Bayesian Inference, Stochastic Simulation and Their Applications in Wireless Communication Systems

Flávio R. Ávila and Michel P. Tcheou

***Abstract*—Bayesian inference has been successfully applied in fields as varied as anti-spam filtering, DNA sequencing, war codebreaking and election forecasting. Founded on the apparently simple Bayes' theorem, which relates the previous distribution of a parameter with its distribution after evidence is collected, Bayesian tools allow for incorporating all existing knowledge about the phenomenon under study in order to improve parameter estimation. Because of the stochastic nature of the wireless channel, Bayesian inference is particularly well suited to the problem of symbol detection in many modern digital communication systems. When combined with Markov Chain Monte Carlo (MCMC) techniques, Bayesian receivers are capable of achieving minimum Bit Error Rate (BER) while avoiding the prohibitively high computational complexity associated with standard Maximum Likelihood (ML) or Maximum A Posteriori (MAP) estimators. In addition, such receivers are capable of numerically integrating out channel coefficients and noise variance, thus avoiding the need to use sub-optimal estimates of these parameters. This tutorial presents the rudiments of Bayesian statistics and MCMC in general, and discusses their applications in wireless communications in particular. The paper also details the design of Bayesian MCMC receiver in a system employing BPSK and subject frequency-selective fading and Gaussian noise. Afterwards, recent advances in Bayesian receivers are surveyed for several important practical wireless transmission schemes, including MIMO, CDMA and OFDM. In addition, the paper addresses the application of Bayesian tools in challenging channel conditions — namely, nonlinear, non-Gaussian, underwater and fast fading channels.**

***Index Terms*—Bayesian inference, Markov Chain Monte Carlo, Wireless communications, Symbol detection.**

I. Introduction

From the most prosaic questions, as to whether bring an umbrella when going outside, to complex ones, such as which profession to choose, life requires the ability to make decisions in an uncertain world. Coping effectively with such problems involves evaluating the likelihood of alternative hypotheses, assessing the risk of a decision, predicting future outcomes and so on. The field of statistical inference helps us to accomplish these tasks in a rigorous and systematic way.

Two schools of statistics, Bayesian and Frequentist [1]–[3], contest for primacy over inference methods and data analysis. The fundamental difference between them lies in the interpretation of probability and, as a consequence, in the treatment of the quantity one wishes to estimate — the parameters, in statistical jargon. The frequentists understand probability as the limit of the relative frequency of occurrence of an event in a universe of possibilities, and thus restrict the probabilistic analysis to repeatable events. In contrast, the Bayesian view is founded on the notion of probability as a subjective measure for the degree of belief in the likelihood of some event happening or on the veracity of a sentence.

As a consequence, frequentists argue that the parameters should be treated as unknown fixed constants, while Bayesians interpret them as random variables, even when they are predefined. Since the latter group understands probability as a degree of belief, the parameters being unknown is a legitimate reason for them to be treated as random variables with their associated probability distributions.

Although it was condemned as anti-scientific by high-caliber statisticians such as Ronald Fisher and Karl Pearson [4] because of its assumed subjective nature, the Bayesian view continued to be adopted during most of the XX century because it allowed solutions to problems for which the frequentist view failed. After decades of scorn, Bayesianism became more palatable for the statistical mainstream thanks to theoretical advances made by mathematicians such as Jeffreys [5], De Finetti [6] and Savage [7]. However, its definitive popularization came only at the end of the 80s, when the appearance of numerical techniques based on Markov Chain Monte Carlo (MCMC) [8] [9] allowed Bayesian tools to be applied in many problems that were previously inaccessible because of intense computational demand. Currently, Bayesian methods are present in the core of myriad applications in machine learning, genetics, anti-spam filtering [10], forensics, election forecasting [11] and other fields.

In engineering problems, including ones in telecommunications [12], effective solutions can be obtained by exploiting the flexibility of the Bayesian paradigm. The ability to introduce any prior knowledge — often subjective and hard to quantify — is valuable in real situations in which data are scarce. In a wireless communication system [13], the previous knowledge about the typical behavior of the channel in a given environment, in addition to the statistical properties of the transmitted signals, can be incorporated in the solution in order to allow more accurate detection of the transmitted symbols. The view of parameters as random variables, prerogative of the Bayesian philosophy, allows for the marginalization of the quantities that are not of immediate interest, yielding optimal solutions to the parameters that one wishes to estimate — for instance, the transmitted signal.

This paper presents the rudiments of Bayesian inference and the techniques of stochastic simulation that allow it to be ap-

The Ad Hoc Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Lisandro Lovisolo.

Flávio R. Ávila and Michel P. Tcheou are with the PROSAICO Lab, Rio de Janeiro State University. E-mails: {flavio.avila,mtcheou}@uerj.br. The authors would like to thank FAPERJ for funding their work.

Digital Object Identifier: 10.14209/jcis.2016.27

plied to complex problems. We will describe their application to the task of symbol detection and channel estimation in digital wireless communications systems, in which the frequency selective and noisy channel produce a problem that is adequate for Bayesian tools. Even though most of these problems can be conveniently tackled by Sequential Monte Carlo [14] methods, especially in fast fading channel conditions, we chose to focus on batch processing, while dedicating one section to sequential solutions. A good exposition of both batch and sequential approaches in signal processing can be found in [15].

The first sections of the tutorial introduce the foundations of Bayesian data analysis, and include an exposition of Bayes' theorem, the formalization of the concept of prior distribution (in particular, the conjugate and the non-informative prior distributions), the Bayesian hierarchical model and parameter elimination. In the latter sections, we turn on the numerical techniques for distribution simulation based on Markov Chain Monte Carlo (MCMC), which allow for the solution of inference problems that would be otherwise intractable. The focus will be on the popular and powerful Metropolis-Hastings algorithm and the Gibbs sampling. The remainder of the paper surveys possible Bayesian approaches to the problem of signal detection in a wireless communication system for important modern wireless communications systems.

II. Bayes' Theorem

We use statistic inference to obtain information about hidden quantities through observable data described by probabilistic models. In a CDMA system (Code Division Multiplex Access) [16], for example, several users share the same communication medium, transmitting in the same frequency band, and having their messages affected by channel distortion and noise. The receiver is responsible for processing this complicated combination of signals — which only depend on the transmitted messages indirectly — and inferring the most probable information sent by each user.

We formalize the concept of inference by defining the possibly multidimensional data set $\mathbf{x}$, whose distribution is referred to as $p(\mathbf{x}|\boldsymbol{\theta})$. Here, $\boldsymbol{\theta}$ congregates the unknown parameters to be estimated — although sometimes, as in the context of wireless communications, such parameters can represent unknown data (e.g. transmitted symbols).

The frequentist school understands the parameters $\boldsymbol{\theta}$ as unknown and fixed values. Consequently, the inference is carried out based on the distribution $p(\mathbf{x}|\boldsymbol{\theta})$, known as the likelihood function, and understood as a probability distribution of the observed data for a certain set of parameters $\boldsymbol{\theta}$. In contrast, the Bayesian school sees the parameters as random variables with certain associated distributions. Before data are observed, the parameters distribution quantifies the confidence level of the expert on their possible values, and is called *a priori* (or simply prior) distribution, $p(\boldsymbol{\theta})$. After considering data, we can build the so-called *a posteriori* (or simply posterior) distribution $p(\boldsymbol{\theta}|\mathbf{x})$, which combines the prior knowledge with newly acquired information.

The connection between these quantities is stated by Bayes' theorem (or Bayes' rule), first proposed by reverend Thomas Bayes (posthumously published in [17]) and later reformulated by Laplace [18]. By using the theorem, we can obtain the *a posteriori* distribution from the *a priori* distribution, considering observed data. The theorem states that the posterior distribution is obtained essentially by the product of the likelihood function and the prior distribution:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}. \tag{1}$$

In this expression $p(\mathbf{x})$ is the probability density of data vector $\mathbf{x}$ considering all possible $\boldsymbol{\theta}$. As $\mathbf{x}$ is known, $p(\mathbf{x})$ is a proportionality constant that makes the quotient integrate to one. Thus, $p(\mathbf{x})$ is given by

$$p(\mathbf{x}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{2}$$

Because $p(\mathbf{x})$ does not affect the shape of function $p(\boldsymbol{\theta}|\mathbf{x})$, one usually removes $p(\mathbf{x})$ from Bayes' theorem expression and rewrites it as a proportionality relation:

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3}$$

Bayes' theorem can be informally summarized as: the posterior probability of an event corresponds to the product between the prior distribution and the likelihood function generated from the observable data:

$$\text{Posterior probability} = \text{Likelihood} \times \text{Prior probability} \tag{4}$$

In summary, the theorem allows us to weigh prior knowledge along with new information after performing the experiment, producing new and refined knowledge.

1) Medical Testing: We illustrate the theorem in a situation of practical importance that unfortunately confuses both laymen and specialists. We present here a concise version of the problem exposed in [4]. A woman received a mammogram with abnormal evaluation. What is the probability that she has breast cancer? To answer this question, we need some basic statistics, namely:

(i) Probability of abnormal mammogram in the absent of disease: 10%.
(ii) Probability of abnormal mammogram in the presence of disease: 80%.
(iii) Probability of any woman suffering from breast cancer: 0.4%.

Many would claim that the probability of a woman having cancer is 80% — the percentage of abnormal results when disease is present. A crucial piece of information is often forgotten: the prior probability — *before* examination has been carried out — of any woman suffering from breast cancer. Considering disease rareness (0.4 %), the evidence needs to be very strong to significantly increase the probability from its initial baseline. Certainly, we expect that the abnormal result increases this initial probability, but due to exam imperfection, we anticipate that the posterior probability is an intermediate value between 0.4 and 80 %.

Bayes' theorem allows us to calculate it exactly. Let A denote the event "woman has cancer" and B, the event "abnormal mammography". We wish to calculate the posterior

probability of a woman having the disease when we know the abnormal test result, that is,

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}. \quad (5)$$

Using the given information, we have

(i) $p(A|B) = 0.8$
(ii) $p(B) = 0.004$
(iii) $p(A) = p(A|B)p(B) + p(A|\overline{B})p(\overline{B}) = 0.8 \times 0.004 + 0.1 \times 0.996 = 0.1028$.

Note that $p(A)$ is the probability of abnormal test result achieved by summing the probabilities of abnormal test results when a woman has cancer and when she is healthy. The bar above the event denotes its complementary, i.e., its negation.

Finally, by replacing these values in Eq. (5), we have

$$p(B|A) = \frac{0.8 \times 0.004}{0.1028} = 0.311 = 3.11\% \quad (6)$$

This counter-intuitive result shows the importance of considering these prior probabilities and recommends that medical tests should be carefully interpreted.

2) Binary Symmetric Channel: Fig. 1 illustrates a communication channel where binary symbols are transmitted (0 or 1) and, due to channel imperfections, there is a probability p usually smaller than 1 that a symbol is correctly detected [12]. Consider p_0 as the probability of sending bit 0, implying, $p_1 = 1 - p_0$.

Let X be the sent symbol and Y, the received one. It is useful to calculate the inverse probability that symbol X has been sent knowing that Y was received. By considering Bayes' theorem, we have

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} \quad (7)$$

where $p(Y = 1) = p_1 p + p_0(1 - p)$ e $p(Y = 0) = p_0 p + p_1(1 - p)$

Thus,

$$p(X = 1|Y = 1) = \frac{pp_1}{p_1 p + p_0(1 - p)} \quad (8)$$

$$p(X = 0|Y = 0) = \frac{pp_0}{p_0 p + p_1(1 - p)} \quad (9)$$

Naturally the probability that there is a transmission error in each case is obtained by calculating the complementary of each of the above expressions.

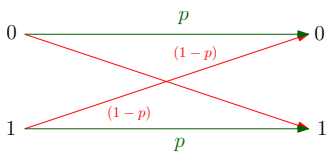

Fig. 1. Binary symmetric binary. Binary symbols are transmitted with error probability $(1 - p)$

III. Bayesians vs. Frequentists

The detailed exposition of the differences and similarities between the Bayesian and frequentist schools is clearly beyond the scope of this paper. Here we will limit ourselves to provide a glimpse of what we judge essential about the subject and we discuss two examples in which the differences between the schools is salient.

Let us informally illustrate the difference between the schools by invoking the familiar experiment of die toss. We wish to calculate, for instance, the probability that the face on the top is an even number. In the early development of probability, the solution would be based on the symmetry of the experiment, which yields to elementary events of equal probabilities. Then, the rules of probabilistic inference would be used to calculate the chances of more complex events. Because of symmetry, the six possible results in one toss should be equally likely, and since it is required that the sum of probabilities equals 1, it follows that the probability of each face turning up is 1/6. The probability of an even face would then be the ratio between the amount of even numbers between 1 and 6 (that is 2, 4 and 6) and the total number of possible results (6), yielding a probability of 1/2.

Notwithstanding its intuitive appeal, this argument would be rejected by a frequentist. Being the relative frequency of events, probability cannot be assigned to events before any experiment is performed. In this example, the probability of an even number is obtained by counting the times when the numbers 2, 4 or 6 were effectively obtained in a long sequence of tosses, and then dividing the result by the total number of tosses. Because of the law of large numbers, it is expected that this ratio will approach the classical solution (1/2) when the number of tosses tends to infinity provided the die is fair. However, one should only judge the fairness of the die after experiments are performed.

The Bayesian view can be seen as lying amidst the classical and the frequentist: symmetry considerations matter, but variations are allowed when new experiments are performed. Before the first toss, all six sides can be reasonably considered equally likely; for each new toss, the probabilities can vary in one direction or another, according to the sequence of results obtained up to that moment. Probability is, thus, a dynamic quantity, which starts with a plausible value and becomes progressively more accurate as more evidence is acquired. In the limit, after infinitely many trials, the Bayesian and the frequentist tend to agree with each other.

To assess the differences more formally, let us explore further the role of the likelihood function on Bayesian and frequentist statistics. As stated in a previous section, the likelihood function, $p(\mathbf{x}|\boldsymbol{\theta})$, consists of the distribution of the data $\mathbf{x}$ when parameters $\boldsymbol{\theta}$ are assumed to be correct. Frequentist inference is usually based solely on $p(\mathbf{x}|\boldsymbol{\theta})$, and $\boldsymbol{\theta}$ is understood as the set of parameters that specify the distribution of $\mathbf{x}$ and not as a conditioning variable. In contrast, Bayesians use the likelihood as the link between the prior and the posterior distributions — via Bayes' theorem as shown in the previous section — and $\boldsymbol{\theta}$ is seen as a variable that helps specifying the distribution of the data $\mathbf{x}$.

The criteria of Maximum Likelihood (ML) [1] is one of the most popular tools in the frequentist repertoire. The idea behind the ML estimate is intuitively appealing: the estimated parameters are those that make the observed data more likely to be obtained. In practice, it is obtained by finding the value of $\boldsymbol{\theta}$ that maximizes the likelihood function.

The Bayesian school also allows for point estimates akin to ML, but the many possible criteria are based on the posterior distribution, not the likelihood. The Bayesian Minimal Squared Error (BMSE) and the Maximum a Posteriori (MAP) [3] [2] are some of the most popular choices. The BMSE is based on minimizing the expected squared error of the estimate, which results in the expected value of the parameter posterior distribution. The MAP results from simply maximizing the posterior distribution, and it is a consequence of minimizing a hit-or-miss cost function.

Next sections illustrate these estimators for a simple and didactic scenario of coin toss and for the linear model, which is widespread in engineering problems.

1) Coin Toss: One wants to estimate the probability θ of heads being obtained in a coin toss. We have seen previously that symmetry considerations are controversial, despite their intuitive appeal. A frequentist would argue that experiments ought to be performed before a meaningful answer can be given. He then tosses the coin three times and results are, say, tails in the three cases. He proceeds to calculate the likelihood $p(\mathbf{x}|\theta)$, that is, the probability that the result $\mathbf{x} = [1\ 1\ 1]$ — representing three tails — will be obtained considering all possible values of $\theta \in [0\ 1]$. Assuming the results to be mutually independent, the probability of three tails is the product of each individual probability, $p(\mathbf{x}|\theta) = (1-\theta)^3$. Since the likelihood is maximized when $\theta = 0$, the frequentist ML criteria produces the estimate $\theta^{\mathrm{ML}} = 0$.

Challenged by the same problem, a Bayesian statistician would start by defining a prior distribution $p(\theta)$ that quantifies the previous knowledge about the experiment. It is reasonable to assume that θ has higher chances of being around 1/2 than in the extremes (0 or 1). One possibility would be the curve in blue in Fig. 2(a). By combining this prior with the likelihood via Bayes' theorem, the Bayesian would obtain the posterior (green in the figure), whose maximization yields a MAP estimate for θ on the left of 1/2. If more experiments could be performed, more informative posteriors would be obtained (Fig 2(b)), resulting in more accurate estimates. Accordingly, the ML estimate would become more accurate with more data, and the ML and MAP estimates would be typically more similar.

2) Linear Model: The so-called linear model is commonly used in many electrical engineering problems [2], including in telecommunications. It describes a situation in which the observed data have a linear relation with the set of parameters, and the measurement is corrupted with additive noise. The model can be used, for instance, to describe a signal composed by sinusoids with unknown characteristics added to noise, or to represent the received signal in a wireless communications system with multi-path and additive noise.

More formally, the N-elements vector containing the observed data, $\mathbf{x}$, is the sum of a product of matrices product

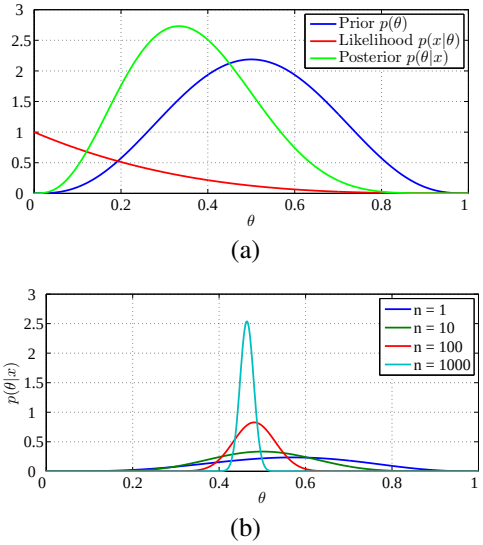

Fig. 2. Calculating the probability of 'head' in a coin toss experiment. (a) Example of a prior distribution, a likelihood function and a posterior distribution. (b) The posterior distribution gets progressively more concentrated around the true value (1/2) when the number of experiments n increases.

and noise $\mathbf{v}$:

$$\mathbf{x} = \mathbf{G}\boldsymbol{\theta} + \mathbf{v} \tag{10}$$

where the product $\mathbf{G}\boldsymbol{\theta}$ determines the linear relation between the data $\mathbf{x}$ and the unknown parameters $\boldsymbol{\theta}$. In many applications, $\mathbf{v}$ can be approximated as white Gaussian noise of zero mean and variance σ_v^2 (often unknown). If the matrix $\mathbf{G}$ is known, the distribution of $\mathbf{x}$ given $\boldsymbol{\theta}$ equals the distribution of $\mathbf{v}$ with mean shifted by $\mathbf{G}\boldsymbol{\theta}$. Hence, the likelihood is:

$$p(\mathbf{x}|\boldsymbol{\theta}) = p_v(\mathbf{x} - \mathbf{G}\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma_v^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_v^2}(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})\right\}. \tag{11}$$

The maximization of this function with respect to $\boldsymbol{\theta}$ is equivalent to minimizing the argument of the exponential. By making the gradient of the argument equal to zero, we obtain the ML estimate:

$$\boldsymbol{\theta}^{\mathrm{ML}} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{x}. \tag{12}$$

This problem would be solved differently in the Bayesian school. The parameters $\boldsymbol{\theta}$ would be treated as random variables with associated prior distribution, reflecting the previous knowledge about the parameters. Considering, for the sake of illustration, a Gaussian prior with mean $\mathbf{m}_{\boldsymbol{\theta}}$ and covariance matrix $\mathbf{C}_{\boldsymbol{\theta}}$, we would have a posterior whose maximization would generate the following estimator for $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{\mathrm{MAP}} = (\mathbf{G}^T\mathbf{G} + \sigma_v^2\mathbf{C}_{\boldsymbol{\theta}}^{-1})^{-1}(\mathbf{G}^T\mathbf{x} + \sigma_v^2\mathbf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{m}_{\boldsymbol{\theta}}). \tag{13}$$

If the elements of $\mathbf{C}_{\boldsymbol{\theta}}$ in the above formula are large, the terms introduced by the prior distribution would be of little relevance, and the MAP solution would approximate the ML. This is plausible, since the large values for $\mathbf{C}_{\boldsymbol{\theta}}$ reflect a vague prior, which implies a larger importance for the observed data in the final result. On the other hand, if a large quantity of data

is obtained, and thus $\mathbf{x}$ is a high dimensional vector, the terms depending on $\mathbf{G}$ and $\mathbf{x}$ would dominate the terms depending on the prior parameters, which are fixed. In this case, with abundance of data, the MAP estimator would approximate the ML, regardless of the prior.

These two examples allow us to conclude that the Bayesian and frequentist schools tend to agree with each other when a lot of data is available. Nevertheless, the schools can wildly disagree when data are scarce and previous knowledge is relevant.

IV. Elements of Bayesian Inference

The following sections present elements that are unique to the Bayesian school and are present in all major textbooks on Bayesian statistics [3] [2].

A. Hierarchical Bayesian Model

In many practical problems in engineering, the observed data depend on many parameters hierarchically related to each other. In wireless communications, for example, the received signal depends on the channel, which can be characterized by a linear filter with unknown coefficients. In this context, a Bayesian would assign a probabilistic model to the channel coefficients, and since the parameters of this model are unknown, the statistician would describe then in a probabilistic fashion in terms of other parameters — which would then be called hyperparameters.

Figure 3 illustrates this concept more generally. The possibly multivariate data sets x_1, x_2, ..., x_n, depend on the parameters θ_1, θ_2, ..., θ_n, respectively, which, in their turn, are instances of a random variable described by the hyperparameter ϕ. This model could represent, for example, the grades of students from n different schools, each one with a certain mean θ_i $i \in \{1, \ldots, n\}$, which depends on a more fundamental distribution that can be written in terms of ϕ (possibly the general mean of the students in that region).

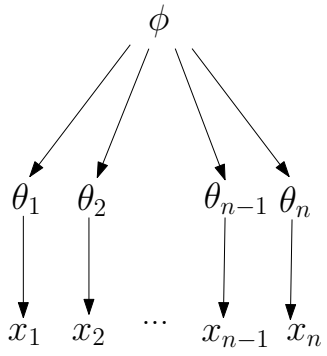

Fig. 3. Example of Bayesian hierarchical model. The data set x_1, x_2, ..., x_n, depends on parameters θ_1, θ_2, ..., θ_n, respectively, which are instances of random variables described by the parameter ϕ.

Bayes' theorem yields the posterior distribution of the unknown parameters as a function of the observed data. In this example, the posterior distribution is denoted by $p(\theta|x_1, \ldots, x_n)$, and the prior distribution of θ depends on ϕ, which is characterized by its prior $p(\phi)$. Hence,

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi})}{p(x_1, \ldots, x_n)}, \quad (14)$$

in which we used $p(\theta, \phi) = p(\theta|\phi)p(\phi)$.

B. Parameter Elimination

It is common for the set of parameters to contain elements that are not of interest for estimation. When modeling a communication system, the channel coefficients are necessary for a complete model of the transmission process; but the main goal is the estimation of the transmitted data. Because it treats the parameters as random variables, the Bayesian school authorizes us to work with some variables of interest, while integrating out the so-called nuisance parameters. By partitioning $\boldsymbol{\theta}$ between the nuisance parameters and those one wishes to estimate, in such a way that $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\psi})$, the posterior for the parameters of interest can be calculated as:

$$p(\boldsymbol{\phi}) = \int_{\psi} p(\boldsymbol{\phi}, \boldsymbol{\psi}|\mathbf{x}) d\boldsymbol{\psi}. \quad (15)$$

This integral cannot always be evaluated analytically or with classical numerical integration techniques. The difficulty is especially severe when the distribution is multivariate and does not have a known form. In these cases, the algorithms based on MCMC to be discussed later are especially recommended.

C. Prior Distribution

The prior distribution quantifies the knowledge about the possible values of the parameters before data have been acquired. Even though frequentist criteria can be applied to guide the choice of $p(\boldsymbol{\theta})$, the knowledge from the subjective experience of the designer is usually the main factor in determining the choice of the prior distribution. Techniques for helping the designer in this task include the histogram method, the distribution function method and the method of relative likelihood [3]. Although they are useful, these techniques can yield to overly complicated distributions that are difficult to handle analytically, thus hindering the later inference procedure.

1) Conjugate Priors: To assure the tractability of the posterior distribution, a popular strategy is to adopt priors with the same algebraic structure of the likelihood, which guarantees that the product between them, from which the posterior depends, have a form that is easy to manipulate. For instance, if the likelihood is Gaussian, the choice of a Gaussian prior would produce a Gaussian posterior, which is easy to analyze and sample from.

Let us return to the linear model of Sec. III-2. Since the noise is Gaussian, the likelihood is Gaussian too. By assuming a Gaussian prior for $\boldsymbol{\theta}$ and covariance matrix $\mathbf{C}_\theta$, the posterior would then be the product of Gaussians:

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p_v(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\mathbf{G}\boldsymbol{\theta}, \sigma_v^2\mathbf{I})\mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_\theta, \mathbf{C}_\theta), \quad (16)$$

which results in a Gaussian with modified parameters:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \mathcal{N}\left(\boldsymbol{\theta}\,|\boldsymbol{\mu}_p, \mathbf{C}_p\right), \tag{17}$$

where

$$\boldsymbol{\mu}_p = (\mathbf{G}^T\mathbf{G} + \sigma_v^2\mathbf{C}_{\boldsymbol{\theta}}^{-1})^{-1}(\mathbf{G}^T\mathbf{x} + \sigma_v^2\mathbf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{m}_{\boldsymbol{\theta}}), \tag{18}$$

and

$$\mathbf{C}_p = \left(\frac{\mathbf{G}^T\mathbf{G}}{\sigma_v^2} + \mathbf{C}_\theta^{-1}\right)^{-1}. \tag{19}$$

If the noise variance σ_v^2 in this model is unknown, we could describe it statistically. By looking at the expression of the likelihood, we see that σ_v^2 appears in a form that resembles the inverse gamma distribution, which has a positive support region and is defined by parameters α and β:

$$p(x|\alpha,\beta) = \mathcal{IG}(x|\alpha,\beta) = \frac{\beta^\alpha}{\tau(\alpha)}x^{-\alpha-1}\exp\left(\frac{-\beta}{x}\right). \tag{20}$$

Thus, the conjugate prior for σ_v^2 is an inverted gamma $p(\sigma_v^2|\alpha_v,\beta_v)$. With this prior, the posterior is also an inverted-gamma with modified parameters.

$$p(\sigma_v^2|\mathbf{x},\boldsymbol{\theta}) = \mathcal{IG}\left(\sigma_v^2\left|\alpha_v + \frac{N}{2}, \beta_v + \frac{(\mathbf{x}-\mathbf{G}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{G}\boldsymbol{\theta})}{2}\right.\right), \tag{21}$$

where N is the number of elements in $\mathbf{x}$.

2) Non-informative Prior: As the name suggests, a non-informative prior is a distribution that ideally does not contain any information about the possible values of the parameters. The apparently more obvious choice for a non-informative prior would be a uniform distribution associating equal probability to all values in the parameter sample space. Convenient for discrete parameters, this choice is problematic for continuous parameters, since the uniform distribution is not invariant to one-to-one transformation. In other words, if $\boldsymbol{\theta}$ is uniform, the distribution of $\boldsymbol{\phi} = f(\boldsymbol{\theta})$ is no longer uniform. However, the absence of knowledge about $\boldsymbol{\theta}$ should imply equal ignorance about any variable obtained by some transformation of $\boldsymbol{\theta}$, and thus the distribution of $\boldsymbol{\phi}$ should be uniform as well.

In order to circumvent this inconsistency of the uniform prior, Jeffreys [5] introduced a class of distributions that are very vague and invariant to one-to-one transformations. The Jeffreys's prior associated to the variable $\boldsymbol{\theta}$ is given by:

$$p(\boldsymbol{\theta}) \propto |\mathbf{I}(\boldsymbol{\theta})|^{1/2}, \tag{22}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher's information matrix [1] of $\boldsymbol{\theta}$, defined as:

$$I(\boldsymbol{\theta}) = E_{\mathbf{X}|\boldsymbol{\theta}}\left[-\frac{\partial^2 \ln\left(p(\mathbf{x}|\boldsymbol{\theta})\right)}{\partial\boldsymbol{\theta}^2}\right], \tag{23}$$

where $E_{\mathbf{X}|\boldsymbol{\theta}}$ denotes the expected value of the variable $\mathbf{X}|\boldsymbol{\theta}$. Note that the Jeffrey's prior depends only on the likelihood. In the case of models that are invariant with respect to the scale, such as the standard-deviation of the Gaussian in the linear model, we obtain the following Jeffrey's prior:

$$p(\sigma_v) \propto \frac{1}{\sigma_v}. \tag{24}$$

There are other possible choices for non-informative priors. Jaynes [19] proposes the criterion of maximum entropy to specify of the prior distribution. Bernardo [20] also argues that the prior distribution should be chosen based on information-theoretic criteria. In most cases, the resulting distribution is improper — i.e., its integral is infinity — like in Eq. (24). This is not always inconvenient, since in many cases the resulting posterior is proper.

V. Markov Chain Monte Carlo (MCMC)

Even though it is possible to solve practical Bayesian inference problems with the concepts seen so far, in applications in which the distributions are multivariate and multimodal those tools are inadequate. Often the realistic modeling of physical systems requires the usage of multi-level hierarchical models, yielding to analytically intractable distributions. Situations like these require more sophisticated numerical techniques, the Expectation-Maximization (EM) algorithm [21], the Gibbs sampling [9] and the Metropolis-Hastings (MH) algorithm [22] [23] being some of the most prominent examples. The latter two are based on Markov Chain Monte Carlo (MCMC) [8], a class of methods that consists of designing a Markov chain to generate samples from a given target distribution, which are later used to perform inferences via Monte Carlo techniques.

A. Monte Carlo Methods

Monte Carlo techniques were introduced in the context of the Manhattan project by a group of researchers that included Enrico Fermi, Stanislaw Ulam, John Von Neumann and Nicholas Metropolis [24], [25]. The puzzling name is an allusion to the Monte Carlo casino in Monaco where Stanislaw Ulam's uncle used to gamble, and it is related to the method being dependent on random numbers (like dice in a casino).

Generally, Monte Carlo methods consist of generating i.i.d. (independent and identically distributed) samples from a given distribution, so that they can be used to obtain an approximation of a distribution characteristic. Having a set of samples $X = \{x^{(1)}, \ldots, x^{(N)}\}$ from a certain distribution $p(x)$, we can approximate an integral such as

$$I(f) = \int f(x)p(x)dx \tag{25}$$

by the following summation

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}). \tag{26}$$

It is possible to show that the estimator above is unbiased and, by the law of large numbers, that it converges to the integral of equation (25) almost surely (that is, with probability 1) when N approaches infinity.

B. Markov Chain

The sampling techniques based on Markov Chains allow for indirect sampling when the target distribution cannot be readily sampled from. We start the exposition considering discrete

Markov Chains and then we present the extension of some result to continuous state-space.

A Markov Chain is a discrete-time random process in which future states of the chain do not depend on past states as long as the present state is known — a condition known as Markov property [26]. More formally, let $X^{(n)}$ be the random variable representing the chain state in time n and let S be the state-space for variables $X^{(n)}$. Hence:

$$P(X^{(n)} \in A^{(n)}|X^{(n-1)} \in A^{(n-1)}, \ldots, X^{(0)} \in A^{(0)}) = P(X^{(n)} \in A^{(n)}|X^{(n-1)} \in A^{(n-1)}), \quad (27)$$

for any $A^{(0)}, \ldots, A^{(n)} \in S$. The above equation was written in terms of probabilities rather than densities in order to cover both discrete and continuous state-spaces. If S is a countable set, the Markov chain is said to be discrete. Let us consider initially a state space $S = \{s_1, \ldots, s_N\}$. The transition probabilities from one state to the other, in time n, define the *transition matrix* $\mathbf{T}_n$, whose element of line i and column j is given by:

$$T_n(i|j) = P(X^{(n)} = s_i|X^{(n-1)} = s_j). \quad (28)$$

Since each column in this matrix contains the probability associated with each one of the elements in the sampling space, the elements of each column sum to 1. Matrices with this property are called *stochastic matrices*, which have properties that help the analysis of MCMC algorithms. One such property is the existence of at least one eigenvalue equal to 1. Furthermore, if all elements of $\mathbf{T}_n$ are positive — when it is always possible to move from a given state to any other state —, it is possible to show that the remaining eigenvalues are distinct and lower than 1.

The probability distribution of the chain's state on time n, denoted $P_n(i)$, $i = \{1, \ldots, N\}$, can be obtained from the distribution on time $(n-1)$ through:

$$P_n(i) = \sum_{j=1}^{N} T_n(i|j) P_{n-1}(j), \quad (29)$$

which can be written in vectorial form:

$$\mathbf{P}_n = \mathbf{T}_n \mathbf{P}_{n-1}, \quad (30)$$

in which $\mathbf{P}_n = [P_n(1) \ldots P_n(N)]^T$.

If $\mathbf{T}_n$ does not depend on n, the chain is said to be *homogeneous* and the transition matrix is denoted simply by $\mathbf{T}$. In this case, applying Eq. (30) n times yields:

$$\mathbf{P}_n = \mathbf{T}^n \mathbf{P}_0, \quad (31)$$

where $\mathbf{P}_0$ is the distribution of the initial state of the chain.

In designing MCMC algorithms, the chain should have the two above properties [8] [27].

- Irreducibility: starting from any state, there should exist a non-zero probability that the chain will move to any other state in a finite number of steps. This is equivalent to having $T^n(i|j) > 0$ for some n.
- Aperiodicity: the chain is not trapped in cycles.

1) Invariant Distribution: A distribution is said to be invariant (or stationary) if it remains fixed under the application of the transition matrix. In designing MCMC algorithms, we are interested in building a Markov Chain that produces a given invariant distribution. By denoting the invariant distribution as $\pi(i)$, we should have:

$$\pi(i) = \sum_{j=1}^{N} T(i|j)\pi(j). \quad (32)$$

This equation can be written in a vectorial form as $\boldsymbol{\pi} = \mathbf{T}\boldsymbol{\pi}$, from which we see that $\boldsymbol{\pi}$ is an eigenvector associated with the eigenvalue $\lambda = 1$. In order to determine $\boldsymbol{\pi}$ uniquely, we make the sum of its elements equal to 1.

In an MCMC algorithm, it is usual to force the *detailed balance* condition to assure a given distribution to be invariant. This condition states that the probability of the chain moving from a state s_i in time $(n-1)$ to state s_j in time (n) is equal to the probability that the inverse transition will occur, that is:

$$\pi(j)T(i|j) = \pi(i)T(j|i). \quad (33)$$

In order to see that $\pi(i)$ is the invariant distribution, it suffices to sum the two sides of the above equality for all possible values of j, and to realize that the result is Eq. (32). Despite being more restrictive than Eq. (32), this condition is simpler to impose for MCMC algorithms.

2) Ergodicity: Besides guaranteeing that $\pi(i)$ is the desired invariant distribution, we should assure that $P_n(i)$ converges to $\pi(i)$ when n tends to infinity, regardless of the initial distribution. In this case, we say that $\pi(i)$ is the limit distribution of the chain.

This property is known as *ergodicity*. In order for a Markov chain to be ergodic, the chain needs to be aperiodic and irreducible. In the discrete case, it suffices that the eigenvalues of $\mathbf{T}$ be all distinct, which allows for probability distribution of the initial state to be written using the eigenvectors of $\mathbf{T}$ as a basis:

$$\mathbf{P}_0 = \boldsymbol{\pi} + c_2\mathbf{v}_2 + \ldots + c_N\mathbf{v}_N, \quad (34)$$

where we used the fact that $\boldsymbol{\pi}$ is the eigenvector associated with the eigenvalue $\lambda = 1$ for any stochastic matrix $\mathbf{T}$. In the above equation $\mathbf{v}_i$, $i \in \{2, \ldots, N\}$, are the remaining eigenvectors, and c_i, $i \in \{2, \ldots, N\}$, are coefficients that specify the vector $\mathbf{P}_0$ in the basis formed by the eigenvectors.

The expression for the state distribution in instant n can be thus obtained as [28]:

$$\mathbf{P}_n = \mathbf{T}^n\mathbf{P}_0 = \boldsymbol{\pi} + c_2\lambda_2^n\mathbf{v}_2 + \ldots + c_N\lambda_N^n\mathbf{v}_N, \quad (35)$$

where $\{\lambda_2, \ldots, \lambda_N\}$ together with $\lambda_1 = 1$ form the set of eigenvalues of matrix T. Since all the eigenvectors but the first are lower than 1, it follows that P_n will approach $\boldsymbol{\pi}$ as n tends to infinity, which means that the chain will converge to the desired invariant distribution.

3) Markov Chain for Continuous State Space: For continuous state spaces, the properties described in the above sections are expressed through probability density functions. The Markov property is defined as:

$$p(x^{(n)}|x^{(n-1)},\ldots,x^{(0)}) = p(x^{(n)}|x^{(n-1)}). \quad (36)$$

The *transition kernel* $K_n(x|y)$ in time n is defined as:

$$K_n(x|y) = p_{X^{(n)}}(x|X^{(n-1)} = y), \quad (37)$$

where $p_{X^{(n)}}(x)$ denotes the probability density of the random variable $X^{(n)}$.

Thus, the distribution of the chain in time n is given by:

$$p_n(x) = \int_{y\in S} K_n(x|y)p_{n-1}(y)dy. \quad (38)$$

If the chain is homogeneous, K_n does not depend on n and the detailed balance condition becomes:

$$\int_A\int_B K(x|y)\pi(y)dydx = \int_B\int_A K(y|x)\pi(x)dxdy, \quad (39)$$

for any set of A and B belonging to S.

As in the discrete case, the *detailed balance* suffices to make $\pi(i)$ the invariant distribution of the chain defined by $K(i|j)$. To guarantee that the invariant distribution is also the limit distribution, the chain should be aperiodic and irreducible.

VI. Gibbs Sampler

The Gibbs Sampler [9] is an MCMC algorithm recommended for cases in which the joint distribution is harder to sample from than the conditional distributions. The method was proposed in 1984 in the context of image restoration [9], and brought to the statistical community in 1990 by Gelfand [29].

In a nutshell, the method consists of partitioning a multidimensional variable into several (possibly multivariate) components and drawing samples from conditional distributions of each component when the others remain fixed. The process is repeated using the last sampled values of each component as conditionals for the distribution of the other components. More formally, let $\pi(\boldsymbol{\theta})$ be the joint distribution from which one desires to obtain samples. Then, variable $\boldsymbol{\theta}$ is partitioned in k components, such that $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k\}$. The Gibbs sampler algorithm requires the specification of initial values for each component $\boldsymbol{\theta}_k$ before the iterative conditional sampling is performed. The whole process is described in Algorithm 1, where the symbol $\sim$ indicates that the variable on the left side is a sample from the distribution on the right side.

Algorithm 1 Gibbs sampler algorithm.

1: Initialization: Generate $\boldsymbol{\theta}_1^{(0)}, \boldsymbol{\theta}_2^{(0)} \ldots, \boldsymbol{\theta}_k^{(0)}$;
2: **for** $i = 1$ to N_{it} **do**
3: $\quad \boldsymbol{\theta}_1^{(i)} \sim \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(i-1)},\ldots,\boldsymbol{\theta}_k^{(i-1)})$
4: $\quad \boldsymbol{\theta}_2^{(i)} \sim \pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(i)},\ldots,\boldsymbol{\theta}_k^{(i-1)})$
5: $\quad \vdots$
6: $\quad \boldsymbol{\theta}_k^{(i)} \sim \pi(\boldsymbol{\theta}_k|\boldsymbol{\theta}_1^{(i)},\ldots,\boldsymbol{\theta}_{k-1}^{(i)})$
7: **end for**

To verify that $\pi(\boldsymbol{\theta})$ is an invariant distribution in each operation inside the loop of the Gibbs sampler, we calculate the distribution of $\boldsymbol{\theta}$ after the first iteration. Assuming that the Markov Chain distribution in the end of iteration $(i-1)$ is $\pi(\boldsymbol{\theta})$, that is, $p(\boldsymbol{\theta}^{(i-1)}) = \pi(\boldsymbol{\theta}^{(i-1)})$, we obtain:

$$\begin{aligned} p(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i-1)},\ldots,\boldsymbol{\theta}_k^{(i-1)}) = \\ p(\boldsymbol{\theta}_1^{(i)}|\boldsymbol{\theta}_2^{(i-1)},\ldots,\boldsymbol{\theta}_k^{(i-1)})p(\boldsymbol{\theta}_2^{(i-1)},\ldots,\boldsymbol{\theta}_k^{(i-1)}) = \\ \pi(\boldsymbol{\theta}_1^{(i)}|\boldsymbol{\theta}_2^{(i-1)},\ldots,\boldsymbol{\theta}_k^{(i-1)})\pi(\boldsymbol{\theta}_2^{(i-1)},\ldots,\boldsymbol{\theta}_k^{(i-1)}) = \\ \pi(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i-1)},\ldots,\boldsymbol{\theta}_k^{(i-1)}). \end{aligned} \quad (40)$$

Thus, the transition produced by the Gibbs sampler preserves the distribution of $\boldsymbol{\theta}$. The same logic applied to the following operations implies that, at each sampling, the resultant distribution remains equal to $\pi(\boldsymbol{\theta})$. Under some mild conditions [8], it can be shown that the chain is ergodic — i.e., it converges to the invariant distribution $\pi(\boldsymbol{\theta})$ regardless of how it is initialized.

A. Practical Issues

While the choice of how to partition the parameters set $\boldsymbol{\theta}$ does not affect the long-term properties of the Gibbs sampler, in practice it can severely influence the time needed for convergence. To prevent the so-called slow-mixing of the chain, the general advice is to avoid partitioning $\boldsymbol{\theta}$ in such a way that $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are highly correlated [8].

Another important issue is how to select the samples to perform Monte Carlo inference. Samples taken sequentially after convergence (see Sec. VIII) may exhibit a high degree of dependence, which can introduce bias in the Monte Carlo inference. In order to prevent this problem, several strategies can be implemented. One possibility is to perform many parallel realizations of the Gibbs sampler starting from different points, and then select the m-th sample of each chain to perform inference, where m is any iteration after convergence. Another possibility which is less computationally demanding is to use a single chain but take only every k-th sample of the chain after convergence, where k is large enough to guarantee low correlation between the samples used for inference.

Finally, the number of iterations of the algorithm (N_{it}) should be specified according to the desired accuracy in the Monte Carlo estimation. There is a trade-off between accuracy and resource usage since a highly accurate estimate would require a large N_{it}, implying higher processing time and increased storage space.

VII. Metropolis-Hastings Algorithm

The conditional distribution needed for the Gibbs sampling is not always easy to obtain. In those cases, the Metropolis-Hastings (MH) algorithm, proposed by Nicholas Metropolis and collaborators in 1953 [22] and generalized by Hastings in 1970 [23], is a convenient choice. The idea of the algorithm is to initially draw samples from an auxiliary distribution that is simpler than the target distribution, and later to decide to accept (or reject) this sample according to a probabilistic criteria.

More specifically, according to the MH algorithm, a sample $\boldsymbol{\theta}^*$ is first obtained from a proposal distribution, denoted by $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})$, in which $\boldsymbol{\theta}^{(i)}$ is the current state of the Markov chain. This sample is accepted with probability α given by:

$$\alpha(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^*) = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{(i)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)})}\right). \quad (41)$$

If the generated sample is accepted, the new state of the chain is $\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^*$; otherwise, the chain remains on its current state — i.e., $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}$. The transition kernel is given by:

$$K(\boldsymbol{\theta}^{(i+1)}|\boldsymbol{\theta}^{(i)}) = q(\boldsymbol{\theta}^{(i+1)}|\boldsymbol{\theta}^{(i)})\alpha(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(i+1)}) + \delta_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{\theta}^{(i+1)})r(\boldsymbol{\theta}^{(i)}), \quad (42)$$

where

$$r(\boldsymbol{\theta}^{(i)}) = \int_{\boldsymbol{\theta}^* \in S} q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)}) \left(1 - \alpha(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^*)\right) d\boldsymbol{\theta}^* \quad (43)$$

is the probability that the chain will stay in its current state. The term $\delta_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{\theta}^{(i+1)})$ denotes the Dirac delta function with energy concentrated at the point $\boldsymbol{\theta}^{(i)}$. This term implies that the distribution of $\boldsymbol{\theta}^{(i+1)}$ has significant probability mass at $\boldsymbol{\theta}^{(i)}$.

It is possible to show that the transition kernel satisfies the detailed balance condition, and thus that $\pi(\boldsymbol{\theta})$ is the invariant distribution of the chain. In order to guarantee that $\pi(\boldsymbol{\theta})$ is also the limit distribution, we have to check the irredutibility and aperiodicity of the chain. Since the algorithm always allows for rejection, it follows that the chain is aperiodic. To ensure irreduciblity, it is necessary that the support region of $q(\cdot)$ includes the support region of $\pi(\cdot)$ [27].

Despite the MH algorithm having appeared a few decades earlier than the Gibbs sampling, the former can be seen as a more general version of the latter. Indeed, the MH reduces to the Gibbs sampling when the proposal distribution is made equal to the posterior distribution itself, a case in which the acceptance probability of Eq. (41) would be equal to 1. While, in principle, both algorithms can be applied to the same problem, in practice one algorithm is often more convenient than the other. When directly sampling of full-conditional distribution is easy, the Gibbs sampling is generally preferred; otherwise, the MH with a convenient proposal is usually the best choice. The advantage of using the Gibbs sampling whenever possible is that the sample is always accepted, which tends to create less correlated sequence of samples and to yield faster convergence. Sec. XI will discuss scenarios in which both algorithms can be mixed in a single receiver.

A. Practical Issues

In addition to those issues discussed in Sec. VI-A, the efficiency of the MH algorithm depends particularly on the choice of the proposal $q(\cdot)$. A common choice for $q(\cdot)$ is a Gaussian centered at the current state — i.e., $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)}) = \mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i)}, \sigma_q^2 I)$. The choice of the variance σ_q^2 is crucial. If $q(\cdot)$ is too narrow, only the region near the mode of $\pi(\boldsymbol{\theta})$ is visited. If it is too broad, the percentage of rejected samples is too high. In both cases, the time needed for convergence would be higher than necessary and the generated samples would be highly correlated. A general rule of thumb for choosing σ_q^2 is provided in [2]. For univariate distributions the proposal variance should be chosen in such a way that approximately 44 % of samples are accepted; in higher dimensions, a lower acceptance rate of 23 % should be aimed. The user can fine tune the proposal parameters until the desired acceptance rate is obtained.

VIII. Assessing Convergence

The algorithms presented so far guarantee that the states of the Markov chain will be distributed according to the target distribution when the number of iterations tends to infinity. Since in practice only a finite number of iterations can be performed, it is important to assess when the chain is sufficiently close to the limit distribution in order for its samples to be used for Monte Carlo estimation. For a discrete state space, it can be shown that the chain converges geometrically, and the rate of convergence depends on the second largest eigenvalue, in accordance with the exposition in Sec. V-B2. For continuous state spaces, the convergence is geometric as well, but the rate of convergence is not expressed in simple terms. In both cases estimating in practice the rate of convergence in an analytic fashion is difficult, and so empirical and informal methods for convergence assessment, which consist of analyzing the data generated by the chain, are preferred.

One possible method is based on running n independent parallel chains. The histogram obtained from n samples at the m-th iteration of each chain is compared with the histogram obtained k iterations later; if the histograms are sufficiently similar, it is decided that convergence occurred. The value of k should be large enough to avoid the histograms appearing similar due to correlation between states of the chain. The method of comparing the histograms is flexible, with those based on the Kullback-Leibler divergence a possible choice.

Another popular method is based on one single chain and on the calculation of the ergodic average of the obtained samples. It is expected that after convergence the average would be close to a constant value. Therefore, by observing the behavior of the average, it is possible to informally assess convergence of the chain.

A third technique consists of inspecting the samples from a selected scalar variable considering different starting points. After convergence, the samples of the chain should coalesce in a certain region, regardless of the initialization. This method is implemented and discussed in Sec. X.

IX. Applications of Bayesian Tools in Wireless Communications

The remainder of the paper considers how the concepts seen so far can help us to solve the perennial problem of symbol detection in digital wireless communication systems. In such a system, the message to be transmitted is digitized, coded, modulated and transmitted through a wireless medium. The receiver captures a noisy and distorted version of the transmitted signal and processes the signal in order to estimate the transmitted message. Fig. 4 shows a block diagram

representation of a baseband equivalent model of a digital wireless communication system, in which the combined effect of the modulation/demodulation and the physical multipath channel have been encapsulated in the linear system $H(z)$. We will focus on Bayesian blind receivers, in which the detection is carried out without previous pilot-based training. It is important to point out that the discussed algorithms can be straightforwardly adapted to training-based receivers.

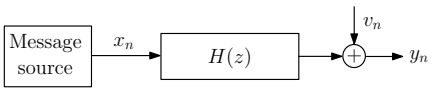

Fig. 4. Baseband equivalent model of a digital communication system with frequency selective channel. A message source generates digital symbols x_n, which are transmitted through the channel defined by the system $H(z)$ and noise v_n, yielding the received signal y_n.

The stochastic nature of the basic elements of this model, namely, the source x_n, the system $H(z)$ and the noise v_n, makes this problem well suited for a Bayesian approach. Since the exact model of a particular communication system depends on many aspects, such as the multiple access scheme, the modulation technique, the transmission environment and so on, a general-purpose receiver is hard to be designed. Instead, each possible combination of the features of the system can lead to a particular Bayesian receiver.

To highlight the most important aspects of Bayesian inference, we detail in next section the design of a Bayesian receiver for digital communication system employing BPSK (*Binary Phase Shift-Keying*) and subject to a frequency selective and Gaussian channel. The goal of this section is to familiarize the reader with the many steps that need to be taken when designing a Bayesian receiver and the results that such an approach is likely to achieve. The model and the MCMC solution described here are a slightly less general version of [30], in which the Gibbs sampling algorithm is adopted to estimate discrete (not necessarily binary) data that have been distorted by a linear system and affected by additive Gaussian noise. In Sec. XI we provide a survey of recent advances in Bayesian approaches in symbol detection for some modern communication systems, emphasizing the difficulties that typically arise in those systems in comparison to the simpler scenario of Sec. X.

X. Bayesian BPSK Symbol Detection for Frequency Selective Channels

In BPSK, phase modulation is used to transmit binary symbols x_1, x_2, ..., x_N, where $x_k \in \{+1, -1\}$. Assuming a linear and time-invariant channel, the received signal y_n is distorted and can be written as a linear combination of the transmitted signal samples, added to the channel noise:

$$y_n = \sum_{l=0}^{L-1} h_l x_{n-l} + v_n, n = 0, \ldots, N-1, \tag{44}$$

where h_l for $l \in \{1, \ldots, L\}$ are the impulse response of the channel — here modeled as real since the input x_n is real — and v_n forms an i.i.d. sequence of samples of Gaussian variables with zero mean and variance σ_v^2. This model can be conveniently written in matrix form as:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}, \tag{45}$$

where vectors $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{v}$ are formed by stacking the samples of y_n, x_n and v_n, respectively, and $\mathbf{H}$ is the channel matrix composed of elements of h_l. This equation is expanded and exemplified below for the case $N = 6$ and $L = 3$:

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} h_0 & 0 & 0 & 0 & 0 & 0 \\ h_1 & h_0 & 0 & 0 & 0 & 0 \\ h_2 & h_1 & h_0 & 0 & 0 & 0 \\ 0 & h_2 & h_1 & h_0 & 0 & 0 \\ 0 & 0 & h_2 & h_1 & h_0 & 0 \\ 0 & 0 & 0 & h_2 & h_1 & h_0 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix}. \tag{46}$$

An alternative way of expressing Eq. (44) that will be useful later is:

$$\mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{v}, \tag{47}$$

where $\mathbf{h} = [h_0 \ldots h_L]^T$ and $\mathbf{X}$ is a matrix defined from elements of x_n. Similarly to the previous case, the expanded form of this equation is given by:

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} x_0 & 0 & 0 \\ x_1 & x_0 & 0 \\ x_2 & x_1 & x_0 \\ x_3 & x_2 & x_1 \\ x_4 & x_3 & x_2 \\ x_5 & x_4 & x_3 \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ h_2 \end{bmatrix} + \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix}. \tag{48}$$

The main simplification in this model is the assumption of a time-invariant channel. In practice it is well known that multipath channels may change over time because of the movement of the transmitter or receiver as well as the movement of the objects in the environment. If these movements are slow and the symbol rate is not too high (a case in which the coherence time T_C is much larger than the symbol period T_S [16]) then the channel can be considered fixed over a certain period of time in which symbol detection takes place. When this is not the case, the time-invariant assumption would be a poor one, and sequential methods such as Particle Filtering should be preferred (see Sec. XI-G). In addition to this time-invariance assumption, the model disregards nonlinear effects that in practice might exist because of power amplifiers (see Sec. XI-E), and it takes into account as disturbances only the ambient noise (typically white and Gaussian), thus ignoring impulsive disturbances (see Sec. XI-D).

From the Bayesian perspective, the problem consists of estimating the transmitted sequence $\mathbf{x}$ based on the received signal $\mathbf{y}$ considering as nuisance the set of parameters $\boldsymbol{\theta} = \{\mathbf{h}, \sigma_v^2\}$. The solution to the problem begins with calculating the *a posteriori* distribution of the unknown quantities, consisting of the message symbols $\mathbf{x}$ and nuisance parameters $\boldsymbol{\theta}$. Using Bayes' theorem we have:

$$p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{49}$$

where

$$p(\mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{50}$$

and $p(\mathbf{x})$ and $p(\boldsymbol{\theta})$ are the prior distributions of transmitted symbols and nuisance parameters, respectively. If symbols +1 and -1 are equally likely, the prior for $\mathbf{x}$ is the same for all combinations of transmitted symbols. The priors for $\mathbf{h} = \{h_1, \ldots, h_L\}$ and σ_v^2 should be chosen in practice based on the channel environment (rural or urban, indoor or outdoor, etc), along with the transmission bandwidth and the expected signal-to-noise ratio. Here we adopt conjugate priors given by:

$$p(\mathbf{h}) = \mathcal{N}(\mathbf{h}|\mathbf{0}_L, \sigma_h^2\mathbf{I}_L), \tag{51}$$

$$p(\sigma_v^2) = \mathcal{IG}(\sigma_v^2|\alpha_v, \beta_v), \tag{52}$$

where σ_h^2 is the variance of each channel tap, and α_v and β_v define the shape of the inverse-gamma distribution.

Besides the priors, the likelihood function is needed to specify the posterior distribution. Exploring the similarity between the wireless channel model of Eq. (45) and the linear model discussed in Section III-2, the likelihood can be calculated as:

$$p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) = p_v\left(\mathbf{y} - \mathbf{H}\mathbf{x}\right), \tag{53}$$

$$p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma_v^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma_v^2}(\mathbf{y}-\mathbf{H}\mathbf{x})^T(\mathbf{y}-\mathbf{H}\mathbf{x})\right\}. \tag{54}$$

A posterior is obtained by replacing the prior and the likelihood in Eq. (49). This is only the first step of the process. The second step is the integration of the nuisance parameters so as to obtain the conditional distribution of the transmitted data with respect to received data:

$$p(\mathbf{x}|\mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{x},\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}. \tag{55}$$

Even in this simplified situation with time-invariant channel and white Gaussian noise, the above integral cannot be solved analytically, and traditional methods of numerical integration are inadequate due to the high dimension of $\boldsymbol{\theta}$. Another difficulty is the estimation of the most likely data sequence: the naive solution, an exhaustive search for all 2^N possible sequences, is clearly impractical for typical values of N.

A. Bayesian MCMC Receiver

MCMC methods allow us to overcome these two difficulties. By sampling the variables from their conditional distributions, the Gibbs sampler is capable of numerically performing the above integral. Moreover, if the transmitted message $\mathbf{x}$ is sampled in a sequential fashion, as explained below, the Gibbs sampling algorithm will have a dramatically lower computational cost in comparison with the exhaustive search solution.

The MCMC procedure is summarized in Algorithm 2, where symbol $\sim$ denotes that the left side variable is a sample from the distribution on the right side. After initialization, the algorithm sequentially samples each unknown variable $\mathbf{x}$, $\mathbf{h}$ and σ_v^2. Each step is detailed next.

Algorithm 2 MCMC algorithm for signal detection in a frequency selective communication channel.

1: Initialization: generate $\mathbf{x}^{(0)}$, $\mathbf{h}^{(0)}$ and $\sigma_v^{2(0)}$;
2: **for** $i = 1$ to I **do**
3: **for** $j = 1$ to b **do**
4: $\mathbf{x}_j^{(i+1)} \sim p(\mathbf{x}_j|\mathbf{y}, [\mathbf{x}_{1:j-1}^{(i+1)}\ \mathbf{x}_{j+1:b}^{(i)}], \mathbf{h}^{(i)}, \sigma_v^{2(i)})$
5: **end for**
6: $\mathbf{h}^{(i+1)} \sim p(\mathbf{h}|\mathbf{y}, \mathbf{x}^{(i+1)}, \sigma_v^{2(i)})$
7: $\sigma_v^{2(i+1)} \sim p(\sigma_v^2|\mathbf{y}, \mathbf{x}^{(i+1)}, \mathbf{h}^{(i+1)})$
8: **end for**

Initialization: MCMC algorithms are flexible in terms of how they are initialized. If the goal is to explore the posterior distribution as fully as possible, then a random initialization using a broad distribution is recommended. Many different random initializations might be tried if one wants to investigate the convergence, using some of the techniques presented in Sec. VIII. On the other hand, if one is concerned with the time needed for convergence, then initializing the algorithm from a favorable point is highly desirable. In practice, initial crude estimates for $\mathbf{x}$, $\mathbf{h}$ and σ_v^2 might be produced by using simpler algorithms, such as the constant modulus algorithm [16], or by simply using the estimates from previous blocks in the case of block processing. In our illustrative implementation, we chose $\mathbf{x}^{(0)} = \text{sign}(\mathbf{y})$, $\mathbf{h}^{(0)} = [1\,0\,\ldots\,0]^T$ and we initialized σ_v^2 as a random sample from its prior distribution.

Sampling of $\mathbf{x}$: Directly sampling of the entire vector $\mathbf{x}$ is impractical because it would require calculating 2^N probabilities, which is prohibitively large for typical values of N. A way to alleviate this burden is to explore the flexibility of the Gibbs sampling and to perform this operation in sub-blocks of small length. The idea is to partition the binary sequence $\mathbf{x}$ of N elements in b blocks of fixed size B in such a way that $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_b\}$. Next we sequentially perform the sampling of each sub-block $\mathbf{x}_j$. When sampling sub-block $\mathbf{x}_j$, the remaining ones are considered to be known and equal to previously sampled values. Since each block contains B elements, now the sampling requires the calculation of conditional probabilities for 2^B possible sub-sequences within each block. If $N = 1024$ and $B = 4$, then $b = 256$, and the number of required computations for sampling each sub-block is $2^4 = 16$, which should be performed 256 times (one for each sub-block). This complexity is much lower than the 2^{1024} computations that would be required if the entire block had to be sampled at once. A particularly convenient choice is $B = 1$, which would lead to a scalar variable x_j being sampled at each iteration. In this case, sampling x_j at the i-th Gibbs iteration would require calculating the ratio below:

$$\frac{p(x_j = +1|\mathbf{y}, [\mathbf{x}_{1:j-1}^{(i+1)}\ \mathbf{x}_{j+1:N}^{(i)}], \mathbf{h}^{(i)}, \sigma_v^{2(i)})}{p(x_j = -1|\mathbf{y}, [\mathbf{x}_{1:j-1}^{(i+1)}\ \mathbf{x}_{j+1:N}^{(i)}], \mathbf{h}^{(i)}, \sigma_v^{2(i)})} = \frac{p(\mathbf{y}|[\mathbf{x}_{1:j-1}^{(i+1)}\ x_j = +1\ \mathbf{x}_{j+1:N}^{(i)}], \mathbf{h}^{(i)}, \sigma_v^{2(i)})}{p(\mathbf{y}|[\mathbf{x}_{1:j-1}^{(i+1)}\ x_j = -1\ \mathbf{x}_{j+1:N}^{(i)}], \mathbf{h}^{(i)}, \sigma_v^{2(i)})}, \tag{56}$$

in which the expression on the right-hand side is the well known likelihood ratio used in detection theory. In this case,

the hypotheses consist of the transmitted bit (0 or 1) at time j. The difference between the Gibbs sampling and traditional hypothesis testing is that a random sample from the hypothesis will be drawn rather than a deterministic choice of the most likely hypothesis. The calculation of the above expression can be done by using Eq. (53) with $\mathbf{x}$ replaced by $[\mathbf{x}_{1:j-1}^{(i+1)}\, x_j = +1\, \mathbf{x}_{j+1:N}^{(i)}]$ in the numerator and by $[\mathbf{x}_{1:j-1}^{(i-1)}\, x_j = -1\, \mathbf{x}_{j+1:N}^{(i)}]$ in the denominator.

Sampling of $\mathbf{h}$: In order to calculate the conditional distribution of $\mathbf{h}$ we use the equivalent channel model of Eq. (47) in which the dependence of the output $\mathbf{y}$ on the channel coefficients $\mathbf{h}$ is made explicit. The model is still an instance of the linear model of Sec. III-2, in which $\mathbf{h}$ and $\mathbf{X}$ play the roles of $\boldsymbol{\theta}$ and $\mathbf{G}$, respectively. Thus, the full-conditional distribution of $\mathbf{h}$ can be written as:

$$p(\mathbf{h}|\mathbf{y}, \mathbf{x}, \sigma_v^2) \propto p_v(\mathbf{y} - \mathbf{X}\mathbf{h}) p(\mathbf{h}) = \mathcal{N}(\mathbf{h}|\mathbf{m}_h, \mathbf{C}_h), \quad (57)$$

where

$$\mathbf{m}_h = \left(\mathbf{X}^T\mathbf{X} + \sigma_v^2 \mathbf{C}_h\right)^{-1} \mathbf{X}^T \mathbf{y} \quad (58)$$

and

$$\mathbf{C}_h = \left(\frac{\mathbf{X}^T\mathbf{X}}{\sigma_v^2} + \mathbf{C}_h^{-1}\right)^{-1}. \quad (59)$$

Sampling of σ_v^2 In the discussion about the parameter elimination (Section IV-B) we saw that an inverted-gamma prior for the variance σ_v^2 yields a posterior of the same shape, but with modified parameters. Since when sampling σ_v^2 the remaining parameters $\mathbf{h}$ and $\mathbf{x}$ are known we can use Eq. (21) to calculate the conditional distribution of σ_v^2. It is necessary to replace $\mathbf{G}$ with $\mathbf{X}$, $\boldsymbol{\theta}$ with $\mathbf{h}$ and $\mathbf{x}$ with $\mathbf{y}$, in order to obtain the full-conditional distribution for σ_v^2:

$$p(\sigma_v^2|\mathbf{y}, \mathbf{x}, \mathbf{h}) = \mathcal{IG}\left(\sigma_v^2 \middle| \alpha_v + \frac{N}{2}, \beta_v + \frac{(\mathbf{y} - \mathbf{H}\mathbf{x})^T(\mathbf{y} - \mathbf{H}\mathbf{x})}{2}\right), \quad (60)$$

B. Simulations

To illustrate the MCMC receiver, we chose a channel of length $L = 11$ with impulse response shown in Fig. 5 and noise variance $\sigma_v^2 = 0.1$, corresponding to an SNR of 10 dB. We initialize the parameter σ_v^2 by sampling from its prior distribution using $\alpha_v = 2$ and $\beta_v = 1$; the vector of channel coefficients $\mathbf{h}$ is initialized as $\mathbf{h}^{(0)} = [1\, 0\, \ldots\, 0]^T$, corresponding to a flat fading channel, and $\mathbf{x}$ is initialized with the sign of each element of vector $\mathbf{y}$. Several realizations of the algorithm are performed, each time with different initialization; in each realization, the algorithm runs for $I = 1000$ iterations. To assess the number of iterations typically required for convergence, we adopt the third procedure described in Sec. VIII, in which samples drawn from a scalar variable are plotted for each realization of the chain. In this application, the most natural choice is variable σ_v^2. The evolution of σ_v^2 over time for each initialization (Fig. 6) allows us to conclude that the chain converges after a few dozens iterations. From this figure, it is safe to consider the first 100 iterations as burn-in and to use the remaining 900 iterations to perform Monte Carlo inference on the quantities we wish to estimate. By analyzing the histograms presented in Fig. 7, one can conclude that the Gibbs sampling receiver accurately estimated the correct parameter value (red square in the figure). From the sampled values for $\mathbf{x}$ we can easily calculate the maximum a posteriori estimate for each transmitted symbol: the MAP estimate for x_j is simply the most frequently sampled value (+1 or -1) among all samples of $x_j^{(i)}$ for $i > 100$. In this example, the symbol error rate was about 0.1 %. For the sake of comparison, a simple detector that only evaluates the sign of $\mathbf{y}$ to estimate $\mathbf{x}$ would produce an error rate of about 15 %.

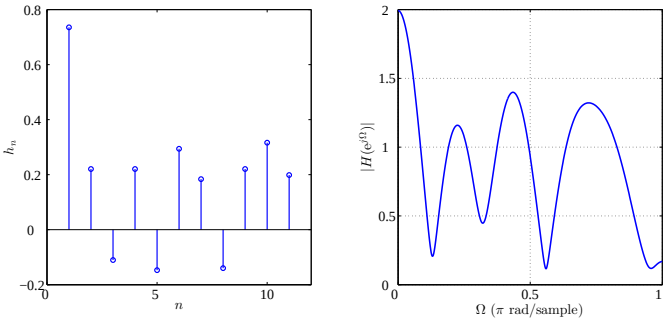

Fig. 5. Characterization of multi-path communication channel. Left: channel impulse response. Right: absolute value of the frequency response of the chanel.

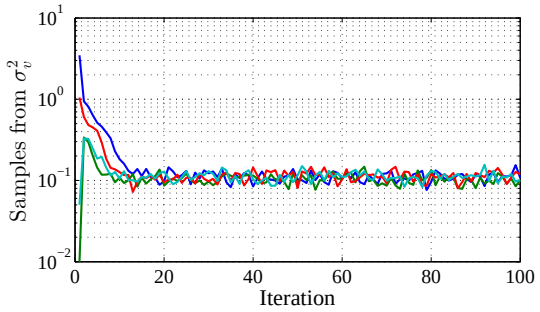

Fig. 6. Samples of the noise variance σ_v^2 via Gibbs sampling, for several initializations. The samples coalesce in the region around the correct values ($\sigma_v^2 = 0.1$) after a few dozen iterations.

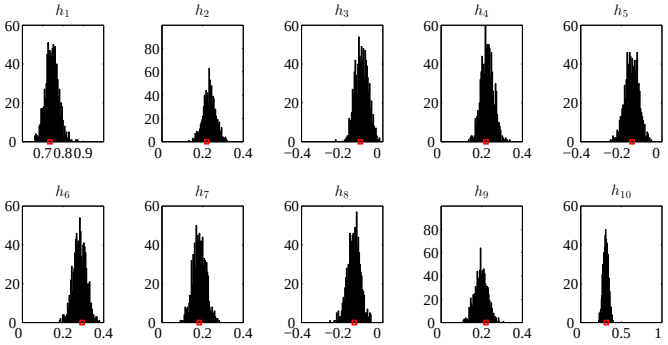

Fig. 7. Histograms of each coefficient of the channel impulse response, generated from the Gibbs sampling. The samples are concentrated around the true value (in red).

XI. Applications of Bayesian Techniques in More Complex Scenarios

The simple communication model of the previous section considers BPSK modulation and assumes static frequency selective channel as well as Gaussian noise. In many practical situations, the modulation scheme is more sophisticated and these simplifying channel conditions are inaccurate, which yield rather complex Bayesian solutions. This section presents a survey of recent advances in Bayesian receivers applied to systems of practical importance. For the sake of organization, we divided the section according to specific aspects of the overall system that might differ from the one considered so far.

A. Code Division Multiple Acces (CDMA)

If a wireless medium is shared by more than one user, Multiple Access Interference (MAI) among users' messages in the receiver might occur. In systems employing TDMA (Time Division Multiple Access) and FDMA (Frequency Division Multiple Access) [16] the level of interference among users is typically low, and thus the receiver can estimate the message of each user in a separate fashion. In contrast, DS-CDMA (Direct Sequence — Code Division Multiple Access) scheme [13] poses difficulties for the receiver because of imperfect orthogonality between users' spreading codes. Multiuser Detection (MUD) in CDMA systems is an important and difficult task that has been extensively researched [31].

For a CDMA system with K users in a flat fading scenario, the received signal can be written as [32]:

$$\mathbf{y}_n = \sum_{k=0}^{K-1} h^{(k)} x_n^{(k)} \mathbf{s}_k + \mathbf{v}_n, n = 0, 1, \ldots, N-1, \qquad (61)$$

where now $h^{(k)}$ is the single tap of the flat channel for user k, $x_n^{(k)}$ is the n-th symbol of user k, $\mathbf{s}_k$ is the k-th user spreading sequence (known by the receiver), and $\mathbf{v}_n$ represents Gaussian white noise. By denoting $\mathbf{Y} = [\mathbf{y}_0\, \mathbf{y}_1\, \ldots\, \mathbf{y}_{N-1}]$, the detection consists of obtaining the posterior probability of each transmitted signal for each user, that is

$$P(x_n^{(k)}|\mathbf{Y}), n = 0, \ldots, N-1, k = 1, \ldots, K. \qquad (62)$$

This problem can be solved using MCMC by treating $h^{(k)}$ and the noise variance σ_v^2 as nuisance parameters, in a procedure similar to the one in the previous section. This is the approach carried out in [33] and [32], where conditional distributions for each parameter were calculated by exploring the noise Gaussianity and the linearity of the model in Eq. (61). A Gibbs sampling-based solution was designed in order to generate samples from the full posterior distribution from which the most likely symbol sequence for each user can be obtained. Similarly to what was done in Alg. 2, the authors explore the flexibility of the Gibbs sampling in order to sample the message data on a symbol-by-symbol basis, thus reducing the computational complexity of the receiver. A similar approach is adopted in [34] to perform multituser detection for a channel subject to Intersymbol Interference (ISI). Still in the MCMC context, the case of impulsive noise in CDMA is addressed in [35], in which a mixture of Gaussians with two components is used to model the noise.

B. Orthogonal Frequency Division Multiplex (OFDM)

Systems employing OFDM modulation allow the transmission of symbols through several overlapping yet orthogonal carriers, which in practice are defined from the vector basis of the Fourier transform. By introducing a cyclic prefix, which repeats a certain number of symbols at the beginning of every block of transmitted symbols, OFDM systems indirectly transform frequency selective channels into a number of parallel flat AWGN channels. Hence, OFDM symbols can be recovered by deploying an algorithm similar to the one presented in Sec. X for a special case of $L = 1$, corresponding to a flat fading channel with one unknown tap.

In practice, however, OFDM systems suffer from imperfections that preclude perfect flattening of the channel. One common problem is the presence of frequency-offset of the carriers caused by the mismatch between the oscilator in the transmitter and that in the receiver, which creates Inter-Subcarriers Interference (ICI). As argued in [36], the common approach of estimating the frequency-offset and then compensating for it in the receiver is sub-optimal. The authors of [36] then propose a Bayesian procedure to estimate the symbol sequence while integrating out in an MCMC context the frequency-offset and channel impulse response coefficients. While the standard model for OFDM is linear, the presence of frequency-offset makes the solution more complicated. By adopting a uniform prior between specified limits for the frequency-offset, they show that the posterior distribution for this parameter cannot be easily sampled. Their solution was to use a Metropolis-Hastings step within the Gibbs sampling framework, in which a simple uniform distribution is chosen as the proposal distribution. This is an example of a common practice called MH within Gibbs [8], in which MH steps are included in a Gibbs sampling framework when direct sampling of particular variables is difficult. They also considered a variation of the MCMC algorithm in which the Gibbs sampling is adopted with local linearization of the model with respect to the frequency-offset, resulting in a Gaussian approximate posterior distribution which can be easily sampled from. Simulations indicate that the solution employing Metropolis-Hastings is superior and close to the one obtained with ideal Channel State Information (CSI).

A related and more sophisticated MCMC solution is adopted in [37], in which phase noise, in addition to multipath fading and frequency-offset, is handled by the model in an MIMO-OFDM scenario. Simulations indicate that the Bayesian solution achieves results that are almost as good as those of a receiver with perfect knowledge of all relevant variables. Another common issue affecting OFDM is the nonlinearity of the High Power Amplifier (HPA), which tends to be severe due to the high Peak-to=Average Power Ratio (PAPR) observed due to the near Gaussian OFDM signal. An early example of Bayesian MCMC receiver can be found in [38], in which the problem of detection of clipped OFDM symbols is addressed.

C. Multiple Input/Multiple Output (MIMO) Systems

MIMO systems [39] employ several transmitting and receiving antennas in order to explore the diversity inherent in wireless channels and to substantially increase spectral efficiency. Such systems have attracted significant attention lately and are expected to be employed in fifth generation (5G) cellular networks [40]. In comparison to the scenario in Sec. X, MIMO systems require modeling the many possible paths between each pair of transmitter and receiver antennas. Even though the number of variables grows substantially, the model is still linear and the Bayesian analysis that is similar to the one in Sec. X applies.

A recent approach that uses a variant of MCMC appears in [41], where a low-complexity multiuser receiver for large-scale MIMO systems is proposed. A modification of the standard Gibbs sampling algorithm is adopted in order to alleviate the so-called *stalling problem* that occurs when SNR is high, a situation in which the algorithm can be trapped in a low quality local maximum. The idea is to adopt a mixed sampling scheme where the algorithm chooses probabilisticaly between sampling from the posterior distribution of the symbol sequence or from a uniform distribution. The authors consider further the use of multiple restart strategies that have proved effective for high-order QAM, in addition to MCMC estimation of the channel parameters. Simulations show that the receiver combining those techniques performs near-optimally in terms of BER with a complexity that scales well with the number of transmitter and receiver antennas.

In [42], an MCMC-based approach for multiuser detection in coded MIMO and CDMA systems is presented. By disregarding the effect of the multipath channel, and considering a BPSK source, the paper focuses on a turbo-based receiver which ultimately depends on calculating the likelihood ratio for each possible transmitted bit. The solution employs the Gibbs sampling algorithm to indirectly implement an analytically complicated summation. The authors show that their approach can be seen as a Rao-Blackwellized [8] version of an existing sub-optimal approach, and argue that their proposal achieves a lower variance estimate. They also tackle the *stalling problem* by modifying the posterior distribution of the data sequence to make it less concentrated on certain values. Simulations indicate that the proposed Bayesian solution outperforms traditional approaches such as the sphere decoding [43] and MMSE (Minimum Mean Square Error) [44] detectors in terms of Bit Error Rate.

D. Impulsive Noise

Despite the ubiquity of white Gaussian noise modeling of real-world electrical disturbances, in many important cases the noise can show more erratic behavior that does not follow a normal distribution. For instance, impulsive noise is prevalent in systems employing Extremely Low Frequency (ELF) or Very Low Frequency (VLF). Because non-Gaussian distributions are analytically less tractable than their Gaussian counterparts, a Bayesian receiver capable of dealing with impulsive noise is more complex than the one described in Algorithm 2.

In [45], impulsive noise is considered in a system employing multiple antennas and that is subject to flat fading. A blind receiver is designed using an MCMC approach similarly to the one previously described. To model the heavy-tailed distribution of impulsive noise, the author adopts a sub-Gaussian density with a free parameter α that controls the shape of the curve. The presented MCMC algorithm has a structure similar to that of Algorithm 2, except that now some parameters cannot be easily sampled from their full-conditional distribution. The Metropolis-Hastings algorithm with a Gaussian proposal is adopted to indirectly obtain samples of those parameters in an MH within Gibbs scheme.

Simulations reported by the authors indicate that the Bayesian blind receiver performs very close to the optimum receiver (in which all the parameters are assumed to be known) in terms of Bit Error Rate. In addition, the receiver can work when noise is Gaussian, a situation corresponding to $\alpha = 2$. The authors argue that the receiver can be implemented in practice considering the low data rate in ELF/VLF systems (for which the algorithm was designed) and the fact that parallelization using Graphical Processing Units (GPUs) can be used.

E. Nonlinear Effects

In addition to the frequency selectivity of the channel, in practice the presence of nonlinear effects in power amplifiers poses additional challenges for receivers. Nonlinear channels can arise in situations with severe fading, which requires high transmission power and thus amplifiers operating outside their linear range. Practical examples include satellite communications, in which severe fading is caused by the long distance between transmitter and receiver, and millimiter wave communication, in which substantial path loss is the result of the system adopting very high frequency carriers [16].

Nonlinear effects have usually been handled by introducing in the transmitter a predistortion, consisting of a nonlinear system that ideally compensates for the nonlinearity in the amplifier. The compensating nonlinear system requires implementing a feedback control loop operating in the basedband, which can be computationally intensive and can produce imperfect cancellation. Another possibility is to implement the nonlinear equalization in the receiver. This approach requires sophisticated nonlinear signal processing tools, since the nonlinear distortion is unknown in the receiver and the received signal is also affected by linear multipath distortion.

A recent approach [46] adopting the latter strategy performs Bayesian equalization of nonlinear channels in millimeter wave communications, which is a technology likely to be used in 5G mobile telephone systems. The overall channel is modeled as a cascade of a fixed memoryless nonlinearity associated with the power amplifier, a time-varying linear system describing the effect of multipath, and an additive white Gaussian term to model ambient noise. The nonlinear element is approximated by a truncated Taylor series while a Finite Impulse Response filter represents the linear part. The nonlinear element in the model makes the posterior distribution analytically intractable — that is, the distribution for most

relevant variables does not have a well-known form. At this point, a batch approach consisting of Gibbs sampling with some MH steps could be adopted, resulting in a structure similar to Algorithm 2. The authors, however, chose to describe the model in a state-space form, and to tackle the problem of symbol detection in a Sequential Monte Carlo framework [15]. The practical solution is based on particle filtering (PF) and adopts local linearization of the nonlinear curve to simplify the computation. Simulation results show that the proposed PF-based solution outperforms traditional transmitter predistortion-based solutions and blind linear equalizers based on Maximum A Posteriori (MAP) estimation. Furthermore, the authors argue that the introduction of a nonlinear element in the model does not increase significantly the energy consumption of the receiver in comparison to a PF solution that deals only with multipath distortion.

F. Underwater Acoustic Communications

Most wireless receivers are designed for transmission through the air, typically for urban or sub-urban environments in which the channel has large bandwidth and changes relatively slowly and the propagation speed is very high. In contast, the transmission medium in underwater acoustic communications (UAC) [47] is characterized by low speed, highly scarce bandwidths, high delay spread causing severe ISI and rapidly changing multipath channels. On the bright side, the impulse response of such channels can be considered sparse, and thus its estimate can be improved by employing sophisticated compressive sensing techniques [48].

In [49], a semi-blind Gibbs sampling receiver is designed for joint symbol detection and channel estimation for Single-Input Single-Output (SISO) uncoded UAC systems. The considered system has the same elements discussed in Sec. X, except that now the channel is highly time-varying. The receiver processes the data on a block basis, and the channel is assumed to be fixed in each block. To handle the interblock interference caused by the channel memory, accommodations on a single-block solution are performed in which data from previous blocks are used for parameter estimation in current block. The proposed Gibbs sampling solution achieves significantly lower BER values in comparison to an alternative two-stage solution in which the symbol and channel are separately estimated, at a cost of a substantially higher computational burden.

A few works in the literature explore the sparsity of the channel to obtain improved parameters estimates. An example of such an approach can be found in [50], in which the compressive sensing estimate of the channel is performed and the result is used as a parameter for a Bayesian detector. A similar procedure is proposed in [51]; the main difference is that the channel model is incorporated in the Bayesian context by using a sparsity-inducing prior for the channel coefficients.

G. Fast Fading Channels

Some of the most common applications of wireless communications aim at providing service to mobile users, leading to offset of the carrier frequency and a channel that changes over time. One way to deal with this type of channel is by employing the modern techniques of sequential Monte Carlo (SMC) [15], [52], a generic class of techniques that includes the increasingly popular particle filters. As a generalization of Kalman filters, these techniques allow for constant update of the channel estimate as well as the transmitted symbols as new data are received. Additionally they are capable of dealing effectively with nonlinear channels and non-Gaussian noise. These techniques are a powerful tool for severely distorted time-varying, nonlinear and frequency selective channels. Examples of applications of particle filtering to varied problems in communications include [53], [54]. This paper has focused on batch processing using MCMC; we intend to address sequential processing in a Bayesian context in future works.

XII. Discussion on the performance of Bayesian/MCMC receivers

This section addresses performance issues of MCMC receivers. We start surveying a paper in which the convergence of three MCMC algorithms is investigated, and we follow with an informal discussion regarding the comparison between Bayesian/MCMC receivers and a few alternative approaches for designing blind receivers.

A. Convergence Analysis

As stated in Sec. VIII, establishing accurate convergence results for MCMC algorithms is difficult. Even when analytic formulas are available, they depend on parameters that are unknown and difficult to estimate. In [55], a comprehensive analysis of convergence of various MCMC algorithms for a few wireless communication systems is provided. The paper resorts to a mix of theoretical and empirical tools to provide approximate yet useful guidelines on how to assess and improve convergence of MCMC algorithms. The authors state that the convergence of MCMC algorithms is geometric, provided the chain is aperiodic and irreducible. While determining the exact convergence rate is difficult, in practice it can be roughly estimated by exploring the fact that the convergence rate is closely linked to the correlation between samples of the chain. They also show that the overall convergence of the chain can be indirectly assessed by analyzing the correlation of a subset of the variables of the chain. In an MCMC receiver, the sequence of sampled digital symbols forms a discrete Markov chain, for which the convergence rate depends on the second largest eigenvalue of the transition matrix (as seen in Sec. VIII). By numerically calculating this transition matrix for a few MCMC receivers the authors managed to obtain estimates for their convergence rates for a few relevant cases.

For an AWGN channel (corresponding to Eq. (44) with $L = 1$) the authors show that it is possible to analytically integrate out the nuisance parameters and thus perform sampling only of the transmitted data sequence. The resulting distribution for the transmitted data is bimodal due to the ambiguity in the model: the same output $y_n = h_1 x_n + v_n$ can be obtained with variables $-h_1$ and $-x_n$. The transition matrix is $2^N \times 2^N$, where N is the number of transmitted symbols. By assuming the noise variance to be known (which tends to be the case in practice), and $N = 5$ symbols, the second largest eigenvalue

λ_2 of the resulting 32×32 matrix can be computed. Their analysis shows that perhaps paradoxically a high SNR leads to slower convergence. The explanation is that the modes of the posterior distribution are farther apart when SNR is high, which makes it harder for the chain to transition from one mode to the other. The authors report that, by using differential coding, this ambiguity is removed and the convergence tends to improve when SNR increases.

For an ISI channel (like the one in Eq. 44), the authors discuss a few variations of MCMC receivers that are closely related to Algorithm 2. They consider three possibilities: (1) alternate sampling of data sequence $\mathbf{x}$ and joint sampling of the remaining nuisance parameters $\mathbf{h}$ and σ_v^2; (2) sampling of data sequence $\mathbf{x}$ in groups of size q, with the remaining steps similar to (1); and (3) sampling of data sequence from its posterior (with nuisance parameters integrated out) one component at a time either through Gibbs sampling or Metropolis-Hastings. Their analysis shows that the Gibbs sampling of one component at a time is superior to all others. The grouping algorithm converges faster than the two components algorithm; yet it requires more computation per iteration. For all algorithms, the convergence gets slower as the SNR increases. The reason now is the shifting ambiguity caused by the channel, which makes the posterior multimodal and more concentrated when SNR is high. The authors suggest introducing constraints in the model to avoid ambiguity and thus improve convergence.

Finally, the paper addresses Bayesian receivers for CDMA channels without ISI. The adopted channel model for CDMA can be seen as a generalization of the AWGN channel in which a vector of binary symbols, rather than a single symbol, is transmitted at every time index. The convergence of a standard Gibbs sampling based multiuser detection algorithm is studied with respect to SNR and the correlation between spreading codes (ρ). The numerical analysis showed that higher SNR and higher ρ are detrimental to the convergence of the algorithm.

In summary, the authors conclude that convergence is typically negatively affected by higher SNR, high chain correlation and ambiguities in the model. In addition, integrating out continuous parameters whenever possible helps to improve convergence, since fewer variables remain in the chain and they become less correlated.

B. Comparison to Alternative Approaches

The MCMC-based receiver presented in the previous section can be categorized as a blind receiver, since it does not require pilot-based training. The advantage of such receivers is that they provide higher throughput since resources that would have been allocated to transmit pilot data can be used to transmit information data. Blind receivers in general rely on known properties of the transmitted signal and the channel in order to perform symbol detection based solely on the received signal. Many approaches for blind receivers exist with varying levels of complexity and performance.

The Constant Modulus Algorithm (CMA), for example, explores the fact that, in a few systems such as FSK, MSK and QPSK, the envelope of the transmitted signal is constant, corresponding to constant modulus constelation [13]. Since the multipath channel and the noise tend to modify this modulus, a receiver can be designed to process the signal in order to force the modulus to be constant again. It is expected that the output of such receiver will produce symbols that are closer to the transmitted ones on average. The CMA can be straightforwardly implemented using gradient-based adaptive approaches, similarly to traditional LMS adaptive filters. However, the various versions of the CMA suffer from two main drawbacks in comparison to traditional adaptive equalizers: longer time needed for convergence and risk of converging to a local minimum.

Another well-researched approach for blind equalization is to use second- or higher-order statistics (HOS) of the transmitted signal. Even though in general the phase information is not captured by second order statistics such as the autocorrelation function (ACF), this is not the case when the the ACF is periodic. A practical way to impose this periodicity, making the signal cyclostationary and allowing the use of the ACF for equalization, is to oversample the signal output. The use of High Order statistics does not require the cyclostationary assumption, and thus can be used without oversampling the received signal; however the accuracy of the HOS-based equalizer is inferior [13].

In comparison to these techniques, the fully-Bayesian approach is expected to produce superior results in terms of symbol error rate, because it takes into account all available knowledge about the relevant variables. In particular, Bayesian methods can easily incorporate noise in the model, as explained in last section, thus avoiding the phenomenon of noise enhancement which is a drawback in many blind equalization techniques, including those mentioned in previous paragraphs. While the posterior distribution needed in a Bayesian solution can be quite complex, the usage of MCMC prevents the equalizer from converging to a local minimum. In contrast, as stated before, traditional gradient-based techniques are prone to sub-optimal convergence. Bayesian methods are typically associated with higher computational complexity, which is the result of using realistic modeling with more parameters. However, when applied to digital communication problems, Bayesian tools jointly with MCMC using the blocking strategy presented earlier can produce receivers of acceptable complexity.

XIII. CONCLUSIONS

In this tutorial we presented the fundamentals of Bayesian inference and the techniques for stochastic simulation, with particular emphasis on their application in symbol detection for several modern digital wireless communication systems. We explored in detail the case of time-invariant frequency selective channel with additive Gaussian noise in a system employing BPSK. This simple scenario allows us to understand the main elements of Bayesian analysis, namely: the choice of prior distributions, the calculation of the posterior distribution via Bayes' theorem and the practical receiver design via MCMC techniques. In contrast to traditional receivers that avoid full statistical modeling, the resulting algorithm naturally yields optimal estimation of users' data as it integrates out all

nuisance parameters. Furthermore, the method can produce joint blind symbol detection and channel estimatation, and it can be adapted to yield channel estimation when a pilot sequence is available. The stochastic nature of MCMC makes the algorithms more robust to local minima, and it can indirectly produce credence intervals for all variables in the model in order to assess the accuracy of the estimates. On the downside, the complexity of the resulting hierarchical Bayesian model may imply a high computational burden, yielding to high energy consumption in the receiver. This burden, however, can be alleviated by implementing some pre-processing to produce initial favorable estimates for key variables in the model, or by using past estimates in the context of block processing. Having been successfully tested in many wireless technologies that are expected to be adopted in the near future, such as massive MIMO and millimeter wave communications, Bayesian tools tend to become more popular as communication systems become more complex and the continuously increasing computational power enables their practical implementation.

REFERENCES

[1] L. Wasserman, *All of Statistics – A Concise Course in Statistical Inference*. Springer, 2010.

[2] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Taylor & Francis, 2014, vol. 2.

[3] H. S. Migon and D. Gamerman, *Statistical Inference: An Integrated Approach*. London, UK: Hodder Arnold, 1999.

[4] S. Mcgrayne, *The Theory that Would not Die – How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press, 2011.

[5] H. Jeffreys, *The Theory of Probability*. OUP Oxford, 1998.

[6] B. de Finetti, *Theory of Probability*. Wiley, 1974.

[7] L. J. Savage, *The Foundations of Statistics*. Courier Corporation, 1972.

[8] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York, USA: Springer, 2004.

[9] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, Nov. 1984, doi = 10.1109/TPAMI.1984.4767596.

[10] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in *Learning for Text Categorization: Papers from the 1998 workshop*, vol. 62, 1998, pp. 98–105.

[11] N. Silver, *The Signal and the Noise: Why So Many Predictions Fail — but Some Don't*. Penguin Publishing Group, 2012.

[12] S. S. Haykin, *Communication Systems*. Wiley, 2001.

[13] A. F. Molisch, *Wireless Communications*. Wiley, 2010.

[14] A. Doucet, N. De Freitas, and N. Gordon, "An introduction to sequential monte carlo methods," in *Sequential Monte Carlo methods in practice*. Springer, 2001, pp. 3–14.

[15] J. V. Candy, *Bayesian signal processing: Classical, modern and particle filtering methods*. John Wiley & Sons, 2011, vol. 54.

[16] T. S. Rappaport, *Wireless Communications: Principles and Practice*. Prentice Hall PTR, 2002.

[17] T. Bayes, "An essay towards solving a problem in the doctrine of chances." *Philosophical Transactions of the Royal Society*, vol. 53, pp. 370–418, 1763.

[18] P. S. Laplace, *Essai Philosophique sur les Probabilités. English; A philosophical essay on probabilities*. J. Wiley, 1902.

[19] E. Jaynes, "Prior probabilities," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 3, pp. 227–241, Set. 1978, doi = 10.1109/TSSC.1968.300117.

[20] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *Journal of the Royal Statistical Society*, vol. 41, no. 2, pp. 113–147, Dez. 1979.

[21] N. M. L. A. P. Dempster and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[22] N. Metropolis, A. W. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[23] W. K. Hastings, "Monte Carlo sampling methods using Markov Chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970. doi = 10.1093/biomet/57.1.97.

[24] R. Eckhardt, "Stan Ulam, John von Neumann, and the Monte Carlo method," *Los Alamos Science*, pp. 131–143, 1987.

[25] C. Robert and G. Casella, "A short history of Markov Chain Monte Carlo: subjective recollections from incomplete data," *Statistical Science*, pp. 102–115, 2011.

[26] A. Papoulis, *Probability, Random Variables, and Stochastic Process*. New York, USA: McGraw-Hill, 1984.

[27] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, Jan.-Fev. 2003.

[28] R. Neal, "Probabilistic inference using Markov Chain Monte Carlo methods," University of Toronto, Department of Computer Science, Toronto, Canada, Technical Report CRG-TR-93-1, 1993.

[29] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 398–409, 1990. doi = 10.2307/2289776.

[30] R. Chen and T.-H. Li, "Blind restoration of linearly degraded discrete signals by gibbs sampling," *IEEE Transactions on Signal Processing*, vol. 43, no. 10, pp. 2410–2413, 1995. doi = 10.1109/78.469847.

[31] S. Verdu, *Multiuser detection*. Cambridge university press, 1998.

[32] X. Wang and H. V. Poor, *Wireless communication systems: Advanced techniques for signal reception*. Prentice Hall Professional, 2004.

[33] X. Wang, R. Chen, and J. S. Liu, "Monte Carlo Bayesian signal processing for wireless communications," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 30, no. 1-3, pp. 89–105, 2002. doi = 10.1023/A:1014094724899.

[34] X. Wang and R. Chen, "Blind turbo equalization in gaussian and impulsive noise," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 4, pp. 1092–1105, 2001. doi = 10.1109/25.938583.

[35] ——, "Adaptive bayesian multiuser detection for synchronous CDMA with gaussian and impulsive noise," *IEEE Transactions on Signal Processing*, vol. 48, no. 7, pp. 2013–2028, 2000. doi = 10.1109/78.847787.

[36] B. Lu and X. Wang, "Bayesian blind turbo receiver for coded OFDM systems with frequency offset and frequency-selective fading," *Selected Areas in Communications, IEEE Journal on*, vol. 19, no. 12, pp. 2516–2527, 2001. doi = 10.1109/49.974616.

[37] F. Z. Merli, X. Wang, and G. M. Vitetta, "A Bayesian multiuser detection algorithm for MIMO-OFDM systems affected by multipath fading, carrier frequency offset, and phase noise," *Selected Areas in Communications, IEEE Journal on*, vol. 26, no. 3, pp. 506–516, 2008. doi = 10.1109/JSAC.2008.080409.

[38] D. Declercq and G. B. Giannakis, "Recovering clipped OFDM symbols with bayesian inference," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 1. IEEE, 2000, pp. 157–160.

[39] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[40] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014. doi = 10.1109/MCOM.2014.6736746.

[41] T. Datta, N. A. Kumar, A. Chockalingam, and B. S. Rajan, "A novel monte-carlo-sampling-based receiver for large-scale uplink multiuser MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3019–3038, 2013. doi = 10.1109/TVT.2013.2260572.

[42] B. Farhang-Boroujeny, D. H. Zhu, and Z. Shi, "Markov Chain Monte Carlo algorithms for CDMA and MIMO communication systems," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1896–1909, 2006. doi = 10.1109/TSP.2006.872539.

[43] B. M. Hochwald and S. Ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE transactions on communications*, vol. 51, no. 3, pp. 389–399, 2003. doi = 10.1109/TCOMM.2003.809789.

[44] X. Wang and H. V. Poor, "Iterative (turbo) soft interference cancellation and decoding for coded CDMA," *IEEE Transactions on communications*, vol. 47, no. 7, pp. 1046–1061, 1999. doi = 10.1109/26.774855.

[45] W. Ying, Y. Jiang, Y. Liu, and P. Li, "A blind receiver with multiple antennas in impulsive noise modeled as the sub-Gaussian distribution via the MCMC algorithm," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3492–3497, 2013. doi = 10.1109/TVT.2013.2250535.

[46] B. Li, C. Zhao, M. Sun, H. Zhang, Z. Zhou, and A. Nallanathan, "A Bayesian approach for nonlinear equalization and signal detection in millimeter-wave communications," *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 3794–3809, 2015. doi = 10.1109/TWC.2015.2412119.

[47] M. Stojanovic and J. Preisig, "Underwater acoustic communication channels: Propagation models and statistical characterization," *IEEE Communications Magazine*, vol. 47, no. 1, pp. 84–89, 2009. doi = 10.1109/MCOM.2009.4752682.

[48] R. G. Baraniuk, "Compressive sensing," *IEEE signal processing magazine*, vol. 24, no. 4, 2007. doi = 10.1109/ALLERTON.2015.7447163.

[49] J. Ling and J. Li, "Gibbs-sampler-based semiblind equalizer in underwater acoustic communications," *IEEE Journal of Oceanic Engineering*, vol. 37, no. 1, pp. 1–13, 2012. doi = 10.1109/JOE.2011.2171132.

[50] J. Huang, S. Zhou, J. Huang, J. Preisig, L. Freitag, and P. Willett, "Progressive MIMO-OFDM reception over time-varying underwater acoustic channels," in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*. IEEE, 2010. doi = 10.1109/ACSSC.2010.5757747, pp. 1324–1329.

[51] J. Ling, X. Tan, T. Yardibi, J. Li, M. L. Nordenvaad, H. He, and K. Zhao, "On Bayesian channel estimation and FFT-based symbol detection in MIMO underwater acoustic communications," *IEEE Journal of Oceanic Engineering*, vol. 39, no. 1, pp. 59–73, 2014. doi = 10.1109/JOE.2012.2234893.

[52] A. Doucet, A. Smith, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, ser. Information Science and Statistics. Springer New York, 2001.

[53] E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald, "Particle filters for demodulation of M-ary modulated signals in noisy fading communication channels," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP'00.*, vol. 5. IEEE, 2000. doi = 10.1109/ICASSP.2000.861087, pp. 2797–2800.

[54] D. Angelosante, E. Biglieri, and M. Lops, "Sequential estimation of multipath MIMO-OFDM channels," *Signal Processing, IEEE Transactions on*, vol. 57, no. 8, pp. 3167–3181, 2009. doi = 10.1109/TSP.2009.2020049.

[55] R. Chen, J. S. Liu, and X. Wang, "Convergence analyses and comparisons of Markov Chain Monte Carlo algorithms in digital communications," *Signal Processing, IEEE Transactions on*, vol. 50, no. 2, pp. 255–270, 2002. doi = 10.1109/78.978381.

Flávio R. Ávila received a D.Sc. degree in Electrical Engineering from the Federal University of Rio de Janeiro (UFRJ), Brazil, in 2012. Since 2013, he has been with the departent of Electronics and Telecommunications Engineering at the Rio de Janeiro State University (UERJ). His main research interest is applications of statistical signal processing to audio and speech processing and wireless communications. He is a member of the IEEE, the SBrT and the AES.

Michel P. Tcheou was born in Rio de Janeiro, Brazil. He received the Engineering degree in electronics from Federal University of Rio de Janeiro (UFRJ) in 2003, the M.Sc. and D.Sc. degrees in Electrical Engineering from COPPE/UFRJ in 2005 and 2011, respectively. He has worked at the Electric Power Research Center (Eletrobras Cepel) in Rio de Janeiro, Brazil, from 2006 to 2011. Since 2012 he has been with the Department of Electronics and Communications Engineering (the undergraduate dept.) at Rio de Janeiro State University (UERJ). He has also been with the Postgraduate in Electronics Program. His research interests are in signal processing, communications, data compression and numerical optimization