



Cost free hyper-parameter selection/averaging for Bayesian inverse problems with vanilla and Rao-Blackwellized SMC samplers

Alessandro Viani¹ · Adam M. Johansen² · Alberto Sorrentino¹

Received: 22 December 2022 / Accepted: 2 September 2023
© The Author(s) 2023

Abstract

In Bayesian inverse problems, one aims at characterizing the posterior distribution of a set of unknowns, given indirect measurements. For non-linear/non-Gaussian problems, analytic solutions are seldom available: Sequential Monte Carlo samplers offer a powerful tool for approximating complex posteriors, by constructing an auxiliary sequence of densities that smoothly reaches the posterior. Often the posterior depends on a scalar hyper-parameter, for which limited prior information is available. In this work, we show that properly designed Sequential Monte Carlo (SMC) samplers naturally provide an approximation of the marginal likelihood associated with this hyper-parameter for free, i.e. at a negligible additional computational cost. The proposed method proceeds by constructing the auxiliary sequence of distributions in such a way that each of them can be interpreted as a posterior distribution corresponding to a different value of the hyper-parameter. This can be exploited to perform selection of the hyper-parameter in *Empirical Bayes* (EB) approaches, as well as averaging across values of the hyper-parameter according to some hyper-prior distribution in *Fully Bayesian* (FB) approaches. For FB approaches, the proposed method has the further benefit of allowing prior sensitivity analysis at a negligible computational cost. In addition, the proposed method exploits particles at all the (relevant) iterations, thus alleviating one of the known limitations of SMC samplers, i.e. the fact that all samples at intermediate iterations are typically discarded. We show numerical results for two distinct cases where the hyper-parameter affects only the likelihood: a toy example, where an SMC sampler is used to approximate the full posterior distribution; and a brain imaging example, where a Rao-Blackwellized SMC sampler is used to approximate the posterior distribution of a subset of parameters in a conditionally linear Gaussian model.

Keywords Bayesian inverse problems · Hyper-parameter estimation · Sequential Monte Carlo samplers · Rao-Blackwellization · Empirical Bayes · Fully Bayesian

1 Introduction

In Bayesian inverse problems, one is interested in approximating the posterior distribution of a set of unobservable quantities, x , conditioned on indirect measurements, y (Stuart 2010). Often the posterior distribution depends on a

scalar hyper-parameter, $\theta \in \Theta \subseteq \mathbb{R}$, e.g. the noise variance: one can either perform hyper-parameter selection with an Empirical Bayes (EB) approach targeting the conditional posterior $p^{\theta^*}(x | y)$ with the hyper-parameter set to the value which maximizes the marginal likelihood, $\theta^* := \arg \max_{\theta \in \Theta} \{p^\theta(y)\}$, sometimes termed type-II maximum likelihood (Good 1965), or else marginalize out the hyper-parameter through a Fully Bayesian (FB) approach, targeting the posterior $p(x | y)$. However, both approaches often result in costly procedures.

One relatively common tool for approximating posterior distributions arising in Bayesian inverse problems are Sequential Monte Carlo (SMC) samplers (Del Moral et al. 2006). SMC samplers construct an artificial sequence of distributions such that the first one can be readily sampled from and the last one coincides with the distribution of interest; a

✉ Alessandro Viani
viani@dima.unige.it

Adam M. Johansen
a.m.johansen@warwick.ac.uk

Alberto Sorrentino
sorrentino@dima.unige.it

¹ Dipartimento di Matematica, Università di Genova, 16146 Genova, Italy

² Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

set of particles is drawn from the first density, and evolves gradually to approximate each distribution in the sequence.

Compared to standard alternatives, SMC samplers provide distinct advantages that make them a suitable choice when analytical solutions are not available. In contrast to Variational Bayesian inference (Fox and Roberts 2012), they need not approximate the posterior distribution with a member of a specific class of functions, often chosen for computational convenience, thus avoiding the problem of misspecification. In comparison to MCMC (Gilks et al. 1995), they feature simultaneous exploration of multiple areas of the state space through the multiple particles, thus improving the chances of fully exploring multi-modal posterior distributions and have often been found to give better performance at given computational cost—see, for example, (Del Moral et al. 2006, Section 4.2) and references therein.

In most implementations of SMC samplers for Bayesian inverse problems, the samples obtained at intermediate iterations are discarded, because intermediate iterations are only used to facilitate the approximation of the target distribution. Not directly using these samples, except perhaps to estimate a normalizing constant, seemingly results in a substantial waste of computational resources. Indeed, we have recently witnessed a growing number of studies that attempt to exploit/recycle particles from previous iterations in the final estimates (Gramacy et al. 2010; Drovandi et al. 2019; Le Thu et al. 2016; Dau and Chopin 2022). Gramacy et al. (2010) propose to recycle particles at different iterations by considering a weighted sum of all the approximated distributions in order to maximise the Effective Sample Size (ESS). Alternatively, Le Thu et al. (2016) propose to combine particles from past SMC samplers iterations considering the so called *Deterministic Mixture Weight estimator*; a solution derived to combine weighted particles drawn from different proposal distributions. Recently Drovandi et al. (2019) developed a method which allows the samples from each generation of the algorithm to be used to approximate integrals over a part of the state space.

In this work we show that, for a large class of hierarchical Bayesian inverse problems featuring scalar hyper-parameters, the intermediate iterations of properly designed SMC samplers can be used to perform selection of the hyper-parameter and/or averaging with respect to it, making EB/FB approaches feasible. All of this has only a negligible additional computational cost and, in the case of averaging, it also entails recycling of the particles at intermediate iterations, thus reducing the typical waste of computational resources.

The key idea underlying the proposed method is to define the auxiliary sequence of distributions in such a way that each distribution is a posterior distribution conditioned on a different value of the hyper-parameter. Such construction turns out to be extremely simple under certain conditions, for instance when the hyper-parameter appears only in the

likelihood and the likelihood belongs to the natural exponential family; under other circumstances, finding the right sequence can be more challenging. Given the sequence, the estimate of the normalizing constant, naturally produced by SMC samplers, corresponds to an estimate of the evidence for the specific value of the hyper-parameter, which then allows maximum likelihood or Bayesian inference on the hyper-parameter.

We provide the right tempering sequence for two different models largely used in inverse problems:

- when the likelihood belongs to the Natural Exponential Family (NEF): here the tempering sequence obtained by raising the likelihood to a growing power between zero and one results in a proper sequence of densities that can be interpreted as posterior distributions;
- when the conditional posterior for a subset of variables x_1 can be analytically computed, and an SMC sampler is used only to approximate the posterior on the remaining variables x_2 . For this class of models, which includes among others Conditionally Linear Gaussian (CLG) models, the auxiliary distribution sequence devised for the first case does not, in general, have the desired properties, therefore we devise alternative sequences that can be used fruitfully in two special sub-cases.

The most straightforward application of the proposed method is the context of additive Gaussian noise inverse problems; here the interest is in the estimation of the joint posterior distribution for the state variables and the noise variance or the posterior distribution for the state variables conditioned on the estimated value for the noise variance.

As a first examples we consider the problem of recovering the mean of a Gaussian distribution from noisy observations, showing that the proposed approach performs as well as alternative approaches but with significant advantages in computational time. Then we show numerical results for a real world problem encountered in source analysis of Magneto/Electro-Encephalography data, in this case we show that the proposed approach provides reliable results and a substantial reduction of computational cost with respect to alternative approaches.

2 Motivating example: source estimation in magneto/ electro-encephalography

Magneto-/Electro-Encephalography (M/E-EG) are two non-invasive medical imaging techniques that record the magnetic/electric field on the scalp; from these recordings, it is possible to estimate the underlying neural currents (Hämäläinen et al. 1993). Using the *dipolar* assumption, this problem

consists of estimating an unknown number of point sources, called *dipoles*, each one defined by two quantities:

- a location in the brain volume, conveniently represented as the index r of a cell of a discretized brain (or *voxel*); dipole location is assumed to be fixed in time;
- a 3-D vector q representing orientation and intensity of the neural current at the specified voxel, and changing dynamically in time.

The inference problem can be formalized as

$$y(t) = \sum_{i=1}^d G(r_i)q_i(t) + \varepsilon(t) \tag{1a}$$

$$\varepsilon(t) \sim \mathcal{N}(0, \theta^2 \Sigma) \tag{1b}$$

where: $t = 1, \dots, T$ is a time index; $y(t)$ is an array containing the data recorded by all M/E-EG sensors at time t ; d is the (unknown) number of dipoles; $G(r_i)$ is the so called lead-field matrix, representing the magnetic/electric field generated by a unitary dipole located at r_i ; $\varepsilon(t)$ is additive Gaussian noise whose (spatial) covariance matrix Σ is known up to a scale factor θ . We observe that the lead-field matrix entries are non-linear functions of the source locations r_i , therefore linear inversion models are not directly applicable to this inverse problem.

This model was originally adopted in Sorrentino et al. (2013, 2014), where all unknown parameters were sampled with an SMC sampler, leading to high computational cost for long time series; in Sommariva and Sorrentino (2014) a Rao-Blackwellized version was presented that imposed a Gaussian prior on the q variables and exploited the CLG structure, allowing to treat long time series with reduced computational cost. Finally, in Viani et al. (2021) a hierarchical model was presented that overcomes the limitations of the Gaussian prior by using a hyper-prior on the prior variance, thus substantially reducing the dependence on this hyper-parameter. Defining $\mathbf{y} := (y(1), \dots, y(T))$ and $\mathbf{q}_{1:d} := (q_{1:d}(1), \dots, q_{1:d}(T))$, the posterior distribution decomposes as:

$$p^\theta(d, r_{1:d}, \mathbf{q}_{1:d}, \lambda \mid \mathbf{y}) = p^\theta(\mathbf{q}_{1:d} \mid \mathbf{y}, d, r_{1:d}, \lambda) p^\theta(d, r_{1:d}, \lambda \mid \mathbf{y}) \tag{2}$$

where the conditional posterior $p^\theta(\mathbf{q}_{1:d} \mid \mathbf{y}, d, r_{1:d}, \lambda)$ can be computed analytically, and only the second factor on the right hand side of (2) has to be approximated via Monte Carlo. Importantly, there remains a dependence on the hyper-parameter θ , namely the overall noise level, whose value has to be estimated.

3 SMC samplers for Bayesian inverse problems

In this section we provide a brief summary of a class of SMC samplers that are often used for the approximation of posterior distributions in Bayesian inference problems. Notice that SMC samplers can be applied in more general situations, not analyzed in this paper; for further details on general SMC samplers algorithms the reader is referred to Del Moral et al. (2006, 2007).

Consider a Bayesian inference problem where the aim is to approximate the posterior distribution

$$p(x \mid y) = \frac{p(x)p(y \mid x)}{p(y)} \tag{3}$$

where y represents the data and x the unknown parameters. The posterior distribution is often a complex distribution in a possibly high-dimensional space and is typically difficult to sample from directly.

SMC samplers provide an effective way to sample such complex distributions, and can be briefly summarized as follows.

The first step is to define a sequence of intermediate densities:

$$\{p_t(x \mid y)\}_{t=0}^T, \tag{4a}$$

$$p_T(x \mid y) = p(x \mid y), \tag{4b}$$

$$p_t(x \mid y) \simeq p_{t+1}(x \mid y), \tag{4c}$$

that “smoothly” transition from an easy-to-sample initial density p_0 to the posterior density p_T . Condition (4c) is required in order to guarantee a smooth transition toward the target density and hence to allow a good approximation of p_{t+1} to be obtained from the corresponding approximation of p_t .

A natural, but not mandatory, choice in Bayesian inference is to reach the posterior density by starting from the prior and increasing the power of the likelihood using the so called geometric bridge, or tempering path (Syed et al. 2021; Chopin and Papaspiliopoulos 2020; Bernton et al. 2019; Neal 2001):

$$p_t(x \mid y) \propto p(x)p(y \mid x)^{\alpha_t}, \tag{5a}$$

$$0 = \alpha_0 < \alpha_1 < \dots < \alpha_T = 1. \tag{5b}$$

Once the sequence of distributions has been selected, SMC samplers work as follows (for convenience we assume that the prior can be sampled from directly although this is not necessary):

- sample a set of N weighted particles $\{\mathbf{x}^{(0)}; \mathbf{W}^{(0)}\}$ from the initial distribution p_0 with $\mathbf{W}^{(0)} \equiv 1/N$.

- for $t = 1, \dots, T$:
 1. perform one, or more, Markov Chain Monte Carlo (MCMC) step/s; such as Metropolis Hastings step/s
 2. perform an Importance Sampling (IS) step from the current distribution p_{t-1} to the next distribution p_t updating the un-normalized importance weights and normalizing them using the relations

$$\mathbf{w}_i^{(t)} = \mathbf{w}_i^{(t-1)} \frac{\gamma_t(\mathbf{x}_i^{(t-1)} | y)}{\gamma_{t-1}(\mathbf{x}_i^{(t-1)} | y)}, \tag{6a}$$

$$\mathbf{W}_i^{(t)} := \frac{\mathbf{w}_i^{(t)}}{\sum_{n=1}^N \mathbf{w}_n^{(t)}}, \tag{6b}$$

for $i = 1, \dots, N$; where γ_t represents the un-normalized t -th distribution of the sequence (4a); in the case of (5a), for example, $\gamma_t(x | y) = p(x)p(y | x)^{\alpha_t}$.

At this point one obtains an approximation of the t -th distribution of the sequence as:

$$\hat{p}_t(x | y) = \sum_{n=1}^N \mathbf{W}_n^{(t)} \delta_{x_n^{(t)}}(x). \tag{7}$$

In this step one also obtains an estimator of the normalizing constant of the distribution p_t , crucial for model selection in general and for the proposed method in particular. It can be easily evaluated; for simplicity, assuming that resampling occurs at every step, as the product over time of the average of the un-normalized importance weights at each time:

$$\hat{p}_t(y) = \prod_{s=1}^t \frac{1}{N} \sum_{n=1}^N \mathbf{w}_n^{(s)}. \tag{8}$$

the expression in the case that resampling is conducted adaptively is the corresponding product over resampling times of the average of the weights accumulated since the last resampling time (see, e.g., (Guarniero et al. 2017, p. 1641) for an explicit expression).

3. perform a resampling step to avoid degeneracy of the importance weights (Douc et al. 2005; Gerber et al. 2019). A widely used strategy is to perform resampling whenever the Effective Sample Size (ESS) (see, e.g. Liu 2008) is under a fixed threshold. If resampling is performed, replace $\{\mathbf{x}^{(t)}\}$ with the collection of resampled particles and set $\mathbf{w}_i^{(t)} = 1$ and $\mathbf{W}_i^{(t)} = 1/N$ for each i .

One important property of SMC samplers comes from Eqs. (6a)–(6b) which allow the evaluation of the importance

weights at time t using only the particles at the previous step. This allows the sequence of steps described in the algorithm to be carried out in a different order, with step 2 and 3 being carried out pre-emptively. By conducting resampling in this way before the mutation step, a more diverse collection of particle values is obtained at no additional cost. This modification further allows an adaptive choice of the actual sequence of densities, as defined in (4a), through an online selection of the next exponent (Del Moral et al. 2012; Sorrentino et al. 2014).

4 Selection/averaging of the hyper-parameter

Let $\Theta \subseteq \mathbb{R}$ and consider a Bayesian inverse problem depending on a hyper-parameter $\theta \in \Theta$. We are now going to show how an SMC sampler can be used both to select a specific value for the hyper-parameter and/or to approximate the joint posterior distribution $p(x, \theta | y)$ at no additional cost with respect to the SMC sampler that approximates the conditional posterior $p^\theta(x | y)$.

The key idea underlying the proposed method is to construct an SMC sampler whose target distribution is $p^{\theta^*}(x | y)$ for some value $\theta^* \in \Theta$, and whose intermediate distributions are posterior distributions corresponding to different values of the hyper-parameter for a set of values $\Theta_{0:T} := \{\theta \in \Theta : \theta = \theta(t); t = 0, \dots, T\}$

$$p_t^{\theta^*}(x | y) = p^{\theta(t)}(x | y) = \frac{p^{\theta(t)}(y | x)p^{\theta(t)}(x)}{p^{\theta(t)}(y)}. \tag{9}$$

Given the sequence above, one can estimate pointwise the evidence for the hyper-parameter $p^\theta(y)$ for $\theta \in \Theta_{0:T}$ through the Importance Sampling step within SMC sampler (8). Under regularity assumptions for $p^\theta(y)$ w.r.t. θ one can interpolate this finite set of values to obtain a smooth approximation of the evidence and, assuming the availability of a hyper-prior $p(\theta)$, that we assume to be negligible outside a compact set $[\theta_{\min}, \theta_{\max}]$, an approximation of the marginal posterior $\hat{p}(\theta | y)$.

For an EB approach, one can first find the mode of the interpolating function properly weighted

$$\bar{\theta} = \arg \max_{\theta \in [\theta_{\min}, \theta_{\max}]} \{\hat{p}(\theta | y)\}, \tag{10}$$

where we assume that the range of $\Theta_{0:T}$ contains θ^* , θ_{\min} and θ_{\max} . This can be done numerically by binary search, using importance sampling to estimate the marginal likelihood of values of θ between those in $\Theta_{0:T}$. We can then apply importance sampling to obtain an approximation of $p^{\bar{\theta}}(x | y)$.

In order to avoid degeneration of importance weights, one should do importance sampling from $p^{\theta(\bar{t})}(x | y)$, where

$\theta(\bar{t})$ is the closest value to $\bar{\theta}$ such that the support and tails of $p^{\theta(\bar{t})}(x | y)$ are larger and heavier, respectively, than those of $p^{\bar{\theta}}(x | y)$; for instance, assuming that $\{\theta(t)\}_{t=0,\dots,T}$ is a decreasing sequence, and that the distributions tails become lighter as θ becomes smaller, we shall select the iteration

$$\bar{t} == \max\{t : \theta(t) > \bar{\theta}\}. \tag{11}$$

For a FB approach one obtains an approximation of the posterior

$$p(\theta | y) \propto p^\theta(y)p(\theta) \tag{12}$$

for $\theta \in \Theta_{0:T}$, allowing to compute estimates such as the posterior mean or mode for the hyper-parameter.

In addition, it is possible to approximate the marginal posterior of the parameters

$$\begin{aligned} p(x | y) &= \int p(x, \theta | y)d\theta \\ &= \int p^\theta(x | y)p(\theta | y)d\theta \\ &\propto \int p^\theta(x | y)p^\theta(y)p(\theta)d\theta. \end{aligned} \tag{13}$$

taking into account uncertainty on parameters deriving from uncertainty on the hyper-parameter. This can be done by considering all particles at all iterations and re-weighting them

$$\begin{aligned} \hat{p}(x | y) &= \sum_{t=0}^T \sum_{n=1}^N \left\{ \mathbf{w}_n^{(t)} \delta_x(x_n^{(t)}) \hat{p}^{\theta^{(t)}}(y) \right. \\ &\quad \left. \times p(\theta^{(t)})g^{(t)}(\Theta_{0:T}) \right\} \end{aligned} \tag{14}$$

where $g^{(t)}$ is a function representing the interpolation weights.

For example, in the case of a standard quadrature method such as the trapezoidal rule we get

$$g_t(\Theta_{0:T}) = \begin{cases} \|\theta_1 - \theta_0\|/2 & t = 0 \\ \|\theta_{t+1} - \theta_{t-1}\|/2 & 1 < t < T \\ \|\theta_T - \theta_{T-1}\|/2 & t = T \end{cases} \tag{15}$$

but of course more sophisticated options are available (Zhou et al. 2016).

The additional computational cost required for calculating (12)–(14) is negligible compared to that of the approximation of $p^{\theta^*}(x | y)$ directly with an SMC sampler employing likelihood tempering.

Moreover, the proposed FB approach has the advantage of making use of particles at all iterations, thus avoiding the usual waste of computational resources.

As a last point we remark that, in the FB case, it is possible to modify the hyper-prior without re-running the SMC sampler, provided that its support does not increase, by simply re-weighting the contributions of each generation of samples: this allows cheap *prior sensitivity analysis*, an important aspect to consider in applied Bayesian analyses, at a very small computational cost.

The construction of sequence (9) is not always straightforward. In the following, we consider an inverse problem whose likelihood belongs to the NEF and the prior does not depend on the hyper-parameter, deriving sequence (9) for two distinct cases:

1. the case where SMC samplers are used to approximate the full posterior distribution;
2. the case where the conditional posterior for a subset of variables x_1 can be analytically computed, and a Rao-Blackwellized SMC sampler is used to approximate the posterior on the remaining variables x_2 .

4.1 Case 1: vanilla SMC samplers for the full posterior distribution

As the likelihood belongs to the Natural Exponential Family (NEF) with natural scalar hyper-parameter $\theta \in \Theta \subseteq \mathbb{R}$, it has the following density

$$p^\theta(y | x) = \exp(\theta T(y | x) - A_\theta) \tag{16}$$

where $T(y | x)$ is a sufficient statistic and A_θ represents the log-normalizing constant.

Proposition 1 *Let $p^\theta \in NEF$ with sufficient statistic T and canonical parameter θ s.t. $p^\theta(x) = \exp(\theta T(x) - A_\theta)$ and $\alpha \neq 0$, then:*

$$[p^\theta(x)]^\alpha = \exp(A_{\alpha\theta} - \alpha A_\theta)p^{\alpha\theta}(x)$$

By the previous proposition, whose trivial proof is provided in Appendix A, it is straightforward to show that the sequence (5a) naturally provides an evaluation of the joint posterior distribution $p(x, \theta | y)$ for the set of values $\Theta_{1:T} = \{\theta \in \Theta : \theta = \theta_{\alpha_i}; t = 1, \dots, T\}$.

As an example, in the case of an inverse problem with additive Gaussian noise of unknown variance, the distributions of the sequence are posterior distributions corresponding to a decreasing variance $\sigma(t) = \sigma(T)/\sqrt{\alpha_t}$ where $\sigma(T)$ represents the noise standard deviation at the very last iteration of the SMC samplers.

The Gaussian distribution with unknown standard deviation is one of the most common distribution for the likelihood in the models for which we envisage this method being most

useful. It is not the case that every distribution that is a member NEF family will exhibit decreasing dispersion as α increases like the Gaussian does. This underlies the difficulties with so-called fractionated priors in the Consensus Monte Carlo algorithm (Scott et al. 2016) as discussed in (Rendell et al. 2021, Section 2.4). However, there are many distributions in the NEF family for which this diminishing dispersion property holds, which are suitable for the tempering sequence (5a), for example:

Exponential distribution with natural parameter $\theta = \lambda$, with density

$$p(x) = \lambda \exp(-\lambda x)$$

for $x \geq 0$;

Laplace distribution with natural parameter $\theta = \lambda$, with density:

$$p(x) = \frac{\lambda}{2} \exp(-\lambda|x - \mu|),$$

for known location μ ;

Lognormal distribution with natural parameter $\theta = 1/2\sigma^2$ and density

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log(x))^2}{2\sigma^2}\right).$$

Pareto distribution with known minimum value, x_m , and rate λ for which the natural parameter is $-(\lambda + 1)$ and the density:

$$\frac{\lambda x_m^\lambda}{x^{\lambda+1}},$$

for $x \geq x_m$.

4.2 Case 2: Rao-Blackwellized SMC samplers

We now consider the case where the unknown variable x can be decomposed into a pair of components $x = (x_1, x_2)$, and:

- the prior on x_1 belongs to the NEF with respect to a hyper-parameter λ

$$p(x_1 | \lambda) = \exp(\lambda S(x_1) - A_\lambda) \tag{17}$$

where $S(x_1)$ is a sufficient statistic and A_λ is the log-normalization constant;

- the conditional posterior $p^\theta(x_1 | x_2, \lambda, y)$ can be computed analytically.

Under these assumptions, in the natural decomposition of the joint posterior density

$$p^\theta(x_1, x_2, \lambda | y) = p^\theta(x_1 | x_2, \lambda, y) p^\theta(x_2, \lambda | y), \tag{18}$$

only the second factor of the right hand side needs to be approximated by an SMC sampler, thus reducing the variance of the importance weights and improving the quality of the approximation. This class of models is widely used and appreciated in applications; in particular, an SMC sampler targeting the marginal posterior $p^\theta(x_2, \lambda | y)$ typically leads to more accurate estimates than an SMC sampler targeting the full posterior and using the same computational resources (Murphy and Russell 2001).

As a consequence of the hypothesis that both the likelihood (16) and the prior on the Rao-Blackwellized variable (17) belong to the NEF, the marginal likelihood turns out to be

$$\begin{aligned} p^\theta(y | x_2, \lambda) &= \int p^\theta(y | x_1, x_2, \lambda) p^\lambda(x_1) dx_1 \\ &= \int \exp\left(\left\langle \begin{bmatrix} \theta \\ \lambda \end{bmatrix}, \begin{bmatrix} T(y | x_1, x_2) \\ S(x_1) \end{bmatrix} \right\rangle - (A_\theta + A_\lambda)\right) dx_1. \end{aligned} \tag{19}$$

In most cases, the marginal likelihood in equation (19) does not have a closed form solution; below we show two special cases in which it does.

4.2.1 Additive statistic for the likelihood

If the statistic $T(y | x_1, x_2)$ of the full likelihood (16) is the sum of two statistics $T(y | x_1)$ and $T(y | x_2)$, then the marginal likelihood also belongs to the NEF with respect to the same parameter

$$\begin{aligned} p^\theta(y | x_2, \lambda) &= \int \exp\left(\theta(T(y | x_1) + T(y | x_2)) \right. \\ &\quad \left. + \lambda S(x_1) - (A_\theta^{(1)} + A_\theta^{(2)} + A_\lambda)\right) dx_1 \tag{20} \\ &\propto \exp\left(\theta T(y | x_2) - A_\theta^{(2)}\right). \end{aligned}$$

For this particular subclass of models, the natural sequence (5a) is still valid, as the marginal likelihood is still in the NEF.

From the practical viewpoint, this case corresponds to a Bayesian inverse problem with multiple unknown parameters, where both the prior and the posterior for two distinct subsets of parameters factorize. One can perform inference analytically for one collection of parameters and resort to Monte Carlo only for the remainder.

4.2.2 Conditionally linear gaussian model

If both the full likelihood and the prior on x_1 have normal distribution

$$p^\theta(y | x_1, x_2) \sim \mathcal{N}(\mu(x_2)x_1, \theta^2 \Sigma) \tag{21a}$$

$$p(x_1 | \lambda) \sim \mathcal{N}(\eta, \Gamma_\lambda). \tag{21b}$$

it is well known that both the marginal likelihood (2) and the conditional posterior (Sommariva and Sorrentino 2014) are Gaussian with known mean and variance

$$p^\theta(y | x_2, \lambda) \sim \mathcal{N}\left(\mu(x_2)\eta, \mu(x_2)^t \Gamma_\lambda \mu(x_2) + \theta^2 \Sigma\right) \tag{22a}$$

$$p^\theta(x_1 | x_2, \lambda, y) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma}); \tag{22b}$$

where

$$\bar{\mu} := \Gamma_\lambda \mu(x_2)^t (\mu(x_2) \Gamma_\lambda \mu(x_2)^t + \theta^2 \Sigma)^{-1} y;$$

$$\bar{\Sigma} := \Gamma_\lambda - \Gamma_\lambda \mu(x_2)^t (\mu(x_2) \Gamma_\lambda \mu(x_2)^t + \theta^2 \Sigma)^{-1} \mu(x_2) \Gamma_\lambda.$$

In this case, the marginal likelihood is not in the NEF with respect to the parameter θ and the natural sequence (5a) does not work. Indeed, by applying to the CLG model the same sequence constructed in the general case, one would get

$$p_t^\theta(x_2, \lambda | y) \propto p(x_2, \lambda) p^\theta(y | x_2, \lambda)^{\alpha_t} \tag{23a}$$

$$p^\theta(y | x_2, \lambda)^{\alpha_t} \propto \mathcal{N}\left(y; \eta \mu(x_2), \frac{1}{\alpha_t} \left(\mu(x_2)^t \Gamma_\lambda \mu(x_2) + \theta^2 \Sigma\right)\right), \tag{23b}$$

since the marginal likelihood also embodies the prior on the marginalized variable x_1 , the exponent also affects the prior for x_1 ; therefore, as already observed in Sommariva and Sorrentino (2014), the distributions of this sequence cannot be considered as (marginals of) posterior distributions under the same prior.

Alternatively, one could consider the sequence of marginals of the natural sequence for the approximation of the complete posterior density:

$$p_t^\theta(x_2, \lambda | y) := \int p_t^\theta(x_1, x_2, \lambda | y) dx_1 \tag{24}$$

However, also this choice leads to a sequence of distributions that cannot be interpreted as posterior distributions under different values of θ ; this happens because, as shown in Appendix A (Corollary 1 and Proposition 2), the integral in (24) is

$$\begin{aligned} p_t^\theta(x_2, \lambda | y) &\propto \int p(x_1, x_2, \lambda) p^\theta(y | x_1, x_2, \lambda)^{\alpha_t} dx_1 \\ &= p(x_2, \lambda) \int p(x_1 | x_2, \lambda) p^\theta(y | x_1, x_2, \lambda)^{\alpha_t} dx_1 \\ &= p(x_2, \lambda) \ell_t(\lambda) \mathcal{N}\left(y; \eta \mu(x_2), \mu(x_2)^t \Gamma_\lambda \mu(x_2) + \frac{\theta^2}{\alpha_t} \Sigma\right), \end{aligned} \tag{25}$$

where the Gaussian distribution can be interpreted as the marginal likelihood of the CLG model, with a different value of θ , but the normalization constant $\ell_t(\lambda)$, defined as in Corollary 1 in Appendix A, depends on the hyper-parameter λ and thus actually modifies the distribution.

However, it is not difficult to devise a proper sequence of intermediate distributions for the case of a CLG model. In fact, it is sufficient to explicitly remove the λ -dependent normalization factor from (25) and construct the sequence as:

$$\begin{aligned} p_t^\theta(x_2, \lambda | y) &\propto p^\theta(x_2, \lambda) \\ &\times \mathcal{N}\left(y; \eta \mu(x_2), \mu(x_2)^t \Gamma_\lambda \mu(x_2) + \frac{\theta^2}{\alpha_t} \Sigma\right). \end{aligned} \tag{26}$$

With this definition we can apply the proposed approach to a CLG model while also exploiting Rao-Blackwellization.

5 Toy example

We proceed with a numerical validation of the proposed approach by first using a toy example¹; following the arguments in Sect. 4, we compare the results with natural alternatives for Fully Bayesian (FB) and Empirical Bayes (EB) approaches. In the following, particularly in the pictures, we denote by PropEB and PropFB the results obtained by the proposed method performing Empirical Bayes and Fully Bayesian approaches, respectively.

5.1 Setup

Consider an inverse problem where the aim is to reconstruct the mean of a Gaussian waveform of known variance σ^2 , given noisy measurements $y(t)$, i.e.

$$y(t) = \mathcal{N}(t; \mu, \sigma^2) + \varepsilon(t) \tag{27a}$$

$$\varepsilon(t) \sim \mathcal{N}(0, \theta^2). \tag{27b}$$

where $\mathcal{N}(t; \mu, \sigma^2)$ is the probability density function of a Gaussian of mean μ and standard deviation σ , evaluated at t .

¹ Code available at: https://github.com/alessandro-viani/2023_cost_free_examples/tree/main/toy_example.

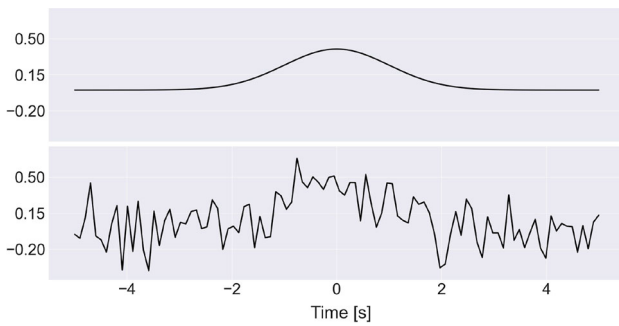


Fig. 1 The figure shows in the first row an example of data without noise while in the second row the same data with the addition of noise

We assume observations are available at I points separated by unit intervals $\{t_i\}_{i=1}^I$ and we want to make inference on the Gaussian mean.

5.2 Data generation

Data $\mathbf{y} = (y(t_1), \dots, y(t_I))$ are generated considering $I = 100$ measurements in the interval $[-5, 5]$ obtained by perturbing the Gaussian density at each observation time independently with additive Gaussian noise of zero mean and standard deviation $\theta_{\text{true}} \sim \mathcal{U}[0.1, 0.2]$.

With these settings, we generate 100 independent realizations of the dataset in order to test the proposed algorithm.

5.3 Prior and likelihood

- We assume $p(\mu) \sim \mathcal{U}([-5, 5])$ as a truncation of the Jeffrey’s prior to the convex hull of the measurements;
- we assume $p(\theta) \sim \Gamma(2, 4\theta^*)$, where θ^* is an estimated value for the hyper-parameter;
- we assume conditional independence between observations given the parameter, obtaining a simple factorization for the likelihood

$$p^\theta(\mathbf{y} | \mu) = \prod_{t=1}^T p^\theta(y(t) | \mu). \tag{28}$$

5.4 Algorithm settings

For each of the 100 generated datasets, we compare the results obtained with the proposed method with those obtained with a FB approach and an EB approach.

Each SMC sampler used has the following settings:

- number of particles set to 100 as a compromise between performances and quality of the approximation;
- $\theta^* = \min\{\theta_{\text{true}}\}/2$; this allows the true value θ_{true} to be within the range of values explored by the proposed method during SMC sampler iterations;

- number of iterations set to 500, with the sequence of exponents from 0 to 1 evenly spaced on a logarithmic scale in order to guarantee a smooth transition between intermediate distributions;
- resampling step performed by means of systematic resampling (Douc et al. 2005) whenever the effective sample size is lower than half of the number of particles;
- Gaussian proposal kernel for the MCMC step.

5.5 Comparison with alternative approaches

We compare the performances of the proposed method with those of two alternatives, one performing an Empirical Bayes approach and the other one performing a Fully Bayesian approach.

5.5.1 Empirical Bayes approach

For the EB approach we first obtain a maximum a posteriori estimate for the hyper-parameter:

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} \{\hat{p}(\theta | \mathbf{y})\}; \tag{29}$$

where $\hat{p}(\theta | \mathbf{y})$ is obtained by considering $M = 100$ evenly spaced samples in the interval $[-5, 5]$ for the mean μ :

$$\hat{p}(\theta | \mathbf{y}) = \frac{1}{M} \sum_{i=1}^M p(\mu_i, \theta | \mathbf{y})$$

and then selecting the maximum value obtained over an evenly spaced grid of 500 points for $\theta \in [\theta^*, 50 \cdot \theta_{\text{true}}]$

Once an estimate for the hyper-parameter is obtained, we consider an SMC sampler targeting the posterior distribution $p^{\hat{\theta}_{\text{MAP}}}(\mu | \mathbf{y})$.

5.5.2 Fully Bayesian approach

For the FB approach we consider an SMC sampler targeting the posterior distribution $p(\mu, \theta | \mathbf{y})$, i.e. the hyper-parameter is sampled by the SMC sampler like all other parameters; the posterior distribution for the hyper-parameter is then obtained by marginalizing the joint distribution. The SMC sampler sequence chosen for the implementation is the tempering sequence given by Eqs. (5a) (5b).

5.6 Results

We analyze the performances in terms of selection of the parameter and hyper-parameter considering the Posterior Mean (PM) and the maximum a posteriori (MAP) estimators, and compute the estimation error as the Euclidean distance between the true and the estimated value of the

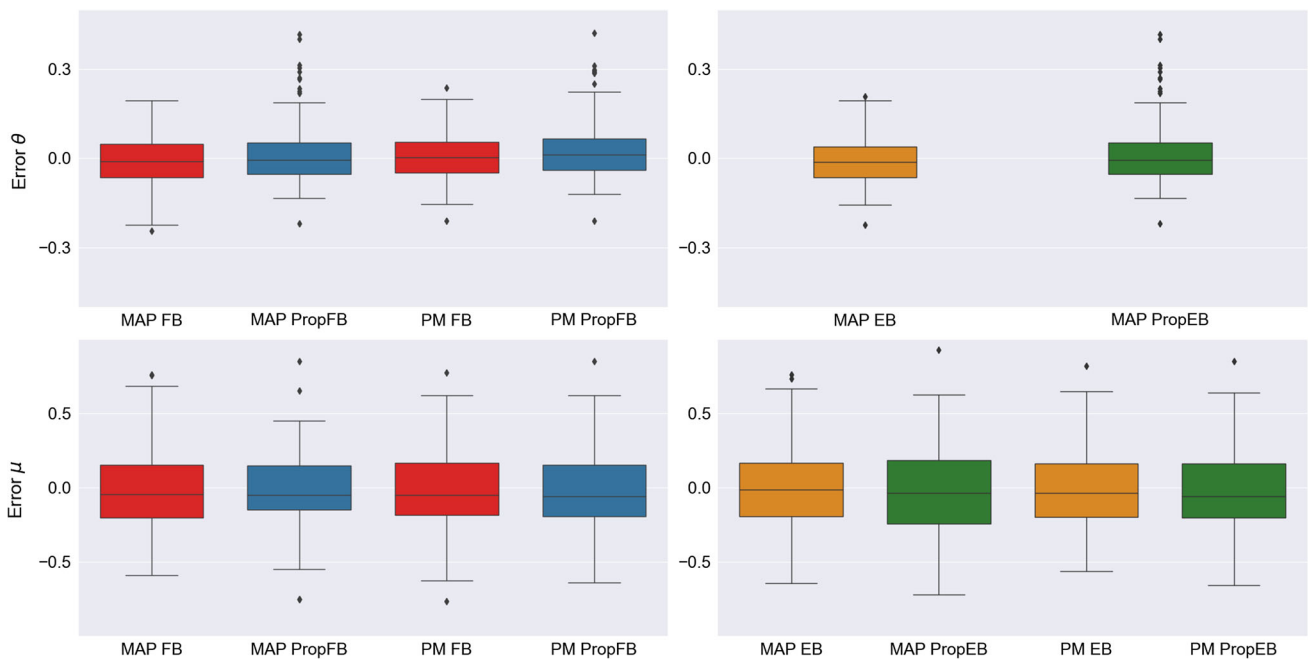


Fig. 2 Estimation error for the hyper-parameter θ (top row) and for the parameter μ (second row). Fully Bayesian, Proposed Fully Bayesian, Empirical Bayes and Proposed Empirical Bayes are shown in red, blue, yellow and green, respectively

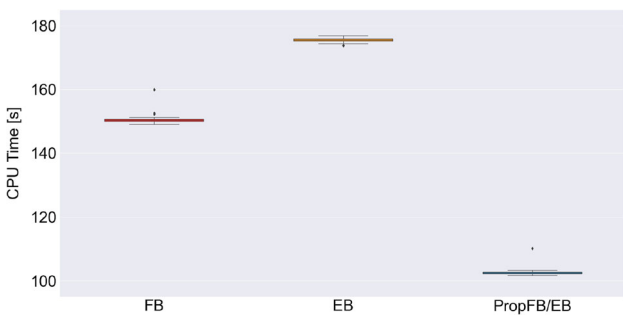


Fig. 3 Computational time for the different utilized methods. The computational time is referred to a MacBook Pro (13-inch, M1, 2020) with 8 GB of memory

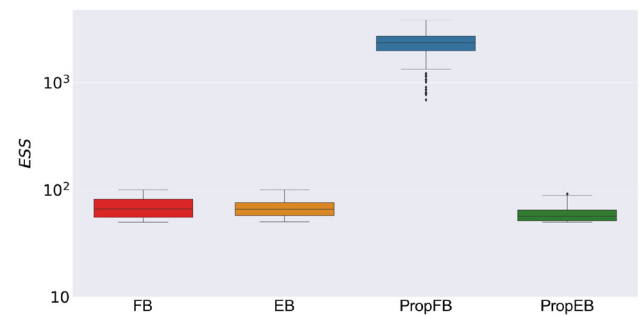


Fig. 4 Effective Sample size for the four different approaches

hyper-parameter and parameter. In Fig. 2 we report the corresponding boxplots.

We notice that the proposed approach features similar performances as the alternative approaches, either the EB and the FB, in terms of estimation error, while keeping a substantially lower computational cost (Fig. 3). In the case of the FB approach, the proposed method also features a larger ESS (Fig. 4).

5.6.1 Sample result

For illustrative purposes, in this section we show results from one specific dataset taken from the 100 simulations used in the previous section.

In Fig. 5 we show the output obtained by the proposed method and by the two alternative approaches, specifically by showing:

- the approximated posterior distribution for the hyper-parameter;
- the approximated posterior distribution for the parameter obtained in a FB approach;
- the approximated posterior distribution for the parameter in an EB approach.

As far as the approximation of the marginal posterior of the hyper-parameter is concerned, both approximations peak around the correct value, i.e. $\theta_{\text{true}} = 0.24$. Regarding the approximations for the posterior of the parameter, we observe that all the approximated distributions peak at a value close to the true value (zero) well within their support.

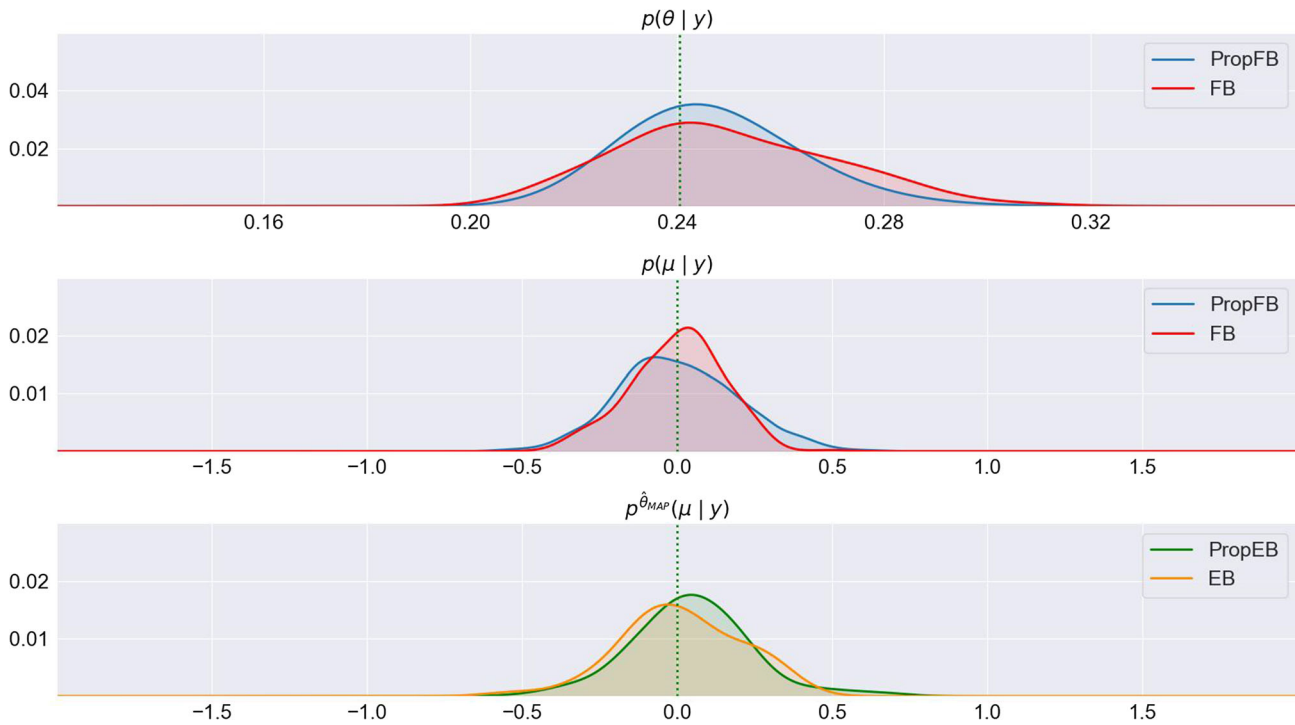


Fig. 5 Illustrative example of the posterior for the hyper-parameter (first row), the marginal of the joint posterior for the parameter (second row) and the conditional posterior for the parameter (third row); the dotted green line specify the true value.

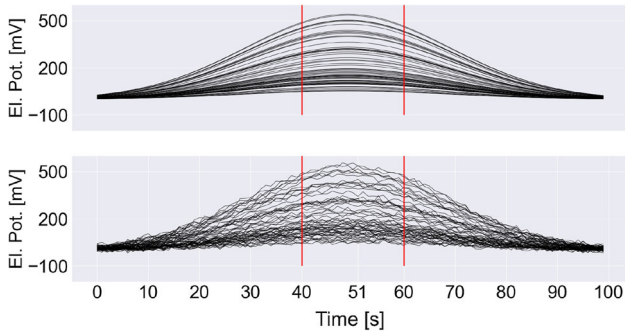


Fig. 6 Example of noise-free (top panel) and noisy (bottom panel) simulated data. Red vertical lines delimit the portion of data actually used for inference

6 Application to source imaging in magneto/electro-encephaloGraphy

In this Section we present the results² obtained with the application of the Rao-Blackwellized SMC samplers with the proposed method described in Sect.4.2 for the resolution of the M/E-EG inverse problem (Sommariva and Sorrentino 2014) introduced as a motivating example in Sect. 2.

6.1 Data generation

Data $y = (y(1), \dots, y(T))$ are generated with the following configuration:

- brain discretization Ω with 8193 voxels;
- number of EEG channels: 59;
- number of dipoles: $d = 4$;
- dipole position r_i : randomly drawn, with uniform distribution among the voxels, with the constraint that the distance between the two dipoles is larger than 3 cm; the constraint was set in order to allow for identifiability of the four dipoles;
- dipole moment q_i : orientation chosen among the three orthogonal directions, as the one that maximizes signal strength; unit dipole strength;
- noise standard deviation: $\theta_{\text{true}} \sim \mathcal{U}[1, 10]$.

With these settings, we generate 100 independent realizations of the dataset in order to test the proposed algorithm; Fig.6 shows one example of the obtained data.

² Code available at: https://github.com/alessandro-viani/2023_cost_free_examples/tree/main/meg_example.

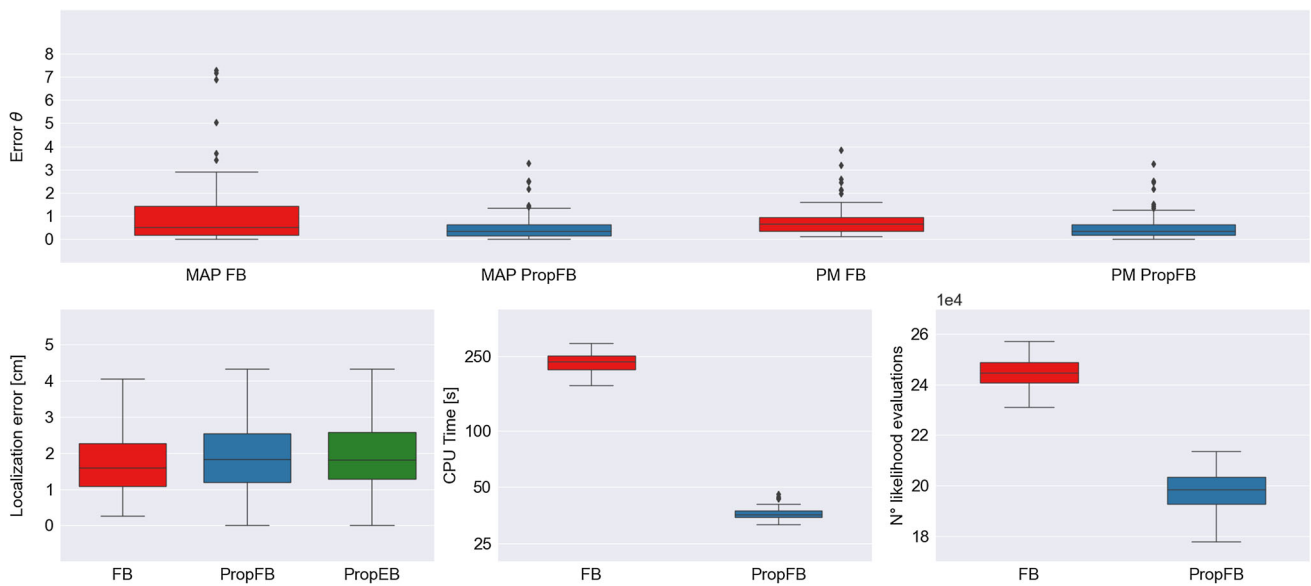


Fig. 7 Estimation error for the hyper-parameter θ (top row), for the parameter μ , computational time and number of likelihood evaluations (second row). Fully Bayesian, Proposed Fully Bayesian and Proposed

Empirical Bayes are shown in red, blue and green, respectively. The computational time is referred to a MacBook Pro (13-inch, M1, 2020) with 8 GB of memory

6.2 Prior and likelihood

We assume that all parameters are a priori independent, being $x = (d, \lambda, r_{1:d})$, the prior density is therefore

$$p(x) = p(d)p(\lambda) \prod_{i=1}^d p(r_i), \tag{30}$$

where we specify:

$$\begin{aligned} p(d) &= \text{Poisson}(d; 1); \\ p(\log(\lambda)) &= \mathcal{U}(\lambda; [-8, -5]); \\ p(r_i) &= \mathcal{U}(r; \Omega). \end{aligned}$$

We recall that the variables q_i can be marginalized out thanks to a Gaussian prior assumption, and therefore are not considered in the SMC algorithm.

We assume that noise is not correlated in time, corresponding to conditional independence between data recorded at different time points; the likelihood thus factorizes

$$p^\theta(\mathbf{y} | x) = \prod_{t=1}^T p^\theta(y(t) | d, \lambda, r_{1:d}). \tag{31}$$

6.3 Algorithm settings

Each SMC sampler was applied with the following settings:

- analysis window corresponding to the interval [40, 60], as shown in Fig. 6, i.e. analysis windows centered in the peak of the signal;
- number of particles set to 200, as a compromise between performances and quality of the approximation;
- $\theta^* = \min\{\theta_{\text{true}}\}/2$; this allows the true value θ_{true} to be within the range of values explored by the Proposed method during SMC sampler iterations; the order of magnitude of noise is typically known in this kind of data, therefore it would not be difficult to apply a similar reasoning to experimental data;
- number of iterations set to 200, with the sequence of exponents from 0 to 1 evenly spaced on a logarithmic scale;
- resampling step performed by means of systematic resampling (Douc et al. 2005) whenever the effective sample size is lower than half of the number of particles;
- MCMC kernels as described in Sommariva and Sorrentino (2014).

6.4 Performance metrics

We consider the performances in terms of selection of the hyper-parameter and in terms of localization of current dipoles.

The estimates considered for the hyper-parameter are the MAP and the PM of the marginal posterior $p(\theta | \mathbf{y})$, while the estimates for the number and the localization are defined as:

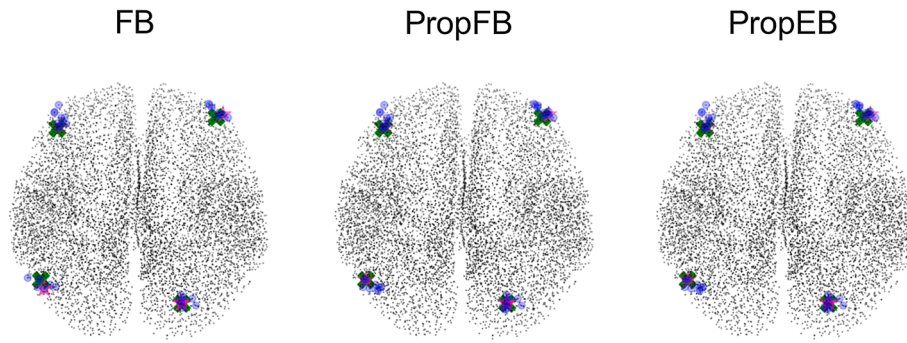


Fig. 8 Posterior probability maps for source localization obtained with the Fully Bayesian (left panel), the Proposed Fully Bayesian (central panel) and Proposed Empirical Bayes(right panel) approaches. Results

are visualized on a discretized brain as black dots, the blue dots represent the probability regions higher than 10%, the purple stars are the estimated dipoles while the green cross are the true ones

- estimator for number of dipoles: $\hat{d} = \arg \max_{d \in \mathbb{N}} (p(d | \mathbf{y}))$
- estimator for dipole location: we construct \hat{d} clusters and then obtain \hat{r}_i , for $i = 1, \dots, \hat{d}$, as the peak of the marginal posterior $p(r | \mathbf{y}, \hat{d})$ in the i -th cluster.

We note that the location estimates are a little-nonstandard in the statistics literature, but this strategy is widespread in the mutiple-object tracking literature (see, e.g., Sorrentino et al. 2013) as a natural solution to the label-switching problem in this context.

As the number of dipoles is estimated from the data, the true and estimated number of dipoles might differ; for this reason, in order to evaluate the localization error we consider the Optimal Sub-Pattern Assignment (OSPA) metric (Ristic et al. 2011), defined as follows:

$$OSPA(\hat{r}_{1:\hat{d}}, r_{1:d}) = \min_{\phi} \sum_{i=1}^{\min\{\hat{d}, d\}} \|\hat{r}_i - r_{\phi(i)}\| \quad (32)$$

where the minimum is taken over all possible permutations, ϕ , of $\{1, \dots, d\}$.

6.5 Results

In Fig. 7 we report the boxplots for the performance metrics and the computational costs of the tested algorithms.

Our results indicate that the proposed approach performs slightly better than the alternative in terms of parameter and hyper-parameter estimation. The computational cost of the proposed approach is considerably lower than the one of the alternative approach; the difference is less evident than in the case of the toy example. This difference can be explained by the combined effect of the variable dimension model, i.e. the SMC sampler exploring spaces with different number of sources, and the sampling of the hyper-parameter: when the sampled hyper-parameter is large, the SMC sampler tends

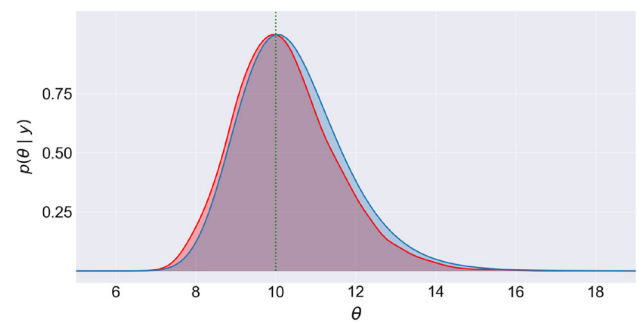


Fig. 9 The figure shows the posterior for the parameter θ approximated with the Fully Bayesian approach (red) and with the Proposed Fully Bayesian (blue). The green dotted line represents the true value for the hyper-parameter $\theta = 10$

to prefer configurations with fewer sources whose likelihood calculation is less expensive.

6.5.1 Sample result

For illustrative purposes, in this Section we show results from one specific dataset taken from the 100 simulations used in the previous Section.

In Fig. 8 we show the posterior distribution for the source location $p(r | \mathbf{y}, \hat{d})$ approximated respectively by the FB, PropFB and the PropEB approaches. All the three posterior are reciprocally similar and both methods estimate four sources in the brain with similar supports.

In Fig. 9 we show the approximated posterior distributions for the hyper-parameter provided by the two algorithms. Again we can observe that the two approximations are similar to each other and peaked around the correct value $\theta_{\text{true}} = 10$. We remark that the set of possible values for the hyper-parameter visited during the SMC iterations for the proposed method goes from 153 to 0.5; the plot focusses on a subset of this range in order to better illustrate the posterior density.

7 Conclusions

We presented a method that allows us to simultaneously conduct Fully Bayesian and Empirical Bayes estimation for a class of problems with a scalar hyperparameter at negligible additional cost.

Experiments show that the method performs slightly better than the natural alternatives, but with important differences. The proposed approach is more versatile in several ways: it allows to compute maximum likelihood/a posteriori estimates of the hyper-parameter; it allows to recycle the SMC samples for a different hyper-prior; it allows hyper-parameter selection via marginal maximum likelihood, and to provide estimates of the unknown parameters for a specific value of the hyper-parameter. In addition, when it comes to averaging across different values of the hyper-parameter, it provides substantially more Monte Carlo samples, potentially allowing better approximations of the posterior and resulting in better estimates of the unknowns.

Importantly, all these advantages are obtained essentially for free, i.e. at no additional computational cost; in addition, the proposed approach exploits samples at all iterations, thus simultaneously overcoming one of the known limitations of SMC samplers, i.e. the fact that intermediate samples are usually discarded.

Finally, although this article is dedicated to exploiting the particular structure present in a class of problems with *scalar* hyper-parameters in a way which yields both standard and empirical Bayesian estimates simultaneously with little overhead, it also suggests a path to efficiently performing empirical Bayesian estimation in a broader class of models. Estimating the gradient of the marginal likelihood with respect to the hyper-parameter using the current particle set would in principle allow the adaptive specification of a sequence of hyper-parameter values (and hence posterior distributions) which converges towards that which maximises the marginal likelihood. Such an approach is in the spirit of the SOUL (De Bortoli et al. 2021) and PGD-type (Kuntz et al. 2023) algorithms but would employ sequential Monte Carlo in order to provide sample approximations rather than Langevin-type dynamics. There are two ways one could view such an algorithm: as a Monte Carlo approximation of a gradient-based optimizer for the hyperparameter; or as an adaptive SMC sampler in which the sequence of distributions is specified by following an approximate gradient direction in the space of parameters. Such an approach would provide a natural strategy for performing empirical Bayes in settings with multivariate hyper-parameters; however, outside the scalar setting explored in this article one would lose the option to extract Fully Bayesian results from the same simulation. Exploring this rather different approach is beyond the scope of this manuscript but provides an interesting avenue for future exploration.

Acknowledgements Adam Johansen acknowledges support from the United Kingdom Engineering and Physical Sciences Research Council, under grant numbers EP/R034710/1 and EP/T004134/1

Author Contributions AV contributed to devising the core idea of the approach, did all numerical tests and contributed to writing the manuscript; AMJ expanded the core idea with several generalizations and contributed to writing the manuscript; AS contributed to devising the core idea of the approach and to writing the manuscript.

Funding Open access funding provided by Università degli Studi di Genova within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix A Proofs

A.1 Proof of Proposition 1

Proof This is direct:

$$[p^\theta(x)]^\alpha = \exp(\alpha\theta T(x) - \alpha A_\theta) = \exp(A_{\alpha\theta} - \alpha A_\theta) p^{\alpha\theta}(x)$$

□

Corollary 1 Let $p(x | \sigma) = \mathcal{N}(x; \mu, \sigma^2\Gamma)$ be an m -dimensional Gaussian density, then for any $\alpha \neq 0$:

$$p(x | \sigma)^\alpha = \sqrt{((2\pi)^m \det(\sigma^2\Gamma))^{1-\alpha} \alpha^{-m}} p\left(x \mid \frac{\sigma}{\sqrt{\alpha}}\right).$$

Proof The proof follows by the consideration that the family of the considered densities is a NEF with:

- $\theta = \frac{1}{\sigma^2}$
- $T(x) = -\frac{1}{2}(x - \mu)^t \Gamma^{-1}(x - \mu)$
- $\exp(A_\theta) = ((2\pi)^m \det(\frac{1}{\theta}\Gamma))^{\frac{1}{2}}$

Therefore the previous result guarantees the thesis because the normalizing constant is given by

$$\begin{aligned} & \exp(A_{\alpha\theta} - \alpha A_\theta) \\ &= \sqrt{\left((2\pi)^m \det\left(\frac{1}{\alpha\theta}\Gamma\right) \right) \left((2\pi)^m \det\left(\frac{1}{\theta}\Gamma\right) \right)^{-\alpha}} \\ &= \sqrt{\left((2\pi)^m \det\left(\frac{1}{\theta}\Gamma\right) \right)^{1-\alpha} \alpha^{-m}} = \sqrt{\left((2\pi)^m \det(\sigma^2\Gamma) \right)^{1-\alpha} \alpha^{-m}} \end{aligned}$$

□

A.2 Gaussian marginalisation

Proposition 2 Let $p(x_1 | \lambda) \sim \mathcal{N}(\eta, \Gamma_\lambda)$ be an m -dimensional Gaussian density and consider a k -dimensional Gaussian density $p^\theta(y | x_1, x_2, \lambda) \sim \mathcal{N}(\mu(x_2)x_1, \Sigma_\theta)$, assuming that x_1 is independent of x_2 :

$$p(x_1 | \lambda) p^\theta(y | x_1, x_2, \lambda) = \mathcal{N}\left(\begin{bmatrix} x_1 \\ y \end{bmatrix}; \begin{bmatrix} \eta \\ \eta\mu(x_2) \end{bmatrix}, \begin{bmatrix} \Gamma_\lambda & \Gamma_\lambda\mu(x_2)^t \\ \mu(x_2)\Gamma_\lambda & \Sigma_\theta + \mu(x_2)\Gamma_\lambda\mu(x_2)^t \end{bmatrix}\right),$$

$$p^\theta(y | x_2, \lambda) \sim \mathcal{N}(y; \eta\mu(x_2), \Sigma_\theta + \mu(x_2)\Gamma_\lambda\mu(x_2)^t).$$

Proof Without loss of generality we assume that $\eta = 0$, therefore the product of the Gaussian densities turns out to be

$$\begin{aligned} & p(x_1 | \lambda) p^\theta(y | x_1, x_2, \lambda) \\ & \propto \exp\left((y - \mu(x_2)x_1)^N \Sigma_\theta^{-1} (y - \mu(x_2)x_1) + x_1^t \Gamma_\lambda^{-1} x_1 \right) \\ & = \exp\left(y^t \Sigma_\theta^{-1} y - x_1^t \mu(x_2)^t \Sigma_\theta^{-1} y + x_1^t \mu(x_2)^t \Sigma_\theta^{-1} \mu(x_2)x_1 \right. \\ & \quad \left. - y^t \Sigma_\theta^{-1} \mu(x_2)x_1 + x_1^t \Gamma_\lambda^{-1} \Sigma_\theta \right) \\ & = \exp\left(y^t \Sigma_\theta^{-1} y - x_1^t \mu(x_2)^t \Sigma_\theta^{-1} y + x_1^t (\mu(x_2)^t \Sigma_\theta^{-1} \mu(x_2) \right. \\ & \quad \left. + \Gamma_\lambda^{-1}) x_1 - y^t \Sigma_\theta^{-1} \mu(x_2)x_1 \right) \\ & = \exp\left(\begin{bmatrix} x_1 \\ y \end{bmatrix}^N \begin{bmatrix} \Gamma_\lambda^{-1} + \mu(x_2)^t \Sigma_\theta^{-1} \mu(x_2) & -\mu(x_2)^t \Sigma_\theta^{-1} \\ -\Sigma_\theta^{-1} \mu(x_2) & \Sigma_\theta^{-1} \end{bmatrix} \begin{bmatrix} x_1 \\ y \end{bmatrix} \right) \end{aligned}$$

with the normalizing constant

$$\begin{aligned} & \left((2\pi)^m \det(\Sigma_\theta) (2\pi)^k \det(\Gamma_\lambda) \right)^{-\frac{1}{2}} \\ & = \left((2\pi)^{m+k} \det(\Sigma_\theta) \det(\Gamma_\lambda) \right)^{-\frac{1}{2}}. \end{aligned} \tag{A1}$$

If we consider the multivariate normal density

$$\mathcal{N}\left(\begin{bmatrix} x_1 \\ y \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Gamma_\lambda & \Gamma_\lambda\mu(x_2)^t \\ \mu(x_2)\Gamma_\lambda & \Sigma_\theta + \mu(x_2)\Gamma_\lambda\mu(x_2)^t \end{bmatrix}\right) \tag{A2}$$

then the inverse of the covariance matrix, thanks to a classical result of block-matrix inversion, turns out to be

$$\begin{aligned} & \begin{bmatrix} \Gamma_\lambda & \Gamma_\lambda\mu(x_2)^t \\ \mu(x_2)\Gamma_\lambda & \Sigma_\theta + \mu(x_2)\Gamma_\lambda\mu(x_2)^t \end{bmatrix}^{-1} \\ & = \begin{bmatrix} \Gamma_\lambda & \Gamma_\lambda\mu(x_2)^t \\ \mu(x_2)\Gamma_\lambda & \Sigma_\theta + \mu(x_2)\Gamma_\lambda\mu(x_2)^t \end{bmatrix}^{-1} \\ & = \begin{bmatrix} \Gamma_\lambda^{-1}(I + \Gamma_\lambda\mu(x_2)^t \Sigma_\theta^{-1} \mu(x_2)\Gamma_\lambda\Gamma_\lambda^{-1}) & -\Gamma_\lambda^{-1} \Gamma_\lambda\mu(x_2)^t \Sigma_\theta^{-1} \\ -\Sigma_\theta^{-1} \mu(x_2)\Gamma_\lambda\Gamma_\lambda^{-1} & \Sigma_\theta^{-1} \end{bmatrix} \\ & = \begin{bmatrix} \Gamma_\lambda^{-1}(I + \Gamma_\lambda\mu(x_2)^t \Sigma_\theta^{-1} \mu(x_2)) & -\mu(x_2)^t \Sigma_\theta^{-1} \\ -\Sigma_\theta^{-1} \mu(x_2) & \Sigma_\theta^{-1} \end{bmatrix} \\ & = \begin{bmatrix} \Gamma_\lambda^{-1} + \mu(x_2)^t \Sigma_\theta^{-1} \mu(x_2) & -\mu(x_2)^t \Sigma_\theta^{-1} \\ -\Sigma_\theta^{-1} \mu(x_2) & \Sigma_\theta^{-1} \end{bmatrix}, \end{aligned}$$

Where the normalizing constant is

$$\left((2\pi)^{\frac{m+k}{2}} \det\left(\begin{bmatrix} \Gamma_\lambda & \Gamma_\lambda\mu(x_2)^t \\ \mu(x_2)\Gamma_\lambda & \Sigma_\theta + \mu(x_2)\Gamma_\lambda\mu(x_2)^t \end{bmatrix} \right) \right)^{-\frac{1}{2}}$$

where the determinant of the covariance matrix is equal to

$$\begin{aligned} & \det(\Sigma_\theta + \mu(x_2)\Gamma_\lambda\mu(x_2)^t) \\ & \det(\Gamma_\lambda - \Gamma_\lambda\mu(x_2)^t (\Sigma_\theta + \mu(x_2)\Gamma_\lambda\mu(x_2)^t)^{-1} \mu(x_2)\Gamma_\lambda) \\ & = \det(\Sigma_\theta (I + \Sigma_\theta^{-1} \mu(x_2)\Gamma_\lambda\mu(x_2)^t)) \det(\Gamma_\lambda) \\ & \det(I - \mu(x_2)^t (\Sigma_\theta + \mu(x_2)\Gamma_\lambda\mu(x_2)^t)^{-1} \mu(x_2)\Gamma_\lambda) \\ & = \det(\Sigma_\theta) \det(I + \Sigma_\theta^{-1} \mu(x_2)\Gamma_\lambda\mu(x_2)^t) \\ & \det(\Gamma_\lambda) \det(I + \mu(x_2)^t \Sigma_\theta^{-1} \mu(x_2)\Gamma_\lambda^{-1}) \\ & = \det(\Sigma_\theta) \det(I + \Sigma_\theta^{-1} \mu(x_2)\Gamma_\lambda\mu(x_2)^t) \det(\Gamma_\lambda) \\ & \det(I + \Sigma_\theta^{-1} \mu(x_2)\Gamma_\lambda\mu(x_2)^t)^{-1} \\ & = \det(\Sigma_\theta) \det(\Gamma_\lambda). \end{aligned}$$

Therefore, from a well known result on Gaussian densities, we obtain the thesis

$$\begin{aligned} p^\theta(y | x_2, \lambda) &= \int p^\theta(y, x_1 | x_2, \lambda) dx_1 \\ &= \int p^\theta(y | x_1, x_2, \lambda) p(x_1 | x_2, \lambda) dx_1 \\ &= \mathcal{N}(y; 0, \Sigma_\theta + \mu(x_2)\Gamma_\lambda\mu(x_2)^t). \end{aligned}$$

□

References

Bernton, E, Heng, J, Doucet, A, Jacob, P. E.: Schrödinger bridge samplers. e-print 1912.13170, ArXiv, (2019)
 Chopin, N., Papaspiliopoulos, O.: An introduction to sequential Monte Carlo. Springer, (2020)
 Dau, H.-D., Chopin, N.: Waste-free sequential Monte Carlo. J. R. Stat. Soc. Ser. B Stat Methodol. 84(1), 114–148 (2022)
 De Bortoli, V., Durmus, A., Pereyra, M., Fernandez Vidal, A.: Efficient stochastic optimisation by unadjusted Langevin Monte Carlo. Stat. Comput. (2021). <https://doi.org/10.1007/s11222-020-09986-y>
 Del Moral, P., Doucet, A., Jasra, A.: Sequential Monte Carlo samplers. J. Royal Stat. Soci: Series B (Stat. Methodol.) 68(3), 411–436 (2006)

- Del Moral, P., Doucet, A., Jasra, A.: Sequential Monte Carlo for Bayesian computation. *Bay. Stat.* **8**(1), 34 (2007)
- Del Moral, P., Doucet, A., Jasra, A.: An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.* **22**(5), 1009–1020 (2012)
- Douc, R., Cappé, O., Moulines, E.: Comparison of resampling schemes for particle filtering. In: ISPA 2005. Proceedings of the 4th International symposium on image and signal processing and analysis, 2005., pp. 64–69. IEEE, (2005)
- Fox, C.W., Roberts, Stephen J.: A tutorial on variational Bayesian inference. *Artif. Intell. Rev.* **38**, 85–95 (2012)
- Gerber, M., Chopin, N., Whiteley, N.: Negative association, ordering and convergence of resampling methods. *Ann. Stat.* **37**(4), 2236–2260 (2019)
- Gilks, W.R., Richardson, S., Spiegelhalter, D.: Markov chain Monte Carlo in practice. CRC Press, Cambridge (1995)
- Good, I.J.: The estimation of probabilities: An essay on modern Bayesian methods. MIT Press, Cambridge (1965)
- Gramacy, R., Samworth, R., King, R.: Importance tempering. *Stat. Comput.* **20**(1), 1–7 (2010)
- Guarniero, P., Johansen, A.M., Lee, A.: The iterated auxiliary particle filter. *J. Am. Stat. Assoc.* **112**(520), 1636–1647 (2017)
- Hämäläinen, M., Hari, R., Ilmoniemi, R.J., Knuutila, J., Lounasmaa, O.V.: Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.* **65**, 413–497 (1993)
- Kuntz, J., Lim, J. N., Johansen, A. M.: Particle algorithms for maximum likelihood training of latent variable models. In Proceedings on 26th International Conference on Artificial Intelligence and Statistics (AISTATS), volume 206 of Proceedings of Machine Learning Research, pages 5134–5180, April (2023)
- Drovandi, C.C., South, L.F., Pettitt, A.N.: Sequential Monte Carlo samplers with independent Markov chain Monte Carlo proposals. *Bay. Anal.* **14**(3), 135–143 (2019)
- Liu, J.S.: Monte Carlo strategies in scientific computing. Springer, Berlin (2008)
- Murphy, Kevin, Russell, Stuart: Rao-Blackwellised particle filtering for dynamic Bayesian networks. In Sequential Monte Carlo methods in practice, pp. 499–515. Springer (2001)
- Neal, R.M.: Annealed importance sampling. *Stat. Comput.* **11**(2), 125–139 (2001)
- Le Thu, T., Nguyen, F.S., Peters, G.W., Delignon, Y.: Efficient sequential Monte Carlo samplers for Bayesian inference. *IEEE Trans. Signal Process.* **64**(5), 1305–1319 (2016)
- Rendell, L.J., Johansen, A.M., Lee, A., Whiteley, N.: Global consensus Monte Carlo. *J. Comput. Graph. Stat.* **30**(2), 249–259 (2021)
- Ristic, B., Vo, B.-N., Clark, D., Vo, B.-T.: A metric for performance evaluation of multi-target tracking algorithms. *IEEE Trans. Signal Process.* **59**(7), 3452–3457 (2011)
- Scott, S.L., Blocker, A.W., Bonassi, F.V., Chipman, H.A., George, E.I., McCulloch, R.E.: Bayes and big data: the consensus Monte Carlo algorithm. *Int. J. Manag. Sci. Eng. Manag.* **11**(2), 78–88 (2016)
- Sommariva, S., Sorrentino, A.: Sequential Monte Carlo samplers for semi-linear inverse problems and application to magnetoencephalography. *Inverse Prob.* **30**(11), 114020 (2014)
- Sorrentino, A., Johansen, A.M., Aston, J.A.D., Nichols, T.E., Kendall, W.S.: Dynamic filtering of static dipoles in magnetoencephalography. *The Annals Appl. Stat.* **7**, 955–988 (2013)
- Sorrentino, A., Luria, G., Aramini, R.: Bayesian multi-dipole modelling of a single topography in MEG by adaptive sequential Monte Carlo samplers. *Inverse Prob.* **30**(4), 045010 (2014)
- Stuart, A.M.: Inverse problems: a Bayesian perspective. *Acta Numerica* **19**, 451–559 (2010)
- Syed, S., Romaniello, V., Campbell, T., Bouchard-C.: Alexandre: parallel tempering on optimized paths. In: International conference on machine learning, pp. 10033–10042. PMLR (2021)
- Viani, A., Luria, G., Bornfleth, H., Sorrentino, A.: Where Bayes tweaks Gauss: conditionally Gaussian priors for stable multi-dipole estimation. *Inverse Probl. Imag.* **15**(5), (2021)
- Zhou, Y., Johansen, A.M., Aston, J.A.D.: Toward automatic model comparison: an adaptive sequential Monte Carlo approach. *J. Comput. Graph. Stat.* **25**(3), 701–726 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.