The Recycling Gibbs Sampler for Efficient Learning

Luca Martino[⋆], Victor Elvira[◇], Gustau Camps-Valls[⋆]

[⋆] Image Processing Laboratory, Universitat de València (Spain).

[◇] Department of Signal Processing, Universidad Carlos III de Madrid, Leganés (Spain).

Abstract

Monte Carlo methods are essential tools for Bayesian inference. Gibbs sampling is a well-known Markov chain Monte Carlo (MCMC) algorithm, extensively used in signal processing, machine learning, and statistics, employed to draw samples from complicated high-dimensional posterior distributions. The key point for the successful application of the Gibbs sampler is the ability to draw efficiently samples from the full-conditional probability density functions. Since in the general case this is not possible, in order to speed up the convergence of the chain, it is required to generate auxiliary samples whose information is eventually disregarded. In this work, we show that these auxiliary samples can be recycled within the Gibbs estimators, improving their efficiency with no extra cost. This novel scheme arises naturally after pointing out the relationship between the standard Gibbs sampler and the chain rule used for sampling purposes. Numerical simulations involving simple and real inference problems confirm the excellent performance of the proposed scheme in terms of accuracy and computational efficiency. In particular we give empirical evidence of performance in a toy example, inference of Gaussian processes hyperparameters, and learning dependence graphs through regression.

Keywords: Bayesian inference, Markov Chain Monte Carlo (MCMC), Gibbs sampling, Gaussian Processes (GP), automatic relevance determination (ARD).

1 Introduction

'Reduce, Reuse, Recycle'
The Greenpeace motto

Monte Carlo algorithms have become very popular over the last decades (Liu, 2004; Robert and Casella, 2004). Many practical problems in statistical signal processing, machine learning and statistics, demand fast and accurate procedures for drawing samples from probability distributions that exhibit arbitrary, non-standard forms (Andrieu et al., 2003; Fitzgerald, 2001), (Bishop, 2006, Chapter 11). One of the most popular Monte Carlo methods are the families of Markov chain Monte Carlo (MCMC) algorithms (Andrieu et al., 2003; Robert and Casella, 2004) and particle filters (Bugallo et al., 2007; Djurić et al., 2003). The MCMC techniques generate a Markov chain

(i.e., a sequence of correlated samples) with a pre-established target probability density function (pdf) as invariant density (Liu, 2004; Liang et al., 2010).

The Gibbs sampling technique is a well-known MCMC algorithm, extensively used in the literature in order to generate samples from multivariate target densities, drawing each component of the samples from the full-conditional densities (Chen et al., 2016; Koch, 2007; Kotecha and Djurić, 1999; Goudie and Mukherjee, 2016; Lucka, 2016; Zhang et al., 2016).[1] In order to draw samples from a multivariate target distribution, the key point for the successful application of the standard Gibbs sampler is the ability to draw efficiently from the univariate conditional pdfs (Liu, 2004; Robert and Casella, 2004). The best scenario for Gibbs sampling occurs when specific direct samplers are available for each full-conditional, e.g. inversion method or, more generally, some transformation of a random variable (Devroye, 1986; Robert and Casella, 2004). Otherwise, other Monte Carlo techniques, such as rejection sampling (RS) and different flavors of the Metropolis-Hastings (MH) algorithms, are typically used *within* the Gibbs sampler to draw from the complicated full-conditionals. The performance of the resulting Gibbs sampler depends on the employed *internal* technique, as pointed out for instance in (Cai et al., 2008; Gilks et al., 1995; Martino et al., 2015a,b).

In this context, some authors have suggested to use more steps of the MH method within the Gibbs sampler (Müller, 1991; Gelfand and Lee, 1993; Fox, 2012). Moreover, other different algorithms have been proposed as alternatives to the MH technique (Cai et al., 2008; Koch, 2007; Shao et al., 2013). For instance, several automatic and self-tuning samplers have been designed to be used primarily *within-Gibbs*: the adaptive rejection sampling (ARS) (Gilks, 1992; Gilks and Wild, 1992), the griddy Gibbs sampler (Ritter and Tanner, 1992), the FUSS sampler (Martino et al., 2015b), the Adaptive Rejection Metropolis Sampling (ARMS) method (Gilks et al., 1995, 1997; Meyer et al., 2008; Zhang et al., 2016), and the Independent Doubly Adaptive Rejection Metropolis Sampling (IA^{2}RMS) technique (Martino et al., 2015a), just to name a few.

Most of the previous solutions require performing several MCMC steps for each full-conditional in order to improve the performance, although only one of them is considered to produce the resulting Markov chain because the rest of samples play the mere role of auxiliary variables. Strikingly, they require an increase in the computational cost that is not completely paid off: several samples are drawn from the full-conditionals, but only a subset of these generated samples is employed in the final estimators. In this work, we show that the rest of generated samples can be directly incorporated within the corresponding Gibbs estimator. We call this approach the *Recycling Gibbs (RG) sampler* since *all* the samples drawn from each full-conditional can be used also to provide a better estimation, instead of discarding them.

The consistency of the proposed RG estimators is guaranteed, as will be noted after considering the connection between the Gibbs scheme and the chain rule for sampling purposes (Devroye, 1986;

[1] In general these full-conditionals are univariate. Nevertheless, block-wise Gibbs sampling approaches where several random variables are updated simultaneously, have been proposed to speed up the convergence of the Gibbs sampler (Roberts and Sahu, 1997). However, unless direct sampling from the multi-variate full-conditionals is feasible, these approaches still result in an increased difficulty of drawing samples and a higher computational cost per iteration. Furthermore, the performance of the overall algorithm can decrease if the blocks are not properly chosen, especially when direct sampling from the multivariate full-conditionals is unfeasible (Liu, 2004; Liang et al., 2010; Robert and Casella, 2004). The novel recycling scheme can also be used in the block approach.

Robert and Casella, 2004). In particular, we show that the standard Gibbs approach is equivalent (after the burn-in period) to the standard chain rule, whereas RG is equivalent to an alternative version of the chain rule presented in this work as well. RG fits particularly well combined with adaptive MCMC schemes where different internal steps are performed also for adapting the proposal density, see e.g. (Gilks et al., 1995; Martino et al., 2015a; Meyer et al., 2008; Zhang et al., 2016). The novel RG scheme allows us to obtain better performance without adding any extra computational cost. This will be shown through intensive numerical simulations. First, we test RG in a simple toy example with a bimodal bivariate target. We also include experiments for hyper-parameter estimation in Gaussian Processes (GPs) regression problems with the so-called *automatic relevance determination* (ARD) kernel function (Bishop, 2006). Finally, we apply the novel scheme in real-life geoscience problems of dependence estimation among bio-geo-physical variables from satellite sensory data. The MATLAB code of the numerical examples is provided at `http://isp.uv.es/code/RG.zip`.

The remainder of the paper is organized as follows. Section 2 fixes notation and recalls the problem statement of Bayesian inference. The standard Gibbs sampler and the chain rule for sampling purposes are summarized in Section 3, highlighting their connections. In the same section, we then introduce an alternative chain rule approach, which is useful for describing the novel scheme. The RG technique proposed here is formalized in Section 4. Sections 5 provides empirical evidence of the benefits of the proposed scheme, considering different multivariate posterior distributions. Finally, Section 6 concludes and outlines further work.

2 Bayesian inference

Machine learning, statistics, and signal processing often face the problem of inference through density sampling of potentially complex multivariate distributions. In particular, Bayesian inference is concerned about doing inference about a variable of interest exploiting the Bayes' theorem to update the probability estimates according to the available information. Specifically, in many applications, the goal is to infer a variable of interest, $\mathbf{x} = [x_1, \ldots, x_D] \in \mathbb{R}^D$, given a set of observations or measurements, $\mathbf{y} \in \mathbb{R}^P$. In Bayesian inference all the statistical information is summarized by means of the posterior pdf, i.e.,

$$\bar{\pi}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})}{Z(\mathbf{y})}, \tag{1}$$

where $\ell(\mathbf{y}|\mathbf{x})$ is the likelihood function, $g(\mathbf{x})$ is the prior pdf and $Z(\mathbf{y})$ is the marginal likelihood (a.k.a., Bayesian evidence). In general, $Z(\mathbf{y})$ is unknown and difficult to estimate in general, so we assume to be able to evaluate the unnormalized target function,

$$\pi(\mathbf{x}) = \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x}). \tag{2}$$

The analytical study of the posterior density $\bar{\pi}(\mathbf{x}) \propto \pi(\mathbf{x})$ is often unfeasible and integrals involving $\bar{\pi}(\mathbf{x})$ are typically intractable. For instance, one might be interested in the estimation of

$$I = \int_{\mathbb{R}^D} f(\mathbf{x})\bar{\pi}(\mathbf{x})d\mathbf{x}, \tag{3}$$

where $f(\mathbf{x})$ is a squared integrable function (with respect to $\bar{\pi}$). In order to compute the integral I numerical approximations are typically required. Our goal here is to approximate this integral by using Monte Carlo (MC) quadrature (Liu, 2004; Robert and Casella, 2004). Namely, considering T independent samples from the posterior target pdf, i.e., $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)} \sim \bar{\pi}(\mathbf{x})$, we can write

$$\widehat{I}_T = \frac{1}{T} \sum_{t=1}^{T} f(\mathbf{x}^{(t)}) \xrightarrow{p} I. \tag{4}$$

This means that for the weak law of large numbers, $\widehat{I}_T$ converges in probability to I: that is, for any positive number $\epsilon > 0$, we have $\lim_{T\to\infty} \Pr(|\widehat{I}_T - I| > \epsilon) = 0$. In general, a direct method for drawing independent samples from $\bar{\pi}(\mathbf{x})$ is not available, and alternative approaches, e.g., MCMC algorithms, are needed. An MCMC method generates an ergodic Markov chain with invariant density $\bar{\pi}(\mathbf{x})$ (a.k.a., stationary pdf). Even though, the generated samples $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}\}$ are then correlated in this case, $\widehat{I}_T$ is still a consistent estimator.

Within the MCMC framework, we can consider a block approach working directly into the D-dimensional space, e.g., using a Metropolis-Hastings (MH) algorithm (Robert and Casella, 2004), or a component-wise approach (Haario et al., 2005; Johnson et al., 2013; Levine et al., 2005) working iteratively in different uni-dimensional slices of the entire space, e.g., using a Gibbs sampler (Liu, 2004; Liang et al., 2010).[2] In many applications, and for different reasons, the component-wise approach is the preferred choice. For instance, this is the case when the full-conditional distributions are directly provided or when the probability of accepting a new state with a complete block approach becomes negligible as the dimension of the problem D increases. In the following section, we outline the standard Gibbs approach, and remark its connection with the chain rule method. The main notation and acronyms of the work are summarized in Table 1.

3 Gibbs sampling and the chain rule method

This section reviews the fundamentals about the standard Gibbs sampler, reviews the recent literature on Gibbs sampling when complicated full-conditional pdfs are involved, and points out the connection between GS and the chain rule. A variant of the chain rule is also described, which is related to the novel scheme introduced in the next section.

3.1 The Standard Gibbs (SG) sampler

The Gibbs sampler is perhaps the most widely used algorithm for inference in statistics and machine learning (Chen et al., 2016; Koch, 2007; Goudie and Mukherjee, 2016; Robert and Casella, 2004). Let us define $\mathbf{x}_{\neg d} := [x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_D]$ and introduce the following equivalent notations

$$\bar{\pi}_d(x_d|x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_D) = \bar{\pi}_d(x_d|x_{1:d-1}, x_{d+1:D}) = \bar{\pi}_d(x_d|\mathbf{x}_{\neg d}).$$

[2]There also exist intermediate strategies where the same subset of variables are jointly updated, which is often called the Blocked Gibbs Sampler.

Table 1: Main notation and acronyms of the work.

D	Dimension of the inference problem, $\mathbf{x} \in \mathbb{R}^D$.
T	Total number of iterations of the Gibbs scheme.
M	Total number of iterations of the MCMC method inside the Gibbs scheme.
t_b	Length of the burn-in period.
$\mathbf{x}$	Variable of interest; parameters to be inferred, $\mathbf{x} = [x_1, \ldots, x_D]$.
$\mathbf{y}$	Collected data: observations or measurements.
$\mathbf{x}_{\neg d}$	$\mathbf{x}$ without the d-th component, i.e., $[x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_D]$.
$x_{a:b}$	The vector $x_{a:b} = [x_a, x_{a+1}, x_{a+2}, \ldots, x_b]$ with $b > a > 0$.
$\bar{\pi}(\mathbf{x})$	Normalized posterior pdf $\bar{\pi}(\mathbf{x}) = p(\mathbf{x}\|\mathbf{y})$.
$\pi(\mathbf{x})$	Posterior function proportional to the posterior pdf, $\bar{\pi}(\mathbf{x}) \propto \pi(\mathbf{x})$.
$\bar{\pi}_d(x_d\|\mathbf{x}_{\neg d})$	d-th full-conditional pdf.
$p_d(x_d)$	d-th marginal pdf.
SG	Standard Gibbs.
TRG	Trivial Recycling Gibbs.
MH	Metropolis-Hastings.
MRG	Multiple Recycling Gibbs.

In order to denote the unidimensional full-conditional pdf of the component $x_d \in \mathbb{R}$, $d \in \{1, \ldots, D\}$, given the rest of variables $\mathbf{x}_{\neg d}$, i.e.

$$\bar{\pi}_d(x_d|\mathbf{x}_{\neg d}) = \frac{\bar{\pi}(\mathbf{x})}{\bar{\pi}_{\neg d}(\mathbf{x}_{\neg d})} = \frac{\bar{\pi}(\mathbf{x})}{\int_{\mathbb{R}} \bar{\pi}(\mathbf{x}) dx_d}. \tag{5}$$

The density $\bar{\pi}_{\neg d}(\mathbf{x}_{\neg d}) = \int_{\mathbb{R}} \bar{\pi}(\mathbf{x}) dx_d$ is the joint pdf of all variables but x_d. The Gibbs algorithm generates a sequence of T samples, and is formed by the steps in Algorithm 1. Note that the main assumption for the application of Gibbs sampling is being able to draw efficiently from these univariate full-conditional pdfs $\bar{\pi}_d$. However, in general, we are not able to draw directly from any arbitrary full-conditional pdf. Thus, other Monte Carlo techniques are needed for drawing from all the $\bar{\pi}_d$.

Algorithm 1 The Standard Gibbs (SG) algorithm.

1: Fix T, D
2: **for** $t = 1, \ldots, T$ **do**
3: **for** $d = 1, \ldots, D$ **do**
4: Draw $x_d^{(t)} \sim \bar{\pi}_d(x_d|x_{1:d-1}^{(t)}, x_{d+1:D}^{(t-1)})$.
5: **end for**
6: Set $\mathbf{x}^{(t)} = [x_1^{(t)}, x_2^{(t)}, \ldots, x_D^{(t)}]$.
7: **end for**

3.2 Monte Carlo-within-Gibbs sampling schemes

In many cases, drawing directly from the full-conditional pdf is not possible, hence the use of another Monte Carlo scheme is needed. Figure 1 summarizes the main techniques proposed in literature for this purpose. In some specific situations, rejection samplers (Caffo et al., 2002; Hörmann, 2002; Hörmann et al., 2007; Marrelec and Benali, 2004; Tanizaki, 1999) and their adaptive version, as the *adaptive rejection sampler* (ARS) Gilks and Wild (1992), are employed to generate one sample from each $\bar{\pi}_d$ per iteration. Since the standard ARS can be applied only to lo-concave densities, several extensions have been introduced Hörmann (1995); Görür and Teh (2011); Martino and Míguez (2011). The ARS algorithms are very appealing techniques since they construct a non-parametric proposal to mimic the shape of the target pdf, yielding in general excellent performance (i.e., independent samples from $\bar{\pi}_d$ with a high acceptance rate).

Monte Carlo schemes used within-Gibbs

IS	ARS	MCMC		
		stand. MH	**automatic**	**adaptive**
(Koch, 2007)	(Gilks and Wild, 1992) (Hörmann, 1995a) (Evans and Swartz, 1998) (Görür and Teh, 2011) (Martino and Míguez, 2011)	(Gelfand and Lee, 1993)	(Ritter and Tanner, 1992) (Shao et al., 2013) (Martino et al., 2015b)	(Gilks et al., 1995) (Haario et al., 2005) (Levine et al., 2005) (Meyer et al., 2008) (Cai et al., 2008) (Martino et al., 2015a)

Figure 1: Summary of the main Monte Carlo algorithms which have been employed within the Gibbs sampling technique.

However, the range of application of the ARS samplers is limited to some specific classes of densities. Thus, in general, other approaches are required. For instance in (Koch, 2007) an approximated strategy is used, considering the application of the importance sampling (IS) scheme within the Gibbs sampler. A more common approach is to apply an additional MCMC sampler to draw samples from $\bar{\pi}_d$ (Gelfand and Lee, 1993). Therefore, in many practical scenarios, we have an MCMC (e.g., an MH sampler) inside another MCMC scheme (i.e., the Gibbs sampler) as shown in Figures 1-2. In the so-called *MH-within-Gibbs* approach[3], only one MH step is often performed within each Gibbs iteration to draw samples from each full-conditional. This hybrid approach preserves the ergodicity of the Gibbs sampler (Robert and Casella, 2004, Chapter 10), and provides good performance in many cases. However, several authors have noted that using a single MH step for the internal MCMC is not always the best solution in terms of performance, c.f. (Brewer and Aitken, 1993).

Using a larger number of iterations of the MH algorithm within-Gibbs can improve the performance (Müller, 1991; Gelfand and Lee, 1993; Fox, 2012). This is the scenario graphically represented in Figure 2. Moreover, different more sophisticated MCMC algorithms to be applied within-Gibbs have been proposed (Gilks et al., 1995; Cai et al., 2008; Haario et al., 2005; Ritter and Tanner, 1992). Some of these techniques employ an automatic construction of the proposal density

[3]Sometimes MH-within-Gibbs is also referred as to the Single Component MH algorithm (Haario et al., 2005) or the Componentwise MH algorithm (Levine et al., 2005).

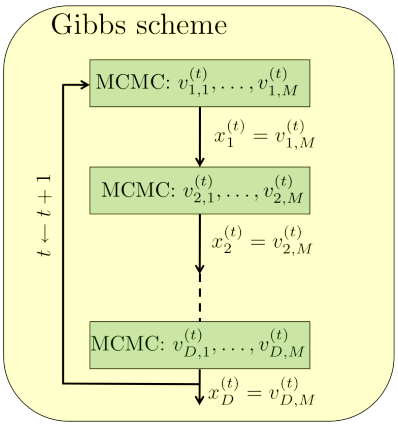


Figure 2: Graphical representation of a generic MCMC-within-Gibbs scheme, where M steps of the internal MCMC algorithm are applied for each full-conditional pdf (see Algorithm 2). Different internal MCMC methods have been proposed in literature.

tailored to the specific full-conditional (Ritter and Tanner, 1992; Shao et al., 2013; Martino et al., 2015b). Other methods use an adaptive parametric proposal pdf (Haario et al., 2005; Levine et al., 2005), while other ones employ adaptive non-parametric proposals (Gilks et al., 1995; Meyer et al., 2008; Martino et al., 2015a), in the same fashion of the ARS schemes. It is important to remark here that performing more steps of an adaptive MH method within a Gibbs sampler can provide better results than a longer Gibbs chain applying only one step of a standard MH method (Gilks et al., 1995). Algorithm 2 describes a generic MCMC-within-Gibbs sampler considering M steps of the internal MCMC at each Gibbs iteration.

While these algorithms are specifically designed to be applied "within-Gibbs" and provide very good performance, they still require an increase in the computational cost that is not completely exploited: several samples are drawn from the full-conditionals and used to adapt the proposal pdf, but only a subset of them is employed within the resulting Gibbs estimator. In this work, we show how they can be incorporated within the corresponding Gibbs estimator to improve performance without jeopardizing its consistency. In the following, we show the relationships between the standard Gibbs scheme and the chain rule. Then, we describe an alternative formulation of the chain rule useful for introducing the novel Gibbs approach described in Section 4.

3.3 Chain rule and the connection with Gibbs sampling

Let us highlight an important consideration for the derivation of the novel Gibbs approach we will introduce in the following section. For the sake of simplicity, let us consider a bivariate target pdf that can be factorized according to the chain rule,

$$\begin{aligned} \bar{\pi}(x_1, x_2) &= \bar{\pi}_2(x_2|x_1)p_1(x_1) \\ &= \bar{\pi}_1(x_1|x_2)p_2(x_2), \end{aligned}$$

where we have denoted with p_1, p_2, the marginal pdfs of x_1 and $\bar{\pi}_2$, $\bar{\pi}_1$, are the conditional pdfs. Let us consider the first equality. Clearly, if we are able to draw from the marginal pdf $p_1(x_1)$ and

Algorithm 2 Generic MCMC-within-SG sampler.

Choose $[x_1^{(0)}, \ldots, x_D^{(0)}]$.
for $t = 1, \ldots, T$ **do**
  for $d = 1, \ldots, D$ **do**
    Perform M steps of an MCMC algorithm with initial state $v_{d,0}^{(t)} = x_d^{(t-1)}$, and target pdf $\bar{\pi}_d(x_d | x_{1:d-1}^{(t)}, x_{d+1:D}^{(t-1)})$, yielding the sequence of samples $v_{d,1}^{(t)}, \ldots, v_{d,M}^{(t)}$.
    Set $x_d^{(t)} = v_{d,M}^{(t)}$.
  end for
  Set $\mathbf{x}^{(t)} = x_{1:D}^{(t)} = [x_1^{(t)}, x_2^{(t)}, \ldots, x_D^{(t)}]$.
end for
Return $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}\}$

from the conditional pdf $\bar{\pi}_2(x_2|x_1)$, we can draw samples from $\bar{\pi}(x_1, x_2)$ following the chain rule procedure in Algorithm 3. Note that, consequently, the T independent random vectors $[x_1^{(t)}, x_2^{(t)}]$, with $t = 1, \ldots, T$, are all distributed as $\bar{\pi}(x_1, x_2)$.

Algorithm 3 Chain rule method

for $t = 1, \ldots, T$ **do**
  Draw $x_1^{(t)} \sim p_1(x_1)$ and $x_2^{(t)} \sim \bar{\pi}_2(x_2|x_1^{(t)})$
end for

3.3.1 Standard Gibbs sampler as the chain rule

Let us consider again the previous bivariate case where the target pdf is factorized as $\bar{\pi}(\mathbf{x}) = \bar{\pi}(x_1, x_2)$. The standard Gibbs sampler in this bivariate case consists of the steps in Algorithm 4. After the burn-in period, the chain converges to the target pdf, i.e., $\mathbf{x}^{(t)} \sim \bar{\pi}(\mathbf{x})$. Therefore, recalling that $\bar{\pi}(x_1, x_2) = \bar{\pi}_2(x_2|x_1)p_1(x_1) = \bar{\pi}_1(x_1|x_2)p_2(x_2)$ for $t \geq t_b$, each component of the vector $\mathbf{x}^{(t)} = [x_1^{(t)}, x_2^{(t)}]$ is distributed as the corresponding marginal pdf, i.e., $x_1^{(t)} \sim p_1(x_1)$ and $x_2^{(t)} \sim p_2(x_2)$. Therefore, after t_b iterations, the standard Gibbs sampler can be interpreted as the application of the chain rule procedure in Algorithm 3. Namely, for $t \geq t_b$, Algorithm 4 is equivalent to generate $x_1^{(t)} \sim p_1(x_1)$, and then draw $x_2^{(t)} \sim \bar{\pi}_1(x_2|x_1^{(t)})$.

Algorithm 4 The standard Gibbs sampler for a bivariate target pdf.

for $t = 1, \ldots, T$ **do**
  Draw $x_2^{(t)} \sim \bar{\pi}_1(x_2|x_1^{(t-1)})$.
  Draw $x_1^{(t)} \sim \bar{\pi}_2(x_1|x_2^{(t)})$.
  Set $\mathbf{x}^{(t)} = [x_1^{(t)}, x_2^{(t)}]$.
end for

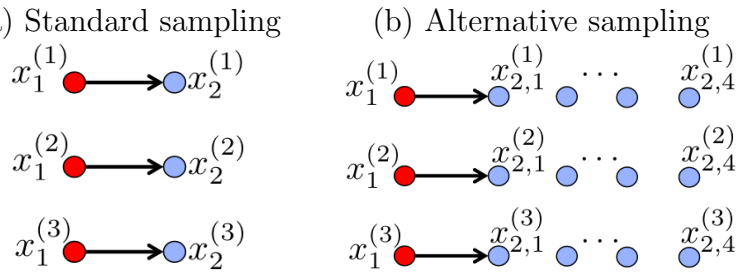


Figure 3: Graphical representation of the (a) standard chain rule sampling ($M = 1$), and (b) the alternative chain rule sampling ($M = 4$). In both cases, $N = 3$. The total number of drawn vectors $[x_1^{(t)}, x_{2,m}^{(t)}] \sim \bar{\pi}(x_1, x_2) = \bar{\pi}_2(x_2|x_1)p_1(x_1)$ is $NM = 3$ and $NM = 12$, respectively.

3.3.2 Alternative chain rule procedure

An alternative procedure is shown in Algorithm 5. This chain rule draws M samples from the full conditional $\bar{\pi}_2(x_2|x_1)$ at each t-th iteration, and generates samples from the joint pdf $\bar{\pi}(x_1, x_2)$.

Algorithm 5 An alternative chain rule procedure.

1: **for** $t = 1, \ldots, T$ **do**
2: Draw $x_1^{(t)} \sim p_1(x_1)$.
3: Draw $x_{2,m}^{(t)} \sim \bar{\pi}_2(x_2|x_1^{(t)})$, with $m = 1, \ldots .M$.
4: **end for**

Note that all the TM vectors, $[x_1^{(t)}, x_{2,m}^{(t)}]$, with $t = 1, \ldots, T$ and $m = 1, \ldots, M$, are samples from $\bar{\pi}(x_1, x_2)$. This scheme is valid and, in some cases, can present some benefits w.r.t. the traditional scheme in terms of performance, depending on some characteristics contained in the joint pdf $\bar{\pi}(x_1, x_2)$. For instance, the correlation between variables x_1 and x_2, and the variances of the marginal pdfs $p_1(x_1)$ and $p_2(x_2)$. Figure 3 shows the graphical representation of the standard chain rule sampling scheme (with $T = 3$ and $M = 1$), and the alternative chain rule sampling procedure described before (with $T = 3$, $M = 4$).

At this point, a natural question arises: is it possible to design a Gibbs sampling scheme equivalent to the alternative chain rule scheme described before? In the next section, we introduce the Multiple Recycling Gibbs Sampler (MRG), which corresponds to the alternative chain rule procedure, as summarized in Fig. 4.

4 The Recycling Gibbs sampler

The previous considerations suggest that we can benefit from some previous intermediate points produced in the Gibbs procedure. More specifically, let us consider the following *Trivial Recycling Gibbs* (TRG) procedure in Algorithm 6.

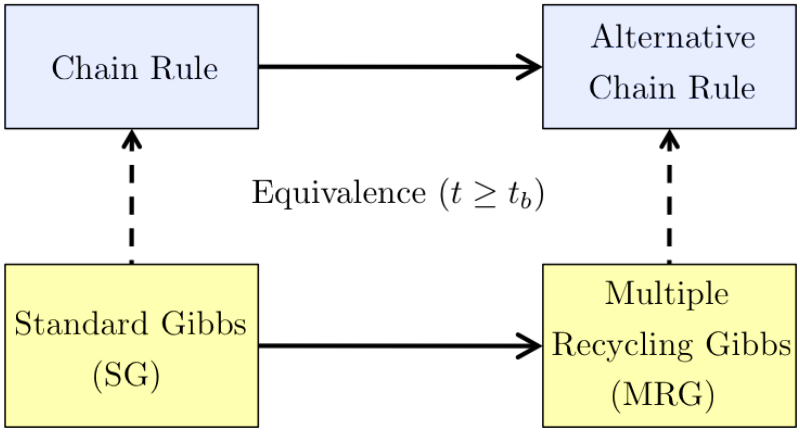


Figure 4: Graphical representation of the relationships between chain rules and Gibbs schemes.

Algorithm 6 Trivial Recycling Gibbs (TRG) procedure.

for $t = 1, \dots, T$ **do**
  for $d = 1, \dots, D$ **do**
    Draw $x_d^{(t)} \sim \bar{\pi}_d(x_d | x_{1:d-1}^{(t)}, x_{d+1:D}^{(t-1)})$.
    Set $\mathbf{x}_d^{(t)} = [x_{1:d-1}^{(t)}, x_d^{(t)}, x_{d+1:D}^{(t-1)}] = [x_{1:d}^{(t)}, x_{d+1:D}^{(t-1)}]$.
  end for
end for
return Return $\{\mathbf{x}_d^{(t)}\}$ for all d and t.

The procedure generates DT samples $\mathbf{x}_d^{(t)}$, with $d = 1, \dots, D$ and $t = 1, \dots, T$, shown in Figure 5(b) with circles and squares. Note that if we consider only the subset of generated vectors

$$\mathbf{x}_D^{(t)}, \quad t = 1, \dots, T,$$

by setting $d = D$, we obtain the outputs of the standard Gibbs (SG) sampler approach in Algorithm 1. Namely, the samples generated by a SG procedure can be obtained by subsampling the samples obtained by the proposed RG. Figure 5(a) depicts with circles $T + 1$ vectors (considering also the starting point) corresponding to a run of SG with $T = 4$. Figure 5(b) shows with squares the additional points used in TRG.

Let us consider the estimation by SG and TRG of a generic moment, i.e., given a function $f(x_d)$, of d-th marginal density, i.e., $p_d(x_d) = \int_{\mathbb{R}^{D-1}} \bar{\pi}(\mathbf{x}) d\mathbf{x}_{\neg d}$. After a closer inspection, we note that both estimators corresponding to the SG and TRG coincide:

$$\int_{\mathbb{R}^D} f(x_d) \bar{\pi}(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}} f(x_d) p_d(x_d) dx_d \approx \frac{1}{DT} \sum_{t=1}^{T} \sum_{d=1}^{D} f(x_d^{(t)}) = \frac{1}{T} \sum_{t=1}^{T} f(x_D^{(t)}), \tag{6}$$

where for the last equality we are assuming (for the sake of simplicity) that the d-th component is the second variable in the Gibbs scan and $T = kD$, $k \in \mathbb{N}$. This is due to the fact that, in

TRG, each component $x_d^{(t)}$ is repeated exactly D times (inside different consecutive samples) and we have D times more samples in TRG than in a standard SG. Hence, in such situation, there are no apparent advantages of using TRG w.r.t. a SG approach. Namely, TRG and SG are equivalent schemes in the approximation of the marginal densities. The advantages of a RG strategy appear clear when more than one sample is drawn from the full-conditional, $M > 1$, as discussed below.

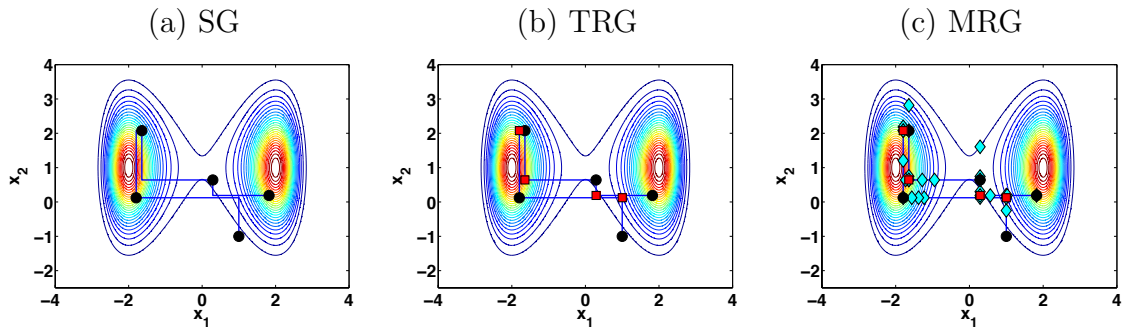


Figure 5: We consider $T = 4$ iterations of a Gibbs sampler and $M = 5$ iterations of the MH for drawing from each full-conditional pdfs. **(a)** With the circles we denote the $T + 1$ points (considering the starting point) used in the standard Gibbs estimators. **(b)** The vectors (denoted with circles and squares) used in the TRG estimators. **(c)** The vectors (denoted with circles, squares and diamonds) used in the MRG estimators.

Based on the previous considerations, we design the *Multiple Recycling Gibbs* (MRG) sampler which draws $M > 1$ samples from each full conditional pdf, as shown in Algorithm 7. Figure 5(c) shows all the samples (denoted with circles, squares and diamonds) used in the MRG estimators. Thus, given a specific function $f(\mathbf{x})$ in the integral in Eq. (3), the MRG estimator is eventually formed by TDM samples, without removing any burn-in period,

$$\widehat{I}_T = \frac{1}{TDM} \sum_{t=1}^{T} \sum_{d=1}^{D} \sum_{m=1}^{M} f(\mathbf{x}_{d,m}^{(t)}). \tag{7}$$

Observe that in order to go forward to sampling from the next full-conditional, we only consider the last generated component, i.e., $z_d^{(t)} = x_{d,M}^{(t)}$. However, an alternative to step 8 of Algorithm 7 is: (a) draw $j \sim \mathcal{U}(1, \ldots, M)$ and (b) set $z_d^{(t)} = x_{d,j}^{(t)}$. Note that choosing the last sample $x_{d,M}^{(t)}$ is more convenient for an MCMC-within-MRG scheme.

As shown in Figure 4, MRG is equivalent to the alternative chain rule scheme described in the previous section, so that the consistency of the MRG estimators is guaranteed. The ergodicity of the generated chain is also ensured since the dynamics of the MRG scheme is identical to the dynamics of the SG sampler (they differ in the construction of final estimators). Note that with $M = 1$, we go back to the TRG scheme.

The MRG approach is convenient in terms of accuracy and computational efficiency, as also confirmed by the numerical results in Section 5. MRG is particularly advisable if an adaptive MCMC is employed to draw from the full-conditional pdfs, i.e., when several MCMC steps are

Algorithm 7 Multiple Recycling Gibbs (MRG) sampler.

Choose a starting point $[z_1^{(0)}, \ldots, z_D^{(0)}]$.
for $t = 1, \ldots, T$ **do**
  for $d = 1, \ldots, D$ **do**
    for $m = 1, \ldots, M$ **do**
      Draw $x_{d,m}^{(t)} \sim \bar{\pi}_d(x_d | z_{1:d-1}^{(t)}, z_{d+1:D}^{(t-1)})$.
      Set $\mathbf{x}_{d,m}^{(t)} = [z_{1:d-1}^{(t)}, x_{d,m}^{(t)}, z_{d+1:D}^{(t-1)}]$.
    end for
    Set $z_d^{(t)} = x_{d,M}^{(t)}$.
  end for
end for
return Return $\{\mathbf{x}_{d,m}^{(t)}\}$ for all d, m and t.

performed for sampling from each full-conditional and adapting the proposal. We can use all the sequence of samples generated by the internal MCMC algorithm in the resulting estimator. Algorithm 8 shows the detailed steps of an MCMC-within-MRG algorithm, when a direct method for sampling the full-conditionals is not available.

Algorithm 8 Generic MCMC-within-MRG sampler.

Choose a starting point $[z_1^{(0)}, \ldots, z_D^{(0)}]$.
for $t = 1, \ldots, T$ **do**
  for $d = 1, \ldots, D$ **do**
    Perform M steps of an MCMC algorithm with target pdf $\bar{\pi}_d(x_d | x_{1:d-1}^{(t)}, x_{d+1:D}^{(t-1)})$, yielding the sequence of samples $x_{d,1}^{(t)}, \ldots, x_{d,M}^{(t)}$, with initial state $x_{d,0}^{(t)} = x_d^{(t-1)}$.
    Set $\mathbf{x}_{d,m}^{(t)} = [z_{1:d-1}^{(t)}, x_{d,m}^{(t)}, z_{d+1:D}^{(t-1)}]$, for $m = 1, \ldots, M$.
    Set $z_d^{(t)} = x_{d,M}^{(t)}$.
  end for
end for
return Return $\{\mathbf{x}_{d,m}^{(t)}\}$ for all d, m and t.

Figure 6(a) depicts the random vectors obtained with one run of an MH-within-Gibbs procedure, with $T = 10^3$ and $M = 5$. Figure 6(b) illustrates all the outputs of the previous run, including all the auxiliary samples generated by the MH algorithm. Hence, these vectors are the samples obtained with a MH-within-MRG approach. The histogram of the samples in Figure 6(b) is depicted Figure 6(c). Note that the histogram of the MH-within-MRG samples reproduces adequately the shape of the target pdf shown in Figure 5. This histogram was obtained with one run of MH-within-MRG fixing $T = 10^4$ and $M = 5$.

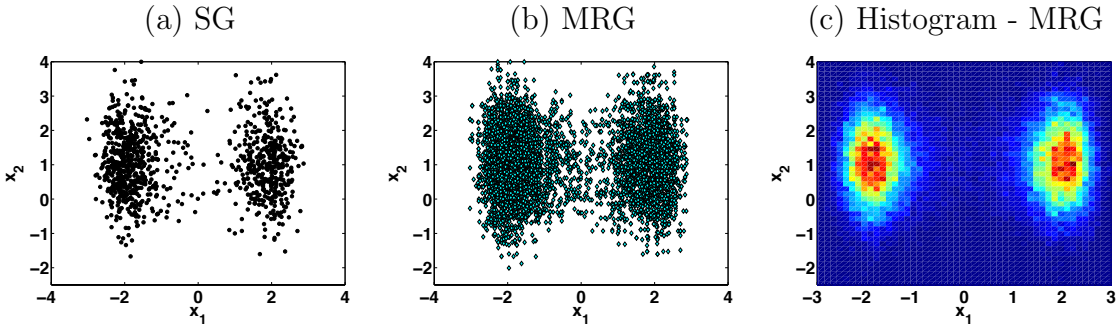


Figure 6: **(a)** Outputs of one MH-within-Gibbs run with $T = 10^3$ and $M = 5$, considering the target with contour plot shown in Fig. 5. **(b)** Outputs of one MH-within-MRG run with $T = 10^3$ and $M = 5$. **(c)** Histograms obtained using all the points in Figure (b), i.e., the MRG outputs with $T = 10^4$ and $M = 5$.

5 Experimental Results

This section gives experimental evidence of performance of the proposed scheme. First we study the efficiency of the proposed scheme in a toy example. Then, we show its use in a hyper-parameter estimation problem using Gaussian process (GP) regression (Rasmussen and Williams, 2006) with a kernel function usually employed for automatic relevance determination (ARD) of the input features. Results show the advantages of the MRG scheme in both scenarios. Furthermore, we apply MRG in a dependence detection problem using both real and simulated remote sensing data. For the sake of reproducibility, the interested reader may find the source code of the experiments in `http://isp.uv.es/code/RG.zip`.

5.1 Experiment 1: A first analysis of the efficiency

We test the new MRG scheme in a simple numerical simulation involving a bi-dimensional target pdf:

$$\bar{\pi}(x_1, x_2) \propto \exp\left(-\frac{(x_1^2 - \mu_1)^2}{2\delta_1^2} - \frac{(x_2 - \mu_2)^2}{2\delta_2^2}\right),$$

with $\mu_1 = 4$, $\mu_2 = 1$, $\delta_1 = \sqrt{\frac{5}{2}}$ and $\delta_2 = 1$. Figure 5 shows the contour plot of $\bar{\pi}(x_1, x_2)$. Our goal is to approximate via Monte Carlo the expected value, $\mathbb{E}[\mathbf{X}]$ where $\mathbf{X} = [X_1, X_2] \sim \bar{\pi}(x_1, x_2)$. We test different Gibbs techniques: the MH (Robert and Casella, 2004) and IA^{2}RMS (Martino et al., 2015a) algorithms within SG and within MRG sampling schemes. For the MH method, we use a Gaussian random walk proposal,

$$q(x_{d,m}^{(t)}) \propto \exp\left(-\frac{(x_{d,m}^{(t)} - x_{d,m-1}^{(t)})^2}{2\sigma^2}\right),$$

for $d \in \{1, 2\}$, $1 \leq m \leq M$ and $1 \leq t \leq T$. We test different values of the σ parameter. For IA^{2}RMS, we start with the set of support points $\mathcal{S}_0 = \{-10, -6, -2, 2, 6, 10\}$, see (Martino et al.,

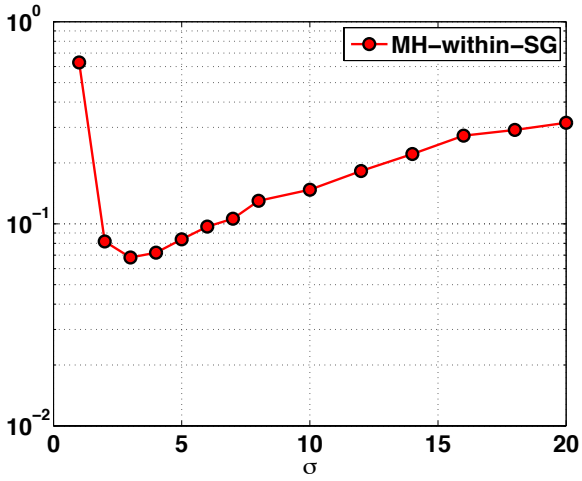
2015a) for further details. When we consider the standard Gibbs (SG) scheme $M = 1$, whereas for MRG $M > 1$. We averaged the Mean Square Error (MSE) over 10^5 independent runs for each Gibbs scheme.

Figure 7(a) shows the MSE (in log-scale) of the MH-within-SG scheme as function of the standard deviation σ of the proposal pdf (we set $M = 1$ and $T = 1000$, in this case). The performance of the Gibbs samplers depends strongly on the choice of σ of the *internal* MH method. The optimal value is approximately $\sigma^* \approx 3$. The use of an adaptive proposal pdf is a possible solution, as shown in Figure 8(a). Figure 7(b) depicts the MSE (in log-scale) as function of T with $M = 1$ and $M = 20$ (for MH-within-SG we also show the case $\sigma = 1$ and $\sigma = 3$). Again we observe the importance of using the optimal value $\sigma^* \approx 3$ and, as a consequence, using an adaptive proposal pdf is recommended, see e.g. (Haario et al., 2001). Moreover, the use $M = 20$ improves the results even without employing all the points in the estimators (i.e., in a SG scheme) since, as M increases, we improve the convergence of the internal chain. Moreover, the MH-within-MRG technique provides the smallest MSE values. We can thus conclude that recycling the internal samples provides more efficient estimators.

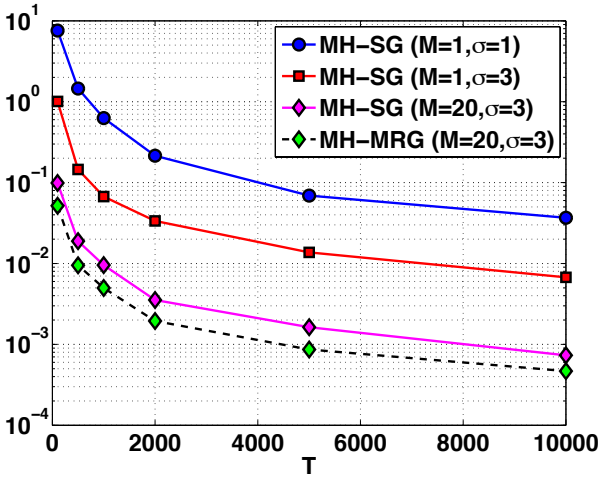
These considerations are confirmed in Figure 8(a) (represented again in log-scale). Here we fix again $T = 1000$ and vary M. As M increases, the MSE becomes smaller when the MRG technique is employed. When a standard Gibbs sampler (SG) is used, the curves show an horizontal asymptote since the internal chains converge after a certain value $M \geq M^*$, and there is not a great benefit from increasing M (recall that in SG we do not recycle the internal samples). Within an MRG scheme the increase of M always yield lower MSE since now we do recycle the internal samples (note the log-scale). Clearly, the benefit of using MRG w.r.t. SG increases as M grows. Figure 8(a) also shows the advantage of using an adaptive MCMC scheme (in this case IA^{2}RMS (Martino et al., 2015a)). The advantage is clearer when the MH and IA^{2}RMS schemes are used within MRG. More specifically note that, as the MH method employed the optimal scale $\sigma^* \approx 3$, Figure 8(a) shows the importance of a non-parametric construction of the proposal pdf employed in IA^{2}RMS. Actually, such construction allows adaptation of the entire shape of the proposal, which becomes closer and closer to the target. The performance of IA^{2}RMS and MH within Gibbs becomes more similar as M increases. This is due to the fact that, in this case, with a high enough value of M, the MH chain is able to exceed its burn-in period and eventually converges. Finally, note that the adaptation speeds up the convergence of the chain generated by IA^{2}RMS. The advantage of using the adaptation is more evident for intermediate values of M, e.g., $10 < M < 30$, where the difference with the use of a standard MH is higher. As M increases and the chain generated by MH converges, the difference between IA^{2}RMS and MH is reduced.

In Figure 8(b), we compare the performance of IA^{2}RMS-within-MRG scheme, setting $M = 20$ and varying T, with MH-within-a standard Gibbs scheme (i.e., $M = 1$) with a longer chain, i.e., a higher value of $T' > T$. In order to provide a comparison as fair as possible, we use the optimal scale parameter $\sigma^* \approx 3$ for the MH method. For each value of T and T', the MSE and computational time (in seconds) is given.[4] We can observe that, for a fixed time, IA^{2}RMS-within-MRG outperforms in MSE the standard MH-within-Gibbs scheme with a longer chain. These observations confirm the advantages of the proposed MRG approach.

[4]The computational times are obtained in a Mac processor 2.8 GHz Intel Core i5.



(a) MSE as function of σ.



(b) MSE as function of T.

Figure 7: **Exp. 1- (a)** MSE (log-scale) as function of σ for MH-within-SG ($M = 1$ and $T = 1000$). **(b)** MSE (log-scale) as function of T for MH-within-SG and MH-within-MRG schemes. We have tested $M \in \{1, 20\}$ and $\sigma = \{1, 3\}$ (we recall that $\sigma = 3$ is the optimal scale parameter for MH; see Figure **(a)**).

5.2 Experiment 2: Learning Hyperparameters in Gaussian Processes

Gaussian processes (GPs) are Bayesian state-of-the-art tools for function approximation and regression (Rasmussen and Williams, 2006). As for any kernel method, selecting the covariance function and learning its hyperparameters is the key to attain significant performance. We here evaluate the proposed approach for the estimation of hyperparameters of the Automatic Relevance Determination (ARD) covariance (Bishop, 2006, Chapter 6). Notationally, let us assume observed data pairs $\{y_j, \mathbf{z}_j\}_{j=1}^P$, with $y_j \in \mathbb{R}$ and

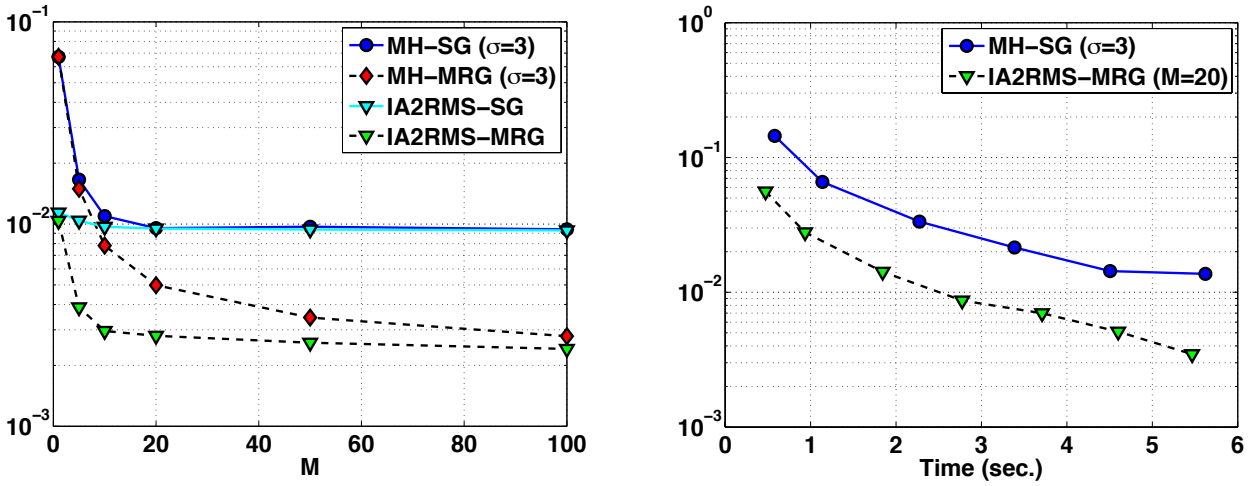
$$\mathbf{z}_j = [z_{j,1}, z_{j,2}, \ldots, z_{j,L}]^\top \in \mathbb{R}^L,$$

where L is the dimension of the input features. We also denote the corresponding $P \times 1$ output vector as $\mathbf{y} = [y_1, \ldots, y_P]^\top$ and the $L \times P$ input matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_P]$. We address the regression problem of inferring the unknown function f which links the variable y and $\mathbf{z}$. Thus, the assumed model is

$$y = f(\mathbf{z}) + e, \tag{8}$$

where $e \sim N(e; 0, \sigma^2)$, and that $f(\mathbf{z})$ is a realization of a Gaussian Process (GP) (Rasmussen and Williams, 2006). Hence $f(\mathbf{z}) \sim \mathcal{GP}(\mu(\mathbf{z}), \kappa(\mathbf{z}, \mathbf{r}))$ where $\mu(\mathbf{z}) = 0$, $\mathbf{z}, \mathbf{r} \in \mathbb{R}^L$, and we consider the ARD kernel function

$$\kappa(\mathbf{z}, \mathbf{r}) = \exp\left(-\sum_{\ell=1}^{L} \frac{(z_\ell - r_\ell)^2}{2\delta_\ell^2}\right), \quad \text{with} \quad \delta_\ell > 0, \quad \ell = 1, \ldots, L. \tag{9}$$



(a) MSE as function of M ($T = 1000$).

(b) MSE as function of spent time (sec.).

Figure 8: **Exp. 1- (a)** MSE (log-scale) as function of M for different MCMC-within-Gibbs schemes (we fix $T = 1000$). **(b)** MSE (log-scale) as function of the spent computational time (seconds). For IA^{2}RMS-within-MRG, we fix $M = 20$ and vary T, computing the MSE in estimation and the corresponding spent computational time. For MH-within-SG, we set $\sigma = 3$, $M = 1$, and vary T' (longer than T) and again we compute MSE and the spent time.

Note that we have a different hyper-parameter δ_ℓ for each input component z_ℓ, hence we also define $\boldsymbol{\delta} = \delta_{1:L} = [\delta_1, \ldots, \delta_L]$. Using ARD allows us to infer the relative importance of different components of inputs: a small value of δ_ℓ means that a variation of the ℓ-component z_ℓ impacts the output more, while a high value of δ_ℓ shows virtually independence between the ℓ-component and the output.

Given these assumptions, the vector $\mathbf{f} = [f(\mathbf{z}_1), \ldots, f(\mathbf{z}_P)]^\top$ is distributed as

$$p(\mathbf{f}|\mathbf{Z}, \boldsymbol{\delta}, \kappa) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K}), \tag{10}$$

where $\mathbf{0}$ is a $P \times 1$ null vector, and $\mathbf{K}_{ij} := \kappa(\mathbf{z}_i, \mathbf{z}_j)$, for all $i, j = 1, \ldots, P$, is a $P \times P$ matrix. Note that, in Eq. (10), we have expressed explicitly the dependence on the input matrix $\mathbf{Z}$, on the vector $\boldsymbol{\delta}$ and on the choice of the kernel family κ. Therefore, the vector containing all the hyper-parameters of the model is

$$\begin{aligned} \boldsymbol{\theta} &= [\theta_{1:L} = \delta_{1:L}, \theta_{L+1} = \sigma], \\ \boldsymbol{\theta} &= [\boldsymbol{\delta}, \sigma] \in \mathbb{R}^{L+1}, \end{aligned}$$

i.e., all the parameters of the kernel function in Eq. (9) and standard deviation σ of the observation noise. Considering the filtering scenario and the tuning of the parameters (i.e., inferring the vectors $\mathbf{f}$ and $\boldsymbol{\theta}$), the full Bayesian solution addresses the study of the full posterior pdf involving $\mathbf{f}$ and

$\boldsymbol{\theta}$,

$$p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa) = \frac{p(\mathbf{y}|\mathbf{f}, \mathbf{Z}, \boldsymbol{\theta}, \kappa)p(\mathbf{f}|\mathbf{z}, \boldsymbol{\theta}, \kappa)p(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{Z}, \kappa)}, \tag{11}$$

where $p(\mathbf{y}|\mathbf{f}, \mathbf{Z}, \boldsymbol{\theta}, \kappa) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \sigma^2\mathbf{I})$ given the observation model in Eq. (8), $p(\mathbf{f}|\mathbf{z}, \boldsymbol{\theta}, \kappa)$ is given in Eq. (10), and $p(\boldsymbol{\theta})$ is the prior over the hyper-parameters. We assume $p(\boldsymbol{\theta}) = \prod_{\ell=1}^{L+1} \frac{1}{\theta_\ell^\beta}\mathbb{I}_{\theta_\ell}$ where $\beta = 1.3$ and $\mathbb{I}_v = 1$ if $v > 0$, and $\mathbb{I}_v = 0$ if $v \leq 0$. Note that the posterior in Eq. (11) is analytically intractable but, given a fixed vector $\boldsymbol{\theta}'$, the marginal posterior of $p(\mathbf{f}|\mathbf{y}, \mathbf{Z}, \boldsymbol{\theta}', \kappa) = \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ is known in closed-form: it is Gaussian with mean $\boldsymbol{\mu}_p = \mathbf{K}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}$ and covariance matrix $\boldsymbol{\Sigma}_p = \mathbf{K} - \mathbf{K}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{K}$ (Rasmussen and Williams, 2006). For the sake of simplicity, in this experiment we focus on the marginal posterior density of the hyperparameters,

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa) = \int p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa)d\mathbf{f} \propto p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{Z}, \kappa)p(\boldsymbol{\theta}),$$

which can be evaluated analytically. Actually, since $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{Z}, \kappa) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathbf{K} + \sigma^2\mathbf{I})$ and $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa) \propto p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{Z}, \kappa)p(\boldsymbol{\theta})$, we have

$$\log\left[p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa)\right] \;\propto\; -\frac{1}{2}\mathbf{y}^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log\left[\det\left[\mathbf{K} + \sigma^2\mathbf{I}\right]\right] - \beta\sum_{\ell=1}^{L+1}\log\theta_\ell, \tag{12}$$

with $\theta_\ell > 0$, where clearly $\mathbf{K}$ depends on $\theta_{1:L} = \delta_{1:L}$ and recall that $\theta_{L+1} = \sigma$ (Rasmussen and Williams, 2006). However, the moments of this marginal posterior cannot be computed analytically. Then, in order to compute the Minimum Mean Square Error (MMSE) estimator, i.e., the expected value $\mathbb{E}[\boldsymbol{\Theta}]$ with $\boldsymbol{\Theta} \sim p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa)$, we approximate $\mathbb{E}[\boldsymbol{\Theta}]$ via Monte Carlo quadrature. More specifically, we apply a Gibbs-type samplers to draw from $\pi(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa)$. Note that dimension of the problem is $D = L + 1$ since $\boldsymbol{\theta} \in \mathbb{R}^D$.

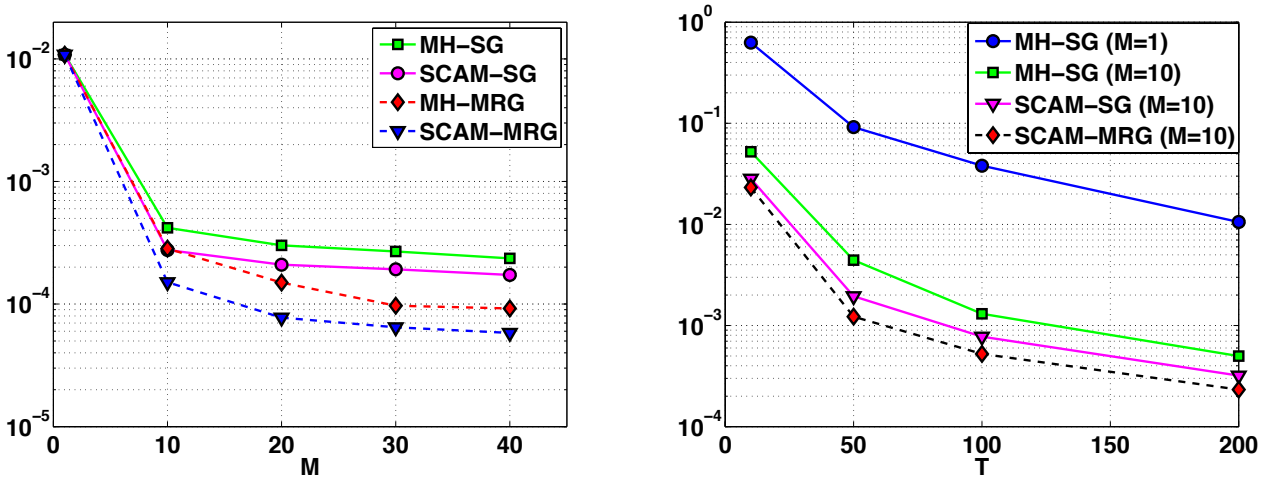
We generated the $P = 500$ pairs of data, $\{y_j, \mathbf{z}_j\}_{j=1}^P$, drawing $\mathbf{z}_j \sim \mathcal{U}([0, 10]^L)$ and $\mathbf{y}_j$ according to the model in Eq. (8), considered $L \in \{1, 3\}$ so that $D \in \{2, 4\}$, and set $\sigma^* = \frac{1}{2}$ for both cases, $\delta^* = 1$ and $\boldsymbol{\delta}^* = [1, 3, 1]$, respectively (recall that $\boldsymbol{\theta}^* = [\boldsymbol{\delta}^*, \sigma^*]$). Keeping fixed the generated data for each scenario, we then computed the ground-truths using an exhaustive and costly Monte Carlo approximation, in order to be able to compare the different techniques.

We tested the standard MH within SG and MRG, and also the Single Component Adaptive Metropolis (SCAM) algorithm (Haario et al., 2005) within SG and MRG. SCAM is a component-wise version of the adaptive MH method (Haario et al., 2001) where the covariance matrix of the proposal is automatically adapted. In SCAM, the covariance matrix of the proposal is diagonal and each element is adapted considering only the corresponding component: that is, the variances of the marginal densities of the target pdf are estimated and used as a scale parameter of the proposal pdf in the corresponding component.[5] We averaged the results using 10^3 independent runs. Figure 9(a) shows the MSE curves (in log-scale) of the different schemes as function of $M \in \{1, 10, 20, 30, 40\}$, while keeping fixed $T = 100$ (in this case, $D = 2$). Figure 9(b) depicts the

[5]More specifically, we have implemented an accelerated version of SCAM which takes more advantage of the MRG scheme, since the variance is also adjusted online during the sampling of the considered full-conditional (for more details, see the code at `http://isp.uv.es/code/RG.zip`).

MSE curves ($D = 4$) as function of T considering in one case $M = 1$ and $M = 10$ for the rest. In both figures, we notice that (1) using an $M > 1$ is advantageous in any case (SG or MRG), (2) using a procedure to adapt the proposal improves the results, and (3) MRG, i.e., recycling all the generated samples, always outperforms the SG schemes.

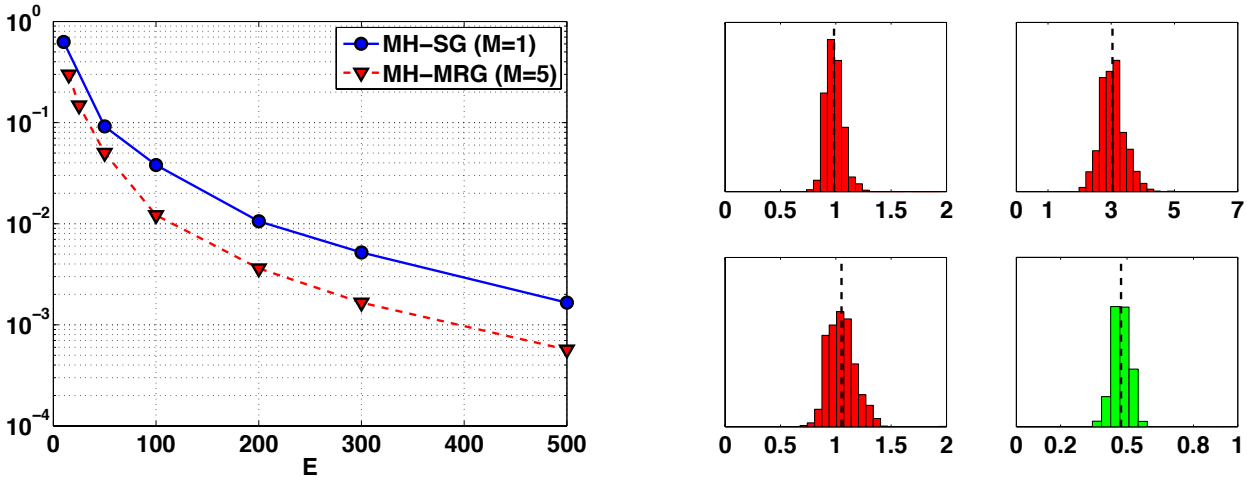
Figure 10(a) compares the MH-within-SG with the MH-within-MRG, showing the MSE as function of the total number of target evaluations $E = MT$. We set $M = 5$, $T \in \{3, 5, 10, 20, 40, 60, 100\}$ for MH-within-MRG, whereas we have $M = 1$ and $T \in \{10, 50, 100, 200, 300, 500\}$ for MH-within-SG. Namely, we used a longer Gibbs chain for MH-within-SG. Note that the MH-within-MRG provides always smaller MSEs, considering the same total number of evaluations E of the target density. Finally, Figure 10(b) depicts the histograms of the samples $\boldsymbol{\theta}^{(t)}$ drawn from the posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa)$ in a specific run, with $D = 4$, generated using MH-within-MRG with $M = 5$ and $T = 2000$. The dashed line represents the mean of the samples (recall that $\delta_1^* = 1$, $\delta_2^* = 3$, $\delta_3^* = 1$ and $(\sigma^*)^2 = 0.5$). Note that all the samples $\boldsymbol{\theta}^{(t)}$ can be employed for approximating the full Bayesian solution of the GP which involves the joint posterior pdf in Eq. (11).



(a) MSE as function of M ($T = 100$).

(b) MSE as function of T.

Figure 9: **Exp. 2- (a)** MSE (log-scale) as function of M (starting with $M = 1$) for different MCMC-within-Gibbs schemes ($T = 100$ and $D = 2$). **(b)** MSE (log-scale) as function of T for different techniques (in this case, $D = 4$), with $M = 1$ for the MH-within-SG method depicted with a solid line and circles, whereas $M = 10$ for the remaining curves. Note that, in both figures, the MRG approaches, shown with dashed lines, always outperform the corresponding standard Gibbs (SG) schemes, shown with solid lines.



(a) MSE versus the number of target evaluations E.

(b) Histograms of the generated samples.

Figure 10: **Exp. 2- (a)** MSE (log-scale) as function of the total number of target evaluations $E = MT$ ($D = 4$). Namely, for MH-within-SG we have $M = 1$ and $T \in \{10, 50, 100, 200, 300, 500\}$, whereas for MH-within-MRG we have $M = 5$ and $T \in \{3, 5, 10, 20, 40, 60, 100\}$. **(b)** Histograms of the samples drawn from the posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa)$ in a specific run, with $\boldsymbol{\theta} = [\delta_1, \delta_2, \delta_3, \sigma]$, i.e., $D = 4$, generated using MH-within-MRG with $M = 5$ and $T = 2000$.

5.3 Experiment 3: Learning Dependencies in Remote Sensing Variables

We now consider the application of the MRG scheme to study the dependence among different geophysical variables. Specifically, we consider the case of temperature estimation from thermal infrared (TIR) remotely sensed data. In this scenario, land surface temperature (T_s) and emissivity (ϵ) are the two main geo-biophysical variables to be retrieved from TIR data, since most of the energy detected by the sensor in this spectral region is directly emitted by the land surface. The atmosphere status can be considered as a mediating variable in the relations between the satellite measured T and T_s, and is here summarized by the integral of the water vapour W through the whole atmospheric column. Both variables T_s and ϵ are coupled and constitute a typical problem in remote sensing referred as to the "temperature and emissivity separation problem". On the one hand, models for estimating land temperature T_s typically involve simple parametrizations of at-sensor brightness temperatures T, the mean and/or differential emissivities ($\bar{\epsilon}$ and $\Delta\epsilon$), and the total atmospheric water vapour content W. On the other hand, a plethora of models for estimating ϵ have been devised (Jimenez-Muñoz and Sobrino, 2008).

Here we focus on the application of the MRG sampler to infer the statistical dependencies between the observed variables. We aim to obtain the dependence graph between the considered geophysical variables, $X_1 = T_s$, $X_2 = W$, $X_3 = \epsilon$ and $X_4 = T$. To assess such relations, we considered synthetic data simulating ASTER sensor conditions (Sobrino et al., 2008). A total of 6588 data points was available. For simplicity, we focused on band 10 ($\sim 8.3\mu$m) for T and ϵ, and subsampled the dataset to finally work with 220 data points. The data was subsequently standardized.

5.3.1 Main procedure

We study the 12 possible regression models of type

$$X_i = f_{j,i}(X_j) + E_{j,i}. \qquad i, j \in \{1, 2, 3, 4\}, \quad \text{with } i \neq j,$$

where $E_{j,i} \sim \mathcal{N}(0, \sigma^2_{i,j})$ and $f_{j,i}(x_j)$ is a realization of a Gaussian Process (GP) (Rasmussen and Williams, 2006), with zero mean and kernel function defined in Eq. (9) (note that in this case $L = 1$). For each regression problem, we trained the corresponding GP model using SCAM-within-MRG with $T = 200$ and $M = 10$. We analyze the empirical distributions of corresponding hyper-parameters $\boldsymbol{\theta}_{i,j} = [\delta_{i,j}, \sigma^2_{i,j}]$ obtained by Monte Carlo. We focus mainly on the distribution of $\delta_{i,j}$, i.e., the hyper-parameter of the kernel in Eq. (9). Hyper-parameter $\delta_{i,j}$ will tend to be higher and its distribution will exhibit heavier right tail if the Signal-Noise-Ratio (SNR) is low and if X_i and X_j are close to independence. However, the spread of $\delta_{i,j}$ also depends on the noise power in the system, the unknown mapping (deterministic or stochastic) linking X_i and X_j, and the asymmetry of the regression functions, i.e. in general $f_{i,j} \neq f_{j,i}$. Hypothesis testing comes into play to determine the significance of the association between all pairs of variables X_i and X_j.

5.3.2 Hypothesis testing and surrogate data

In order to find significance levels and thresholds about the existence of any possible dependence (strong or weak), we perform an hypothesis test with the null-hypothesis "$\mathcal{H}_0$: independence between X_i and X_j". We build the sampling distribution of $\mathcal{H}_0$ by the surrogate data method in (Theiler et al., 1992). Under the null-hypothesis $\mathcal{H}_0$ of independence, some proper surrogate data can be generated by shuffling the output values (i.e., we permute the outputs keeping fixed the inputs) while keeping fixed the input features. This way we have new data points sharing the same input and output values with the true data, but breaking any structure which links the inputs with the outputs. Clearly, this procedure considers different values for each pair of indices i and j (i.e., each variables X_i and X_j).

Given a set of surrogate data, we applied SCAM-within-MRG with $T = 200$ and $M = 10$ and obtain the empirical distribution of the hyper-parameter $\delta_{i,j}$. We repeated this procedure 500 times, generating different surrogate data at each run. We computed different empirical moments of the distribution of the hyper-parameter $\delta_{i,j}$, as mean, median and variance from the empirical distributions obtained via Monte Carlo with the true data and the surrogate data. Averaged results over 500 runs are shown in Table 2. We show mean $\mathbb{E}[\delta]$, median $\bar{\delta}_{0.5}$ and standard deviation $\sqrt{\text{var}[\delta]}$, obtained analyzing the true data, and the p-values obtained comparing the estimated statistics with the corresponding distributions acquired by the surrogate data method.

Figure 11 shows the undirected graphs with significance level set to $\alpha = 0.1$. The width of the lines represents the significance of the link according to the estimated p-values. If one of the two p-values corresponding the two possible regressions between the variable X_i and X_j is greater than 0.1 the link is depicted in dashed line. The obtained graphs are consistent with a priori physical knowledge about the problem. In particular, it is common sense that the surface temperature T_s and ϵ depend on W and T. After all, remote sensing data processing mainly deals with the estimation of the surface parameters from the acquired satellite brightness temperatures (T) and the atmosphere status (W)[6]. In addition, while emissivity ϵ is generally used for retrieving T_s, some simpler methods using only T are successfully used (Jimenez-Munoz and Sobrino, 2007). It is also worth noting that the used data considered only natural surfaces, hence the database was biased towards high values of ϵ, thus explaining why the relationship between T_s and ϵ was not captured.

6 Conclusions

The Gibbs sampling method is a well-known Markov chain Monte Carlo (MCMC) algorithm, extensively applied in statistics, signal processing and machine learning, in order to obtain samples from complicated a posteriori distributions. A Gibbs sampling approach is particularly useful in high-dimensional inference problems, since the generated samples are constructed component by component. In this sense, it can be considered the MCMC counterpart of the particle filtering methods, for static (i.e., non-sequential inference) and batch (i.e., all the data are processed jointly)

[6]While one could be tempted to infer causal relations, the proposed approach cannot cope with asymmetries in the pdfs of variables through GP hyperparameter estimation.

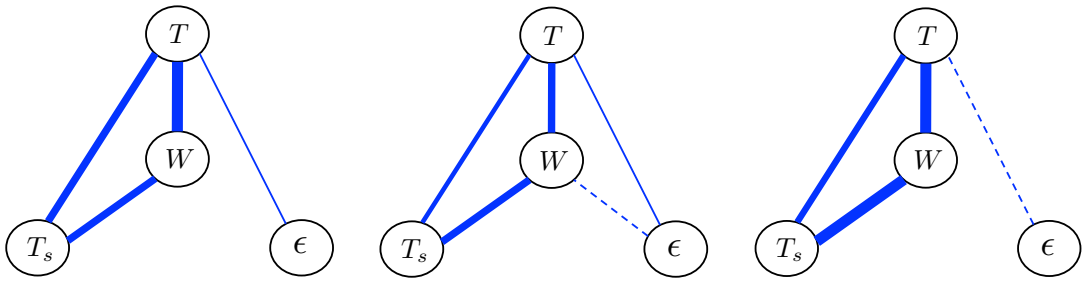
Table 2: **Exp. 3-** Results for the mean $\mathbb{E}[\delta]$, the median $\bar{\delta}_{0.5}$ and the std $\sqrt{\text{var}[\delta]}$.

Link	**In**	**Out**	**Mean** ($\mathbb{E}[\delta]$)	p**-value**	**Median** ($\bar{\delta}_{0.5}$)	p**-value**	**Std** ($\sqrt{\text{var}[\delta]}$)	p**-value**
$T_s - W$	T_s	W	0.68	0.004	0.67	0.002	0.15	0.001
	W	T_s	0.32	0.001	0.32	0.002	0.15	0.001
$T_s - \epsilon$	T_s	ϵ	11.85	0.22	5.61	0.20	33.16	0.56
	ϵ	T_s	11.58	0.23	5.67	0.23	47.36	0.68
$T_s - T$	T_s	T	2.56	0.03	2.53	0.08	0.66	0.006
	T	T_s	1.83	0.02	1.76	0.03	0.48	0.002
$W - T$	W	T	0.33	0.001	0.34	0.001	0.14	0.001
	T	W	0.40	0.001	0.39	0.004	0.14	0.001
$W - \epsilon$	W	ϵ	11.77	0.22	4.04	0.09	32.76	0.54
	ϵ	W	11.20	0.21	4.48	0.13	47.97	0.71
$T - \epsilon$	T	ϵ	5.22	0.06	3.34	0.09	6.88	0.10
	ϵ	T	6.32	0.09	3.95	0.10	10.24	0.14

frameworks. The key point for the successful application of the SG sampler is the ability to draw efficiently from each the full-conditional densities. However, in real-world applications, drawing from complicated full-conditionals is required, and no direct methods are available in these cases. For solving this issue, several specifically-designed MCMC algorithms has been proposed to be employed within the SG sampler. Most of them require the generation of auxiliary samples that are not included in the resulting estimators. The use of more auxiliary samples accelerates the convergence of the generated Gibbs chain and improves the performance, at the expense of an additional computational effort.

In this work, we have shown that these auxiliary samples can be included within the Gibbs estimators improving their efficiency without any extra computational cost. The consistency of the resulting estimators is ensured since the novel MRG scheme is equivalent to an alternative formulation of the well-known chain rule method. This alternative chain rule procedure has been also described and discussed in this work. Numerical simulations have confirmed the benefits of the novel scheme. First, we have compared the SG and MRG schemes in a toy example, considering the use of several parameter values and the application of different internal MCMC algorithms. MRG yielded clear improvements of the performance w.r.t. the SG approach. Then, we tested the SG and MRG schemes in a hyperparameter estimation problem for GP regression, considering a kernel for automatic relevance determination (ARD). The MRG approach provided the smallest MSEs in estimation of the hyperparameters in all cases. Finally, we studied the application of the proposed MRG sampler to unveil the dependence between different geophysical variables considering remote sensing satellite data.

As future lines, we plan to investigate the use of different number of samples M_1,..., M_d to be drawn from full-conditionals, and the possible design of an automatic tuning strategy for adapting the number of samples to improve the performance given a specific posterior distribution. This will imply efforts in parallel MRG samplers for scalable learning. We also plan to extend the numerical study with remote sensing data using the MRG scheme in order to infer causal dependences among the geophysical variables, and for that we will consider applying the MRG in (conditional) independence estimation schemes.



(a) Mean. (b) Median. (c) std.

Figure 11: **Exp. 3-** Undirected graphs with significant level $\alpha = 0.1$. The width of the lines represents the significance of the link, shown in Table 2. If one of the two p-values corresponding the two possible regressions is greater than 0.1 the link is depicted in dashed line. The graphs are obtained considering **(a)** the mean, **(b)** the median and **(c)** the standard deviation of the distribution of δ.

Acknowledgements

We thank Dr. J. C. Jiménez at IPL for the remote sensing dataset and the fruitful discussions. This work has been supported by the European Research Council (ERC) through the ERC Consolidator Grant SEDAL ERC-2014-CoG 647423.

References

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, 50(1):5–43.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Brewer, M. and Aitken, C. (1993). Discussion on the meeting on the Gibbs sampler and other Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B*, 55(1):69–70.

Bugallo, M. F., Xu, S., and Djurić, P. M. (2007). Performance comparison of EKF and particle filtering methods for maneuvering targets. *Digital Signal Processing*, 17:774–786.

Caffo, B. S., Booth, J. G., and Davison, A. C. (2002). Empirical supremum rejection sampling. *Biometrika*, 89(4):745–754.

Cai, B., Meyer, R., and Perron, F. (2008). Metropolis-Hastings algorithms with adaptive proposals. *Statistics and Computing*, 18:421–433.

Chen, Y., Bornn, L., De Freitas, N., Eskelin, M., Fang, J., and Welling, M. (2016). Herded Gibbs sampling. *Journal of Machine Learning Research*, 17(1):263–291.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer.

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

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

Fox, C. (2012). A Gibbs sampler for conductivity imaging and other inverse problems. *Proc. of SPIE, Image Reconstruction from Incomplete Data VII*, 8500:1–6.

Gelfand, A. E. and Lee, T. M. (1993). Discussion on the meeting on the Gibbs sampler and other Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B*, 55(1):72–73.

Gilks, W. R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. *Bayesian Statistics*, 4:641–649.

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

Gilks, W. R., Neal, R., Best, N. G., and Tan, K. K. C. (1997). Corrigidum: Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, 46(4):541–542.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41(2):337–348.

Görür, D. and Teh, Y. W. (2011). Concave convex adaptive rejection sampling. *Journal of Computational and Graphical Statistics*, 20(3):670–691.

Goudie, R. J. B. and Mukherjee, S. (2016). A Gibbs sampler for learning DAGs. *Journal of Machine Learning Research*, 17(2):1–39.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Haario, H., Saksman, E., and Tamminen, J. (2005). Component-wise adaptation for high dimensional MCMC. *Computational Statistics*, 20(2):265–273.

Hörmann, W. (1995). A rejection technique for sampling from T-concave distributions. *ACM Transactions on Mathematical Software*, 21(2):182–193.

Hörmann, W. (2002). A note on the performance of the Ahrens algorithm. *Computing*, 69:83–89.

Hörmann, W., Leydold, J., and Derflinger, G. (2007). Inverse transformed density rejection for unbounded monotone densities. *Research Report Series/ Department of Statistics and Mathematics (Economy and Business), Vienna University*.

Jimenez-Munoz, J. C. and Sobrino, J. A. (2007). Feasibility of retrieving land-surface temperature from aster tir bands using two-channel algorithms: A case study of agricultural areas. *IEEE Geoscience and Remote Sensing Letters*, 4(1):60–64.

Jimenez-Muñoz, J. C. and Sobrino, J. A. (2008). Split-window coefficients for land surface temperature retrieval from low resolution thermal infrared sensors. *IEEE Geoscience and Remote Sensing Letters*, 5(4):806–809.

Johnson, A. A., Jones, G. L., and Neath, R. C. (2013). Component-wise Markov Chain Monte Carlo: uniform and geometric ergodicity under mixing and composition. *Statistical Science*, 28(3):360–375.

Koch, K. R. (2007). Gibbs sampler by sampling-importance-resampling. *Journal of Geodesy*, 81(9):581–591.

Kotecha, J. and Djurić, P. M. (1999). Gibbs sampling approach for generation of truncated multivariate Gaussian random variables. *Proceedings of Acoustics, Speech, and Signal Processing, (ICASSP)*.

Levine, R. A., Yu, Z., Hanley, W. G., and Nitao, J. J. (2005). Implementing component-wise Hastings algorithms. *Computational Statistics and Data Analysis*, 48(2):363–389.

Liang, F., Liu, C., and Caroll, R. (2010). *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. Wiley Series in Computational Statistics, England.

Liu, J. S. (2004). *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag.

Lucka, F. (2016). Fast Gibbs sampling for high-dimensional Bayesian inversion. *arXiv:1602.08595*.

Marrelec, G. and Benali, H. (2004). Automated rejection sampling from product of distributions. *Computational Statistics*, 19(2):301–315.

Martino, L. and Míguez, J. (2011). A generalization of the adaptive rejection sampling algorithm. *Statistics and Computing*, 21(4):633–647.

Martino, L., Read, J., and Luengo, D. (2015a). Independent doubly adaptive rejection Metropolis sampling within Gibbs sampling. *IEEE Transactions on Signal Processing*, 63(12):3123–3138.

Martino, L., Yang, H., Luengo, D., Kanniainen, J., and Corander, J. (2015b). A fast universal self-tuned sampler within Gibbs sampling. *Digital Signal Processing*, 47:68 – 83.

Meyer, R., Cai, B., and Perron, F. (2008). Adaptive rejection Metropolis sampling using Lagrange interpolation polynomials of degree 2. *Computational Statistics and Data Analysis*, 52(7):3408–3423.

Müller, P. (1991). A generic approach to posterior integration and Gibbs sampling. *Technical Report 91-09, Department of Statistics of Purdue University*.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87(419):861–868.

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

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B*, 59(2):291–317.

Shao, W., Guo, G., Meng, F., and Jia, S. (2013). An efficient proposal distribution for Metropolis-Hastings using a b-splines technique. *Computational Statistics and Data Analysis*, 53:465–478.

Sobrino, J. A., Jimenez-Muñoz, J. C., Soria, G., Romaguera, M., Guanter, L., Moreno, J., Plaza, A., and Martinez, P. (2008). Land surface emissivity retrieval from different VNIR and TIR sensors. *IEEE Transactions on Geoscience and Remote Sensing*, 46(2):316–327.

Tanizaki, H. (1999). On the nonlinear and non-normal filter using rejection sampling. *IEEE Transaction on automatic control*, 44(3):314–319.

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., and Farmer, J. D. (1992). Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, 58(2):77–94.

Zhang, H., Wu, Y., Cheng, L., and Kim, I. (2016). Hit and run ARMS: adaptive rejection Metropolis sampling with hit and run random direction. *Journal of Statistical Computation and Simulation*, 86(5):973–985.