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

E. Curbelo*, L. Martino†, F. Llorente**, D. Delgado-Gmez*,

† Universidad Rey Juan Carlos (URJC), Madrid, Spain.

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

** Stony Brook University, New York, USA.

October 6, 2023

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 AIS (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 for obtaining 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

are 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 in 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 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].

In this work, we focus on the problem of make 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 (specially 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 be also 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 the idea of splitting the inference space in two blocks, $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ (as in a block Gibbs sampling).

Main proposed scheme - Complete ATAIS. Firstly, we extend and generalize the approach presented in [14, 15]. The proposed inference scheme is divided in two main parts. In the first part the we approximate the conditional posterior of $\boldsymbol{\theta}$ given the data and the maximum likelihood estimator $\boldsymbol{\Sigma}_{\texttt{ML}}$ 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 adaptive importance sampler that addresses a sequence of different conditional posteriors and a post-process that allows a Bayesian inference over $\boldsymbol{\Sigma}$ as well. We refer to the overall scheme (first and second part) as *complete ATAIS*.

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

The conditional posteriors addressed by ATAIS differ in the use of different covariance matrices: this procedure can resemble a tempering of the posterior distribution [16, 17, 18, 19].

Auxiliary competitive scheme - ILIS. As also remarked in different works [14, 20, 11], the application of a Monte Carlo sampling methods directly in the complete space $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$ 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 [21, 22] for this inference context. With respect to the main algorithm in [21, 22], we switch the positions of the importance sampling (IS) method and Markov Chain Monte Carlo (MCMC) techniques [23, 24, 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 is 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 final observation, we highlight that ATAIS could be also 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 be also estimated.
- The model considers a vectorial non-linear function $\mathbf{f}(\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ so that, we can perform uncertainty analysis over $\boldsymbol{\theta}$ and/or $\boldsymbol{\Sigma}$, obtaining credible intervals. Moreover, we can perform hypothesis testing or model selection approximating the marginal likelihood [3, 25, 26]. Hence, we remark that the proposed scheme is *much more* than an optimizer: it is a sampler which allows a complete Bayesian inference over $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$.
- In its second inference part, ATAIS recycles all the samples (w.r.t. $\boldsymbol{\theta}$) and the posterior evaluations from the first part. Therefore, this second part does not require any additional evaluation of the possibly complex and costly nonlinearity $\mathbf{f}$.
- We also introduce several extensions as addressing models with t-student noise (or, more generally, with other elliptical distributions) and/or the possible use of mini-batches (that is allowed by ATAIS).
- Since, in ATAIS, we consider a sequence of adaptive conditional posteriors, ATAIS can be considered as an adaptive importance sampler where *both* proposal and target pdfs are adapted.

- 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 in two blocks, $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$, where different numbers of samples is considered for each block, denoted as NT for $\boldsymbol{\theta}$ and J for $\boldsymbol{\Sigma}$.
- A discussion, with practical suggestions, regarding the tuning of hyper-parameters of the prior densities is provided.
- We also design a competitive sampling scheme, denoted as ILIS, for comparing the performance of ATAIS.

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. The second part of the main proposed inference scheme is described in Section 4. The alternative scheme, inverted layered importance sampling (ILIS), is given in Section 5. Finally, Section 6 contains several numerical experiments and Section 7 provides some final conclusions.

2 Problem Statement

Let us denote as $\boldsymbol{\theta} = [\theta_1, ..., \theta_M]^\top \in \boldsymbol{\Theta} \subseteq \mathbb{R}^M$, 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,

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

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

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 $\boldsymbol{\Sigma}$ 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

[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 6.2.

function is

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

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 $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$, obtaining at least some point estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Sigma}}$. 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 application 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 $\mathbf{x}_{k,r} \in \mathbb{R}^L$, i.e.,

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

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. Hence, 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.,

$$\mathbf{y}_r = \mathbf{f}(\boldsymbol{\theta}, \tau_{1,r}, ..., \tau_{K,r}) + \mathbf{v}_r, \tag{8}$$

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

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).

2.2 Bayesian inference in the complete space

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

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathbf{Y}) = \frac{1}{p(\mathbf{Y})} p(\boldsymbol{\theta}, \boldsymbol{\Sigma}, \mathbf{Y}) = \frac{1}{p(\mathbf{Y})} \ell(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) g_{\theta}(\boldsymbol{\theta}) g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}), \tag{10}$$

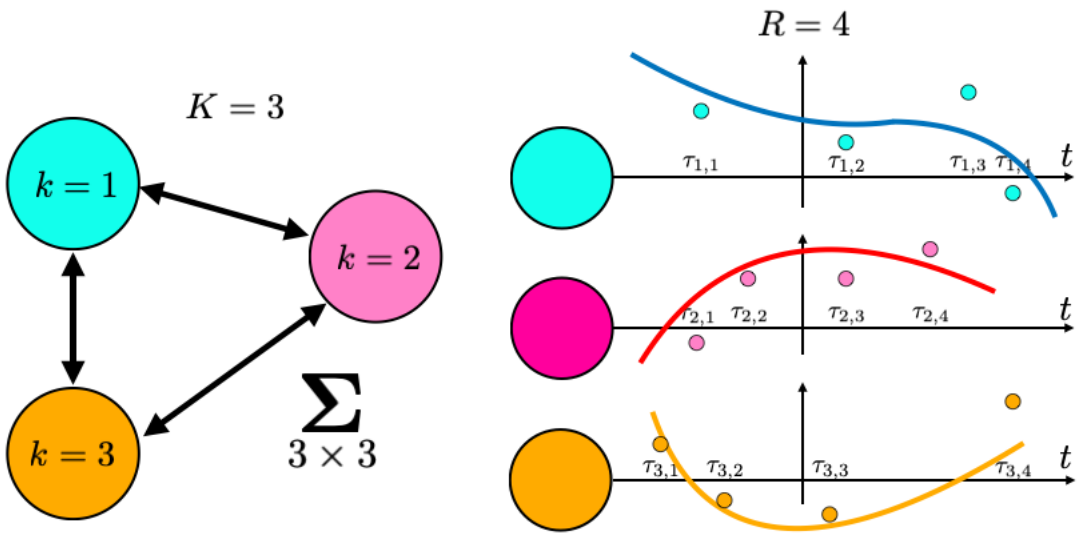


Figure 1: Graphical representation of the considered multioutput model with $K = 3$ output signals and $R = 4$ time instants for each signal. One can suppose that $\boldsymbol{\Sigma}$ represents a 3×3 covariance matrix of three possible nodes in a graph.

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

Main observation. Generally, generating random samples from a complicated posterior in Eq. (10) and computing efficiently the integrals involving $p(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathbf{Y})$ 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 have also the constrains regarding $\boldsymbol{\Sigma}$, 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 $\boldsymbol{\Sigma}$ and parameters of the nonlinearity $\boldsymbol{\theta}$. Indeed, "wrong choices" of $\boldsymbol{\Sigma}$ can easily jeopardize the sampling of $\boldsymbol{\theta}$.
Below, we describe an inference scheme formed by *two main parts*. First, we tackle the problem of drawing from conditional posterior of $\boldsymbol{\theta}$ given the data the maximum likelihood estimator of $\boldsymbol{\Sigma}$. With this goal, the maximum likelihood estimator of $\boldsymbol{\Sigma}$ must be obtained. Therefore, in this first part, we apply a Bayesian inference over $\boldsymbol{\theta}$ and a frequentist approach over $\boldsymbol{\Sigma}$. 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 (samples and other information) obtained in the first part.

3 First part of the proposed inference scheme

Main idea. The main idea underlying the complete ATAIS algorithm is to take advantage of the split of the inference space (working firstly in smaller portions of the entire space). In a 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 $\boldsymbol{\Sigma}_{\texttt{ML}}$). Whereas, a known analytic formula is employed for obtaining a sequence of optimized matrices $\boldsymbol{\Sigma}_{\texttt{ML}}$. In the second part, described in Section 4, we generate sampling random matrices from an auto-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

$$p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma}_{\texttt{ML}}) = \frac{\ell(\mathbf{Y}|\boldsymbol{\Sigma}_{\texttt{ML}}, \boldsymbol{\theta}) g_\theta(\boldsymbol{\theta})}{p(\mathbf{Y}|\boldsymbol{\Sigma}_{\texttt{ML}})} \propto \ell(\mathbf{Y}|\boldsymbol{\Sigma}_{\texttt{ML}}, \boldsymbol{\theta}) g_\theta(\boldsymbol{\theta}). \tag{11}$$

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

$$\boldsymbol{\Sigma}_{\texttt{ML}} = \arg\max_{\boldsymbol{\Sigma}} \ell(\mathbf{Y}|\boldsymbol{\Sigma}, \boldsymbol{\theta}_{\texttt{MAP}}), \tag{12}$$

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

$$\begin{aligned} \boldsymbol{\theta}_{\texttt{MAP}} &= \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma}_{\texttt{ML}}), \\ &= \arg\min_{\boldsymbol{\theta}} \left[\sum_{r=1}^{R} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}_{\texttt{ML}}^{-1} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta})) + \log g_\theta(\boldsymbol{\theta}) \right]. \end{aligned} \tag{13}$$

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

$$\boldsymbol{\Sigma}_{\texttt{ML}} = \frac{1}{R} \sum_{r=1}^{R} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}_{\texttt{MAP}})) (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}_{\texttt{MAP}}))^\top. \tag{14}$$

Note that $\boldsymbol{\Sigma}_{\texttt{ML}}$ depends on $\boldsymbol{\theta}_{\texttt{MAP}}$, and $\boldsymbol{\theta}_{\texttt{MAP}}$ depends on $\boldsymbol{\Sigma}_{\texttt{ML}}$. Similar approaches for dealing with unknown covariance can be found in [20, 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 $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}$, $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t)}$ of $\boldsymbol{\theta}_{\texttt{MAP}}$, $\boldsymbol{\Sigma}_{\texttt{ML}}$, respectively [12, 13]. Clearly, we desire the convergence as the number of iterations grow, $t \to \infty$, i.e.,

$$\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)} \longrightarrow \boldsymbol{\theta}_{\texttt{MAP}}, \tag{15}$$

$$\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t)} \longrightarrow \boldsymbol{\Sigma}_{\texttt{ML}}. \tag{16}$$

Table 1: Alternating optimization.

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

1 Estimate, by Monte Carlo,

$$\boldsymbol{\theta}_{\texttt{MAP}}^{(t)} = \arg\min_{\boldsymbol{\theta}} \left[\sum_{r=1}^{R} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}))^\top \left[\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t-1)} \right]^{-1} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta})) - \log g_\theta(\boldsymbol{\theta}) \right], \tag{17}$$

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

2 Compute

$$\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t)} = \frac{1}{R} \sum_{r=1}^{R} \left(\mathbf{y}_r - \mathbf{f}_r(\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}) \right) \left(\mathbf{y}_r - \mathbf{f}_r(\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}) \right)^\top. \tag{18}$$

The suggested iterative approach is summarized briefly by two steps. Starting with an initial matrix $\boldsymbol{\Sigma}_{\texttt{ML}}^{(0)}$, that is as a rough approximation of $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}$, the alternating optimization procedure is given in Table 1.
Since, we employ IS scheme for obtaining $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}$, 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 iteration, we can build a particle approximation of $p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)})$, i.e.,

$$\widehat{p}(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)}) = \sum_{t=1}^{T} \sum_{n=1}^{N} \widetilde{w}_t^{(n)} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)}), \qquad \sum_{t=1}^{T} \sum_{n=1}^{N} \widetilde{w}_t^{(n)} = 1. \tag{19}$$

By Eq. (19), we can approximate all the moments associate to the conditional posterior $p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)})$ 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 $\boldsymbol{\theta}_{\texttt{MAP}}$ can be bounded increasing N, i.e., even with a bad choice of $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t-1)}$ we can obtain a reasonable vector $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}$, and the estimator $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t)}$ in Eq. (18) approaches the matrix $\boldsymbol{\Sigma}_{\texttt{ML}}$ in Eq. (14), as $t \to \infty$. Moreover, as the number of realizations R grows the matrix $\boldsymbol{\Sigma}_{\texttt{ML}}$ in Eq. (14) converges to the true covariance matrix of the data.
Note that the pair $\boldsymbol{\theta}_{\texttt{MAP}}$ and $\boldsymbol{\Sigma}_{\texttt{ML}}$ are *fixed points* of *the iterative (dynamical) system* formed by Eqs. (17)-(18). Namely, the key point of the convergence is to be able to find a good approximation of $\boldsymbol{\theta}_{\texttt{MAP}}$ (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 [27, 28, 29, 30]. It has the same convergence rate of a Monte Carlo method for stochastic optimization, as a standard simulated

annealing [16].

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 say $T_0 < T$) of the algorithm considering

$$\boldsymbol{\theta}_{\texttt{MAP}}^{(t)} = \arg\min_{\boldsymbol{\theta}} \left[\sum_{r=1}^{R} \|\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta})\|^2 - \log g_\theta(\boldsymbol{\theta}) \right], \qquad t = 1, ..., T_0, \tag{20}$$

that is equivalent to set $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t)} = \mathbf{I}_K$ for $t = 0, ..., T_0 - 1$ in Eq. (17), where $\mathbf{I}_K$ is a $K \times K$ unit matrix. Thus, in the first T_0 iterations, we focus only in finding a good point $\boldsymbol{\theta}_{\texttt{MAP}}^{(T_0)}$. Indeed, note that if there exists a point $\boldsymbol{\theta}^*$ such that $\sum_{r=1}^{R} \|\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}^*)\|^2 = 0$, then this point $\boldsymbol{\theta}^*$ is also *a root* for $\sum_{r=1}^{R} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}^*))^\top \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{y}_r - \mathbf{f}_r(\boldsymbol{\theta}^*)) = 0$ for any possible covariance matrix $\widehat{\boldsymbol{\Sigma}}$.

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

Note that, in the iterative procedure, we have a sequence conditional posteriors $p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t)})$. For this reason, we call the algorithm as *adaptive target adaptive importance sampling* (ATAIS)[3] 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 the Step 1 of the alternating procedure described above. More generally, we will provide all the details of the ATAIS algorithm. For simplifying the notation, we denote the unnormalized conditional posterior at the t-th iteration,

$$\pi_t(\boldsymbol{\theta}) = \ell(\mathbf{Y}|\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t-1)}, \boldsymbol{\theta}) g_\theta(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t-1)}). \tag{21}$$

At each iteration, we consider $\pi_t(\boldsymbol{\theta})$ as the target distribution. Finally, we are able to approximate $\pi_{T+1}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)})$, without any additional evaluation of the likelihood function. The dependence on the iteration t is due to $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(t)}$ varies with t. The ATAIS algorithm is outlined in Table 6, 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 $q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)$ with mean $\boldsymbol{\mu}_t$ and a covariance matrix $\boldsymbol{\Lambda}_t$. An importance weight

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

is assigned to each sample $\boldsymbol{\theta}_t^{(n)}$, for all n and t.

[3] Another reason is that it is also an extension of the techniques in [14, 15], that use the acronym ATAIS as well.

Optimal denominator in IS weights. Since we adapt the proposal density during the iterations, we are actually in a multiple IS scenario [31, 22]. 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.,

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

This procedure provides the lowest variance of the final IS estimators but requires an 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 [32]. The simplest solution among the proposed one is to build a *compressed* denominator [33, 22]. Here, for avoiding instabilities in the results, we discard the samples in the first iterations when the proposal density changes substantially. For instance, one can discard the samples in the first iterations t such that $\|\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)} - \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t-1)}\| > \epsilon$ where ϵ is a small positive value.

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

$$\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}. \tag{22}$$

Note that, we set $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}$ 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, his choice helps in the search of 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 $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}$ due to (18).
The covariance matrix $\boldsymbol{\Lambda}_t$ is adapted by considering the empirical covariance of the weighted samples at the t-th iteration, plus a diagonal matrix controlled by a parameter $\delta > 0$ which determines the elements in the diagonal. The value of δ must be always greater than zero, since it helps the IS performance (see , e.g., [25, Numerical Example 1]) and avoids catastrophic scenarios. For a robust implementation, we suggest to use a greater value of δ specially in the first iterations of the algorithm. The value of δ could be decreased as the iterations grow.

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

$$\widetilde{w}_t^{(n)} = w_t^{(n)} \frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})}, \qquad \text{for all } n, t, \tag{23}$$

in order to obtain a particle approximation of the measure of the final conditional posterior $\pi_{T+1}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)})$. Thus, the algorithm returns the final estimators $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(T)}$, $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)}$, 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. (23) does not require any additional evaluations of the model, if the 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 $p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma})$.

3.2 Possible use of mini-batches

When the number of vectors $\mathbf{y}_r$ of data R grows, the calculation of the likelihood can become costly. ATAIS allows the direct use of mini-batches of data (see [34, 35]). 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,

$$\tilde{\pi}_t(\boldsymbol{\theta}) \propto \left[\prod_{i=1}^{L} \ell(\mathbf{y}_i|\boldsymbol{\Sigma}_{\texttt{ML}}^{(t-1)}, \boldsymbol{\theta})\right] g_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \tag{24}$$

where $\{y_i\}_{i=1}^{L}$ are $L < R$ vectors selected randomly over the R possible vectors. As stated in [34], ATAIS can use the subposteriors (24) in step **a.ii** in (6) reaching asymptotically the true values of $\boldsymbol{\theta}_{\texttt{map}}$ and $\boldsymbol{\Sigma}_{\texttt{ML}}$. 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, the complete ATAIS method can consider the use of mini-batches only in the first part, whereas, in the second part, the full-posterior must be evaluated.

3.3 Elliptically contoured distributions as 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 [36, 37, 38], i.e.,

$$\mathbf{v}_r \sim p(\mathbf{v}), \tag{25}$$

where represents the density below

$$p(\mathbf{v}) = \frac{k_p}{|\boldsymbol{\Sigma}|^{1/2}} h\left(\mathbf{v}^\top \boldsymbol{\Sigma}^{-1} \mathbf{v}\right),$$

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

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

where $\nu > 0$ represents the degrees of freedom.

4 Second part of the proposed inference scheme

Here, we describe the second part of the ATAIS procedure, which allow 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 6.

4.1 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 any 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.,

$$\widehat{p}(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma}) = \sum_{t=1}^{T}\sum_{n=1}^{N} \bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}) \cdot \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)}), \qquad \sum_{t=1}^{T}\sum_{n=1}^{N} \bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}) = 1, \tag{26}$$

where

$$\rho_t^{(n)}(\boldsymbol{\Sigma}) = \frac{\ell(\mathbf{Y}|\boldsymbol{\theta}_t^{(n)}, \boldsymbol{\Sigma}) g_\theta(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)}, \quad \text{and} \tag{27}$$

$$\bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}) = \frac{\rho_t^{(n)}(\boldsymbol{\Sigma})}{\sum_{\tau=1}^{T}\sum_{i=1}^{N} \rho_\tau^{(i)}(\boldsymbol{\Sigma})}. \tag{28}$$

Given a new matrix $\boldsymbol{\Sigma}$, to compute the likelihood

$$\ell(\mathbf{Y}|\boldsymbol{\theta}_t^{(n)}, \boldsymbol{\Sigma}) = \left(\frac{1}{(2\pi)^{K/2}\det(\boldsymbol{\Sigma})^{1/2}}\right)^R \exp\left(-\frac{1}{2}\sum_{r=1}^{R}\left(\mathbf{y}_r - \mathbf{f}_r\left(\boldsymbol{\theta}_t^{(n)}\right)\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_r - \mathbf{f}_r\left(\boldsymbol{\theta}_t^{(n)}\right)\right)\right), \tag{29}$$

$$= \left(\frac{1}{(2\pi)^{K/2}\det(\boldsymbol{\Sigma})^{1/2}}\right)^R \exp\left(-\frac{1}{2}\sum_{r=1}^{R}\left(\mathbf{e}_{t,r}^{(n)}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{e}_{t,r}^{(n)}\right),\right) \tag{30}$$

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

4.2 Approximation of the complete posterior distribution and marginal likelihood

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

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathbf{Y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\Sigma}, \mathbf{Y})}{p(\mathbf{Y})} = \frac{\ell(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) g_\theta(\boldsymbol{\theta}) g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})}{p(\mathbf{Y})}, \tag{31}$$

$$\propto \ell(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) g_\theta(\boldsymbol{\theta}) g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}), \tag{32}$$

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

$$\beta_{t,j}^{(n)} = \frac{\ell(\mathbf{Y}|\boldsymbol{\theta}_t^{(n)}, \boldsymbol{\Sigma}^{(j)}) g_\theta(\boldsymbol{\theta}_t^{(n)}) g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t) q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}, \tag{33}$$

$$= \rho_t^{(n)}(\boldsymbol{\Sigma}^{(j)}) \frac{g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}{q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})} = \rho_t^{(n)}(\boldsymbol{\Sigma}^{(j)}) \gamma_j = \rho_{t,j}^{(n)} \gamma_j, \tag{34}$$

where we have set $\gamma_j = \frac{g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}{q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}$ and $\rho_{t,j}^{(n)} = \rho_t^{(n)}(\boldsymbol{\Sigma}^{(j)})$ are given in Eq. (27). Clearly, if $q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) = g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$ then $\gamma_j = 1$. The complete posterior approximation is given by

$$\widehat{p}(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathbf{Y}) = \sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{n=1}^{N} \bar{\beta}_{t,j}^{(n)} \cdot \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)})\delta\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(j)}\right) \tag{35}$$

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 recall the recycling Gibbs sampling idea in [7], where the space is divided in blocks and different numbers of samples is considered for each block.
The marginal likelihood $p(\mathbf{Y})$ can be approximated as

$$p(\mathbf{Y}) \approx \widehat{p}(\mathbf{Y}) = \frac{1}{JNT}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{n=1}^{N}\beta_{t,j}^{(n)}. \tag{36}$$

4.3 Approximation of the marginal posteriors

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

$$p(\boldsymbol{\theta}|\mathbf{Y}) = \int_{\Sigma} p(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathbf{Y})d\boldsymbol{\Sigma} \approx \widehat{p}(\boldsymbol{\theta}|\mathbf{Y}) = \sum_{t=1}^{T}\sum_{n=1}^{N}\bar{\alpha}_t^{(n)} \cdot \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)}), \tag{37}$$

where

$$\bar{\alpha}_t^{(n)} = \frac{\sum_{j=1}^{J}\beta_{t,j}^{(n)}}{\sum_{i=1}^{J}\sum_{v=1}^{T}\sum_{m=1}^{N}\beta_{v,i}^{(m)}}. \tag{38}$$

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

$$p(\boldsymbol{\Sigma}^{(j)}|\mathbf{Y}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}, \boldsymbol{\Sigma}^{(j)}|\mathbf{Y})d\boldsymbol{\theta} \approx \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\beta_{t,j}^{(n)}}{\sum_{i=1}^{J}\sum_{v=1}^{T}\sum_{m=1}^{N}\beta_{v,i}^{(m)}}\delta\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(j)}\right), \tag{39}$$

$$= \bar{\lambda}_j \cdot \delta\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(j)}\right), \tag{40}$$

where

$$\bar{\lambda}_j = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\beta_{t,j}^{(n)}}{\sum_{i=1}^{J}\sum_{v=1}^{T}\sum_{m=1}^{N}\beta_{v,i}^{(m)}}. \tag{41}$$

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

$$p(\boldsymbol{\Sigma}|\mathbf{Y}) \approx \widehat{p}(\boldsymbol{\Sigma}|\mathbf{Y}) = \sum_{j=1}^{J}\bar{\lambda}_j \cdot \delta\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(j)}\right). \tag{42}$$

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 be also 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
$p(\boldsymbol{\theta}\|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_{\mathtt{ML}}^{(T)})$	$\widetilde{w}_t^{(n)}$ – See Eqs. (19) and (23)	
$p(\boldsymbol{\theta}\|\mathbf{Y}, \boldsymbol{\Sigma})$	$\bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}) = \frac{\ell(\mathbf{Y}\|\boldsymbol{\theta}_t^{(n)}, \boldsymbol{\Sigma}) g_\theta(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}\|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)}$	
$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}\|\mathbf{Y})$	$\bar{\beta}_{t,j}^{(n)} = \frac{\beta_{t,j}^{(n)}}{\sum_{i=1}^{J}\sum_{v=1}^{T}\sum_{m=1}^{N}\beta_{v,i}^{(m)}}$	$\beta_{t,j}^{(n)} = \frac{\ell(\mathbf{Y}\|\boldsymbol{\theta}_t^{(n)}, \boldsymbol{\Sigma}^{(j)}) g_\theta(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}\|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)} \cdot \frac{g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}{q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}$
$p(\boldsymbol{\theta}\|\mathbf{Y})$	$\bar{\alpha}_t^{(n)} = \frac{\sum_{j=1}^{J}\beta_{t,j}^{(n)}}{\sum_{i=1}^{J}\sum_{v=1}^{T}\sum_{m=1}^{N}\beta_{v,i}^{(m)}}$	$\beta_{t,j}^{(n)} = \rho_t^{(n)}(\boldsymbol{\Sigma}^{(j)}) \cdot \gamma_j$
$p(\boldsymbol{\Sigma}\|\mathbf{Y})$	$\bar{\lambda}_j = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\beta_{t,j}^{(n)}}{\sum_{i=1}^{J}\sum_{v=1}^{T}\sum_{m=1}^{N}\beta_{v,i}^{(m)}}$	
$p(\mathbf{Y})$	$\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{n=1}^{N}\beta_{t,j}^{(n)}$	

4.4 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

$$g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) = g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}|\boldsymbol{\Phi}, \nu) \propto |\boldsymbol{\Sigma}|^{\frac{\nu-K-1}{2}} \exp\left(-\frac{1}{2}\mathrm{trace}(\boldsymbol{\Phi}^{-1}\boldsymbol{\Sigma})\right), \tag{43}$$

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 an $K \times K$ *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

$$\boldsymbol{\Phi} = \frac{1}{\nu}\widehat{\boldsymbol{\Sigma}}_{\mathtt{ML}}^{(T)}. \tag{44}$$

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 the data by ATAIS (see also data-based priors in [26]). 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. (36), 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 Eq. (34).

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

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

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

1. Draw ν multivariate Gaussian samples $\mathbf{s}_i = [s_{i,1}, \ldots, s_{i,K}]^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi})$, with $i = 1, \ldots, \nu$.
2. Set $\boldsymbol{\Sigma}' = \sum_{i=1}^{\nu} \mathbf{s}_i \mathbf{s}_i^\top$.

5 Inverted layered importance sampling (ILIS)

The direct application of Monte Carlo methods in the complete space of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\theta}$ samples and $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}$ are obtained. Then the samples $\boldsymbol{\theta}$ are re-weighted and other weighted sample matrices $\boldsymbol{\Sigma}$ are generated.
In this section, we introduce a method, conceptually simpler than ATAIS, called inverted layered importance sampling (ILIS). See Table 3 for a detailed description. This method starts by generating N covariance matrices and running N different (parallel and indipendent) MCMC chains with target density $\bar{\pi}(\boldsymbol{\theta}|\boldsymbol{\Sigma}^{(n)}, \mathbf{Y})$, i.e., the conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{\Sigma}^{(n)}$. Some important considerations are provided below:

- 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 weight $\gamma_n = \dfrac{g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(n)})}{q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(n)})}$.
- ILIS can be seen as a Monte Carlo scheme which combines IS and MCMC techniques, based on *two layers*. With respect to the works in [21, 22], we can interpreted that the IS part forms the

upper layer, whereas the MCMC chains are generated in the lower layer of ILIS. and all these vectors are weighted with the weight $\gamma_n = \dfrac{g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(n)})}{q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(n)})}$.

- 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

$$\widehat{p}(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathbf{Y}) = \sum_{t=1}^{T}\sum_{n=1}^{N} \bar{\gamma}_n \cdot \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)})\delta\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(n)}\right), \quad \text{with} \quad \bar{\gamma}_n = \frac{\gamma_n}{\sum_{i=1}^{N}\gamma_i}. \tag{45}$$

Even if ILIS seems much simpler than ATAIS, the results are not particularly good compared to the obtained performance by ATAIS, as we show in the numerical simulations.

Joint use of ATAIS and ILIS. One interested practitioner could use the first part of ATAIS in Table 6 "to feed" with good parameters the proposal density $q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$ 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 an 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.

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

6 Simulations

We test the proposed scheme in three different numerical examples. It is important to remark that, in all the numerical experiments, any attempts of using a Monte Carlo approach (such an MCMC algorithm)

directly of the joint space $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$ produces much higher errors in estimation, which makes difficult to compare and visualize with respect to the results obtained by the proposed schemes. Therefore, we mainly compare complete ATAIS with ILIS.

6.1 Localization in wireless sensor network

To test the proposed methods, we aim to solve the task of determining 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. Lets recall that each observations has the form

$$y_r = \mathbf{f}_r(\boldsymbol{\theta}) + \mathbf{v}_r \tag{46}$$

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

$$\mathbf{f}_r(\boldsymbol{\theta}) = [-A \log(\| \boldsymbol{\theta} - \mathbf{s}_1 \|^2),\ -A \log(\| \boldsymbol{\theta} - \mathbf{s}_2 \|^2),\ -A \log(\| \boldsymbol{\theta} - \mathbf{s}_3 \|^2)] \tag{47}$$

i.e.,

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

for $i = 1, 2, 3$. The error term is $\mathbf{v_r} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\texttt{true}})$, with $\boldsymbol{\Sigma}_{\texttt{true}} \in \mathbb{R}^{3\times3}$ 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 $\boldsymbol{\theta}_{\texttt{true}} = [2.5, 2]$. In this scenario 50 observation vectors were generated. For comparison purposes, the prior over $\boldsymbol{\theta}$ was set as uniform, i.e., $g(\boldsymbol{\theta}) \propto 1$.

We test 3 different algorithms: (a) complete ATAIS, (b) ILIS using Metropolis Hastings (MH) chains and (c) a unique MH chain working only in the $\boldsymbol{\theta}$-space, keeping fixed the covariance matrix to the maximum likelihood estimation obtained by ATAIS in its first part (i.e., $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)}$). The results are given in Figure 3. Our goal is to approximate the MAP estimation of $\boldsymbol{\theta}$. All the results are averaged over 1000 independent runs.
In ATAIS, we set $T = 40$ iterations with $N = 50$ particles. 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 $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(0)}$ is set to the identity matrix $\mathbf{I}_3$. 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 generates $J = 10$ 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). Here, we consider all the chains and we estimate a unique estimation of the MAP of $\boldsymbol{\theta}$. Finally, for the unique/single MH chain addressing the conditional posterior keeping fixed $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)}$ (obtained by ATAIS), we consider again a random walk Gaussian proposal density (with inital $[0, 0]^\top$ and diagonal covariance $0.05\mathbf{I_2}$) and with length $T = 2000$.
Note that the number of evaluations of the non-linear model $\mathbf{f}$ differs 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.
In Figure 3(a), we can see the final ATAIS approximates of MAP of $\boldsymbol{\theta}$, 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 3. Looking at Figure 3 (a) it is clear that ATAIS (green squares) provides the best performance estimating $\boldsymbol{\theta}_{\texttt{true}}$ (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 worst than ATAIS. Additionally, we provide the mean absolute error (MAE) versus N, obtained by ATAIS in estimating the mean in $\boldsymbol{\theta}$ in Figure 4 (left). The error decreases as we increase the number of samples for both values of $T = 50$ and $T = 100$. Other MAE values (also in estimating the covariance matrix $\boldsymbol{\Sigma}$) obtained by ATAIS are given in Tables 4 and 5, in column "Localization". In Tables 4 and 5, $\boldsymbol{\Sigma}_{\texttt{ML}}$ is the true maximum of the likelihood function for the covariance matrix fixing $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_{\texttt{true}}$ (i.e., using $\boldsymbol{\theta}_{\texttt{true}}$ in Eq (18)). As expected, the best results are obtained increasing the number of particles and the iterations, allowing a better exploration of the parameter space.
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 6 is quite important. In order 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.

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 ($\boldsymbol{\Phi}$) equals 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. (39). 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 $\boldsymbol{\Sigma}$, as shown below (where we have averaged over 100 independent runs) in Eq (48). 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,

$$\begin{pmatrix} [0.6697, 1.3594] & [-0.3206, 0.2958] & [-0.6946, 0.3115] \\ [-0.3206, 0.2958] & [1.2584, 2.4827] & [-0.4709, 0.8096] \\ [-0.6946, 0.3115] & [-0.4709, 0.8096] & [2.9975, 5.8437] \end{pmatrix} \tag{48}$$

Note that the covariance matrix $\boldsymbol{\Sigma}$ represents the covariance among the different sensors in the network. In Figure 2, we can see the histograms obtained by the resampled particles of each components of $\boldsymbol{\Sigma}$ (after performing resampling according to the weights $\{\bar{\lambda}_j\}_{j=1}^{J}$). We must remark how the histograms corresponding to the null components of $\boldsymbol{\Sigma}_{\texttt{true}}$ have the mode very close to the value 0.

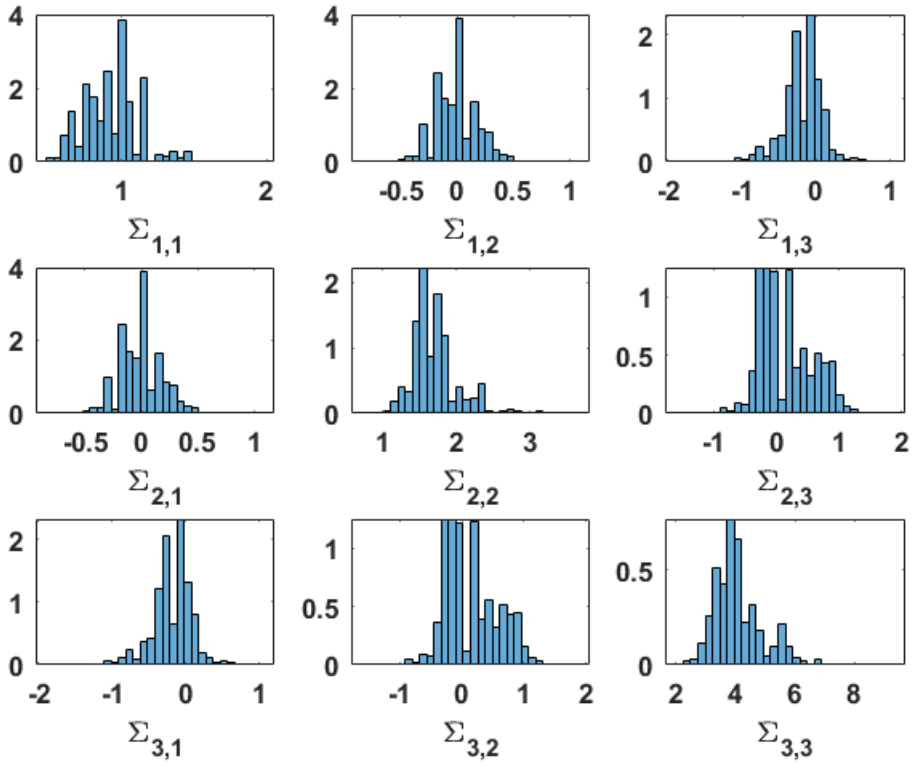
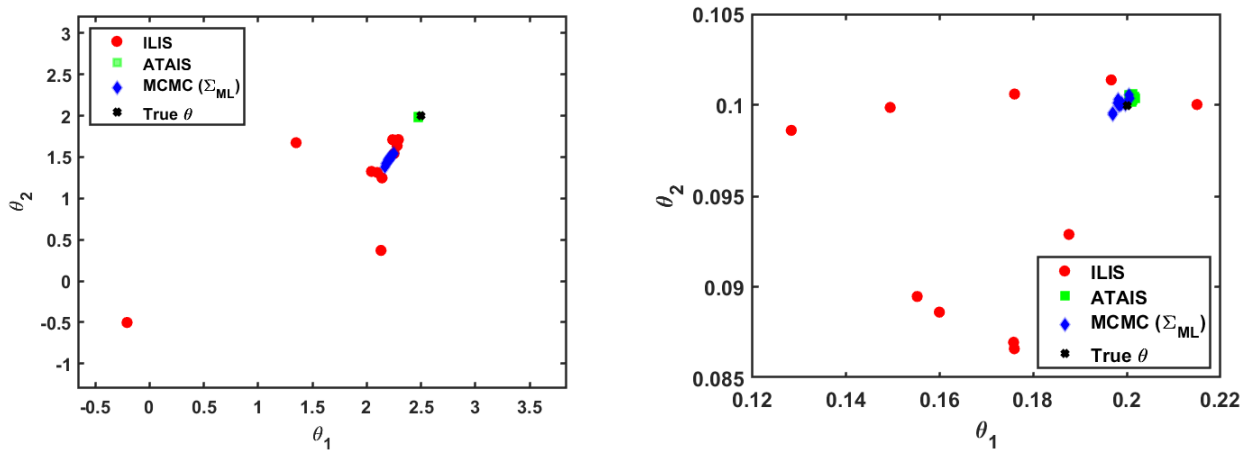


Figure 2: Histogram of the components, denoted as $[\boldsymbol{\Sigma}]_{i,j} = \boldsymbol{\Sigma}_{i,j}$, of the covariance matrices after resampling according to the weights $\bar{\lambda}_j$, for the location example.



(a) (b)

Figure 3: Green squares represent the estimation with ATAIS, red circles stand for the estimations of ILIS with a covariance matrix generated from an uninformative wishart distribution. The blue diamonds represent the estimations of the MCMC using the estimation of $\boldsymbol{\Sigma}_{\mathtt{ML}}$ found by ATAIS. The black cross stands for the true MAP. **(a)** location example; **(b)** multi-output example.

6.2 Multi-output model

In this second example, we take a multi-output model given by

$$\mathbf{y}_r = \mathbf{f}_r(\boldsymbol{\theta}) + \mathbf{v}_r \tag{49}$$

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

$$\begin{aligned} f_{r,1}(\theta_1) &= \theta_1 \sin(t)t, \\ f_{r,2}(\theta_2) &= \theta_2 \cos(t)t^2, \\ f_{r,3}(\boldsymbol{\theta}) &= f_{r,3}(\theta_1, \theta_2) = (\theta_1 + \theta_2)\sin(t)\cos(t), \\ f_{r,4}(\theta_2) &= \theta_2 t^2, \end{aligned} \tag{50}$$

Where the error term $\mathbf{v_r} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathtt{true}})$ with

$$\boldsymbol{\Sigma}_{\mathtt{true}} = \begin{pmatrix} 0.1 & 0.3 & 0.16 & 0 \\ 0.3 & 1.05 & 0 & 0 \\ 0.16 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2.95 \end{pmatrix}$$

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

$$D = M + \frac{K(K+1)}{2} = 2 + \frac{4(4+1)}{2} = 10.$$

This example states clearly the difficulty of performing an estimation $\boldsymbol{\Sigma}$ directly from the vectors observed, $\{\mathbf{y}_r\}_{r=1}^R$. This difficulty comes from the vectors not sharing the same theoretical mean. The prior for $\boldsymbol{\theta}$ was also taken as $g(\boldsymbol{\theta}) \propto 1$.

For this example we also test three algorithms: (a) complete ATAIS, (b) ILIS using Metropolis Hastings (MH) and (c) a single MH chain working exclusively in the $\boldsymbol{\theta}$-space, maintaining fixed the covariance matrix and equal to the maximum likelihood estimation obtained by ATAIS, as in the previous example. The results are presented in Figure 3(b). We aim to approximate the MAP estimate of $\boldsymbol{\theta}$. All the results are averaged over 1000 independent runs.

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, $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(0)}$, is the identity matrix, $\mathbf{I}_4$. 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 that fix the covariance matrix to the estimated $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)}$ by ATAIS we consider a Gaussian random walk proposal density. This proposal density has an initial mean $[0, 0]$ and has diagonal covariance matrix $0.005\mathbf{I}_2$. The length of the chain is $T = 2000$.

In Figure 3(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 single MH chain are displayed using blue diamonds. Looking at Figure 3(b) it is clear that the ATAIS gives the best estimations of $\boldsymbol{\theta}_{\texttt{true}}$ (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 $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)}$ that is the estimation provided by ATAIS. Then, the success of this third method is mainly due to an ATAIS ability.

Additionally, we provide the MAE versus the number of particles N, obtained by ATAIS in estimating the mean of $\boldsymbol{\theta}$ in Figure 4 (right). The error decreases as we increase the number of samples for both values of $T = 50$ and $T = 100$. Other MAE values (also in estimating the covariance matrix $\boldsymbol{\Sigma}$) obtained by ATAIS are given in Tables 4 and 5, in column "Multi-output". 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 $\boldsymbol{\Sigma}_{\texttt{ML}}$ is very small, the error estimating the $\boldsymbol{\Sigma}_{\texttt{true}}$ of the process can be high, since the difference between $\boldsymbol{\Sigma}_{\texttt{ML}}$ and $\boldsymbol{\Sigma}_{\texttt{true}}$ depends on the amount of data (they become closer with more data). This numerical example shows the strength of ATAIS, since in this multi-output problem the covariance matrix $\boldsymbol{\Sigma}$ cannot be approximately in advance directly from the data.

Credible interval with 95% **of probability for the matrix**. In order to perform a complete Bayesian inference over the covariance matrix $\boldsymbol{\Sigma}$, 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) equals to $\widehat{\Sigma}_{\mathtt{ML}}^{(T)}$ (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. (39). 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 (51) (where we have averaged over 100 independent runs). The first part of ATAIS was performed with $T = 50$ and $N = 50$.

$$\begin{pmatrix} [0.0614, 0.1586] & [0.1669, 0.4483] & [0.0703, 0.4736] & [-0.2578, 0.1985] \\ [0.1669, 0.4483] & [0.6125, 1.5325] & [-0.4739, 0.7620] & [-0.7571, 0.7682] \\ [0.0703, 0.4736] & [-0.4739, 0.7620] & [1.5868, 3.9349] & [-1.3594, 1.0586] \\ [-0.2578, 0.1985] & [-0.7571, 0.7682] & [-1.3594, 1.0586] & [2.4684, 9.1016] \end{pmatrix} \tag{51}$$

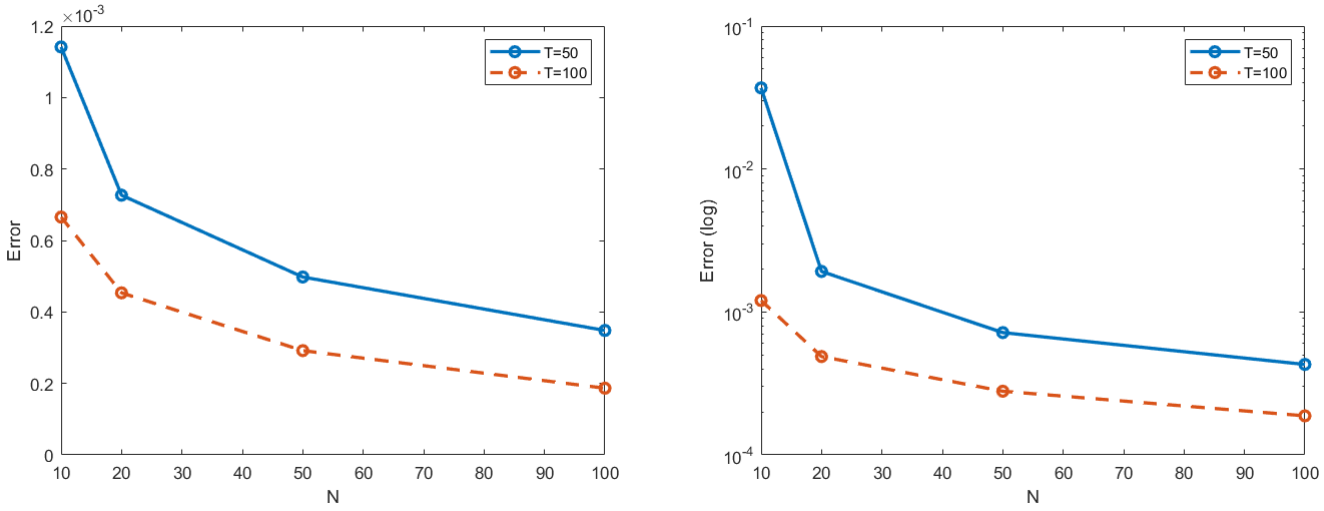


Figure 4: Mean absolute error (MAE) for the estimation of the true mean of the posterior with different number of particles by ATAIS. (Left: location example. Right: multi-output example with the scale for y-axis in logarithm).

Table 4: Mean absolute error (MAE) averaged over 1000 simulations of ATAIS for estimating the $\theta_{\mathtt{MAP}}$, $\Sigma_{\mathtt{ML}}$ and $\Sigma_{\mathtt{true}}$. For every value N, the number of iterations is fixed at 50.

	Localization			Multi-output		
N	$\theta_{\mathtt{map}}$	$\Sigma_{\mathtt{ML}}$	$\Sigma_{\mathtt{true}}$	$\theta_{\mathtt{map}}$	$\Sigma_{\mathtt{ML}}$	$\Sigma_{\mathtt{true}}$
5	0.0377	0.8934	1.0720	0.2325	0.7666	0.9094
12	0.0207	0.0443	0.2322	0.0102	0.0136	0.1789
25	0.0205	0.0443	0.2323	0.0013	0.0026	0.1709
50	0.0205	0.0442	0.2324	0.0012	0.0023	0.1713
100	0.0204	0.0442	0.2325	0.0009	0.0017	0.1712

Table 5: Mean absolute error over 100 simulations of ATAIS for estimating the $\boldsymbol{\theta}_{\texttt{MAP}}$, $\boldsymbol{\Sigma}_{\texttt{ML}}$ and $\boldsymbol{\Sigma}_{\texttt{true}}$. For every value T, the number of particles is fixed at 100.

	Location			Multi-output		
T	$\boldsymbol{\theta}_{\texttt{map}}$	$\boldsymbol{\Sigma}_{\texttt{ML}}$	$\boldsymbol{\Sigma}_{\texttt{true}}$	$\boldsymbol{\theta}_{\texttt{map}}$	$\boldsymbol{\Sigma}_{\texttt{ML}}$	$\boldsymbol{\Sigma}_{\texttt{true}}$
5	0.1740	2.4068	2.4644	0.2100	0.4648	0.5526
10	0.0758	0.4292	0.5360	0.1219	0.1293	0.2328
20	0.0328	0.5835	0.6933	0.0355	0.2663	0.3594
30	0.0205	0.0441	0.2326	0.0015	0.0031	0.1711
50	0.0205	0.0443	0.2324	0.0010	0.0021	0.1714

6.3 Application to a biology system

In this third example we focus on an inference problem in a biology system [39]. 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 $\boldsymbol{\theta} = (k_{12}, k_{21}, k_{1e}, b)$. The dynamics of the system is rule by the following differential equations:

$$\begin{aligned} \frac{dx_1(t)}{dt} &= -(k_{1e} + k_{12}) \cdot x_1(t) + k_{21} \cdot x_2(t) + b \cdot u(t), \\ \frac{dx_2(t)}{dt} &= k_{12} \cdot x_1(t) - k_{21} \cdot x_2(t), \end{aligned} \tag{52}$$

where

$$u(t) = \begin{cases} t + 0.5 & \text{if } 0 \leq t \leq 1 \\ 1.5e^{1-t} & \text{if } t > 1, \end{cases} \tag{53}$$

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

$$\boldsymbol{\Sigma}_{\texttt{true}} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 2 \end{pmatrix}.$$

The prior for the components of $\boldsymbol{\theta}$ was set as uniform in the interval $[0, 5]$ (for all the components) as suggested in [39]. The system above has not analytically close solution. Hence, the generation of the data according to the system is obtained using a Runge-Kutta Matlab solver [40]. We assume that the solution is perturbed by Gaussian noise according to the distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_{\texttt{true}})$. Then, 100 data points are sampled at equidistant times from the solution of the system (52).

Remark. Note that, in this experiment, the function $\mathbf{f}$ is evaluated only approximately since it is the solution of the system (52) which can not be evaluated exactly. Thus, even considering the vector of true values $\boldsymbol{\theta}_{\texttt{true}}$, we need to approximate $\mathbf{f}(\boldsymbol{\theta}_{\texttt{true}})$ by a differential equation solver (discretizing it) obtaining $\widehat{\mathbf{f}}(\boldsymbol{\theta}_{\texttt{true}})$. Since also the vector of $\boldsymbol{\theta}$ must be estimated, here we can interpret that we are doing a double approximation of $\mathbf{f}(\boldsymbol{\theta})$.

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, $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(0)}$, was the identity, $\mathbf{I}_2$.
In the second part of ATAIS, we generate $J = 1000$ matrices from a Whishart distribution (degrees of freedom, $\nu = 100$) with reference matrix equal to $\widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)}$, the final estimate of the first part of ATAIS. To each of the generated matrices we assign a weight following Eq (39). 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 (54). It can be seen that our interval contains the true components of the covariance matrix.

$$\begin{pmatrix} [0.7820, 1.3876] & [0.7099, 1.4614] \\ [0.7099, 1.4614] & [1.6518, 2.8760] \end{pmatrix} \tag{54}$$

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 (52), even if the distance to the $\boldsymbol{\theta}_{\texttt{true}}$ could be large.

7 Conclusions

In this work, we have introduced an adaptive importance sampling (AIS) method for robust inference in complex Bayesian inversion problems with unknown parameters $\boldsymbol{\theta}$ of the non-linear mapping and unknown covariance matrix $\boldsymbol{\Sigma}$ of the noise perturbation. The variables of interest are split in two blocks, the parameters $\boldsymbol{\theta}$ of the non-linear model and the covariance matrix $\boldsymbol{\Sigma}$, are handled in different ways. The main proposed inference scheme is divided in two main parts. The first part is devoted to approximate a conditional posterior $\boldsymbol{\theta}$ given the data and the maximum likelihood estimator of $\boldsymbol{\Sigma}$. This first part allows of finding regions of high probability about $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ (working alternately in subsets of the complete space, with reduced dimensions).
In the second part, a Bayesian approach is also performed over $\boldsymbol{\Sigma}$ re-using and re-weighting the samples of $\boldsymbol{\theta}$ previously generated. Then, an approximation of the complete posterior of $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$ is provided. This second part does not required 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 adaptive importance sampler that addresses a sequence of different conditional posteriors and a post-process that allows a Bayesian inference also over $\boldsymbol{\Sigma}$. Additionally, a simpler compelling scheme, called ILIS, has been introduced in order to compare with ATAIS. In the numerical simulations presented in this work we can see the good performance of ATAIS providing a complete Bayesian analysis of the complete space $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$.

Acknowledgments

The work was partially supported by the Young Researchers R&D Project, ref. num. F861 (AUTO-BA-GRAPH) funded by Community of Madrid and Rey Juan Carlos University, and by Agencia Estatal de Investigación AEI (project SP-GRAPH, ref. num. PID2019-105032GB-I00).

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] 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.

[15] 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.

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

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

[18] 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.

[19] 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.

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

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

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

[23] 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.

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

[25] 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.

[26] 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.

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

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

[29] 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.

[30] 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.

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

[32] 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.

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

[34] 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.

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

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

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

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

[39] 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.

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

Table 6: ATAIS: an adaptive IS scheme with a sequence of adaptive target pdfs

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

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

$$w_t^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)}, \qquad n = 1, ..., N. \tag{55}$$

   (b) **Current maximum estimations:**

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

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

   (c) **Global maximum estimations:**

   - If $\pi_t(\boldsymbol{\theta}_{\mathtt{max}}^{(t)}) > \pi_{\mathtt{MAP}}$:
     i. $\widehat{\boldsymbol{\theta}}_{\mathtt{MAP}}^{(t)} = \boldsymbol{\theta}_{\mathtt{max}}^{(t)}$,
     ii. $\widehat{\boldsymbol{\Sigma}}_{\mathtt{ML}}^{(t)} = \widehat{\boldsymbol{\Sigma}}_t$,
     iii. Update according to $\widehat{\boldsymbol{\theta}}_{\mathtt{MAP}}^{(t)}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathtt{ML}}^{(t)}$, i.e, $\pi_{\mathtt{MAP}} = \pi_{t+1}(\widehat{\boldsymbol{\theta}}_{\mathtt{MAP}}^{(t)})$.
   - Otherwise $\widehat{\boldsymbol{\theta}}_{\mathtt{MAP}}^{(t)} = \widehat{\boldsymbol{\theta}}_{\mathtt{MAP}}^{(t-1)}$, and $\widehat{\boldsymbol{\Sigma}}_{\mathtt{ML}}^{(t)} = \widehat{\boldsymbol{\Sigma}}_{\mathtt{ML}}^{(t-1)}$.

   (d) **Adaptation:** Set

$$\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\mathtt{MAP}}^{(t)}, \tag{56}$$

$$\boldsymbol{\Lambda}_t = \sum_{n=1}^{N} \bar{w}_t^{(n)}(\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t)^\top(\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t) + \delta\mathbf{I}_M, \tag{57}$$

   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 $\delta > 0$.

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

$$\widetilde{w}_t^{(n)} = w_t^{(n)} \frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})}. \tag{58}$$