y_{i-1} \rangle h_{\theta,i:n} \otimes L_\theta^{\varphi_\theta} \langle y_{i-1} \rangle h_{\theta,i:n-1})}{(\mu \otimes \mu) \bar{L}_\theta \langle y_{-m:i-2} \rangle \, (h_{\theta,i-1:n} \otimes h_{\theta,i-1:n-1})} \\
&= \frac{(\mu \otimes \mu) \bar{L}_\theta \langle y_{-m:i-2} \rangle (L_\theta^{\varphi_\theta} \langle y_{i-1} \rangle \otimes L_\theta \langle y_{i-1} \rangle)(h_{\theta,i:n} \otimes h_{\theta,i:n-1} - h_{\theta,i:n-1} \otimes h_{\theta,i:n})}{(\mu \otimes \mu) \bar{L}_\theta \langle y_{-m:i-2} \rangle \, (h_{\theta,i-1:n} \otimes h_{\theta,i-1:n-1})}.
\end{aligned} \tag{3.92}$$

We can write

$$(h_{\theta,i:n} \otimes h_{\theta,i:n-1} - h_{\theta,i:n-1} \otimes h_{\theta,i:n}) = \bar{L}_\theta \langle y_i \rangle (h_{\theta,i+1:n} \otimes h_{\theta,i+1:n-1} - h_{\theta,i+1:n-1} \otimes h_{\theta,i+1:n}) \tag{3.93}$$

where we have

$$(h_{\theta,i+1:n} \otimes h_{\theta,i+1:n-1} - h_{\theta,i+1:n-1} \otimes h_{\theta,i+1:n})(x, x') = (\delta_x \otimes \delta_{x'}) \bar{L}_\theta \langle y_{i+1:n-2} \rangle (g_{\theta,n-1} \otimes \mathbb{1}_{\mathbb{X}} - \mathbb{1}_{\mathbb{X}} \otimes g_{\theta,n-1})$$

for all $(x, x') \in \mathbb{X}^2$, thanks to (3.48). Under Condition C4, we may assume without loss of generality that the observation y_{n-1} is such that $g_{\theta,n-1}$ is bounded, as per Remark 14. This boundedness combined with the condition $(n - i - 2)^{-1} \sum_{l=i+1}^{n-2} \mathbb{1}_K(y_l) \geq \max\left(1 - \gamma^-, (1 + \gamma^+)/2\right)$

for $n > i + l_0 + 1$ from (3.61) allows us to apply Lemma 17, which yields

$$\begin{aligned}
&|h_{\theta,i+1:n} \otimes h_{\theta,i+1:n-1} - h_{\theta,i+1:n-1} \otimes h_{\theta,i+1:n}|\,(x,x') \\
&\qquad \leq \quad \rho^{n-i-2}(\delta_x \otimes \delta_{x'})\bar{L}_\theta\langle y_{i+1:n-2}\rangle\,(g_{\theta,n-1} \otimes \mathbb{1}_{\mathbb{X}} + \mathbb{1}_{\mathbb{X}} \otimes g_{\theta,n-1}) \\
&\qquad\qquad + \quad 2\,\eta^{n-i-2}\,\|g_{\theta,n-1}\|_\infty \prod_{l=i+1}^{n-2} \|L_\theta\langle y_l\rangle(\cdot,\mathbb{X})\|_\infty^2
\end{aligned}$$

for some $\rho \in (0,1)$, all $\eta > 0$, and all $(x,x') \in \mathbb{X}^2$. This implies

$$\begin{aligned}
&|h_{\theta,i+1:n} \otimes h_{\theta,i+1:n-1} - h_{\theta,i+1:n-1} \otimes h_{\theta,i+1:n}| \\
&\leq \quad \rho^{n-i-2}\,(h_{\theta,i+1:n} \otimes h_{\theta,i+1:n-1} + h_{\theta,i+1:n-1} \otimes h_{\theta,i+1:n}) \;+\; 2\,\eta^{n-i-2}\,G_{n-1}\left(\prod_{l=i+1}^{n-2} G_l^2\right)(\mathbb{1}_{\mathbb{X}} \otimes \mathbb{1}_{\mathbb{X}})\,.
\end{aligned}$$

Combining this inequality with (3.93) yields

$$\begin{aligned}
&|h_{\theta,i:n} \otimes h_{\theta,i:n-1} - h_{\theta,i:n-1} \otimes h_{\theta,i:n}| \qquad (3.94)\\
&\leq \quad \rho^{n-i-2}\,(h_{\theta,i:n} \otimes h_{\theta,i:n-1} + h_{\theta,i:n-1} \otimes h_{\theta,i:n}) \;+\; 2\,\eta^{n-i-2}\,G_{n-1}\left(\prod_{l=i+1}^{n-2} G_l^2\right)(h_{\theta,i:i+1} \otimes h_{\theta,i:i+1})
\end{aligned}$$

since $\bar{L}_\theta\langle y_i\rangle(\mathbb{1}_{\mathbb{X}} \otimes \mathbb{1}_{\mathbb{X}}) = (h_{\theta,i:i+1} \otimes h_{\theta,i:i+1})$. Plugging (3.94) back into (3.92) leads to

$$\begin{aligned}
&I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) - I^{\theta,\mu,i}_{-m:n-2}(\varphi_\theta) \quad \leq \quad \rho^{n-i-2}\left(I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) + I^{\theta,\mu,i}_{-m:n-2}(\varphi_\theta)\right) \\
&\quad + \; 2\,\eta^{n-i-2}\,G_{n-1}\left(\prod_{l=i+1}^{n-2} G_l^2\right)\frac{(\mu\otimes\mu)\bar{L}_\theta\langle y_{-m:i-2}\rangle(L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle \otimes L_\theta\langle y_{i-1}\rangle)\,(h_{\theta,i:i+1} \otimes h_{\theta,i:i+1})}{(\mu\otimes\mu)\bar{L}_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n} \otimes h_{\theta,i-1:n-1})}
\end{aligned}$$

which simplifies to

$$\begin{aligned}
&I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) \quad \leq \quad \left(\frac{1+\rho^{n-i-2}}{1-\rho^{n-i-2}}\right) I^{\theta,\mu,i}_{-m:n-2}(\varphi_\theta) \qquad (3.95)\\
&\quad + \quad \left(\frac{2\,\eta^{n-i-2}\,G_{n-1}}{1-\rho^{n-i-2}}\right)\left(\prod_{l=i+1}^{n-2} G_l^2\right)\frac{(\mu\otimes\mu)\bar{L}_\theta\langle y_{-m:i-2}\rangle(L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:i+1} \otimes L_\theta\langle y_{i-1}\rangle h_{\theta,i:i+1})}{(\mu\otimes\mu)\bar{L}_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n} \otimes h_{\theta,i-1:n-1})}.
\end{aligned}$$

The last ratio on the right hand side can be written as

$$\begin{aligned}
&\frac{(\mu\otimes\mu)\bar{L}_\theta\langle y_{-m:i-2}\rangle(L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:i+1}\otimes L_\theta\langle y_{i-1}\rangle h_{\theta,i:i+1})}{(\mu\otimes\mu)\bar{L}_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n}\otimes h_{\theta,i-1:n-1})}\\
&= \frac{\mu L_\theta\langle y_{-m:i-2}\rangle\,(L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:i+1})}{\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:i+1})}\;\frac{\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:i+1})}{\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n})}\;\frac{\mu L_\theta\langle y_{-m:i-2}\rangle\,(L_\theta\langle y_{i-1}\rangle h_{\theta,i:i+1})}{\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n-1})}\\
&= \frac{\mu L_\theta\langle y_{-m:i-2}\rangle\,(L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:i+1})}{\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:i+1})}\;\frac{\mu L_\theta\langle y_{-m:i}\rangle\,(\mathbb{1}_{\mathbb{X}})}{\mu L_\theta\langle y_{-m:i}\rangle\,(h_{\theta,i+1:n})}\;\frac{\mu L_\theta\langle y_{-m:i}\rangle\,(\mathbb{1}_{\mathbb{X}})}{\mu L_\theta\langle y_{-m:i}\rangle\,(h_{\theta,i+1:n-1})}\\
&= I_{-m:i}^{\theta,\mu,i}(\varphi_\theta)\;\frac{1}{\mu_{\theta,-m:i}\,(h_{\theta,i+1:n})}\;\frac{1}{\mu_{\theta,-m:i}\,(h_{\theta,i+1:n-1})} \qquad (3.96)
\end{aligned}$$

for which we have

$$\begin{aligned}
\mu_{\theta,-m:i}\,(h_{\theta,i+1:n}) &= \int \mu_{\theta,-m:i}\,(dx)\,L_\theta\langle y_{i+1:n-1}\rangle(x,\mathbb{X})\\
&\geq \mu_{\theta,-m:i}\,(D)\left(\prod_{l=i+1}^{n-2}F_l\right)F_{n-1} \qquad (3.97)
\end{aligned}$$

and

$$\begin{aligned}
\mu_{\theta,-m:i}\,(h_{\theta,i+1:n-1}) &= \int \mu_{\theta,-m:i}\,(dx)\,L_\theta\langle y_{i+1:n-2}\rangle(x,\mathbb{X})\\
&\geq \mu_{\theta,-m:i}\,(D)\left(\prod_{l=i+1}^{n-2}F_l\right) \qquad (3.98)
\end{aligned}$$

Plugging (3.96), (3.97), and (3.98) into (3.95) yields

$$I_{-m:n-1}^{\theta,\mu,i}(\varphi_\theta)\;\leq\;\left(\frac{1+\rho^{n-i-2}}{1-\rho^{n-i-2}}\right)I_{-m:n-2}^{\theta,\mu,i}(\varphi_\theta)\;+\;\frac{2\,\eta^{n-i-2}\left(\prod_{l=i+1}^{n-2}V_l^2\right)V_{n-1}}{(1-\rho^{n-i-2})\,\mu_{\theta,-m:i}\,(D)^2}\,I_{-m:i}^{\theta,\mu,i}(\varphi_\theta).$$

Under condition (3.60), we may apply Lemma 16, so that

$$I_{-m:n-1}^{\theta,\mu,i}(\varphi_\theta)\;\leq\;\left(\frac{1+\rho^{n-2-i}}{1-\rho^{n-2-i}}\right)I_{-m:n-2}^{\theta,\mu,i}(\varphi_\theta)\;+\;\frac{2\,\eta^{n-2-i}\left(\prod_{l=i+1}^{n-2}V_l^2\right)V_{n-1}}{(1-\rho^{n-2-i})\,Z_1^2\,\mu(D)^2}\,I_{-m:i}^{\theta,\mu,i}(\varphi_\theta) \qquad (3.99)$$

where $Z_1 > 0$ is a $\mathbb{P}_\star$-almost surely finite random variable. This derivation holds for any $\eta > 0$, hence we may freely choose $\eta \in (0,1)$ satisfying $\eta < \exp(-2\,\mathbb{E}_\star[\log^+ V_0])$, so that $-\log\eta > \mathbb{E}_\star[\log^+ V_0^2]$. Since $(V_l)_{l\in\mathbb{Z}}$ and $(V_l^2)_{l\in\mathbb{Z}}$ are stationary sequences satisfying $\mathbb{E}_\star[\log^+(V_0)] < +\infty$ and $\mathbb{E}_\star[\log^+(V_0^2)] = 2\,\mathbb{E}_\star[\log^+ V_0] < +\infty$ by assumption, Lemma 6 from

Douc and Moulines (2012) guarantees the existence of a constant $\eta_2 \in (0,1)$ and a $\mathbb{P}_\star$-almost surely finite random variable $Z_2 > 0$ such that

$$\eta^{n-i-2} \left(\prod_{l=i+1}^{n-2} V_l^2 \right) V_{n-1} \;\leq\; \eta_2^{n-i-2}\, Z_2$$

for all $(n,i) \in \mathbb{Z}^2$ with $n-2 \geq i+1$. By defining $Z_3 = 2Z_2(Z_1^2\,\mu(D)^2)^{-1}$, this implies

$$I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) \;\leq\; \left(\frac{1+\rho^{n-i-2}}{1-\rho^{n-i-2}} \right) I^{\theta,\mu,i}_{-m:n-2}(\varphi_\theta) \;+\; Z_3 \left(\frac{\eta_2^{n-i-2}}{1-\rho^{n-i-2}} \right) I^{\theta,\mu,i}_{-m:i}(\varphi_\theta). \tag{3.100}$$

Condition (3.61) guarantees that we have $k^{-1} \sum_{l=i+1}^{i+k} \mathbb{1}_K(y_l) \geq \max\left(1-\gamma^-,\, (1+\gamma^+)/2\right)$ for all $k \in [\![l_0, n-i]\!]$, hence we may apply (3.100) recursively on $I^{\theta,\mu,i}_{-m:l}(\varphi_\theta)$ in a backward fashion for all $l \in [\![i+l_0, n]\!]$, which leads to

$$I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) \;\leq\; \left(\prod_{l=l_0}^{n-i-2} \frac{1+\rho^l}{1-\rho^l} \right) I^{\theta,\mu,i}_{-m:i+l_0}(\varphi_\theta) \;+\; Z_3 \left(\sum_{l=l_0}^{n-i-2} \frac{\eta_2^l}{1-\rho^l} \prod_{j=l+1}^{n-i-2} \frac{1+\rho^j}{1-\rho^j} \right) I^{\theta,\mu,i}_{-m:i}(\varphi_\theta). \tag{3.101}$$

Using Lemma 21, we get

$$\begin{aligned}
I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) \;&\leq\; \exp\left(\frac{2}{(1-\rho)^2} \right) \left(I^{\theta,\mu,i}_{-m:i+l_0}(\varphi_\theta) \;+\; Z_3 \left(\sum_{l=l_0}^{n-i-2} \frac{\eta_2^l}{1-\rho^l} \right) I^{\theta,\mu,i}_{-m:i}(\varphi_\theta) \right) \\
&\leq\; \exp\left(\frac{2}{(1-\rho)^2} \right) \left(I^{\theta,\mu,i}_{-m:i+l_0}(\varphi_\theta) \;+\; \frac{Z_3}{1-\rho^{l_0}} \left(\sum_{l=l_0}^{+\infty} \eta_2^l \right) I^{\theta,\mu,i}_{-m:i}(\varphi_\theta) \right) \\
&\leq\; \exp\left(\frac{2}{(1-\rho)^2} \right) \left(I^{\theta,\mu,i}_{-m:i+l_0}(\varphi_\theta) \;+\; \frac{Z_3\,\eta_2^{l_0}}{(1-\rho^{l_0})(1-\eta_2)}\, I^{\theta,\mu,i}_{-m:i}(\varphi_\theta) \right)
\end{aligned} \tag{3.102}$$

If $i - l_0 - 1 \leq -m$, then we have $\max(I^{\theta,\mu,i}_{-m:i+l_0}(\varphi_\theta), I^{\theta,\mu,i}_{-m:i}(\varphi_\theta)) \leq Z_0$, so that

$$I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) \;\leq\; Z_4 \tag{3.103}$$

for all $i - l_0 - 1 \leq -m < i + l_0 + 1 < n$, where

$$Z_4 \;=\; \exp\left(\frac{2}{(1-\rho)^2} \right) \left(1 \;+\; \frac{Z_3\,\eta_2^{l_0}}{(1-\rho^{l_0})(1-\eta_2)} \right) Z_0.$$

Note that $Z_4 > Z_0$. If $-m < i - l_0 - 1$, we bound terms of the form $I^{\theta,\mu,i}_{-m:i_0}(\varphi_\theta)$ for $i \leq i_0 \leq i + l_0 + 1$ by using a similar recursive argument. We have

$$
\begin{aligned}
& I^{\theta,\mu,i}_{-m:i_0}(\varphi_\theta) - I^{\theta,\mu,i}_{-m+1:i_0}(\varphi_\theta) \qquad (3.104)\\
&= \frac{\mu L_\theta\langle y_{-m:i-2}\rangle (L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1})}{\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:i_0+1})} - \frac{\mu L_\theta\langle y_{-m+1:i-2}\rangle (L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1})}{\mu L_\theta\langle y_{-m+1:i-2}\rangle\,(h_{\theta,i-1:i_0+1})}\\
&= \frac{(\mu L_\theta\langle y_{-m}\rangle\otimes\mu)\bar{L}_\theta\langle y_{-m+1:i-2}\rangle\,(L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1}\otimes h_{\theta,i-1:i_0+1} - h_{\theta,i-1:i_0+1}\otimes L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1})}{(\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:i_0+1}))\,(\mu L_\theta\langle y_{-m+1:i-2}\rangle\,(h_{\theta,i-1:i_0+1}))}.
\end{aligned}
$$

Since the functions $L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1}$ and $h_{\theta,i-1:i_0+1}$ are non-negative and bounded, we may again apply Lemma 17 thanks to $(m+i-2)^{-1}\sum_{l=-m+1}^{i-2}\mathbb{1}_K(y_l) \geq \max\left(1-\gamma^-,\,(1+\gamma^+)/2\right)$ from condition (3.61), which yields

$$
\begin{aligned}
&\left|(\mu L_\theta\langle y_{-m}\rangle\otimes\mu)\bar{L}_\theta\langle y_{-m+1:i-2}\rangle\,(L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1}\otimes h_{\theta,i-1:i_0+1} - h_{\theta,i-1:i_0+1}\otimes L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1})\right|\\
&\leq \tilde{\rho}^{m+i-2}(\mu L_\theta\langle y_{-m}\rangle\otimes\mu)\bar{L}_\theta\langle y_{-m+1:i-2}\rangle\,(L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1}\otimes h_{\theta,i-1:i_0+1} + h_{\theta,i-1:i_0+1}\otimes L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1})\\
&\qquad + \tilde{\eta}^{m+i-2}\left(\prod_{l=-m+1}^{i-2}\|L_\theta\langle y_l\rangle(\bullet,\mathbb{X})\|^2_\infty\right) 2\,\|L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1}\|_\infty\,\|h_{\theta,i-1:i_0+1}\|_\infty\\
&\leq \tilde{\rho}^{m+i-2}(\mu L_\theta\langle y_{-m}\rangle\otimes\mu)\bar{L}_\theta\langle y_{-m+1:i-2}\rangle\,(L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1}\otimes h_{\theta,i-1:i_0+1} + h_{\theta,i-1:i_0+1}\otimes L^{\varphi_\theta}_\theta\langle y_{i-1}\rangle h_{\theta,i:i_0+1})\\
&\qquad + \tilde{\eta}^{m+i-2}\left(\prod_{l=-m+1}^{i-2}G_l^2\right) 2\left(G^{\varphi}_{i-1}\prod_{l=i}^{i_0}G_l\right)\left(G_{i-1}\prod_{l=i}^{i_0}G_l\right) \qquad (3.105)
\end{aligned}
$$

for some $\tilde{\rho}\in(0,1)$ and all $\tilde{\eta}>0$. On the other hand, we have

$$
(\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:i_0+1}))\,(\mu L_\theta\langle y_{-m+1:i-2}\rangle\,(h_{\theta,i-1:i_0+1})) \;\geq\; \mu(D)^2\left(\prod_{l=-m+1}^{i_0}F_l^2\right)F_{-m}. \qquad (3.106)
$$

Plugging (3.105) and (3.106) back into (3.104) leads to

$$
\begin{aligned}
&\left|I^{\theta,\mu,i}_{-m:i_0}(\varphi_\theta) - I^{\theta,\mu,i}_{-m+1:i_0}(\varphi_\theta)\right|\\
&\leq\; \tilde{\rho}^{m+i-2}\left(I^{\theta,\mu,i}_{-m:i_0}(\varphi_\theta) + I^{\theta,\mu,i}_{-m+1:i_0}(\varphi_\theta)\right) \;+\; \frac{2}{\tilde{\eta}^{\,i_0-i}\,\mu(D)^2}\,\tilde{\eta}^{m+i_0}\left(\prod_{\substack{l=-m+1\\ l\neq i-1}}^{i_0}V_l^2\right)\frac{V^{\varphi}_{i-1}V_{i-1}}{F_{-m}}\\
&\leq\; \tilde{\rho}^{m+i-2}\left(I^{\theta,\mu,i}_{-m:i_0}(\varphi_\theta) + I^{\theta,\mu,i}_{-m+1:i_0}(\varphi_\theta)\right) \;+\; \frac{2}{\tilde{\eta}^{\,l_0+1}\,\mu(D)^2}\,\tilde{\eta}^{m+i_0}\left(\prod_{l=-m+1}^{i_0}\widetilde{V}_l^2\right)\frac{1}{F_{-m}}. \qquad (3.107)
\end{aligned}
$$

As previously, this derivation holds for any $\tilde{\eta} > 0$, so we may freely choose $\tilde{\eta} \in (0,1)$ with $\tilde{\eta} < \exp(-2\,\mathbb{E}_\star[\log^+ \widetilde{V}_0])$, so that $-\log\tilde{\eta} > \mathbb{E}_\star[\log^+ \widetilde{V}_0^2]$. Since $(\widetilde{V}_l^2)_{l\in\mathbb{Z}}$, and $(1/F_l)_{l\in\mathbb{Z}}$ are stationary sequences satisfying $\mathbb{E}_\star[\log^+(\widetilde{V}_l^2)] \leq 2(\mathbb{E}_\star[\log^+ V_0^\varphi] + \mathbb{E}_\star[\log^+ V_0]) < +\infty$ and $\mathbb{E}_\star[\log^+(1/F_l)] = \mathbb{E}_\star[\log^- F_l] < +\infty$, Lemma 6 from Douc and Moulines (2012) guarantees the existence of a constant $\tilde{\eta}_2 \in (0,1)$ and a $\mathbb{P}_\star$-a.s. finite random variable $\widetilde{Z}_2 > 0$ such that

$$\tilde{\eta}^{m+i_0}\left(\prod_{l=-m+1}^{i_0}\widetilde{V}_l^2\right)\frac{1}{F_{-m}} \;\leq\; \tilde{\eta}_2^{m+i_0-1}\,\widetilde{Z}_2 \;\leq\; \tilde{\eta}_2^{m+i-2}\,\widetilde{Z}_2 \tag{3.108}$$

for all $m \in \mathbb{N}$. By defining $\widetilde{\widetilde{Z}}_3 = 2\widetilde{Z}_2(\tilde{\eta}^{l_0+1}\mu(D)^2)^{-1}$, combining (3.108) with (3.107) implies

$$I^{\theta,\mu,i}_{-m:i_0}(\varphi_\theta) \;\leq\; \left(\frac{1+\tilde{\rho}^{m+i-2}}{1-\tilde{\rho}^{m+i-2}}\right) I^{\theta,\mu,i}_{-m+1:i_0}(\varphi_\theta) \;+\; \widetilde{\widetilde{Z}}_3\left(\frac{\tilde{\eta}_2^{m+i-2}}{1-\tilde{\rho}^{m+i-2}}\right). \tag{3.109}$$

Under condition (3.61), we have $k^{-1}\sum_{l=i-1-k}^{i-2}\mathbb{1}_K(y_l) \geq \max\left(1-\gamma^-, (1+\gamma^+)/2\right)$ for all $l_0 \leq k < m+i-1$, so that we may apply (3.109) recursively on $I^{\theta,\mu,i}_{l:i_0}(\varphi_\theta)$ for $-m \leq l < i-l_0$ in a forward fashion, which leads to

$$I^{\theta,\mu,i}_{-m:i_0}(\varphi_\theta) \;\leq\; \left(\prod_{l=l_0}^{m+i-2}\frac{1+\tilde{\rho}^l}{1-\tilde{\rho}^l}\right) I^{\theta,\mu,i}_{i-l_0-1:i_0}(\varphi_\theta) \;+\; \widetilde{\widetilde{Z}}_3\left(\sum_{l=l_0}^{m+i-2}\frac{\tilde{\eta}_2^l}{1-\tilde{\rho}^l}\prod_{j=l+1}^{m+i-2}\frac{1+\tilde{\rho}^j}{1-\tilde{\rho}^j}\right).$$

Using Lemma 21 again, along with (3.91), we get

$$\begin{aligned} I^{\theta,\mu,i}_{-m:i_0}(\varphi_\theta) \;&\leq\; \exp\left(\frac{2}{(1-\tilde{\rho})^2}\right)\left(I^{\theta,\mu,i}_{i-l_0-1:i_0}(\varphi_\theta) \;+\; \frac{\widetilde{\widetilde{Z}}_3\,\tilde{\eta}_2^{l_0-1}}{(1-\tilde{\rho}^{l_0})(1-\tilde{\eta}_2)}\right) \\ &\leq\; \exp\left(\frac{2}{(1-\tilde{\rho})^2}\right)\left(Z_0 \;+\; \frac{\widetilde{\widetilde{Z}}_3\,\tilde{\eta}_2^{l_0-1}}{(1-\tilde{\rho}^{l_0})(1-\tilde{\eta}_2)}\right) \end{aligned} \tag{3.110}$$

for $i \leq i_0 \leq i+l_0+1$. Plugging this back into (3.102) leads to

$$I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) \;\leq\; Z_5 \tag{3.111}$$

for all $(m,n) \in \mathbb{N}^2$ with $-m < i-l_0-1 < i+l_0+1 < n$, where

$$Z_5 = \exp\left(\frac{4}{(1-\rho)^2}\right)\left(1 \;+\; \frac{Z_3\,\eta_2^{l_0}}{(1-\rho^{l_0})(1-\eta_2)}\right)\left(Z_0 \;+\; \frac{\widetilde{\widetilde{Z}}_3\,\tilde{\eta}_2^{l_0-1}}{(1-\tilde{\rho}^{l_0})(1-\tilde{\eta}_2)}\right).$$

Note that (3.110) readily yields

$$I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) \;\leq\; Z_6 \tag{3.112}$$

for all $(m,n) \in \mathbb{N}^2$ with $-m < i - l_0 - 1 < n \leq i + l_0 + 1$, where

$$Z_6 = \exp\left(\frac{2}{(1-\rho)^2}\right)\left(Z_0 \,+\, \frac{\widetilde{Z}_3\,\tilde{\eta}_2^{\,l_0-1}}{(1-\tilde{\rho}^{\,l_0})(1-\tilde{\eta}_2)}\right).$$

By putting together (3.91), (3.103), (3.111), and (3.112), we have

$$I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) \;\leq\; Z_{i,l_0}$$

with $Z_{i,l_0} = \max(Z_4, Z_5, Z_6)$, which is positive and finite $\mathbb{P}_\star$-almost surely. This concludes the proof of Lemma 18. □

Proof of Lemma 19. Let $(\gamma^-, \gamma^+) \in (0,1)^2$ with $\gamma^- < \gamma^+$ such that under Assumption A8, there exists a $\mathbb{P}_\star$-almost surely finite random integer L_0 that ensures

$$\frac{1}{k}\sum_{l=i+1}^{i+k} \mathbb{1}_K(y_l) \;\geq\; \max\left(1-\gamma^-,\, \frac{1+\gamma^+}{2}\right) \tag{3.113}$$

for all $k \geq L_0$ and all $i \in \mathbb{Z}$.

Let $N = 2\,(L_0+1)$ so that $(n+m)/2 > L_0$ for all $m \in \mathbb{N}$ and $n \geq N$. Let's fix such $m \in \mathbb{N}$ and $n \geq N$. Using the same notation as in the proof of Lemma 18, we have

$$\begin{aligned}
&I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) - I^{\theta,\mu',i}_{-m:n-1}(\varphi_\theta) \qquad (3.114)\\
&= \frac{\mu L_\theta\langle y_{-m:i-2}\rangle(L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n})}{\mu L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n})} - \frac{\mu' L_\theta\langle y_{-m:i-2}\rangle(L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n})}{\mu' L_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n})}\\
&= \frac{(\mu\otimes\mu')\bar{L}_\theta\langle y_{-m:i-2}\rangle\,(L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n}\otimes h_{\theta,i-1:n} - h_{\theta,i-1:n}\otimes L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n})}{(\mu\otimes\mu')\bar{L}_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n}\otimes h_{\theta,i-1:n})}\\
&= \frac{(\mu\otimes\mu')\bar{L}_\theta\langle y_{-m:i-2}\rangle\,(L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n}\otimes L_\theta\langle y_{i-1}\rangle h_{\theta,i:n} - L_\theta\langle y_{i-1}\rangle h_{\theta,i:n}\otimes L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n})}{(\mu\otimes\mu')\bar{L}_\theta\langle y_{-m:i-2}\rangle\,(h_{\theta,i-1:n}\otimes h_{\theta,i-1:n})}.
\end{aligned}$$

On the one hand, we have

$$(\mu \otimes \mu')\bar{L}_\theta\langle y_{-m:i-2}\rangle \, (h_{\theta,i-1:n} \otimes h_{\theta,i-1:n}) \;\geq\; \mu(D)\mu'(D) \left(\prod_{l=-m}^{n-1} F_l^2 \right). \tag{3.115}$$

On the other hand, if $\lceil (n-m)/2 \rceil \leq i \leq n$, then $m+i-1 \geq (n+m)/2 - 1 \geq L_0$, so that

$$\frac{1}{m+i-1} \sum_{l=-m}^{i-2} \mathbb{1}_K(y_l) \;\geq\; \max\left(1-\gamma^-, \, \frac{1+\gamma^+}{2}\right)$$

thanks to (3.113), which allows us to apply Lemma 17 and get

$$\begin{aligned}
&\left|(\mu \otimes \mu')\bar{L}_\theta\langle y_{-m:i-2}\rangle \, (L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n} \otimes L_\theta\langle y_{i-1}\rangle h_{\theta,i:n} - L_\theta\langle y_{i-1}\rangle h_{\theta,i:n} \otimes L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n})\right| \\
&\leq \;\; \rho^{m+i-1}(\mu \otimes \mu')\bar{L}_\theta\langle y_{-m:i-2}\rangle \, (L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n} \otimes L_\theta\langle y_{i-1}\rangle h_{\theta,i:n} + L_\theta\langle y_{i-1}\rangle h_{\theta,i:n} \otimes L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n}) \\
&\quad + \;\; \eta^{m+i-1} \left(\prod_{l=-m}^{i-2} \|L_\theta\langle y_l\rangle(\bullet, \mathbb{X})\|_\infty^2 \right) 2 \, \|L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n}\|_\infty \, \|L_\theta\langle y_{i-1}\rangle h_{\theta,i:n}\|_\infty . \\
&\leq \;\; \rho^{m+i-1}(\mu \otimes \mu')\bar{L}_\theta\langle y_{-m:i-2}\rangle \, (L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n} \otimes L_\theta\langle y_{i-1}\rangle h_{\theta,i:n} + L_\theta\langle y_{i-1}\rangle h_{\theta,i:n} \otimes L_\theta^{\varphi_\theta}\langle y_{i-1}\rangle h_{\theta,i:n}) \\
&\quad + \;\; \eta^{m+i-1} \left(\prod_{l=-m}^{i-2} G_l^2 \right) 2 \left(G_{i-1}^{\varphi} \prod_{l=i}^{n-1} G_l \right) \left(G_{i-1} \prod_{l=i}^{n-1} G_l \right)
\end{aligned} \tag{3.116}$$

for any $\eta > 0$ and some associated $\rho \in (0,1)$. Plugging (3.115) and (3.116) into (3.114) yields

$$\begin{aligned}
&\left|I_{-m:n-1}^{\theta,\mu,i}(\varphi_\theta) - I_{-m:n-1}^{\theta,\mu',i}(\varphi_\theta)\right| \\
&\quad \leq \quad \rho^{m+i-1} \left|I_{-m:n-1}^{\theta,\mu,i}(\varphi_\theta) + I_{-m:n-1}^{\theta,\mu',i}(\varphi_\theta)\right| \;+\; \frac{2\,\eta^{m+i-1}}{\mu(D)\mu'(D)} \left(\prod_{\substack{l=-m \\ l \neq i-1}}^{n-1} V_l^2 \right) V_{i-1}^{\varphi} V_{i-1}.
\end{aligned} \tag{3.117}$$

Regarding the first term in (3.117), equation (3.113) ensures (3.60) and (3.61), so that

$$\left|I_{-m:n-1}^{\theta,\mu,i}(\varphi_\theta) + I_{-m:n-1}^{\theta,\mu',i}(\varphi_\theta)\right| \;\leq\; 2\, Z_{1,i} \tag{3.118}$$

for some $\mathbb{P}_\star$-almost surely finite random variable $Z_{1,i} > 0$, thanks to Lemma 18. Regarding the second term in (3.117), we are free to choose $\eta \in (0,1)$ with $\eta < \exp(-2\,\mathbb{E}_\star[\log^+ \widetilde{V}_0])$, so that $-\log \eta > \mathbb{E}_\star[\log^+ \widetilde{V}_0^2]$. Since $(\widetilde{V}_l^2)_{l\in\mathbb{Z}}$ is a stationary sequence satisfying $\mathbb{E}_\star[\log^+(\widetilde{V}_0^2)] \leq$

$2(\mathbb{E}_\star[\log^+ V_0^\varphi] + \mathbb{E}_\star[\log^+ V_0]) < +\infty$, Lemma 6 from Douc and Moulines (2012) and its proof guarantee the existence of a $\mathbb{P}_\star$-almost surely finite random variable $Z_2 > 0$ such that

$$\eta^{m+n-1} \left(\prod_{\substack{l=-m \\ l \neq i-1}}^{n-1} V_l^2 \right) V_{i-1}^\varphi V_{i-1} \;\leq\; \eta^{m+n-1} \left(\prod_{l=-m}^{n-1} \widetilde{V}_l^2 \right) \;\leq\; \eta_2^{m+n-1} Z_2 \tag{3.119}$$

for all $-m \leq 0 < i \leq n$, and all $\eta_2 = \eta/(\alpha\tilde{\alpha}) \in (0,1)$ with $\eta < \alpha < \exp(-2\,\mathbb{E}_\star[\log^+ V_0]) < 1$ and $(\eta/\alpha) < \tilde{\alpha} < 1$. In particular, by judiciously choosing $\eta = \exp(-18\,\mathbb{E}_\star[\log^+ V_0])$, $\alpha = \eta^{1/6}$, $\tilde{\alpha} = \eta^{1/6}$, we get $\eta_2 = \eta^{2/3}$ and thus

$$\eta^{m+i-1} \left(\prod_{\substack{l=-m \\ l \neq i-1}}^{n-1} V_l^2 \right) V_{i-1}^\varphi V_{i-1} \;\leq\; \frac{\eta_2^{m+n-1}}{\eta^{n-i}} Z_2 \;\leq\; (\eta^{1/3})^{2(m+n-1)-3(n-i)} Z_2 \;\leq\; (\eta^{1/3})^{m+i} Z_2 \tag{3.120}$$

since $\eta^{1/3} \in (0,1)$, and $2(m+n-1) - 3(n-i) = (m+i) + [m+n-1-2(n-i)] \geq (m+i)$ thanks to $\lceil (n-m)/2 \rceil \leq i \leq n$. Plugging (3.118) and (3.120) into (3.117) yields

$$\left| I_{-m:n-1}^{\theta,\mu,i}(\varphi_\theta) - I_{-m:n-1}^{\theta,\mu',i}(\varphi_\theta) \right| \;\leq\; \rho^{m+i} \frac{2\,Z_{1,i}}{\rho} \;+\; (\eta^{1/3})^{m+i} \frac{2\,Z_2}{\mu(D)\mu'(D)}$$

which, for $\lfloor (n-m)/2 \rfloor < i \leq n$, can be rewritten as

$$\left| I_{-m:n-1}^{\theta,\mu,i}(\varphi_\theta) - I_{-m:n-1}^{\theta,\mu',i}(\varphi_\theta) \right| \;\leq\; \rho_1^{m+i} Z_{3,i} \tag{3.121}$$

where $\rho_1 = \max(\rho, \eta^{1/3}) \in (0,1)$ and $Z_{3,i} = 2\,(Z_{1,i}\,\rho^{-1} + Z_2(\mu(D)\mu'(D))^{-1})$, with $Z_{3,i} < +\infty$, $\mathbb{P}_\star$-almost surely. This part will later lead us to (3.65).

For the second part of the Lemma 19, we have

$$\begin{aligned}
&I_{-m:n-1}^{\theta,\mu,i}(\varphi_\theta) - I_{-m:n-2}^{\theta,\mu,i}(\varphi_\theta) \\
&= \frac{\mu L_\theta \langle y_{-m:i-2} \rangle (L_\theta^{\varphi_\theta} \langle y_{i-1} \rangle h_{\theta,i:n})}{\mu L_\theta \langle y_{-m:i-2} \rangle \, (h_{\theta,i-1:n})} - \frac{\mu L_\theta \langle y_{-m:i-2} \rangle (L_\theta^{\varphi_\theta} \langle y_{i-1} \rangle h_{\theta,i:n-1})}{\mu L_\theta \langle y_{-m:i-2} \rangle \, (h_{\theta,i-1:n-1})} \\
&= \frac{(\mu \otimes \mu) \bar{L}_\theta \langle y_{-m:i-2} \rangle \, (L_\theta^{\varphi_\theta} \langle y_{i-1} \rangle \otimes L_\theta \langle y_{i-1} \rangle)(h_{\theta,i:n} \otimes h_{\theta,i:n-1} - h_{\theta,i:n-1} \otimes h_{\theta,i:n})}{(\mu \otimes \mu) \bar{L}_\theta \langle y_{-m:i-2} \rangle \, (h_{\theta,i-1:n} \otimes h_{\theta,i-1:n-1})}.
\end{aligned} \tag{3.122}$$

If $-m \leq i < \lceil (n-m)/2 \rceil$, then $n-i-2 \geq (n+m)/2 - 1 \geq L_0$, so that (3.113) ensures

$$\frac{1}{n-i-2} \sum_{l=i+1}^{n-2} \mathbb{1}_K(y_l) \;\geq\; \max\left(1-\gamma^-,\, \frac{1+\gamma^+}{2}\right)$$

which allows us to apply Lemma 17 and repeat the computation in (3.94) to get

$$\begin{aligned}
&|h_{\theta,i:n} \otimes h_{\theta,i:n-1} - h_{\theta,i:n-1} \otimes h_{\theta,i:n}| \\
&\leq \;\; \rho^{n-i-2} \left(h_{\theta,i:n} \otimes h_{\theta,i:n-1} + h_{\theta,i:n-1} \otimes h_{\theta,i:n}\right) \;+\; 2\,\eta^{n-i-2}\, G_{n-1} \left(\prod_{l=i+1}^{n-2} G_l^2\right) \left(h_{\theta,i:i+1} \otimes h_{\theta,i:i+1}\right).
\end{aligned}$$

Combining this with (3.122) and Lemma 16 under condition (3.113) leads to

$$\left|I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) - I^{\theta,\mu,i}_{-m:n-2}(\varphi_\theta)\right| \;\leq\; \rho^{n-i-2} \left(I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) + I^{\theta,\mu,i}_{-m:n-2}(\varphi_\theta)\right) \;+\; \frac{2\,\eta^{n-i-2} \left(\prod_{l=i+1}^{n-2} V_l^2\right) V_{n-1}}{Z_4\,\mu\,(D)^2}\, I^{\theta,\mu,i}_{-m:i}(\varphi_\theta)$$

for some positive $\mathbb{P}_\star$-a.s. finite Z_4, similarly to (3.99). Using (3.118) and (3.119), this implies

$$\left|I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) - I^{\theta,\mu,i}_{-m:n-2}(\varphi_\theta)\right| \;\leq\; 2\,\rho^{n-i-2}\, Z_{1,i} \;+\; \frac{2\,\eta_2^{n-i-2}\, Z_2}{Z_4\,\mu\,(D)^2}\, Z_{1,i}$$

which, for $-m \leq i < \lceil (n-m)/2 \rceil$, further gives

$$\left|I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) - I^{\theta,\mu,i}_{-m:n-2}(\varphi_\theta)\right| \;\leq\; \rho_2^{n-i}\, Z_{5,i} \tag{3.123}$$

where $\rho_2 = \max(\rho, \eta_2) \in (0,1)$ and $Z_{5,i} = 2\,Z_{1,i}(\rho^{-2} \,+\, Z_2 Z_4^{-1}(\mu(D)\,\eta_2)^{-2})$.

By combining equations (3.121) and (3.123) while defining $\rho_* = \max(\rho_1, \rho_2) \in (0,1)$ and $Z_i^* = \max(Z_{3,i}, Z_{5,i})$, we finally get, for all $(m,n) \in \mathbb{N}^2$ with $n > N$,

$$\left|I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) - I^{\theta,\mu',i}_{-m:n-1}(\varphi_\theta)\right| \;\leq\; \rho_*^{m+i}\, Z_i^*$$

if $\lceil (n-m)/2 \rceil \leq i \leq n$, and

$$\left|I^{\theta,\mu,i}_{-m:n-1}(\varphi_\theta) - I^{\theta,\mu,i}_{-m:n-2}(\varphi_\theta)\right| \;\leq\; \rho_*^{n-i}\, Z_i^*$$

if $-m \leq i < \lceil (n-m)/2 \rceil$, which concludes the proof of Lemma 19. □

Proof of Lemma 20. The main challenge is to prove (3.67), as (3.68) follows directly from applying (3.67) to the probability measures μ and $\mu' = \mu L_\theta \langle y_{-m':-m-1} \rangle$. We have

$$\begin{aligned}
&\left(\Delta^{\mathbb{E},\theta}_{k,m,\mu}(\varphi_\theta) - \Delta^{\mathbb{E},\theta}_{k,m,\mu'}(\varphi_\theta)\right) \\
&\quad = \quad \left(I^{\theta,\mu,k}_{-m:k}(\varphi_{\theta,k}) - I^{\theta,\mu',k}_{-m:k}(\varphi_{\theta,k})\right) \\
&\qquad + \sum_{i=-m}^{k-1} \left(I^{\theta,\mu,i}_{-m:k}(\varphi_{\theta,i}) - I^{\theta,\mu',i}_{-m:k}(\varphi_{\theta,i}) + I^{\theta,\mu',i}_{-m:k-1}(\varphi_{\theta,i}) - I^{\theta,\mu,i}_{-m:k-1}(\varphi_{\theta,i})\right)
\end{aligned}$$

where $I^{\theta,\mu,i}_{-m:k}(\varphi_{\theta,l}) = \mathbb{E}_{\theta,\mu}\left[\varphi_{\theta,l}(X_{i-1}, X_i) \,|\, y_{-m:k}\right]$. By the triangle inequality, we get

$$\begin{aligned}
&\left|\Delta^{\mathbb{E},\theta}_{k,m,\mu}(\varphi_\theta) - \Delta^{\mathbb{E},\theta}_{k,m,\mu'}(\varphi_\theta)\right| \\
&\quad \leq \quad \left|I^{\theta,\mu,k}_{-m:k}(\varphi_{\theta,k}) - I^{\theta,\mu',k}_{-m:k}(\varphi_{\theta,k})\right| \\
&\qquad + \sum_{i=-m}^{\lceil (k-m)/2 \rceil - 1} \left(\left|I^{\theta,\mu,i}_{-m:k}(\varphi_{\theta,i}) - I^{\theta,\mu,i}_{-m:k-1}(\varphi_{\theta,i})\right| + \left|I^{\theta,\mu',i}_{-m:k-1}(\varphi_{\theta,i}) - I^{\theta,\mu',i}_{-m:k}(\varphi_{\theta,i})\right|\right) \\
&\qquad + \sum_{i=\lceil (k-m)/2 \rceil}^{k-1} \left(\left|I^{\theta,\mu,i}_{-m:k}(\varphi_{\theta,i}) - I^{\theta,\mu',i}_{-m:k}(\varphi_{\theta,i})\right| + \left|I^{\theta,\mu',i}_{-m:k-1}(\varphi_{\theta,i}) - I^{\theta,\mu,i}_{-m:k-1}(\varphi_{\theta,i})\right|\right)
\end{aligned}$$

By using Lemma 19, there exists a constant $\rho \in (0,1)$ and a $\mathbb{P}_\star$-almost surely finite random variable $K \in \mathbb{N}$ with $K \geq 2$, such that for all $(m,k,i) \in \mathbb{N}^3$ with $k > K$, we have

$$\left|I^{\theta,\mu,i}_{-m:k}(\varphi_{\theta,i}) - I^{\theta,\mu,i}_{-m:k-1}(\varphi_{\theta,i})\right| + \left|I^{\theta,\mu',i}_{-m:k-1}(\varphi_{\theta,i}) - I^{\theta,\mu',i}_{-m:k}(\varphi_{\theta,i})\right| \;\leq\; 2\,\rho^{k-i}\,Z_i$$

if $-m \leq i < \lceil (k-m)/2 \rceil$, and

$$\left|I^{\theta,\mu,i}_{-m:k}(\varphi_{\theta,i}) - I^{\theta,\mu',i}_{-m:k}(\varphi_{\theta,i})\right| + \left|I^{\theta,\mu',i}_{-m:k-1}(\varphi_{\theta,i}) - I^{\theta,\mu,i}_{-m:k-1}(\varphi_{\theta,i})\right| \;\leq\; 2\,\rho^{m+i}\,Z_i$$

if $\lceil (k-m)/2 \rceil \leq i \leq k$, for some $\mathbb{P}_\star$-almost surely finite random variable $Z_i \in \mathbb{R}_+$. These

inequalities lead to

$$\begin{aligned}
\left|\Delta^{\mathbb{E},\theta}_{k,m,\mu}(\varphi_\theta) - \Delta^{\mathbb{E},\theta}_{k,m,\mu'}(\varphi_\theta)\right| &\leq \rho^{m+k}\, Z_k \,+\, 2\left(\sum_{i=-m}^{\lceil (k-m)/2\rceil - 1} \rho^{k-i}\, Z_i\right) \,+\, 2\left(\sum_{i=\lceil (k-m)/2\rceil}^{k-1} \rho^{m+i}\, Z_i\right) \\
&\leq 2\left(\max_{j\in[\![-m,k]\!]} Z_j\right)\left(\sum_{i=-m}^{\lceil (k-m)/2\rceil - 1} \rho^{k-i} \,+\, \sum_{i=\lceil (k-m)/2\rceil}^{k} \rho^{m+i}\right) \\
&\leq 2\left(\sum_{j=-m}^{k} Z_j\right)\left(\sum_{i=-\infty}^{\lceil (k-m)/2\rceil - 1} \rho^{k-i} \,+\, \sum_{i=\lceil (k-m)/2\rceil}^{+\infty} \rho^{m+i}\right) \\
&\leq 2\left(\sum_{j=-m}^{k} \frac{Z_j \max(1,j^2)}{\max(1,j^2)}\right)\frac{2\,\rho^{\frac{k+m}{2}}}{1-\rho} \\
&\leq \frac{4\,\max(k^2,m^2)}{1-\rho}\left(\sum_{j=-\infty}^{+\infty} \frac{Z_j}{\max(1,j^2)}\right)\rho^{\frac{k+m}{2}} \\
&\leq \tilde{Z}\,\max(k^2,m^2)\,\tilde{\rho}^{\,k+m}
\end{aligned}$$

where $\tilde{\rho} = \rho^{1/2} \in (0,1)$ and

$$\tilde{Z} \;=\; \frac{4}{1-\rho}\left(\sum_{j=-\infty}^{+\infty} \frac{Z_j}{\max(1,j^2)}\right).$$

By Assumption A9, we have $\mathbb{E}_\star[Z_i] = \mathbb{E}_\star[Z_0] < +\infty$ for all $i \in \mathbb{Z}$. The non-negativity of $Z_j/\max(1,j^2)$ for all $j \in \mathbb{Z}$ allows us to apply the Fubini-Tonelli theorem, so that

$$\mathbb{E}_\star\left[\sum_{j=-\infty}^{+\infty} \frac{Z_j}{\max(1,j^2)}\right] \;=\; \sum_{j=-\infty}^{+\infty} \frac{\mathbb{E}_\star\left[Z_j\right]}{\max(1,j^2)} \;=\; \mathbb{E}_\star\left[Z_0\right]\left(1 + 2\sum_{j=1}^{+\infty}\frac{1}{j^2}\right) \;<\; +\infty.$$

Thus, we obtain $\mathbb{E}_\star[\tilde{Z}] < +\infty$, which implies that $\tilde{Z}$ is $\mathbb{P}_\star$-almost surely finite. This concludes the proof of Lemma 20. □

3.7 Numerical experiments

Despite the technical limitations of the conditions presented in Section 3.6, the consistency and asymptotic Normality of posterior distributions are fairly general phenomena that one can reasonably expect to witness in practice for a wide range of state-space models. In this

section, we illustrate such phenomena numerically by looking at the posterior distributions of the population dynamics and stochastic volatility models introduced in Chapter 1.

3.7.1 Diffusion models for population dynamics of red kangaroos

This illustration complements the example presented in Section 1.4.2 of Chapter 1. For each population model M_1, M_2, and M_3, the respective posteriors of the parameters are estimated via SMC^2 across 5 replications. The marginal posterior densities under models M_1 to M_3 are respectively shown in Figures 3.1, 3.2, and 3.3. These estimated posterior densities appear concentrated when contrasted with the vague independent priors σ, τ, $b \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 10)$, and $r \sim \text{Unif}(-10, 10)$. The plots suggest that consistency and asymptotic Normality of the posterior may still occur, even when the technical conditions of Section 3.6 are not met.

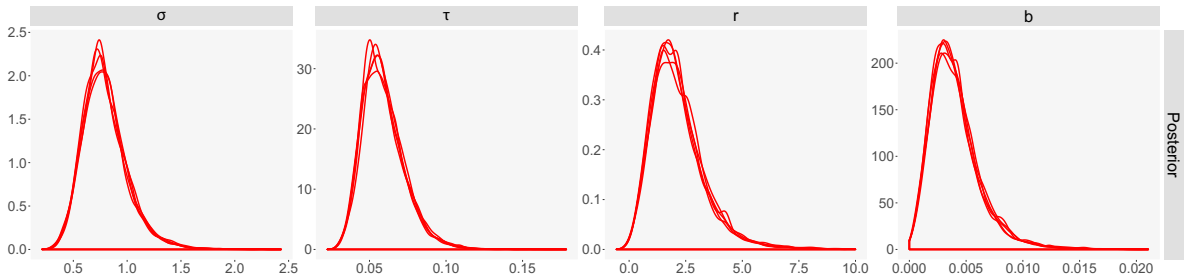

Figure 3.1. *Estimated marginal posterior densities of* (σ, τ, r, b) *under model* M_1*, given 41 observations, with independent priors* σ, τ, $b \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 10)$ *and* $r \sim \text{Unif}(-10, 10)$*, plotted for 5 replications of* SMC^2 *(solid lines). See Section 3.7.1.*

3.7.2 Lévy-driven stochastic volatility models

This illustration complements the example presented in Section 1.3.3 of Chapter 1. For each Lévy-driven stochastic volatility model M_1 and M_2, the respective posterior densities of the parameters are estimated via SMC^2 across 5 replications. The estimated marginal posterior densities under models M_1 and M_2 are respectively shown in Figures 3.4 and 3.5, along with the corresponding marginal prior densities for comparison. To facilitate the assessment of

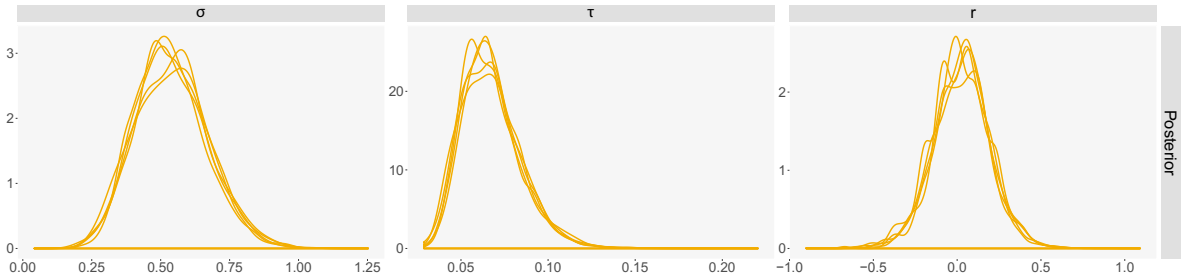

Figure 3.2. *Estimated marginal posterior densities of* (σ, τ, r) *under model* M_2*, given 41 observations, with independent priors* $\sigma, \tau \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 10)$ *and* $r \sim \text{Unif}(-10, 10)$*, plotted for 5 replications of* SMC^2 *(solid lines). See Section 3.7.1.*

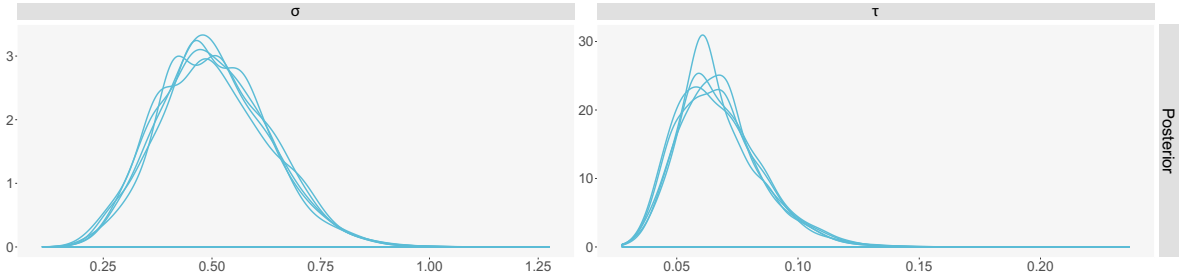

Figure 3.3. *Estimated marginal posterior densities of* (σ, τ) *under model* M_3*, given 41 observations, with independent priors* $\sigma, \tau \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 10)$*, plotted for 5 replications of* SMC^2 *(solid lines). See Section 3.7.1.*

posterior concentration, the marginal prior densities are plotted over the same support as their corresponding marginal posterior densities. Similarly to the previous example, the concentration of the posterior seems to be occurring, at least marginally on each parameter. The only exception is for λ_2 under model M_2, whose posterior after 1000 observations appear virtually identical to the prior. This can be explained by the posterior of w concentrating near 1 as the observations are generated from M_1, thus making the second factor irrelevant in model M_2. The parameter λ_2 associated with the second factor is then not identifiable.

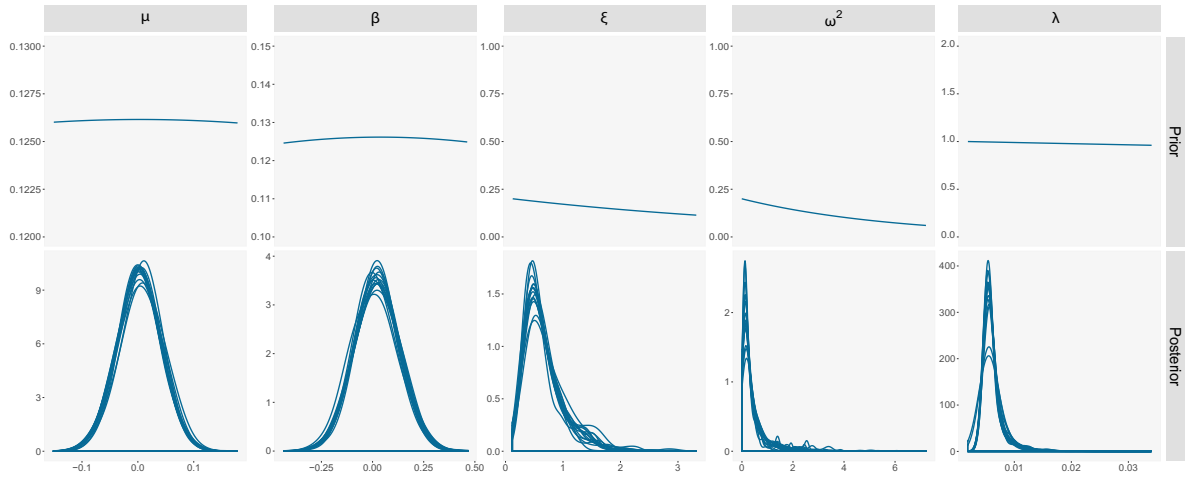

Figure 3.4. *Top panels: marginal prior densities* $\mu, \beta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 10)$; $\xi, \omega^2 \overset{\text{i.i.d.}}{\sim} \text{Exp}(1/5)$; $\lambda \sim \text{Exp}(1)$, *plotted over the support of the posterior densities. Bottom panels: estimated marginal posterior densities under model* M_1, *given 1000 observations, plotted for 5 replications of* SMC^2 *(solid lines). See Section 3.7.2.*

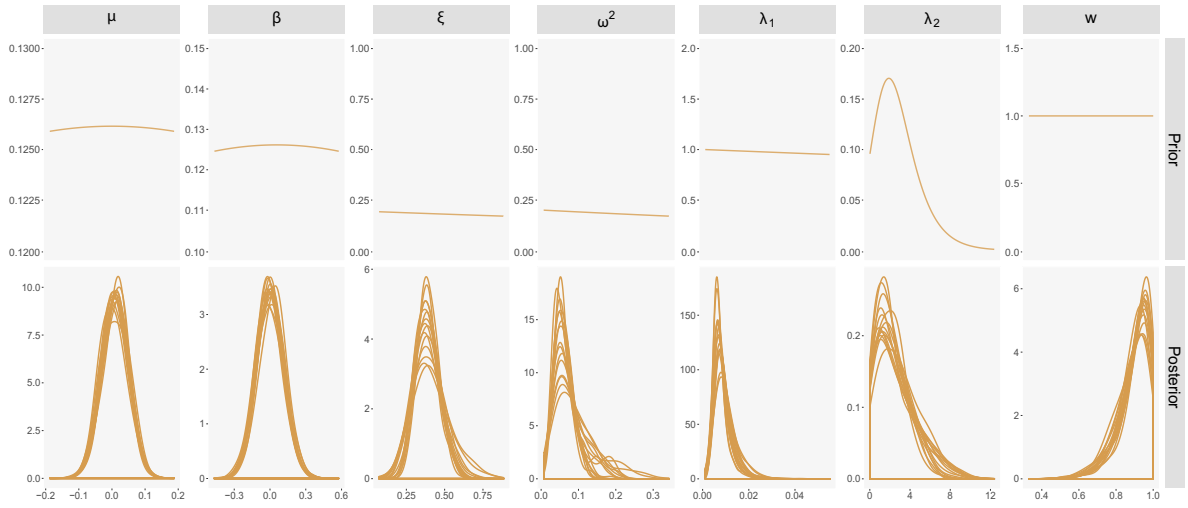

Figure 3.5. *Top panels: marginal prior densities* $\mu, \beta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 10)$; $\xi, \omega^2 \overset{\text{i.i.d.}}{\sim} \text{Exp}(1/5)$; $\lambda_1 \sim \text{Exp}(1)$; $\lambda_2 - \lambda_1 \sim \text{Exp}(1/2)$; $\text{w} \sim \text{Unif}(0, 1)$, *plotted over the support of the posterior densities. Bottom panels: estimated marginal posterior densities under model* M_2, *given 1000 observations, plotted for 5 replications of* SMC^2 *(solid lines). See Section 3.7.2.*

3.8 Discussion

The recent advances on the asymptotic theory of the MLE in possibly misspecified state-space models pave the way for envisioning a proof of posterior consistency and Bernstein-von Mises type of results under satisfyingly weak conditions. Many of the ingredients needed for the latter can be obtained as direct byproducts of the former, with the exception of the uniform control of the observed Fisher information around the limit of the MLE. This control over the Hessian of the log-likelihood virtually constitutes the only additional obstacle to overcome, and leads to — by re-expressing that Hessian in terms of the full likelihood using the standard Fisher and Louis identities — studying the asymptotic behavior of conditional expectations of possibly unbounded test functions. In contrast with the study of the MLE, our attempt suggests that the need for stronger conditions might be unavoidable, especially regarding the mixing behavior of the sequence of observations. Our quest for a general proof remains unfulfilled and will be the object of further research.

Most of the technical conditions under consideration are likely to be sufficient but far from necessary, and despite the remaining elusiveness of a complete proof, one may still reasonably hope for posterior consistency and asymptotic Normality to occur in practical setting for a wide range of state-space models. In practice, one can rely on SMC samplers (Chopin, 2002; Del Moral et al., 2006; Chopin et al., 2013) to sequentially obtain approximations of all the successive posterior distributions, thus providing an empirical way to heuristically assess the concentration of the posterior distribution and monitor its closeness to Normality as the number of observations increases.

Appendices

A

Pseudo-code and algorithms

Contents

This appendix provides implementation details about the various sampling algorithms used in this thesis, including SMC samplers (Appendix A.1), SMC^2 (Appendix A.2), and unbiased MCMC methods (Appendix A.3). For the sake of making this manuscript self-contained, we provide simplified versions of these algorithms, along with some pseudo-code. More extensive discussions about implementation choices can be found in the corresponding references.

A.1 Implementation of SMC methods

This section presents general guidelines to implement SMC samplers (Chopin, 2002; Del Moral et al., 2006; Del Moral, Doucet and Jasra, 2012; Duan and Fulop, 2015). Section A.1.1 and Algorithm 1 describe a non-adaptive version of SMC, which will be used when getting unbiased estimators of marginal likelihoods is of importance, whereas Section A.1.2 and Algorithm 2 present an adaptive version that will be used when efficiency and consistency of such estimators are the main concerns. The descriptions in Algorithms 1 and 2 are abstracted in order to encapsulate both SMC and more sophisticated SMC^2 samplers.

A.1.1 SMC with fixed temperatures and kernels

Given T observations $y_{1:T} \in \mathbb{Y}^T$ and a chosen number of particles $N_\theta \in \mathbb{N}^*$, an SMC sampler produces, at each step $t \in [\![0, T]\!]$, a set of N_θ particles $\theta_t^{(1:N_\theta)} = (\theta_t^{(1)}, ..., \theta_t^{(N_\theta)})$ with associated normalized weights $W_t^{(1:N_\theta)} = (W_t^{(1)}, ..., W_t^{(N_\theta)})$, targeting the posterior distribution $p(d\theta|y_{1:t})$ (Chopin, 2002; Del Moral et al., 2006). These particles are initialized as i.i.d. draws $\theta_0^{(1:N_\theta)}$ from a proposal distribution $q_0(d\theta)$. When the prior $p(d\theta)$ is proper and can be sampled from, one can chose $q_0(d\theta) = p(d\theta)$. Otherwise, $q_0(d\theta)$ should be chosen as an approximation of the first proper posterior distribution. Going from an approximate sample of the posterior $p(d\theta|y_{1:t})$ to an approximate sample of the next posterior $p(d\theta|y_{1:t+1})$ is achieved by successively targeting the intermediate bridging distributions whose densities are given by $p_{\gamma_{t,j}}(\theta) \propto p(\theta|y_{1:t})p(y_{t+1}|y_{1:t}, \theta)^{\gamma_{t,j}}$, where the $\gamma_{t,j}$'s are well-chosen *temperatures* satisfying $0 = \gamma_{t,0} < \gamma_{t,1} < ... < \gamma_{t,J_t} = 1$ for some $J_t \in \mathbb{N}^*$.

Algorithm 1 describes a non-adaptive version of SMC, where the inputs include a fixed temperature schedule $\{0 = \gamma_{t,0} < \gamma_{t,1} < ... < \gamma_{t,J_t} = 1 : t \in [\![0, T-1]\!]\}$, a fixed resampling schedule $\{(r_{t,1}, ..., r_{t,J_t}) \in \{\texttt{True}, \texttt{False}\}^{J_t} : t \in [\![0, T-1]\!]\}$, and some corresponding proposal densities $q_{t,j}$ each respectively approximating the targets $p_{\gamma_{t,j}}$. We also assume that given

any particular $\theta_t^{(m)}$, we have a procedure to construct a non-negative estimator $\hat{G}_{t+1}^{(m)}$ of the incremental likelihood $p(y_{t+1}|y_{1:t}, \theta_t^{(m)})$, such that $\hat{Z}_{t+1}^{(m)} = \prod_{s=1}^{t+1} \hat{G}_s^{(m)}$ is an unbiased estimator of the likelihood $p(y_{1:t+1}|\theta_t^{(m)})$. This construction could possibly involve generating additional random variables, of which one needs to appropriately keep track when resampling and moving the corresponding particles in Step (B). When the likelihood is tractable, one can directly use $\hat{G}_{t+1}^{(m)} = p(y_{t+1}|y_{1:t}, \theta_t^{(m)})$ and Algorithm 1 reduces to a tempered version of the iterated batch importance sampling algorithm (Chopin, 2002) with non-adaptive rejuvenation. The use of unbiased estimators in place of exact likelihoods can be justified in the spirit of random-weight importance sampling (Fearnhead, Papaspiliopoulos, Roberts and Stuart, 2010), and will be the key idea of the SMC^2 samplers reviewed in Section A.2. When resampling is triggered, the moves in Step (B) can be performed as many times as desired in order to improve the rejuvenation rate of the particles, albeit at the expense of additional computations. These move steps correspond to applying a particle Markov chain Monte Carlo (PMCMC) kernel (Andrieu et al., 2010) that leaves the ongoing target $p_{\gamma_{t,j}}$ invariant. Possible choices of resampling schemes for Step (A) of Algorithm 1 are discussed in Gerber et al. (2017), and we will mainly use either multinomial or SSP resampling.

The output of SMC samplers consists of successive weighted samples $(W_t^{(1:N_\theta)}, \theta_t^{(1:N_\theta)})$ which can be used to form the estimators $\sum_{m=1}^{N_\theta} W_t^{(m)} h(\theta_t^{(m)})$. As $N_\theta \to +\infty$, these estimators consistently estimate conditional expectations of the form $\mathbb{E}[h(\Theta) \,|\, y_{1:t}]$ with respect to the posterior distributions $\Theta \sim p(d\theta \,|\, y_{1:t})$ for all $t \in [\![1, T]\!]$ and all suitably integrable test functions h. Under appropriate resampling schemes (i.e. satisfying Assumption 2 in Andrieu et al. (2010), e.g. multinomial or SSP), the non-adaptive SMC sampler presented in Algorithm 1 also sequentially produces the following quantities

$$\hat{Z}_t = \prod_{s=1}^{t} \prod_{j=0}^{J_s-1} \left(\sum_{m=1}^{N_\theta} W_{s,j}^{(m)} w_{s,j+1}^{(m)} \right) = \hat{Z}_{t-1} \prod_{j=0}^{J_t-1} \left(\sum_{m=1}^{N_\theta} W_{t,j}^{(m)} w_{t,j+1}^{(m)} \right) \tag{A.1}$$

which are unbiased estimators of the marginal likelihoods $p(y_{1:t})$ at every step $t \in [\![1, T]\!]$, for any fixed number of particles N_θ. The properties of such estimators have been extensively studied in e.g. Del Moral (2004), Cérou et al. (2011), and references therein.

A.1.2 SMC with adaptive tempering and moves

For a fixed number of particles N_θ, the efficiency of the SMC sampler in Algorithm 1 may be improved by choosing the temperatures, resampling schedule, and rejuvenation kernels in an adaptive fashion, as described in Algorithm 2. Let N_θ/c be the minimum *effective sample size* (ESS), with $c \geq 1$ fixed at some desired value (e.g. $c = 2$ in our numerical experiments). Given the current temperature $\gamma_{t,j} < 1$ and weighted particles $(W_{t,j}^{(1:N_\theta)}, \theta_{t,j}^{(1:N_\theta)})$, step (ii) adaptively determines the next temperature $\gamma_{t,j+1} \in (\gamma_{t,j+1}, 1]$ by setting it to its largest possible value, while mitigating the degeneracy of the weights by maintaining the ESS above the desired threshold N_θ/c. Starting from $(\gamma_{0,0}\,, G_0^{(1:N_\theta)}, \theta_{0,j}^{(1:N_\theta)})$ with $\mathrm{ESS}(\gamma_{0,0}) = N_\theta$, then at all time t we have $\mathrm{ESS}(\gamma_{t,j}) \geq N_\theta/c$ given the current $(\gamma_{t,j}\,, G_{t+1}^{(1:N_\theta)}, \theta_{t,j}^{(1:N_\theta)})$, by construction. Besides, given $(\gamma_{t,j}\,, G_{t+1}^{(1:N_\theta)}, \theta_{t,j}^{(1:N_\theta)})$, the function $\gamma \mapsto \mathrm{ESS}(\gamma)$ is continuous. These two facts guarantee the existence and uniqueness of the $\gamma_{t,j+1}$ defined in Step (ii) of Algorithm 2. Beside the effective sample size, other diagnostics could be used to monitor the degeneracy of the weights.

Regarding the proposal density $q_{t,j}$ in Step (A) of Algorithm 2, we use a mixture of Normals fitted to the latest set of weighted particles $(W_{t,j+1}^{(1:N_\theta)}, \theta_{t,j}^{(1:N_\theta)})$, using five components by default throughout our numerical experiments. Using only one component in that mixture corresponds to the Normal proposal discussed in Section 4.2 of Chopin (2002).

The output of this adaptive SMC sampler can be used in a similar fashion as described in Section A.1.1. One should be warned that although the estimator of the marginal likelihood given in (A.1) is still consistent as $N_\theta \to +\infty$, it is no longer guaranteed to be unbiased when using such an adaptive implementation of SMC. In practice, one could address this issue by performing a preliminary run of adaptive SMC to store the relevant temperatures, resampling

A. Pseudo-code and algorithms

Algorithm 1 Non-adaptive SMC sampler

1. For $m \in [\![1, N_\theta]\!]$:
   - Sample $\theta_{0,0}^{(m)} \sim p(d\theta)$ and set $\theta_0^{(m)} = \theta_{0,0}^{(m)}$ (initialize)
   - Set $W_0^{(m)} = W_{0,0}^{(m)} = 1/N_\theta$
2. For $t \in [\![0, T-1]\!]$:
   (a) For $m \in [\![1, N_\theta]\!]$: construct a non-negative estimator $\hat{G}_{t+1}^{(m)}$ of $p(y_{t+1}|y_{1:t}, \theta_t^{(m)})$ such that $\hat{Z}_{t+1}^{(m)} = \prod_{s=1}^{t+1} \hat{G}_s^{(m)}$ unbiasedly estimates the likelihood $p(y_{1:t+1}|\theta_t^{(m)})$
   (b) For $j \in [\![0, J_t - 1]\!]$:
      (i) For $m \in [\![1, N_\theta]\!]$, compute: $w_{t,j+1}^{(m)} = (\hat{G}_{t+1}^{(m)})^{\gamma_{t,j+1}-\gamma_{t,j}} W_{t,j}^{(m)}$ (re-weight)
      (ii) For $m \in [\![1, N_\theta]\!]$, compute: $W_{t,j+1}^{(m)} = w_{t,j+1}^{(m)} \,/\, \sum_{i=1}^{N_\theta} w_{t,j+1}^{(i)}$ (normalize weights)
      (iii) If $r_{t,j+1} = \texttt{False}$, then for $m \in [\![1, N_\theta]\!]$, set: $\theta_{t,j+1}^{(m)} = \theta_{t,j}^{(m)}$
      (iv) If $r_{t,j+1} = \texttt{True}$, then: (rejuvenate)
         (A) Sample indexes $A_{t,j}^{(1)}, ..., A_{t,j}^{(N_\theta)}$ from $[\![1, N_\theta]\!]$ with respective weights $W_{t,j+1}^{(1:N_\theta)}$ and with replacement using a prespecified resampling scheme (sample ancestors)
         (B) For $m \in [\![1, N_\theta]\!]$:
            - Set $W_{t,j+1}^{(m)} = 1/N_\theta$ (reset weights)
            - Set $\theta_{t,j+1}^{(m)} = \theta_{t,j}^{(A_{t,j}^{(m)})}$ and $\hat{Z}_{t+1}^{(m)} = \hat{Z}_{t+1}^{(A_{t,j}^{(m)})}$ (resample particles)
            - Propose new particles $\tilde{\theta}_{t,j+1}^{(m)} \sim q_{t,j+1}(\theta_{t,j+1}^{(m)}, d\theta)$ and construct non-negative estimators $\hat{G}_s(\tilde{\theta}_{t,j+1}^{(m)})$ of $p(y_{s+1}|y_{1:s}, \tilde{\theta}_{t,j+1}^{(m)})$ for $s \in [\![0, T]\!]$ such that $\hat{Z}_{t+1}(\tilde{\theta}_{t,j+1}^{(m)}) = \prod_{s=1}^{t+1} \hat{G}_s(\tilde{\theta}_{t,j+1}^{(m)})$ unbiasedly estimates $p(y_{1:t+1}|\tilde{\theta}_{t,j+1}^{(m)})$
            - Accept $\theta_{t,j+1}^{(m)} = \tilde{\theta}_{t,j+1}^{(m)}$, $\hat{G}_{1:t+1}^{(m)} = \hat{G}_{1:t+1}(\tilde{\theta}_{t,j+1}^{(m)})$, and $\hat{Z}_{t+1}^{(m)} = \hat{Z}_{t+1}(\tilde{\theta}_{t,j+1}^{(m)})$ with probability: $\min\left(1, \frac{p(\tilde{\theta}_{t,j+1}^{(m)})}{p(\theta_{t,j+1}^{(m)})} \frac{\hat{Z}_{t+1}(\tilde{\theta}_{t,j+1}^{(m)})^{\gamma_{t,j+1}}}{\hat{Z}_{t+1}^{(m)\,\gamma_{t,j+1}}} \frac{q_{t,j+1}(\tilde{\theta}_{t,j+1}^{(m)}, \theta_{t,j+1}^{(m)})}{q_{t,j+1}(\theta_{t,j+1}^{(m)}, \tilde{\theta}_{t,j+1}^{(m)})}\right)$
   (c) When $\gamma_{t,j+1} = 1$ (i.e. $j = J_t - 1$), then for $m \in [\![1, N_\theta]\!]$:
      - Set $\theta_{t+1}^{(m)} = \theta_{t+1,0}^{(m)} = \theta_{t,J_t}^{(m)}$ and $W_{t+1}^{(m)} = W_{t+1,0}^{(m)} = W_{t,J_t}^{(m)}$
3. Output the weighted samples $(W_t^{(1:N_\theta)}, \theta_t^{(1:N_\theta)})_{t \in [\![1,T]\!]}$ respectively targeting the successive posterior distributions $p(d\theta \,|\, y_{1:t})$ for all $t \in [\![1, T]\!]$.

times, and fitted kernels during the course of the algorithm. One would then perform a second run using non-adaptive SMC using the previously stored algorithmic parameters.

A.2 Implementation of SMC^2 methods

For state-space models, the incremental likelihoods of the form $p(y_{t+1} \,|\, y_{1:t}, \theta_t^{(m)})$ correspond to typically intractable integrals with respect to the latent states and cannot be directly evaluated (except in simple cases, e.g. finite discrete space for the latent states, linear Gaussian state-space models), so that precautions must be taken. Fortunately, particle filters — reviewed in Appendix A.2.1 — can provide unbiased estimators of the likelihoods in state-space models, which can in turn be used in place of the unavailable likelihoods to yield the SMC^2 algorithm reviewed in Section A.2.2.

A.2.1 Particle filters

In the context of a state-space model with fixed parameter $\theta \in \mathbb{T}$, initial distribution μ_θ, transition kernel f_θ, and observation density g_θ, some notable instances of SMC are known as particle filters. Particle filters (e.g. see Section 2.1 in Chopin et al., 2013) take as inputs a desired number N_x of particles, a collection of observations $y_{1:T} \in \mathbb{Y}^T$, a parameter $\theta \in \mathbb{T}$, and a collection of proposal distributions $\{q_{t,\theta}(dx_t \,|\, x_{t-1})\}$ which can be sampled from. It outputs a sequence of properly weighted particles $(W_{t,\theta}^{(1:N_x)}, X_t^{(1:N_x)})$ targeting the *filtering distribution* $p(x_t \,|\, y_{1:t}, \theta)$ for all $t \in [\![1, T]\!]$, as described in Algorithm 3. Similarly to the discussion of generic SMC samplers in Section A.1.1, various resampling schemes may be considered in Step 3(a) (e.g. see Liu and Chen, 1998; Pitt and Shephard, 1999), although we will limit ourselves to multinomial resampling to facilitate the analysis. As long as one can sample from the transition kernel f_θ, particle filters can be run by choosing $q_{t,\theta}(dx_t \,|\, x_{t-1}) = f_\theta(dx_t \,|\, x_{t-1})$ to simplify the importance weights appearing in Step 3(b). Such a choice leads to the standard

Algorithm 2 Adaptive SMC sampler

1. For $m \in [\![1, N_\theta]\!]$:
   - Sample $\theta_{0,0}^{(m)} \sim p(d\theta)$ and set $\theta_0^{(m)} = \theta_{0,0}^{(m)}$ (initialize)
   - Set $W_0^{(m)} = W_{0,0}^{(m)} = 1/N_\theta$
   - Set $j = 0$ and $\gamma_{0,0} = 0$
2. For $t \in [\![0, T-1]\!]$:
   - For $m \in [\![1, N_\theta]\!]$: construct a non-negative estimator $\hat{G}_{t+1}^{(m)}$ of $p(y_{t+1}|y_{1:t}, \theta_t^{(m)})$ such that $\hat{Z}_{t+1}^{(m)} = \prod_{s=1}^{t+1} \hat{G}_s^{(m)}$ unbiasedly estimates the likelihood $p(y_{1:t+1}|\theta_t^{(m)})$
   - While $\gamma_{t,j} < 1$:
     (a) Find the next temperature adaptively as follows:
       (i) Define the effective sample size as a function of γ :
       $$\mathrm{ESS}(\gamma) = \frac{\left(\sum_{m=1}^{N_\theta} W_{t,j}^{(m)} \left(\hat{G}_{t+1}^{(m)}\right)^{\gamma-\gamma_{t,j}}\right)^2}{\sum_{m=1}^{N_\theta} \left(W_{t,j}^{(m)} \left(\hat{G}_{t+1}^{(m)}\right)^{\gamma-\gamma_{t,j}}\right)^2} \quad \text{(assess degeneracy)}$$
       (ii) Set $\gamma_{t,j+1} = \max\{\gamma \in (\gamma_{t,j}, 1] : \mathrm{ESS}(\gamma) \geq N_\theta/c\}$ (adapt temperature)
     (b) For $m \in [\![1, N_\theta]\!]$, set $w_{t,j+1}^{(m)} = (\hat{G}_{t+1}^{(m)})^{\gamma_{t,j+1}-\gamma_{t,j}} W_{t,j}^{(m)}$ (re-weight)
     (c) For $m \in [\![1, N_\theta]\!]$, set $W_{t,j+1}^{(m)} = w_{t,j+1}^{(m)} / \sum_{i=1}^{N_\theta} w_{t,j+1}^{(i)}$ (normalize weights)
     (d) If $\gamma_{t,j+1} < 1$, then:
       (A) Use the weighted sample $(W_{t,j+1}^{(1:N_\theta)}, \theta_{t,j}^{(1:N_\theta)})$ to adaptively construct a proposal density $q_{t,j+1}$ that approximates $p_{\gamma_{t,j+1}}$.
       (B) Rejuvenate particles by performing step (iv) of Algorithm 1
     (e) If $\gamma_{t,j+1} = 1$, then for $m \in [\![1, N_\theta]\!]$:
       - Set $\theta_{t+1}^{(m)} = \theta_{t+1,0}^{(m)} = \theta_{t,j+1}^{(m)}$ and $W_{t+1}^{(m)} = W_{t+1,0}^{(m)} = W_{t,j+1}^{(m)}$
3. Output the weighted samples $(W_t^{(1:N_\theta)}, \theta_t^{(1:N_\theta)})_{t \in [\![1,T]\!]}$ respectively targeting the successive posterior distributions $p(d\theta \,|\, y_{1:t})$ for all $t \in [\![1, T]\!]$.

bootstrap particle filter (Gordon, Salmond and Smith, 1993). As a byproduct, non-adaptive particle filters also provide unbiased estimators of the likelihood $p(y_{1:t}|\theta)$ for all $t \in [\![1, T]\!]$.

Algorithm 3 Particle filter

1. For $n \in [\![1, N_x]\!]$: (assimilate first observation)
   - Sample $X_1^{(n)} \sim q_{1,\theta}(dx_1)$ independently (initialize)
   - Set $w_{t,\theta}^{(n)} = \mu_\theta(X_1^{(n)})\, g_\theta(y_1 \,|\, X_1^{(n)}) \,/\, q_{1,\theta}(X_1^{(n)})$ (re-weight)
2. For $n \in [\![1, N_x]\!]$:
   - Set $W_{1,\theta}^{(n)} = w_{1,\theta}^{(n)} \,/\, \sum_{i=1}^{N_x} w_{1,\theta}^{(i)}$ (normalize weights)
3. For $t \in [\![2, T]\!]$: (assimilate next observations)
   (a) Sample indexes $A_{t-1}^{(1)}, ..., A_{t-1}^{(N_x)}$ from $[\![1, N_x]\!]$ with respective weights $W_{t-1,\theta}^{(1:N_x)}$ and with replacement using a prespecified resampling scheme (sample ancestors)
   (b) For $n \in [\![1, N_x]\!]$:
   - Sample $X_t^{(n)} \sim q_{t,\theta}(dx_t \,|\, X_{t-1}^{(A_{t-1}^{(n)})})$ (move particles)
   - Set $w_{t,\theta}^{(n)} = f_\theta(X_t^{(n)} \,|\, X_{t-1}^{(A_{t-1}^{(n)})})\, g_\theta(y_t \,|\, X_t^{(n)}) \,/\, q_{t,\theta}(X_t^{(n)} \,|\, X_{t-1}^{(A_{t-1}^{(n)})})$ (re-weight)
   (c) For $n \in [\![1, N_x]\!]$:
   - Set $W_{t,\theta}^{(n)} = w_{t,\theta}^{(n)} \,/\, \sum_{i=1}^{N_x} w_{t,\theta}^{(i)}$ (normalize weights)
4. Output the weighted samples $(W_{t,\theta}^{(1:N_x)}, X_t^{(1:N_x)})_{t \in [\![1,T]\!]}$ respectively targeting the successive filtering distributions $p(dx_t \,|\, y_{1:t}, \theta)$ for all $t \in [\![1, T]\!]$. As a byproduct,

$$\hat{Z}_{t,\theta} \;=\; \prod_{s=1}^{t} \left(\frac{1}{N_x} \sum_{n=1}^{N_x} w_{t,\theta}^{(n)} \right)$$

is an unbiased estimator of the likelihood $p(y_{1:t}|\theta)$ for all $t \in [\![1, T]\!]$.

A.2.2 SMC^2 algorithm

Although the incremental likelihoods $p(y_{t+1} \,|\, y_{1:t}, \theta_t^{(m)})$ — needed for the re-weighting step in SMC algorithms — cannot be directly evaluated in the context of state-space models, we

can run a particle filter conditional on $\theta_t^{(m)}$ (Section A.2.1 and Algorithm 3) to obtain non-negative estimators $\hat{G}_{t+1}^{(m)}$ of the incremental likelihoods $p(y_{t+1}|y_{1:t}, \theta_t^{(m)})$, while ensuring that $\hat{Z}_{t+1}^{(m)} = \prod_{s=1}^{t+1} \hat{G}_s^{(m)}$ is an unbiased estimator of the likelihood $p(y_{1:t+1}|\theta_t^{(m)})$ for all $t \in [\![1, T]\!]$. The SMC^2 algorithm (Chopin et al., 2013) then amounts to an SMC sampler on an extended target space, where each particle $\theta_t^{(m)}$ is augmented to form $(\theta_t^{(m)}, \mathrm{PF}_t^{(m)})$, with $\mathrm{PF}_t^{(m)}$ denoting all the variables generated by running a particle filter with observations $y_{1:t}$ and parameter $\theta_t^{(m)}$. In particular, these variables contain a system of N_x weighted particles $(W_{x,t}^{(1:N_x,m)}, X_t^{(1:N_x,m)})$ that is used to get the required unbiased estimators of the likelihood $p(y_{1:t} \,|\, \theta_t^{(m)})$. Our implementation uses bootstrap particle filters (Gordon et al., 1993) for simplicity, but more efficient versions — such as the auxiliary particle filter (Pitt and Shephard, 1999) — could be used within SMC^2, as illustrated in Golightly and Kypraios (2017). When the $\theta_t^{(m)}$'s get rejuvenated, their respective $\mathrm{PF}_t^{(m)}$'s are resampled and re-generated accordingly. As a particular case of an SMC sampler, the discussion from Section A.1.2 readily applies to SMC^2 and allows for adaptive choices of temperatures, resampling times, and proposal distributions. An additional layer of adaption consists in allowing the size N_x of the particle system attached to each $\theta_t^{(m)}$ to dynamically grow during the course of the algorithm, i.e. one starts with a small initial number N_x to ease the computational burden, and adaptively doubles it whenever the acceptance rate of the moves in the rejuvenation phase are below some desired threshold. Increasing N_x is achieved by using conditional SMC steps, which sample new $\mathrm{PF}_t^{(m)}$'s made of a larger number of particles conditional on an existing path from the current $\mathrm{PF}_t^{(m)}$'s, as detailed in Section 3.6.2 of Chopin et al. (2013). Other relevant considerations are discussed e.g. in Chopin, Ridgway, Gerber and Papaspiliopoulos (2015) and Duan and Fulop (2015).

The output of SMC^2 consists of a succession of weighted samples $(W_{\theta,t}^{(1:N_\theta)}, \theta_t^{(1:N_\theta)})$ targeting the posterior distributions $p(d\theta \,|\, y_{1:t})$ for all $t \in [\![1, T]\!]$, with each $\theta_t^{(m)}$ carrying its attached weighted particles $(W_{x,t}^{(1:N_x,m)}, X_t^{(1:N_x,m)})$ targeting the filtering distributions $p(dx_t \,|\, y_{1:t}, \theta_t^{(m)})$. For all $t \in [\![1, T]\!]$ and all suitably integrable test functions h,

conditional expectations $\mathbb{E}[h(\Theta, X_t) \,|\, y_{1:t}]$ with respect to the joint posterior distributions $(\Theta, X_t) \sim p(d\theta \,|\, y_{1:t}) p(dx_t \,|\, y_{1:t}, \theta)$ can be consistently estimated by estimators of the form $\sum_{m=1}^{N_\theta} W_{\theta,t}^{(m)} \sum_{n=1}^{N_x} W_{x,t}^{(n,m)} h(\theta_t^{(m)}, X_t^{(n,m)})$, as $N_\theta \to +\infty$. As an SMC sampler, the non-adaptive version of SMC^2 also produces unbiased estimators of the marginal likelihoods $p(y_{1:t})$ for all $t \in [\![1, T]\!]$, similarly to the ones from (A.1).

A.3 Implementation of unbiased MCMC methods

Many popular MCMC algorithms (see e.g. Robert and Casella, 1999; Liu, 2008; Brooks et al., 2011; Green et al., 2015) have been successfully debiased by the use of well-designed couplings (Jacob et al., 2017; Middleton et al., 2018; Heng and Jacob, 2019; Middleton et al., 2019). In this appendix, we focus on unbiased variants of the Metropolis-Hastings (MH) algorithm targeting posterior distributions, which are used in Section 2.3 of Chapter 2.

When targeting posterior distributions, the standard MH algorithm (Hastings, 1970) requires evaluations of the likelihood to compute the acceptance ratio of new proposals. In the context of state-space models where likelihoods are intractable, the MH algorithm can still be run by simply substituting the unavailable likelihoods by unbiased estimators obtained from particle filters (Appendix A.2.1), leading to new variants known as the particle independent Metropolis-Hastings (PIMH) and the particle marginal Metropolis-Hastings (PMMH) algorithms (Andrieu et al., 2010, Sections 2.4.1 and 2.4.2 respectively). The substitution of the likelihood by an unbiased estimator can be justified in the spirit of pseudo-marginal methods (Andrieu and Roberts, 2009) and these particle MH algorithms turn out to be exact MH algorithms targeting an extended distribution, for which one of the marginals coincide with the original target distribution of interest. Bringing these algorithms into the unbiased MCMC toolbox has been recently achieved, leading to the coupled PMMH (Middleton et al., 2018) and the coupled PIMH (Middleton et al., 2019) algorithms. We

are interested in the latter, and more specifically in the use of SMC samplers as proposal distributions (Appendix A.3.1) within coupled PIMH algorithms (Appendix A.3.2).

A.3.1 SMC samplers as joint proposal distributions

As reviewed in Appendices A.1 and A.2, one run of SMC over T observations produces a collection of weighted particles $(W_{\theta,t}^{(m)}, \theta_t^{(m)})_{(t,m)\in[\![1,T]\!]\times[\![1,N_\theta]\!]}$ approximating the successive posterior distributions $(p(d\theta|y_{1:t}))_{t\in[\![1,T]\!]}$, and a collection of non-negative unbiased estimators $(\hat{Z}_t^{(m)})_{(t,m)\in[\![1,T]\!]\times[\![1,N_\theta]\!]}$ of the successive evidences $(Z_t)_{t\in[\![1,T]\!]} = (p(y_{1:t}))_{t\in[\![1,T]\!]}$. In the case of SMC^2 samplers, each $\theta_t^{(m)}$ also carries its own system of weighted particles $(W_{x,t}^{(n,m)}, X_t^{(n,m)})_{n\in[\![1,N_x]\!]}$ targeting the filtering distributions $p(dx_t \,|\, y_{1:t}, \theta_t^{(m)})$. Algorithm 4 in Appendix A.3.2 explains how to turn the output of an SMC sampler into a joint proposal distribution that can be used in coupled PIMH algorithms. The unbiasedness of the $\hat{Z}_t^{(m)}$'s will be crucial to ensure the validity of the coupled PIMH, as the latter relies on a pseudo-marginal MH kernel, for which unbiasedness of the likelihood estimators is paramount (Andrieu and Roberts, 2009; Andrieu et al., 2010). This prohibits the use of any adaptive resampling schemes when using SMC-based proposals within coupled PIMH algorithms. As discussed in Appendix A.1.2, adaptation can still be used in a preliminary run, as a way to tune the algorithmic parameters before performing the final run with non-adaptively.

A.3.2 Coupled PIMH

As inspired by Section 2.4 from Middleton et al. (2018), our implementation of coupled PIMH concurrently runs T pairs of Markov chains, such that, for $t \in [\![1, T]\!]$, pair t evolves according to a coupled particle MH kernel targeting the posterior $p(d\theta|y_{1:t})$ at horizon t. The proposals for the MH kernels at each horizon $t \in [\![1, T]\!]$ can be generated jointly by a run of SMC or SMC^2 over the whole time horizon $[\![1, T]\!]$. Using Algorithm 4, one run of SMC provides T proposed draws that can be fed to the respective T pairs of chains, as

Algorithm 4 SMC as a joint proposal distribution

1. Perform one run of SMC over all T observations to produce weighted particles $(W_{\theta,1:T}^{(1:N_\theta)}, \theta_{1:T}^{(1:N_\theta)})$, non-negative unbiased estimators $\hat{Z}_{1:T}^{(1:N_\theta)}$ of evidences [and accessory weigthed particles $(W_{x,1:T}^{(1:N_x,1:N_\theta)}, X_{1:T}^{(1:N_x,1:N_\theta)})$ for SMC^2] as described in Appendix A.3.1.
2. For $t \in [\![1, T]\!]$:
   - Draw m from $[\![1, N_\theta]\!]$ with respective probabilities $W_{\theta,t}^{(1:N_\theta)}$.
   - Set $\theta_t^* = \theta_t^{(m)}$.
   - Set $Z_t^* = \hat{Z}_t^{(m)}$ [and $(W_{x,t}^{*(1:N_x)}, X_t^{*(1:N_x)}) = (W_{x,t}^{(1:N_x,m)}, X_t^{(1:N_x,m)})$ for SMC^2].
3. Jointly output $(\theta_t^*, Z_t^*)_{t\in[\![1,T]\!]}$ [and $(W_{x,t}^{*(1:N_x)}, X_t^{*(1:N_x)})_{t\in[\![1,T]\!]}$ for SMC^2] as proposed draws for each corresponding horizon $t \in [\![1, T]\!]$.

   By construction the θ_t^*'s are drawn from the respective particle approximations $\sum_{m=1}^{N_\theta} W_{\theta,t}\delta_{\theta_t^{(m)}}$ of the successive posterior distributions $p(d\theta|y_{1:t})$ and the Z_t^*'s unbiasedly estimate $p(y_{1:t})$ for all $t \in [\![1, T]\!]$.

explained in Algorithm 5. Under mild conditions, the coupled PIMH algorithm produces estimators (2.13) that are unbiased, while having both finite variances and expected costs (Middleton et al., 2019, Proposition 3). The distribution of the meeting times is derived in Proposition 8 of Middleton et al. (2019), and can be conveniently simulated so as to help anticipate and tune the behavior of the couple PIMH algorithm, as explained in Section 2.3.4 of Chapter 2. Further guidelines and theoretical guarantees for coupled PIMH algorithms can be found in Middleton et al. (2019).

Algorithm 5 Coupled PIMH

1. Initialization: set $n = 1$ and $\tau_t = +\infty$ for all $t \in [\![1, T]\!]$.
2. Use SMC [or SMC^2] as a proposal via Algorithm 4 to sample $(\theta_t^{(0)}, \hat{Z}_t^{(0)})_{t\in[\![1,T]\!]}$ [and $(W_{x,t}^{(1:N_x,0)}, X_t^{(1:N_x,0)})_{t\in[\![1,T]\!]}$ for SMC^2]. (initialize chain 1 for all horizons)
3. While $n < \max(m, \tau_{1:T})$: (if $n < m$ or one pair of chains still haven't met)
   - Use SMC [or SMC^2] as a proposal via Algorithm 4 to propose $(\theta_t^*, \hat{Z}_t^*)_{t\in[\![1,T]\!]}$ [and $(W_{x,t}^{(1:N_x,*)}, X_t^{(1:N_x,*)})_{t\in[\![1,T]\!]}$ for SMC^2]. (generate proposals for all horizons)
   - For $t \in [\![1, T]\!]$: (grow one pair of coupled chains for each horizon t)
     (a) Draw $U_t \sim \text{Unif}(0, 1)$.
     (b) If $U_t \leq \min(1, \hat{Z}_t^*/\hat{Z}_t^{(n-1)})$: set $(\theta_t^{(n)}, \hat{Z}_t^{(n)}) = (\theta_t^*, \hat{Z}_t^*)$ (chain 1 accepts proposal)
     [and $(W_{x,t}^{(1:N_x,n)}, X_t^{(1:N_x,n)}) = (W_{x,t}^{(1:N_x,*)}, X_t^{(1:N_x,*)})$ for SMC^2].

     Else: set $(\theta_t^{(n)}, \hat{Z}_t^{(n)}) = (\theta_t^{(n-1)}, \hat{Z}_t^{(n-1)})$ (chain 1 rejects proposal)
     [and $(W_{x,t}^{(1:N_x,n)}, X_t^{(1:N_x,n)}) = (W_{x,t}^{(1:N_x,n-1)}, X_t^{(1:N_x,n-1)})$ for SMC^2].

     (c) If $n = 1$: set $\tilde{Z}_t^{(n-2)} = \hat{Z}_t^*$ by convention. (chain 2 initialized as first proposal)
     (d) If $U_t \leq \min(1, \hat{Z}_t^*/\tilde{Z}_t^{(n-2)})$: (chain 2 accepts proposal)
     set $(\tilde{\theta}_t^{(n-1)}, \tilde{Z}_t^{(n-1)}) = (\theta_t^*, \hat{Z}_t^*)$.
     [and $(\tilde{W}_{x,t}^{(1:N_x,n-1)}, \tilde{X}_t^{(1:N_x,n-1)}) = (W_{x,t}^{(1:N_x,*)}, X_t^{(1:N_x,*)})$ for SMC^2].

     Else: (chain 2 rejects proposal)
     set $(\tilde{\theta}_t^{(n-1)}, \tilde{Z}_t^{(n-1)}) = (\tilde{\theta}_t^{(n-2)}, \tilde{Z}_t^{(n-2)})$.
     [and $(\tilde{W}_{x,t}^{(1:N_x,n-1)}, \tilde{X}_t^{(1:N_x,n-1)}) = (\tilde{W}_{x,t}^{(1:N_x,n-2)}, \tilde{X}_t^{(1:N_x,n-2)})$ for SMC^2].

     (e) If $U_t \leq \min(1, \hat{Z}_t^*/\hat{Z}_t^{(n-1)}, \hat{Z}_t^*/\tilde{Z}_t^{(n-2)})$: (both chains have accepted)
     set $\tau_t = n$ (chains for horizon t have met after n steps)
     (f) Set $n = n + 1$
4. Output the pair $\mathcal{V}_t = (\theta_t^{(n)})_{n\in[\![0,\max(m,\tau_{1:T})-1]\!]}$ and $\mathcal{W}_t = (\tilde{\theta}_t^{(n)})_{n\in[\![0,\max(m,\tau_{1:T})-2]\!]}$ of coupled chains for each time horizon $t \in [\![1, T]\!]$.

   [along with the corresponding systems of particles $(W_{x,t}^{(1:N_x,n)}, X_t^{(1:N_x,n)})_{n\in[\![0,\max(m,\tau_{1:T})-1]\!]}$ and $(\tilde{W}_{x,t}^{(1:N_x,n)}, \tilde{X}_t^{(1:N_x,n)})_{n\in[\![0,\max(m,\tau_{1:T})-2]\!]}$ for SMC^2]

B

Supplementary material for Chapter 1

Contents

This appendix provides additional examples to illustrate the robustness of the H-score to vague priors (Appendix B.1) and its consistency for linear time series models (Appendix B.2). We also present a rigorous construction of the H-score for discrete observations (Appendix B.3), and we provide detailed proofs (Appendices B.4 and B.5) of the results from Chapter 1.

B.1 Illustration of robustness with Normal models

This section complements Section 1.2.3 of Chapter 1. We consider the two Normal models

$$M_1: \quad Y_1, ..., Y_T \,|\, \theta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta_1, 1\right), \quad \theta_1 \sim \mathcal{N}\left(0, \sigma_0^2\right),$$
$$M_2: \quad Y_1, ..., Y_T \,|\, \theta_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \theta_2\right), \quad \theta_2 \sim \text{Inv-}\chi^2\left(\nu_0, s_0^2\right).$$

The positive hyperparameters are chosen as $\nu_0 = 0.1$ and $s_0^2 = 1$. We compare M_1 and M_2 using observations generated as $Y_1, ..., Y_T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(1, 1)$, for different $\log(\sigma_0) \in \{0, 150, 350\}$, i.e. for increasingly vague priors on μ. In this setting, M_1 is well-specified whereas M_2 is misspecified.

Under M_1, we have $Y_t \,|\, Y_{1:t-1} \sim \mathcal{N}\left(\mu_{t-1}, \sigma_{t-1}^2 + 1\right)$ for all $t \in [\![0, T]\!]$ by conjugacy, with $\sigma_t^2 = (t + \sigma_0^{-2})^{-1}$ and $\mu_t = \sigma_t^2 \sum_{i=1}^t Y_i$ for all $t \in [\![1, T]\!]$. Under model M_2, we have $Y_t \,|\, Y_{1:t-1} \sim t_{\nu_{t-1}}\left(0, s_{t-1}^2\right)$ for all $t \in [\![0, T]\!]$ by conjugacy, with $\nu_t = \nu_0 + t$ and $s_t = (\nu_0 s_0^2 + \sum_{i=1}^t Y_i^2)/\nu_t$ for all $t \in [\![1, T]\!]$. Given $Y_{1:T}$, these conjugacy results allow us to compute the log-Bayes factor and H-factor analytically.

We generate 100 independent samples, each consisting of $T = 1000$ i.i.d. draws from $\mathcal{N}(1, 1)$. For each sample $Y_{1:T}$ and every $\log(\sigma_0) \in \{0, 150, 350\}$, we compute the log-Bayes factor and H-score of M_1 against M_2. The results are shown in Figure B.1. For any fixed value of σ_0, both the H-factor and log-Bayes factor are consistent, in the sense of asymptotically choosing the correct model M_1, as $T \to +\infty$. However, for any fixed T, no matter how large, there always exists a large enough σ_0 such that the log-Bayes factor chooses the wrong model M_2 with arbitrarily high probability. This is because $\log p_{M_1}(Y_{1:T})$ behaves equivalently to $-\log(\sigma_0)$ as $\sigma_0 \to +\infty$. This sensitivity of the log-Bayes factor to the vagueness of priors is arguably undesirable, as it can lead to choosing a misspecified model over a well-specified one. This artifact is even more unsettling when considering that, with $T = 1000$ observations, all three prior specifications $\log(\sigma_0) \in \{0, 150, 350\}$ essentially lead to the same posterior on

μ, i.e. the same model fit in some sense. By contrast, the H-factor is virtually unchanged when σ_0 is increased beyond a certain value.

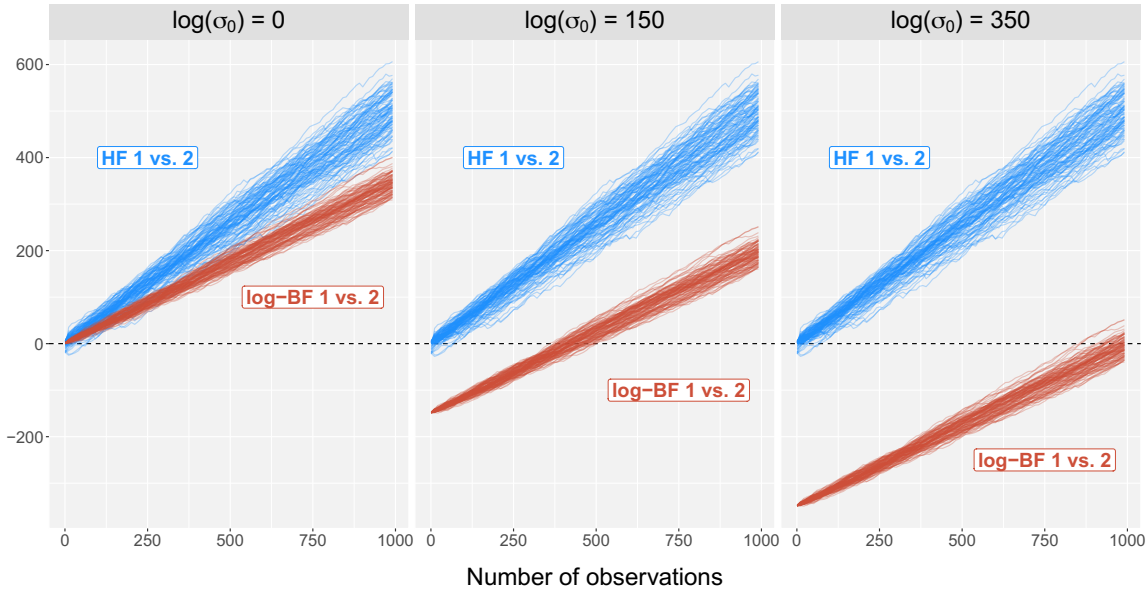

Figure B.1. *Exact log-Bayes factors (log-BF) and H-factors (HF) of M_1 against M_2, computed for 100 independent samples (thin solid lines) generated as i.i.d. $\mathcal{N}(1,1)$, under three increasingly vague priors on μ:* $\log(\sigma_0) = 0$ *(left panel),* $\log(\sigma_0) = 150$ *(middle panel),* $\log(\sigma_0) = 350$ *(right panel). See Appendix B.1.*

B.2 Illustration of consistency with ARMA models

Define the stationarity triangle $\mathbb{S} = \{(\phi_1, \phi_2) \in \mathbb{R}^2 : |\phi_2| < 1,\ \phi_2 - \phi_1 < 1,\ \phi_2 + \phi_1 < 1\}$. Let Unif($\mathbb{S}$) denote the bivariate uniform distribution on the set $\mathbb{S}$ and let $(\varepsilon_t)_{t\in\mathbb{N}}$ denote a sequence of i.i.d. standard Normal variables. We consider the following time series models, corresponding respectively to AR(1), AR(2), and MA(1) models.

M_1: $Y_1 \,|\, \phi, \sigma^2 \sim \mathcal{N}\left(0, \sigma^2/(1-\phi^2)\right)$; $\quad Y_t = \phi Y_{t-1} + \sigma\varepsilon_t\,,\ \forall t \geq 2$;

with independent priors $\phi \sim \text{Unif}(-1,1)$ *and* $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2)$.

M_2: $Y_1, Y_2 \,|\, \phi_1, \phi_2, \sigma^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{(1-\phi_2)\sigma^2/(1+\phi_2)}{(1-\phi_2-\phi_1)(1-\phi_2+\phi_1)}\right)$; $\quad Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \sigma\varepsilon_t\,,\ \forall t \geq 3$;

with independent priors $(\phi_1, \phi_2) \sim \text{Unif}(\mathbb{S})$ *and* $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2)$.

M_3: $Y_t = \sigma(\varepsilon_t + \theta\varepsilon_{t-1})$, $\forall t \geq 1$;

with independent priors $\theta \sim \text{Unif}(-1, 1)$ *and* $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2)$.

The positive hyperparameters are set to $\nu_0 = 1$ and $s_0^2 = 1$. First, we consider a non-nested setting by comparing M_1 and M_3 under the following two data-generating processes:

(1) AR(1): $Y_1 \sim \mathcal{N}(0, 1)$ and $Y_t = 0.6\, Y_{t-1} + 0.8\, \varepsilon_t$, i.e. M_1 is well-specified while M_3 is not.

(2) MA(1): $Y_t = \varepsilon_t + 0.5\, \varepsilon_{t-1}$, i.e. M_3 is well-specified while M_1 is not.

ARMA models can be regarded as particular cases of linear Gaussian state-space models, whose likelihood can be computed using Kalman filters. Thus, H-scores of ARMA models can be estimated by directly using SMC in conjunction with Kalman filters, instead of more sophisticated SMC^2 algorithms. For each data-generating process, we generate $T = 1000$ observations and estimate the H-score of M_1 and M_3 via SMC with $N_\theta = 1024$ particles. The estimated H-factors and log-Bayes factors of M_1 against M_3 are shown in Figure B.2. We see that the H-factor asymptotically chooses the correct model.

We now consider a nested setting by comparing M_1 and M_2 under the following two data-generating processes:

(3) AR(1): $Y_1 \sim \mathcal{N}(0, 1)$ and $Y_t = 0.6\, Y_{t-1} + 0.8\, \varepsilon_t$, i.e. both M_1 and M_2 are well-specified.

(4) AR(2): $Y_1, Y_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $Y_t = 0.25\, Y_{t-1} + 0.5\, Y_{t-2} + 0.75\, \varepsilon_t$, i.e. M_2 is well-specified but M_1 is not.

The data-generating processes are initialized at their respective stationary distributions. For each case, we generate $T = 1000$ observations and estimate the H-score of M_1 and M_2 via SMC with $N_\theta = 1024$ particles. The respective H-factors and log-Bayes factors of M_1 against M_2 are shown in Figure B.3. Case 3 suggests that, when dealing with nested well-specified models, the H-factor asymptotically favors the model of smallest dimension.

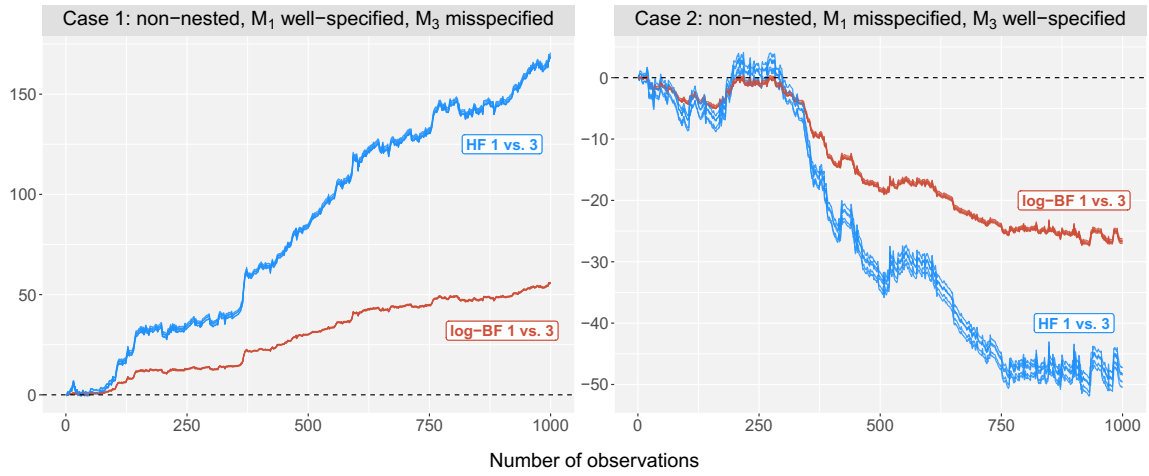

Figure B.2. *Estimated log-Bayes factors (log-BF) and H-factors (HF) of M_1 against M_3, computed for 5 replications (thin solid lines), under two data-generating processes: AR(1) (Case 1) and MA(1) (Case 2). See Appendix B.2.*

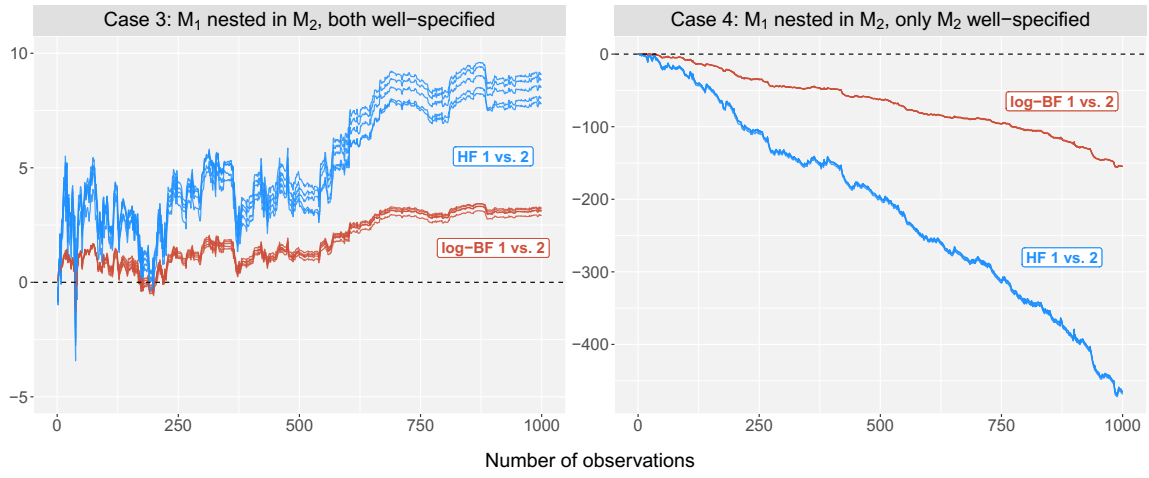

Figure B.3. *Estimated log-Bayes factors (log-BF) and H-factors (HF) of M_1 against M_2, computed for 5 replications (thin solid lines), under two data-generating processes: AR(1) (Case 3) and AR(2) (Case 4). See Appendix B.2.*

B.3 Propriety of the H-score for discrete observations

0-homogeneous score functions for discrete observations are proper if and only if they are super-gradients of 1-homogeneous concave entropy functions (McCarthy, 1956; Hendrickson and Buehler, 1971). It follows that we can construct a proper 0-homogeneous scoring rule in terms of a collection of homogeneous functions over the cliques of an undirected graph on the space $\mathbb{Y} = [\![a_1, b_1]\!] \times ... \times [\![a_{d_y}, b_{d_y}]\!]$ (Dawid et al., 2012). More precisely, let $\mathcal{G}$ denote an undirected graph with a set of nodes equal to $\mathbb{Y}$ and a set of edges defined as $\{(y_1, y_2) \in \mathbb{Y}^2 : y_1 - y_2 \in \{-2e_k, -e_k, e_k, 2e_k\} \text{ for some } k \in [\![1, d_y]\!]\}$. Here e_k denotes the canonical vector of $\mathbb{Z}^{d_y}$ that has all coordinates equal to 0 except for its k-th coordinate that equals 1. The cliques (maximal complete subsets) of this graph are of the form $\{y-e_k, y, y+e_k\}$. Define the function $H : (0,\infty)^3 \to \mathbb{R}$ as $H(p_1, p_2, p_3) = -(p_3 - p_1)^2/p_2$. This function is 1-homogeneous and concave. Indeed, for any $\lambda > 0$, we have $H(\lambda p_1, \lambda p_2, \lambda p_3) = \lambda H(p_1, p_2, p_3)$. Besides, the Hessian of H at any $(p_1, p_2, p_3) \in (0,\infty)^3$ is given by

$$\begin{pmatrix} -\frac{2(p_3-p_1)^2}{p_2^3} & \frac{2(p_3-p_1)}{p_2^2} & -\frac{2(p_3-p_1)}{p_2^2} \\ \frac{2(p_3-p_1)}{p_2^2} & -\frac{2}{p_2} & \frac{2}{p_2} \\ -\frac{2(p_3-p_1)}{p_2^2} & \frac{2}{p_2} & -\frac{2}{p_2} \end{pmatrix}.$$

For all $(p_1, p_2, p_3) \in (0,\infty)^3$, the determinants of the extracted matrices

$$\left(-\frac{2(p_3-p_1)^2}{p_2^3}\right), \left(-\frac{2}{p_2}\right), \begin{pmatrix} -\frac{2(p_3-p_1)^2}{p_2^3} & \frac{2(p_3-p_1)}{p_2^2} \\ \frac{2(p_3-p_1)}{p_2^2} & -\frac{2}{p_2} \end{pmatrix}, \begin{pmatrix} -\frac{2}{p_2} & \frac{2}{p_2} \\ \frac{2}{p_2} & -\frac{2}{p_2} \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} -\frac{2(p_3-p_1)^2}{p_2^3} & -\frac{2(p_3-p_1)}{p_2^2} \\ -\frac{2(p_3-p_1)}{p_2^2} & -\frac{2}{p_2} \end{pmatrix}$$

are respectively negative, negative, 0, 0, and 0. The determinant of the Hessian is also equal to 0. In other words, all the principal minors of the negative Hessian are non-negative. By Sylvester's criterion (Horn and Johnson, 1985), this implies that the negative Hessian of H at (p_1, p_2, p_3) is positive semi-definite, for all $(p_1, p_2, p_3) \in (0,\infty)^3$, which proves that the function H is concave.

Following the construction from Section 3.3 of Dawid et al. (2012), we can define, for all probability mass functions p on $\mathbb{Y}$, the concave entropy function

$$\mathcal{E}_{\mathcal{H}^D}(p) = -\sum_{k=1}^{d_y} \sum_{\substack{y \in \mathbb{Y} \text{ s.t.} \\ a_k < y_{(k)} < b_k}} p(y) \left(\frac{p(y+e_k) - p(y-e_k)}{2\,p(y)} \right)^2, \tag{B.1}$$

whose associated score function is given by

$$\mathcal{H}^D(y,p) = \sum_{k=1}^{d_y} \mathcal{H}_k^D(y,p)\,,$$

where

$$\mathcal{H}_k^D(y,p) = \begin{cases} \frac{p(y+2\,e_k)-p(y)}{2\,p(y+e_k)} & \text{if } y_{(k)} = a_k, \\ \frac{p(y+2\,e_k)-p(y)}{2\,p(y+e_k)} + \left(\frac{p(y+e_k)-p(y-e_k)}{2\,p(y)}\right)^2 & \text{if } y_{(k)} = a_k + 1, \\ \frac{p(y+2\,e_k)-p(y)}{2\,p(y+e_k)} - \frac{p(y)-p(y-2\,e_k)}{2\,p(y-e_k)} + \left(\frac{p(y+e_k)-p(y-e_k)}{2\,p(y)}\right)^2 & \text{if } a_k + 1 < y_{(k)} < b_k - 1, \\ -\frac{p(y)-p(y-2\,e_k)}{2\,p(y-e_k)} + \left(\frac{p(y+e_k)-p(y-e_k)}{2\,p(y)}\right)^2 & \text{if } y_{(k)} = b_k - 1, \\ -\frac{p(y)-p(y-2\,e_k)}{2\,p(y-e_k)} & \text{if } y_{(k)} = b_k. \end{cases}$$

The concavity of the entropy function guarantees that $\mathcal{H}^D$ is a proper scoring rule. The entropy in (B.1) can be interpreted as a discrete analog of the entropy function of the H-score for continuous observations, which is given by $-\int_{\mathbb{Y}} \|\nabla_y \log p(y)\|^2\, p(y) dy$ under mild regularity assumptions (Hyvärinen, 2005; Dawid and Musio, 2015).

The alternative definition using forward differences, given by

$$\begin{cases} 2\left(\frac{p(y+e_k)-p(y)}{p(y)}\right) + \left(\frac{p(y+e_k)-p(y)}{p(y)}\right)^2 & \text{if } y_{(k)} = a_k, \\ 2\left(\frac{p(y+e_k)-p(y)}{p(y)} - \frac{p(y)-p(y-e_k)}{p(y-e_k)}\right) + \left(\frac{p(y+e_k)-p(y)}{p(y)}\right)^2 & \text{if } a_k < y_{(k)} < b_k, \\ -2\left(\frac{p(y)-p(y-e_k)}{p(y-e_k)}\right) & \text{if } y_{(k)} = b_k, \end{cases}$$

is a particular case of the pair scoring rule from Example 4.1 in Dawid et al. (2012), where we choose the concave function G to be $u \mapsto -(u-1)^2$.

B.4 Identities for the H-score

In this section, we fix a model M and drop the dependence on the model in the notation. Equations (1.4), (1.11), and (1.12) result from algebraic manipulations, under assumptions guaranteeing the existence of all the relevant derivatives and integrals, as well as enabling differentiation under the integral sign. Such assumptions can be stated as follows.

Assumption A10. *For all $t \in \mathbb{N}^*$, the following conditions hold:*

(a) $\theta \mapsto p(y_t|y_{1:t-1}, \theta)\, p(\theta|y_{1:t-1})$ *is integrable on* $\mathbb{T}$ *for all* $y_{1:t} \in \mathbb{Y}^t$.

(b) $y_t \mapsto p(y_t|y_{1:t-1}, \theta)$ *is twice differentiable on* $\mathbb{Y}$ *for all* $(y_{1:t-1}, \theta) \in \mathbb{Y}^{t-1} \times \mathbb{T}$.

(c) For all $k \in [\![1, d_y]\!]$, *there exist integrable functions* $h_{1,k,t}$ *and* $h_{2,k,t}$ *such that*

$$\left|\frac{\partial p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}}\, p(\theta|y_{1:t-1})\right| \leq h_{1,k,t}(\theta) \qquad \text{and} \qquad \left|\frac{\partial^2 p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}^2}\, p(\theta|y_{1:t-1})\right| \leq h_{2,k,t}(\theta)$$

for all $(y_{1:t}, \theta) \in \mathbb{Y}^t \times \mathbb{T}$.

Assumption A11. *For all $t \in \mathbb{N}^*$, the following conditions hold:*

(a) $x_t \mapsto p(x_t|y_{1:t-1}, \theta)\, g_\theta(y_t|x_t)$ *is integrable on* $\mathbb{X}$ *for all* $(y_{1:t}, \theta) \in \mathbb{Y}^t \times \mathbb{T}$.

(b) $y_t \mapsto g_\theta(y_t|x_t)$ *is twice differentiable on* $\mathbb{Y}$ *for all* $(\theta, x_t) \in \mathbb{T} \times \mathbb{X}$.

(c) For all $k \in [\![1, d_y]\!]$, *there exist integrable functions* $h_{3,k,t}$ *and* $h_{4,k,t}$ *such that*

$$\left|\frac{\partial g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}}\, p(x_t|y_{1:t-1}, \theta)\right| \leq h_{3,k,t}(x_t) \qquad \text{and} \qquad \left|\frac{\partial^2 g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}^2}\, p(x_t|y_{1:t-1}, \theta)\right| \leq h_{4,k,t}(x_t)$$

for all $(y_{1:t}, \theta, x_t) \in \mathbb{Y}^t \times \mathbb{T} \times \mathbb{X}$.

B.4.1 Proof of (1.4)

Consider some generic prior $p(\theta)$ and likelihood $p(y|\theta)$. Assume that $\theta \mapsto p(y|\theta)p(\theta)$ is integrable for every $y \in \mathbb{Y}$, $y \mapsto p(y|\theta)$ is twice differentiable on $\mathbb{Y}$ for every $\theta \in \mathbb{T}$, and, for all $k \in \{1, ..., d_y\}$, both $\theta \mapsto \left|\frac{\partial\, p(y|\theta)}{\partial y_{(k)}}\, p(\theta)\right|$ and $\theta \mapsto \left|\frac{\partial^2\, p(y|\theta)}{\partial {y_{(k)}}^2}\, p(\theta)\right|$ are dominated by integrable functions on $\mathbb{T}$. Let $p(y) = \int_{\mathbb{T}} p(y|\theta)p(\theta)d\theta$. The previous assumptions allow us to partially differentiate $y \mapsto p(y)$ twice under the integral sign with respect to each coordinate. Recall from (1.1) the definition of the H-score,

$$\mathcal{H}(y, p) = \sum_{k=1}^{d_y} \left(2\, \frac{\partial^2 \log p(y)}{\partial y_{(k)}^2} + \left(\frac{\partial \log p(y)}{\partial y_{(k)}} \right)^2 \right).$$

For all $k \in [\![1, d_y]\!]$, partial differentiation under the integral sign yields, on the one hand,

$$\begin{aligned} \frac{\partial \log p(y)}{\partial y_{(k)}} &= \frac{1}{p(y)} \int \left(\frac{\partial\, p(y|\theta)}{\partial y_{(k)}} \right) p(\theta) d\theta \\ &= \int \left(\frac{\partial\, \log p(y|\theta)}{\partial y_{(k)}} \right) p(\theta|y) d\theta = \mathbb{E}\left[\left. \frac{\partial\, \log p(y|\Theta)}{\partial y_{(k)}} \right| y \right]. \end{aligned}$$

On the other hand, partially differentiating twice under the integral sign yields

$$\begin{aligned} \frac{\partial^2 \log p(y)}{\partial {y_{(k)}}^2} &= -\left(\frac{\partial\, \log p(y)}{\partial y_{(k)}} \right)^2 + \frac{1}{p(y)} \frac{\partial^2\, p(y)}{\partial {y_{(k)}}^2} \\ &= -\left(\frac{\partial\, \log p(y)}{\partial y_{(k)}} \right)^2 + \frac{1}{p(y)} \int \left(\frac{\partial^2\, p(y|\theta)}{\partial {y_{(k)}}^2} \right) p(\theta) d\theta. \end{aligned}$$

Regarding the integrand in the last term, we have

$$\frac{\partial^2\, p(y|\theta)}{\partial {y_{(k)}}^2} = p(y|\theta) \left[\frac{\partial^2 \log p(y|\theta)}{\partial {y_{(k)}}^2} + \left(\frac{\partial\, \log p(y|\theta)}{\partial y_{(k)}} \right)^2 \right].$$

This leads to

$$\begin{aligned} \frac{\partial^2 \log p(y)}{\partial {y_{(k)}}^2} &= -\left(\frac{\partial\, \log p(y)}{\partial y_{(k)}} \right)^2 + \int p(\theta|y) \left[\frac{\partial^2 \log p(y|\theta)}{\partial {y_{(k)}}^2} + \left(\frac{\partial\, \log p(y|\theta)}{\partial y_{(k)}} \right)^2 \right] d\theta \\ &= -\left(\frac{\partial\, \log p(y)}{\partial y_{(k)}} \right)^2 + \mathbb{E}\left[\left. \frac{\partial^2 \log p(y|\Theta)}{\partial {y_{(k)}}^2} + \left(\frac{\partial\, \log p(y|\Theta)}{\partial y_{(k)}} \right)^2 \right| y \right]. \end{aligned}$$

By putting everything together we finally get

$$\mathcal{H}(y,p) = \sum_{k=1}^{d_y} \left(\mathbb{E}\left[2\,\frac{\partial^2 \log p(y|\Theta)}{\partial {y_{(k)}}^2} + 2\left(\frac{\partial \log p(y|\Theta)}{\partial y_{(k)}}\right)^2 \middle| y \right] - \left(\mathbb{E}\left[\frac{\partial \log p(y|\Theta)}{\partial y_{(k)}} \middle| y \right] \right)^2 \right). \quad \text{(B.2)}$$

For a given model M with parameter $\theta \in \mathbb{T}$, we have

$$p(y_t|y_{1:t-1}) = \int_{\mathbb{T}} p(y_t|y_{1:t-1},\theta)\, p(\theta|y_{1:t-1}).\, d\theta \quad \text{(B.3)}$$

Thus, under Assumption A10, applying (B.2) and (B.3) to each term of the sum in (1.3) yields

$$\mathcal{H}_T(M) = \sum_{t=1}^{T} \sum_{k=1}^{d_y} \left(2\,\mathbb{E}\left[\frac{\partial^2 \log p(y_t|y_{1:t-1},\Theta)}{\partial y_{t_{(k)}}^2} + \left(\frac{\partial \log p(y_t|y_{1:t-1},\Theta)}{\partial y_{t_{(k)}}}\right)^2 \middle| y_{1:t} \right] - \left(\mathbb{E}\left[\frac{\partial \log p(y_t|y_{1:t-1},\Theta)}{\partial y_{t_{(k)}}} \middle| y_{1:t} \right] \right)^2 \right)$$

which proves (1.4). □

B.4.2 Proof of (1.11) and (1.12)

Under Assumption A11, we can partially differentiate under the integral sign, so that for all $k \in [\![1, d_y]\!]$, we have

$$\begin{aligned}
\frac{\partial \log p(y_t|y_{1:t-1},\theta)}{\partial y_{t_{(k)}}} &= \frac{1}{p(y_t|y_{1:t-1},\theta)} \int p(x_t|y_{1:t-1},\theta) \left(\frac{\partial\, g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}} \right) dx_t \\
&= \frac{1}{p(y_t|y_{1:t-1},\theta)} \int p(x_t|y_{1:t-1},\theta)\, g_\theta(y_t|x_t) \left(\frac{\partial \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}} \right) dx_t \\
&= \int \left(\frac{\partial \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}} \right) p(x_t|y_{1:t},\theta)\, dx_t\,,
\end{aligned}$$

where the last equality comes from the fact that

$$\frac{p(x_t|y_{1:t-1},\theta)\, g_\theta(y_t|x_t)}{p(y_t|y_{1:t-1},\theta)} = \frac{p(x_t,y_t|y_{1:t-1},\theta)}{p(y_t|y_{1:t-1},\theta)} = p(x_t|y_{1:t},\theta). \quad \text{(B.4)}$$

This proves (1.11).

Regarding (1.12), we proceed similarly and have, for all $k \in [\![1, d_y]\!]$,

$$\frac{\partial^2 \log p(y_t|y_{1:t-1},\theta)}{\partial y_{t_{(k)}}^2} = -\left(\frac{\partial \log p(y_t|y_{1:t-1},\theta)}{\partial y_{t_{(k)}}} \right)^2 + \frac{1}{p(y_t|y_{1:t-1},\theta)} \frac{\partial^2 p(y_t|y_{1:t-1},\theta)}{\partial y_{t_{(k)}}^2}. \quad \text{(B.5)}$$

The second term can be rewritten as

$$\frac{1}{p(y_t|y_{1:t-1},\theta)}\frac{\partial^2 p(y_t|y_{1:t-1},\theta)}{\partial y_{t(k)}^2} = \frac{1}{p(y_t|y_{1:t-1},\theta)}\int p(x_t|y_{1:t-1},\theta)\left(\frac{\partial^2 g_\theta(y_t|x_t)}{\partial y_{t(k)}^2}\right)dx_t\,, \quad \text{(B.6)}$$

where the integrand can be written as

$$\frac{\partial^2 g_\theta(y_t|x_t)}{\partial y_{t(k)}^2} = g_\theta(y_t|x_t)\left(\frac{\partial^2 \log g_\theta(y_t|x_t)}{\partial y_{t(k)}^2} + \left(\frac{\partial \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}}\right)^2\right). \quad \text{(B.7)}$$

By plugging (B.7) into (B.6) and using again (B.4), we get

$$\frac{1}{p(y_t|y_{1:t-1},\theta)}\frac{\partial^2 p(y_t|y_{1:t-1},\theta)}{\partial y_{t(k)}^2} = \int p(x_t|y_{1:t},\theta)\left(\frac{\partial^2 \log g_\theta(y_t|x_t)}{\partial y_{t(k)}^2} + \left(\frac{\partial \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}}\right)^2\right)dx_t\,.$$

By plugging this back into (B.5), we finally get

$$\frac{\partial^2 \log p(y_t|y_{1:t-1},\theta)}{\partial y_{t(k)}^2} = -\left(\frac{\partial \log p(y_t|y_{1:t-1},\theta)}{\partial y_{t(k)}}\right)^2 + \int\left(\frac{\partial^2 \log g_\theta(y_t|x_t)}{\partial y_{t(k)}^2} + \left(\frac{\partial \log g_\theta(y_t|x_t)}{\partial y_{t(k)}}\right)^2\right)p(x_t|y_{1:t},\theta)\,dx_t\,,$$

which proves (1.12). □

B.5 Consistency of the H-score

Without much loss of generality, we prove the results in the case of continuous univariate observations ($d_y = 1$). Thanks to (1.3), the proofs can be generalized to multivariate observations by working on each dimension separately. Unless stated otherwise, we assume that A10 and A11 hold, so that we may use (1.13) and (1.5).

Section B.5.1 should be read as a proof of concept: we prove Theorem 1 and 2 by using intermediary results as high-level assumptions (A12 to A21). This allows us to highlight the key steps of the proofs. In Section B.5.2, we present explicit conditions (C6 to C11) that are sufficient for these assumptions to hold. Some of these conditions are strong, which enables intuitive proofs; our simulation studies suggest that the consistency of the H-score is likely to hold under weaker conditions. Detailed proofs of all the intermediary results are provided in Section B.5.3. Proofs under weaker conditions or discrete observations are left for future work.

B.5.1 Proofs of Theorem 1 and 2

The first ingredient is the $\mathbb{P}_\star$-almost sure concentration of the posterior distribution $p(d\theta|Y_{1:t})$ around some limit value $\theta^\star \in \mathbb{T}$, as the number of observations increases (Assumption A12).

Assumption A12. *$\mathbb{P}_\star$-almost surely, there exists $\theta^\star \in \mathbb{T}$ such that, if $\Theta_t \sim p(d\theta|Y_{1:t})$ for all $t \in \mathbb{N}^*$, then $\Theta_t \xrightarrow[t\to+\infty]{\mathcal{D}} \theta^\star$.*

Posterior concentration in i.i.d. settings can be formally enforced by explicit regularity conditions (e.g. Condition C6 in Section B.5.2.1). In the case of state-space models with dependent observations, we treat posterior concentration as a working assumption. From now on, we assume that Assumption A12 holds, so that we can unambiguously refer to the limit point $\theta^\star$ around which the posterior distribution concentrates.

In addition to concentration of the posterior distribution, we also want the posterior moments of specific test functions to converge, $\mathbb{P}_\star$-almost surely. In particular, as the posterior distribution concentrates to a point mass, we want the posterior expectations and variances appearing in (1.5) to respectively converge to a finite limit and to 0, as the number of observations increases (Assumption A13).

Assumption A13. *The following limits hold:*

$$(a) \quad \mathbb{E}\left[\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \Theta)\right) \,\middle|\, Y_{1:t}\right] - \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right) \xrightarrow[t\to+\infty]{\mathbb{P}_\star - a.s.} 0.$$

$$(b) \quad \mathrm{Var}\left(\frac{\partial \log p(Y_t|Y_{1:t-1}, \Theta)}{\partial y_t} \,\middle|\, Y_{1:t}\right) \xrightarrow[t\to+\infty]{\mathbb{P}_\star - a.s.} 0.$$

By Stolz-Cesàro's theorem, the $\mathbb{P}_\star$-a.s. convergence of the posterior moments in Assumption A13 implies the $\mathbb{P}_\star$-a.s. convergence of their Cesàro means. This leads to the convergence of the prequential quantities, so that

$$\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \Theta)\right) \,\middle|\, Y_{1:t}\right]\right) - \left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} 0 \quad \text{(B.8)}$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{Var}\left(\frac{\partial \log p(Y_t|Y_{1:t-1},\Theta)}{\partial y_t}\middle| Y_{1:t}\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star-a.s.} 0\,. \tag{B.9}$$

At this stage, the proof starts to differ depending on which setting we consider.

B.5.1.1 Models for i.i.d. data

For i.i.d. models, we have $\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1},\theta^\star)\right) = \mathcal{H}\left(Y_t, p(dy|\theta^\star)\right)$ for all $t \in \mathbb{N}^*$. If the Y_t's are generated as i.i.d. from $p_\star$ (Assumption A14), then the integrability of $\mathcal{H}\left(Y, p(dy|\theta^\star)\right)$ with respect to $Y \sim p_\star$ (Assumption A15) enables the application of the law of large numbers to the quantity $T^{-1}\sum_{t=1}^{T}\mathcal{H}\left(Y_t, p(dy|\theta^\star)\right)$.

Assumption A14. *The observations $(Y_t)_{t\in\mathbb{N}^*}$ are i.i.d. draws from $p_\star$.*

Assumption A15. *The H-score of $p(dy|\theta^\star)$ is integrable: $\mathbb{E}_\star\left[\left|\mathcal{H}\left(Y, p(dy|\theta^\star)\right)\right|\right] < +\infty$.*

Under Assumptions A14 and A15, the law of large numbers reduces (B.8) to

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathcal{H}\left(Y_t, p(dy|\Theta)\right)\,\middle|\, Y_{1:t}\right] \xrightarrow[T\to+\infty]{\mathbb{P}_\star-a.s.} \mathbb{E}_\star\left[\mathcal{H}\left(Y, p(dy|\theta^\star)\right)\right], \tag{B.10}$$

where the expectation is taken with respect to $Y \sim p_\star$. If M_1 and M_2 are both i.i.d. models satisfying A10, A12, A13, and A15, then combining (1.5), (B.9) and (B.10) leads to

$$\frac{1}{T}\mathcal{H}_T(M_j) \xrightarrow[T\to+\infty]{\mathbb{P}_\star-a.s.} \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_j(dy|\theta_j^\star)\right)\right],$$

for each $j \in \{1,2\}$. Taking the difference of the respective scores yields

$$\frac{1}{T}\left(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star-a.s.} \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_2(dy|\theta_2^\star)\right)\right] - \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_1(dy|\theta_1^\star)\right)\right], \tag{B.11}$$

In order to interpret the consistency of the H-score in terms of an appropriate divergence, we impose further regularity assumptions on the models and the data-generating process itself (Assumption A16). Assumption A16(a) allows us to define the H-score of $p_\star$, assumed to be integrable by A16(b). Assumption A16(c) ensures the strict propriety of the H-score.

Assumption A16. *The data-generating process and the model satisfy the following conditions:*

(a) $y \mapsto p_\star(y)$ *is twice differentiable.*

(b) $\mathbb{E}_\star\left[\left|\mathcal{H}\left(Y, p_\star(dy)\right)\right|\right] < +\infty$.

(c) $\dfrac{\partial \log p(y|\theta^\star)}{\partial y} p_\star(y) \xrightarrow[|y|\to+\infty]{} 0$.

Under Assumptions A12, A15, and A16, we can define the divergence $\mathrm{D}_{\mathcal{H}}(p_\star, M_j)$ as in (1.7). By adding and subtracting $\mathbb{E}_\star\left[\mathcal{H}\left(Y, p_\star(dy)\right)\right]$ in (B.11), we get

$$\frac{1}{T}\left(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \mathrm{D}_{\mathcal{H}}(p_\star, M_2) - \mathrm{D}_{\mathcal{H}}(p_\star, M_1).$$

Under A16(c), integration by parts (Hyvärinen, 2005; Dawid and Musio, 2015) leads to

$$\mathrm{D}_{\mathcal{H}}(p_\star, M_j) = \int \left(\frac{\partial \log p_\star(y)}{\partial y} - \frac{\partial \log p_j(y|\theta_j^\star)}{\partial y}\right)^2 p_\star(y) dy.$$

Therefore, we have $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) \geq 0$.

If $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) = 0$, then $\partial \log p_\star(y)/\partial y = \partial \log p_j(y|\theta_j^\star)/\partial y$ for all $y \in \mathbb{Y}$. Hence, $\log p_\star(y) = \log p_j(y|\theta_j^\star) + \log(c)$ for all $y \in \mathbb{Y}$ and some constant $c > 0$. This leads to $p_\star(y) = c\, p_j(y|\theta_j^\star)$ for all $y \in \mathbb{Y}$. Since probability densities integrate to 1, we necessarily have $c = 1$, i.e. $p_\star(y) = p_j(y|\theta_j^\star)$ for all $y \in \mathbb{Y}$. This concludes the proof of Theorem 1.

B.5.1.2 State-space models

In the case of state-space models and dependent observations, more subtle arguments are needed since we can no longer apply the standard law of large numbers to the term $T^{-1}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}\theta^\star)\right)$ in (B.8). Instead, we approximate this term by a stationary analog, to which ergodic theorems will apply.

To this end, we assume the process $(Y_t)_{t\in\mathbb{N}^*}$ is strongly stationary and ergodic (Assumption A17). Under strong stationarity, we can artificially extend the index set to negative integers

and consider the two-sided process $(Y_t)_{t\in\mathbb{Z}}$. We also need the dependence of the H-score on the initial distribution of the latent Markov chain to vanish quickly enough. This will be referred to as the *forgetting property* of the H-score (Assumption A18).

Assumption A17. *The process $(Y_t)_{t\in\mathbb{N}^*}$ is strongly stationary and ergodic.*

Assumption A18. *There exist $\rho \in (0,1)$ and $\gamma > 0$ such that, for all $t \in \mathbb{N}^*$, all $m \in \mathbb{N}$, and all $y_{-m:t} \in \mathbb{Y}^{m+t+1}$,*

$$|\mathcal{H}(y_t, p(dy_t|y_{-m+1:t-1}, \theta^\star)) - \mathcal{H}(y_t, p(dy_t|y_{-m:t-1}, \theta^\star))| \leq \gamma\, \rho^{t+m-1}.$$

Under Assumptions A17 and A18, we can prove that, $\mathbb{P}_\star$-a.s., for all $t \in \mathbb{N}^*$, the sequence $(\mathcal{H}(Y_t, p(dy_t|Y_{-m+1:t-1}, \theta^\star)))_{m\in\mathbb{N}}$ is a real-valued Cauchy sequence, and thus converges to a $\mathbb{P}_\star$-a.s. limit denoted by $\mathcal{H}(Y_t, p(dy_t|Y_{-\infty:t-1}, \theta^\star))$. In other words, $\mathbb{P}_\star$-almost surely, and for all $t \in \mathbb{N}^*$, we have

$$\mathcal{H}(Y_t, p(dy_t|Y_{-m+1:t-1}, \theta^\star)) \xrightarrow[m\to+\infty]{} \mathcal{H}(Y_t, p(dy_t|Y_{-\infty:t-1}, \theta^\star))\,. \tag{B.12}$$

Using (B.12) and the forgetting property in A18, we can prove that

$$\left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right)\right) - \left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{-\infty:t-1}, \theta^\star)\right)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star-a.s.} 0\,. \tag{B.13}$$

The proofs of (B.12) and (B.13) are provided in Section B.5.3. Equation (B.13) implies that, $\mathbb{P}_\star$-almost surely, the term $T^{-1}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}\theta^\star)\right)$ in (B.8) can be asymptotically approximated by the stationary quantity $T^{-1}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{-\infty:t-1}, \theta^\star)\right)$, to which ergodic theorems can be applied under adequate integrability conditions (Assumption A19).

Assumption A19. *The limit in (B.12) is integrable:* $\mathbb{E}_\star\left[\left|\mathcal{H}\left(Y_1, p(dy_1|Y_{-\infty:0}, \theta^\star)\right)\right|\right] < +\infty$.

Under Assumption A19, Birkhoff's ergodic theorem after combining (B.8) and (B.13) yields

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \Theta)\right) \,\middle|\, Y_{1:t}\right] \xrightarrow[T\to+\infty]{\mathbb{P}_\star-a.s.} \mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p(dy_1|Y_{-\infty:0}, \theta^\star)\right)\right]. \tag{B.14}$$

Under Assumption A17, and assuming both models M_1 and M_2 satisfy A10, A12, A13, A18, and A19, we can piece together (1.5), (B.9), and (B.14). This leads to

$$\frac{1}{T}\left(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p_2(dy_1|Y_{-\infty:0}, \theta_2^\star)\right)\right] - \mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p_1(dy_1|Y_{-\infty:0}, \theta_1^\star)\right)\right]. \quad \text{(B.15)}$$

In order to interpret this consistency result in terms of a divergence, we need the quantity $\mathcal{H}\left(Y_1, p_j(dy_1|Y_{-\infty:0}, \theta_j^\star)\right)$ to correspond to the actual H-score at Y_1 of a twice differentiable probability density function $y_1 \mapsto p_j(y_1|Y_{-\infty:0}, \theta_j^\star)$, conditional on $Y_{-\infty:0}$, $\mathbb{P}_\star$-almost surely (Assumption A20).

Assumption A20. *$\mathbb{P}_\star$-almost surely, we can define the conditional density $p(y_1|Y_{-\infty:0}, \theta^\star)$ of Y_1 given $Y_{-\infty:0}$, and the limit $\mathcal{H}(Y_1, p(dy_1|Y_{-\infty:0}, \theta^\star))$ in (B.12) corresponds to the actual H-score at Y_1 of $p(dy_1|Y_{-\infty:0}, \theta^\star)$.*

In order to define the divergence $\mathrm{D}_{\mathcal{H}}$, we make further regularity assumptions on the models and the data-generating process (Assumption A21). Similarly to Assumption A16 in the i.i.d. setting, A21(a) allows us to define the H-score of $p_\star(dy_1|Y_{-\infty:0})$, conditional on $Y_{-\infty:0}$, $\mathbb{P}_\star$-almost surely, while A21(b) enforces its integrability. Assumption A21(c) ensures the strict propriety of the H-score, conditional on $Y_{-\infty:0}$, $\mathbb{P}_\star$-almost surely.

Assumption A21. *The data-generating process and the model satisfy the following:*

(a) *The conditional density $y_1 \mapsto p_\star(y_1|Y_{-\infty:0})$ of Y_1 given $Y_{-\infty:0}$ is well-defined and twice differentiable on $\mathbb{Y}$.*

(b) $\mathbb{E}_\star\left[\left|\mathcal{H}\left(Y_1, p_\star(dy_1|Y_{-\infty:0})\right)\right|\right] < +\infty$.

(c) $\dfrac{\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1} p_\star(y_1|Y_{-\infty:0}) \xrightarrow[|y_1|\to+\infty]{\mathbb{P}_\star - a.s.} 0.$

Under Assumptions A12, A19, and A21, we can define the divergence $\mathrm{D}_{\mathcal{H}}(p_\star, M_j)$ as in (1.14). By adding and subtracting $\mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p_\star(dy_1|Y_{-\infty:0})\right)\right]$ in (B.15), we get

$$\frac{1}{T}\left(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \mathrm{D}_{\mathcal{H}}(p_\star, M_2) - \mathrm{D}_{\mathcal{H}}(p_\star, M_1).$$

The tower property of conditional expectations, combined with an integration by parts using A21(c), leads to

$$\mathrm{D}_{\mathcal{H}}(p_\star, M_j) = \mathbb{E}_\star\left[\int\left(\frac{\partial \log p_\star(y_1|Y_{-\infty:0})}{\partial y_1} - \frac{\partial \log p_j(y_1|Y_{-\infty:0}, \theta_j^\star)}{\partial y_1}\right)^2 p_\star(y_1|Y_{-\infty:0})dy_1\right].$$

Thus, we have $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) \geq 0$. If $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) = 0$, the same reasoning as in the proof of (1.6) shows that, $\mathbb{P}_\star$-almost surely, we have $p_j(y_1|Y_{-\infty:0}, \theta_j^\star) = p_\star(y_1|Y_{-\infty:0})$ for all $y_1 \in \mathbb{Y}$. This concludes the proof of Theorem 2.

B.5.2 Sufficient conditions and intermediary results

B.5.2.1 Assumption A12: Concentration of the posterior distribution

For a generic i.i.d. model $\{p(dy|\theta) : \theta \in \mathbb{T}\}$ with prior distribution $p(d\theta)$, if the observations are assumed to be i.i.d. from $p_\star$, then Theorem 1.3.4. in Ghosh and Ramamoorthi (2003) shows that the following set of regularity conditions (Condition C6) ensures the concentration of the posterior. In other words, for i.i.d. models and data, Condition C6 and Assumption A14 guarantee Assumption A12.

Condition C6. *The model satisfies the following conditions:*

(a) $\mathbb{T}$ *is a compact metric space, and* $p(\theta) > 0$ *for all* $\theta \in \mathbb{T}$.

(b) $y \mapsto p(y|\theta)$ *is measurable for all* $\theta \in \mathbb{T}$, *and* $\theta \mapsto p(y|\theta)$ *is continuous for all* $y \in \mathbb{Y}$.

(c) $\int_{\mathbb{Y}} \sup_{\theta\in\mathbb{T}} |\log p(y|\theta)|\, p_\star(y)dy < +\infty$.

Condition C6 can be relaxed to allow for semi-continuity and non-compact parameter spaces, as discussed in Remark 1.3.5 of Ghosh and Ramamoorthi (2003) and its references (e.g. Wald, 1949; Le Cam, 1953; Kiefer and Wolfowitz, 1956; Huber, 1967; Perlman, 1972).

Posterior concentration for general state-space models with dependent data is less standard, especially when allowing for misspecification. Some concentration results have been proved in specific cases (e.g. Lijoi et al., 2007; De Gunst and Shcherbakova, 2008; Shalizi, 2009; Gassiat and Rousseau, 2014; Douc et al., 2019, and references therein). However, as far as we know, a formal proof of posterior concentration with explicit conditions on possibly misspecified state-space models has yet to be derived.

B.5.2.2 Assumption A13: Convergence of specific posterior moments

Concentration of the posterior distribution does not guarantee convergence of any posterior moments. The latter can be ensured by further imposing equicontinuity (Condition C7) and uniform integrability (Condition C8).

Condition C7. *$\mathbb{P}_\star$-almost surely, the following statements hold:*

(a) $\{\theta \mapsto \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta)\right) : t \in \mathbb{N}^*\}$ *is equicontinuous at* $\theta^\star$.

(b) $\left\{\theta \mapsto \dfrac{\partial \log p(Y_t|Y_{1:t-1}, \theta)}{\partial y_t} : t \in \mathbb{N}^*\right\}$ *is equicontinuous at* $\theta^\star$.

Equicontinuity at $\theta^\star$ of a family of functions $\{\theta \longmapsto h_t(\theta) : t \in \mathbb{N}^*\}$ means that all the functions in the family share a common (i.e. not depending on t) modulus of continuity at $\theta^\star$. Equicontinuity can be enforced by the stronger but more explicit condition that there exists a neighborhood $\mathcal{U}_{\theta^\star}$ of $\theta^\star$, on which the functions $\theta \longmapsto h_t(\theta)$ are differentiable for all $t \in \mathbb{N}^*$, and such that $\sup_{(t,\theta)\in\mathbb{N}^*\times\mathcal{U}_{\theta^\star}} \|\nabla_\theta h_t(\theta)\| = L < +\infty$. Indeed, by the mean value theorem, such uniform boundedness of the gradients ensures that the functions $\theta \longmapsto h_t(\theta)$ are L-Lipschitz on $\mathcal{U}_{\theta^\star}$ for all $t \in \mathbb{N}^*$, where L does not depend on t. Then, for any arbitrary $\varepsilon > 0$, we can find $\delta_\varepsilon > 0$ not depending on t (e.g. $\delta_\varepsilon = \varepsilon/L$ if $L > 0$, or else any $\delta_\varepsilon > 0$ if

$L = 0$) such that, for all $\theta \in \mathcal{U}_{\theta^\star}$, $\|\theta - \theta^\star\| < \delta_\varepsilon$ implies $\sup_{t\in\mathbb{N}^*} \|h_t(\theta) - h_t(\theta^\star)\| < \varepsilon$, which proves the equicontinuity at $\theta^\star$ of the family $\{\theta \longmapsto h_t(\theta) : t \in \mathbb{N}^*\}$.

Condition C8. *$\mathbb{P}_\star$-almost surely, if $\Theta_t \sim p(d\theta|Y_{1:t})$ for all $t \in \mathbb{N}^*$, then the following statements hold:*

(a) $\{\mathcal{H}(Y_t, p(dy_t|Y_{1:t-1}, \Theta_t)) : t \in \mathbb{N}^*\}$ *is uniformly integrable given* $(Y_t)_{t\in\mathbb{N}^*}$.

(b) $\left\{\left(\frac{\partial \log p(Y_t|Y_{1:t-1}, \Theta_t)}{\partial y_t}\right)^2 : t \in \mathbb{N}^*\right\}$ *is uniformly integrable given* $(Y_t)_{t\in\mathbb{N}^*}$.

Uniform integrability of a family of random variables $\{H_t : t \in \mathbb{N}^*\}$ can be enforced by the stronger but more explicit condition of L_α-boundedness: if there exists $\alpha > 1$ such that $\sup_{t\in\mathbb{N}^*} \mathbb{E}[\,|H_t|^\alpha\,] < +\infty$, then $\{H_t : t \in \mathbb{N}^*\}$ is uniformly integrable (e.g. see Theorem 25.12 and its corollary in Billingsley, 1995).

Convergence of the relevant posterior moments in Assumption A13 can be obtained as a consequence of Assumption A12 combined with Conditions C7 and C8. This is summarized by the following lemma.

Lemma 22. *Assume A12, C7, and C8. Then Assumption A13 holds and we have:*

(a) $\mathbb{E}\left[\mathcal{H}(Y_t, p(dy_t|Y_{1:t-1}, \Theta)) \,\middle|\, Y_{1:t}\right] - \mathcal{H}(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)) \xrightarrow[t\to+\infty]{\mathbb{P}_\star - a.s.} 0.$

(b) $\mathrm{Var}\left(\frac{\partial \log p(Y_t|Y_{1:t-1}, \Theta)}{\partial y_t}\,\middle|\, Y_{1:t}\right) \xrightarrow[t\to+\infty]{\mathbb{P}_\star - a.s.} 0.$

The proof of Lemma 22 is provided in Section B.5.3.3.

B.5.2.3 Assumption A18: Forgetting property of the H-score

For state-space models, the forgetting property of the H-score can be obtained as a consequence of the forgetting property of the latent Markov chain stated in (B.16) (following from Condition C9) and some appropriate boundedness conditions on the first two derivatives of the observation log-density (Condition C10). Condition C9 corresponds to a simplified version of Assumption A13.1 in Douc et al. (2014).

Condition C9. *The model satisfies the following conditions:*

(a) There exists a dominating probability measure η *on* $\mathbb{X}$ *such that the transition kernel* $f_{\theta^\star}(dx_{t+1}|x_t)$ *has density* $\nu_{\theta^\star}(x_{t+1}|x_t) = (df_{\theta^\star}(\cdot|x_t)/d\eta)(x_{t+1})$ *with respect to* η.

(b) There exist positive constants σ^- *and* σ^+ *such that, for all* $(x_t, x_{t+1}) \in \mathbb{X} \times \mathbb{X}$, *the transition density* $\nu_{\theta^\star}(x_{t+1}|x_t)$ *satisfies* $0 < \sigma^- < \nu_{\theta^\star}(x_{t+1}|x_t) < \sigma^+ < +\infty$.

(c) For all $y_t \in \mathbb{Y}$, *the integral* $\int_{\mathbb{X}} g_{\theta^\star}(y_t, x_t)\, \eta(dx_t)$ *is bounded away from* 0 *and* $+\infty$.

Under strong stationarity of the process $(Y_t)_{t\in\mathbb{N}^*}$, Lemma 13.2 in Douc et al. (2014) guarantees that for all $t \in \mathbb{N}^*$, all $m \in \mathbb{N}$, and all realizations $y_{-m:t} \in \mathbb{Y}^{m+t+1}$, the filtering distributions of the latent states satisfy

$$d_{TV}\left(p(dx_t|y_{-m+1:t}, \theta^\star), p(dx_t|y_{-m:t}, \theta^\star)\right) \leq \rho^{t+m-1}, \tag{B.16}$$

where d_{TV} stands for the total variation distance and $\rho = 1 - (\sigma^-/\sigma^+) \in (0, 1)$. Condition C9 would typically require the latent space $\mathbb{X}$ to be finite or compact, and ensures that the transition kernel is geometrically ergodic. Such a condition can generally be weakened to allow for non-finite and non-compact spaces (e.g. Douc et al., 2011; Douc and Moulines, 2012).

When (1.11) and (1.12) hold, H-scores for a fixed $\theta^\star$ can be written in terms of expectations with respect to the corresponding filtering distributions. Differences of H-scores can then be related to total variation distance between filtering distributions by assuming the integrands in (1.11) and (1.12) are bounded (Condition C10).

Condition C10. *The model satisfy the following domination conditions:*

(a) $b = \sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} \left| \frac{\partial^2 \log g_{\theta^\star}(y|x)}{\partial y^2} + \left(\frac{\partial \log g_{\theta^\star}(y|x)}{\partial y} \right)^2 \right| < +\infty.$

(b) $c = \sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} \left| \frac{\partial \log g_{\theta^\star}(y|x)}{\partial y} \right| < +\infty.$

Condition C10 could be enforced by the stronger conditions that $\mathbb{X}$ and $\mathbb{Y}$ are compact, and the first two derivatives of $y \mapsto \log g_{\theta^\star}(y|x)$ are continuous with respect to (x, y). Compactness conditions may look quite restrictive, since most well-known continuous distributions have non compact supports. In reality, for all practical purposes, we could always envision a sufficiently large compact space in which all our numerical values would lie. As stated earlier, Conditions C9 and C10 should only be regarded as mere sufficient conditions that allow for more straightforward proofs. The models of our simulation studies do not satisfy these conditions, and yet we observe the consistency of H-scores. This indicates that consistency is likely to hold under weaker conditions.

Under Assumptions A11, A12, and A17, Conditions C9 and C10 combined with (1.11)-(1.12) guarantee that the forgetting property of the H-score in Assumption A18 holds, as stated by the following lemma.

Lemma 23. *Assume A11, A12, A17, C9, and C10. Then, for all $t \in \mathbb{N}^*$, all $m \in \mathbb{N}$, and all $y_{-m:t} \in \mathbb{Y}^{m+t+1}$,*

$$|\mathcal{H}(y_t, p(dy_t|y_{-m+1:t-1}, \theta^\star)) - \mathcal{H}(y_t, p(dy_t|y_{-m:t-1}, \theta^\star))| \leq 2(b + c^2)\, \rho^{t+m-1}, \tag{B.17}$$

$$\sup_{m \in \mathbb{N}} |\mathcal{H}(y_t, p(dy_t|y_{-m+1:t-1}, \theta^\star))| \leq 2\, b + c^2, \tag{B.18}$$

where $\rho = 1 - \frac{\sigma^-}{\sigma^+} \in (0, 1)$.

Equation (B.17) in Lemma 23 enforces Assumption A18 with $\gamma = 2(b+c^2)$, while (B.18) directly buys us Assumptions A15 and A19. The proof of Lemma 23 is provided in Section B.5.3.

B.5.2.4 Assumption A20: H-score of conditional density given the infinite past

Ensuring that we may define $y_1 \mapsto p(y_1|Y_{-\infty:0}, \theta^\star)$ as an actual probability density function can be done under further domination and integrability conditions on the observation density (Condition C11).

Condition C11. *Let $\nu_{\theta^\star}$ be the probability measure from Condition C9. Then we have:*

(a) $\sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} g_{\theta^\star}(y|x) < +\infty$.

(b) $\mathbb{E}_\star\left[\left|\log\left(\int_{\mathbb{X}} g_{\theta^\star}(Y_1|x)\nu_{\theta^\star}(dx)\right)\right|\right] < +\infty$.

Condition C11 corresponds to a simplified statement of Assumption A13.3 in Douc et al. (2014). Under Assumption A17 with Conditions C9 and C11, Lemma 13.12 and Proposition 13.5 from Douc et al. (2014) show that $y_1 \mapsto \log p(y_1|Y_{-\infty:0}, \theta^\star) = \lim_{m\to+\infty} \log p(y_1|Y_{-m+1:0}, \theta^\star)$ exists and defines an actual log-density, $\mathbb{P}_\star$-almost surely. The $\mathbb{P}_\star$-almost sure twice differentiability of $y_1 \mapsto \log p(y_1|Y_{-\infty:0}, \theta^\star)$ follows from the uniform convergence of the first two derivatives of $y_1 \mapsto \log p(y_1|Y_{-m+1:0}, \theta^\star)$ as $m \to +\infty$ (e.g. Theorem 7.17 from Rudin, 1964), which can be proved using (1.11)-(1.12) and the domination conditions from C10. In other words, under Assumptions A11 and A17, Conditions C9 to C11 ensure that Assumption A20 holds. This is stated by the following lemma.

Lemma 24. *Assume A11, A17, C9, C10, and C11. Then, $\mathbb{P}_\star$-almost surely, there exists a continuous probability density function $x_1 \mapsto p(x_1|Y_{-\infty:0}, \theta^\star) = \lim_{m\to+\infty} p(x_1|Y_{-m+1:0}, \theta^\star)$ with respect to $\nu_{\theta^\star}$. Define the function*

$$y_1 \mapsto p(y_1|Y_{-\infty:0}, \theta^\star) = \int g_{\theta^\star}(y_1|x_1)p(x_1|Y_{-\infty:0}, \theta^\star)\nu_{\theta^\star}(dx_1).$$

Then, $\mathbb{P}_\star$-almost surely, $p(Y_1|Y_{-\infty:0}, \theta^\star) = p(y_1|Y_{-\infty:0}, \theta^\star)|_{y_1=Y_1}$, and $y_1 \mapsto p(y_1|Y_{-\infty:0}, \theta^\star)$ is the conditional density with respect to the Lebesgue measure of Y_1 given the σ-algebra generated by $(Y_{-m})_{m\in\mathbb{N}}$ under $\mathbb{P}_\star$. Moreover, $\mathbb{P}_\star$-almost surely, the limit function

$$y_1 \mapsto \log p(y_1|Y_{-\infty:0}, \theta^\star) = \lim_{m\to+\infty} \log p(y_1|Y_{-m+1:0}, \theta^\star)$$

exists and is twice differentiable on $\mathbb{Y}$, with

$$\frac{\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1} = \lim_{m\to+\infty} \frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1},$$

$$\frac{\partial^2 \log p(y_1|Y_{-\infty:0},\theta^\star)}{\partial y_1^2} = \lim_{m\to+\infty} \frac{\partial^2 \log p(y_1|Y_{-m+1:0},\theta^\star)}{\partial y_1^2},$$

and

$$\mathcal{H}(y_1, p(dy_1|Y_{-\infty:0},\theta^\star)) = 2\,\frac{\partial^2 \log p(y_1|Y_{-\infty:0},\theta^\star)}{\partial y_1^2} + \left(\frac{\partial^2 \log p(y_1|Y_{-\infty:0},\theta^\star)}{\partial y_1^2}\right)^2$$

for all $y_1 \in \mathbb{Y}$.

The proof of Lemma 24 is provided in Section B.5.3.

B.5.3 Proofs of intermediary results

B.5.3.1 Proof of (B.12)

Fix some arbitrary $\varepsilon > 0$ and $t \in \mathbb{N}^*$. Since $\rho \in (0,1)$, we have $\rho^N \to 0$ as $N \to +\infty$, so there exists some $N \in \mathbb{N}$ large enough such that $\gamma\,\rho^{t+N}(1-\rho)^{-1} < \varepsilon$. Using Assumption A18, we get, $\mathbb{P}_\star$-almost surely, for any $n > m > N$,

$$\begin{aligned}
&|\mathcal{H}(Y_t, p(dy_t|Y_{-m+1:t-1},\theta^\star)) - \mathcal{H}(Y_t, p(dy_t|Y_{-n+1:t-1},\theta^\star))| \\
&\leq \sum_{k=m}^{n-1} |\mathcal{H}(Y_t, p(dy_t|Y_{-k+1:t-1},\theta^\star)) - \mathcal{H}(Y_t, p(dy_t|Y_{-k:t-1},\theta^\star))| \\
&\leq \gamma\,\rho^{t-1} \sum_{k=m}^{n-1} \rho^k \\
&\leq \gamma\,\rho^{t-1} \sum_{k=N+1}^{+\infty} \rho^k \\
&\leq \varepsilon.
\end{aligned}$$

Therefore $(\mathcal{H}(Y_t, p(dy_t|Y_{-m+1:t-1},\theta^\star))_{m\in\mathbb{N}}$ is a Cauchy sequence for every $t \in \mathbb{N}^*$, $\mathbb{P}_\star$-almost surely. Since $\mathbb{R}$ is complete, this sequence converges $\mathbb{P}_\star$-almost surely to a limit, denoted by

$$\mathcal{H}(Y_t, p(dy_t|Y_{-m+1:t-1},\theta^\star)) \xrightarrow[m\to+\infty]{\mathbb{P}_\star - a.s.} \mathcal{H}(Y_t, p(dy_t|Y_{-\infty:t-1},\theta^\star)).$$

□

B.5.3.2 Proof of (B.13)

We have, $\mathbb{P}_\star$-almost surely, for every $T \in \mathbb{N}^*$,

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1},\theta^\star)\right) - \mathcal{H}\left(Y_t, p(dy_t|Y_{-\infty:t-1},\theta^\star)\right)\right)\right| \\
\leq\ & \frac{1}{T}\sum_{t=1}^{T}\sum_{m=0}^{+\infty}\left|\mathcal{H}\left(Y_t, p(dy_t|Y_{-m+1:t-1},\theta^\star)\right) - \mathcal{H}\left(Y_t, p(dy_t|Y_{-m:t-1},\theta^\star)\right)\right| \\
\leq\ & \frac{\gamma}{T}\sum_{t=1}^{T}\sum_{m=0}^{+\infty}\rho^{t+m-1},
\end{aligned}$$

where $\rho \in (0,1)$ and $\gamma > 0$ are given by Assumption A18. Properties of geometric series lead to

$$\begin{aligned}
&\left|\left(\frac{1}{T}\sum_{t=1}^{T}\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1},\theta^\star)\right)\right) - \left(\frac{1}{T}\sum_{t=1}^{T}\mathcal{H}\left(Y_t, p(dy_t|Y_{-\infty:t-1},\theta^\star)\right)\right)\right| \\
\leq\ & \frac{\gamma}{T}\sum_{t=1}^{+\infty}\rho^{t-1}\sum_{m=0}^{+\infty}\rho^{m} \\
\leq\ & \frac{\gamma}{T(1-\rho)^2}\,.
\end{aligned}$$

The upper bound goes to 0 as $T \to +\infty$, therefore

$$\left(\frac{1}{T}\sum_{t=1}^{T}\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1},\theta^\star)\right)\right) - \left(\frac{1}{T}\sum_{t=1}^{T}\mathcal{H}\left(Y_t, p(dy_t|Y_{-\infty:t-1},\theta^\star)\right)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} 0.$$

□

B.5.3.3 Proof of Lemma 22

Any finite intersection of almost sure events is an almost sure event, thus we can find a common event A such that $\mathbb{P}_\star(A) = 1$, and on which all the assumptions and conditions hold simultaneously. Fix some arbitrary $\omega \in A$. For all $t \in \mathbb{N}^*$, define $y_t = Y_t(\omega)$ and let $\Theta_t \sim p(d\theta|y_{1:t})$. By Assumption A12, we have $\Theta_t \xrightarrow[t\to+\infty]{\mathcal{D}} \theta^\star$. The space $\mathbb{T}$ is a metric space and the support of the limit distribution δ_{θ^*} is the singleton $\{\theta^*\}$, which is separable, so by Skorokhod's representation theorem (e.g. see Theorem 6.7 in Billingsley, 1968), we can construct random variables $(\Theta'_t)_{t\in\mathbb{N}^*}$ on some instrumental probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such

that $\Theta'_t \sim \Theta_t$ for all $t \in \mathbb{N}^*$ and $\Theta'_t \xrightarrow[t\to+\infty]{\mathbb{P}-a.s.} \theta^\star$, where $\mathbb{P}$ captures the randomness of $(\Theta'_t)_{t\in\mathbb{N}^*}$ conditional on the realizations $(y_t)_{t\in\mathbb{N}^*}$. We have, $\mathbb{P}$-almost surely, for any arbitrary $\varepsilon > 0$ and the corresponding $\delta_\varepsilon > 0$ given by the equicontinuity stated in Condition C7(a), the existence of some $t_0 \in \mathbb{N}^*$ such that, for every $t > t_0$, we have $d(\Theta'_t, \theta^\star) < \delta_\varepsilon$ and

$$|\mathcal{H}(y_t, p(dy_t|y_{1:t-1}, \Theta'_t)) - \mathcal{H}(y_t, p(dy_t|y_{1:t-1}, \theta^\star))| \;\leq\; \varepsilon.$$

Therefore, we have

$$\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \Theta'_t)\right) - \mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \theta^\star)\right) \quad \xrightarrow[t\to+\infty]{\mathbb{P}-a.s.} \quad 0. \tag{B.19}$$

Similarly, using C7(b), we get

$$\frac{\partial \log p(y_t|y_{1:t-1}, \Theta'_t)}{\partial y_t} - \frac{\partial \log p(y_t|y_{1:t-1}, \theta^\star)}{\partial y_t} \quad \xrightarrow[t\to+\infty]{\mathbb{P}-a.s.} \quad 0. \tag{B.20}$$

The family $\{\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \Theta'_t)\right) - \mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \theta^\star)\right)\}_{t\in\mathbb{N}^*}$ is uniformly integrable by Condition C8(a) and the fact that $\Theta'_t \sim \Theta_t \sim p(d\theta|y_{1:t})$ for all $t \in \mathbb{N}^*$, so that the convergence from (B.19) implies the convergence of the first moments (e.g. see Theorem 25.12 in Billingsley, 1995). In other words, we get

$$\mathbb{E}\left[\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \Theta'_t)\right) \mid y_{1:t}\right] - \mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \theta^\star)\right) \quad \xrightarrow[t\to+\infty]{} \quad 0.$$

By construction, we have $\Theta'_t \sim \Theta_t$, thus

$$\mathbb{E}\left[\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \Theta_t)\right) \mid y_{1:t}\right] - \mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \theta^\star)\right) \quad \xrightarrow[t\to+\infty]{} \quad 0.$$

Since this holds for all $\omega \in A$ and $\mathbb{P}_\star(A) = 1$, we conclude that

$$\mathbb{E}\left[\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \Theta)\right) \mid Y_{1:t}\right] - \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right) \quad \xrightarrow[t\to+\infty]{\mathbb{P}_\star-a.s.} \quad 0\,,$$

where the expectation is taken with respect to the posterior distribution of Θ given $Y_{1:t}$, which proves A13(a).

Similarly, the family $\left\{\left(\partial \log p(y_t|y_{1:t-1},\Theta'_t)/\partial y_t - \partial \log p(y_t|y_{1:t-1},\theta^\star)/\partial y_t\right)^2\right\}_{t\in\mathbb{N}^*}$ is uniformly integrable by Condition C8(b) and the fact that $\Theta'_t \sim \Theta_t \sim p(d\theta|y_{1:t})$ for all $t \in \mathbb{N}^*$, so that the convergence from (B.20) implies the convergence of the first two moments, and a fortiori the convergence of the variance. Thus,

$$\mathrm{Var}\left(\frac{\partial \log p(y_t|y_{1:t-1},\Theta'_t)}{\partial y_t} - \frac{\partial \log p(y_t|y_{1:t-1},\theta^\star)}{\partial y_t}\,\middle|\, y_{1:t}\right) \xrightarrow[t\to+\infty]{} 0.$$

By construction, we have $\Theta'_t \sim \Theta_t$. Besides, $\partial \log p(y_t|y_{1:t-1},\theta^\star)/\partial y_t$ is constant given $y_{1:t}$. Therefore,

$$\mathrm{Var}\left(\frac{\partial \log p(y_t|y_{1:t-1},\Theta_t)}{\partial y_t}\,\middle|\, y_{1:t}\right) \xrightarrow[t\to+\infty]{} 0.$$

Since this holds for all $\omega \in A$ and $\mathbb{P}_\star(A) = 1$, we conclude that

$$\mathrm{Var}\left(\frac{\partial \log p(Y_t|Y_{1:t-1},\Theta)}{\partial y_t}\,\middle|\, Y_{1:t}\right) \xrightarrow[t\to+\infty]{\mathbb{P}_\star-a.s.} 0\,,$$

where the variance is taken with respect to the posterior distribution of Θ given $Y_{1:t}$, which proves A13(b). □

B.5.3.4 Proof of Lemma 23

By (1.11)-(1.12) under Assumption A11, the H-score satisfies

$$\begin{aligned}\mathcal{H}(y_t, p(dy_t|y_{-m+1:t-1},\theta^\star)) \;=\; 2\int \left[\frac{\partial^2 \log g_{\theta^\star}(y_t|x_t)}{\partial y_t^2} + \left(\frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t}\right)^2\right] p(dx_t|y_{-m+1:t},\theta^\star) \\ + \left(\int \frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t} p(dx_t|y_{-m+1:t},\theta^\star)\right)^2. \qquad \text{(B.21)}\end{aligned}$$

Under Condition C10, the triangular inequality and the fact that probability densities integrate to 1 lead to

$$\begin{aligned}&|\mathcal{H}(y_t, p(dy_t|y_{-m+1:t-1},\theta^\star)) - \mathcal{H}(y_t, p(dy_t|y_{-m:t-1},\theta^\star))| \\ &\le 2\left|\int \left[\frac{\partial^2 \log g_{\theta^\star}(y_t|x_t)}{\partial y_t^2} + \left(\frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t}\right)^2\right] (p(dx_t|y_{-m+1:t},\theta^\star) - p(dx_t|y_{-m:t},\theta^\star))\right|\end{aligned}$$

$$
\begin{aligned}
&+ \left| \left(\int \frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t} p(dx_t|y_{-m+1:t}, \theta^\star) \right)^2 - \left(\int \frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t} p(dx_t|y_{-m+1:t}, \theta^\star) \right)^2 \right| \\
&\leq 2\,b \left| \int \left(p(dx_t|y_{-m+1:t}, \theta^\star) - p(dx_t|y_{-m:t}, \theta^\star) \right) \right| \\
&\quad + c^2 \left| \int \left(p(dx_t|y_{-m+1:t}, \theta^\star) - p(dx_t|y_{-m:t}, \theta^\star) \right) \right| \left| \int \left(p(dx_t|y_{-m+1:t}, \theta^\star) + p(dx_t|y_{-m:t}, \theta^\star) \right) \right| \\
&\leq 2\,b\, d_{TV}\Big(p(dx_t|y_{-m+1:t}, \theta^\star), p(dx_t|y_{-m:t}, \theta^\star) \Big) + 2\,c^2\, d_{TV}\Big(p(dx_t|y_{-m+1:t}, \theta^\star), p(dx_t|y_{-m:t}, \theta^\star) \Big) \\
&\leq 2\left(b + c^2\right) d_{TV}\left(p(dx_t|y_{-m+1:t}, \theta^\star), p(dx_t|y_{-m:t}, \theta^\star) \right) \\
&\leq 2\left(b + c^2\right) \rho^{t+m-1} \qquad \text{(B.22)}
\end{aligned}
$$

where the last inequality comes from (B.16) under Condition C9. This proves (B.17). From (B.21) and Condition C10, the triangular inequality and the fact that probability densities integrate to 1 yield (B.18). □

B.5.3.5 Proof of Lemma 24

We closely follow the proof of Lemma 13.12 in Douc et al. (2014). We have

$$
p(x_1|Y_{-m+1:0}, \theta^\star) = \int \nu_{\theta^\star}(x_1|x_0)\, p(dx_0|Y_{-m+1:0}, \theta^\star), \qquad \text{(B.23)}
$$

for all $x_1 \in \mathbb{X}$ and all $m \in \mathbb{N}^*$, $\mathbb{P}_\star$-almost surely. By Condition C9 and (B.16), we get

$$
\begin{aligned}
|p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star)| &\leq \sigma^+\, d_{TV}\left(p(dx_0|Y_{-m+1:0}, \theta^\star), p(dx_0|Y_{-m:0}, \theta^\star) \right) \\
&\leq \sigma^+ \rho^{m-1}
\end{aligned}
$$

for all $x_1 \in \mathbb{X}$ and all $m \in \mathbb{N}^*$, $\mathbb{P}_\star$-almost surely. The upper bound does not depend on x_1, hence

$$
\sup_{x_1 \in \mathbb{X}} |p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star)| \leq \sigma^+ \rho^{m-1},
$$

for all $m \in \mathbb{N}^*$, $\mathbb{P}_\star$-almost surely. The geometric series $\sum_m \rho^m$ converges, since $\rho \in (0, 1)$, thus

$$
\sum_{m=1}^{+\infty} \sup_{x_1 \in \mathbb{X}} |p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star)| < +\infty,
$$

$\mathbb{P}_\star$-almost surely. In other words, we have

$$\mathbb{P}_\star\left(\sum_{m=1}^{+\infty}\sup_{x_1\in\mathbb{X}}|p(x_1|Y_{-m+1:0},\theta^\star)-p(x_1|Y_{-m:0},\theta^\star)|<+\infty\right)=1. \tag{B.24}$$

For any $\varepsilon>0$, the convergence of the series in (B.24) guarantees that, $\mathbb{P}_\star$-almost surely, there exists some $N\in\mathbb{N}^*$, such that $\sum_{m=N}^{+\infty}\sup_{x_1\in\mathbb{X}}|p(x_1|Y_{-m+1:0},\theta^\star)-p(x_1|Y_{-m:0},\theta^\star)|<\varepsilon$. Then, for all $r>s>N$,

$$\begin{aligned}\sup_{x_1\in\mathbb{X}}|p(x_1|Y_{-s:0},\theta^\star)-p(x_1|Y_{-r:0},\theta^\star)| &= \sup_{x_1\in\mathbb{X}}\left|\sum_{m=s+1}^{r}p(x_1|Y_{-m+1:0},\theta^\star)-p(x_1|Y_{-m:0},\theta^\star)\right|\\ &\leq \sum_{m=s+1}^{r}\sup_{x_1\in\mathbb{X}}|p(x_1|Y_{-m+1:0},\theta^\star)-p(x_1|Y_{-m:0},\theta^\star)|\\ &\leq \sum_{m=N}^{+\infty}\sup_{x_1\in\mathbb{X}}|p(x_1|Y_{-m+1:0},\theta^\star)-p(x_1|Y_{-m:0},\theta^\star)|\\ &\leq \varepsilon.\end{aligned}$$

This implies that, $\mathbb{P}_\star$-almost surely, the sequence of non-negative continuous functions $(x_1\mapsto p(x_1|Y_{-m:0},\theta^\star))_{m\in\mathbb{N}}$ converges uniformly to a limit function $x_1\mapsto p(x_1|Y_{-\infty:0},\theta^\star)=\lim_{m\to+\infty}p(x_1|Y_{-m:0},\theta^\star)$, which is itself necessarily non-negative and continuous, as a uniform limit of such functions. We can now check that $x_1\mapsto p(x_1|Y_{-\infty:0},\theta^\star)$ is indeed a probability density function.

On the one hand, applying Fatou's Lemma to the collection of non-negative functions $(x_1\mapsto p(x_1|Y_{-m:0},\theta^\star))_{m\in\mathbb{N}}$ yields

$$\int p(x_1|Y_{-\infty:0},\theta^\star)\eta(dx_1)=\int\liminf_{m\to+\infty}p(x_1|Y_{-m:0},\theta^\star)\eta(dx_1)\leq\liminf_{m\to+\infty}\int p(x_1|Y_{-m:0},\theta^\star)\eta(dx_1)=1,$$

where η is the dominating measure introduced in Condition C9(a).

On the other hand, (B.23) and Condition C9 imply $0\leq p(x_1|Y_{-m:0},\theta^\star)\leq\sigma^+$. Applying Fatou's Lemma to the non-negative functions $(x_1\mapsto\sigma^+-p(x_1|Y_{-m:0},\theta^\star))_{m\in\mathbb{N}}$ yields

$$1=\limsup_{m\to+\infty}\int p(x_1|Y_{-m:0},\theta^\star)\eta(dx_1)\leq\int\limsup_{m\to+\infty}p(x_1|Y_{-m:0},\theta^\star)\eta(dx_1)=\int p(x_1|Y_{-\infty:0},\theta^\star)\eta(dx_1).$$

These two inequalities hold $\mathbb{P}_\star$-almost surely, and prove that, $\mathbb{P}_\star$-almost surely, the function $x_1 \mapsto p(x_1|Y_{-\infty:0}, \theta^\star)$ is a probability density with respect to η.

Furthermore, for all $y_1 \in \mathbb{Y}$, all $x_1 \in \mathbb{X}$, and all $m \in \mathbb{N}^*$, we have, $\mathbb{P}_\star$-almost surely,

$$p(y_1|Y_{-m+1:0}, \theta^\star) = \int g_{\theta^\star}(y_1|x_1)\nu_{\theta^\star}(x_1|x_0)\, p(dx_0|Y_{-m+1:0}, \theta^\star)dx_1.$$

By using again (B.16), we get

$$\begin{aligned}
&|p(y_1|Y_{-m+1:0}, \theta^\star) - p(y_1|Y_{-m:0}, \theta^\star)| \\
&\leq \sigma^+ \sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} g_{\theta^\star}(y|x)\ \ d_{TV}\left(p(dx_0|Y_{-m+1:0}, \theta^\star), p(dx_0|Y_{-m:0}, \theta^\star)\right) \\
&\leq \sigma^+ \sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} g_{\theta^\star}(y|x)\ \rho^{m-1}
\end{aligned}$$

for all $y_1 \in \mathbb{Y}$ and all $m \in \mathbb{N}^*$, $\mathbb{P}_\star$-almost surely. The supremum is finite thanks to Condition C11. Using a similar reasoning as in the first part of the proof, we get

$$\mathbb{P}_\star\left(\sum_{m=1}^{+\infty} \sup_{y_1\in\mathbb{Y}} |p(y_1|Y_{-m+1:0}, \theta^\star) - p(y_1|Y_{-m:0}, \theta^\star)| < +\infty\right) = 1, \tag{B.25}$$

so that, $\mathbb{P}_\star$-almost surely, the sequence of functions $(y_1 \mapsto p(y_1|Y_{-m:0}, \theta^\star))_{m\in\mathbb{N}}$ converges uniformly to a limit function $y_1 \mapsto p(y_1|Y_{-\infty:0}, \theta^\star)$, and $p(Y_1|Y_{-\infty:0}, \theta^\star) = p(y_1|Y_{-\infty:0}, \theta^\star)|_{y_1=Y_1}$.

Consider an event $K \subseteq \mathbb{Y}$ such that $\lambda(K) < +\infty$, where λ denotes the Lebesgue measure. On the one hand, martingale convergence theorems (e.g. Corollary B.13 in Douc et al., 2014) guarantee that, $\mathbb{P}_\star$-almost surely,

$$\mathbb{E}\left[\mathbb{1}_K(Y_1)|Y_{-\infty:0}, \theta^\star\right] = \lim_{m\to+\infty} \mathbb{E}\left[\mathbb{1}_K(Y_1)|Y_{-m:0}, \theta^\star\right]. \tag{B.26}$$

On the other hand, the uniform convergence of the functions $(y_1 \mapsto p(y_1|Y_{-m:0}, \theta^\star))_{m\in\mathbb{N}}$ and the finiteness of $\lambda(K)$ allow us to interchange the order of limits and integration. This implies that, $\mathbb{P}_\star$-almost surely, we have

$$\lim_{m\to+\infty} \mathbb{E}\left[\mathbb{1}_K(Y_1)|Y_{-m:0}, \theta^\star\right] = \lim_{m\to+\infty} \int \mathbb{1}_K(y_1)p(y_1|Y_{-m:0}, \theta^\star)\lambda(dy_1)$$

$$= \int \mathbb{1}_K(y_1) \lim_{m\to+\infty} p(y_1|Y_{-m:0},\theta^\star)\lambda(dy_1)$$
$$= \int \mathbb{1}_K(y_1) p(y_1|Y_{-\infty:0},\theta^\star)\lambda(dy_1). \tag{B.27}$$

Combining (B.26) and (B.27) leads to

$$\mathbb{E}\left[\mathbb{1}_K(Y_1)|Y_{-\infty:0},\theta^\star\right] = \int \mathbb{1}_K(y_1) p(y_1|Y_{-\infty:0},\theta^\star)\lambda(dy_1),$$

for any event $K \subseteq \mathbb{Y}$ with $\lambda(K) < +\infty$, $\mathbb{P}_\star$-almost surely. This proves that, $\mathbb{P}_\star$-almost surely, $y_1 \mapsto p(y_1|Y_{-\infty:0},\theta^\star)$ is the conditional density of Y_1 given $Y_{-\infty:0}$. Finally, we get $\log p(y_1|Y_{-\infty:0},\theta^\star) = \lim_{m\to+\infty} \log p(y_1|Y_{-m+1:0},\theta^\star)$ for all $y_1 \in \mathbb{Y}$, $\mathbb{P}_\star$-almost surely, by applying Proposition 13.5 from Douc et al. (2014).

Under Assumption A11, the function $y_1 \mapsto \log p(y_1|Y_{-m+1:0},\theta^\star)$ is $\mathbb{P}_\star$-almost surely twice differentiable for all $m \in \mathbb{N}$. $\mathbb{P}_\star$-almost surely, for all $y_1 \in \mathbb{Y}$, the first derivative is

$$\frac{\partial \log p(y_1|Y_{-m+1:0},\theta^\star)}{\partial y_1} = \int \left(\frac{\partial \log g_{\theta^\star}(y_1|x_1)}{\partial y_1}\right) p(x_1|Y_{-m+1:0},\theta^\star)\, dx_1\,,$$

and the second derivative satisfies

$$\frac{\partial^2 \log p(y_1|Y_{-m+1:0},\theta^\star)}{\partial y_1^2} = -\left(\frac{\partial \log p(y_1|Y_{-m+1:0},\theta^\star)}{\partial y_1}\right)^2 + \int \left[\frac{\partial^2 \log g_{\theta^\star}(y_1|x_1)}{\partial y_1^2} + \left(\frac{\partial \log g_{\theta^\star}(y_1|x_1)}{\partial y_1}\right)^2\right] p(x_1|Y_{-m+1:0},\theta^\star)\, dx_1.$$

We will prove the $\mathbb{P}_\star$-almost sure twice differentiability of $y_1 \mapsto \log p(y_1|Y_{-\infty:0},\theta^\star)$ by proving that the sequences of derivatives $(y_1 \mapsto \partial \log p(y_1|Y_{-m+1:0},\theta^\star)/\partial y_1)_{m\in\mathbb{N}}$ and $(y_1 \mapsto \partial^2 \log p(y_1|Y_{-m+1:0},\theta^\star)/\partial y_1^2)_{m\in\mathbb{N}}$ converge uniformly to well-defined limit functions, $\mathbb{P}_\star$-almost surely. Such uniform convergences imply the twice differentiability of the limit of $(y_1 \mapsto \log p(y_1|Y_{-m+1:0},\theta^\star))_{m\in\mathbb{N}}$ by virtue of Theorem 7.17 from Rudin (1964).

From Condition C10 and (B.16), we have, $\mathbb{P}_\star$-almost surely, for all $m \in \mathbb{N}$ and all $y_1 \in \mathbb{Y}$,

$$\left|\frac{\partial \log p(y_1|Y_{-m+1:0},\theta^\star)}{\partial y_1} - \frac{\partial \log p(y_1|Y_{-m:0},\theta^\star)}{\partial y_1}\right| \leq c\, d_{TV}\left(p(dx_1|Y_{-m+1:0},\theta^\star), p(dx_1|Y_{-m:0},\theta^\star)\right)$$
$$\leq c\,\rho^m.$$

As the upper bound does not depend on $y_1 \in \mathbb{Y}$, we have, $\mathbb{P}_\star$-almost surely, for all $m \in \mathbb{N}$,

$$\sup_{y_1\in\mathbb{Y}} \left|\frac{\partial \log p(y_1|Y_{-m+1:0},\theta^\star)}{\partial y_1} - \frac{\partial \log p(y_1|Y_{-m:0},\theta^\star)}{\partial y_1}\right| \leq c\,\rho^m,$$

where $\rho \in (0,1)$. By using the triangle inequality, we have, $\mathbb{P}_\star$-almost surely,

$$\begin{aligned}
&\sup_{y_1 \in \mathbb{Y}} \left| \sum_{k=m}^{+\infty} \left(\frac{\partial \log p(y_1|Y_{-k+1:0}, \theta^\star)}{\partial y_1} - \frac{\partial \log p(y_1|Y_{-k:0}, \theta^\star)}{\partial y_1} \right) \right| \\
&\leq \sum_{k=m}^{+\infty} \sup_{y_1 \in \mathbb{Y}} \left| \frac{\partial \log p(y_1|Y_{-k+1:0}, \theta^\star)}{\partial y_1} - \frac{\partial \log p(y_1|Y_{-k:0}, \theta^\star)}{\partial y_1} \right| \\
&\leq c \sum_{k=m}^{+\infty} \rho^k \\
&\leq c \, \frac{\rho^m}{1-\rho}.
\end{aligned}$$

Using telescopic sums, and $\rho^m \to 0$ when $m \to +\infty$ since $\rho \in (0,1)$, we get

$$\sup_{y_1 \in \mathbb{Y}} \left| \frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} - \frac{\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1} \right| \xrightarrow[m \to +\infty]{\mathbb{P}_\star - a.s.} 0 \,,$$

where

$$\frac{\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1} = \lim_{\substack{m \to +\infty \\ \mathbb{P}_\star\text{-a.s.}}} \frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} \,.$$

In other words, $\mathbb{P}_\star$-a.s., the sequence of derivatives $(y_1 \mapsto \partial \log p(y_1|Y_{-m+1:0}, \theta^\star)/\partial y_1)_{m \in \mathbb{N}}$ converges uniformly to the function $y_1 \mapsto \partial \log p(y_1|Y_{-\infty:0}, \theta^\star)/\partial y_1$. Besides, we have proved earlier that the sequence of functions $(y_1 \mapsto \log p(y_1|Y_{-m+1:0}, \theta^\star))_{m \in \mathbb{N}}$ converges pointwise to the limit function $y_1 \mapsto \log p(y_1|Y_{-\infty:0}, \theta^\star)$. By using Theorem 7.17 from Rudin (1964), the limit function $y_1 \mapsto \log p(y_1|Y_{-\infty:0}, \theta^\star)$ is $\mathbb{P}_\star$-almost surely differentiable and its derivative is given $\mathbb{P}_\star$-almost surely by

$$\frac{\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1} = \lim_{m \to +\infty} \frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} \,.$$

Regarding the second derivative, we can follow the approach used to derive (B.22) in the proof of Lemma 23, so that, $\mathbb{P}_\star$-almost surely, for all $m \in \mathbb{N}$ and all $y_1 \in \mathbb{Y}$, we have

$$\left| \left(\frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} \right)^2 - \left(\frac{\partial \log p(y_1|Y_{-m:0}, \theta^\star)}{\partial y_1} \right)^2 \right| \leq 2\, c^2\, \rho^m,$$

By using again the triangle inequality, telescopic sums, and the fact that $\rho \in (0,1)$ so that $\rho^m \to 0$ when $m \to +\infty$, we get

$$\sup_{y_1 \in \mathbb{Y}} \left| \sum_{k=m}^{+\infty} \left(\left(\frac{\partial \log p(y_1|Y_{-k+1:0}, \theta^\star)}{\partial y_1} \right)^2 - \left(\frac{\partial \log p(y_1|Y_{-k:0}, \theta^\star)}{\partial y_1} \right)^2 \right) \right| \xrightarrow[m \to +\infty]{\mathbb{P}_\star - a.s.} 0 \,,$$

which implies that, $\mathbb{P}_\star$-a.s., the sequence of functions $\left(y_1 \mapsto \left(\partial \log p(y_1|Y_{-k+1:0}, \theta^\star)/\partial y_1\right)^2 \right)_{m \in \mathbb{N}}$ converges uniformly to some limit function

$$y_1 \mapsto \lim_{m \to +\infty} \left(\frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} \right)^2 . \tag{B.28}$$

By following again the derivation of (B.22) in the proof of Lemma 23, we get, $\mathbb{P}_\star$-a.s., for all $m \in \mathbb{N}$ and all $y_1 \in \mathbb{Y}$,

$$\left| \int \left(\frac{\partial^2 \log g_{\theta^\star}(y_1|x_1)}{\partial y_1^2} + \left(\frac{\partial \log g_{\theta^\star}(y_1|x_1)}{\partial y_1} \right)^2 \right) \left(p(dx_1|Y_{-m+1:0}, \theta^\star) - p(dx_1|Y_{-m:0}, \theta^\star) \right) \right| \leq b\, \rho^m .$$

As previously, the triangle inequality, telescopic sums, and $\rho \in (0,1)$ imply that, $\mathbb{P}_\star$-almost surely, the sequence

$$\left(y_1 \mapsto \int \left(\frac{\partial^2 \log g_{\theta^\star}(y_1|x_1)}{\partial y_1^2} + \left(\frac{\partial \log g_{\theta^\star}(y_1|x_1)}{\partial y_1} \right)^2 \right) p(dx_1|Y_{-m+1:0}, \theta^\star) \right)_{m \in \mathbb{N}}$$

converges uniformly to some limit function

$$y_1 \mapsto \lim_{m \to +\infty} \int \left(\frac{\partial^2 \log g_{\theta^\star}(y_1|x_1)}{\partial y_1^2} + \left(\frac{\partial \log g_{\theta^\star}(y_1|x_1)}{\partial y_1} \right)^2 \right) p(dx_1|Y_{-m+1:0}, \theta^\star). \tag{B.29}$$

Since a sum of two uniformly convergent sequences of functions is still uniformly convergent, with the limit function being the sum of the two limit functions, the previous results imply that the sequence of second derivatives $(y_1 \mapsto \partial^2 \log p(y_1|Y_{-m+1:0}, \theta^\star)/\partial y_1^2)_{m \in \mathbb{N}}$ converges uniformly to the function $y_1 \mapsto \partial^2 \log p(y_1|Y_{-\infty:0}, \theta^\star)/\partial y_1^2$ defined as the sum of the limit functions in (B.28) and (B.29), $\mathbb{P}_\star$-almost surely. By using again Theorem 7.17 from Rudin (1964), the function $y_1 \mapsto \log p(y_1|Y_{-\infty:0}, \theta^\star)$ is twice differentiable with second derivative equal to $y_1 \mapsto \partial^2 \log p(y_1|Y_{-\infty:0}, \theta^\star)/\partial y_1^2$, $\mathbb{P}_\star$-almost surely.

By (B.12) and the previous results, we get, $\mathbb{P}_\star$-almost surely, for all $y_1 \in \mathbb{Y}$,

$$\begin{aligned}
&\mathcal{H}(y_1, p(dy_1|Y_{-\infty:0}, \theta^\star)) \\
&= \lim_{m\to+\infty} \mathcal{H}(y_1, p(dy_1|Y_{-m+1:0}, \theta^\star)) \\
&= \lim_{m\to+\infty} \left(2\,\frac{\partial^2 \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1^2} + \left(\frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} \right)^2 \right) \\
&= 2 \lim_{m\to+\infty} \left(\frac{\partial^2 \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1^2} \right) + \left(\lim_{m\to+\infty} \frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} \right)^2 \\
&= 2\,\frac{\partial^2 \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1^2} + \left(\frac{\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1} \right)^2 .
\end{aligned}$$

□

B.5.4 Heuristic proof for well-specified nested i.i.d. models

In this section, we go back to the i.i.d. setting of Section B.5.1.1, but we now assume that model M_1 is nested in model M_2, in the sense of Eq. (9) in Berger and Pericchi (1996). In other words, with $k_1, k_2 \in \mathbb{N}$ denoting the dimensions of the parameter spaces $\mathbb{T}_1$ and $\mathbb{T}_2$ with $k_2 > k_1 > 0$, we have $\mathbb{T}_2 = \{(\theta_1, \eta) \in \Xi_1 \times \Xi_2\} \subseteq \mathbb{R}^{k_1} \times \mathbb{R}^{k_2-k_1}$ and $\mathbb{T}_1 \subseteq \Xi_1$, and there exists $\eta_1^\star \in \Xi_2$ such that for all $y \in \mathbb{Y}$ and all $\theta_1 \in \mathbb{T}_1$, we have $p_1(y|\theta_1) = p_2(y|\theta_1, \eta_1^\star)$. We also assume that both M_1 and M_2 are well-specified, so that there exists $\theta_1^\star \in \mathbb{T}_1$ such that, for all $y \in \mathbb{Y}$, we have $p_\star(y) = p_1(y|\theta_1^\star) = p_2(y|\theta_2^\star)$, where $\theta_2^\star = (\theta_1^\star, \eta_1^\star)$. For simplicity, we assume $\theta_1^\star$ is in the interior of $\mathbb{T}_1$. In this setting, the H-score is said to be consistent if it asymptotically chooses the model of smaller dimension, i.e. M_1. It is enough to show that $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) \to +\infty$ as $T \to +\infty$, in $\mathbb{P}_\star$-probability. The rest of this section is meant as a mere proof of concept with some heuristic arguments for this consistency to hold. We define

$$\begin{aligned}
\Lambda_T(M_2, M_1) &= \sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p_2(dy|\Theta_2)\right) \,\middle|\, Y_{1:t} \right] - \sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p_1(dy|\Theta_1)\right) \,\middle|\, Y_{1:t} \right], \\
\Delta_T(M_2, M_1) &= \sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_2(Y_t|\Theta_2)}{\partial y_t} \,\middle|\, Y_{1:t} \right) - \sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_1(Y_t|\Theta_1)}{\partial y_t} \,\middle|\, Y_{1:t} \right),
\end{aligned}$$

so that we can write, using (1.5),

$$\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) = \Lambda_T(M_2, M_1) + \Delta_T(M_2, M_1). \tag{B.30}$$

In the non-nested misspecified setting, the first term $\Lambda_T(M_2, M_1)$ would typically dominate and drive the difference $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)$, which would then behave asymptotically as $T\left(\mathrm{D}_{\mathcal{H}}(p_\star, M_2) - \mathrm{D}_{\mathcal{H}}(p_\star, M_1)\right)$. However, in the nested well-specified setting, we have $\mathrm{D}_{\mathcal{H}}(p_\star, M_2) - \mathrm{D}_{\mathcal{H}}(p_\star, M_1) = 0$, so that we would expect the second term $\Delta_T(M_2, M_1)$ to take over, and act as a penalty reflecting the difference in dimensions. This penalty term will dictate the asymptotic behavior of $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)$, provided that the first term $\Lambda_T(M_2, M_1)$ does not grow too fast (Assumption A22).

This can be informally motivated by the following arguments. Under appropriate regularity conditions, for each $j \in \{1, 2\}$, the posterior of Θ_j given $Y_{1:T}$ under model M_j concentrates in $\mathbb{P}_\star$-probability around the maximum likelihood estimator (MLE) defined as $\hat{\theta}^{\mathrm{MLE}}_{j,T} = \arg\min_{\theta \in \mathbb{T}_j} -\sum_{t=1}^T \log p_j(Y_t|\theta)$ (e.g. see Theorem 1.3.4 in Ghosh and Ramamoorthi, 2003). Under conditions ensuring the consistency of M-estimators (e.g. see Theorem 5.7 in Van der Vaart, 2000), the minimum H-score estimator (mHE) defined as $\hat{\theta}^{\mathrm{mHE}}_{j,T} = \arg\min_{\theta \in \mathbb{T}_j} \sum_{t=1}^T \mathcal{H}\left(Y_t, p_j(dy|\theta)\right)$ gets asymptotically close to the MLE, in the sense that $\lim_{T\to+\infty}(\hat{\theta}^{\mathrm{MLE}}_{j,T} - \hat{\theta}^{\mathrm{mHE}}_{j,T}) = (\theta^\star_j - \theta^\star_j) = 0$ in $\mathbb{P}_\star$-probability, where the limits of the MLE and the mHE coincide by well-specification and identifiability of the models, and strict propriety of the H-score. We can then write

$$\Lambda_T(M_2, M_1) = \sum_{t=1}^T \mathcal{H}\left(Y_t, p_2(dy|\hat{\theta}^{\mathrm{mHE}}_{2,T})\right) - \sum_{t=1}^T \mathcal{H}\left(Y_t, p_1(dy|\hat{\theta}^{\mathrm{mHE}}_{1,T})\right) + \sum_{j=1}^2 R^{\mathrm{MLE}}_{j,T} + \sum_{j=1}^2 R^{\mathrm{mHE}}_{j,T},$$

where the remainder terms for each $j \in \{1, 2\}$ are defined as

$$R^{\mathrm{MLE}}_{j,T} = \sum_{t=1}^T \left(\mathbb{E}\left[\mathcal{H}\left(Y_t, p_j(dy|\Theta_j)\right) \,\middle|\, Y_{1:t}\right] - \mathcal{H}\left(Y_t, p_j(dy|\hat{\theta}^{\mathrm{MLE}}_{j,T})\right)\right),$$

$$R^{\mathrm{mHE}}_{j,T} = \sum_{t=1}^T \left(\mathcal{H}\left(Y_t, p_j(dy|\hat{\theta}^{\mathrm{MLE}}_{j,T})\right) - \mathcal{H}\left(Y_t, p_j(dy|\hat{\theta}^{\mathrm{mHE}}_{j,T})\right)\right).$$

The term $R_{j,T}^{\mathrm{MLE}}$ is controlled by how fast the posterior of Θ_j concentrates around $\hat{\theta}_{j,T}^{\mathrm{MLE}}$ and how well the posterior expectation of $\mathcal{H}(Y_t, p_j(dy|\Theta_j)$ can be approximated by its expectation with respect to a Dirac mass at that MLE, which would typically require uniform integrability conditions for the posterior moments to converge. The term $R_{j,T}^{\mathrm{mHE}}$ is controlled by how fast $\hat{\theta}_{j,T}^{\mathrm{mHE}}$ approaches $\hat{\theta}_{j,T}^{\mathrm{MLE}}$ and how smooth the function $(y, \theta_j) \mapsto \mathcal{H}(y, p_j(dy|\theta_j))$ is. Using the Landau notation (i.e. small and big O), we assume that $R_{1,T}^{\mathrm{MLE}} + R_{2,T}^{\mathrm{MLE}} + R_{1,T}^{\mathrm{mHE}} + R_{2,T}^{\mathrm{mHE}} = o(\log T)$ as $T \to +\infty$, in $\mathbb{P}_\star$-probability, for ease of exposition. Finally, we can reasonably assume that

$$\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p_2(dy|\hat{\theta}_{2,T}^{\mathrm{mHE}})\right) - \sum_{t=1}^{T} \mathcal{H}\left(Y_t, p_1(dy|\hat{\theta}_{1,T}^{\mathrm{mHE}})\right) = O(1) \tag{B.31}$$

as $T \to +\infty$, in $\mathbb{P}_\star$-probability. Indeed, the difference between H-score minima is analogous to the difference between log-likelihood maxima $\sum_{t=1}^{T} \log p_1(Y_t|\hat{\theta}_{1,T}^{\mathrm{MLE}}) - \sum_{t=1}^{T} \log p_2(Y_t|\hat{\theta}_{2,T}^{\mathrm{MLE}})$ appearing when trying to prove the consistency of the log-Bayes factor (e.g. see Chib and Kuffner, 2016). Under suitable conditions, the difference between log-likelihood maxima converges in distribution to a scaled $\chi^2_{k_2-k_1}$ distribution (e.g. see Vuong, 1989). This is known as Wilks's theorem in the likelihood ratio test literature (Wilks, 1938), and its proof essentially relies on a Taylor expansion combined with the asymptotic Normality of the MLE. Under differentiability assumptions on the functions $\theta_j \mapsto \sum_{t=1}^{T} \mathcal{H}\left(Y_t, p_j(dy|\theta_j)\right)$ and asymptotic Normality of the M-estimators $\hat{\theta}_{j,T}^{\mathrm{mHE}}$ (e.g. see Theorem 5.23 in Van der Vaart, 2000), we may assume that a similar distributional result holds for the difference of H-score minima, leading to (B.31). All the previous heuristic arguments motivate Assumption A22 stated below.

Assumption A22. $\Lambda_T(M_2, M_1) = o(\log T)$ *as* $T \to +\infty$, *in* $\mathbb{P}_\star$*-probability.*

Looking now at the penalty term $\Delta_T(M_2, M_1)$, we can define, for each $j \in \{1, 2\}$, the function

$$G_j : \mathbb{Y} \times \mathbb{T}_j \longrightarrow \mathbb{R}$$

$$(y_t, \theta_j) \longmapsto \frac{\partial \log p_j(y_t|\theta_j)}{\partial y_t}.$$

Under conditions enabling the Bernstein-von Mises theorem to hold (e.g. see Theorem 1.4.2 in Ghosh and Ramamoorthi, 2003), as $t \to +\infty$, the posterior distribution of $\sqrt{t}(\Theta_{j,t} - \hat{\theta}_{j,t}^{\mathrm{MLE}})$ would approach a Normal distribution with mean 0 and covariance matrix $V_j^\star = I_j(\theta_j^\star)^{-1}$, where $I_j(\theta_j^\star) = \mathbb{E}_\star[-\nabla_{\theta_j}^2 \log p_j(Y|\theta_j^\star)]$ corresponds to the Fisher information matrix at $\theta_j^\star$ when M_j is well-specified. Provided that the posterior second moments converge and G_j is differentiable with respect to θ_j such that $\nabla_{\theta_j} G_j(y_t, \theta_j^\star) \neq 0$ for all $y_t \in \mathbb{Y}$, applying the Delta method to the previous statement motivates Assumption A23.

Assumption A23. $\mathrm{Var}\left(\frac{\partial \log p_j(Y_t|\Theta_j)}{\partial y_t}\middle| Y_{1:t}\right) = \nabla_{\theta_j} G_j(Y_t, \theta_j^\star)^\top (V_j^\star/t) \nabla_{\theta_j} G_j(Y_t, \theta_j^\star) + o\left(\frac{1}{t}\right)$ *as* $t \to +\infty$, $\mathbb{P}_\star$*-almost surely, for each* $j \in \{1, 2\}$.

Using $\sum_{t=1}^T (1/t) = O(\log T)$ as $T \to +\infty$, and Assumption A23, we have, $\mathbb{P}_\star$-almost surely,

$$\Delta_T(M_2, M_1) = \sum_{t=1}^T \frac{1}{t} K_t + o(\log T), \tag{B.32}$$

where $K_t = \left(\nabla_{\theta_2} G_2(Y_t, \theta_2^\star)^\top V_2^\star \nabla_{\theta_2} G_2(Y_t, \theta_2^\star) - \nabla_{\theta_1} G_1(Y_t, \theta_1^\star)^\top V_1^\star \nabla_{\theta_1} G_1(Y_t, \theta_1^\star)\right)$. The K_t's are i.i.d., so that if we assume $\mathbb{E}_\star[K_t^2] < +\infty$ and define $\sigma^2 = \mathrm{Var}_\star(K_t)$, then we have

$$\mathrm{Var}_\star\left(\sum_{t=1}^T \frac{1}{t} K_t\right) = \sigma^2 \sum_{t=1}^T \frac{1}{t^2} \xrightarrow[T \to +\infty]{} \sigma^2 \frac{\pi^2}{6} < +\infty. \tag{B.33}$$

On the other hand, if we define $\delta_{21} = \mathbb{E}_\star[K_t]$, then we have

$$\mathbb{E}_\star\left(\sum_{t=1}^T \frac{1}{t} K_t\right) = \delta_{21} \sum_{t=1}^T \frac{1}{t} = \delta_{21} \log T + o(\log T). \tag{B.34}$$

Using Chebyshev's inequality, combining (B.32), (B.33), and (B.34) leads to

$$\Delta_T(M_2, M_1) = \delta_{21} \log T + o(\log T),$$

in $\mathbb{P}_\star$-probability. Piecing everything together, we finally get

$$\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) = \delta_{21} \log T + o(\log T), \tag{B.35}$$

in $\mathbb{P}_\star$-probability, where

$$\delta_{21} = \mathbb{E}_\star \left[\nabla_{\theta_2} G_2(Y, \theta_2^\star)^\top V_2^\star \nabla_{\theta_2} G_2(Y, \theta_2^\star) - \nabla_{\theta_1} G_1(Y, \theta_1^\star)^\top V_1^\star \nabla_{\theta_1} G_1(Y, \theta_1^\star) \right]. \tag{B.36}$$

This implies that as $T \to +\infty$, the sign of $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)$ will be determined by the sign of δ_{21}. If $\theta_2 = (\theta_1, \eta)$ consists of orthogonal parameters, in the sense that $\mathbb{E}_\star[\nabla_\eta \nabla_{\theta_1} \log p_2(Y|\theta_1^\star, \eta_1^\star)] = 0$ so that the Fisher information matrix $I_2(\theta_2^\star)$ is block-diagonal, then $V_2^\star$ has the following block-diagonal structure

$$V_2^\star = \left(\begin{array}{c|c} V_1^\star & 0 \\ \hline 0 & V_{22}^\star \end{array} \right),$$

where $V_{22}^\star = \mathbb{E}_\star[-\nabla_\eta^2 \log p_2(Y|\theta_2^\star)]^{-1}$ is a symmetric positive definite matrix of size $(k_2 - k_1)$. Using (B.5.4) and the fact that $\nabla_{\theta_2} G_2(y_t, \theta_2^\star)^\top = (\nabla_{\theta_1} G_1(y_t, \theta_1^\star)^\top, \nabla_\eta G_2(y_t, \theta_2^\star)^\top)$, we get $\delta_{21} = \mathbb{E}_\star[\nabla_\eta G_2(Y, \theta_2^\star)^\top V_{22}^\star \nabla_\eta G_2(Y, \theta_2^\star)] > 0$, where the positivity comes from the positive definiteness of $V_{22}^\star$. In other words, (B.36) becomes

$$\delta_{21} = \mathbb{E}_\star \left[\left(\nabla_\eta \frac{\partial \log p_2(Y|\theta_2^\star)}{\partial y} \right)^\top \mathbb{E}_\star[-\nabla_\eta^2 \log p_2(Y|\theta_2^\star)]^{-1} \left(\nabla_\eta \frac{\partial \log p_2(Y|\theta_2^\star)}{\partial y} \right) \right] > 0. \tag{B.37}$$

This implies $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) \xrightarrow[T \to +\infty]{} +\infty$, so that we asymptotically choose the smaller model M_1, as desired.

B.5.5 Illustration of parsimony with nested Normal models

We consider the following nested Normal models

$$\begin{aligned} M_1: &\quad Y_1, ..., Y_T \,|\, \mu_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu_1, 1\right), \quad \mu_1 \sim \mathcal{N}\left(0, \sigma_0^2\right), \\ M_2: &\quad Y_1, ..., Y_T \,|\, \mu_2, \sigma_2^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu_2, \sigma_2^2\right), \quad \mu_2 \,|\, \sigma_2^2 \sim \mathcal{N}\left(0, \sigma_2^2\right), \quad \sigma_2^2 \sim \text{Inv-}\chi^2\left(\nu_0, s_0^2\right). \end{aligned}$$

The positive hyperparameters are chosen as $\sigma_0^2 = 10$, $\nu_0 = 0.1$, and $s_0^2 = 1$. We compare M_1 and M_2, using data generated as $Y_1, ..., Y_T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_\star, \sigma_\star^2)$, in the following settings: (1)

$(\mu_\star, \sigma_\star^2) = (0, 5)$, i.e. M_2 is well-specified while M_1 is not; (2) $(\mu_\star, \sigma_\star^2) = (0, 1)$, i.e. both M_1 and M_2 are well-specified. For each case, we generate $T = 10^5$ observations and perform 5 runs of SMC with $N_\theta = 1024$ particles to estimate the log-Bayes factors and H-factors of M_1 against M_2. Each run averages the factors over 100 different orderings of the data, sampled uniformly from all the possible permutations. The results are shown in Figure B.4.

In case 1, both factors correctly select the well-specified model M_2, as expected. In case 2, M_1 is nested in M_2 with respective dimensions $k_1 = 1$ and $k_2 = 2$. Besides, the Fisher information under the Normal model M_2 is diagonal. Using the same notation as in (B.37), we get $\theta_2^\star = (0, 1)$ and $\delta_{21} = 2$, so that our postulated result in (B.35) becomes $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) = 2 \log T + o(\log T)$. Regarding the log-Bayes factor, standard approximation via the BIC yields $\log p_2(Y_{1:T}) - \log p_1(Y_{1:T}) = \left(\frac{k_2 - k_1}{2}\right) \log T + o(\log T) = (1/2) \log T + o(\log T)$.

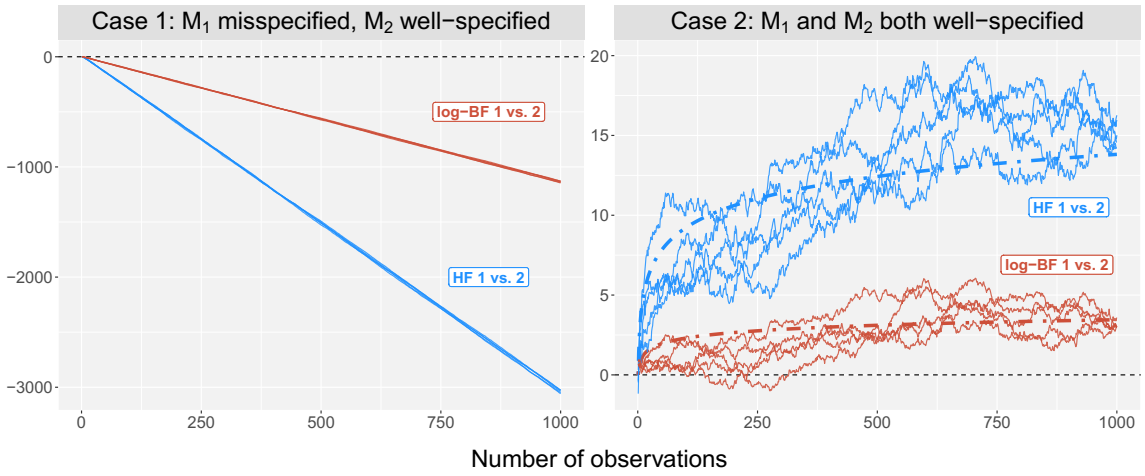

Figure B.4. *Estimated log-Bayes factors (log-BF) and H-factors (HF) of M_1 against M_2, computed for 5 replications (thin solid lines), under two i.i.d. data-generating processes: $\mathcal{N}(0, 5)$ (Case 1) and $\mathcal{N}(0, 1)$ (Case 2). Each run averages the factors over 100 random orderings of the data. The variability within each factor is due to these random orderings and Monte Carlo error. In case 1 (left panel), this error is negligible relative to the magnitude of the factors, which makes the 5 replications hard to distinguish. In case 2 (right panel), the dot-dashed lines indicate the theoretical logarithmic growth of each factor: they respectively correspond to $t \mapsto (1/2) \log(t)$ (log-BF) and $t \mapsto 2 \log(t)$ (HF). See Appendix B.5.5.*

C

Supplementary material for Chapter 2

Contents

This appendix provides additional computational details for the derivations presented in Section 2.3.1 of Chapter 2. It also provides additional plots to accompany the numerical diagnostics of Section 2.3.4.

C.1 H-scores as functions of posterior expectations

We complement Sections 2.3.1.2 and 2.3.1.3 of Chapter 2 by providing explicit expressions of the H-score as a function of posterior expectations, both in the case where the observations are continuous (Appendix C.1.1) and discrete (Appendix C.1.2).

C.1.1 Continuous observations

Using equation (2.21) of Section 2.3.1.2, we can write the H-score at horizon $s \in [\![1, T]\!]$ as

$$\mathcal{H}_s(M) \;=\; h_s\left((E_{t,d,j})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![1,2]\!]}\right)$$

where $(E_{t,d,j})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![1,2]\!]}$ is the vector enumerating the $E_{t,d,j}$'s in the lexicographic order of $[\![1,s]\!] \times [\![1,d_y]\!] \times [\![1,2]\!]$ induced by the natural order on $\mathbb{N}$ i.e.

$$(1,1,1) < (1,1,2) < (1,2,1) < ... < (1,d_y,2) < (2,1,1) < ..... < (s,d_y,2)$$

and h_s is a continuously differentiable function defined for all $\xi = (\xi_j)_{j\in[\![1,2d_y s]\!]} \in \mathbb{R}^{s\times d_y\times 2}$ by

$$h_s(\xi) \;=\; \sum_{t=1}^{s}\sum_{d=1}^{d_y}\left(2\,\xi_{2d_y(t-1)+2d-1} - \left(\xi_{2d_y(t-1)+2d}\right)^2\right) \tag{C.1}$$

whose gradient can be concisely obtained in terms of its subcomponents as

$$\nabla h_s(\xi) \;=\; \bigsqcup_{t=1}^{s}\bigsqcup_{d=1}^{d_y}\begin{pmatrix}2\\ -2\,\xi_{2d_y(t-1)+2d}\end{pmatrix} \tag{C.2}$$

where $\bigsqcup$ denotes the concatenation operator that vertically stacks from top to bottom an ordered collection of vectors.

C.1.2 Discrete observations

Using equation (2.23) of Section 2.3.1.3, the H-score at horizon $s \in [\![1, T]\!]$ can be written as

$$\mathcal{H}_s(M) \;=\; h_s\left((p_{t,d,j})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![-2,2]\!]}\right)$$

where $(p_{t,d,j})_{(t,d,j)\in[\![1,s]\!]\times[\![1,d_y]\!]\times[\![-2,2]\!]}$ is the vector enumerating the $p_{t,d,j}$'s in the lexicographic order of $[\![1,s]\!] \times [\![1,d_y]\!] \times [\![-2,2]\!]$ induced by the natural order on $\mathbb{Z}$, and h_s is a continuously differentiable function defined for all $\xi = (\xi_j)_{j\in[\![1,5d_y s]\!]} \in \mathbb{R}^{s\times d_y\times 5}$ by

$$h_s(\xi) \;=\; \sum_{t=1}^{s}\sum_{d=1}^{d_y}\left(\frac{v_{t,d,2}-v_{t,d,0}}{2\,v_{t,d,1}} - \frac{v_{t,d,0}-v_{t,d,-2}}{2\,v_{t,d,-1}} + \left(\frac{v_{t,d,1}-v_{t,d,-1}}{2\,v_{t,d,0}}\right)^2\right) \tag{C.3}$$

with $v_{t,d,j} = \xi_{5d_y(t-1)+5d+(j-2)}$ for $j \in [\![-2,2]\!]$, and its gradient can be written as

$$\nabla h_s(\xi) = \bigsqcup_{t=1}^{s} \bigsqcup_{d=1}^{d_y} \begin{pmatrix} \frac{1}{2\,v_{t,d,-1}} \\ \frac{v_{t,d,0}-v_{t,d,-2}}{2\,v_{t,d,-1}^2} - \frac{v_{t,d,1}-v_{t,d,-1}}{2\,v_{t,d,0}^2} \\ -\frac{1}{2\,v_{t,d,1}} - \frac{1}{2\,v_{t,d,-1}} - \frac{(v_{t,d,1}-v_{t,d,-1})^2}{2\,v_{t,d,0}^3} \\ -\frac{v_{t,d,2}-v_{t,d,0}}{2\,v_{t,d,1}^2} + \frac{v_{t,d,1}-v_{t,d,-1}}{2\,v_{t,d,0}^2} \\ \frac{1}{2\,v_{t,d,1}} \end{pmatrix} \tag{C.4}$$

or more explicitly

$$\nabla h_s(\xi) = \bigsqcup_{t=1}^{s} \bigsqcup_{d=1}^{d_y} \begin{pmatrix} \frac{1}{2\,\xi_{5d_y(t-1)+5d-3}} \\ \frac{\xi_{5d_y(t-1)+5d-2}-\xi_{5d_y(t-1)+5d-4}}{2\,\xi_{5d_y(t-1)+5d-3}^2} - \frac{\xi_{5d_y(t-1)+5d-1}-\xi_{5d_y(t-1)+5d-3}}{2\,\xi_{5d_y(t-1)+5d-2}^2} \\ -\frac{1}{2\,\xi_{5d_y(t-1)+5d-1}} - \frac{1}{\xi_{5d_y(t-1)+5d-3}} - \frac{(\xi_{5d_y(t-1)+5d-1}-\xi_{5d_y(t-1)+5d-3})^2}{2\,\xi_{5d_y(t-1)+5d-2}^3} \\ -\frac{\xi_{5d_y(t-1)+5d}-\xi_{5d_y(t-1)+5d-2}}{2\,v_{t,d,1}^2} + \frac{\xi_{5d_y(t-1)+5d-1}-\xi_{5d_y(t-1)+5d-3}}{2\,\xi_{5d_y(t-1)+5d-2}^2} \\ \frac{1}{2\,\xi_{5d_y(t-1)+5d-1}} \end{pmatrix}$$

for all $\xi = (\xi_j)_{j\in[\![1,5d_y s]\!]} \in \mathbb{R}^{s\times d_y\times 5}$.

C.2 Assessing the variance of log-evidence estimators

In the context of Section 2.3.4.1, we use $R = 1000$ independent runs of SMC^2 with $N_\theta = 8192$ and $N_x = 128$ to produce R independent estimators $(\log \hat{Z}^{(r)}_{M_2,t}, \log \hat{Z}^{(r)}_{M_3,t})_{(t,r)\in[\![1,T]\!]\times[\![1,R]\!]}$ of the log-evidences of models M_2 and M_3. To assess the variance of the estimators $\log \hat{Z}_{M_2,t}$ and $\log \hat{Z}_{M_3,t}$ at every horizon $t \in [\![1, T]\!]$, we compute the sample variances of $(\log \hat{Z}^{(r)}_{M_2,t})_{r\in[\![1,R]\!]}$ and $(\log \hat{Z}^{(r)}_{M_3,t})_{r\in[\![1,R]\!]}$ across the R replications, for each horizon $t \in [\![1, T]\!]$, as shown in Figure C.1. This plot complements Figure 2.5 by quantifying the spread of each histogram displayed there. From Figure C.1, the variance of $\log \hat{Z}_t$ seems to grow linearly as t increases, which is in line with the expected behavior of SMC estimators for a fixed number of particles (Del Moral et al., 2006; Chopin et al., 2013). There appears to be a peculiar excess in variance at horizon $t = 4$, perhaps due to the seemingly unusual informativeness of the corresponding observation at that particular time (see Figure 1.4 in Chapter 1).

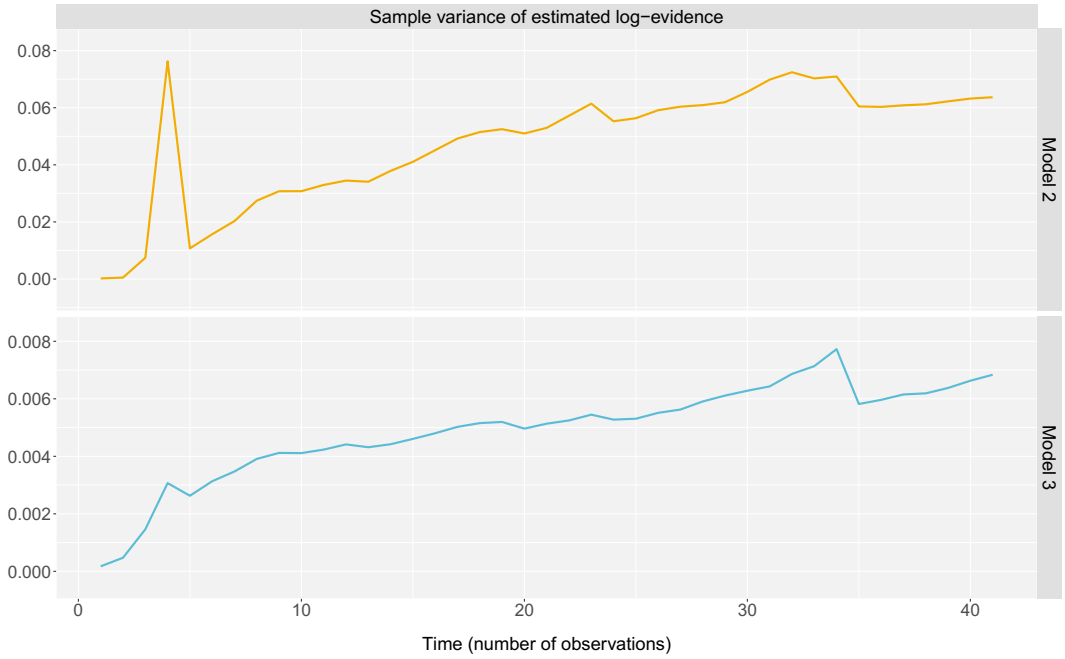

Figure C.1. *Sample variance of the estimated log-evidences* $\log \hat{Z}_t$ *of models* M_2 *(top panel, yellow line) and* M_3 *(bottom panel, blue line) for each time horizon* $t \in [\![1, 41]\!]$*, obtained from* $R = 1000$ *independent runs of SMC*^{2} *with* $N_\theta = 8192$ *and* $N_x = 128$*. See Appendix C.2.*

D

Supplementary material for Chapter 3

Contents

This appendix provides detailed proofs of Lemmas 6 to 10 from Chapter 3. Recall that for all positive semi-definite symmetric matrices $A \in \mathbb{R}^{d\times d}$ and all vectors $x \in \mathbb{R}^d$, we have the inequalities $\lambda_1(A)\|x||^2 \leq x^T Ax \leq \lambda_d(A)\|x\|^2$ and $|x^T Ax| \leq \|A\|\,\|x\|^2$, where $\|A\|$ denotes the spectral norm of A.

D.1 Proof of Lemma 6

Proof. We place ourselves on an intersection of events of $\mathbb{P}_\star$-probability equal to 1, on which Assumptions A2 to A6 simultaneously hold. Let's fix any arbitrary $\varepsilon > 0$. By Assumption A6, there exists $\delta_\varepsilon^\star > 0$ and $N_\varepsilon^\star \in \mathbb{N}$ such that $\theta \mapsto \ell_n(\theta)$ is twice continuously differentiable

on $\{\|\theta-\theta^\star\|<\delta_\varepsilon^\star\}$ and such that, for all $\delta\leq\delta_\varepsilon^\star$ and all $N\geq N_\varepsilon^\star$, we have

$$\sup_{\substack{\|\theta-\theta^\star\|<\delta\\ n>N}}\left\|\frac{\nabla^2\ell_n(\theta)}{n}\right\| \;\leq\; \sup_{\substack{\|\theta-\theta^\star\|<\delta_\varepsilon^\star\\ n>N_\varepsilon^\star}}\left\|\frac{\nabla^2\ell_n(\theta)}{n}\right\| \;<\; \|-J^\star\|+\varepsilon. \tag{D.1}$$

Let's define $\delta_\varepsilon=\min\left(\sqrt{\varepsilon\left[\frac{5}{2}\left(\|-J^\star\|+\varepsilon\right)\right]^{-1}},\ \delta_\varepsilon^\star\right)$, so that $0<\delta_\varepsilon\leq\delta_\varepsilon^\star$.

By Assumption A4, there exists $N_{\delta_\varepsilon}\in\mathbb{N}$ such that $\|\hat\theta_n-\theta^\star\|<\delta_\varepsilon$ for all $n>N_{\delta_\varepsilon}$.

Let's define $N_\varepsilon=\max(N_{\delta_\varepsilon},N_\varepsilon^\star)$.

Since $\theta\mapsto\ell_n(\theta)$ is twice continuously differentiable on $\{\|\theta-\theta^\star\|<\delta_\varepsilon\}\subseteq\{\|\theta-\theta^\star\|<\delta_\varepsilon^\star\}$, the triangle inequality and 2nd order Taylor's theorem with $\nabla\ell_n(\hat\theta_n)=0$ gives us

$$\begin{aligned}\left|\frac{1}{n}\ell_n(\theta)-\frac{1}{n}\ell_n(\theta^\star)\right| \;&\leq\; \left|\frac{1}{n}\ell_n(\theta)-\frac{1}{n}\ell_n(\hat\theta_n)\right|+\left|\frac{1}{n}\ell_n(\hat\theta_n)-\frac{1}{n}\ell_n(\theta^\star)\right|\\ &\leq\; \frac{1}{2}\left|(\theta-\hat\theta_n)^T\frac{\nabla^2\ell_n(\theta_n')}{n}(\theta-\hat\theta_n)\right|+\frac{1}{2}\left|(\theta^\star-\hat\theta_n)^T\frac{\nabla^2\ell_n(\theta_n'')}{n}(\theta^\star-\hat\theta_n)\right|\end{aligned}$$

for all $\theta\in\{\|\theta-\theta^\star\|<\delta_\varepsilon\}$ and all $n>N_\varepsilon$, where we implicitly define $\theta_n'=\gamma'\theta+(1-\gamma')\hat\theta_n$ and $\theta_n''=\gamma''\theta^\star+(1-\gamma'')\hat\theta_n$ for some $(\gamma',\gamma'')\in[0,1]^2$.

Since $\|\theta_n'-\theta^\star\|<\delta_\varepsilon$ and $\|\theta_n''-\theta^\star\|<\delta_\varepsilon$ with $\delta_\varepsilon\leq\delta_\varepsilon^\star$, and $N_\varepsilon>N_\varepsilon^\star$, we use (D.1) to get

$$\begin{aligned}\left|\frac{1}{n}\ell_n(\theta)-\frac{1}{n}\ell_n(\theta^\star)\right| \;&\leq\; \frac{1}{2}\left\|\frac{\nabla^2\ell_n(\theta_n')}{n}\right\|\|\theta-\hat\theta_n\|^2+\frac{1}{2}\left\|\frac{\nabla^2\ell_n(\theta_n'')}{n}\right\|\|\theta^\star-\hat\theta_n\|^2\\ &\leq\; \frac{1}{2}\sup_{\substack{\|\theta-\theta^\star\|<\delta_\varepsilon\\ n>N_\varepsilon}}\left\|\frac{\nabla^2\ell_n(\theta)}{n}\right\|(4\,\delta_\varepsilon^2)+\frac{1}{2}\sup_{\substack{\|\theta-\theta^\star\|<\delta_\varepsilon\\ n>N_\varepsilon}}\left\|\frac{\nabla^2\ell_n(\theta)}{n}\right\|(\delta_\varepsilon^2)\\ &\leq\; \frac{5}{2}\left(\|-J^\star\|+\varepsilon\right)\delta_\varepsilon^2\\ &\leq\; \varepsilon.\end{aligned}$$

This holds for all $\theta\in\{\|\theta-\theta^\star\|<\delta_\varepsilon\}$ and all $n>N_\varepsilon$, so that we finally get

$$\sup_{\substack{\|\theta-\theta^\star\|<\delta_\varepsilon\\ n>N_\varepsilon}}\left|\frac{1}{n}\ell_n(\theta)-\frac{1}{n}\ell_n(\theta^\star)\right|<\varepsilon$$

which concludes the proof of Lemma 6. □

D.2 Proof of Lemma 7

Proof. We place ourselves on an intersection of events of $\mathbb{P}_\star$-probability equal to 1, on which Assumptions A2 to A6 and Lemma 6 simultaneously hold. Let's fix any arbitrary $\varepsilon > 0$. By Lemma 6, there exists $\delta_{\varepsilon/2} > 0$ and $N_{\varepsilon/2} \in \mathbb{N}$ such that

$$\sup_{\substack{\|\theta-\theta^\star\|<\delta_{\varepsilon/2} \\ n>N_{\varepsilon/2}}} \left|\frac{1}{n}\ell_n(\theta) - \frac{1}{n}\ell_n(\theta^\star)\right| < \frac{\varepsilon}{2}. \tag{D.2}$$

By Assumptions A2 and A4, there exists $N_\varepsilon > N_{\varepsilon/2}$ such that, for all $n > N_\varepsilon$, we have both

$$\|\hat{\theta}_n - \theta^\star\| < \delta_{\varepsilon/2} \qquad \text{and} \qquad \left|\frac{1}{n}\ell_n(\theta^\star) - \ell(\theta^\star)\right| < \frac{\varepsilon}{2}. \tag{D.3}$$

Using (D.2) and (D.3), the triangle inequality gives us, for all $n > N_\varepsilon > N_{\varepsilon/2}$,

$$\begin{aligned}
\left|\frac{1}{n}\ell_n(\hat{\theta}_n) - \ell(\theta^\star)\right| &\leq \left|\frac{1}{n}\ell_n(\hat{\theta}_n) - \frac{1}{n}\ell_n(\theta^\star)\right| + \left|\frac{1}{n}\ell_n(\theta^\star) - \ell(\theta^\star)\right| \\
&\leq \sup_{\substack{\|\theta-\theta^\star\|<\delta_{\varepsilon/2} \\ n>N_{\varepsilon/2}}} \left|\frac{1}{n}\ell_n(\theta) - \frac{1}{n}\ell_n(\theta^\star)\right| + \left|\frac{1}{n}\ell_n(\theta^\star) - \ell(\theta^\star)\right| \\
&< \varepsilon
\end{aligned}$$

This holds for all $n > N_\varepsilon$, therefore we get

$$\sup_{n>N_\varepsilon} \left|\frac{1}{n}\ell_n(\hat{\theta}_n) - \ell(\theta^\star)\right| < \varepsilon$$

which concludes the proof of Lemma 7. □

D.3 Proof of Lemma 8

Proof. We place ourselves on an intersection of events of $\mathbb{P}_\star$ probability 1, on which Assumptions A2 to A5 and Lemma 7 simultaneously hold. Let's fix any arbitrary $\delta > 0$.

By consistency of the MLE (Assumption A4), there exists $N_{\delta/2} \in \mathbb{N}$ such that we have $\|\hat{\theta}_n - \theta^\star\| \leq \delta/2$ for all $n > N_{\delta/2}$. By the reverse triangle inequality, if $\|\theta - \hat{\theta}_n\| \geq \delta$, then

$$\|\theta - \theta^\star\| \;\geq\; \|\theta - \hat{\theta}_n\| - \|\hat{\theta}_n - \theta^\star\| \;\geq\; \delta - \delta/2 \;=\; \delta/2.$$

Thus, for all $n > N_{\delta/2}$, we have

$$\{\|\theta - \hat{\theta}_n\| \geq \delta\} \subseteq \{\|\theta - \theta^\star\| \geq \delta/2\}. \tag{D.4}$$

By using (D.4) with Assumption A5, there exists $\eta^\star_{\delta/2} > 0$ and $N_{\eta^\star_{\delta/2}} > N_{\delta/2}$ such that

$$\sup_{\|\theta - \hat{\theta}_n\| \geq \delta} \frac{1}{n}\ell_n(\theta) \;\leq\; \sup_{\|\theta - \theta^\star\| \geq \delta/2} \frac{1}{n}\ell_n(\theta) \;<\; \ell(\theta^\star) - \eta^\star_{\delta/2} \tag{D.5}$$

for all $n > N_{\eta^\star_{\delta/2}}$. By Lemma 7, there exists $N_\delta > N_{\eta^\star_{\delta/2}}$ such that, for all $n > N_\delta$, we have

$$\left|\frac{1}{n}\ell_n(\hat{\theta}_n) - \ell(\theta^\star)\right| < \frac{\eta^\star_{\delta/2}}{2}$$

which implies

$$\frac{1}{n}\ell_n(\hat{\theta}_n) \;>\; \ell(\theta^\star) - \frac{\eta^\star_{\delta/2}}{2}. \tag{D.6}$$

Let's fix $\eta_\delta = \frac{\eta^\star_{\delta/2}}{2} > 0$. Then, by combining (D.5) and (D.6), we obtain, for all $n > N_\delta$,

$$\sup_{\|\theta - \hat{\theta}_n\| \geq \delta} \frac{1}{n}\ell_n(\theta) \;<\; \frac{1}{n}\ell(\hat{\theta}_n) - \eta_\delta$$

which concludes the proof of Lemma 8. □

D.4 Proof of Lemma 10

Proof. We place ourselves on an intersection of events of $\mathbb{P}_\star$ probability 1, on which Assumptions A4 and A6 simultaneously hold. Let's fix any arbitrary $\varepsilon > 0$.

By Assumption A6, there exists $\delta^\star_\varepsilon > 0$ and $N^\star_\varepsilon \in \mathbb{N}$ such that

$$\sup_{\substack{\|\theta - \theta^\star\| < \delta^\star_\varepsilon \\ n > N^\star_\varepsilon}} \left\|\frac{1}{n}\nabla^2\ell_n(\theta) + J^\star\right\| < \varepsilon. \tag{D.7}$$

By consistency of the MLE (Assumption A4), there exists $N_\varepsilon \geq N_\varepsilon^\star$ such that $\|\hat{\theta}_n - \theta^\star\| < \delta_\varepsilon^\star$ for all $n > N_\epsilon$. Combining this with (D.7) gives us, for all $n > N_\varepsilon \geq N_\varepsilon^\star$,

$$\left|\frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n) + J^\star\right| \leq \sup_{\substack{\|\theta-\theta^\star\|<\delta_\varepsilon^\star \\ n>N_\varepsilon^\star}} \left\|\frac{1}{n}\nabla^2\ell_n(\theta) + J^\star\right\| < \varepsilon.$$

Since this holds for all $n > N_\varepsilon$, we finally get

$$\sup_{n>N_\varepsilon} \left|\frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n) + J^\star\right| < \varepsilon.$$

which concludes the proof of Lemma 10. □

Bibliography

Alqallaf, F. and Gustafson, P. (2001). On cross-validation of Bayesian models. *Canadian Journal of Statistics*, 29 (2), 333–340.

Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 72 (3), 269–342.

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

Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B*, 63 (2), 167—241.

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B*, 64 (2), 253–280.

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44 (3-4), 533–534.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37 (6), 1554–1563.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91 (433), 109–122.

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: introduction and comparison. *Model Selection, IMS Lecture Notes – Monograph Series*, 68, 135–207.

Berger, J. O., Pericchi, L. R. and Varshavsky, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya A: The Indian Journal of Statistics*, 60, 307–321.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. John Wiley & Sons.

Bhattacharya, P. (1967). Estimation of a probability density function and its derivatives. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 373–382.

Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, 40 (1), 206–237.

Bickel, P. J., Ritov, Y. and Ryden, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26 (4), 1614–1635.

Billingsley, P. (1968). *Convergence of probability measures*. Wiley Series in Probability and Statistics.

Billingsley, P. (1995). *Probability and measure*. Wiley Series in Probability and Mathematical Statistics.

Bobkov, S. G., Gozlan, N., Roberto, C. and Samson, P.-M. (2014). Bounds on the deficit in the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 267 (11), 4110–4138.

Bontemps, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *The Annals of Statistics*, 39 (5), 2557–2584.

Bornn, L., Doucet, A. and Gottardo, R. (2010). An efficient computational approach for prior sensitivity analysis and cross-validation. *Canadian Journal of Statistics*, 38 (1), 47–64.

Boucheron, S. and Gassiat, E. (2009). A Bernstein-von Mises theorem for discrete probability distributions. *Electronic journal of Statistics*, 3, 114–148.

Bretó, C., He, D., Ionides, E. L. and King, A. A. (2009). Time series analysis via mechanistic models. *The Annals of Applied Statistics*, pp. 319–348.

Brooks, S., Gelman, A., Jones, G. and Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC press.

Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Series in Statistics.

Castillo, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152 (1-2), 53–99.

Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics*, 41 (4), 1999–2028.

Castillo, I. and Nickl, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics*, 42 (5), 1941–1969.

Castillo, I. and Rousseau, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43 (6), 2353–2383.

Caughley, G., Shepherd, N. and Short, J. (1987). *Kangaroos, their ecology and management in the sheep rangelands of Australia*. Cambridge University Press.

Cérou, F., Del Moral, P. and Guyader, A. (2011). A nonasymptotic theorem for unnormalized Feynman-Kac particle models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 47 (3), 629–649.

Chib, S. and Kuffner, T. A. (2016). Bayes factor consistency. *Preprint*, arXiv:1607.00292.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89 (3), 539–552.

Chopin, N., Jacob, P. E. and Papaspiliopoulos, O. (2013). SMC^2: an efficient algorithm for sequential analysis of state-space models. *Journal of the Royal Statistical Society: Series B*, 75 (3), 397–426.

Chopin, N., Ridgway, J., Gerber, M. and Papaspiliopoulos, O. (2015). Towards automatic calibration of the number of state particles within the SMC^2 algorithm. *Preprint*, arXiv:1506.00570.

Chopin, N. et al. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics*, 32 (6), 2385–2411.

Dawid, A. P. (1984). The prequential approach. *Journal of the Royal Statistical Society: Series A*, 147 (2), 278–292.

Dawid, A. P. (2011). Posterior model probabilities. *Handbook of the Philosophy of Science*, 7, 607–630.

Dawid, A. P. and Lauritzen, S. L. (2005). The geometry of decision theory. *In Proceedings of the Second International Symposium on Information Geometry and Its Applications*, pp. 22–28.

Dawid, A. P., Lauritzen, S. and Parry, M. (2012). Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40 (1), 593–608.

Dawid, A. P. and Musio, M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, 10 (2), 479–499.

Dawid, A. P., Musio, M. and Columbu, S. (2017). A note on Bayesian model selection for discrete data using proper scoring rules. *Statistics and Probability Letters*, 129, 101–106.

Dawid, A. P., Musio, M. and Ventura, L. (2016). Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43 (1), 123–138.

De Blasi, P. and Hjort, N. L. (2009). The Bernstein–von Mises theorem in semiparametric competing risks models. *Journal of Statistical Planning and Inference*, 139 (7), 2316–2328.

De Gunst, M. C. M. and Shcherbakova, O. (2008). Asymptotic behavior of Bayes estimators for hidden Markov models with application to ion channels. *Mathematical Methods of Statistics*, 17 (4), 342–356.

Del Moral, P. (2004). *Feynman-Kac formulae*. Springer. New York.

Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68 (3), 411—436.

Del Moral, P., Doucet, A. and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22 (5), 1009–1020.

Dennis, B. and Costantino, R. (1988). Analysis of steady-state populations with the gamma-abundance model: application to tribolium. *Ecology*, 69, 1200–1213.

Douc, R. and Cappé, O. (2005). Comparison of resampling schemes for particle filtering. *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pp. 64–69.

Douc, R. and Matias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, 7 (3), 381–420.

Douc, R. and Moulines, E. (2012). Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *The Annals of Statistics*, 40 (5), 2697–2732.

Douc, R., Moulines, E., Olsson, J. and Van Handel, R. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 39 (1), 474–513.

Douc, R., Moulines, E. and Ryden, T. (2004). Asymptotic properties of the maximum likelihood estimator in the autoregressive models with Markov regime. *The Annals of Statistics*, 32, 2254-2304.

Douc, R., Moulines, E. and Stoffer, D. (2014). *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. 1st edn. Chapman and Hall/CRC.

Douc, R., Olsson, J. and Roueff, F. (2019). Posterior consistency for partially observed Markov models. *Stochastic Processes and their Applications*, in print.

Doucet, A., Pitt, M. K., Deligiannidis, G. and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102 (2), 295–313.

Duan, J.-C. and Fulop, A. (2015). Density-tempered marginalized sequential Monte Carlo samplers. *Journal of Business & Economic Statistics*, 33 (2), 192–202.

Ehm, W. and Gneiting, T. (2012). Local proper scoring rules of order two. *The Annals of Statistics*, 40 (1), 609–637.

Fearnhead, P., Papaspiliopoulos, O., Roberts, G. O. and Stuart, A. (2010). Random-weight particle filtering of continuous time processes. *Journal of the Royal Statistical Society: Series B*, 72 (4), 497–512.

Fearnhead, P. and Taylor, B. M. (2013). An adaptive sequential Monte Carlo sampler. *Bayesian Analysis*, 8 (2), 411–438.

Freedman, D. (1999). Wald lecture: On the Bernstein–von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27 (4), 1119–1141.

Fulop, A. and Li, J. (2013). Efficient learning via simulation: A marginalized resample-move approach. *Journal of Econometrics*, 176 (2), 146–161.

Gassiat, E. and Rousseau, J. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli*, 20 (4), 2039–2075.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B*, 56 (3), 501–514.

Gerber, M., Chopin, N. and Whiteley, N. (2017). Negative association, ordering and convergence of resampling methods. *Preprint*, arXiv:1707.01845.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. Springer Science & Business Media.

Glynn, P. W. and Heidelberger, P. (1990). Bias properties of budget constrained simulations. *Operations Research*, 38 (5), 801–814.

Glynn, P. W. and Whitt, W. (1992*a*). The asymptotic efficiency of simulation estimators. *Operations research*, 40 (3), 505–520.

Glynn, P. W. and Whitt, W. (1992*b*). The asymptotic validity of sequential stopping rules for stochastic simulations. *The Annals of Applied Probability*, 2 (1), 180–198.

Golightly, A. and Kypraios, T. (2017). Efficient SMC^2 schemes for stochastic kinetic models. *Preprint*, arXiv:1704.02791.

Gordon, N., Salmond, J. and Smith, A. (1993). A novel approach to non-linear/non-Gaussian Bayesian state estimation. *IEE Proceedings on Radar and Signal Processing*, 140, 107–113.

Green, P. J., Łatuszyński, K., Pereyra, M. and Robert, C. P. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25 (4), 835–862.

Hardle, W., Marron, J. and Wand, M. (1990). Bandwidth choice for density derivatives. *Journal of the Royal Statistical Society: Series B*, pp. 223–232.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 (1), 97–109.

Hendrickson, A. D. and Buehler, R. J. (1971). Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42 (6), 1916–1921.

Heng, J. and Jacob, P. E. (2019). Unbiased Hamiltonian Monte Carlo with couplings. To appear in: *Biometrika*, DOI: 10.1093/biomet/asy074.

Holmes, C. and Walker, S. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104 (2), 497–503.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1 (1), 221–233.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6, 695–709.

Jacob, P. E., O'Leary, J. and Atchadé, Y. F. (2017). Unbiased Markov chain Monte Carlo with couplings. *Preprint*, arXiv:1708.03625v4.

Jeffreys, H. (1961). *The theory of probability*. Oxford University Press.

Jewson, J., Smith, J. Q. and Holmes, C. (2018). Principled Bayesian minimum divergence inference. *Preprint*, arXiv:1802.09411.

Johnstone, I. M. (2010). High dimensional Bernstein-von Mises: simple examples. *Institute of Mathematical Statistics collections*, 6, 87.

Kamary, K., Mengersen, K., Robert, C. P. and Rousseau, J. (2014). Testing hypotheses via a mixture estimation model. *Preprint*, arXiv:1412.2044v2.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90 (430), 773–795.

Kaufmann, E., Cappé, O. and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17 (1), 1–42.

Key, J. T., Pericchi, L. R. and Smith, A. F. (1999). Bayesian model choice: what and why. *Bayesian statistics*, 6, 343–370.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pp. 887–906.

Kim, Y. (2006). The Bernstein–von Mises theorem for the proportional hazard model. *The Annals of Statistics*, 34 (4), 1678–1700.

Kim, Y. and Lee, J. (2004). A Bernstein–von Mises theorem in the nonparametric right-censoring model. *The Annals of Statistics*, 32 (4), 1492–1512.

Knape, J. and de Valpine, P. (2012). Fitting complex population models by combining particle filters with Markov chain Monte Carlo. *Ecology*, 93 (2), 256–263.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related results. *University of California publications in Statistics*, 1, 277–330.

Le Cam, L. and Yang, G. (2012). *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media.

Lee, A. and Whiteley, N. (2018). Variance estimation in the particle filter. *Biometrika*, 105 (3), 609–625.

Lee, J. and MacEachern, S. N. (2011). Consistency of Bayes estimators without the assumption that the model is correct. *Journal of Statistical Planning and Inference*, 141 (2), 748–757.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40 (1), 127–143.

Lijoi, A., Prünster, I. and Walker, S. G. (2007). Bayesian consistency for stationary models. *Econometric Theory*, 23 (4), 749–759.

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93 (443), 1032–1044.

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, 42 (9), 654–655.

Middleton, L., Deligiannidis, G., Doucet, A. and Jacob, P. E. (2018). Unbiased Markov chain Monte Carlo for intractable target distributions. *Preprint*, arXiv:1807.08691.

Middleton, L., Deligiannidis, G., Doucet, A. and Jacob, P. E. (2019). Unbiased smoothing using particle independent Metropolis-Hastings. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS*, 89.

Moreno, E., Girón, F. J. and Casella, G. (2010). Consistency of objective Bayes factors as the dimension grows. *The Annals of Statistics*, 38 (4), 1937–1952.

Murray, L. M., Lee, A. and Jacob, P. E. (2016). Parallel resampling in the particle filter. *Journal of Computational and Graphical Statistics*, 25 (3), 789–805.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11 (2), 125–139.

O'Hagan, A. (1995). Fractional Bayes factor for model comparison. *Journal of the Royal Statistical Society: Series B*, 57 (1), 99–138.

Parry, M., Dawid, A. P. and Lauritzen, S. (2012). Proper local scoring rules. *The Annals of Statistics*, 40 (1), 561–592.

Perlman, M. D. (1972). On the strong consistency of approximate maximum likelihood estimators. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Theory of Statistics.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94 (446), 590–599.

Rischard, M., Jacob, P. E. and Pillai, N. (2018). Unbiased estimation of log normalizing constants with applications to Bayesian cross-validation. *Preprint*, arXiv:1810.01382.

Rivoirard, V. and Rousseau, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *The Annals of Statistics*, 40 (3), 1489–1523.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Robert, C. and Casella, G. (1999). *Monte Carlo statistical methods*. 1st edn. Springer Science & Business Media.

Rousseau, J. and Taeryon, C. (2012). Bayes factor consistency in regression problems. *HAL, archives ouvertes*, hal-00767469.

Rudin, W. (1964). *Principles of mathematical analysis*. McGraw–Hill.

Santis, F. and Spezzaferri, F. (1999). Methods for default and robust Bayesian model comparison: the fractional Bayes factor approach. *International Statistical Review*, 67 (3), 267–286.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6 (2), 461–464.

Shalizi, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, 3, 1039–1074.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Vol. 3. Cambridge University Press.

Vergé, C., Dubarry, C., Del Moral, P. and Moulines, E. (2015). On parallel implementation of sequential Monte Carlo methods: the island particle model. *Statistics and Computing*, 25 (2), 243–260.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pp. 307–333.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20 (4), 595–601.

Walker, S. G. (2004). Modern Bayesian asymptotics. *Statistical Science*, pp. 111–117.

Walker, S. G. (2013). Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, 143 (10), 1621–1633.

Walker, S. G. (2016). Bayesian information in an experiment and the Fisher information distance. *Statistics & Probability Letters*, 112, 5–9.

Wang, Y. and Blei, D. M. (2018). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, pp. 1–15.

Whiteley, N. (2013). Stability properties of some particle filters. *The Annals of Applied Probability*, 23 (6), 2500–2537.

Whiteley, N., Lee, A. and Heine, K. (2016). On the role of interaction in sequential Monte Carlo algorithms. *Bernoulli*, 22 (1), 494–529.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9 (1), 60–62.

Yonekura, S., Beskos, A. and Singh, S. (2018). Asymptotic analysis of model selection criteria for general hidden Markov models. *Preprint*, arXiv:1811.11834.

Zhou, Q. and Guan, Y. (2018). On the null distribution of Bayes factors in linear regression. *Journal of the American Statistical Association*, 113 (523), 1362–1371.

Zhou, Y., Johansen, A. M. and Aston, J. A. (2016). Toward automatic model comparison: an adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25 (3), 701–726.