Automatic tempered posterior distributions for inverse problems

L. Martino[†], J. López-Santiago[*], J. Míguez[*]

[†] Universidad rey Juan Carlos (URJC), Madrid, Spain.

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

April 17, 2020

Abstract

We propose a new Monte Carlo technique for Bayesian inversion problem. The power of the noise perturbation in the observation model is also estimated jointly with the rest of parameters. Moreover, it is also used as a tempered parameter. Hence, a sequence of tempered posterior densities is considered where the tempered parameter is automatically selected according to the actual estimation of the power of the noise perturbation.

1 Introduction

UNDER CONSTRUCTION

2 Problem Statement

Let denote the observed measurements as $\mathbf{y} = [y_1, ..., y_K]^\top \in \mathbb{R}^K$, and the variable of interest that we desire to infer, as $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M]^\top \in \boldsymbol{\Theta} \subseteq \mathbb{R}^M$. Furthermore, let consider the observation model

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}) + \mathbf{e}, \tag{1}$$

where we have a nonlinear mapping,

$$\mathbf{f}(\boldsymbol{\theta}) = [f_1(\boldsymbol{\theta}), ..., f_K(\boldsymbol{\theta})]^\top : \boldsymbol{\Theta} \subseteq \mathbb{R}^M \to \mathbb{R}^K, \tag{2}$$

and a Gaussian perturbation noise,

$$\mathbf{e} = [e_1, ..., e_K]^\top \sim \mathcal{N}(\mathbf{e}|\mathbf{0}, \sigma^2 \mathbf{I}_K), \tag{3}$$

with $\sigma > 0$, and we have denoted the K-dimensional unit matrix as $\mathbf{I}_K$. The noise variance σ^2 is unknown, in general. The mapping $\mathbf{f}$ could be analytically unknown: the only assumption is

that we are able to evaluate pointwise the nonlinear mapping $\mathbf{f}(\boldsymbol{\theta})$. The likelihood function is

$$\ell(\mathbf{y}|\boldsymbol{\theta},\sigma) = \frac{1}{(2\pi\sigma^2)^{K/2}}\exp\left(-\frac{1}{2\sigma^2}||\mathbf{y}-\mathbf{f}(\boldsymbol{\theta})||^2\right), \quad (4)$$

$$= \frac{1}{(2\pi\sigma^2)^{K/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{K}(y_k - f_k(\boldsymbol{\theta}))^2\right). \quad (5)$$

Note that we have two types of variables of interest: the vector $\boldsymbol{\theta}$ contains the parameters of the nonlinear mapping $\mathbf{f}(\boldsymbol{\theta})$, whereas σ is a scale parameter of the likelihood function.

Goal. Given the vector of measurements $\mathbf{y}$, we desire to make infer regarding the hidden parameters $\boldsymbol{\theta}$ and the noise power σ^2, obtaining at least a point estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}^2$. Note also that the vector

$$\mathbf{r} = \mathbf{f}(\boldsymbol{\theta}) \in \mathbb{R}^K, \quad (6)$$

is a multivariate random variable obtained by the transformation of the random vector $\boldsymbol{\theta}$ trough the nonlinear mapping $\mathbf{f}$. Hence, an additional possible outcome is to obtain an smoothing version of the given observation vector $\mathbf{y} \in \mathbb{R}^K$, i.e., $\widehat{\mathbf{r}} \in \mathbb{R}^K$ (as well as uncertainty and correlation analysis between different y_k's). Finally, we are also interested in design efficient schemes in order to perform model selection, i.e., to compare, select or properly average different models.

Bayesian inference. We consider prior densities $p(\boldsymbol{\theta})$ and $p(\sigma)$ over the unknowns. Hence, the complete posterior density is

$$\bar{\pi}(\boldsymbol{\theta},\sigma|\mathbf{y}) = \frac{1}{Z(\mathbf{y})}\pi(\boldsymbol{\theta},\sigma|\mathbf{y}), \quad (7)$$

where $\pi(\boldsymbol{\theta},\sigma|\mathbf{y}) = Z(\mathbf{y}|\boldsymbol{\theta},\sigma)p(\boldsymbol{\theta})p(\sigma)$ and note that $\bar{\pi}(\boldsymbol{\theta},\sigma|\mathbf{y}) \propto \pi(\boldsymbol{\theta},\sigma|\mathbf{y})$. The marginal likelihood $Z(\mathbf{y})$ is

$$Z(\mathbf{y}) = \int_{\mathbb{R}^+}\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta},\sigma|\mathbf{y})d\boldsymbol{\theta}d\sigma, \quad (8)$$

This quantity is often needed for model selection. Since $Z(\mathbf{y})$ is generally unknown, we can usually evaluate pointwise the unnormalized posterior $\pi(\boldsymbol{\theta},\sigma|\mathbf{y})$. From now on, we remove the dependence on $\mathbf{y}$ to simplify the notation, using $\bar{\pi}(\boldsymbol{\theta},\sigma)$, $\bar{\pi}(\boldsymbol{\theta},\sigma)$, and Z. More generally, the computation of integrals of the form

$$I = \int_{\mathbb{R}^+}\int_{\boldsymbol{\Theta}} g(\boldsymbol{\theta},\sigma)\bar{\pi}(\boldsymbol{\theta},\sigma)d\boldsymbol{\theta}d\sigma, \quad (9)$$

where $g(\cdot)$ is an integrable function, is usually required. We consider a Monte Carlo quadrature approach for approximating the integral above and, more generally, provide a particle approximation of the posterior $\bar{\pi}(\boldsymbol{\theta},\sigma|\mathbf{y})$.

Problem. Generally, generating efficiently samples from a complicated posterior in Eq. (7) and computing efficiently the integrals as in Eqs. (8)-(9) is an hard task. Moreover, this task becomes often more difficult when we try to perform a joint inference where are involved scale parameters, i.e., σ, and parameters of the nonlinearity, i.e., $\boldsymbol{\theta}$. Indeed, "wrong choices" of σ values can easily jeopardize the sampling of $\boldsymbol{\theta}$. Below, we describe the strategy that we propose to tackle this problem.

3 Generic suggested approach

The idea underlying the proposed scheme is to split the space $[\boldsymbol{\theta}, \sigma]$, restricting the sampling problem only with respect to $\boldsymbol{\theta}$ and considering an optimization problem for with respect to σ. The proposed scheme described in the next section obtains the following three aims:

1. **MAP estimation.** We provide an approximation of the maximum a-posteriori estimator

$$[\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}, \widehat{\sigma}_{\texttt{MAP}}] = \arg\max_{\boldsymbol{\theta},\sigma} p(\boldsymbol{\theta}, \sigma|\mathbf{y}). \tag{10}$$

   More specifically, in the proposed algorithms that we describe below, we use a "coordinate ascent" approach (a.k.a., alternating optimization) considering the conditional posteriors, i.e.,

$$\widehat{\boldsymbol{\theta}}^{(t)}_{\texttt{MAP}} = \arg\max_{\sigma} p(\boldsymbol{\theta}|\widehat{\sigma}^{(t-1)}_{\texttt{MAP}}, \mathbf{y}), \tag{11}$$

$$\widehat{\sigma}^{(t)}_{\texttt{MAP}} = \arg\max_{\sigma} p(\sigma|\widehat{\boldsymbol{\theta}}^{(t)}_{\texttt{MAP}}, \mathbf{y}). \tag{12}$$

   where $t \in \mathbb{N}$ is the iteration index of the algorithm.

2. **Particle approximation.** We generate (weighted or unweighted) samples $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N)}$ from approximating the measure of the *conditional posterior* density

$$\bar{\pi}(\boldsymbol{\theta}|\widehat{\sigma}_{\texttt{MAP}}) = p(\boldsymbol{\theta}|\mathbf{y}, \widehat{\sigma}_{\texttt{MAP}}) \propto \pi(\boldsymbol{\theta}|\widehat{\sigma}_{\texttt{MAP}}) = p(\mathbf{y}|\boldsymbol{\theta}, \widehat{\sigma}_{\texttt{MAP}}) p(\boldsymbol{\theta}) p(\widehat{\sigma}_{\texttt{MAP}}). \tag{13}$$

3. **Marginal likelihood estimator.** We provide an estimator of the *conditional* marginal likelihood,

$$\widehat{Z} \approx p(\mathbf{y}|\widehat{\sigma}_{\texttt{MAP}}). \tag{14}$$

These three objectives are obtained by an iterative computational procedure. Thus, the resulting schemes are adaptive Monte Carlo algorithms which combines sampling schemes ad stochastic optimization. However, some part the conditional posterior of σ can be analytically maximized as shown below (jointly with some important considerations).

3.1 Analysis of the Conditional Posterior of σ^2

For the sake of simplicity, let us consider uniform improper priors $p(\sigma) \propto 1$ and $p(\boldsymbol{\theta}) \propto 1$. Note that the conditional marginal posterior with respect to σ^2 is

$$p(\sigma^2|\boldsymbol{\theta}, \mathbf{y}) \propto \frac{1}{\sigma^K} \exp\left(-\frac{||\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})||^2}{2\sigma^2}\right) \tag{15}$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{K}{2}} \exp\left(-\frac{||\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})||^2}{2\sigma^2}\right), \tag{16}$$

that has the form of an *Inverse Gamma* density,

$$p(\sigma^2|\boldsymbol{\theta}, \mathbf{y}) = \mathcal{IG}(\sigma^2|\alpha, \beta) \propto \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{\beta}{\sigma^2}\right), \tag{17}$$

where $\alpha = \frac{K}{2} - 1$ and $\beta = \frac{1}{2}||\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})||^2$. The Inverse Gamma density has expected value $\frac{\beta}{\alpha-1}$ (for $\alpha > 1$) and has a *unique* mode

$$\sigma^2_{\texttt{MAP}}|\boldsymbol{\theta} = \frac{\beta}{\alpha+1} = \frac{1}{K}||\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})||^2, \tag{18}$$

$$\sigma_{\texttt{MAP}}|\boldsymbol{\theta} = \sqrt{\frac{1}{K}||\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})||^2}, \tag{19}$$

where we have remarked that the expression above represents a MAP estimator *conditioned* to a specific value of $\boldsymbol{\theta}$. For $\alpha > 2$, the variance can be written as

$$\text{var}(\sigma^2|\boldsymbol{\theta}) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \leq \frac{\beta^2}{(\alpha-2)^3} \tag{20}$$

Note that $\alpha = \frac{K}{2} - 1$ depends only on the number of data K. If we fix the number of data K, we can see that

$$\text{var}(\sigma^2|\boldsymbol{\theta}) \propto \beta^2 = \frac{1}{K^2}||\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})||^4 = (\texttt{MSE-in-Obs-Space})^2.$$

The expression above shows that if Monte Carlo methods (MCMC or IS) is exploring a region where $\boldsymbol{\theta}$ does not provide a good fit with the data $\mathbf{y}$ (through the model $\mathbf{f}$) then the variance is proportional to the square of the MSE, i.e., it can be huge. This intuitively explains the issue of dealing with the joint sampling of $\boldsymbol{\theta}$ and σ.

4 Automatic Tempering Adaptive Importance Sampling (ATAIS)

In this section, we describe an adaptive importance sampler with an *automatic tempering* approach which follows the suggestions previously described. At each iteration t of the algorithm,

we have an approximation of the conditioned MAP of σ, i.e., $\widehat{\sigma}_{\texttt{MAP}}^{(t-1)}$. Let us define the *tempered conditional posterior* at the t-th iteration,

$$\pi_t(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\widehat{\sigma}_{\texttt{MAP}}^{(t-1)}) = \ell(\mathbf{y}|\boldsymbol{\theta}, \widehat{\sigma}_{\texttt{MAP}}^{(t-1)}) p(\boldsymbol{\theta}) p(\widehat{\sigma}_{\texttt{MAP}}^{(t-1)}). \tag{21}$$

At each iteration, we consider $\pi_t(\boldsymbol{\theta})$ as a target distribution. A set of N samples $\{\boldsymbol{\theta}_t^{(n)}\}_{n=1}^N$ are drawn from a proposal density $q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ with mean $\boldsymbol{\mu}_t$ and a covariance matrix $\boldsymbol{\Sigma}_t$. An importance weight $w_t^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}$ is assigned to each sample. A particle approximation of the conditional MAP estimator of $\boldsymbol{\theta}$ is given by $\widehat{\boldsymbol{\theta}}_t = \arg\max_n \pi_t(\boldsymbol{\theta}_t^{(n)})$. Then, the conditional MAP estimator of σ can be obtained analytically,

$$\widehat{\sigma}_t = \sqrt{\frac{1}{K}||\mathbf{y} - \widehat{\mathbf{r}}_t||^2}, \tag{22}$$

where $\widehat{\mathbf{r}}_t = \mathbf{f}(\widehat{\boldsymbol{\theta}}_t)$ (some alternatives are given in Table 2). The value of current MAP approximation $\pi_t(\widehat{\boldsymbol{\theta}}_t)$ is then compared with the global MAP estimator obtained so far denoted as $\pi_{\texttt{MAP}}^{(t-1)}$. If $\pi_t(\widehat{\boldsymbol{\theta}}_t) \geq \pi_{\texttt{MAP}}^{(t-1)}$, all the global MAP estimators are updated, i.e., we set

$$\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)} = \widehat{\boldsymbol{\theta}}_t, \quad \widehat{\sigma}_{\texttt{MAP}}^{(t)} = \widehat{\sigma}_t, \quad \pi_{\texttt{MAP}}^{(t)} = \pi_t(\widehat{\boldsymbol{\theta}}_t). \tag{23}$$

Otherwise, we keep the previous values $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)} = \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t-1)}$, $\widehat{\sigma}_{\texttt{MAP}}^{(t)} = \widehat{\sigma}_{\texttt{MAP}}^{(t-1)}$, and $\pi_{\texttt{MAP}}^{(t)} = \pi_{\texttt{MAP}}^{(t-1)}$. Finally, the parameters of the proposal density q are updated according to the global MAP estimator $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}$ and the empirical covariance of the weighted samples. Note that, we set $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}$ instead of using the empirical mean of the samples. This is due to we have notice that this choice provide better and more robust results, specially as the dimension of the problem grows. The ATAIS algorithm is completely described in Table 1. Table 2 contains further details. After T iterations, a final correction of the weights is needed, i.e.,

$$\widetilde{w}_t^{(n)} = w_t^{(n)} \frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})} = w_t^{(n)} \frac{\pi(\boldsymbol{\theta}_t^{(n)}|\widehat{\sigma}_{\texttt{MAP}}^{(T)})}{\pi(\boldsymbol{\theta}_t^{(n)}|\widehat{\sigma}_{\texttt{MAP}}^{(t-1)})}, \tag{24}$$

in order to obtain a particle approximation of the measure of the final conditional posterior $\bar{\pi}(\boldsymbol{\theta}|\widehat{\sigma}_{\texttt{MAP}}^{(T)})$.

5 Numerical Simulations

UNDER CONSTRUCTION

6 Conclusions

UNDER CONSTRUCTION

Table 1: ATAIS: AIS with automatic tempering

1. **Initializations:** Choose N, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$, and obtain an initialization for $\widehat{\sigma}_{\texttt{MAP}}$, $\pi_{\texttt{MAP}}$ (it can be done using a particle approximation at step $t = 0$).

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

   (a) **Sampling:**

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

   ii. Assign to each samples the weights

$$w_t^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)} = \frac{\pi(\boldsymbol{\theta}_t^{(n)}|\widehat{\sigma}_{\texttt{MAP}})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}, \qquad n = 1, ..., N. \tag{25}$$

   (b) **Current MAP estimation:**

   i. Obtain $\widehat{\boldsymbol{\theta}}_t = \arg\max_n \pi_t(\boldsymbol{\theta}_t^{(n)})$, and compute $\widehat{\mathbf{r}}_t = \mathbf{f}(\widehat{\boldsymbol{\theta}}_t)$ (for alternatives see Table 2).

   ii. Compute $\widehat{\sigma}_t = \sqrt{\frac{1}{K}||\mathbf{y} - \widehat{\mathbf{r}}_t||^2}$.

   (c) **Global MAP estimation:**

   i. If $\widehat{\sigma}_t \leq \widehat{\sigma}_{\texttt{MAP}}$, then set $\widehat{\sigma}_{\texttt{MAP}} = \widehat{\sigma}_t$.

   ii. If $\pi_t(\widehat{\boldsymbol{\theta}}_t) \geq \pi_{\texttt{MAP}}$, then set $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}} = \widehat{\boldsymbol{\theta}}_t$ and $\pi_{\texttt{MAP}} = \pi_t(\widehat{\boldsymbol{\theta}}_t)$.

   (d) **Adaptation:** Set

$$\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}, \tag{26}$$

$$\boldsymbol{\Sigma}_t = \sum_{n=1}^{N} \bar{w}_t^{(n)} (\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t)^\top (\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t) + \delta \mathbf{I}_M, \tag{27}$$

   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 MAP estimators, and all the weighted samples $\{\boldsymbol{\theta}_t^{(n)}, \widetilde{w}_t^{(n)}\}$ with the corrected weights

$$\widetilde{w}_t^{(n)} = w_t^{(n)} \frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})}. \tag{28}$$

Table 2: Possible model approximations

1. **MAP:** Given $\widehat{\boldsymbol{\theta}}_t = \arg\max_n \pi_t(\boldsymbol{\theta}_t^{(n)})$, then set

$$\widehat{\mathbf{r}}_t = \mathbf{f}(\widehat{\boldsymbol{\theta}}_t), \tag{29}$$

2. **MMSE:** Given $\bar{w}_t^{(n)} \frac{w_t^{(n)}}{\sum_{i=1}^N w_t^{(i)}}$ and $\bar{\boldsymbol{\theta}}_t = \sum_{n=1}^N \bar{w}_t^{(n)} \boldsymbol{\theta}_t^{(n)}$, then set

$$\widehat{\mathbf{r}}_t = \mathbf{f}(\bar{\boldsymbol{\theta}}_t), \tag{30}$$

3. **Fully-Bayesian solution:** Given $\bar{w}_t^{(n)} \frac{w_t^{(n)}}{\sum_{i=1}^N w_t^{(i)}}$, then set

$$\widehat{\mathbf{r}}_t = \sum_{n=1}^N \bar{w}_t^{(n)} \mathbf{f}(\boldsymbol{\theta}_t^{(n)}). \tag{31}$$