Bayesian learning of probability density functions: a Markov chain Monte Carlo approach

Simone Del Favero, Damiano Varagnolo, Gianluigi Pillonetto

***Abstract*— The paper considers the problem of reconstructing a probability density function from a finite set of samples independently drawn from it. We cast the problem in a Bayesian setting where the unknown density is modeled via a nonlinear transformation of a Bayesian prior placed on a Reproducing Kernel Hilbert Space. The learning of the unknown density function is then formulated as a minimum variance estimation problem. Since this requires the solution of analytically intractable integrals, we solve this problem by proposing a novel algorithm based on the Markov chain Monte Carlo framework. Simulations are used to corroborate the goodness of the new approach.**

***Index Terms*— stochastic regularization, regularization parameter, Reproducing Kernel Hilbert Spaces, Metropolis-Hastings algorithm, stochastic processes.**

I. Introduction

In many fields ranging from basic science to engineering one is often confronted with reconstructing the stochastic mechanism generating some observational data. Examples of applications abound and we cite e.g. pattern classification, clustering, time series prediction, characterization of materials, spatial modeling [1], [2], [3], [4].

Reconstructing a probability density function is in general intricate. The problem is in fact intrinsically nonlinear, since it includes nonnegative and unitary constraints. In addition, it is subject to the so-called bias/variance dilemma [5], [6], [7]. If the hypothesis space where the unknown function is searched is too large, the estimate may turn out close to the maximum likelihood one, i.e. a sum of delta function spikes centered at the observations. This estimate is in general poor, since a priori information about the smoothness of the function is typically available. On the other hand, if the hypothesis space is too narrow, the solution could turn out too few adherent to experimental data, with still a poor predictive capability on new data.

Parametric approaches, e.g., [8], tackle these difficulties assuming finite-dimensional hypothesis spaces. This is done by imposing a known parametric form to the unknown density, e.g. a mixture of Gaussians. Regularity and nonnegativity assumptions on the unknown function can thus be easily included in the estimation process, and the problem can be solved by just fitting parameters against data, e.g., via standard nonlinear least squares algorithms. However, the model designer has often too few information to specify so strong a priori assumptions on the density shape. This represents the major drawback of parametric techniques.

A more powerful alternative is represented by nonparametric approaches, which have a wider range of applicability since they do not require to postulate a fixed-in-advance functional form. Examples of nonparametric techniques are penalized likelihood methods [5], [9], [10], [11], ad-hoc penalized likelihood methods, smoothed histograms [12], kernel methods [13], [14], regularized Gaussian Mixtures [15] and orthogonal series estimates [16]. In particular, among the most employed approaches, we cite Parzen's window estimator, the k-nearest neighbor approach and smoothing spline density estimation. Although applied with success in many applications, all these methods have however some limitations as how they handle the bias/variance dilemma. In fact, key parameters controlling the complexity of the hypothesis space and having a major effect on the final estimate, e.g., the kernel width in Parzen's approach, is in practice chosen empirically. Methods used to estimate the optimal values of such parameters are often asymptotic [17], thus prone to error when dealing with small data sets, or are based on cross validation techniques [18], [9], thus possibly subject to statistical error.

In practice, because of the nonlinearity of the problem, it is hard to define rigorous statistical criteria determining the right amount of regularization to be included in the estimation problem. In this paper we propose a statistical modeling approach to overcome this problem.

In particular, we embed density estimation within a stochastic framework, by interpreting Tikhonov regularization as placing an opportune Bayesian prior on a Reproducing Kernel Hilbert Space (RKHS) [19], [18], [7]. We then solve the resulting Bayesian estimation problem with a novel algorithm based on the Markov chain Monte Carlo (MCMC) framework [20], [21]. In particular *we jointly learn the regularization parameter and the unknown density function determining their minimum variance estimates*. The paper is organized as follows: in Section II we provide the statement of the problem, then formulate our statistical assumptions on the unknown density function and provide a brief overview on RKHS theory in Section III. In Section IV we connect our statistical model with Tikhonov regularization and compare our approach with some other literature. In Section V, after briefly

All the authors are with the Department of Information Engineering, University of Padova, via Gradenigo 6/b, 35131 Padova, Italy. Emails: { simone.delfavero | varagnolo | giapi } @dei.unipd.it.

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement n°257462 HYCON2 Network of excellence and n°223866 FeedNetBack, by Progetto di Ateneo CPDA090135/09 funded by the University of Padova, and by the Italian PRIN Project "New Methods and Algorithms for Identification and Adaptive Control of Technological Systems".

introducing the MCMC framework, we propose our novel numerical algorithm, In Section VI we use simulations to test the relative performance of the approach. Conclusions are finally offered in Section VII.

II. Statement of the problem

In the following, given a vector w, we use w_i to refer to the i-th component of w. Moreover, all vectors will be column vectors.

We are given n random samples $\{y_i\}$ collected in the vector y and independently drawn from an unknown probability density function $f(x)$. Such density is assumed to have support on the compact set $X \subset \mathbf{R}^d$. Our aim is to estimate f from y.

III. Statistical assumptions on the unknown density

Before specifying our statistical assumptions on f, we first briefly sketch some properties of RKHS which are relevant in the context of this paper.

A. A brief overview on RKHS theory

In the sequel, let $\mathbf{L}^2(X)$ the classical Lebesgue space of square integrable functions on X, equipped with the inner product $\langle \cdot, \cdot \rangle_2$, and let also $K : X \times X \mapsto \mathbf{R}$.

Definition 1. We say that K is definite positive if for all finite sets $\{x_1, x_2, \ldots, x_k\} \subset X$ the $k \times k$ matrix whose (i,j)-th entry is $K(x_i, x_j)$ is semi-definite positive. Moreover, we say that K is a Mercer kernel if it is continuous, symmetric and definite positive.

The following proposition can be obtained by combining the Spectral Theorem for compact operators and Mercer's theorem (see e.g. [22], [23]).

Proposition 2. Let $K(s,t)$ a Mercer kernel. Then there exist a sequence $\{\lambda_j \geq 0 : \lambda_{j+1} \geq \lambda_j, j = 1, \ldots, \infty\}$ and a basis in $\mathbf{L}^2(X)$ of continuous functions $\{\phi_j : j = 1, \ldots, \infty\}$ such that

$$\begin{aligned} \langle \phi_j, \phi_k \rangle_2 &= \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases} \\ \int_X K(s,t)\phi_j(t)dt &= \lambda_j \phi_j(s) \\ K(s,t) &= \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t) \end{aligned}$$

where the above convergence is uniform in $X \times X$.

The following proposition characterizes the Hilbert space associated to the Mercer Kernel K (see e.g. [24], [7]).

Proposition 3. Assigned a Mercer kernel K there exists a unique Hilbert space H such that

- $K(x,y) \in H \quad \forall x \in X$;
- the span of the set $\{K(x,\cdot), x \in X\}$ is dense in H;
- $f(x) = \langle f(y), K(y,x) \rangle_H \quad \forall f \in H$.

In particular, the space H takes the following form

$$H = \left\{ f \in \mathbf{L}^2(X) \;\middle|\; f = \sum_{j=1}^{\infty} a_j \phi_j \text{ and } \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j} < \infty \right\}$$

equipped with the inner product $\langle \cdot, \cdot \rangle_H$ where, given $f, g \in H$ with $f = \sum_{j=1}^{\infty} a_j \phi_j$ and $g = \sum_{j=1}^{\infty} b_j \phi_j$, we have

$$\langle f, g \rangle_H = \sum_{j=1}^{\infty} \frac{a_j b_j}{\lambda_j} \; .$$

The space H is also known in literature as the RKHS associated to the reproducing kernel K. Remarkably, the above Proposition enables us to interpret H as a certain subset of smooth functions in $\mathbf{L}^2(X)$ determined by the eigenvalues and eigenvectors of K.

Example 4. As an example of RKHS, let's define the Green's function G_{W_m} and the reproducing kernel K_{W_m} on $[0,T] \times [0,T]$, $T \in \mathbf{R}$ as

$$\begin{aligned} G_{W_m}(x,y) &:= \begin{cases} 0 & \text{if } x \leq y \\ 1 & \text{if } x > y \text{ and } m = 1 \\ \frac{(x-y)^{m-1}}{(m-1)!} & \text{otherwise} \end{cases} \\ K_{W_m}(x,y) &:= \int_0^T G_{W_m}(x,\tau) G_{W_m}(y,\tau) d\tau \; . \end{aligned}$$

Given a function $f : [0,T] \mapsto \mathbf{R}$, we use $f^{(i)}$ to denote the i-th derivative of f. The RKHS associated to K_{W_m} is then

$$\begin{aligned} W_m = \; & \big\{ f : [0,T] \mapsto \mathbf{R} \;\big|\; f^{(m)} \in \mathbf{L}^2[0,T], \\ & f^{(j)} \text{ absolutely continuous and} \\ & f^{(j)}(0) = 0 \text{ for } j = 0, \ldots, m-1 \big\} \end{aligned}$$

equipped with the inner product (see e.g. [18])

$$\langle f, g \rangle_{W_m} = \left\langle f^{(m)}, g^{(m)} \right\rangle_2 \; .$$

The eigenvalues and eigenvectors of K_{W_m} can be in general numerically computed (see e.g. Lemma 9 in [25]). In particular, if m equals 1, $\phi_{W_1,j}$ and $\lambda_{W_1,j}$ admit the following closed forms [26]:

$$\begin{aligned} \lambda_{W_1,j} &= T^2/[(j-1)\pi + \pi/2]^2 \\ \phi_{W_1,j}(t) &= \sqrt{2/T} \sin\left[(x/T)(j\pi - \pi/2)\right] \; . \end{aligned}$$

B. Stochastic modeling of the unknown probability density function

We now cast our density estimation problem in a stochastic framework. We start defining a Bayesian prior for the unknown function f. In the sequel, we denote with Ψ a certain deterministic, continuous and nonlinear transformation mapping the space of continuous functions on X into itself.

Assumption 5. There exist a positive real number γ and numerable collections of functions $\{\phi_i(x)\}$ on X and non-negative real numbers $\{\lambda_i\}$ such that

- $K(x,y) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(y)$ is a Mercer kernel of a RKHS H;

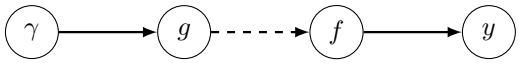

Fig. 1. Bayesian network describing the nonlinear stochastic model for density estimation.

- the function f is a random field of the form

$$f(x) = \frac{\Psi(g)}{\int_X \Psi(g)dx} \tag{1}$$

where

$$g(x) = \sum_{j=1}^{\infty} a_j \phi_j(x)$$

and $\{a_j\}$ are independent Gaussian random variables, being the variance of a_j equal to λ_j/γ.

A graphical description of our model is depicted in Fig. 1 by using the formalism of Bayesian networks (see [27]). Here, random fields/vectors are represented by nodes, stochastic relationships by arrows and deterministic relationships by dashed arrows. One can thus note that in our framework also the regularization parameter is modeled as random variable. In particular, following a standard statistical choice (see e.g. [27]) we specify the entire network by assigning to γ a Gamma distribution $\Gamma(\alpha, \beta)$ of mean α/β.

IV. Relationship with Tikhonov regularization theory

A. Connection with Tikhonov nonlinear regularization

In this sub-section we derive a connection between the probabilistic model of Fig. 1 and Tikhonov nonlinear regularization. Given a function $g \in H$, where $g = \sum_{j=1}^{\infty} a_j \phi_j(x)$, let a^N the vector containing the first N components of $\{a_j\}$. We define as g_a^N the following finite-dimensional approximation of g,

$$g_a^N(x) := \sum_{j=1}^{N} a_j^N \phi_j(x) \ ,$$

where $x \in X$. We define the following prior distribution for a^N:

$$\begin{aligned} p_{a^N}(a^N) &= \frac{1}{\left(2\pi\gamma^{-1}\right)^{N/2}\sqrt{\lambda_1 \cdots \lambda_N}} \exp\left(-\frac{\gamma}{2}\sum_{j=1}^{N}(a_j^N)^2/(\lambda_j)\right) \\ &= \frac{1}{\left(2\pi\gamma^{-1}\right)^{N/2}\sqrt{\lambda_1 \cdots \lambda_N}} \exp\left(-\frac{\gamma}{2}\|g_a^N(x)\|_H^2\right) \ . \end{aligned}$$

Let $\Psi(g_a^N)_{y_j}$ be the function $\Psi(g_a^N)$ evaluated at y_j. Then the conditional density for y given a^N and γ is

$$p_{y|a^N,\gamma}\left(y|a^N,\gamma\right) = \frac{\prod_{j=1}^{n}\Psi(g_a^N)_{y_j}}{\int_X \Psi(g_a^N)dx} \ .$$

The corresponding negative log of the likelihood for a given $y \in \mathbf{R}^n$ and a^N is

$$\begin{aligned} l^N(y, a^N \mid \gamma) = \ & \frac{1}{2}\sum_{j=1}^{N}\log\left(\frac{2\pi\lambda_j}{\gamma}\right) - \sum_{j=1}^{n}\log\left(\Psi(g_a^N)_{y_j}\right) \\ & + \log\left(\int_X \Psi(g_a^N)dx\right) + \gamma\frac{\|g_a^N\|_H^2}{2} \ . \end{aligned}$$

We point out that H, being a RKHS, is a subset of the space of continuous functions and convergence in the topology induced by $\|\cdot\|_H$ implies uniform convergence, see e.g. [7]. Then we easily have that

$$l^N\left(y, g_a^N \mid \gamma\right) - \frac{1}{2}\sum_{j=1}^{N}\log\left(\frac{2\pi\lambda_j}{\gamma}\right) \xrightarrow{N\to\infty} l(y, g \mid \gamma)$$

where

$$\begin{aligned} l\left(y, g \mid \gamma\right) := \ & -\sum_{j=1}^{n}\log\left(\Psi(g)_{y_j}\right) \\ & + \gamma\frac{\|g\|_H^2}{2} + \log\left(\int_X \Psi(g)dx\right) \ . \end{aligned}$$

Given the model of Fig. 1, we can interpret

$$\hat{g} = \arg\min_{g\in H} \ l\left(y, g \mid \gamma\right) \tag{2}$$

as the maximum a posteriori (MAP) estimate of g given y and γ. MAP estimate is thus provided by a Tikhonov nonlinear variational problem which contains two contrasting terms. The first one, equal to $-\sum_{j=1}^{n}\log(\Psi(g)_{y_j})$, takes into account the experimental evidence, the second one, equal to $\|g\|_H^2$, the a priori information about the regularity of the solution. The trade-off between these two components is then established by the regularization parameter γ. Finally, a third term is also present in the estimator, equal to $\log(\int_X \Psi(g)dx)$, which is to enforce the nonnegative and unitary constraint on the unknown probability density function.

B. Connection with other density estimation approaches

We notice that Estimator (2) has already been proposed in the literature with different choices of Ψ. For instance, [28] introduces a penalty on the second derivative of the squared root of the function of interest. Under our framework, this corresponds to define $\Psi : f \mapsto f^{(2)}$, embedding the problem in W_2. [9] instead assumes Ψ to be exponential, i.e. $\Psi : f \mapsto e^f$, and proposes efficient iterative algorithms to implement this model (asymptotic properties are described in [29]).

However, rigorous statistical criteria to determine γ in (2) *have not been so far proposed.* In addition it is worth pointing out that, even if γ were known, the MAP estimate of g given y and γ is in general less robust than its a posteriori expected value. In the next Section we then show how to compute the minimum variance estimates of f and γ via a MCMC approach.

We finally remark that density estimation through infinite Gaussian mixtures [30] correspond to MCMC approaches using more vague priors: in (2) designers can directly

encode assumptions on the regularity of the density through appropriate $K(\cdot,\cdot)$'s, while in infinite Gaussian mixtures this design opportunity is missing.

V. Numerical algorithms

Recovering minimum variance estimates and confidence intervals from the a posteriori probability density function of f and γ graphically described in Fig. 1 requires the solution of analytically intractable integrals. Here, we derive a novel MCMC algorithm for density estimation which circumvents this difficulty. We start providing a brief overview on the Metropolis-Hastings algorithm [20] on which our numerical procedure relies on.

A. A brief overview on the MCMC framework

The goal of a MCMC algorithm is to simulate realizations from a certain posterior distribution so that empirical estimates for any statistics of interest can be determined. A MCMC procedure thus consists of two steps. Firstly, a Markov process with limiting quantities that follow the invariant distribution of interest, in our case the a posteriori probability density function of f and γ, is designed. This first step is used to recover the target distribution of interest in sampled form. Secondly, a Monte Carlo integration is done to obtain the integrals of interest. The common mechanism by which the first step can be performed is the Metropolis/Hastings algorithm. A variant of this procedure, named the single-component Metropolis/Hastings algorithm, will be in particular used in this paper. To describe it, let $\pi(\theta)$ the target density, being θ a finite-dimensional vector containing the parameters of interest. We denote with $q_i\left(Z_{t+1}|Z_t\right)$ a proposal density from which a candidate value Z_{t+1} is drawn when the current state of the chain is Z_t. The scheme suggests to divide Z into h portions of desired dimension, then specifying h proposal density functions $q_i(\cdot|\cdot)$ with $i = 1, 2, \ldots, h$. Every iteration of the algorithm is composed by h distinct phases where Metropolis Hastings updates are employed to explore the parameter space by proposing moves which are subsequently either accepted or rejected. To be specific, at step i of iteration $t+1$ the i-th portion of Z, denoted with $Z_{t+1.i}$, is drawn by the kernel $q_i(Z_{t+1.i}|Z_{t.i}, Z_{t.-i})$, where

$$Z_{t.-i} := \{Z_{t+1.1}, \ldots, Z_{t+1.i-1}, Z_{t.i+1}, \ldots, Z_{t.h}\} \ .$$

In practice, the first $i-1$ components of $Z_{t.-i}$ come from the first $i-1$ steps computed at instant $t+1$. The candidate is then accepted with probability

$$\delta(Z_{t.-i}, Z_{t.i}, Z_{t+1.i}) = \min\left(1, \frac{\pi(Z_{t+1.i}|Z_{t.-i})q_i(Z_{t.i}|Z_{t+1.i}, Z_{t.-i})}{\pi(Z_{t.i}|Z_{t.-i})q_i(Z_{t+1.i}|Z_{t.i}, Z_{t.-i})}\right) \ .$$

If the proposal is rejected then the chain remains in the current state. This scheme guarantees, under mild additional conditions, π to be the limiting distribution of the Markov chain generated (see e.g. [21]). This virtually holds independently of the particular proposal densities $\{q_i(\cdot|\cdot)\}$ employed. Even if their choice is essentially arbitrary, it has however a crucial influence on the rate of convergence of the algorithm. In other words, it can be often problematic to design an efficient MCMC scheme which obtains an accurate reconstruction (in sampled form) of π after a reasonable number of iterations.

B. MCMC algorithm for density estimation

Following [31], we define a MCMC procedure which relies upon a representation of g in terms of a finite subset of eigenvectors $\{\phi_j\}$. To be specific, we assume

$$g(x) = \sum_{j=1}^{N} a_j^N \phi_j(x) \tag{3}$$

where N depends on the specific problem and has to be chosen large enough so as to provide an accurate approximation of the original infinite-dimensional model.

We then block the parameter space into two groups, i.e. γ and a^N. As concerns the updating of γ, let Λ the $N \times N$ diagonal matrix with (j,j)-th entry equal to λ_j. After some computations we obtain

$$p_{\gamma|a^N,y}\left(\gamma \mid a^N, y\right) = \Gamma\left(\frac{N}{2} + \alpha \ , \ \frac{1}{2}\left(a^N\right)' \Lambda^{-1}\left(a^N\right) + \beta\right).$$

We then choose as the proposal density for γ its conditional distribution given a^N and y, thus defining a Gibbs sampler update, see [21].

As concerns a^N, the same kind of strategy can no more be exploited, since the corresponding posterior does not take a standard form. We then propose a move in which these parameters are updated by sampling a value in a symmetric interval around the current position, in accordance with a Gaussian distribution with covariance matrix Σ. In particular, we adapt the proposal scales as follows. We begin a pilot-tuning run from some arbitrary values γ_0 and a_0^N. We preliminarily set Σ to a matrix proportional to $\gamma_0^{-1}\Lambda$, with the scale factor chosen in order to make the acceptance ratio for a^N to be around 0.2–0.4. This preliminary stage is used in rder to let the algorithm approximately learn the a posteriori correlation of the components of a^N given y. Then, after a certain number of iterations, we set once and for all Σ as proportional to the covariance matrix of the generated samples of a^N, still choosing a scale factor which ensures an acceptance ratio of the proposed moves around 0.3 [21].

We summarize the MCMC procedure by means of the following Algorithm 1. The symbol $\mathcal{N}(\mu, \Sigma)$ is therein used to denote a Gaussian probability density function having mean μ and covariance matrix Σ.

Remark 6. Ψ, as well as the RKHS H, should be chosen in accordance with available information on f. Moreover, as concerns the transformation $\Psi(g)/\int \Psi(g)dx$, it appears important to define it as injective. Otherwise, the posterior probability of a^N given y could turn out multimodal with many peaks (or also improper), thus making difficult the convergence of the generated Markov chain.

Algorithm 1

(*initialization*) set (γ_0, a_0^N) and $k = 1$

for $k = 1, 2, \ldots$ **do**

sample γ_k from

$$\Gamma\left(\frac{N}{2} + \alpha, \frac{1}{2}\left(a_{k-1}^N\right)' \Lambda^{-1}\left(a_{k-1}^N\right) + \beta\right)$$

sample s from $\mathcal{N}\left(a_{k-1}^N, \Sigma\right)$

accept s with probability

$$\delta\left(s, a_{k-1}^N, \gamma_k\right) = \min\left(1, \frac{\rho(s, \gamma_k)}{\rho\left(a_{k-1}^N, \gamma_k\right)}\right)$$

where

$$\rho\left(a^N, \gamma\right) := \frac{\prod_{i=1}^n \Psi(g_a^N)_{y_i}}{\int_X \Psi(g_a^N) dx} \exp\left(-\frac{\gamma}{2}(a^N)' \Lambda^{-1} (a^N)\right)$$

if s is accepted then set $a_k^N = s$, otherwise $a_k^N = a_{k-1}^N$

VI. Numerical experiments

We present two set of univariate simulations on $[0, 1]$ to examine the relative effectiveness of the proposed methodology. Before introducing the examples in detail, it is worth pointing out that in every case study the binary control of Raftery and Lewis has been used in order to assess the convergence of the generated Markov chains, see [32]. In particular we have always required to estimate the quantiles 0.025, 0.25, 0.5, 0.75, 0.975 with precision respectively 0.005, 0.01, 0.01, 0.01, 0.005, and with probability 0.95.

A. Bayesian learning of an exponential probability density function

We start considering the reconstruction of an exponential density, with mean 0.15, from 100 samples independently drawn from it. Data are displayed by means of a histogram in Fig. 2 (top panel).

In (1) we include information about the smoothness of the unknown function f by setting Ψ to an exponential transform and assuming that g belongs to W_1, with support on[1] $[0, 1]$. We recall that this exponential transformation is the nonparametric counterpart of the classic log-normal choice in parametric frameworks.

As concerns γ, α and β are chosen so as to define a poorly informative prior on this hyper-parameter. After setting N to 100 in (3), the model in Fig. 1 has been solved by reconstructing the joint posterior density of γ and f via a Monte Carlo run where 4000 samples were generated. In Fig. 3 (top panel) the (unnormalized) posterior of γ is depicted. In Fig. 4 (top panel) we report the minimum variance estimate of f (solid line), together with its 95% confidence interval (shaded area) and the true function

[1]It is easy to assess that this kind of choice makes $\Psi(g)/\int \Psi(g)dx$ injective. This in particular holds thanks to the side condition at 0 present in the definition of W_1.

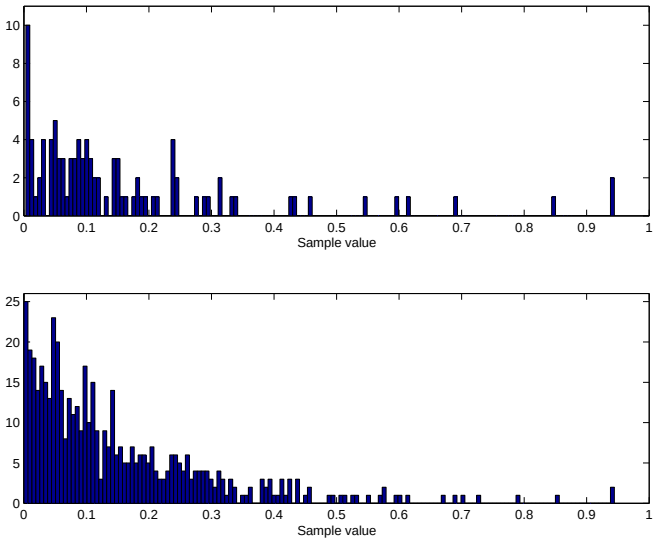

Fig. 2. Reconstruction of an exponential density function. *Top:* histogram of 100 samples independently drawn from the unknown density. *Bottom:* histogram of 500 samples independently drawn from the unknown density.

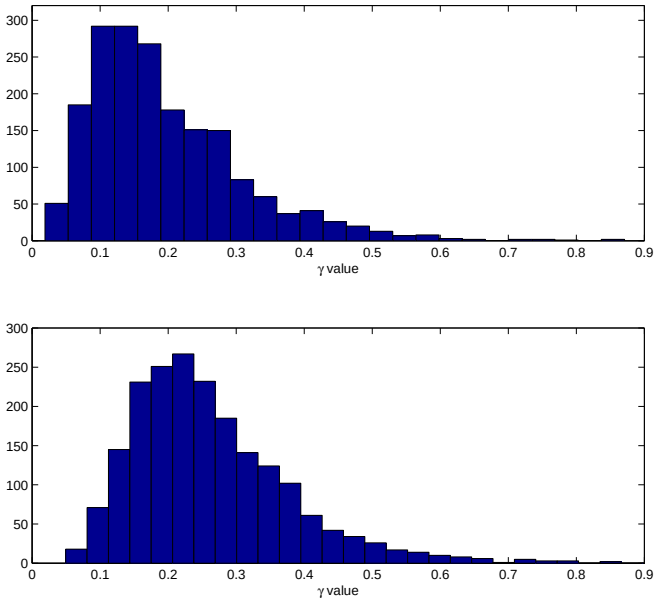

Fig. 3. Reconstruction of an exponential density function. *Top:* posterior of γ obtained (in sampled form) by MCMC using the training set of 100 samples. *Bottom:* posterior of γ obtained (in sampled form) by MCMC using the training set of 500 samples.

(dashed line). Even though the training set size is small, the density estimate is somewhat close to the true one. It thus appears that a suitable amount of regularization has been introduced by the nonlinear estimator.

We then repeated the entire estimation process by adding to the training set other 400 samples independently drawn from the exponential density. Data are displayed by means of a histogram in Fig. 2 (bottom panel). After setting N to 100 in (3), a Monte Carlo run of 3000 samples were generated. The posterior of γ (in sampled form) is visible in Fig. 3 (bottom panel). Results regarding the estimate of f are then

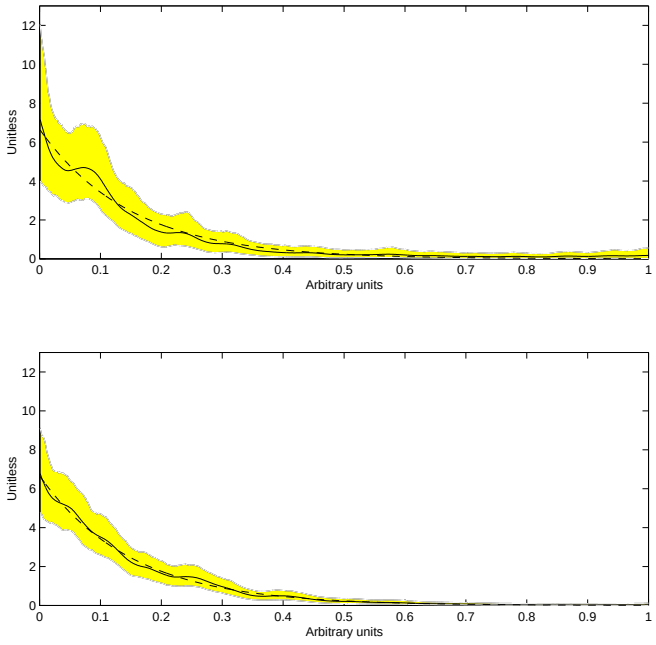

Fig. 4. Reconstruction of an exponential density function. *Top:* minimum variance estimate (continuous line) with 95% confidence interval (shaded area) and true density (dashed line) using the training set of 100 samples. *Bottom:* minimum variance estimate (continuous line) with 95% confidence interval (shaded area) and true density (dashed line) using the training set of 500 samples.

reported in Fig. 4 (bottom panel), with the same rationale followed in the top panel of the same figure. One can note that the minimum variance estimate is very close to the truth. Moreover, comparing the confidence intervals depicted in the first and top panel, one can appreciate how the uncertainty related to the estimate has been reduced by augmenting the size of the data set. This illustrates the capability of the proposed approach in clearly making the investigator assess the amount of information that different training sets provide.

B. Bayesian learning of a mixture of Gaussians

As a second example, we consider a benchmark problem proposed in [9] which consists of reconstructing a density on $[0, 1]$ proportional to

$$\frac{1}{3}e^{-50(x-0.3)^2} + \frac{2}{3}e^{-50(x-0.7)^2} \ .$$

This function is depicted in Fig. 5 (dashed line), and it can be noticed that it virtually corresponds to a mixture of Gaussians.

In (1) we include information about the smoothness of the unknown function f by setting Ψ as in Sec. VI-A and assuming that g belongs to W_2, with support on $[0, 1]$. As concerns α and β, they are chosen so as to define a poorly informative prior on γ. After setting N to 100 in (3), we consider 300 replicates of this problem. For each of these 300 simulations we generate a new training set of 200 samples and then use our MCMC scheme to obtain the minimum variance estimate of f, in accordance with the model depicted in Fig. 1. In Fig. 5 we report the mean of

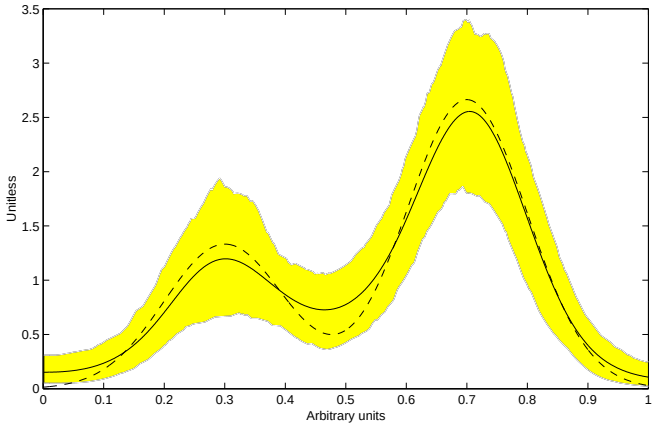

Fig. 5. Reconstruction of a mixture of Gaussians through Monte Carlo simulations: estimates mean (continuous line), with 95% bands of variability (shaded area), and true density (dashed line).

the 300 estimates (continuous line) together with the 95% variability band (shaded area), i.e. the interval between the 2.5 and 97.5 percentiles of the Monte Carlo distribution of the density estimates at each point of their support. Even though the size of the data set is really small, one can note that the mean is quite close to the true profile. This is particularly evident near the second peak since, on average, the most part of the samples is generated from it at every of the 300 Monte Carlo runs. Moreover, 95% variability bands show that the variance of the error is not high.

VII. Conclusions

The choice of the regularization parameter is a crucial issue in learning theory and, more in general, when dealing with ill-posed problems [33], [18], [34], [7]. In particular, when reconstructing a probability density function, its determination presents formidable difficulties due to the nonlinear nature of the problem. In this paper, we have proposed a new technique which tackles this difficulty by casting the density estimation problem within a Bayesian framework. An MCMC approach is then used to implement the resulting stochastic model.

The power of our method consists in providing estimates that take into account all the sources of uncertainty present in the problem. In particular, the proposed algorithm is able to return minimum variance estimates of both the unknown density and the regularization parameter. In addition, our scheme can associate to the estimate a confidence interval, thus allowing the investigator to assess how informative the training set is.

In future work, we will test the methodology to reconstruct multivariate densities, also improving its computational efficiency.

References

[1] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer Academic Publishers, 1992.

[2] A. S. Weigend and A. N. Srivastava, "Predicting conditional probability distributions: a connectionist approach," *International Journal of Neural Systems*, vol. 6, no. 2, pp. 109 – 118, June 1995.

[3] S. Fiori, "Nonsymmetric pdf estimation by artificial neurons: Application to statistical characterization of reinforced composites," *IEEE Transactions on Neural Networks*, vol. 14, no. 4, pp. 959 – 962, July 2003.

[4] F. Liang, "Continuous contour monte carlo for marginal density estimation with an application to spatial statistical model," *Journal of Computational and Graphical Statistics*, vol. 16, no. 3, pp. 608 – 632, September 2007.

[5] B. Silverman, "On the estimation of a probability density function by the maximum penalized likelihood method," *Annals of Statistics*, vol. 10, pp. 795 – 810, 1982.

[6] V. N. Vapnik, *Statistical learning theory*. New York: Wiley, 1998.

[7] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American mathematical society*, vol. 39, pp. 1–49, 2001.

[8] V. Krylov, G. Moser, S. B. Serpico, and J. Zerubia, "On the method of logarithmic cumulants for parametric probability density function estimation," INRIA Sophia Antipolis, Tech. Rep. RR-7666, 2011.

[9] C. Gu, "Smoothing spline density estimation: A dimensionless automatic algorithm," *Journal of the American Statistical Association*, vol. 88, pp. 495 – 504, June 1993.

[10] A. Komárek and E. Lesaffre, "Generalized linear mixed model with a penalized gaussian mixture as a random effects distribution," *Computational Statistics & Data Analysis*, vol. 52, no. 7, pp. 3441 – 3458, March 2008.

[11] C. Schellhase and G. Kauermann, "Density estimation and comparison with a penalized mixture approach," *Computational Statistics*, vol. –, pp. 1 – 21, 2011.

[12] G. Wahba, "Histosplines with knots which are order statistics," *Journal of the Royal Statistical Association*, vol. 38, pp. 140 – 151, 1976.

[13] E. Parzen, "On the estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.

[14] E. López-Rubio and J. M. O. de Lazcano-Lobato, "Soft clustering for nonparametric probability density function estimation," *Pattern Recognition Letters*, vol. 29, pp. 2085 – 2091, 2008.

[15] C. Archambeau and M. Verleysen, "Fully nonparametric probability density function estimation with finite gaussian mixture models," in *International Conference on Advances in Pattern Recognition*, 2003.

[16] M. Girolami, "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Computation*, vol. 14, pp. 669 – 688, 2002.

[17] Y. Yang, "Penalized semiparametric density estimation," *Statistics and Computing*, vol. 19, no. 4, pp. 355 – 366, 2009.

[18] G. Wahba, *Spline models for observational data*. SIAM, Philadelphia, 1990.

[19] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. Washington, D.C.: Winston/Wiley, 1977.

[20] W. Hastings, "Monte Carlo sampling methods using Markov chain and their applications," *Biometrika*, vol. 57, pp. 97–109, 1970.

[21] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov chain Monte Carlo in Practice*. London: Chapman and Hall, 1996.

[22] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral eqations," *Phil. Trans. Roy. Soc. London Ser.*, vol. 209, pp. 415–446, 1909.

[23] R. Courant and D. Hilbert, *Methods of mathematical physics*. Interscience, 1962.

[24] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337 – 404, 1950.

[25] G. De Nicolao and G. Ferrari-Trecate, "Consistent identification of NARX models via regularization networks," *IEEE Transactions on Automatic Control*, vol. 44, pp. 2045–2049, 1999.

[26] A. M. Yaglom, *Correlation theory of stationary and related random functions*. Springer-Verlag, New York, 1987, vol. 1.

[27] P. Magni, R. Bellazzi, and G. De Nicolao, "Bayesian function learning using MCMC methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1219–1331, 1998.

[28] L. Liu, M. Levine, and Y. Zhu, "A functional em algorithm for mixing density estimation via nonparametric penalized likelihood maximization," *Journal of Computational and Graphical Statistics*, vol. 18, no. 2, pp. 481 – 504, 2009.

[29] C. Gu and C. Qiu, "Smoothing spline density estimation: theory," *Annals of Statistics*, vol. 21, pp. 227–234, 1993.

[30] C. E. Rasmussen, "The infinite Gaussian mixture model." in *Advances in Neural Information Processing Systems (NIPS)*, vol. 12, 2000, pp. 554–560.

[31] G. Pillonetto and B. Bell., "Bayes and empirical Bayes semi-blind deconvolution using eigenfunctions of a prior covariance," *Automatica*, vol. 43, no. 10, pp. 1698–1712, 2007.

[32] A. E. Raftery and S. M. Lewis, *Implementing MCMC*. Markov Chain Monte Carlo in Practice. W.R. Gilks, S.Richardson, and D.J. Spiegelhalter, eds. London: Chapman and Hall, 1996, pp. 115–130.

[33] J. A. Rice, "Choice of smoothing parameter in deconvolution problems," *Contemporary Mathematics*, vol. 59, pp. 137–151, 1986.

[34] P. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM Review*, vol. 34, pp. 561 – 580, 1992.