Working paper

2023-05

Statistics and Econometrics
ISSN 2387-0303

Adaptive posterior distributions for covariance matrix learning in Bayesian inversion problems for multioutput signals

Ernesto Curbelo, Luca Martino, Fernando Llorente, David Delgado-Gómez

uc3m | Universidad Carlos III de Madrid

Adaptive posterior distributions for covariance matrix learning 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.

May 29, 2023

Abstract

In this work, we propose an adaptive importance sampling (AIS) scheme for multivariate Bayesian inversion problems, which is based in two main ideas: the inference procedure is divided in two parts and the variables of interest are split in two blocks. We assume that the observations are generated from a complex multivariate non-linear function perturbed by correlated Gaussian noise. 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. Two numerical examples are presented that show the benefits of the proposed approach.
Keywords: Bayesian inversion; importance Sampling; 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 the posterior density over the unknown parameters. The computation of complicated integrals involving these posterior distributions are often needed (e.g., any moments 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 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 [1, 2, 3, 4]. The last two classes are the most used by the user, 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 [3, 5, 6, 7].

In this work, we extend the approach in [8, 9]. 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, $\mathbf{v}_r \sim \mathcal{N}(\mathbf{v}_r|\mathbf{0}, \boldsymbol{\Sigma})$. 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 that no assumptions have been made over $\mathbf{f}$, and usually it represents some complex physical process. For instance, $\mathbf{f}(\boldsymbol{\theta})$ could be also non-differentiable. The unique requirement is to be able to evaluate point-wise $\mathbf{f}(\boldsymbol{\theta})$. The inference task on the complete space $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$ can be particularly challenging, firstly due to the dimension of the space and, secondly, since wrong choices for the covariance matrix $\boldsymbol{\Sigma}$ can jeopardize the inference over $\boldsymbol{\theta}$ (and viceversa). Therefore, the proposed idea is to split the variables of interest in two blocks, $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ (as in a block Gibbs sampling) and to separate the inference problem 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 $\boldsymbol{\Sigma}$. In the second part, we also perform Bayesian inference of the covariance matrix $\boldsymbol{\Sigma}$ and approximate the complete posterior of pair of variables of interest $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$. The resulting scheme is a more robust inference approach for Bayesian inversion, based on adaptive importance sampler that addresses a sequence of different conditional posteriors.

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. The likelihood 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.

Application to time series or spatial processes. In the case of having K different time series in continuous time, or K spatial processes we can have more explicit notation, where have a one-to-one correspondence between each $r \in \{1, ..., R\}$ and a real time instant $\tau_{k,r} \in \mathbb{R}$ or a point $\mathbf{x}_{k,r} \in \mathbb{R}^L$, i.e.,

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

More generally, a more explicit notation, instead of $\mathbf{f}_r(\boldsymbol{\theta})$, in these scenarios would be the following:

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

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

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

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).

Bayesian inference in the complete space. The full Bayesian solution considers the study of the complete posterior density

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

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

In this first stage, we consider a sub-optimal (in Bayesian sense) but substantially more efficient inference scheme, studying only a conditional posterior distribution. 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 $\boldsymbol{\Sigma}_{\texttt{ML}}$.

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 $\widehat{\boldsymbol{\theta}}^{(t)}_{\texttt{MAP}}$, $\widehat{\boldsymbol{\Sigma}}^{(t)}_{\texttt{ML}}$ of $\boldsymbol{\theta}_{\texttt{MAP}}$, $\boldsymbol{\Sigma}_{\texttt{ML}}$, respectively. Clearly, we desire the convergence as the number of iterations grow, $t \rightarrow \infty$, i.e.,

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

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

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

Table 1: Alternating optimization.

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

1 Estimate $\widehat{\boldsymbol{\theta}}^{(t)}_{\texttt{MAP}}$ by Monte Carlo, e.g., by an importance sampling (IS) scheme, working with respect to $p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}^{(t-1)}_{\texttt{ML}})$, i.e., equivalently

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

2 Compute

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

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

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

By Eq. (28), we can approximate all the moments associate to the conditional posterior $p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}^{(T)}_{\texttt{ML}})$ 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 $\widehat{\boldsymbol{\Sigma}}^{(t-1)}_{\texttt{ML}}$ we can

obtain a reasonable vector $\widehat{\boldsymbol{\theta}}^{(t)}_{\texttt{MAP}}$), and the estimator $\widehat{\boldsymbol{\Sigma}}^{(t)}_{\texttt{ML}}$ 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. Moreover, 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).

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

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

that is equivalent to set $\widehat{\boldsymbol{\Sigma}}^{(t)}_{\texttt{ML}} = \mathbf{I}_K$ for $t = 0, ..., T_0 - 1$ in Eq. (17), where $\mathbf{I}_K$ is a $K \times K$ unit matrix. Thus, we avoid the alternating optimization in the first T_0 iterations, and we use them to find a good point $\boldsymbol{\theta}^{(T_0)}_{\texttt{MAP}}$. 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 $p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}^{(T)}_{\texttt{ML}})$. With respect to $\boldsymbol{\Sigma}$, we only provide a frequentist estimator $\widehat{\boldsymbol{\Sigma}}^{(T)}_{\texttt{ML}}$.

Note that, in the iterative procedure, we have a sequence conditional posteriors $p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}^{(t)}_{\texttt{ML}})$. For this reason, we call the algorithm as *adaptive target adaptive importance sampling* (ATAIS)[1] The details of the ATAIS algorithm which performs this scheme are given in the next section.

4 Adaptive Target Adaptive Importance Sampling (ATAIS)

This section is devoted to provide more details with respect to 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,

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

At each iteration, we consider $\pi_t(\boldsymbol{\theta})$ as the target distribution but, finally, we are able to approximate $\pi_{T+1}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}^{(T)}_{\texttt{ML}})$, without any additional evaluation of the likelihood function. The dependence on the iteration t is due to $\widehat{\boldsymbol{\Sigma}}^{(t)}_{\texttt{ML}}$ varies with t. The ATAIS algorithm is outlined in Table 2, whereas the main feature of ATAIS are described below.

IS steps. A set of N samples $\{\boldsymbol{\theta}^{(n)}_t\}_{n=1}^{N}$ 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

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

[1] Another reason is that it is also an extension of the techniques in [8, 9], that use the acronym ATAIS as well.

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

Optimal denominator in IS weights. Since we adapt the proposal density during the iterations, we are actually in a multiple IS scenario [10, 11]. 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 [12]. The simplest solution among the proposed one is to build a *compressed* denominator [13, 11]. 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., [14, 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 $p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma})$.

5 Second part of the proposed inference scheme

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 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 generates 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.,

$$\widehat{p}(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Sigma}) = \sum_{t=1}^{T}\sum_{n=1}^{N} \bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}) \cdot \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)}), \qquad \sum_{t=1}^{T}\sum_{n=1}^{N} \bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}) = 1, \tag{28}$$

where

$$\bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}) \propto \rho_t^{(n)}(\boldsymbol{\Sigma}) = \frac{\ell(\mathbf{Y}|\boldsymbol{\theta}_t^{(n)}, \boldsymbol{\Sigma}) g_\theta(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)}, \tag{29}$$

$$\bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}) = \frac{\rho_t^{(n)}(\boldsymbol{\Sigma})}{\sum_{\tau=1}^{T}\sum_{i=1}^{N} \bar{\rho}_\tau^{(i)}(\boldsymbol{\Sigma})}. \tag{30}$$

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

$$\ell(\mathbf{Y}|\boldsymbol{\theta}_t^{(n)}, \boldsymbol{\Sigma}) = \left(\frac{1}{(2\pi)^{K/2}\det(\boldsymbol{\Sigma})^{1/2}}\right)^R \exp\left(-\frac{1}{2}\sum_{r=1}^{R}\left(\mathbf{y}_r - \mathbf{f}_r\left(\boldsymbol{\theta}_t^{(n)}\right)\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_r - \mathbf{f}_r\left(\boldsymbol{\theta}_t^{(n)}\right)\right)\right), \tag{31}$$

$$= \left(\frac{1}{(2\pi)^{K/2}\det(\boldsymbol{\Sigma})^{1/2}}\right)^R \exp\left(-\frac{1}{2}\sum_{r=1}^{R}\left(\mathbf{e}_{t,r}^{(n)}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{e}_{t,r}^{(n)}\right),\right) \tag{32}$$

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

5.2 Approximating the complete posterior

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

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

and a proposal density factorized as $q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t) q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$. 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

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

$$= \bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}^{(j)}) \frac{g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}{q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})} = \bar{\rho}_t^{(n)}(\boldsymbol{\Sigma}^{(j)}) \gamma_j, \tag{34}$$

where we have set $\gamma_j = \frac{g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}{q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{(j)})}$. 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 $\bar{\beta}_{t,j}^{(n)} = \frac{\beta_{t,j}^{(n)}}{\sum_{i=1}^{J} \sum_{v=1}^{T} \sum_{m=1}^{N} \bar{\beta}_{v,i}^{(m)}}$. 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 [5], where the space is divided in blocks and different numbers of samples is considered for each block.

5.3 Approximation of the marginal posterior of the covariance matrix

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

$$p(\boldsymbol{\Sigma}^{(j)}|\mathbf{Y}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}, \boldsymbol{\Sigma}^{(j)}|\mathbf{Y}) d\boldsymbol{\theta} \approx \frac{\sum_{t=1}^{T} \sum_{n=1}^{N} \beta_{t,j}^{(n)}}{\sum_{i=1}^{J} \sum_{v=1}^{T} \sum_{m=1}^{N} \beta_{v,i}^{(m)}} \delta\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(j)}\right), \tag{36}$$

$$= \bar{\lambda}_j \cdot \delta\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(j)}\right). \tag{37}$$

For instance, an 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.

5.4 Prior over covariance matrices

Consider a positive definite $K \times K$ matrix $\boldsymbol{\Sigma} = \{\sigma_{i,j}\}$ with $i, j = 1, \ldots, K$. The Wishart distribution is defined on the space $\mathbb{R}^K \times \mathbb{R}^K$ of positive definite matrices. The corresponding pdf is

$$g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{\frac{\nu-K-1}{2}} \exp\left(-\frac{1}{2}\text{trace}(\boldsymbol{\Phi}^{-1}\boldsymbol{\Sigma})\right), \tag{38}$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of the matrix $\boldsymbol{\Sigma}$, $\nu \geq K$ is the number of degrees of freedom and $\boldsymbol{\Phi}$ is an $K \times K$ *reference* covariance matrix. The Wishart distribution is often interpreted as a multivariate extension of the Gamma distribution. We choose here

$$\boldsymbol{\Phi} = \frac{1}{\nu} \widehat{\boldsymbol{\Sigma}}_{\texttt{ML}}^{(T)}. \tag{39}$$

In this sense, our approach is an *empirical Bayes approach* since this parameter of the prior is chosen after looking the data by ATAIS (see also data-based priors in [15]). For simplicity, we assume $q_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) = g_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$ then, as a consequence, $\gamma_j = 1$ in Eq. (34).

Generation of random matrices according to the Wishart prior. When ν is an integer, the Wishart distribution represents the sums of squares (and cross-products) of n draws from a multivariate Gaussian distribution. Specifically, given ν random $K \times 1$ vectors $\mathbf{s}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi})$, $i = 1, \ldots, \nu$, of dimension K, the 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, trivially we can derived 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 $\boldsymbol{\Sigma}' = \sum_{i=1}^{\nu} \mathbf{s}_i \mathbf{s}_i^{\top}$.

6 Simulations

We test the proposed scheme in two numerical examples. The first numerical experiment is a simple bidimensional location example where the function $\mathbf{f}$ does no depend on the time instant. The second experiment considers a four dimensional signal depending on the instant of time with the unknown parameters lying on a 2D space.

6.1 First numerical analysis

To test our proposed method, we aim to solve the task of determining the location of a target in a two-dimensional space 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 $\boldsymbol{s}_1, ..., \boldsymbol{s}_K$. We collect R measurements from each sensor, and these measurements follow a certain distribution. Lets recall that each observations has the from

$$y_r = \mathbf{f}_r(\boldsymbol{\theta}) + \mathbf{v}_r \tag{40}$$

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

$$\mathbf{f}_r(\boldsymbol{\theta}) = [-A \log(\| \boldsymbol{\theta} - \boldsymbol{s}_1 \|^2),\ -A \log(\| \boldsymbol{\theta} - \boldsymbol{s}_2 \|^2),\ -A \log(\| \boldsymbol{\theta} - \boldsymbol{s}_3 \|^2)] \tag{41}$$

The error term is $\mathbf{v_r} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\texttt{true}})$, with $\Sigma_{\texttt{true}} \in \mathbb{R}^{3\times 3}$ 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.
To facilitate simulation, we consider a scenario where there are $K = 3$ 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: $\boldsymbol{s}_1 = [0.5, 1]$, $\boldsymbol{s}_1 = [3.5, 1]$ and $\boldsymbol{s}_3 = [2, 3]$. 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., $g(\boldsymbol{\theta}) \propto 1$. In table 3 and 4 we present the results for this example under the column "Location". As a remark, we must highlight that we are measuring the error in the estimation of two matrices, one is the real covariance matrix which we know because we user simulated data and the other is the covariance matrix we obtain if we replaced the true value of $\boldsymbol{\theta}$ when calculating the covariance matrix of the observations. As expected, the best results are obtained increasing the number of particles and the iterations, allowing a better exploration of the parameter space. Enhancing the exploration of the algorithm is a hard task in this example, given that the posterior is very narrow, which makes it hard for the generated samples to be in regions with high posterior evaluation, this is why the adaptation of the proposal covariance in step 2d of Table 2 is very important. In order to improve even more the exploration, at some iteration the covariance matrix of the proposal density can be increased to allow exploration of areas away from the narrow mode. For the task of finding the mode the algorithm does no need many particles, e.g., when using 50 iterations it is enough to use 10 particles. The importance of the number of particles is evidenced in Figure 1 (left), where we can see the mean absolute error when calculating the mean of the posterior using the samples from the iterative process. In this case, even when the algorithm performs well with a few particles per iteration, the error continue to decrease as we increase the number of samples for both, $T = 50$ and $T = 100$.

Credible interval with 95% **of probability for the matrix**. In order to perform a Bayesian inference over the covariance matrix, we consider a Whisart proposal with $\nu = 100$ (degrees of freedom) with a reference matrix ($\mathbf{\Phi}$) equals to $\widehat{\mathbf{\Sigma}}^{(T)}_{\texttt{ML}}$. With this proposal distribution, we generate J matrices and assign a weight to each of the following Eq. (36). Applyng resampling according to the 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 $\mathbf{\Sigma}$, as shown below (where we have averaged over 50 indipendent runs),

$$\begin{pmatrix} [0.6361; 1.1075] & [-0.2343; 0.2230] & [-0.5306; 0.1851] \\ [-0.2343; 0.2230] & [1.1698; 2.0466] & [-0.3207; 0.6487] \\ [-0.5306; -0.5306] & [-0.3207; 0.6487] & [2.8710; 5.0023] \end{pmatrix}. \tag{42}$$

Remark. The particularity of this simple example makes it possible to estimate the desired covariance matrix by calculating the covariance matrix of the observations. This is just a particular case where

the observation share the same theoretical mean, but our method is able to accomplish the estimation in a more general scenario where each observation has its own mean. As example of this, we present an example where the signal is given by a four dimension vector, and the components of such vector are not uncorrelated.

6.2 Multi-output model

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

$$\mathbf{y}_r = \mathbf{f}_r(\boldsymbol{\theta}) + \mathbf{v}_r \tag{43}$$

where the vector function $\mathbf{f}_r : \mathbb{R}^2 \to \mathbb{R}^4$ is given by the components

$$\begin{array}{rcl} f_{r,1}(\boldsymbol{\theta}) & = & \theta_1 \sin(t)t, \\ f_{r,2}(\boldsymbol{\theta}) & = & \theta_2 \cos(t)t^2, \\ f_{r,3}(\boldsymbol{\theta}) & = & (\theta_1 + \theta_2)\sin(t)\cos(t), \\ f_{r,4}(\boldsymbol{\theta}) & = & \theta_2 t^2, \end{array} \tag{44}$$

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

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

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 $g(\boldsymbol{\theta}) \propto 1$.

The results for this example are shown in Tables 3 and 4 under the column multi-output. The importance of using a large number of particles and iterations is also evidenced. It is fair to comment that the even when the error of the estimation of $\boldsymbol{\Sigma}_{\texttt{ML}}$ is very small, the error estimating the true $\boldsymbol{\Sigma}$ of the process can be high, because this estimation depends on the amount of data we have, being better when there are many data. We must highlight that this example shows the strength of ATAIS, since in this problem the covariance matrix cannot be calculated directly from the observations. Even though it is evidenced that if the number of particles is great enough then there is no need of a very large number of iterations, which can reduce the computational time of the algorithm. Once more, the particles obtained through ATAIS are used to calculate the mean of the posterior. The errors are shown in Figure 1 (right), in log scale, for a better appreciation. As in the previous example, for both cases ($T = 50$ and $T = 100$) increasing the number of samples will lead to a better exploration of the sample space which results in better approximations of the true conditional mean of the posterior.

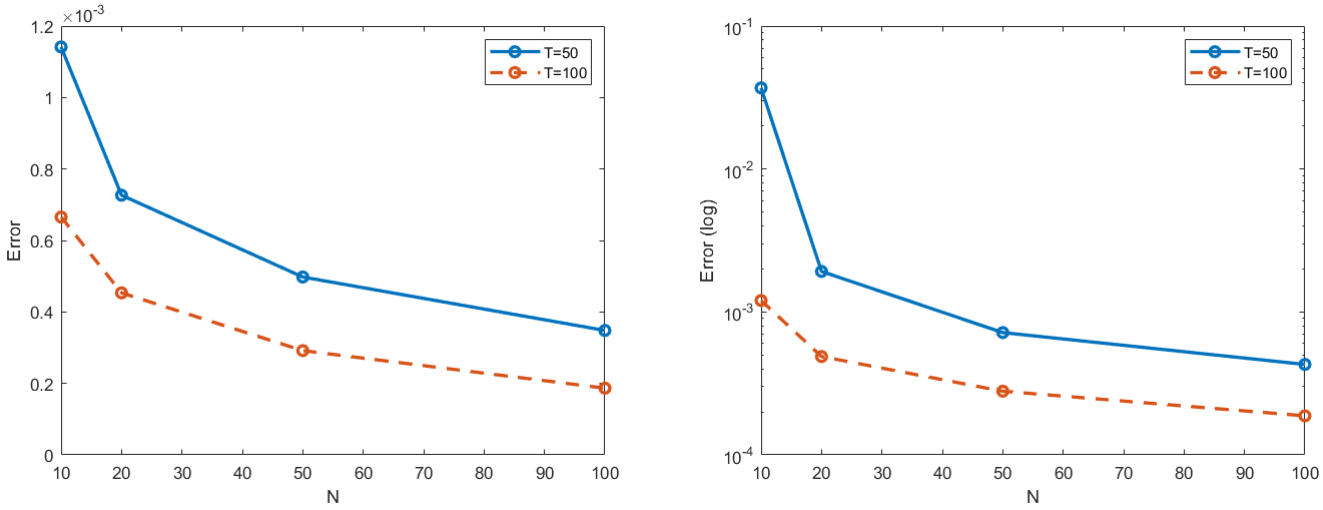

Figure 1: Mean absolute error for the estimation of the true mean of the posterior with different number of particles. (Left: location example. Right: multi-output example with the scale for y-axis in logarithm).

7 Conclusions

In this paper we introduce 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. Moreover, the 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 $\boldsymbol{\Sigma}$. In the second part, a Bayesian approach is also performed over $\boldsymbol{\Sigma}$ and an approximation of the complete posterior of $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$ is provided.
The proposed AIS algorithm, called ATAIS, is employed in the first part of the inference scheme. In each iteration, ATAIS deals with a sequence of posterior distributions depending on the current estimation of the covariance matrix. All the particles yielded by ATAIS are finally reused for obtaining a the complete posterior of $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$. Differente numerical simulations shows the good performance of ATAIS.

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

[2] 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.

[3] 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.

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

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

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

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

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

[14] 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.

[15] 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.

Table 2: ATAIS: AIS with adaptive target pdf

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

$$w_t^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t)}, \qquad n = 1, ..., N. \tag{24}$$

   (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

$$\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\mathtt{MAP}}^{(t)}, \tag{25}$$

$$\boldsymbol{\Lambda}_t = \sum_{n=1}^{N} \bar{w}_t^{(n)}(\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t)^\top(\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t) + \delta\mathbf{I}_M, \tag{26}$$

   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

$$\widetilde{w}_t^{(n)} = w_t^{(n)} \frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})}. \tag{27}$$

Table 3: 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 N, the number of iterations is fixed at 50.

	Location			Multi-output		
N	$\boldsymbol{\theta}_{\texttt{map}}$	$\boldsymbol{\Sigma}_{\texttt{ML}}$	$\boldsymbol{\Sigma}_{\texttt{true}}$	$\boldsymbol{\theta}_{\texttt{map}}$	$\boldsymbol{\Sigma}_{\texttt{ML}}$	$\boldsymbol{\Sigma}_{\texttt{true}}$
5	0.0377	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 4: 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