Adaptive posterior distributions for uncertainty analysis of covariance matrices in Bayesian inversion problems for multioutput signals

E. Curbelo[*], L. Martino[†], F. Llorente[**], D. Delgado-Gomez[*],

[*] Universidad Carlos III de Madrid (UC3M), Madrid, Spain.

[†] Università di Catania, Catania, Italia.

[**] Stony Brook University, New York, USA.

Abstract

In this paper we address the problem of performing Bayesian inference for the parameters of a nonlinear multioutput model and the covariance matrix of the different output signals. We propose an adaptive importance sampling (AIS) scheme for multivariate Bayesian inversion problems, which is based in two main ideas: the variables of interest are split in two blocks and the inference takes advantage of known analytical optimization formulas. We estimate both the unknown parameters of the multivariate non-linear model and the covariance matrix of the noise. In the first part of the proposed inference scheme, a novel AIS technique called adaptive target adaptive importance sampling (ATAIS) is designed, which alternates iteratively between an IS technique over the parameters of the non-linear model and a frequentist approach for the covariance matrix of the noise. In the second part of the proposed inference scheme, a prior density over the covariance matrix is considered and the cloud of samples obtained by ATAIS are recycled and re-weighted to obtain a complete Bayesian study over the model parameters and covariance matrix. ATAIS is the main contribution of the work. Additionally, the inverted layered importance sampling (ILIS) is presented as a possible compelling algorithm (but based on a conceptually simpler idea). Different numerical examples show the benefits of the proposed approaches.

Keywords: Bayesian inversion; importance sampling; uncertainty analysis; covariance matrix; tempering; sequence of posteriors

1 Introduction

The estimation of parameters from noisy observations is at the center of areas such as signal processing, statistics and machine learning. Looking at this problem from a Bayesian perspective, the inference problem becomes the construction and analysis of the posterior density over the unknown parameters [1, 2]. The computation of complicated integrals involving these posterior distributions is often needed (e.g., any moment of the random variable distributed as the posterior density). Monte Carlo sampling methods are able to draw samples from the posterior probability density function

(pdf) and hence those integrals can be approximated by stochastic quadrature formulas employing the generated samples. The Monte Carlo techniques can be divided into four main families: direct transformation methods, rejection sampling, importance sampling and Markov Chain Monte Carlo (MCMC) algorithms [3, 4, 5, 6]. The last two classes are the most used by the users, since they are universal methods, i.e., they can always be applied.

However, the Monte Carlo techniques find several difficulties that jeopardize their performance in many scenarios, for instance, when working in high-dimensional spaces, and with narrow, tight posteriors. Both issues are related to the problem of the exhaustive exploration of the state space. For these reasons, many Monte Carlo algorithms try to work in sub-dimensional spaces (step by step, with iterative or sequential procedures), such as the Gibbs sampling and the particle filtering schemes [5, 7, 8, 9]. Furthermore, the inference on a positive definite matrix requires to fulfill certain restrictions about the entries of the matrix. This issue makes also difficult the use of gradient approaches since many solutions can be outside the allowed support domain.

In this work, we focus on the problem of making a joint inference on a covariance matrix and a vector of parameters [3, 10]. This is a particularly complex inference problem since bad choices of the covariance matrix can jeopardize the sampling of the vector of interest [11, 12, 13]. This problem can suffer both issues previously described: it is often high - dimensional (especially if the dimension matrix is big) and the posterior is often tight. More specifically, we address a generic multidimensional Bayesian inversion problem, where each vector observation $\mathbf{y_r}$ is the output of a multidimensional, nonlinear *vectorial* mapping $\mathbf{f}(\boldsymbol{\theta})$ of the parameter of interest $\boldsymbol{\theta}$, perturbed by an error vector with correlated components that, e.g., can be Gaussian $\mathbf{v}_r \sim \mathcal{N}(\mathbf{v}_r|\mathbf{0}, \boldsymbol{\Sigma})$.[1] The goal is to make inference in the joint space of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$. The dimension of the entire space grows linearly with the dimension of the vector $\boldsymbol{\theta}$ and quadratically with the dimension of the matrix $\boldsymbol{\Sigma}$. We consider virtually no assumptions over the vectorial non-linearity $\mathbf{f}$, and usually it represents some complex physical process. For instance, $\mathbf{f}(\boldsymbol{\theta})$ could also be non-differentiable. In this work, the unique requirement about $\mathbf{f}$ is to be able to evaluate point-wise $\mathbf{f}(\boldsymbol{\theta})$. Since, the inference task on the complete space $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$ is particularly challenging, we introduce two different compelling Monte Carlo schemes based on the idea of splitting the inference space into two blocks, $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ (as in a block Gibbs sampling [3, 7], and/or other similar approaches [14]).

Main proposed scheme - ATAIS. Firstly, we extend and generalize the approach presented in [15, 16]. The proposed inference scheme is divided into two main parts. In the first part, we approximate the conditional posterior of $\boldsymbol{\theta}$ given the data and the maximum likelihood estimator $\boldsymbol{\Sigma}_{\mathrm{ML}}$ of the matrix $\boldsymbol{\Sigma}$. This first part is called *adaptive target adaptive importance sampling* (ATAIS), since we perform an adaptive importance sampling on a sequence of adaptive posteriors (due to the variation of $\boldsymbol{\Sigma}$). The ATAIS method is then completed by a second part which allows a complete Bayesian inference also over $\boldsymbol{\Sigma}$. Indeed, in this second inference part, we approximate the complete posterior of pair of variables of interest $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$, without any additional generation of samples over $\boldsymbol{\theta}$. The resulting scheme is a robust inference approach for Bayesian inversion, based on an adaptive importance sampler that

---

[1]We assume Gaussianity in the first part of the work, only for clarity and simplicity in the explanation.

addresses a sequence of different conditional posteriors and a post-process that allows a Bayesian inference over Σ as well. We refer to the overall scheme (first and second part) as *complete ATAIS*. The conditional posteriors addressed by ATAIS differ in the use of different covariance matrices: this procedure can resemble a tempering of the posterior distribution [17, 18, 19, 20].

Auxiliary competitive scheme - ILIS. As also remarked in different works [15, 21, 11], the application of a Monte Carlo sampling method directly in the complete space $\{\theta, \Sigma\}$ is particularly challenging and the resulting performance is quite poor. Hence, at least with our current knowledge of the literature, it is also difficult to find a competitive alternative to ATAIS, which can provide errors in estimation of the same magnitude. However, in our practical experience, we have designed another Monte Carlo scheme (conceptually simpler than ATAIS) that can also obtain reasonable results. We call this competitive scheme, *inverse layered importance sampling* (ILIS) since we adapt the idea given [22, 23] for this inference context. With respect to the main algorithm in [22, 23], we switch the positions of the importance sampling (IS) method and Markov Chain Monte Carlo (MCMC) techniques [24, 25, 7, 10]: in ILIS the upper layer is formed by an IS procedure, and the lower layer is formed by *weighted* MCMC chains. Conceptually speaking, ILIS can be considered simpler than ATAIS, but the ILIS performance is more sensible on choice of certain proposal parameters (e.g., covariance reference matrix in the upper layer proposal), whereas the complete ATAIS procedure is able to auto-tune some auxiliary parameters, reducing the number of parameters decided by the user. In this sense, ATAIS is more automatic and robust than ILIS. As a final observation, we highlight that ATAIS could also be combined and jointly employed.

A summary of the main contributions of the work and related important considerations are given below:

- We propose a robust and efficient inference scheme for complex Bayesian inversion problems, where a scale (covariance) matrix must also be estimated. The model considers a vectorial nonlinear function $\mathbf{f}(\theta)$ to invert, that can represent complex dynamical systems, a set of time series models, or a statistical spatial model for instance.

- The proposed method allows a complete Bayesian analysis of θ and Σ so that, we can perform uncertainty analysis over θ and/or Σ, obtaining credible intervals. Moreover, we can perform hypothesis testing or model selection approximating the marginal likelihood [3, 26, 27]. Hence, we remark that the proposed scheme is *much more* than an optimizer: it is a sampler that allows a complete Bayesian inference over θ and Σ. In its second inference part, ATAIS recycles all the samples (w.r.t. θ) and the posterior evaluations from the first part. Thus, this second part does not require any additional evaluation of the possibly complex and costly nonlinearity $\mathbf{f}(\theta)$.

- ATAIS can be considered as an adaptive importance sampler where *both* proposal and target pdfs are adapted. Indeed, in ATAIS, we consider a sequence of adaptive conditional posteriors. The complete ATAIS method can be also interpreted as an IS version of the *recycling Gibbs sampling* scheme in [7] (with two blocks). Indeed, the complete space is divided into two blocks, θ and Σ, where different numbers of samples are considered for each block, denoted as NT for θ and J for Σ.

- We also introduce several extensions as addressing models with non-Gaussian noises, e.g., with t-Student's noise (or, more generally, with other elliptical distributions) and/or the possible use of mini-batches (that is allowed by ATAIS). A discussion with practical suggestions regarding the tuning of hyper-parameters of the prior densities is provided. Furthermore, a detailed discussion about alternatives and improvements (for speeding up or reducing the computational cost) is given as well.

- We also designed a competitive sampling scheme, denoted as ILIS, for comparing the performance of ATAIS. Several comparisons with other benchmark techniques are also provided.

The range of applications includes the inference of any temporal and spatial dataset such as: systems of differential equations explaining the behavior of a disease (for instance, SIR models for COVID) or time series in different cities, weather prediction in different regions, medical signals with multiple sensors, topology graph estimation and, more generally, any inference problem with time-varying signals defined in different spatial locations [28, 29, 30, 31].

The paper is structured as follows. We start with the description of the problem statement in Section 2. The first part of the main proposed inference scheme is introduced in Sections 3 and 3.1. Section 4 provides a detailed discussion about possible alternatives and improvements. The second part of the main proposed inference scheme is described in Section 5. The alternative scheme, inverted layered importance sampling (ILIS), is given in Section 6. Finally, Section 7 contains several numerical experiments and Section 8 provides some final conclusions.

2 Problem Statement

Let us denote as $\boldsymbol{\theta} = [\theta_1, ..., \theta_M]^\top \in \Theta \subseteq \mathbb{R}^M$, a variable of interest that we desire to infer. Moreover, related to $\boldsymbol{\theta}$, we observe

- R values in different time instants (or spatial points) of

- K different signals (time series), i.e.,

$\mathbf{y}_r = [y_{r,1}, ..., y_{r,K}] \in \mathbb{R}^{K \times 1}$ for $r = 1, ..., R$. Hence, all received data can be stored in a matrix $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_R] \in \mathbb{R}^{K \times R}$. Furthermore, let us consider the observation model

$$\mathbf{y}_r = \mathbf{f}_r(\boldsymbol{\theta}) + \mathbf{v}_r, \qquad r = 1, ..., R, \tag{1}$$

$$\mathbf{Y} = \mathbf{F}(\boldsymbol{\theta}) + \mathbf{V}, \tag{2}$$

where we have a nonlinear mapping for each time instant and each time series,

$$\mathbf{f}_r(\boldsymbol{\theta}) = [f_{r,1}(\boldsymbol{\theta}), ..., f_{r,K}(\boldsymbol{\theta})]^\top : \Theta \subseteq \mathbb{R}^M \rightarrow \mathbb{R}^{K \times 1}, \tag{3}$$

$$\mathbf{F}(\boldsymbol{\theta}) = [\mathbf{f}_1(\boldsymbol{\theta}), ..., \mathbf{f}_R(\boldsymbol{\theta})] : \Theta \subseteq \mathbb{R}^M \rightarrow \mathbb{R}^{K \times R}, \tag{4}$$

and a $K \times 1$ vector of Gaussian noise perturbation for each time instant,

$$\mathbf{v}_r = [v_{r,1}, ..., v_{r,K}]^\top \sim \mathcal{N}(\mathbf{v}_r | \mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^{K \times 1}, \tag{5}$$

$$\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_R] \in \mathbb{R}^{K \times R}, \tag{6}$$

where Σ is $K \times K$ covariance matrix, which generally is unknown. The mapping $\mathbf{f}_r(\boldsymbol{\theta})$ could be analytically unknown, the only assumption is that we are able to evaluate it pointwise.[2] The likelihood function is

$$\ell(\mathbf{Y}|\boldsymbol{\theta}, \Sigma) = \left(\frac{1}{(2\pi)^{K/2}\det(\Sigma)^{1/2}}\right)^R \exp\left(-\frac{1}{2}\left[\sum_{r=1}^{R}(\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}))^\top \Sigma^{-1}(\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}))\right]\right), \tag{7}$$

Note that we have two types of variables of interest for an inference point of view:

- the vector $\boldsymbol{\theta}$ contains the parameters of the nonlinear mapping $\mathbf{f}_r(\boldsymbol{\theta})$, for $r = 1, ..., R$,

- and Σ is a scale matrix of the likelihood function.

Given the complete matrix of measurements $\mathbf{Y}$, we desire to make inferences regarding the hidden parameters $\boldsymbol{\theta}$ and the noise matrix Σ, obtaining at least some point estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\Sigma}$. We are also interested in performing uncertainty and correlation analysis among the components of $\boldsymbol{\theta}$. Furthermore, we aim to perform model selection, i.e., to compare, select or properly average different models.

2.1 Application to time series and spatial processes

The range of applications of the considered model is very broad. For instance, in the case of having K different time series (in continuous or discrete time), or K spatial processes we can have more explicit notation, where there is a one-to-one correspondence between each index $r \in \{1, ..., R\}$ and a real time instant $\tau_{k,r} \in \mathbb{R}$ or a point $\mathbf{x}_{k,r} \in \mathbb{R}^{d \times 1}$, i.e.,

$$r \in \{1, ..., R\} \longleftrightarrow \{\tau_{k,r} \in \mathbb{R}\}_{k=1}^{K}, \qquad r \in \{1, ..., R\} \longleftrightarrow \{\mathbf{x}_{k,r} \in \mathbb{R}^d\}_{k=1}^{K}.$$

Each vector $\mathbf{y}_r$ (of dimension $K \times 1$) contains the measurements at time instants $\tau_{1,r}, ..., \tau_{K,r}$ (or $\mathbf{x}_{1,r}, ..., \mathbf{x}_{K,r}$) each one corresponding to a different time series. Generally, an alternative notation is

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}, \boldsymbol{\tau}) + \mathbf{v}, \qquad \mathbf{y} = \mathbf{f}(\boldsymbol{\theta}, \mathbf{X}) + \mathbf{v}, \tag{8}$$

where $\boldsymbol{\tau} = [\tau_1, ..., \tau_K]$ and $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_K]$ is $d \times K$ matrix.[3] More specifically, recalling the observation equation $\mathbf{y}_r = \mathbf{f}_r(\boldsymbol{\theta}) + \mathbf{v}_r$, we could use a more explicit notation, instead of $\mathbf{f}_r(\boldsymbol{\theta})$, i.e.,

$$\mathbf{y}_r = \mathbf{f}(\boldsymbol{\theta}, \tau_{1,r}, ..., \tau_{K,r}) + \mathbf{v}_r, \tag{9}$$

$$\mathbf{y}_r = \mathbf{f}(\boldsymbol{\theta}, \mathbf{x}_{1,r}, ..., \mathbf{x}_{K,r}) + \mathbf{v}_r, \qquad r = 1, ..., R, \tag{10}$$

where $\tau_{1,r}, ..., \tau_{K,r}$, or $\mathbf{x}_{1,r}, ..., \mathbf{x}_{K,r}$ play the role of auxiliary known parameters (or vectors of parameters). A graphical representation is given in Figure 1.

Remark. Note that the vector $\boldsymbol{\theta} = [\theta_1, ..., \theta_M]^\top$ contains the parameters that are possibly shared from the models representing the K different time series (or K spatial processes), or all the parameters that only affect one series (or just a subset of time series).

---

[2]Each component $f_{r,k}$, for $k = 1, ...K$, can be a function of the complete vector $\boldsymbol{\theta}$ or only a subset of components of this vector. See for instance the simulation experiment in Section 7.2.

[3] Note that in the isotopic scenario, where $\tau_1 = \tau_2... = \tau_K = \tau$, the notation can be simplified as $\mathbf{f}(\boldsymbol{\theta}, \tau)$ where τ is a scalar.
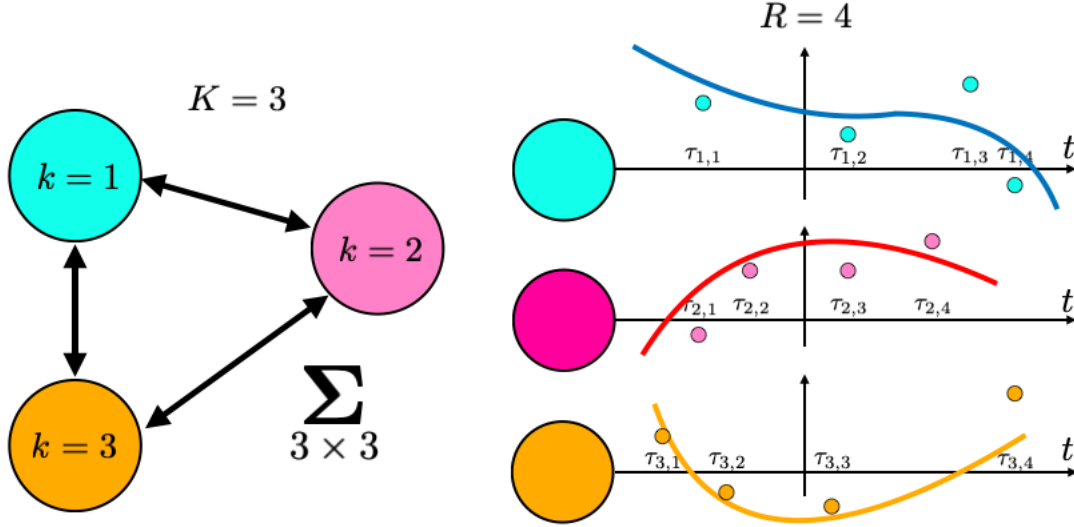
Figure 1: Graphical representation of the considered multioutput model with $K = 3$ output signals and $R = 4$ time instants for each signal. One can suppose that Σ represents a 3×3 covariance matrix of three possible nodes in a graph.

2.2 Bayesian inference in the complete space

The full Bayesian solution considers the study of the complete posterior density

$$p(\boldsymbol{\theta}, \Sigma | \mathbf{Y}) = \frac{1}{p(\mathbf{Y})} p(\boldsymbol{\theta}, \Sigma, \mathbf{Y}) = \frac{1}{p(\mathbf{Y})} \ell(\mathbf{Y} | \boldsymbol{\theta}, \Sigma) g_{\theta}(\boldsymbol{\theta}) g_{\Sigma}(\Sigma), \qquad (11)$$

where $g_{\theta}(\boldsymbol{\theta})$ and $g_{\Sigma}(\Sigma)$ represent the prior densities over the vector $\boldsymbol{\theta}$ and the matrix Σ. Usually, complex integrals involving $p(\boldsymbol{\theta}, \Sigma | \mathbf{Y})$ should be computed in order to perform the inference.

Main observation. Generally, generating random samples from a complicated posterior in Eq. (11) and computing efficiently the integrals involving $p(\boldsymbol{\theta}, \Sigma | \mathbf{Y})$ is a hard task. Note that the complete dimension of the inference problem D is

$$D = M + \frac{K(K + 1)}{2},$$

i.e., the number of parameters to infer is exactly D. With $M = 2$ and $K = 5$, we have $D = 17$ and with $M = 2$, $K = 10$ we have $D = 57$. The dimension D grows linearly with M and quadratic with respect to K. Moreover, we also have the constraints regarding Σ, since it must be a covariance matrix. This task becomes even more difficult when we try to perform a joint inference, learning jointly the covariance matrix Σ and parameters of the nonlinearity $\boldsymbol{\theta}$. Indeed, "wrong choices" of Σ can easily jeopardize the sampling of $\boldsymbol{\theta}$. Note also that the inference of a covariance matrix requires to fulfill certain constraints on the entries of the matrix.

Proposed approach. Below, we describe an inference scheme formed by *two main parts*. First,

6

we tackle the problem of drawing from the conditional posterior of $\boldsymbol{\theta}$, conditioned to the data and the maximum likelihood (ML) estimator of $\boldsymbol{\Sigma}$ (that is generally unknown and must be approximated as well). Therefore, in this first part, we apply a Bayesian inference over $\boldsymbol{\theta}$ and a frequentist approach over $\boldsymbol{\Sigma}$ (obtaining an approximation of the ML estimator). In the second part, we assume also a prior density over the covariance matrix $\boldsymbol{\Sigma}$, and perform a Bayesian inference over $\boldsymbol{\Sigma}$ as well, recycling the outputs (weighted samples and other information) obtained in the first part.

3   First part of the proposed inference scheme

The main idea underlying the proposed inference scheme is to take advantage of the split of the inference space (working firstly in smaller portions of the entire space). In the first part, described in this section, we search for high probability regions in the complete space, sampling from a *sequence of adaptive conditional posterior distributions* with respect to $\boldsymbol{\theta}$ (given a covariance matrix $\boldsymbol{\Sigma}_{\texttt{ML}}$). An analytic formula is employed for obtaining a sequence of optimized matrices $\boldsymbol{\Sigma}_{\texttt{ML}}$. In the second part, described in Section 5, we generate random matrices from a *tuned* prior pdf (or possibly other proposal density) and we re-weight all the previously generated samples w.r.t. $\boldsymbol{\theta}$, in order to allow a complete Bayesian inference (hence including uncertainty analysis, etc.) for both $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$.

More specifically, in the first stage, we consider a sub-optimal (in Bayesian sense) but substantially more efficient inference scheme (since we work in a reduced - much smaller - dimensional space), studying only a sequence of conditional posterior distributions. More precisely, we study the following conditional posterior

$$p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma}_{\texttt{ML}}) = \frac{\ell(\mathbf{Y}|\boldsymbol{\Sigma}_{\texttt{ML}}, \boldsymbol{\theta})g_{\theta}(\boldsymbol{\theta})}{p(\mathbf{Y}|\boldsymbol{\Sigma}_{\texttt{ML}})} \propto \ell(\mathbf{Y}|\boldsymbol{\Sigma}_{\texttt{ML}}, \boldsymbol{\theta})g_{\theta}(\boldsymbol{\theta}). \tag{12}$$

Furthermore, we have denoted the *conditioned* maximum likelihood estimator of $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma}_{\texttt{ML}} = \arg\max_{\boldsymbol{\Sigma}} \ell(\mathbf{Y}|\boldsymbol{\Sigma}, \boldsymbol{\theta}_{\texttt{MAP}}), \tag{13}$$

where $\boldsymbol{\theta}_{\texttt{MAP}}$ denotes the global maximum of $p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma}_{\texttt{ML}})$, i.e.,

$$\boldsymbol{\theta}_{\texttt{MAP}} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma}_{\texttt{ML}}),$$

$$= \arg\min_{\boldsymbol{\theta}} \left[\sum_{r=1}^{R} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}))^{\top} \boldsymbol{\Sigma}_{\texttt{ML}}^{-1} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta})) + \log g_{\theta}(\boldsymbol{\theta}) \right]. \tag{14}$$

It is important to observe that, given $\boldsymbol{\theta}_{\texttt{MAP}}$, we have the analytic form of $\boldsymbol{\Sigma}_{\texttt{ML}}$, i.e.,

$$\boldsymbol{\Sigma}_{\texttt{ML}} = \frac{1}{R} \sum_{r=1}^{R} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}_{\texttt{MAP}})) (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}_{\texttt{MAP}}))^{\top}. \tag{15}$$

Note that $\boldsymbol{\Sigma}_{\texttt{ML}}$ depends on $\boldsymbol{\theta}_{\texttt{MAP}}$, and $\boldsymbol{\theta}_{\texttt{MAP}}$ depends on $\boldsymbol{\Sigma}_{\texttt{ML}}$. For high dimensional estimators of $\boldsymbol{\Sigma}_{\texttt{ML}}$, see the procedure described in [32]. Moreover, in case of heavy-tailed distributions and presence of

outliers, see the relevant survey [33]. Finally, similar approaches for dealing with unknown covariance can be found in [21, 11].

Remark. The key idea to implement this inference scheme is to perform an alternating optimization procedure where, at each iteration t, we produce two estimations $\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}$, $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(t)}$ of $\boldsymbol{\theta}_{\text{MAP}}$, $\boldsymbol{\Sigma}_{\text{ML}}$, respectively [12, 13]. Clearly, we desire the convergence as the number of iterations grows, $t \to \infty$, i.e.,

$$\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)} \longrightarrow \boldsymbol{\theta}_{\text{MAP}}, \tag{16}$$

$$\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(t)} \longrightarrow \boldsymbol{\Sigma}_{\text{ML}}. \tag{17}$$

The suggested iterative approach is summarized briefly in two steps. Starting with an initial matrix $\boldsymbol{\Sigma}_{\text{ML}}^{(0)}$, that is as a rough approximation of $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}$, the alternating optimization procedure is given in Table 1.

Table 1: Alternating optimization.

For $t = 1, \ldots, T$:

   1 Estimate, by Monte Carlo,

$$\boldsymbol{\theta}_{\text{MAP}}^{(t)} = \arg \min_{\boldsymbol{\theta}} \left[\sum_{r=1}^{R} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}))^\top \left[\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(t-1)} \right]^{-1} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta})) - \log g_{\theta}(\boldsymbol{\theta}) \right], \tag{18}$$

     obtaining $\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}$, e.g., using an importance sampling (IS) scheme with respect to $p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(t-1)})$.

   2 Compute

$$\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(t)} = \frac{1}{R} \sum_{r=1}^{R} \left(\mathbf{y}_r - \mathbf{f}_r(\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}) \right) \left(\mathbf{y}_r - \mathbf{f}_r(\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}) \right)^\top. \tag{19}$$

Since, we employ IS scheme for obtaining $\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}$, at each t-th iteration, we have also a cloud of particles $\{\boldsymbol{\theta}_t^{(n)}\}_{n=1}^{N}$ that can be used for performing Bayesian inference over $\boldsymbol{\theta}$. Namely, after T iterations, we can build a particle approximation of $p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(T)})$, i.e.,

$$\widehat{p}(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(T)}) = \sum_{t=1}^{T} \sum_{n=1}^{N} \widetilde{w}_t^{(n)} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)}), \qquad \sum_{t=1}^{T} \sum_{n=1}^{N} \widetilde{w}_t^{(n)} = 1. \tag{20}$$

By Eq. (20), we can approximate all the moments associated with the conditional posterior $p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(T)})$ hence, for instance, we can also provide an uncertainty estimation over the vector of $\boldsymbol{\theta}$.

On the convergence of the alternating optimization. Due to the error in step 1 of the alternating optimization (described above) can be controlled by the number of particles N (i.e., the error in the approximation of θ_{MAP} can be bounded by increasing N, i.e., even with a bad choice of $\widehat{\Sigma}_{\text{ML}}^{(t-1)}$ we can obtain a reasonable vector $\widehat{\theta}_{\text{MAP}}^{(t)}$, and the estimator $\widehat{\Sigma}_{\text{ML}}^{(t)}$ in Eq. (19) approaches the matrix Σ_{ML} in Eq. (15), as $t \to \infty$. Moreover, as the number of realizations R grows the matrix Σ_{ML} in Eq. (15) converges to the true covariance matrix of the data.

Note that the pair θ_{MAP} and Σ_{ML} are *fixed points* of *the iterative (dynamical) system* formed by Eqs. (18)-(19). Namely, the key point of the convergence is to be able to find a good approximation of θ_{MAP} (placing us close to the fixed point). This is possible since we are working in a reduced portion of the complete space, and more efficient Monte Carlo scheme can be applied [34, 35, 36, 37]. It has the same convergence rate of a Monte Carlo method for stochastic optimization, as a standard simulated annealing [17]. The rate of convergence of an optimizer driven by sampling has been also recently studied in [38, 39]. The authors in [38] show that, in a specific sampler (that iteratively optimizes a posterior pdf), the error bound is $O\left(N^{-\frac{1}{2(M+1)}}\right)$, where M is the dimension of the θ-space.

Accelerating the convergence of the global optimization problem. In other to find a good region of the space for starting the alternating optimization, we can use some iterations (let's say $T_0 < T$) of the algorithm considering

$$\theta_{\text{MAP}}^{(t)} = \arg\min_{\theta} \left[\sum_{r=1}^{R} \|\mathbf{y}_r - \mathbf{f}_r(\theta)\|^2 - \log g_{\theta}(\theta) \right], \qquad t = 1, ..., T_0, \tag{21}$$

that is equivalent to set $\widehat{\Sigma}_{\text{ML}}^{(t)} = \mathbf{I}_K$ for $t = 0, ..., T_0 - 1$ in Eq. (18), where $\mathbf{I}_K$ is a $K \times K$ unit matrix. Thus, in the first T_0 iterations, we focus only on finding a good point $\theta_{\text{MAP}}^{(T_0)}$. Indeed, note that if there exists a point θ^* such that $\sum_{r=1}^{R} \|\mathbf{y}_r - \mathbf{f}_r(\theta^*)\|^2 = 0$, then this point θ^* is also *a root* for $\sum_{r=1}^{R} (\mathbf{y}_r - \mathbf{f}_r(\theta^*))^\top \widehat{\Sigma}^{-1} (\mathbf{y}_r - \mathbf{f}_r(\theta^*)) = 0$ for any possible covariance matrix $\widehat{\Sigma}$.

Outputs of this first part of the inference scheme. With the procedure above, we perform a Bayesian inference over the vector θ, but *only* analyzing and approximating the conditional posterior $p(\theta|\mathbf{Y}, \widehat{\Sigma}_{\text{ML}}^{(T)})$. In this first part, with respect to Σ, we only provide a frequentist estimator $\widehat{\Sigma}_{\text{ML}}^{(T)}$.

Note that, in the iterative procedure, we have a sequence conditional posteriors $p(\theta|\mathbf{Y}, \widehat{\Sigma}_{\text{ML}}^{(t)})$. For this reason, we call the algorithm as *adaptive target adaptive importance sampling* (ATAIS).[4] The details of the ATAIS algorithm which performs this scheme are given in the next section.

3.1 Adaptive Target Adaptive Importance Sampling (ATAIS)

This section provides more details about Step 1 of the alternating procedure described above. More generally, we will provide all the details of the ATAIS algorithm. To simplify the notation, we denote the unnormalized conditional posterior at the t-th iteration,

$$\pi_t(\theta) = \ell(\mathbf{Y}|\widehat{\Sigma}_{\text{ML}}^{(t-1)}, \theta) g_\theta(\theta) \propto p(\theta|\mathbf{Y}, \widehat{\Sigma}_{\text{ML}}^{(t-1)}). \tag{22}$$

---

[4]Another reason is that it is also an extension of the techniques in [15, 16], that use the acronym ATAIS as well.

At each iteration, we consider $\pi_t(\boldsymbol{\theta})$ as the target distribution. Finally, we are able to approximate $\pi_{T+1}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\Sigma}_{\mathrm{ML}}^{(T)})$, without any additional evaluation of the likelihood function. The dependence on the iteration t is due to $\widehat{\Sigma}_{\mathrm{ML}}^{(t)}$ varies with t. The ATAIS algorithm is outlined in Table 8, whereas the main features of ATAIS are described below.

IS steps. A set of N samples $\{\boldsymbol{\theta}_t^{(n)}\}_{n=1}^{N}$ are drawn from a (normalized) proposal density $q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \Lambda_t)$ with mean $\boldsymbol{\mu}_t$ and a covariance matrix Λ_t. An importance weight

$$w_t^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \Lambda_t)},$$

is assigned to each sample $\boldsymbol{\theta}_t^{(n)}$, for all n and t. Note that we assume that $q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \Lambda_t)$ normalized, i.e., $\int_\Theta q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \Lambda_t)d\boldsymbol{\theta} = 1$ and with heavier tails than π_t.

Proposal adaptation. The location parameter of the proposal density is moved to $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(t)}$, i.e.,

$$\boldsymbol{\mu}_{t+1} = \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(t)}. \tag{23}$$

Note that, we set $\boldsymbol{\mu}_{t+1} = \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(t)}$ instead of using the empirical mean of the samples (as in other classical AIS schemes). This is because we have noticed that this choice provides better and more robust results, especially as the dimension of the problem grows. Indeed, this choice helps in the search for the global maximum (since the next cloud of particles will be around the current MAP estimation) and, as a consequence, helps also the estimation of $\widehat{\Sigma}_{\mathrm{ML}}$ due to (19). Regarding the adaptation of the covariance matrix Λ_t of the proposal density, see Section 4.

ATAIS outputs. After T iterations, a final correction of the weights is needed, i.e.,

$$\widetilde{w}_t^{(n)} = w_t^{(n)} \frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})}, \qquad \text{for all } n, t, \tag{24}$$

in order to obtain a particle approximation of the measure of the final conditional posterior $\pi_{T+1}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\Sigma}_{\mathrm{ML}}^{(T)})$. Thus, the algorithm returns the final estimators $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(T)}$, $\widehat{\Sigma}_{\mathrm{ML}}^{(T)}$, and all the weighted samples $\{\boldsymbol{\theta}_t^{(n)}, \widetilde{w}_t^{(n)}\}$, for all $n = 1, ..., N$ and $t = 1, ..., T$. Other outputs can be obtained with a post-processing of the weighted samples, as shown below. Note that Eq. (24) does not require any additional evaluations of the model, if we save the computation of the error vectors $\mathbf{e}_{t,r}^{(n)} = \mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}_t^{(n)})$. Moreover, we can also use $\{\mathbf{e}_{t,r}^{(n)}\}$ and $\{\boldsymbol{\theta}_t^{(n)}\}$ for building a particle approximation of any other conditional posterior $p(\boldsymbol{\theta}|\mathbf{Y}, \Sigma)$, as proposed in the second part of the inference. Note that ATAIS is an adaptive importance sampler. Alternatively, if we would like to employ an adaptive MCMC algorithm in Table 8 (instead of an importance sampler), - assuming we use a sufficiently large T - we should set $w_t^{(n)} = 1$ in Eq. (24). In any case, the weights would be $\widetilde{w}_t^{(n)} = \frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})}$ in Eq. (24), keeping still the IS nature.

3.2 Elliptically contoured distributions in the observation models

The ATAIS algorithm described above (including the alternating optimization above) can be also applied for more general models. When the noise has an elliptically contoured distribution [40, 41, 42],

10

i.e.,

$$\mathbf{v}_r \sim p(\mathbf{v}), \quad \text{with} \quad p(\mathbf{v}) = \frac{k_p}{|\Sigma|^{1/2}} h\left(\mathbf{v}^\top \Sigma^{-1} \mathbf{v}\right), \tag{25}$$

where $k_p > 0$ is a constant, $h(s) : \mathbb{R} \rightarrow \mathbb{R}^+$ is a one-dimensional positive function. The $K \times K$ matrix Σ is a scale matrix related to the corresponding covariance matrix. An example is the multivariate t-distribution:

$$p(\mathbf{v}) \propto \frac{1}{|\Sigma|^{1/2}} \left[1 + \frac{1}{\nu} \mathbf{v}^\top \Sigma^{-1} \mathbf{v} \right]^{-(\nu+K)/2},$$

where $\nu > 0$ represents the degrees of freedom. However, unlike in the Gaussian case, in this scenario the empirical covariance matrix is not an estimator of the scale matrix Σ, but of a scaled version of the scale matrix, i.e., $\alpha \Sigma$ where $\alpha > 0$. Generally, the estimator depends on the specific function $h(s)$ [43].[5] Let us consider the problem of estimating only the scale parameter keeping fixed $\mathbf{f}(\boldsymbol{\theta})$, as in step 2(b)ii of ATAIS. When the conditions expressed in [43] are satisfied (which are satisfied for absolutely continuous distributions such as Student's-t densities), the scaling matrix Σ can be estimated by a fixed-point method as follows. Let define an initialization $\widehat{\Sigma}_0 > 0$, the function $\eta(s) = -2\frac{h'(s)}{h(s)}$ and $\mathbf{e}_r = \mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta})$, with $r = 1, \ldots, R$. We estimate Σ by the iterative formula:

$$\widehat{\Sigma}_k = \frac{1}{R} \sum_{r=1}^{R} \eta\left(\mathbf{e}_r^\top \left[\widehat{\Sigma}_{k-1}\right]^{-1} \mathbf{e}_r\right) \mathbf{e}_r \mathbf{e}_r^\top,$$

where k represents the index of this iterative process. Note that $\eta(s)$ depends on $h(s)$. More sophisticated procedures for estimating Σ can be found in [44, 33]. Finally, we remark that an ATAIS scheme could be also designed when the noise has a generalized Gaussian density. In one dimension, this pdf has the form $p(v) = \frac{\beta}{2s\Gamma(1/\beta)} \exp\left(-\frac{|v-\mu|^\beta}{s^\beta}\right)$ where, fixing μ, β, we know the estimator $\widehat{s} = \left(\frac{\beta}{N} \sum_{i=1}^{N} |x_i - \mu|^\beta\right)^{1/\beta}$.

4 Speeding up, robustness and possible alternatives

4.1 About the choice of the type of proposal density, N and T

Regarding the choice of the proposal, we need to be able to evaluate it pointwise (it must be normalized) and be able to draw samples from it. The other additional requirement is that the proposal density must have heavier tails than the target pdf, i.e., the conditional posterior $\pi_t(\boldsymbol{\theta})$. The presence of multimodality in the target pdf is not an issue, especially if the cyclic adaptation of the covariance matrix of the proposal is employed (see Section 4.2). Moreover, multiple proposal schemes can be used, to deal with multimodality (see Section 4.5).

The consistency of the ATAIS estimators is ensured when $N \rightarrow \infty$ (with finite T) or $T \rightarrow \infty$ (with

---

[5]In [43, Chapter 13], the interested reader can find the properties that the elliptical distribution must satisfy for the existence and uniqueness of the estimators of the location and scale parameters.

finite N). However, at least theoretically, in any adaptive importance sampler the choice of the value of N is more critical than T, since the adaptation of the proposal pdf is driven by the N samples for each iteration. In ATAIS, the value N is even more important, since the conditional posterior is also adapted (due to the estimation and update of $\widehat{\Sigma}_{\mathrm{ML}}^{(t)}$). Namely, the N weighted samples per iteration affect (a) the proposal adaptation and also (b) the first step of the alternating optimization in Table 1 i.e., more specifically, in step 2(b)i of Table 8. However, the simulations in Section 7.2 and Figure 8(a) show the importance of the adaptation in ATAIS and as a consequence of the number of iterations T.

4.2 Adaptation of the covariance matrix of the proposal density

The adaptation of the parameters of the proposal density is described in Eqs. (23)-(77) and (78) of Table 8. Regarding the covariance matrix Λ_t we consider Eq. (78) that is formed by two parts as follows:

$$\Lambda_{t+1} = \underbrace{\sum_{n=1}^{N} \bar{w}_t^{(n)} (\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t)^\top (\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t)}_{\text{empirical}} + \underbrace{\delta_t \mathbf{I}_M}_{\text{cyclic}}, \qquad \delta_t > 0. \tag{26}$$

The first term is the empirical covariance of the weighted samples at the t-th iteration. However, this term can be harmful to the performance, especially at the beginning of the adaptation, often yielding to empirical covariance matrices with very small traces (depending on the degeneracy of the weighted samples). Robust alternatives to this first term are discussed in [45].

The second term is a diagonal matrix controlled by a positive parameter $\delta_t > 0$ which determines the elements in the diagonal. This term helps the IS performance (see , e.g., [26, Numerical Example 1]) and avoids catastrophic scenarios. Moreover, this second term always ensures to have positive definite matrices Λ_t and avoid numerical issues. Generally, one should use a greater value of δ_t especially in the first iterations, and the value of δ_t could be decreased as the iterations grow. Generally, one should use a large value of δ_t in the initial iterations, and gradually decrease the value of δ_t as the iterations progress. For a robust implementation, here we suggest to apply a cyclic/periodic strategy, i.e.,

$$\delta_{t+1} = \begin{cases} a\, \delta_t, & \text{if } \delta_t \geq \delta_{\min}, \\ \delta_0, & \text{if } \delta_t < \delta_{\min}, \end{cases} \tag{27}$$

with $0 < a < 1$, where δ_0 is the maximum value that the user desires to use for exploration (that is also an initial value) and $\delta_{\min} > 0$ is the minimum possible value that can be employed avoiding numerical problems. The goal is to combine iterations devoted to the *exploration* (associated with greater values of δ_t, close to δ_0) and other iterations devoted to the *exploitation* (associated with smaller values of δ_t, close to $\delta_{\min}$). The coefficient a determines the speed of decrease of δ_t and, as a consequence, the number of cycles/periods. Note that, in any case, the first term which depends on the weighted samples can drive the final covariance matrix Λ_t.

4.3 Saving and re-weighting only few particles

The first part of ATAIS yields NT weighted samples. These NT samples must be re-weighted in the last step as in Eq. (24) or (79). At each iteration, we should save N weighted particles. This can

produce an important computational load. However, most of these samples have a negligible weight, especially in the first iterations, when ATAIS is still looking for the regions of high probability. The use of the resampling procedure (i.e., bootstrap) over N samples at each iteration or NT final samples is not advisable, since (if N large) it can be very slow.

We suggest to save, at each iteration t, only the (extremely) relevant samples with a normalized weight $\bar{w}_t^{(n)} \geq \frac{1}{N}$. This criterion is also connected with a possible effective sample size (ESS) measure proposed and discussed in [46]. Note that if only a few samples have a non-negligible weight (i.e., the - very frequent - degenerate case), we just save the relevant ones. On the other hand, if all samples have similar weights (and, as a consequence, all the normalized weights $\bar{w}_t^{(n)}$ are around $1/N$), this is owing to two possible scenarios:

- All the samples are located in a tail of distribution (an almost-flat region that yields similar weights). In this case, most of the samples are not relevant, and we do not lose information if saving only some of them.

- The samples are all relevant, and located in high probability regions. In this case, we lose some information (saving only some of them) but, since the proposal density is already well-adapted (i.e., well-located), in the next iterations we will draw more and more relevant samples around these high probability regions.

Moreover, for the second part of the inference in Section 5, we need to save only the computation of the error vectors $\mathbf{e}_{t,r}^{(n)} = \mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}_t^{(n)})$, e.g., see Eq. (45). Thus, no additional evaluation of the model are required in this second part. Hence, at each iteration, we suggest just to save the error vectors $\mathbf{e}_{t,r}^{(n)}$ of the particles with a normalized weight such that $\bar{w}_t^{(n)} \geq \frac{1}{N}$.

4.4 Optimal denominator in the IS weights

Since we adapt the proposal density during the iterations, we are actually in a multiple IS scenario [47, 23]. It is well-known that the standard IS denominator (using just the unique proposal $q(\theta|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)$) provides instability and high variance in the final IS estimators. The correct way of avoiding this behavior is to employ a mixture of all proposals used during the iterations, i.e.,

$$w_t^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_t^{(n)})}{\frac{1}{t}\sum_{i=1}^{t} q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)}. \tag{28}$$

This procedure provides the lowest variance of the final IS estimators but requires a high computational cost. Indeed, for each sample $\boldsymbol{\theta}_t^{(n)}$, we have to evaluate a mixture where the number of components grows with the iterations. Moreover, *at least* in the final iteration T decided by the user, all the previous weights must be updated recomputing a new denominator for each sample. Alternatives for reducing the computational cost have been proposed [48]. The simplest solution among the proposed ones is to build a *compressed* denominator [49, 23]. To avoid instabilities in the results, we can discard the samples and the proposals in the first iterations (hence, we do not use the means $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}$), when the parameters of the proposal density change substantially. For instance, one can discard the corresponding means $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}$ in the first iterations t such that $\frac{\|\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)} - \widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t-1)}\|}{\max\left[\|\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}\|, \|\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t-1)}\|\right]} > \epsilon$ where $\epsilon \in (0, 1]$.

By our empirical experience, we suggest the use of $\epsilon = 0.3$. However, we remark that the use of the weights of type (28) is not mandatory. ATAIS provides very good performance also with the standard IS weights, as shown by the numerical experiments in Section 7.

4.5   Possible use of multiple proposal densities

In order to foster the robustness of the algorithm, one possibility is to use more than one proposal density at each iteration. The simplest way for extending ATAIS with H multiple proposals is described next. Denoting each proposal as $q_h(\boldsymbol{\theta}|\boldsymbol{\mu}_{t,h}, \boldsymbol{\Lambda}_{t,h})$ with $h = 1, ..., H$, the idea is to draw N samples for each proposal pdf,

$$\boldsymbol{\theta}_{t,h}^{(1)}, ..., \boldsymbol{\theta}_{t,h}^{(N)} \sim q_h(\boldsymbol{\theta}|\boldsymbol{\mu}_{t,h}, \boldsymbol{\Lambda}_{t,h}), \qquad h = 1, ..., H,$$

and weigh them keeping a *unique* (adaptive) posterior distribution $\pi_t(\boldsymbol{\theta}) = \ell(\mathbf{Y}|\widehat{\Sigma}_{\text{ML}}^{(t-1)}, \boldsymbol{\theta})g_\theta(\boldsymbol{\theta})$, as

$$w_{t,h}^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_{t,h}^{(n)})}{q_h(\boldsymbol{\theta}_{t,h}^{(n)}|\boldsymbol{\mu}_{t,h}, \boldsymbol{\Lambda}_{t,h})}, \qquad \text{or} \qquad w_{t,h}^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_{t,h}^{(n)})}{\frac{1}{H}\sum_{j=1}^{H} q_j(\boldsymbol{\theta}_{t,j}^{(n)}|\boldsymbol{\mu}_{t,j}, \boldsymbol{\Lambda}_{t,j})}, \tag{29}$$

where, in the second option, we have employed the multiple IS approach [47, 23], similarly as in Eq. (28). Then, the idea is to perform an independent adaptation of the mean (as in [50]) considering the local MAP estimations, i.e.,

$$\boldsymbol{\mu}_{t+1,h} = \widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t,h)}, \tag{30}$$

where we consider only the samples generated by the *m*-th proposal. The same approach can be employed for adapting $\boldsymbol{\Lambda}_{t,m}$, i.e., independently to each other (see Section 4.2). Hence, in this sense, the steps described so far could be parallelized. Now, denoting as

$$\overline{\boldsymbol{\theta}}_{\text{MAP}}^{(t)} = \arg\max_h \pi_t(\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t,h)}), \tag{31}$$

the best MAP estimator at the t-*th* iteration, we can obtain a *unique* estimation of the covariance matrix,

$$\widehat{\Sigma}_{\text{ML}}^{(t)} = \frac{1}{R}\sum_{r=1}^{R}\left(\mathbf{y}_r - \mathbf{f}_r(\overline{\boldsymbol{\theta}}_{\text{MAP}}^{(t)})\right)\left(\mathbf{y}_r - \mathbf{f}_r(\overline{\boldsymbol{\theta}}_{\text{MAP}}^{(t)})\right)^\top, \tag{32}$$

so that we can have a unique posterior $\pi_{t+1}(\boldsymbol{\theta}) = \ell(\mathbf{Y}|\widehat{\Sigma}_{\text{ML}}^{(t)}, \boldsymbol{\theta})g_\theta(\boldsymbol{\theta})$ for the next iteration of the algorithm. In summary, all the proposal pdfs are adapted independently but they face the same posterior. The rest of the algorithm remains unaltered as in Table 8. More sophisticated adaptive strategies could include a "repulsion" among the proposals, as suggested in [51, 52].

4.6   Possible use of mini-batches

When the number of vectors $\mathbf{y}_r$ of data R grows, different issues appear:

- the calculation of the likelihood becomes costly/slow;

- numerical problems due to the finite precision of the employed machine;

- the exploration of the high probability regions becomes more difficult since the posterior is extremely tight.

All these issues are mitigated by using mini-batches of data. ATAIS allows the direct use of mini-batches (see [53, 54]). Namely, we can use a sub-set of data (e.g., formed by $L < R$ vectors $\mathbf{y}_i$ of data) to create sub-posteriors,

$$\widetilde{\pi}_t(\boldsymbol{\theta}) \propto \left[\prod_{i=1}^{L} \ell(\mathbf{y}_{k_i}|\Sigma_{\text{ML}}^{(t-1)}, \boldsymbol{\theta})\right] g_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \tag{33}$$

where $\mathbf{y}_{k_i} \in \{\mathbf{y}_1, ..., \mathbf{y}_R\}$, i.e., $k_i \in \{1, ..., R\}$ with $i = 1, ..., L$, are $L < R$ vectors that can be selected randomly over the R possible vectors [53, 54]. ATAIS can use the subposteriors, defined as in Eq. (33), in each step 2(b)i of Table 8, i.e.,

$$\boldsymbol{\theta}_{\text{max}}^{(t)} = \arg\max_n \widetilde{\pi}_t(\boldsymbol{\theta}_t^{(n)}), \qquad \widehat{\mathbf{r}}_t = \mathbf{f}_r(\boldsymbol{\theta}_{\text{max}}^{(t)}), \qquad \widehat{\Sigma}_t = \frac{1}{L}\sum_{j=1}^{L}(\mathbf{y}_j - \widehat{\mathbf{r}}_t)(\mathbf{y}_j - \widehat{\mathbf{r}}_t)^\top, \tag{34}$$

using only the L selected data vectors. Below, we describe two possible schemes for modifying ATAIS with mini-batches.

First possible strategy. In the standard ATAIS, we can perform the check in step 2c, obtaining a sequence of $\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}$ and $\widehat{\Sigma}_{\text{ML}}^{(t)}$. Clearly, the proposal density $q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{MAP}}^{(t)}, \Lambda_t)$ is adapted as in Eqs. (77)-(78). After all the T iterations, we consider the computation of the complete conditional posterior in Eq. (22) (with *all* the data) is feasible (at least for a few points). We can compute

$$\pi_t(\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}) \propto p(\boldsymbol{\theta}_{\text{MAP}}^{(t)}|\widehat{\Sigma}_{\text{ML}}^{(t)}, \mathbf{Y}),$$

and

$$\overline{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max\left\{\pi_t\left(\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}\right), \text{ for all } t = 1, ..., T\right\}. \tag{35}$$

The final estimation of the matrix can be done

$$\overline{\Sigma}_{\text{ML}} = \frac{1}{R}\sum_{r=1}^{R}\left(\mathbf{y}_r - \mathbf{f}_r(\overline{\boldsymbol{\theta}}_{\text{MAP}})\right)\left(\mathbf{y}_r - \mathbf{f}_r(\overline{\boldsymbol{\theta}}_{\text{MAP}})\right)^\top, \tag{36}$$

considering all the R data vectors. Moreover, in the second part of the proposed inference scheme (see below in the next section), the final re-weighting step must be made according to the full posterior, so that the final estimations are performed considering all the dataset. Hence, in this strategy, the complete ATAIS method can consider the use of mini-batches only in the T iterations of the first part, whereas, after the T iterations for computing above and in the second part, the full-posterior must be evaluated.

Second possible strategy. Otherwise, if the computation of the complete conditional posteriors is too costly even for a few points, we can follow the approach suggested in [55]. Let us consider the following assumptions:

15

- Assume that all the data are divided into $\frac{R}{L}$ disjoint subsets (so that the union of these subsets includes all the data) so that the complete posterior can be expressed as the product of the $\frac{R}{L}$ sub-posteriors, and to select the total number of iterations as $T = \frac{R}{L}$.

- Assume that, at least after a certain iteration, the variations in $\widehat{\Sigma}_{\mathrm{ML}}^{(t)}$ are small, so that $\widehat{\pi}_{t-1}(\boldsymbol{\theta}) \approx \widehat{\pi}_t(\boldsymbol{\theta}) \approx \widehat{\pi}_{t+1}(\boldsymbol{\theta})$... and so on, i.e., the conditional full-posterior is virtually not changing. This assumption is verified when $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(t)}$ is close to the true vector $\boldsymbol{\theta}_{\mathrm{MAP}}$.

- Assume as prior in each sub-posterior, $g_{\mathrm{sub}}(\boldsymbol{\theta}) = g(\boldsymbol{\theta})^{1/T}$.

Then, the conditional full-posterior (that virtually is not changing) can be expressed as the product of the $T = \frac{R}{L}$ (conditional) sub-posteriors. At each iteration of ATAIS, we can consider a Gaussian approximation of each sub-posterior, $\widehat{p}_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}_{\mathrm{max}}^{(t)}, \Lambda_t)$ (with δ_t in Section 4.2 is small, close to zero), with mean $\boldsymbol{\theta}_{\mathrm{max}}^{(t)}$ and covariance matrix Λ_t, both computed according to the sub-posterior $\widetilde{\pi}_t(\boldsymbol{\theta})$ in Eq. (33) considering the current mini-batch. Hence an approximation of conditional full-posterior can be also a Gaussian distribution,

$$\widehat{\pi}_t(\boldsymbol{\theta}) = \prod_{\tau}^{t} \widehat{p}_t(\boldsymbol{\theta}) \propto \mathcal{N}\left(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(t)}, \Lambda_{\mathrm{tot}}^{(t)}\right), \tag{37}$$

with covariance matrix

$$\Lambda_{\mathrm{tot}}^{(t)} = \left(\sum_{\tau=1}^{t} \Lambda_{\tau}^{-1}\right)^{-1}, \tag{38}$$

and mean

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(t)} = \Lambda_{\mathrm{tot}} \sum_{\tau=1}^{t} \Lambda_{\tau}^{-1} \boldsymbol{\theta}_{\mathrm{max}}^{(\tau)}. \tag{39}$$

Note that, we can also have

$$\widehat{\Sigma}_{\mathrm{ML}}^{(t)} = \frac{1}{R} \sum_{r=1}^{R} \left(\mathbf{y}_r - \mathbf{f}_r(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(t)})\right)\left(\mathbf{y}_r - \mathbf{f}_r(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(t)})\right)^{\top}. \tag{40}$$

Therefore, we can replace the check in step 2c of Table 8 with the computation of $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(t)}$ above $\widehat{\Sigma}_{\mathrm{ML}}^{(t)}$. Note that, unlike the proposal $q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\mathrm{MAP}}^{(t)}, \Lambda_t)$ is still adapted as in Eqs. (77)-(78). The final Gaussian approximation with $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(T)}$, $\Lambda_{\mathrm{tot}}^{(T)}$ and the final estimation $\widehat{\Sigma}_{\mathrm{ML}}^{(T)}$, can be also employed in the second part of the inference scheme instead of evaluating the true full-posterior. Thus, in this second approach, we never need the evaluation of the (true) full-posterior.

4.7 Possible use of the gradient information if the likelihood is differentiable

The current version of ATAIS can be applied even when the model $\mathbf{f}(\boldsymbol{\theta})$ is not differentiable. Moreover, the Monte Carlo approach ensures the convergence to the global maximum as $N \to \infty$ or $T \to \infty$. However, if $\mathbf{f}(\boldsymbol{\theta})$ is differentiable and the practitioner desires to incorporate the gradient information several alternatives strategies can be employed. The simplest idea is to replace the step 2(b)i of Table 8, by a gradient descent step over the function

$$C(\boldsymbol{\theta}) = -\log \pi_t(\boldsymbol{\theta}) = -\log \ell(\mathbf{Y}|\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(t-1)}, \boldsymbol{\theta}) g_\theta(\boldsymbol{\theta}),$$

i.e.,

$$\boldsymbol{\theta}_{\text{max}}^{(t)} = \boldsymbol{\theta}_{\text{MAP}}^{(t-1)} - \gamma \nabla C\left(\boldsymbol{\theta}_{\text{MAP}}^{(t-1)}\right) + e_t,$$

with $\gamma > 0$ and e_t is a noise perturbation, avoiding to remain stuck in a local minimum. Thus, $\boldsymbol{\theta}_{\text{max}}^{(t+1)}$ would be a candidate as a possible new mean of the proposal pdf at step 2c of Table 8, i.e., $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)} = \boldsymbol{\theta}_{\text{max}}^{(t)}$, if $\pi_t(\boldsymbol{\theta}_{\text{max}}^{(t)}) > \pi_{\text{MAP}}$, (where π_{MAP} is the greatest value so far obtained in Eq. (22), considering also the corresponding estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$ is also changing). See [56], for all the formulas required in the case of Gaussian noise perturbation, as in Eq. (5). Clearly, it is also possible to use a stochastic gradient descent using mini-batches [57]. More sophisticated gradient-based IS schemes have been discussed in literature [51], also involving multiple proposal densities and repulsion [51, 58, 59, 56].

5 Second part of the proposed inference scheme

Here, we describe the second part of the ATAIS procedure, which allows a complete Bayesian analysis of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$. It is important to remark that this second part of ATAIS does not require any additional sample generation and likelihood evaluation. Indeed, ATAIS recycles and reweights the samples $\boldsymbol{\theta}_t^{(n)}$ obtained in the first part by Table 8.

5.1 Approximating different conditional posteriors

The idea here is to re-use all the generated samples since if we have saved the computation of the error vectors $\mathbf{e}_{t,r}^{(n)} = \mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}_t^{(n)})$ no additional evaluation of the model are required. Note that the cloud of particles $\{\boldsymbol{\theta}_t^{(n)}\}$ is well-located since ATAIS works to generate samples around the MAP and ML estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$. Moreover, we can also use $\{\mathbf{e}_{t,r}^{(n)}\}$ and $\{\boldsymbol{\theta}_t^{(n)}\}$ for building a particle approximation of any other conditional posterior $p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma})$, i.e.,

$$\widehat{p}(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma}) = \sum_{t=1}^{T} \sum_{n=1}^{N} \bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}) \cdot \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)}), \qquad \sum_{t=1}^{T} \sum_{n=1}^{N} \bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}) = 1, \qquad (41)$$

where

$$\rho_t^{(n)}(\Sigma) = \frac{\ell(\mathbf{Y}|\boldsymbol{\theta}_t^{(n)}, \Sigma)g_\theta(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \Lambda_t)}, \quad \text{and} \tag{42}$$

$$\bar{\rho}_t^{(n)}(\Sigma) = \frac{\rho_t^{(n)}(\Sigma)}{\sum_{\tau=1}^T \sum_{i=1}^N \rho_\tau^{(i)}(\Sigma)}. \tag{43}$$

Given a new matrix Σ, to compute the likelihood

$$\ell(\mathbf{Y}|\boldsymbol{\theta}_t^{(n)}, \Sigma) = \left(\frac{1}{(2\pi)^{K/2}\det(\Sigma)^{1/2}}\right)^R \exp\left(-\frac{1}{2}\sum_{r=1}^R \left(\mathbf{y}_r - \mathbf{f}_r\left(\boldsymbol{\theta}_t^{(n)}\right)\right)^\top \Sigma^{-1} \left(\mathbf{y}_r - \mathbf{f}_r\left(\boldsymbol{\theta}_t^{(n)}\right)\right)\right), \tag{44}$$

$$= \left(\frac{1}{(2\pi)^{K/2}\det(\Sigma)^{1/2}}\right)^R \exp\left(-\frac{1}{2}\sum_{r=1}^R \left(\mathbf{e}_{t,r}^{(n)}\right)^\top \Sigma^{-1} \left(\mathbf{e}_{t,r}^{(n)}\right), \right) \tag{45}$$

we need the vectors $\mathbf{e}_{t,r}^{(n)}$, the inverse matrix of Σ and the determinant of Σ.

5.2 Approximation of the complete posterior distribution and marginal likelihood

We can apply an IS scheme with the complete target pdf,

$$p(\boldsymbol{\theta}, \Sigma|\mathbf{Y}) = \frac{p(\boldsymbol{\theta}, \Sigma, \mathbf{Y})}{p(\mathbf{Y})} = \frac{\ell(\mathbf{Y}|\boldsymbol{\theta}, \Sigma)g_\theta(\boldsymbol{\theta})g_\Sigma(\Sigma)}{p(\mathbf{Y})}, \tag{46}$$

$$\propto \ell(\mathbf{Y}|\boldsymbol{\theta}, \Sigma)g_\theta(\boldsymbol{\theta})g_\Sigma(\Sigma), \tag{47}$$

and employing a proposal density that can be factorized as $q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \Lambda_t)q_\Sigma(\Sigma)$ where the piece of proposal $q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \Lambda_t)$ is the same used in ATAIS at the t-th iteration. Recycling the NT samples produced by ATAIS, i.e., $\boldsymbol{\theta}_t^{(n)} \sim q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \Lambda_t)$ and drawing J random matrices from the proposal $q_\Sigma(\Sigma)$, i.e., $\Sigma^{(j)} \sim q_\Sigma(\Sigma)$, the complete IS weights are

$$\beta_{t,j}^{(n)} = \frac{\ell(\mathbf{Y}|\boldsymbol{\theta}_t^{(n)}, \Sigma^{(j)})g_\theta(\boldsymbol{\theta}_t^{(n)})g_\Sigma(\Sigma^{(j)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \Lambda_t)q_\Sigma(\Sigma^{(j)})}, \tag{48}$$

$$= \rho_t^{(n)}(\Sigma^{(j)})\frac{g_\Sigma(\Sigma^{(j)})}{q_\Sigma(\Sigma^{(j)})} = \rho_t^{(n)}(\Sigma^{(j)})\gamma_j = \rho_{t,j}^{(n)}\gamma_j, \tag{49}$$

where we have set $\gamma_j = \frac{g_\Sigma(\Sigma^{(j)})}{q_\Sigma(\Sigma^{(j)})}$ and $\rho_{t,j}^{(n)} = \rho_t^{(n)}(\Sigma^{(j)})$ are given in Eq. (42). Clearly, if $q_\Sigma(\Sigma) = g_\Sigma(\Sigma)$ then $\gamma_j = 1$. The complete posterior approximation is given by

$$\widehat{p}(\boldsymbol{\theta}, \Sigma|\mathbf{Y}) = \sum_{j=1}^J \sum_{t=1}^T \sum_{n=1}^N \bar{\beta}_{t,j}^{(n)} \cdot \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)})\delta\left(\Sigma - \Sigma^{(j)}\right), \tag{50}$$

where $\bar{\beta}_{t,j}^{(n)} = \frac{\beta_{t,j}^{(n)}}{\sum_{i=1}^{J} \sum_{v=1}^{T} \sum_{m=1}^{N} \beta_{v,i}^{(m)}}$. Note that we have different numbers of samples about $\boldsymbol{\theta}$ (i.e., NT) and $\boldsymbol{\Sigma}$ (i.e., J). This recalls the recycling Gibbs sampling idea in [7], where the space is divided into blocks, and different numbers of samples are considered for each block.

The marginal likelihood $p(\mathbf{Y})$ can be approximated as

$$p(\mathbf{Y}) \approx \widehat{p}(\mathbf{Y}) = \frac{1}{JNT} \sum_{j=1}^{J} \sum_{t=1}^{T} \sum_{n=1}^{N} \beta_{t,j}^{(n)}. \tag{51}$$

5.3 Approximation of the marginal posteriors

An approximation of the marginal posterior distribution of $\boldsymbol{\theta}$ can be obtained

$$p(\boldsymbol{\theta}|\mathbf{Y}) = \int_{\Sigma} p(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathbf{Y})d\boldsymbol{\Sigma} \approx \widehat{p}(\boldsymbol{\theta}|\mathbf{Y}) = \sum_{t=1}^{T} \sum_{n=1}^{N} \bar{\alpha}_t^{(n)} \cdot \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)}), \tag{52}$$

where

$$\bar{\alpha}_t^{(n)} = \frac{\sum_{j=1}^{J} \beta_{t,j}^{(n)}}{\sum_{i=1}^{J} \sum_{v=1}^{T} \sum_{m=1}^{N} \beta_{v,i}^{(m)}}. \tag{53}$$

Moreover, we can assign a weight to each drawn matrix above $\boldsymbol{\Sigma}^{(j)}$, approximating the marginal posterior of the covariance matrix

$$p(\boldsymbol{\Sigma}^{(j)}|\mathbf{Y}) = \int_{\Theta} p(\boldsymbol{\theta}, \boldsymbol{\Sigma}^{(j)}|\mathbf{Y})d\boldsymbol{\theta} \approx \frac{\sum_{t=1}^{T} \sum_{n=1}^{N} \beta_{t,j}^{(n)}}{\sum_{i=1}^{J} \sum_{v=1}^{T} \sum_{m=1}^{N} \beta_{v,i}^{(m)}} \delta\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(j)}\right), \tag{54}$$

$$= \bar{\lambda}_j \cdot \delta\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(j)}\right), \tag{55}$$

where

$$\bar{\lambda}_j = \frac{\sum_{t=1}^{T} \sum_{n=1}^{N} \beta_{t,j}^{(n)}}{\sum_{i=1}^{J} \sum_{v=1}^{T} \sum_{m=1}^{N} \beta_{v,i}^{(m)}}. \tag{56}$$

Then, the marginal posterior of the covariance matrix is approximated as

$$p(\boldsymbol{\Sigma}|\mathbf{Y}) \approx \widehat{p}(\boldsymbol{\Sigma}|\mathbf{Y}) = \sum_{j=1}^{J} \bar{\lambda}_j \cdot \delta\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(j)}\right). \tag{57}$$

For instance, a minimum mean square error estimator of $\boldsymbol{\Sigma}$ can be approximated as

$$\widehat{\boldsymbol{\Sigma}} = \sum_{j=1}^{J} \bar{\lambda}_j \boldsymbol{\Sigma}^{(j)},$$

and approximations of high-order moments $p(\boldsymbol{\Sigma}|\mathbf{Y})$ can also be obtained. Table 2 summarizes all the weights and the corresponding distributions.

Table 2: Summary of the weights and the corresponding distributions.

Distribution to approximate	Normalized weights	Additional information
$p(\boldsymbol{\theta}\|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}}^{(T)})$	$\widetilde{w}_t^{(n)}$ – See Eqs. (20) and (24)	
$p(\boldsymbol{\theta}\|\mathbf{Y}, \boldsymbol{\Sigma})$	$\bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}) = \frac{\ell(\mathbf{Y}\|\boldsymbol{\theta}_t^{(n)}, \boldsymbol{\Sigma}) g_\theta(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}\|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)}$	
$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}\|\mathbf{Y})$	$\bar{\beta}_{t,j}^{(n)} = \frac{\beta_{t,j}^{(n)}}{\sum_{i=1}^{J}\sum_{v=1}^{T}\sum_{m=1}^{N}\beta_{v,i}^{(m)}}$	$\beta_{t,j}^{(n)} = \frac{\ell(\mathbf{Y}\|\boldsymbol{\theta}_t^{(n)}, \boldsymbol{\Sigma}^{(j)}) g_\theta(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}\|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)} \cdot \frac{g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}{q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}$
$p(\boldsymbol{\theta}\|\mathbf{Y})$	$\bar{\alpha}_t^{(n)} = \frac{\sum_{j=1}^{J}\beta_{t,j}^{(n)}}{\sum_{i=1}^{J}\sum_{v=1}^{T}\sum_{m=1}^{N}\beta_{v,i}^{(m)}}$	$\beta_{t,j}^{(n)} = \rho_t^{(n)}(\boldsymbol{\Sigma}^{(j)}) \cdot \gamma_j$
$p(\boldsymbol{\Sigma}\|\mathbf{Y})$	$\bar{\lambda}_j = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\beta_{t,j}^{(n)}}{\sum_{i=1}^{J}\sum_{v=1}^{T}\sum_{m=1}^{N}\beta_{v,i}^{(m)}}$	
$p(\mathbf{Y})$	$\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{n=1}^{N}\beta_{t,j}^{(n)}$	

5.4 Prior and proposal densities over covariance matrices

Consider a positive definite $K \times K$ matrix $\boldsymbol{\Sigma}$. The Wishart distribution is defined on the space $\mathbb{R}^K \times \mathbb{R}^K$ of positive definite matrices. The corresponding pdf is

$$g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) = g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}|\boldsymbol{\Phi}, \nu) \propto |\boldsymbol{\Sigma}|^{\frac{\nu - K - 1}{2}} \exp\left(-\frac{1}{2}\text{trace}(\boldsymbol{\Phi}^{-1}\boldsymbol{\Sigma})\right), \tag{58}$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of the matrix $\boldsymbol{\Sigma}$, $\nu \geq K - 1$ is the number of degrees of freedom and $\boldsymbol{\Phi}$ is a $K \times K$ *reference* covariance matrix. It is possible to see $E_g[\boldsymbol{\Sigma}] = \nu\boldsymbol{\Phi}$. The Wishart distribution is often interpreted as a multivariate extension of the χ^2 distribution.

Choice of $\boldsymbol{\Phi}$ and ν. We choose

$$\boldsymbol{\Phi} = \frac{1}{\nu}\widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}}^{(T)}. \tag{59}$$

Recall that $\mathbb{E}_g[\boldsymbol{\Sigma}] = \nu\boldsymbol{\Phi}$. In this sense, our approach is an *empirical Bayes scheme* since this parameter of the prior is chosen after looking at the data by ATAIS (see also data-based priors in [27]). The parameter ν represents the degrees of freedom of the distribution. This value must be $\nu \geq K - 1$, but for the generated matrices to be non-singular with probability 1 we need $\nu \geq K$. For learning ν, we can use again an empirical Bayes approach maximizing the marginal likelihood $p(\mathbf{Y}) = p(\mathbf{Y}|\nu)$ in Eq. (51), i.e., we can find the ν^* such that $\nu^* = \arg\max p(\mathbf{Y}|\nu)$.

Choice of the proposal pdf. For simplicity, we assume $q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) = g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$, i.e., we choose a proposal density equal to the prior density. As a consequence, with this choice, we have $\gamma_j = 1$ in (49).

Generation of random matrices according to a Wishart density. When v is an integer, the Wishart distribution represents the sums of squares (and cross-products) of v vectors drawn from a multivariate Gaussian distribution. Specifically, given v random vectors of dimension $K \times 1$, i.e. $\mathbf{s}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Phi})$, $i = 1, \ldots, v$, the generated matrix

$$\Sigma' = \sum_{i=1}^{v} \mathbf{s}_i \mathbf{s}_i^\top,$$

is distributed as a Wishart density with v degrees of freedom and $K \times K$ scale matrix $\mathbf{\Phi}$. Then, we can employ the following sampling method:

1. Draw v multivariate Gaussian samples $\mathbf{s}_i = [s_{i,1}, \ldots, s_{i,K}]^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{\Phi})$, with $i = 1, \ldots, v$.

2. Set $\Sigma' = \sum_{i=1}^{v} \mathbf{s}_i \mathbf{s}_i^\top$.

6 Inverted layered importance sampling (ILIS)

The direct application of Monte Carlo methods in the complete space of $\boldsymbol{\theta}$ and Σ generally does not provide good results. This is the reason why we propose the ATAIS algorithm where the inference is carried out in two phases, first a set of weighted samples $\boldsymbol{\theta}_t^{(n)}$ and $\widehat{\Sigma}_{\text{ML}}$ are obtained. Then the samples $\boldsymbol{\theta}_t^{(n)}$ are re-weighted and other weighted samples of matrices $\Sigma^{(j)}$ are generated to perform a complete Bayesian inference.

In this section, we introduce an alternative method, conceptually simpler than ATAIS, called inverted layered importance sampling (ILIS), based also on the division of the space, $\boldsymbol{\theta}$ and Σ. ILIS can perform better than other Monte Carlo methods working in the complete space, but also presents some clear difficulties: this also emphasizes the relevance of ATAIS. See Table 3 for a detailed description. ILIS starts by generating N covariance matrices $\Sigma^{(n)}$ and then it runs N different (parallel and independent) MCMC chains with target density $p(\boldsymbol{\theta}|\Sigma^{(n)}, \mathbf{Y})$, i.e., the conditional posterior of $\boldsymbol{\theta}$ given $\Sigma^{(n)}$. The underlying idea is to apply an IS scheme on the complete space $\{\boldsymbol{\theta}, \Sigma\}$, i.e., considering the complete posterior $p(\boldsymbol{\theta}, \Sigma|\mathbf{Y})$ in Eq. (46), that is

$$p(\boldsymbol{\theta}, \Sigma|\mathbf{Y}) = \frac{1}{p(\mathbf{Y})} \ell(\mathbf{Y}|\boldsymbol{\theta}, \Sigma) g_{\theta}(\boldsymbol{\theta}) g_{\Sigma}(\Sigma) \propto \ell(\mathbf{Y}|\boldsymbol{\theta}, \Sigma) g_{\theta}(\boldsymbol{\theta}) g_{\Sigma}(\Sigma).$$

Moreover, we assume to employ a proposal density of type $q(\boldsymbol{\theta}, \Sigma) = q_{\theta}(\boldsymbol{\theta}|\Sigma) q_{\Sigma}(\Sigma)$, Recall that q, q_{θ} and q_{Σ} must be normalized. The corresponding importance weight function in the full space is

$$w(\boldsymbol{\theta}, \Sigma) = \frac{\ell(\mathbf{Y}|\boldsymbol{\theta}, \Sigma) g(\boldsymbol{\theta}) g(\Sigma)}{q_{\theta}(\boldsymbol{\theta}|\Sigma) q_{\Sigma}(\Sigma)}. \tag{60}$$

One good choice for q_{θ} would be

$$q_{\theta}(\boldsymbol{\theta}|\Sigma) = p(\boldsymbol{\theta}|\Sigma, \mathbf{Y}) = \frac{1}{Z(\Sigma)} \ell(\mathbf{Y}|\boldsymbol{\theta}, \Sigma) g_{\theta}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) = \ell(\mathbf{Y}|\boldsymbol{\theta}, \Sigma) g_{\theta}(\boldsymbol{\theta}),$$

where $Z(\Sigma) = \int_{\Theta} \ell(\mathbf{Y}|\boldsymbol{\theta}, \Sigma) g_{\theta}(\boldsymbol{\theta}) d\boldsymbol{\theta}$. This proposal is suggested and inspired by the success of the ATAIS approach. However, the choice above presents two issues:

- We need to be able to draw from. We can use an MCMC chain for generating samples distributed as $p(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \mathbf{Y})$, after a burn-in period [3, 60, 61];

- We need to be able to evaluate $Z(\boldsymbol{\Sigma})$, which is generally unknown. Thus, an estimation of this normalizing constant, $\widehat{Z}(\boldsymbol{\Sigma}) \approx Z(\boldsymbol{\Sigma})$, is required [62, 26].

Indeed, in this case, the importance weight would be:

$$w(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{\ell(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma})g(\boldsymbol{\theta})g(\boldsymbol{\Sigma})}{\frac{1}{Z(\boldsymbol{\Sigma})}\ell(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma})g_\theta(\boldsymbol{\theta})q_\Sigma(\boldsymbol{\Sigma})}, \tag{61}$$

$$= Z(\boldsymbol{\Sigma})\frac{g(\boldsymbol{\Sigma})}{q_\Sigma(\boldsymbol{\Sigma})}, \tag{62}$$

and the evaluation of $Z(\boldsymbol{\Sigma})$ is required. Other important considerations are remarked below:

- For each $\boldsymbol{\Sigma}^{(n)} \sim q_\Sigma(\boldsymbol{\Sigma})$, we need an MCMC chain to draw from $q_\theta(\boldsymbol{\theta}|\boldsymbol{\Sigma}) = p(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \mathbf{Y})$. Each MCMC algorithm produces a chain of T vectors, i.e., $\boldsymbol{\theta}_1^{(n)}, ...\boldsymbol{\theta}_T^{(n)}$. All these vector are weighted with the same weight $\gamma_n = Z(\boldsymbol{\Sigma}^{(n)})\frac{g_\Sigma(\boldsymbol{\Sigma}^{(n)})}{q_\Sigma(\boldsymbol{\Sigma}^{(n)})}$.

- ILIS can be seen as a Monte Carlo scheme that combines IS and MCMC techniques, based on *two layers* (a global IS scheme that employs internally MCMC chains). With respect to the techniques in [22, 23], we can interpret that the IS part forms the upper layer (deciding the parameter $\boldsymbol{\Sigma}^{(n)}$ to be used in $q_\theta(\boldsymbol{\theta}|\boldsymbol{\Sigma}^{(n)}) = p(\boldsymbol{\theta}|\boldsymbol{\Sigma}^{(n)}, \mathbf{Y})$), whereas the MCMC chains are generated in the lower layer of ILIS with target/invariant densities $p(\boldsymbol{\theta}|\boldsymbol{\Sigma}^{(n)}, \mathbf{Y})$ (note that this pdf plays the role of a proposal q_θ in the global IS scheme). All the states of one chain are weighted with the weight $\gamma_n = Z(\boldsymbol{\Sigma}^{(n)})\frac{g_\Sigma(\boldsymbol{\Sigma}^{(n)})}{q_\Sigma(\boldsymbol{\Sigma}^{(n)})}$ (i.e., the layers are switched with respect to [22, 23]).

- Note that again, as in ATAIS, we have a different number of samples with respect to $\boldsymbol{\theta}$ (i.e., NT), and with respect to $\boldsymbol{\Sigma}$ (i.e., N), which recalls the recycling Gibbs scheme in [7].

The complete posterior approximation by ILIS is given by

$$\widehat{p}(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathbf{Y}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\bar{\gamma}_n \cdot \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)})\delta\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(n)}\right), \quad \text{with} \quad \bar{\gamma}_n = \frac{\gamma_n}{\sum_{i=1}^{N}\gamma_i}. \tag{63}$$

Even if ILIS seems at least theoretically simpler than ATAIS, we need to compute the approximations of $Z(\boldsymbol{\Sigma}^{(n)})$ and its performance is not particularly good compared to ATAIS, as we show in the numerical simulations. The main reason is that ATAIS, in the first part of the inference, finds the regions of high probability and also uses this information to tune the proposal $q_\Sigma(\boldsymbol{\Sigma})$ for the matrix generation, in the second part.

Joint use of ATAIS and ILIS. One interested practitioner could use the first part of ATAIS in Table 8 "to feed" with good parameters the proposal density $q_\Sigma(\boldsymbol{\Sigma})$ and the proposals used inside the N parallel MCMC chains (i.e., design good proposals for ILIS). However, with respect to the complete ATAIS, this scheme has a higher computational cost in terms of generated samples and likelihood evaluation. Indeed, recall that the second part of ATAIS does not require any additional sample generation and likelihood evaluation.

22

Table 3: Inverted layered importance sampling (ILIS)

---

1. Choose N, T, $\boldsymbol{\mu}_0$ and $q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$ as well as the the proposal densities and structure of M possibly different MCMC methods.

2. For $j = 1, \ldots, J$:

   (a) Generate $\boldsymbol{\Sigma}^{(j)} \sim q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$.

   (b) Generate N different MCMC chains of length T obtaining $\{\boldsymbol{\theta}_t^{(n)}\}_{t=1}^T$, with the conditional distribution
   $$\bar{\pi}(\boldsymbol{\theta}|\boldsymbol{\Sigma}^{(j)}, \mathbf{Y}) \propto \ell(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma}^{(j)})g_\theta(\boldsymbol{\theta}),$$
   as a target density.

   (c) Approximate $Z(\boldsymbol{\Sigma}^{(j)})$, using some procedure for estimation normalizing constants [62, 26], i.e., obtain $\widehat{Z}(\boldsymbol{\Sigma}^{(j)}) \approx Z(\boldsymbol{\Sigma}^{(j)}) = \int_\Theta \ell(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma}^{(j)})g_\theta(\boldsymbol{\theta})d\boldsymbol{\theta}$.

   (d) Assign to each pair $\{\boldsymbol{\theta}_t^{(n)}, \boldsymbol{\Sigma}^{(n)}\}$ the weight $\gamma_j = \widehat{Z}(\boldsymbol{\Sigma}^{(j)})\dfrac{g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}{q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}$, for all $j = 1, ..., J$ and $t = 1, ..., T$.

---

7   Simulations

We test the proposed scheme in numerous different numerical examples, comparing it with different benchmark schemes. We have also tested several possible scenarios and variants, such as different noise perturbations in the observation model and the use of mini-batches. The dimension of the space is:

- $D = 8$ for the model in Section 7.1,

- $D = 12$ for the model in Section 7.2,

- $D = 7$ for the model in Section 7.5,

- and $D = 59$ for the model in Section 7.6.

We recall that, when a Monte Carlo method is tested, we should know some groundtruths of the target density, i.e., some true features of the posterior distribution (such as modes, expected values, variances, and any other moments). However, they are generally unknown even in experiments with artificial data. Hence, in many cases, the only possibility for the authors/practitioners is to consider as groundtruths the true values used for generating the data, i.e., $\boldsymbol{\theta}_{\text{true}}$ and $\boldsymbol{\Sigma}_{\text{true}}$ here. Clearly, if the number of data is not enough $\boldsymbol{\theta}_{\text{true}}$ and $\boldsymbol{\Sigma}_{\text{true}}$ can be quite different from the true groundtruth values (i.e., features of the posterior). When the example is low-dimensional, an alternative is to employ an expensive grid (or another determinist approach) for approximating the true features of the posterior

(modes, expected values etc.), and use those approximated values as groundtruths. In this work, for Σ, we have another possible groundtruth for the ML estimation which is the following:

$$\Sigma_{\text{ML}} \approx \frac{1}{R} \sum_{r=1}^{R} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}_{\text{true}})) (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}_{\text{true}}))^\top, \tag{64}$$

where $\boldsymbol{\theta}_{\text{true}}$ are the true values used for generating the data, according to the model considered in the specific example (recall that we do not know the true values of $\boldsymbol{\theta}_{\text{MAP}}$). Clearly, as stated before, Σ_{true} can be also employed as groundtruth, and the difference between Σ_{ML} and Σ_{true} is minimal, if we are analyzing enough data.

We average all the results over different independent runs. At each run, the initializations of the ATAIS parameters (and other schemes) are chosen randomly around a vector of zeros for $\boldsymbol{\theta}$ and an identity matrix for Σ (with a standard deviation of 4; for the covariance matrix Σ, we always consider an initial diagonal matrix with the same random element in the diagonal).

7.1 Localization in wireless sensor network - dimension of the space $D = 8$

In this experiment, we aim to determine the location of a target based on wireless sensor measurements. We can represent the target position as a random vector $\boldsymbol{\theta} \in \mathbb{R}^2$. We have a wireless network of $K = 3$ sensors, whose positions are known and labeled as $\mathbf{s}_1, ..., \mathbf{s}_K$. We collect R measurements from each sensor, and these measurements follow a certain distribution. Let's recall that each observation has the following form:

$$y_r = \mathbf{f}_r(\boldsymbol{\theta}) + \mathbf{v}_r, \tag{65}$$

with $\mathbf{f}_r : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ given by:

$$\mathbf{f}_r(\boldsymbol{\theta}) = \left[-A \log(\| \boldsymbol{\theta} - \mathbf{s}_1 \|^2), \ -A \log(\| \boldsymbol{\theta} - \mathbf{s}_2 \|^2), \ -A \log(\| \boldsymbol{\theta} - \mathbf{s}_3 \|^2) \right], \tag{66}$$

i.e.,

$$f_{r,i}(\boldsymbol{\theta}) = -A \log(\| \boldsymbol{\theta} - \mathbf{s}_i \|^2),$$

for $i = 1, 2, 3$. The error term is $\mathbf{v_r} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\text{true}})$, with $\Sigma_{\text{true}} \in \mathbb{R}^{3 \times 3}$ being a diagonal matrix with diagonal elements 1, 2 and 3. The parameter A is a constant that determines the rate at which the signal strength decreases with distance and is fixed at 10. This value can be influenced by various factors, such as environmental conditions or manufacturing processes. The values of variance of the sensors, as stated before, is unknown for each sensor.

We consider a scenario with $K = 3$ sensors, which makes the complete dimension of the problem to be

$$D = M + \frac{K(K+1)}{2} = 2 + \frac{3(3+1)}{2} = 8.$$

The positions of these sensors are given by: $\mathbf{s}_1 = [0.5, 1]$, $\mathbf{s}_1 = [3.5, 1]$ and $\mathbf{s}_3 = [2, 3]$. The positions of the target (and parameter we want to estimate) is $\boldsymbol{\theta}_{\text{true}} = [2.5, 2]$. In this scenario, 50 observation vectors were generated. For comparison purposes, the prior over $\boldsymbol{\theta}$ was set as uniform, i.e., $g(\boldsymbol{\theta}) \propto 1$.

Testing ATAIS. In ATAIS, we set $T = 40$ iterations and $N = 50$ particles per iteration. We consider a Gaussian proposal density for the $\boldsymbol{\theta}$-space with initial mean $[0, 0]^\top$, with a diagonal initial covariance matrix of $6\mathbf{I}_2$. The initial covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(0)}$ is set to the identity matrix $\mathbf{I}_3$. All the results are averaged over 10^3 independent runs. First, in Figure 2(a), we provide the mean absolute error (MAE) versus N with $T = 50$ and $T = 100$ (two curves), obtained by ATAIS in estimating in the estimation of $\boldsymbol{\theta}_{\text{MAP}}$ (using $\boldsymbol{\theta}_{\text{true}}$ as the groundtruth). The error decreases as we increase the number of samples for both values of $T = 50$ and $T = 100$. In Figure 2(b), we give the MAE versus T with $N = 50$ and $N = 100$. Again, we obtain the expected behavior: the error decreases as T grows and/or N increases. Additionally, other MAE values (also in estimating the covariance matrix $\boldsymbol{\Sigma}$) obtained by ATAIS are given in Table 4. Recall that we consider $\boldsymbol{\theta}_{\text{true}}$ and $\boldsymbol{\Sigma}_{\text{ML}}$ - computed as in Eq (64) - as the groundtruths of this experiment. In this table, we show the two partial MAEs for $\boldsymbol{\theta}_{\text{true}}$ and $\boldsymbol{\Sigma}_{\text{ML}}$, and the MAE in the complete space, as well. Figure 3 illustrates the evolution of the components of $\widehat{\boldsymbol{\theta}}_{\text{MAP}}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}$ with the iterations, in one run of ATAIS.

Comparison with ILIS and a "helped" MH. We compare the ATAIS results with (a) ILIS using Metropolis Hastings (MH) chains and (b) a unique MH chain working only in the $\boldsymbol{\theta}$-space, keeping fixed the covariance matrix to the maximum likelihood estimation $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(T)}$ obtained by ATAIS in its first part. Namely, this MH is helped by using and keeping $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(T)}$ given by ATAIS and works only in the $\boldsymbol{\theta}$-space. For ILIS, we consider as prior and proposal density for the covariance matrices a Wishart distribution with $\nu = 4$ degrees of freedom and a reference matrix $\Phi = 3\mathbf{I}_3$. Then, for $\boldsymbol{\theta}$-space, we use MH chains with random walk Gaussian proposal density (starting in $[0, 0]$ and diagonal covariance $0.05\mathbf{I}_2$) and with length $T = 200$. We set $J = 10$, i.e., we generate $J = 10$ possible matrices and we have $J = 10$ parallel chains considering different target densities (i.e., different conditional posterior pdfs, each one considering a different covariance matrix). Finally, for the unique/single MH chain addressing the conditional posterior keeping fixed $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(T)}$ (obtained by ATAIS), we consider again a random walk Gaussian proposal density (with initial $[0, 0]^\top$ and diagonal covariance $0.05\mathbf{I}_2$) and with length $T = 2000$.

Note that the number of evaluations of the non-linear model $\mathbf{f}$ does not differ in the different methods: in ATAIS is only $NT = 2000$, in ILIS, is $JT = 2000$, whereas in the single MH we have $T = 2000$ evaluations of $\mathbf{f}$. Therefore, all techniques have the same evaluations of the non-linear model $\mathbf{f}(\boldsymbol{\theta})$ that corresponds to the main cost in the likelihood evaluation.

Our goal here is to approximate the MAP estimation of $\boldsymbol{\theta}$ by all these schemes. The results are given in Figure 5. We can see the final ATAIS approximates of MAP of $\boldsymbol{\theta}$ (using $\boldsymbol{\theta}_{\text{true}}$ as the groundtruth), represented with green squares, whereas the estimations of ILIS are presented with red circles. The results provided by the single MH chain are depicted with blue diamonds in Figure 5. Looking at Figure 5, it is clear that ATAIS (green squares) provides the best performance estimating $\boldsymbol{\theta}_{\text{true}}$ (black cross) better than ILIS (red circles). In addition, we can see that the single MH chain using the covariance matrix estimated by ATAIS (blue diamonds) shows better results in most of the cases than ILIS, but they are still worse than ATAIS.

Note that, in this example, the posterior is very narrow, which can make it hard for the generation of samples in regions with high posterior evaluation, this is why the adaptation of the proposal scale matrix in step 2d of Table 8 is quite important. To improve even more the exploration, at some iteration the scale matrix of the proposal densities can be periodically increased to allow exploration of areas

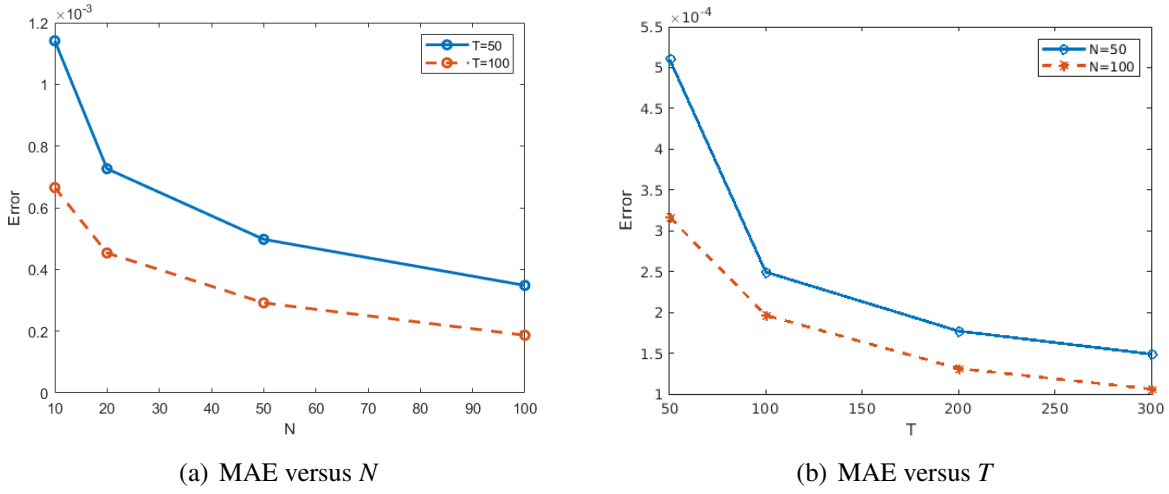away from the narrow mode as described in Section 4.2.

(a) MAE versus N

(b) MAE versus T

Figure 2: Location Example. MAE in the estimation of $\boldsymbol{\theta}_{\texttt{MAP}}$ (using $\boldsymbol{\theta}_{\texttt{true}}$ as the groundtruth), with different number of particles by ATAIS versus (a) N with fixed T and with (b) T with fixed N.

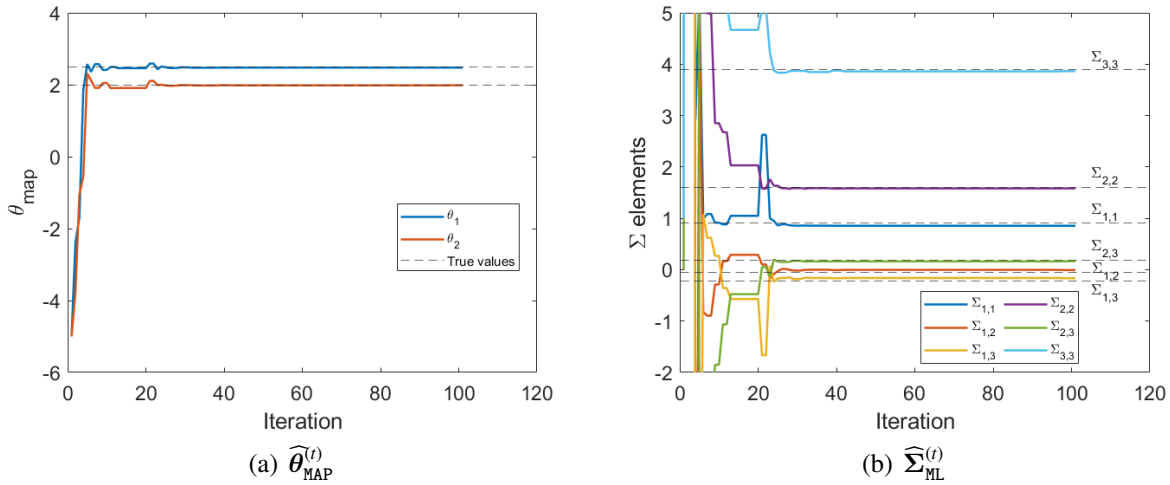(a) $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}$

(b) $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t)}$

Figure 3: Location example. Convergence of the components in $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}$ and $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t)}$ in one run of ATAIS algorithm.

Credible interval with 95% **of probability for the matrix**. In order to show how to perform a complete Bayesian inference over the covariance matrix $\boldsymbol{\Sigma}$ as well, we consider a Wishart proposal with $\nu = 100$ (degrees of freedom) with a reference matrix ($\boldsymbol{\Phi}$) equal to $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)}$ (i.e., we apply the second part of ATAIS). With this proposal distribution, we generate $J = 1000$ matrices and assign a weight to each of them following Eq. (54). Applying resampling (exactly J times) according to the normalized weights $\{\bar{\lambda}_j\}_{j=1}^{J}$, we calculate the percentiles 0.025 and 0.975 for each component to get a credible

interval for the covariance matrix Σ, as shown below (where we have averaged over 100 independent runs) in Eq (67). The first part of ATAIS was performed with $T = 50$ and $N = 50$, obtaining a confidence interval with $\alpha = 0.05$ for each value of the matrix,

$$\begin{pmatrix} [0.6697, 1.3594] & [-0.3206, 0.2958] & [-0.6946, 0.3115] \\ [-0.3206, 0.2958] & [1.2584, 2.4827] & [-0.4709, 0.8096] \\ [-0.6946, 0.3115] & [-0.4709, 0.8096] & [2.9975, 5.8437] \end{pmatrix} \tag{67}$$

Note that the covariance matrix Σ represents the covariance among the different sensors in the network. In Figure 4, we can see the histograms obtained by the resampled particles of each component of Σ (after performing resampling according to the weights $\{\bar{\lambda}_j\}_{j=1}^J$). We must remark on how the histograms corresponding to the null components of Σ_{true} have the mode very close to the value 0.
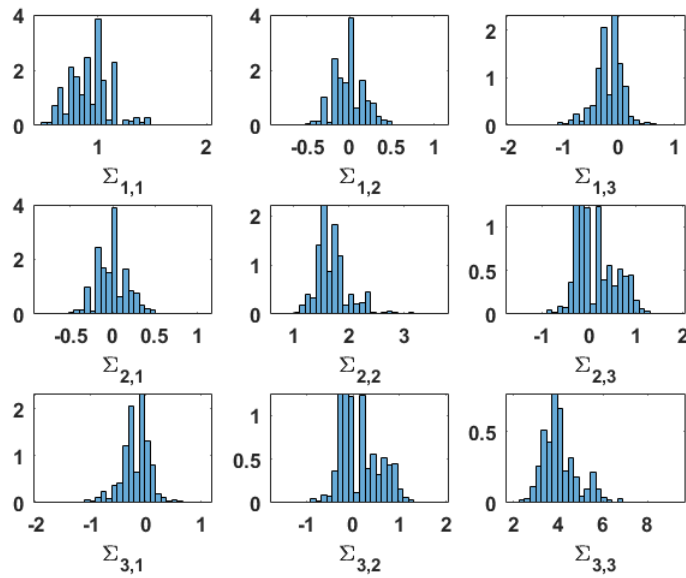

Figure 4: Location example: histogram of the components of Σ, denoted as $[\Sigma]_{i,j} = \Sigma_{i,j}$, of the covariance matrices after resampling according to the weights $\bar{\lambda}_j$, for the location example.
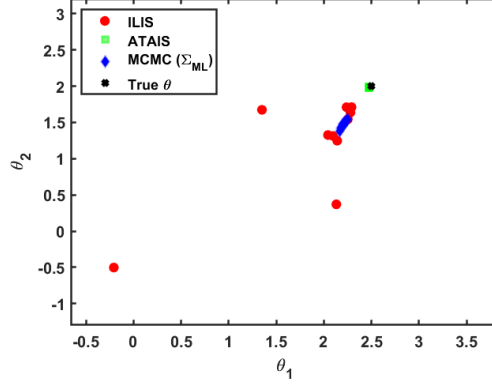
Figure 5: Location example: Estimations of θ_{MAP} with different schemes. the green squares represent the estimation with ATAIS; the red circles depict for the estimations of ILIS with a covariance matrix generated from a Wishart distribution. The blue diamonds represent the estimations of the single MH addressing a conditional posterior as invariant density, where Σ is fixed to the value of estimation Σ_{ML} obtained by ATAIS. The black cross shows θ_{true}.

Table 4: Localization example: MAE averaged over 100 simulations of ATAIS for estimating the $\theta_{true}, \Sigma_{ML}$. Recall that its groundtruth is computed as in Eq. (64). The complete MAE in the whole space is also given.

Varying in N ($T = 50$)				Varying in T ($N = 100$)			
N	θ_{true}	Σ_{ML}	Complete MAE	T	θ_{true}	Σ_{ML}	Complete MAE
5	0.0377	0.8934	0.7378	5	0.1740	2.4068	2.0009
12	0.0207	0.0443	0.0400	10	0.0758	0.4292	0.4834
25	0.0205	0.0443	0.0399	20	0.0328	0.3002	0.3649
50	0.0205	0.0442	0.0399	30	0.0205	0.0441	0.0398
100	0.0204	0.0440	0.0400	50	0.0204	0.0440	0.0400

7.2   Multi-output model - dimension of the space $D = 12$

In this second experiment, we consider the following isotopic multi-output model, using the notation in Eq. (8),

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}, \tau) + \mathbf{v}, \tag{68}$$

where the vector function $\mathbf{f}(\boldsymbol{\theta}, \tau) : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^4$ with $\boldsymbol{\theta} = [\theta_1, \theta_2]^\top$ is given by the components

$$
\begin{aligned}
f_1(\boldsymbol{\theta}, \tau) &= f_1(\theta_1, \tau) = \theta_1 \sin(\tau)\tau, \\
f_2(\boldsymbol{\theta}, \tau) &= f_2(\theta_2, \tau) = \theta_2 \cos(\tau)\tau^2, \\
f_3(\boldsymbol{\theta}, \tau) &= (\theta_1 + \theta_2) \sin(\tau) \cos(\tau), \\
f_4(\boldsymbol{\theta}, \tau) &= f_4(\theta_2, \tau) = \theta_2 \tau^2.
\end{aligned}
\tag{69}
$$

The error term $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\text{true}})$ with

$$
\Sigma_{\text{true}} = \begin{pmatrix} 0.1 & 0.3 & 0.16 & 0 \\ 0.3 & 1.05 & 0 & 0 \\ 0.16 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2.95 \end{pmatrix}.
$$

The true value of the vector $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta}_{\text{true}} = [0.2, 0.1]^\top$. In this case de dimension of the observations remains at $K = 4$, with $\boldsymbol{\theta}$ of dimension $M = 2$, which makes a total dimension of inference space of

$$
D = M + \frac{K(K+1)}{2} = 2 + \frac{4(4+1)}{2} = 12.
$$

Remark. This example shows clearly the difficulty of performing an estimation of Σ directly from the vectors observed, $\{\mathbf{y}_r\}_{r=1}^R$. This is due to the vectors not sharing the same mean (i.e., the functions f_i which also depends on τ).

The prior for $\boldsymbol{\theta}$ is assumed to be uniform and improper, i.e., $g(\boldsymbol{\theta}) \propto 1$. As in the previous example, we also test three algorithms: (a) complete ATAIS, (b) ILIS using Metropolis-Hastings (MH) chains and (c) a single MH chain working exclusively in the $\boldsymbol{\theta}$-space, maintaining fixed the covariance matrix and equal to $\widehat{\Sigma}_{\text{ML}}^{(T)}$ given by ATAIS, as in the previous experiment. We aim to approximate the MAP estimate of $\boldsymbol{\theta}$. All the results are averaged over 1000 independent runs.

ATAIS results. We employ in ATAIS the same specifications as in the previous example, i.e., $N = 50, T = 40$. We use a Gaussian proposal with the initial mean at $[0, 0]$ and the initial covariance matrix is $6\mathbf{I_2}$. The initial covariance matrix for the observations, $\widehat{\Sigma}_{\text{ML}}^{(0)}$, is the identity matrix, $\mathbf{I}_4$. Figure 6 depicts the evolution of the estimators $\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}$ and $\widehat{\Sigma}_{\text{ML}}^{(t)}$ versus iteration t, in one run of ATAIS algorithm. In Figures 7(a) and 7(b), we also provide the MAE versus the number of particles N and T, obtained by ATAIS in estimating $\boldsymbol{\theta}_{\text{MAP}}$, using $\boldsymbol{\theta}_{\text{true}}$ as groundtruth.

Table 5 provides other MAE values also in estimating the covariance matrix Σ, obtained by ATAIS. As expected, the best results are obtained increasing the number of particles and the iterations, allowing a better exploration of the parameter space. Even if the error of the estimation of Σ_{ML} is quite small. This numerical example shows the strength of ATAIS, since in this multi-output problem the covariance matrix Σ cannot be approximately in advance directly from the data. Additionally, considering the MAE in the whole state space, we also perform different simulations where the product NT remains fixed, while the value of T changes in $\{100, 200, 500, 1000, 2000\}$, so that N changes among these values $\{200, 100, 40, 20, 10\}$. The results obtained are given in Figure 8(a). We can observe the robustness of ATAIS: we obtain very small MAE values in all cases.
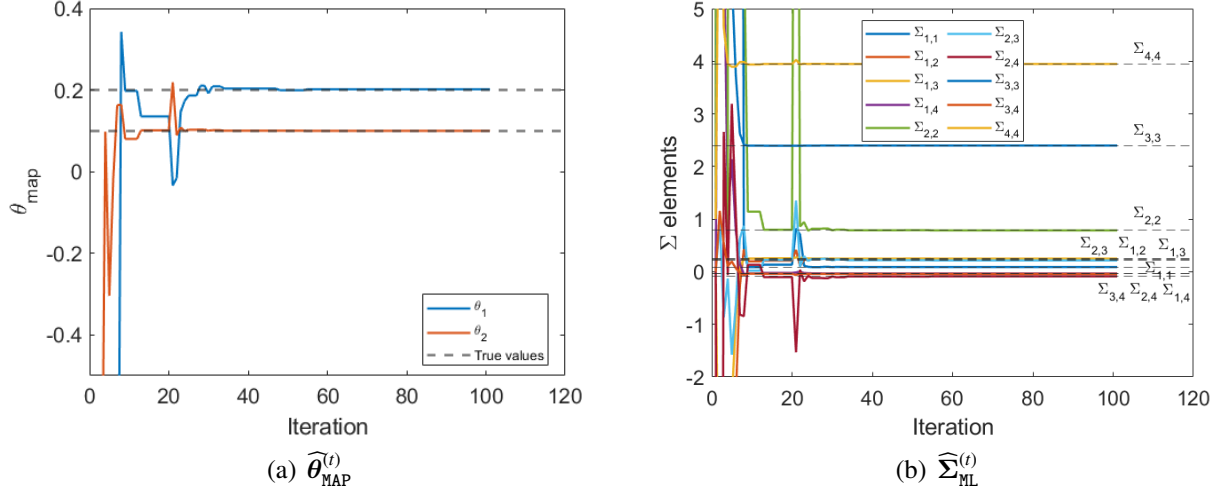
(a) $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(t)}$

(b) $\widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}}^{(t)}$

Figure 6: Multi-output example. Convergence of the components in the estimators $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}^{(t)}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}}^{(t)}$ in one run of ATAIS algorithm.

Comparison with ILIS and an "helped"-MH. We compare again the ATAIS results with (a) ILIS using Metropolis Hastings (MH) chains and (b) a unique MH chain working only in the $\boldsymbol{\theta}$-space, keeping fixed the covariance matrix to the maximum likelihood estimation $\widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}}^{(T)}$ obtained by ATAIS in its first part. Namely, this MH is helped by using and keeping $\widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}}^{(T)}$ given by ATAIS, and works only n the $\boldsymbol{\theta}$-space. For ILIS, we consider as prior and proposal density for the covariance matrix a Wishart distribution with $\nu = 5$ degrees of freedom and a reference matrix $\Phi = \frac{1}{\nu}\mathbf{I}_4$. For working in the $\boldsymbol{\theta}$-space we use MH chains with a Gaussian random walk proposal density with initial mean [0,0] and diagonal covariance matrix $0.005\mathbf{I}_2$. The length of the chains is $T = 200$. We generate $J = 10$ possible matrices, thus we have $J = 10$ parallel chains considering different target densities (each target considers a different covariance matrix). We consider that all the chains provide a unique estimation of the MAP of $\boldsymbol{\theta}$. Finally, for the single MH chain addressing the conditional posterior fixing $\widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}}^{(T)}$ (obtained by ATAIS), we consider a Gaussian random walk proposal density. This proposal density has an initial mean [0, 0] and has a diagonal covariance matrix $0.005\mathbf{I}_2$. The length of the chain is $T = 2000$.

In Figure 8(b), we can see the final ATAIS estimates of the MAP of $\boldsymbol{\theta}$ represented by green squares, while red circles represent the estimations of ILIS. The results provided by a single MH chain are displayed using blue diamonds. Looking at Figure 8(b), it is clear that the ATAIS gives the best estimations of $\boldsymbol{\theta}_{\mathrm{true}}$ (black cross) than the estimations of ILIS. Once again, we see how using the covariance matrix estimated by ATAIS (blue diamonds) gives better results than ILIS. In addition, we can see that the single MH chain using the covariance matrix estimated by ATAIS (blue diamonds) shows better results than ILIS in most of the cases, and some of them are comparable to ATAIS. Recall that, this third method has the advantage of using exactly $\widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}}^{(T)}$ that is the estimation provided by ATAIS. Then, the success of this third method is mainly due to an ATAIS ability.

Credible interval with 95% **of probability for the matrix**. In order to perform a complete Bayesian

inference over the covariance matrix Σ, we apply the second part of ATAIS. We consider a Wishart density as proposal (and prior) with $\nu = 100$ (degrees of freedom) with a reference matrix (Φ) equal to $\widehat{\Sigma}_{ML}^{(T)}$ (i.e., we apply the second part of ATAIS). With this proposal distribution, we generate $J = 1000$ matrices and assign a weight to each of the following Eq. (54). Applying resampling according to the normalized weights $\{\bar{\lambda}_j\}_{j=1}^{J}$, we calculate the percentiles 0.025 and 0.975 for each component to get a credible interval for the covariance matrix Σ, as shown below in Eq (70) (where we have averaged over 100 independent runs). The first part of ATAIS was performed with $T = 50$ and $N = 50$.

$$
\begin{pmatrix}
[0.0614, 0.1586] & [0.1669, 0.4483] & [0.0703, 0.4736] & [-0.2578, 0.1985] \\
[0.1669, 0.4483] & [0.6125, 1.5325] & [-0.4739, 0.7620] & [-0.7571, 0.7682] \\
[0.0703, 0.4736] & [-0.4739, 0.7620] & [1.5868, 3.9349] & [-1.3594, 1.0586] \\
[-0.2578, 0.1985] & [-0.7571, 0.7682] & [-1.3594, 1.0586] & [2.4684, 9.1016]
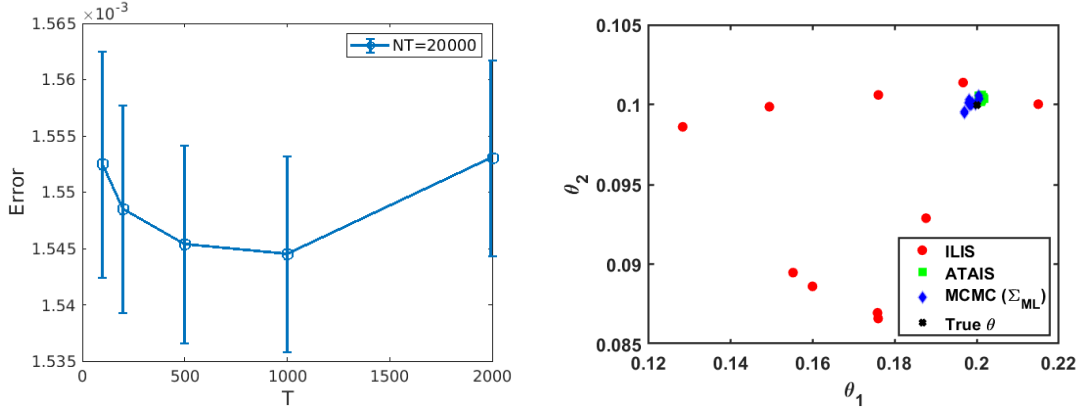\end{pmatrix}
\tag{70}
$$
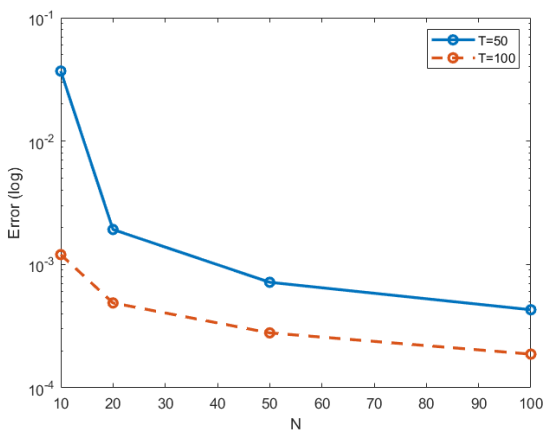
(a) MAE versus N.
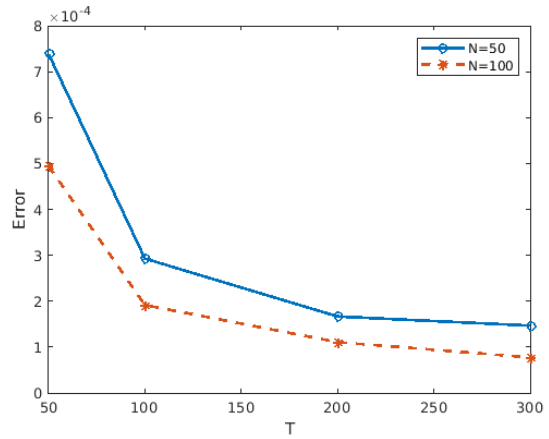
(b) MAE versus T.

Figure 7: Multi-output example. MAE in the estimation of the MAP of the posterior (considering θ_{true} as the groundtruth) by ATAIS; **(a)** versus the number of particles N with fixed T and **(b)** versus iterations T with fixed N.

(a) Keeping constant the product NT. As T grows, N decreases, e.g., in $T = 500$, we have $N = 40$, and in $T = 2000$ we have $N = 10$.

(b) $\widehat{\theta}_{\text{MAP}}$ with different schemes.

Figure 8: Multioutput example. **(a)** Complete MAE in estimating the θ_{MAP} and the Σ_{ML}, keeping fixed the number of posterior evaluations $NT = 20000$, with $T \in \{100, 200, 500, 1000, 2000\}$, hence $N \in \{200, 100, 40, 20, 10\}$. **(b)** Estimations of θ_{MAP}. The green squares represent the estimation with ATAIS; the red circles depict the estimations of ILIS with a covariance matrix generated from a Wishart distribution. The blue diamonds represent the estimations of the single MH addressing a conditional posterior as invariant density, where Σ is fixed to the value of estimation Σ_{ML} obtained by ATAIS. The black cross depicts θ_{true}.

Table 5: Multi-output example: MAE averaged over 1000 simulations of ATAIS for estimating the θ_{true}, Σ_{ML} (its groundtruth is computed as in Eq. (64)), and the complete MAE in the whole space.

	Varying in N (with fixed $T = 50$)				**Varying in T (with fixed $N = 100$)**		
N	θ_{true}	Σ_{ML}	**Complete MAE**	T	θ_{true}	Σ_{ML}	**Complete MAE**
5	0.2325	0.7666	0.7073	5	0.2100	0.4648	0.4365
12	0.0102	0.0136	0.0132	10	0.1219	0.1293	0.1285
25	0.0013	0.0026	0.0024	20	0.0355	0.2663	0.2407
50	0.0012	0.0023	0.0021	30	0.0015	0.0031	0.0029
100	0.009	0.0017	0.0025	50	0.0010	0.0021	0.0019

7.3 Use of mini-batches

In this experiment, we test the possible use of mini-batches in ATAIS. More specifically, we test the second strategy in Section 4.6. We consider the multi-output model in Section 7.2. We consider different number of data in the batches, $L \in \{2, 5, 10, 25\}$. We have $R = 50$ observations. Following Section 4.6, the data are randomly divided into $\frac{R}{L}$ disjoint subsets, with L elements in each batch. Namely, we have exactly $\frac{R}{L}$ batches. The number of iterations in ATAIS, T, is also equal to $\frac{R}{L}$, which forces to take fewer ATAIS steps as we increase the size of the subsets. Hence, we have $T = \frac{R}{L}$ that represents both the number of iterations and the number of batches. Since in this example the total number of observations is $R = 50$, we have $T \in \{25, 10, 5, 2\}$ (according to the values of L given above). The number of particles in each iteration of ATAIS is set to $N \in \{5000, 10^4\}$.

Figure 9 shows the MAE in estimating the full parameter space $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$ (considering $\boldsymbol{\Sigma}_{\text{ML}}$ in Eq. (64) as groundtruth), averaged over 10^5 independent runs. We can see that the MAE is always very small in all cases. Hence, ATAIS works well even using mini-batches. Moreover, since the use of mini-batches has a tempering effect that fosters the exploration of the space, ATAIS works even better with batches when the number of data in each batch, L, is not too small (in Figure 9, the choice $L = 2$ degrades slightly the performance).
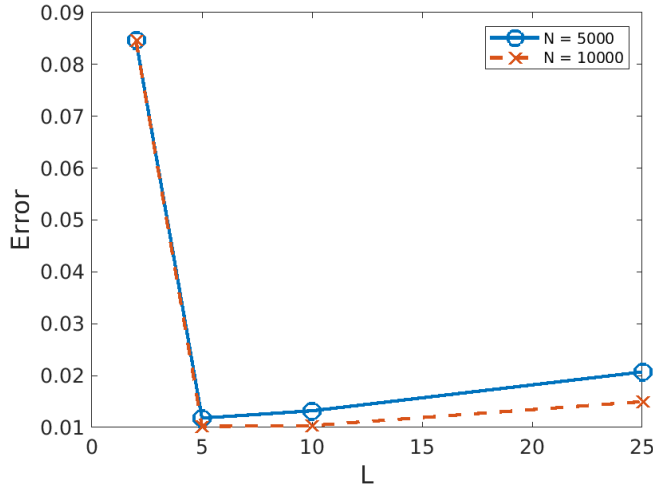

Figure 9: Testing the use of mini-batches in the multi-output model in Section 7.2: MAE in estimating the full parameter space $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$ with $R = 50$ (number of data), $L \in \{2, 5, 10, 25\}$ (number of data in the batches), $T = \frac{R}{L} \in \{50, 10, 5, 2\}$ (number of iterations which coincides with the number of batches). Note that the curve of $N = 5000$ is depicted with a solid line, and the results with $N = 10^4$ are shown with a dashed line.

7.4 Testing ATAIS in models with a Student's-t noise

As described in Section 3.2, ATAIS can be also applied when the noise perturbation in the observation model follows an elliptical distribution. In this experiment, we consider a noise distribution with heavier tails with respect to the Gaussian case. More specifically, we consider a noise variable $\mathbf{v}_r$ in

Eqs. (1)-(25) is distributed as a multivariate Student's-t distribution with 10 degrees of freedom. In this case, the estimation of Σ is obtained by the iterative process described in Section 3.2. We test the case of the Student's-t noise in both models described in the experiments above: the localization model in Section 7.1 and the multi-output model in Section 7.2.

In Table 6, we provide the averaged MAE over 10^3 runs in the estimations of ATAIS for $\boldsymbol{\theta}_{\text{true}}$, and Σ_{ML}, computed as in Eq. (64). We set the number of iterations to $T = 50$ and the number of particles N is taken as $\{5, 12, 25, 50, 100\}$. Table 6 shows the results for the localization model in Section 7.1 and the multi-output model in Section 7.2 both with the Student's-t noise. Again we can observe that ATAIS provides good results. The MAE decreases as the number of particles N increases, as expected. It is worth noting that in this scenario with a noise with a heavier tail the estimation of the matrix Σ_{ML} seems to be more difficult than with the Gaussian noise scenario.

Table 6: Experiment with Student's-t noise: MAE averaged over 1000 simulations of ATAIS for estimating the $\boldsymbol{\theta}_{\text{true}}$ and Σ_{ML} (its groundtruth is computed as in Eq. (64)). The number of iterations is fixed to $T = 50$.

	Localization model with Student's-t noise			Multi-output model with Student's-t noise		
N	$\boldsymbol{\theta}_{\text{true}}$	Σ_{ML}	Complete MAE	$\boldsymbol{\theta}_{\text{true}}$	Σ_{ML}	Complete MAE
5	0.1732	12.3840	10.1639	0.1481	0.3517	0.3291
12	0.0752	2.4801	2.0428	0.0747	0.1157	0.1111
25	0.0470	0.6286	0.5229	0.0484	0.0579	0.0568
50	0.0372	0.1094	0.0963	0.0337	0.0377	0.0366
100	0.0327	0.0753	0.0676	0.0246	0.0284	0.0279

7.5 Application to a biology system - dimension of the space $D = 7$

In this third example of application, we focus on an inference problem in a biology system [63]. We aim to make inference on the covariance matrix of the observations of a model. This model represents a physiological system with two states and one input variable. The model is governed by a set of four parameters denoted as $\boldsymbol{\theta} = [k_{12}, k_{21}, k_{1e}, b]^\top$. The dynamics of the system is ruled by the following differential equations. using the notation in Eq. (8):

$$
\begin{aligned}
\frac{df_1}{d\tau} &= -(k_{1e} + k_{12}) \cdot f_1(\boldsymbol{\theta}, \tau) + k_{21} \cdot f_2(\boldsymbol{\theta}, \tau) + b \cdot u(\tau), \\
\frac{df_2}{d\tau} &= k_{12} \cdot f_1(\boldsymbol{\theta}, \tau) - k_{21} \cdot f_2(\boldsymbol{\theta}, \tau),
\end{aligned}
\tag{71}
$$

where

$$
u(\tau) = \begin{cases} \tau + 0.5 & \text{if } 0 \leq \tau \leq 1 \\ 1.5e^{1-\tau} & \text{if } \tau > 1, \end{cases}
\tag{72}
$$

is an input of the system. We set the true parameters to $\boldsymbol{\theta}_{\text{true}} = [1, 1, 1, 2]^\top$ and

$$
\Sigma_{\text{true}} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 2 \end{pmatrix}.
$$

The prior for the components of $\boldsymbol{\theta}$ was set as uniform in the interval $[0, 5]$ (for all the components) as suggested in [63]. The system of differential equations above has no analytically close solution. Hence, the generation of the data according to the system is obtained using a Runge-Kutta Matlab solver [64]. We assume that the solution is perturbed by Gaussian noise according to the distribution $\mathcal{N}(0, \Sigma_{\text{true}})$. Then, 100 data points are sampled at equidistant times from the solution of the system (71).

Remark. Note that, in this experiment, the function $\mathbf{f}$ is evaluated only approximately since it is the solution of the system (71) which cannot be evaluated exactly. Thus, even considering the vector of true values $\boldsymbol{\theta}_{\text{true}}$, we need to approximate $\mathbf{f}(\boldsymbol{\theta}_{\text{true}}, \tau)$ by a differential equation solver (discretizing it) obtaining $\widehat{\mathbf{f}}(\boldsymbol{\theta}_{\text{true}}, \tau)$.

In order to make inference we use ATAIS with $N = 300$ particles during $T = 100$ iterations. The initial value for the mean of the proposal is set at $[0, 0, 0, 0]$ and for the covariance matrix is $6\mathbf{I}_4$. The initial estimate for the covariance matrix of the data, $\widehat{\Sigma}_{\text{ML}}^{(0)}$, was the identity, $\mathbf{I}_2$. The convergence in one run of ATAIS estimators can be observed in Figure 7.5.
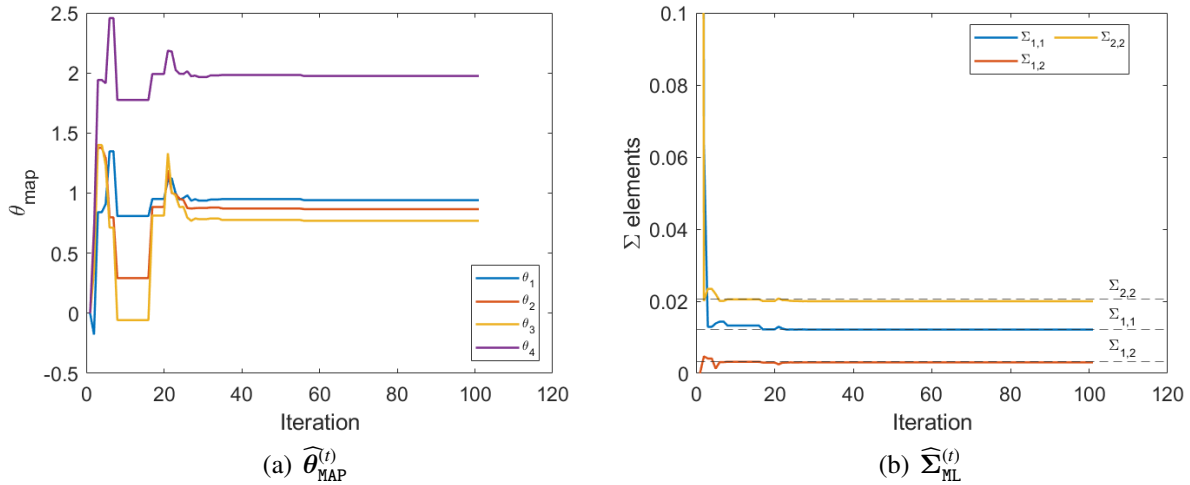

(a) $\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}$

(b) $\widehat{\Sigma}_{\text{ML}}^{(t)}$

Figure 10: Application to a system biology example. Evolution of the components in the estimators $\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}$ and $\widehat{\Sigma}_{\text{ML}}^{(t)}$ versus t, in one run of ATAIS algorithm.

In the second part of ATAIS, we generate $J = 1000$ matrices from a Wishart distribution (degrees of freedom, $\nu = 100$) with reference matrix equal to $\widehat{\Sigma}_{\text{ML}}^{(T)}$, the final estimate of the first part of ATAIS. To each of the generated matrices we assign a weight following Eq (54). After performing resampling. Applying resampling according to the weights $\{\bar{\lambda}_j\}$ we calculate the percentiles 0.025 and 0.975 to get a 95% credible interval (averaged over 50 independent runs) in Eq (73). It can be seen that our interval contains the true components of the covariance matrix.

$$\begin{pmatrix} [0.7820, 1.3876] & [0.7099, 1.4614] \\ [0.7099, 1.4614] & [1.6518, 2.8760] \end{pmatrix} \tag{73}$$

35

As a final remark, in this example, the posterior has a particularly flat shape in some regions, so that many values of the parameters $\boldsymbol{\theta}$ provide acceptable results for the solution of the system (71), even if the distance to the $\boldsymbol{\theta}_{\mathrm{true}}$ could be large.

7.6 Application to graph topology estimation - dimension of the space $D = 59$

In this section, the goal is to apply ATAIS to estimate a graph topology with 10 nodes. The graph is represented by a 10×10 graph shift operator that is usually considered a sparse matrix [65, 66]. In this experiment, we consider that the precision matrix $\mathbf{P} = \Sigma^{-1}$ is the graph shift operator. Thus, we assume $\mathbf{P}$ to be a sparse matrix and hence represent the graph [65, 66]. If the nodes i and j are connected, then it follows that $\mathbf{P}_{i,j} \neq 0$. Estimating the graph topology consists mainly of knowing the connections between the nodes, which is determined by the matrix $\mathbf{P} = \Sigma^{-1}$. The goal is to apply ATAIS to estimate the graph topology. The corresponding adjacent matrix (binary matrix showing the connection, that is the groundtruth in this experiment) is depicted in Figure 11(a). Furthermore, we assume the following multi-output nonlinear model,

$$
\begin{aligned}
f_1(\boldsymbol{\theta}) &= -\theta_4\tau + 5\theta_1^2 \\
f_2(\boldsymbol{\theta}) &= 2\theta_3 \sin -\theta_2\tau \\
f_3(\boldsymbol{\theta}) &= \theta_1 - \theta_3 + \theta_1 \cos(2\tau) \\
f_4(\boldsymbol{\theta}) &= 3\theta_4 + 3\theta_2 + \theta_1 \exp(0.1\tau) \\
f_5(\boldsymbol{\theta}) &= \theta_3^2 - 2\theta_1 + 3\theta_2 - \exp(0.8\theta_3)\exp(1-\tau) \\
f_6(\boldsymbol{\theta}) &= 5(\theta_4 + \theta_3) - \theta_2 \log(1 + 2\tau) \\
f_7(\boldsymbol{\theta}) &= 3\theta_2 - 0.2\tau \sin(\theta_3) \\
f_8(\boldsymbol{\theta}) &= 3\theta_1 + 5\theta_3 - 20\sin(\theta_4)\cos(2\tau + \frac{\pi}{4}) \\
f_9(\boldsymbol{\theta}) &= \theta_2 + 4\theta_4 + 5\exp\left(\frac{1}{1+\theta_3}\right)\tau \\
f_{10}(\boldsymbol{\theta}) &= 5\theta_1 + 10\theta_3 - 5\theta_4 \sin(\tau),
\end{aligned}
\tag{74}
$$

where $f_{r,i}(\boldsymbol{\theta})$ is i-th signal contaminated by Gaussian noise and is emitted by at the i-th node. The model in Eq. (74) is characterized by a vector parameter $\boldsymbol{\theta} \in \mathbb{R}^4$ with the following true values $\boldsymbol{\theta}_{\mathrm{true}} = [0.5, 2, 5, 3]^\top$. We assume an improper uniform prior over $\boldsymbol{\theta}$. Note that, in this scenario, we have $M = 4$ and $K = 10$, hence the complete space has dimension,

$$
D = M + \frac{K(K+1)}{2} = 59.
$$

We estimate Σ with ATAIS, then compute $\mathbf{P} = \Sigma^{-1}$ and apply a threshold value of 0.3 for obtaining a corresponding binary matrix (entries of $\mathbf{P}$ smaller than 0.3 are set to 0, whereas entries of $\mathbf{P}$ greater than 0.3 are set to 1). More sophisticated strategies could be considered, for instance, using ATAIS jointly with a graphical lasso approach. We set $T = 10$ and $N = 20000$ and consider 500 independent runs. For the adaptation of the covariance of the proposal density in Eqs. (26)-(78), we apply a cyclic/periodic strategy in Eq. (27)

$$
\delta_{t+1} = \begin{cases} a\,\delta_t, & \text{if } \delta_t \geq \delta_{\min}, \\ \delta_0, & \text{if } \delta_t < \delta_{\min}, \end{cases}
\tag{75}
$$

with $\delta_0 = 1$, $a = 0.1$ and $\delta_{\min} = 0.05$.

The results in estimation of $\mathbf{P} = \boldsymbol{\Sigma}^{-1}$ are given in percentages in Figure 11. In 69.7% of the runs, we obtain exactly the groundtruth matrix in **(a)**. In the rest of the runs, we get very similar adjacency matrices, shown in **(b)**-**(c)**.

Figure 12 depicts the evaluations of the final conditional log-posterior of the samples $\boldsymbol{\theta}_t^{(n)}$ such that $\bar{w}_t^{(n)} \geq 1/N$ at each iteration. Hence, the number of these relevant samples varies with iteration t and with the specific run, depending also on the adaptation of the proposal and its covariance matrix in Eqs. (26)-(78). For instance, in Figure 12(a), we can observe the evaluation in 3 different runs where the number of extremely relevant samples - with weights such that $\bar{w}_t^{(n)} \geq \frac{1}{N}$ - is 70, 73 and 111, respectively. These numbers (of the extremely relevant samples) can be considered small, but actually they are very well-located as shown in Figure 12(b), which illustrates the histograms of the four components of these relevant samples. Note that the histograms are localized around the true values of $\boldsymbol{\theta}_{\text{true}} = [0.5, 2, 5, 3]^\top$. Indeed, the values of the log-posterior evaluations are close to the log-posterior evaluation at $\boldsymbol{\theta}_{\text{true}}$. Computing an averaged (standard) ESS measure $\widehat{ESS} = \frac{1}{T} \sum_{t=1}^{T} \text{ESS}(t)$ (averaged over the iterations, and also over the different runs) where $\text{ESS}(t) = 1/\sum_{n=1}^{N} \bar{w}_t^{(n)}$ [46], we obtain $\frac{\widehat{ESS}}{N} \approx 0.10$, i.e., 10% of the N particles. However, we can see they are well located, in fact, the MAE in estimation of $\boldsymbol{\theta}$ is only 0.01. Therefore, ATAIS provides very good performance, remarking that the dimension of the space is $D = 59$.

(a) True matrix, obtained in estimation 69.7% of times.

(b) Obtained in estimation 22.1% of times.
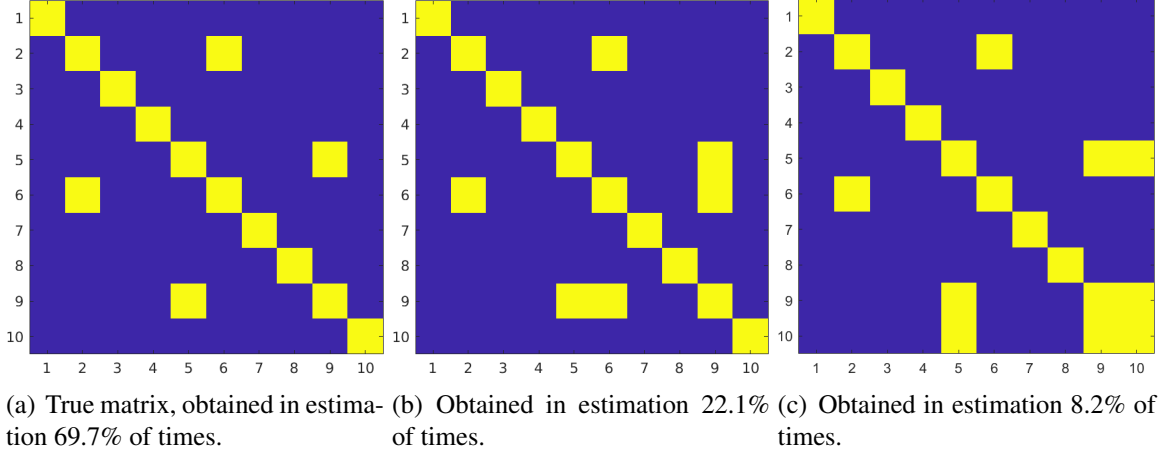
(c) Obtained in estimation 8.2% of times.

Figure 11: Graph topology estimation with ATAIS algorithm (with $T = 10$ and $N = 2 \cdot 10^4$). In the 69.7% of the runs, we obtain exactly the groundtruth matrix in **(a)**. In the rest of other runs, we get very similar adjacency matrices, shown in **(b)**-**(c)**.
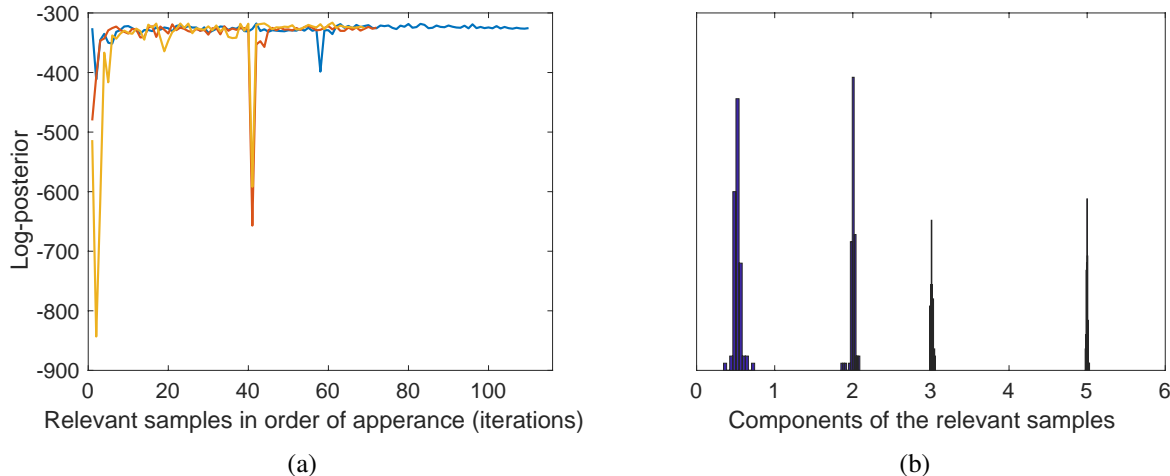
Figure 12: **(a)** Evaluations of the final conditional log-posterior of the samples $\boldsymbol{\theta}_t^{(n)}$ such that $\bar{w}_t^{(n)} \geq 1/N$ at each iteration. **(b)** Histograms of the four components of the relevant samples. They are localized around the true values of $\boldsymbol{\theta}_{\text{true}} = [0.5, 2, 5, 3]^\top$.

7.7 Comparison with other benchmark schemes

First of all, it is important to remark that, in all the numerical experiments above, the application of other Monte Carlo approaches directly on the joint space $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$ often produces much higher errors in estimation, which makes it difficult to compare and visualize with respect to the results obtained by ATAIS. In this section, we make an effort to compare with other techniques. Specifically, we apply different MCMC algorithms on the complete parameter space $\boldsymbol{\theta}, \boldsymbol{\Sigma}$, in order to compare to ATAIS. Since ATAIS can be applied even when $\mathbf{f}$ is not differentiable, in other to perform a fair comparison, we compare with benchmark algorithms in the literature that do not use the gradient information. We consider the multi-output model with Gaussian noise in Section 7.2. Hence, as shown in Section 7.2, the dimension of the complete space is $D = 12$.

Firstly, we apply a Metropolis-Hastings (MH) algorithm, testing different parameters. We use different random walk Gaussian proposal densities for the $\boldsymbol{\theta}$ with covariance matrix $a\mathbf{I}_M$ with $a \in \{0.1, 1, 10\}$ and for the proposal for generating $\boldsymbol{\Sigma}$ is a random walk Wishart with $\nu \in \{5, 50\}$ degrees of freedom, with a length of the chain of $T = 2 \cdot 10^4$. Secondly, we apply an adaptive version of the previously described MH schemes where the random walk Gaussian proposal density with respect to $\boldsymbol{\theta}$ is tuned using the empirical mean and variance of the generated past samples [3, 5]. Then, the parameter a is auto-tuned. Finally, we also applied also MH-within-Gibbs sampler (that is a more adequate scheme for sampling tight posteriors in high-dimensional spaces) where, for each internal Gibbs iteration for sampling from each full-conditional, we apply 10 MH internal steps [67]. The internal MH considers random walk Gaussian proposal pdf with unit variance. Moreover, since the elements of the matrix $\boldsymbol{\Sigma}$ required to fulfill certain constraints we discard the matrices that are not positive definite. Hence, this scheme is much more costly than ATAIS and the other MCMC algorithms above. At each run, the initializations of the chains of all the MCMC schemes (as in ATAIS) described above are chosen

randomly around a vector of zeros for θ and an identity matrix for Σ (with a standard deviation of 4; for Σ, we always consider an initial diagonal matrix with the same random element in the diagonal). Table 7 shows the MAE (averaged over 10^3 independent runs) for different schemes in the multi-output example. We include also the MAE corresponding to ATAIS for facilitating the comparison. Note that, we consider always more evaluations of the model for the MCMC schemes than ATAIS, except for the last line of ATAIS. In any case, ATAIS clearly outperforms the rest of the compared techniques.

Table 7: Comparison with other techniques in the multi-output model of Section 7.2: MAE in the complete space by different methods.

Method	Parameters	Posterior evaluations	MAE
MH	$a = 0.1, \nu = 5, T = 2 \cdot 10^4$		154.71
MH	$a = 0.1, \nu = 50, T = 2 \cdot 10^4$		0.6635
MH	$a = 1, \nu = 5, T = 2 \cdot 10^4$	$T = 2 \cdot 10^4$	6.0458
MH	$a = 1, \nu = 50, T = 2 \cdot 10^4$		0.8170
MH	$a = 10, \nu = 5, T = 2 \cdot 10^4$		3.6669
MH	$a = 10, \nu = 50, T = 2 \cdot 10^4$		1.4905
adaptive MH	$\nu = 5 \ T = 2 \cdot 10^4$	$T = 2 \cdot 10^4$	2.9072
adaptive MH	$\nu = 50 \ T = 2 \cdot 10^4$		0.7481
MH-within-Gibbs	10 internal MH steps, $T = 2 \cdot 10^4$	$T = 2 \cdot 10^4$	0.1351
ATAIS	$N = 100, T = 5$	$NT = 500$	0.4365
ATAIS	$N = 100, T = 20$	$NT = 2000$	0.2407
ATAIS	$N = 100, T = 100$	$NT = 10^4$	0.0015
ATAIS	$N = 100, T = 200$	$NT = 2 \cdot 10^4$	0.0012

8 Conclusions

In this work, we have introduced an adaptive importance sampling (AIS) method for robust inference in complex Bayesian inversion problems with unknown parameters θ of the non-linear mapping and unknown covariance matrix Σ of the noise perturbation. The variables of interest are split in two blocks, the parameters θ of the non-linear model and the covariance matrix Σ, are handled in different ways. The main proposed inference scheme is divided into two main parts. The first part is devoted to approximating a conditional posterior θ given the data and the maximum likelihood estimator of Σ. This first part allows of finding regions of high probability about θ and Σ (working alternately in subsets of the complete space, with reduced dimensions).
In the second part, a Bayesian approach is also performed over Σ re-using and re-weighting the samples of θ previously generated. Then, an approximation of the complete posterior of $\{\theta, \Sigma\}$ is provided. This second part does not require of additional evaluation of the possibly costly non-linear vectorial model $\mathbf{f}$. The resulting scheme is a robust inference approach for Bayesian inversion, based on an adaptive importance sampler that addresses a sequence of different conditional posteriors and a post-process that allows a Bayesian inference also over Σ. Several variants and computational details have

been discussed. Additionally, a simpler compelling scheme, called ILIS, has been introduced in order to compare with ATAIS. Several numerical experiments have been presented. We can observe the good performance obtained by ATAIS, providing a complete Bayesian analysis of the complete space $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$, and outperforming other benchmark algorithms.

Acknowledgments

References

[1] W. J. Fitzgerald. Markov chain Monte Carlo methods with applications to signal processing. *Signal Processing*, 81(1):3–18, January 2001.

[2] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

[3] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[4] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017.

[5] D. Luengo, L. Martino, M. Bugallo, V. Elvira, and S. Sarkka. A Survey of Monte Carlo Methods for Parameter Estimation. *EURASIP Journal on Advances in Signal Processing,*, 25:1–62, 2020.

[6] L. Martino. A review of multiple try MCMC algorithms for signal processing. *Digital Signal Processing*, 75:134 – 152, 2018.

[7] L. Martino, V. Elvira, and G. Camps-Valls. The recycling Gibbs sampler for efficient learning. *Digital Signal Processing*, 74:1–13, 2018.

[8] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: fifteen years later. *technical report*, 2008.

[9] P. M. Djurić, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Míguez. Particle filtering. *IEEE Signal Processing Magazine*, 20(5):19–38, September 2003.

[10] D. P. Liu, Q. T. Zhang, and Q. Chen. Structures and performance of noncoherent receivers for unitary space-time modulation on correlated fast-fading channels. *IEEE Transactions Vwhicular Technology*, 53(4):1116–1125, July 2004.

[11] K. Myers and B. Tapley. Adaptive sequential estimation with unknown noise statistics. *IEEE Transactions on Automatic Control*, 21(4):520–523, 1976.

[12] O. Bodnar and T. Bodnar. Bayesian estimation in multivariate inter-laboratory studies with unknown covariance matrices. *Metrologia*, 60(5):054003, 2023.

[13] M. S. Sinay and J. S. J. Hsu. Bayesian inference of a multivariate regression model. *Journal of Probability and Statistics*, 2014, 2014.

[14] A. J. Rothman, E. Levina, and J. Zhu. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962, 2010.

[15] L. Martino, F. Llorente, E. Curbelo, J. Lopez-Santiago, and J. Miguez. Automatic tempered posterior distributions for bayesian inversion problems. *Mathematics*, 9(7):1–17, 2021.

[16] J. Lopez-Santiago, L. Martino, M. A. Vazquez, and J. Miguez. A Bayesian inference and model selection algorithm with an optimization scheme to infer the model noise power. *Monthly Notices of the Royal Astronomical Society*, 507(3):3351–3361, 2021.

[17] S. K. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[18] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, 19(6):451–458, July 1992.

[19] N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(3):589–607, 2008.

[20] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[21] B. Kugler, F. Forbes, and S. Douté. Fast bayesian inversion for high dimensional inverse problems. *Statistics and Computing*, 32(2):31, 2022.

[22] L. Martino, V. Elvira, D. Luengo, and J. Corander. Layered adaptive importance sampling. *Statistics and Computing*, 27(3):599–623, 2017.

[23] F. Llorente, E. Curbelo, L. Martino, V. Elvira, and D. Delgado. MCMC-driven importance samplers. *Applied Mathematical Modelling*, 11:310–331, 2022.

[24] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, December 2012.

[25] M. F. Bugallo, L. Martino, and J. Corander. Adaptive importance sampling in signal processing. *Digital Signal Processing*, 47:36–49, 2015.

[26] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *SIAM review (SIREV)*, 65(1):3–58, 2023.

[27] F. Llorente, L. Martino, E. Curbelo, J. Lopez-Santiago, and D. Delgado. On the safe use of prior densities for Bayesian model selection. *WIREs Computational Statistics*, 15(1):e1595, 2022.

[28] H. Huang, A. Handel, and X. Song. A Bayesian approach to estimate parameters of ordinary differential equation. *Computational statistics*, 35:1481–1499, 2020.

[29] M. J. Beira and P. J. Sebastião. A differential equations model-fitting analysis of covid-19 epidemiological data to explain multi-wave dynamics. *Scientific Reports*, 11(1):16312, 2021.

[30] M. B. I. Reaz, M. S. Hussain, and F. Mohd-Yasin. Techniques of EMG signal analysis: detection, processing, classification and applications. *Biological procedures online*, 8:11–35, 2006.

[31] Y. Qu, M. A. Bhouri, and P. Gentine. Joint parameter and parameterization inference with uncertainty quantification through differentiable programming. *arXiv preprint arXiv:2403.02215*, 2024.

[32] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

[33] Y. Ke, S. Minsker, Z. Ren, Q. Sun, and W. Zhou. User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, 34(3):454–471, 2019.

[34] L. Martino, R. Casarin, F. Leisen, and D. Luengo. Adaptive independent sticky MCMC algorithms. *EURASIP J. Adv. Signal Process.*, 5:1–28, 2018.

[35] W. R. Gilks and P. Wild. Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*, 41(2):337–348, 1992.

[36] W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Applied Statistics*, 44(4):455–472, 1995.

[37] F. Llorente, L. Martino, D. Delgado-Gmez, and G. Camps-Valls. Deep importance sampling based on regression for model inversion and emulation. *Digital Signal Processing*, 116:103104, 2021.

[38] Ö. D. Akyildiz, D. Crisan, and J. Míguez. Parallel sequential Monte Carlo for stochastic gradient-free nonconvex optimization. *Statistics and computing*, 30(6):1645–1663, 2020.

[39] Ö. D. Akyildiz, I. P. Marino, and J. Míguez. Adaptive noisy importance sampling for stochastic optimization. In *2017 IEEE 7th international workshop on computational advances in multi-sensor adaptive processing (CAMSAP)*, pages 1–5. IEEE, 2017.

[40] J. Dagpunar. *Principles of random variate generation*. Clarendon Press (Oxford and New York), New York, 1988.

[41] L. Martino, D. Luengo, and J. Miguez. *Independent Random Sampling Methods*. Springer, 2018.

[42] K. T. Fang, Z. H. Yang, and S. Kotz. Generation of multivariate distributions by vertical density representation. *Statistics*, 35(3):281–293, 2001.

[43] M. Bilodeau and D. Brenner. *Theory of multivariate statistics*. Springer Science & Business Media, 1999.

[44] S. Hediger, J. Nf, and M. Wolf. R-nl: Covariance matrix estimation for elliptical distributions based on nonlinear shrinkage. *IEEE Transactions on Signal Processing*, 71:16571668, 2023.

[45] Y. El-Laham, V. Elvira, and M. F. Bugallo. Robust covariance adaptation in adaptive importance sampling. *IEEE Signal Processing Letters*, 25:1049–1053, 2018.

[46] L. Martino, V. Elvira, and M. F. Louzada. Effective Sample Size for importance sampling based on the discrepancy measures. *Signal Processing*, 131:386–401, 2017.

[47] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Generalized multiple importance sampling. *Statistical Sicence*, 34(1):129–155, 2019.

[48] Y. El-Laham, L. Martino, V. Elvira, and M. Bugallo. Efficient adaptive multiple importance sampling. *27th European Signal Processing Conference (EUSIPCO)*, pages 1–4, 2019.

[49] L. Martino and V. Elvira. Compressed Monte Carlo with application in particle filtering. *Information Sciences*, 553:331–352, 2021.

[50] L. Martino, V. Elvira, D. Luengo, and J. Corander. An adaptive population importance sampler: Learning from the uncertanity. *IEEE Transactions on Signal Processing*, 63(16):4422–4437, 2015.

[51] V. Elvira, E. Chouzenoux, O. D. Akyildiz, and L. Martino. Gradient-based adaptive importance samplers. *Journal of the Franklin Institute*, 360(13):9490–9514, 2023.

[52] L. Martino, V. Elvira, D. Luengo, A. Artes, and J. Corander. Smelly parallel MCMC chains. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1–5, 2015.

[53] D. O. Akyildiz, I. P. Mario, and J. Mguez. Adaptive noisy importance sampling for stochastic optimization. In *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5, 2017.

[54] F. Llorente, L. Martino, J. Read, and D. Delgado-Gmez. Optimality in noisy importance sampling. *Signal Processing*, 194:108455, 2022.

[55] W. Neiswanger, C. Wang, and E. Xing. Asymptotically exact, embarrassingly parallel MCMC. *preprint, arXiv:1311.4780*, 2014.

[56] C. D. Rodgers. *Inverse methods for atmospheric sounding: theory and practice*, volume 2. World scientific, 2000.

[57] X. Qian and D. Klabjan. The impact of the mini-batch size on the variance of gradients in stochastic gradient descent. *arXiv preprint arXiv:2004.13146*, 2020.

[58] I. Schuster. Gradient importance sampling. *preprint, arXiv:1507.05781*, 2015.

[59] M. Fasiolo, F.E. de Melo, and S. Maskell. Langevin incremental mixture importance sampling. *Statistic Computing*, 28:9490–9514, 2018.

[60] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2004.

[61] L. Martino and V. Elvira. Metropolis sampling. *Wiley StatsRef: Statistics Reference Online*, pages 1–18, 2017.

[62] S. Chib and I. Jeliazkov. Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association*, 96:270–281, 2001.

[63] N. J. Linden, B. Kramer, and P.i Rangamani. Bayesian parameter estimation for dynamical models in systems biology. *PLOS Computational Biology*, 18(10):e1010651, 2022.

[64] Cleve Moler. *Numerical Computing with MATLAB*. The MathWorks, Inc., Natick, MA (USA), 2004.

[65] A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro. Stationary Graph Processes and Spectral Estimation. *IEEE Transactions on Signal Processing*, 65(22):5911–5926, 2017.

[66] C. Uhler. Gaussian graphical models. In *Handbook of graphical models*, pages 217–238. CRC Press, 2018.

[67] L. Martino, R. Casarin, F. Leisen, and D. Luengo. Adaptive independent sticky MCMC algorithms. *EURASIP Journal on the advances in Signal Processing*, 5:1–28, 2018.

Table 8: ATAIS: an adaptive IS scheme with a sequence of adaptive target pdfs

---

1. **Initializations:** Choose N, $\boldsymbol{\mu}_1$, $\boldsymbol{\Lambda}_1$, $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(0)}$, and set $\pi_{\text{MAP}} = 0$. Recall $\pi_t(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(t-1)})$.

2. **For** $t = 1, \ldots, T$:

   (a) **Sampling:**

      i. Draw $\boldsymbol{\theta}_t^{(1)}, ..., \boldsymbol{\theta}_t^{(N)} \sim q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)$.

      ii. Assign to each sample the weights

$$w_t^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)}, \qquad n = 1, ..., N. \tag{76}$$

   (b) **Current maximum estimations:**

      i. Obtain $\boldsymbol{\theta}_{\max}^{(t)} = \arg\max_n \pi_t(\boldsymbol{\theta}_t^{(n)})$, and compute $\widehat{\mathbf{r}}_t = \mathbf{f}_r(\boldsymbol{\theta}_{\max}^{(t)})$.

      ii. Compute $\widehat{\boldsymbol{\Sigma}}_t = \dfrac{1}{R}\sum_{r=1}^{R}(\boldsymbol{y}_r - \widehat{\mathbf{r}}_t)(\boldsymbol{y}_r - \widehat{\mathbf{r}}_t)^\top$.

   (c) **Global maximum estimations:**

      - If $\pi_t(\boldsymbol{\theta}_{\max}^{(t)}) > \pi_{\text{MAP}}$:
         i. $\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)} = \boldsymbol{\theta}_{\max}^{(t)}$,
         ii. $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(t)} = \widehat{\boldsymbol{\Sigma}}_t$,
         iii. Update $\pi_{\text{MAP}} = \pi_{t+1}(\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)})$ (note that $\pi_{t+1}(\boldsymbol{\theta})$ takes into account the new $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(t)}$).
      - Otherwise $\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)} = \widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t-1)}$, and $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(t)} = \widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(t-1)}$.

   (d) **Adaptation:** Set

$$\boldsymbol{\mu}_{t+1} = \widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}, \tag{77}$$

$$\boldsymbol{\Lambda}_{t+1} = \sum_{n=1}^{N} \bar{w}_t^{(n)}(\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t)^\top(\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t) + \delta_t \mathbf{I}_M, \tag{78}$$

   where $\bar{w}_t^{(n)} \frac{w_t^{(n)}}{\sum_{i=1}^{N} w_t^{(i)}}$ are the normalized weights, $\bar{\boldsymbol{\theta}}_t = \sum_{n=1}^{N} \bar{w}_t^{(n)}\boldsymbol{\theta}_t^{(n)}$ and $\delta_t > 0$.

3. **Output:** Return the final estimators $\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(T)}$, $\widehat{\boldsymbol{\Sigma}}_{\text{ML}}^{(T)}$, and all the weighted samples $\{\boldsymbol{\theta}_t^{(n)}, \widetilde{w}_t^{(n)}\}$, for all t and n, with the corrected weights

$$\widetilde{w}_t^{(n)} = w_t^{(n)}\frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})}. \tag{79}$$