Assessment of Nonlinear Dynamic Models by Kolmogorov–Smirnov Statistics

Petar M. Djuric, *Fellow, IEEE*, and Joaquín Míguez, *Member, IEEE*

***Abstract*—Model assessment is a fundamental problem in science and engineering and it addresses the question of the validity of a model in the light of empirical evidence. In this paper, we propose a method for the assessment of dynamic nonlinear models based on empirical and predictive cumulative distributions of data and the Kolmogorov–Smirnov statistics. The technique is based on the generation of discrete random variables that come from a known discrete distribution if the entertained model is correct. We provide simulation examples that demonstrate the performance of the proposed method.**

***Index Terms*—Cumulative distributions, Kolomogorov–Smirnov statistics, model assessment, particle filtering, predictive distributions.**

I. Introduction

In many science and engineering applications, we use models, which we consider correct, and we make decisions and/or draw conclusions based on the models. If we have a set of models to choose from and we do not know which model is the best one (given a certain criterion), we proceed with a model selection procedure [2], [20]. Often, however, we have only one model, and then we may simply want to know whether the model is good or not. We refer to this problem as *model assessment*. If the model is not good, we continue by proposing a modified one. It is clear that the assessment of the plausibility of a model is a fundamental problem. Areas where it is of great interest include robotics, telecommunications, control, speech processing, climatology, epidemiology, geology, astrophysics, econometrics, and finance.

Suppose that we have a model $\mathcal{M}_0$ for a vector of observations $\boldsymbol{y}$ described by the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta},\ \mathcal{M}_0)$ and the prior $p(\boldsymbol{\theta}|\mathcal{M}_0)$, where $\boldsymbol{\theta}$ is a vector of model parameters. For the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{\theta}$ given the model $\mathcal{M}_0$, we can write

$$p(\boldsymbol{y},\boldsymbol{\theta}|\mathcal{M}_0) \propto p(\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M}_0)\ p(\boldsymbol{y}|\mathcal{M}_0) \tag{1}$$

where the symbol $\propto$ denotes proportionality. The first factor on the right side of the proportionality sign in (1) is the posterior of $\boldsymbol{\theta}$ given the observed data $\boldsymbol{y}$ and the model $\mathcal{M}_0$, and the second one is the predictive distribution of the data given the model $\mathcal{M}_0$. In much of the literature the emphasis has been on the first factor, which is essential for *parameter estimation*. We point out that even if the model $\mathcal{M}_0$ were incorrect, it would be impossible to detect its incorrectness from the first factor [5]. The second factor $p(\boldsymbol{y}|\mathcal{M}_0)$, however, can be used for model assessment. In particular, one would seek some function of the data $\boldsymbol{y}$, $f(\boldsymbol{y})$,[1] and would construct statistics of that function under the assumption that the model is correct. For instance, the function could simply be the sample average or a moment coefficient. The underlying idea is that $f(\boldsymbol{y})$ would have unusual values if the model is not correct [5].

There is a class of tests for model assessment which are based on statistics derived from empirical distribution functions [28]. In this paper we propose to use one of them, the Kolmogorov–Smirnov (KS) statistic, which we apply to the construction of $f(\boldsymbol{y})$. Our method is based on processing the observations one at a time. After an observation becomes available, we transform it to produce a KS distance measure, and all the distances computed up to that time are used for testing whether the model should be rejected.

Mathematically, we formulate the problem as follows. We define the model $\mathcal{M}_0$ by using the set of equations

$$\boldsymbol{x}_t = g(\boldsymbol{x}_{t-1}, \boldsymbol{u}_t) \tag{2}$$

$$y_t = h(\boldsymbol{x}_t, \boldsymbol{v}_t) \tag{3}$$

where

$t \in \mathbb{N}$	is a discrete time index;
$\boldsymbol{x}_t \in \mathbb{R}^{n_x}$	is the unknown state of the system at time instant t;
$y_t \in \mathbb{R}$	is a scalar observation;
$\boldsymbol{u}_t \in \mathbb{R}^{n_u}$	is a state noise vector with known distribution;
$\boldsymbol{v}_t \in \mathbb{R}^{n_v}$	is an observation noise vector with known distribution;

[1]We note that $f(\cdot)$ is a function of the data $\boldsymbol{y}$ and not a probability distribution function of $\boldsymbol{y}$.

Manuscript received August 15, 2009; accepted May 22, 2010. Date of publication June 21, 2010; date of current version September 15, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tryphon Georgiou. Most of the work of P. M. Djuric was carried out while he held the position of Chair of Excellence of Universidad Carlos III de Madrid-Banco de Santander. His work has also been supported by the National Science Foundation under Award CCF-0515246 and the Office of Naval Research under Award N000140610012. The work of J. Míguez has been supported by the Ministry of Science and Innovation of Spain (Project TEC-2006-13514-C02-01/TCM MONIN, Project TEC2009-14504-C02-01 DEIPRO and program Consolider-Ingenio 2010 project CSD2008-00010 COMONSENS). This work is a continuation of the work presented at the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, 2009.

P. M. Djuric is with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: djuric@ece.sunysb.edu).

J. Míguez is with the Departamento de Teoría de la Señal y Comunicaciones, Universidad Carlos III de Madrid, Leganés, 28911 Madrid, Spain (e-mail: joaquin.miguez@uc3m.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2010.2053707

$g : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \longrightarrow \mathbb{R}^{n_x}$ is a known function;

$h : \mathbb{R}^{n_x} \times \mathbb{R}^{n_v} \longrightarrow \mathbb{R}^{n_y}$ is a known function.

We also assume that the initial distribution of the state $p(\boldsymbol{x}_0)$ is known. The objective is to develop a sequential methodology for assessing the validity of $\mathcal{M}_0$.

We point out that a more general description of the model would include unknown constant parameters both in the state and observation equations. Here we do not include them for presentation purposes and for avoiding distraction from the main contribution of the paper.

The remaining of the paper is organized as follows. In Section II, we briefly review some standard model validation methods from the literature and give some basic background on the Kolmogorov-Smirnov family of tests. We introduce the proposed method in Section III and follow up with a discussion of some relevant implementation issues in Section IV. In Section V, we show illustrative computer simulation results and, finally, in Section VI we provide conclusions of our work.

II. Background

In this section, we first review some methods for model assessment and then recall some basics of the KS statistics.

A. Methods for Model Assessment

The classical approaches for model assessment are based on standard test statistics derived from reference distributions that are independent of the unknown model parameters. The literature on model assessment is rich, especially on assessing linear regression and generalized linear models, where the test is often known as test for goodness of fit.

In this paper, the objective is to develop a methodology for assessing dynamic models. Unlike the assessment of linear regression and generalized linear models, the assessment of dynamic models has received much less attention. A quantity that plays an important role in assessments is the predictive distribution of the observations. In [14], the statistics for assessment are given by the normalized and recursive residuals defined by

$$r_t = \frac{y_t - E(Y_t|y_{1:t-1}, \mathcal{M}_0)}{\sigma_{y_t|y_{1:t-1},\mathcal{M}_0}}$$

where $E(Y_t|y_{1:t-1}, \mathcal{M}_0)$ is the expectation of Y_t given the past observations $y_{1:t-1}$ and the model $\mathcal{M}_0$, and $\sigma_{y_t|y_{1:t-1},\mathcal{M}_0}$ is the conditional standard deviation of Y_t.[2] Under the hypothesis that the model $\mathcal{M}_0$ generates the data, the residuals are independent and identically distributed (i.i.d.) samples from a standard Gaussian distribution. This entails that one can use one of the numerous tests that exist in the literature for deciding whether the r_ts are indeed distributed as predicted by the model. For example, if $\mathcal{M}_0$ is correct, the sum of squared residuals

$$\rho_t = \sum_{\tau=1}^{t} r_\tau^2$$

has a χ^2 distribution with t degrees of freedom. In testing ρ_t, one can then apply one of several tests, for instance, Pearson's χ^2 test [21]. For more on the subject of testing the residuals r_t, see [6].

In [27], an assessment method was proposed based on the one-to-one transformation of the distribution $F(y_{1:t}|\mathcal{M}_0)$ into a uniform distribution on the $t-$ dimensional hypercube via [23]

$$\psi_t = F(y_t|y_{1:t-1}, \mathcal{M}_0) \tag{4}$$

where the ψ_ts are outcomes of independent uniform random variables on [0, 1], and $F(y_t|y_{1:t-1}, \mathcal{M}_0) = P(Y_t \leq y_t|y_{1:t-1}, \mathcal{M}_0)$ is the one-step-ahead predictive cumulative distribution function (CDF) of Y_t given the observations $y_{1:t-1}$ and the model $\mathcal{M}_0$.[3] The uniform random variables are then converted to Gaussians by the transformation $Z_t = \Phi^{-1}(\Psi_t)$, where $\Phi(\cdot)$ stands for the standard Gaussian CDF. Thus, if the model is correct, we have a sequence of i.i.d. standard normal random variables [23]. It is important to note that this is true even if the CDF of Y_t in (4) is not Gaussian. For testing the Gaussianity of Z_t one may apply the Shapiro–Wilk [24], [25], or the Bowman–Shenton [4] tests, or, for testing independence, the Ljung–Box test [17].

The work from [27] was also extended in [15], where it was assumed that the model had additional unknown parameters. The unknowns had to be integrated out, and for integration, the authors proposed the use of Markov chain Monte Carlo (MCMC) methods. MCMC sampling had to be employed at every time instant, which is impractical for large data sets.

The method from [27] was further exploited in [11], where the standard normal random variables used for testing the adequacy of the model were estimated by MCMC and importance sampling. The advantage of [11] over [15] is that MCMC sampling is run only for a small percentage of the data.

Further use of the method from [27] can be found in [15] and [29] in the context of modeling stochastic volatility and speech signals, respectively. There, the estimation of $P(Y_t \leq y_t|y_{1:t-1}, \mathcal{M}_0)$ was implemented by sequential Monte Carlo methods. A similar approach was also suggested in [1].

B. Kolmogorov–Smirnov Tests

The classical one-dimensional KS tests belong to the class of nonparametric tests, and they measure disagreement between two CDFs by using the largest absolute difference between them [3], [22]. These distances are referred to as KS statistics. In general, there are two types of KS tests, known as one-sample and two-sample KS tests. The former create an empirical CDF from available data and compare it to a reference CDF. The latter form empirical CDFs from two sets of data and compare the obtained CDFs. Under the null hypothesis, one can obtain the distribution of the KS statistic and then apply it for testing the null hypothesis.

In our work, we use the two-sample KS test and therefore we explain it in more detail. Let us consider two sets of samples $y^{(1)}, y^{(2)}, \ldots, y^{(L)}$ and $\tilde{y}^{(1)}, \tilde{y}^{(2)}, \ldots, \tilde{y}^{(K)}$, which are i.i.d. according to some continuous probability distributions $F(y)$ and $G(y)$, where $F(\cdot)$ and $G(\cdot)$ denote the CDFs of Y and $\tilde{Y}$, respectively. We want to test the null hypothesis $\mathcal{H}_0$, where

$$\mathcal{H}_0 : F(y) = G(y).$$

[2]We denote scalar random variables with capital letters and their values with small letters. The notation $y_{1:t-1}$ signifies the set of observations $\{y_1, y_2, \ldots, y_{t-1}\}$.

[3]The notation $P(A)$ indicates the probability of the event A.

We note that the test does not postulate anything about the distributions of the data except that they are continuous distributions.

From the samples, we construct the empirical distributions $\hat{F}^{[L]}(\cdot)$ and $\hat{G}^{[K]}(\cdot)$, where L and K are the numbers of samples available for constructing the respective distributions, and test for their agreement by using the KS statistic defined by

$$D^{[L,K]} = \sup_y |\hat{F}^{[L]}(y) - \hat{G}^{[K]}(y)|.$$

The hypothesis of equal CDFs is rejected at level α if [22]

$$\sqrt{\frac{LK}{L+K}} D^{[L,K]} \geq \gamma_\alpha$$

where γ_α is a threshold that satisfies

$$P\left(\sqrt{\frac{LK}{L+K}} D^{[L,K]} \geq \gamma_\alpha\right) = \alpha.$$

One can obtain the relevant thresholds from tables. In [18] there are tables for small L and K, and from [26] one can use limiting results to obtain thresholds for large L and K.

In the rest of the paper, we use data sets with $L = 1$ and $K > 1$, and so we study this case. Even though it may seem disadvantageous to apply the KS test in this extreme scenario, we will see that, in fact, it provides some benefits. Since in the sequel we always have $L = 1$ and K is known from the context, we simplify the notation and denote the KS distance $D^{[1,K]}$ simply as D.

III. Model Assessment

A. Preliminary Results

Before we proceed, we state two propositions, where we assume that the data $y^{(1)}$ and $\tilde{y}^{(1)}, \tilde{y}^{(2)}, \ldots, \tilde{y}^{(K)}$ come from continuous distributions.

Proposition 1: The support of the KS statistic D is given by $0.5 \leq D \leq 1$.

Proof: Let $y^{(1)}$ denote the single sample from which the empirical CDF $\hat{F}^{[1]}(y)$ is constructed. Since $\hat{F}^{[1]}(y) = 0$ for all $y < y^{(1)}$ and $\hat{F}^{[1]}(y) = 1$ for all $y \geq y^{(1)}$, it is apparent that

$$\begin{aligned} D &= \sup_y \left|\hat{F}^{[1]}(y) - \hat{G}^{[K]}(y)\right| \\ &= \max\left\{\hat{G}^{[K]}(y^{(1)}), 1 - \hat{G}^{[K]}(y^{(1)})\right\}. \end{aligned}$$

The equation above directly implies that $D \leq 1$, because $0 \leq \hat{G}^{[K]}(y) \leq 1$ for all y. Moreover, $\hat{G}^{[K]}(y^{(1)})$ and $1 - \hat{G}^{[K]}(y^{(1)})$ cannot be smaller than 0.5 simultaneously, hence $D \geq 0.5$. □

Proposition 2: If $y^{(1)}$ and $\tilde{y}^{(k)}$, $k = 1, 2, \ldots, K$ are i.i.d. samples from the same distribution, we have the equation shown at the bottom of the page.

Proof: Let $\mathsf{A} = \{\tilde{y}^{(k)}, k \in \{1, \ldots, K\} : \tilde{y}^{(k)} < y^{(1)}\}$ be the set of samples that are smaller than $y^{(1)}$ and let $a = |\mathsf{A}|$ be the number of such samples (obviously, $a \in \{0, 1, \ldots, K\}$). Since $\tilde{y}^{(1)}, \ldots, \tilde{y}^{(K)}, y^{(1)}$ are i.i.d. samples, the values of a are, *a priori*, equally probable. In particular, there are $(K+1)!$ different ways in which the samples $\tilde{y}^{(1)}, \ldots, \tilde{y}^{(K)}, y^{(1)}$ can be ordered and the value of a depends on the relative position of $y^{(1)}$ after the samples are sorted out (e.g., if $y^{(1)}$ is the smallest sample, then $a = 0$ and if there is exactly one $\tilde{y}^{(i)} < y^{(1)}$, then $a = 1$). However, given the relative position of $y^{(1)}$, there are still $K!$ different ways in which the remaining samples can be ordered. Therefore, $P(a = n) = K!/(K+1)! = 1/(K+1)$ for every $n \in \{0, 1, \ldots, K\}$, i.e., a is a realization of a discrete uniform random variable, A.

Moreover, the realization $A = a$ determines the value of $\hat{G}^{[K]}(y^{(1)})$, i.e., $\hat{G}^{[K]}(y^{(1)}) = a/K$, and, as a consequence, the value of D. To be precise, assume first that K is odd. Then

$$\begin{aligned} D &= \max\left\{\hat{G}^{[K]}(y^{(1)}), 1 - \hat{G}^{[K]}(y^{(1)})\right\} \\ &= \max\left\{\frac{a}{K}, \frac{K-a}{K}\right\} \end{aligned}$$

which implies that there are two different outcomes of A that result in the same value of D. Specifically, for every $n \geq (K+1)/2$, both $A = n$ and $A = K - n$ yield $D = n/K$. Therefore,

$$P\left(D = \frac{n}{K}\right) = P(A = n) + P(A = K - n) = \frac{2}{K+1}.$$

If we assume that K is even, the same argument can be applied for any $n \geq K/2+1$, hence $P(D = n/K) = 2/(K+1)$, but the value $D = 0.5$ can only be achieved when $A = K/2$, which implies $P(D = 0.5) = 1/(K+1)$. □

Thus, under $\mathcal{H}_0$ and $K > 1$ and odd, according to Proposition 2, the KS statistic is a uniform discrete random variable on the support $\mathcal{S} = \{(K+1)/2K, (K+3)/2K, \ldots, K\}$, whereas if $K > 1$ and even, it is almost uniform. These distributions provide us with various possibilities for testing the hypothesis $\mathcal{H}_0$, as will be shown in the sequel.

B. Use of the KS Statistics for Model Assessment

In our problem, we process the data $y_{1:t}$ sequentially as they are observed. The data are modeled by (2) and (3), and typical processing, using stochastic filtering methods, amounts to tracking the posterior distribution of the state, $p(\boldsymbol{x}_{0:t}|y_{1:t}, \mathcal{M}_0)$. Let us assume for a moment that we can generate independent

$$K = 2l+1, l \geq 1: \quad P\left(D = \frac{n}{K}\right) = \frac{2}{K+1}, \quad n = \frac{K+1}{2}, \frac{K+3}{2}, \ldots, K-1, K$$

$$K = 2l, l \geq 1: \quad P\left(D = \frac{n}{K}\right) = \begin{cases} \frac{2}{K+1}, & \text{if } n = \frac{K}{2}+1, \ldots, K-1, K \\ \frac{1}{K+1}, & \text{if } n = \frac{K}{2} \end{cases}.$$

TABLE I
SUMMARY OF THE PROPOSED ASSESSMENT METHOD

At time instant t,
1. generate particles $\boldsymbol{x}_t^{(j)}$, $j = 1, ..., J$, from the predictive distribution $p(\boldsymbol{x}_t|y_{1:t-1}, \mathcal{M}_0)$ according to (6),
2. generate samples $\tilde{y}_t^{(k)}$, $k = 1, ..., K$, from the predictive distribution $p(y_t|y_{1:t-1}, \mathcal{M}_0)$ according to (7),
3. sort all the $\tilde{y}_t^{(k)}$s, and compute the KS statistic d_t according to (5), and
4. test the validity of the model using $d_{1:t}$.

samples from $p(y_t|y_{1:t-1}, \mathcal{M}_0)$ (which we view as fictitious observation data), i.e.,

$$\tilde{y}_t^{(k)} \sim p(y_t|y_{1:t-1}, \mathcal{M}_0), \quad k = 1, 2, \ldots, K$$

where

$$p(y_t|y_{1:t-1}, \mathcal{M}_0) = \int p(y_t|\boldsymbol{x}_t, \mathcal{M}_0) p(\boldsymbol{x}_t|y_{1:t-1}, \mathcal{M}_0) d\boldsymbol{x}_t.$$

We reiterate that we have only one actual observation at time t, while $\tilde{y}_t^{(1)}, \ldots, \tilde{y}_t^{(K)}$ are simulated. We note that the observed y_t stands for $y^{(1)}$.

Let us compute

$$d_t = \sup_y \left| \hat{F}_t^{[1]}(y), \hat{G}_t^{[K]}(y) \right| = \max \left\{ \hat{G}_t^{[K]}, 1 - \hat{G}_t^{[K]} \right\} \quad (5)$$

where $\hat{F}_t^{[1]}$ is the empirical CDF generated by the single observation y_t, while $\hat{G}_t^{[K]}$ is the empirical CDF obtained from the simulated samples $\tilde{y}_t^{(1)}, \ldots, \tilde{y}_t^{(K)}$. When the model is correct, d_t is a realization of the random variable D_t, with known probability distribution given by Proposition 2. Therefore, we can assess the degree of compatibility of the assumed model $\mathcal{M}_0$ with the data $y_{1:t}$ by performing various tests on $d_{1:t}$, which form a sequence of i.i.d. samples when $\mathcal{M}_0$ is correct. One conceptually straightforward possibility is to apply a one-sample test, such as, e.g., Pearson's χ^2 test, to the samples $d_{1:t}$ (note that the reference distribution given by Proposition 2 is discrete). However, in practice it may be simpler and more useful to design a test that can be applied, at every time t, to a scalar statistic computed from the samples $d_{1:t}$ available up to that time. In [9], we used the central limit theorem and the fact that, for even moderate t, the random variable

$$\bar{D}_t = \frac{1}{t} \sum_{\tau=1}^{t} D_\tau$$

is distributed approximately as a Gaussian with known mean and variance. Alternative approaches exist, however, that do not involve approximations. In Section V, we show that the distribution of $\bar{D}_t$ can be found exactly and we design a simple sequential test for obtaining exact p-values of the model hypothesis.

C. Simulation of Samples From $\mathcal{M}_0$

All that remains is to show how we can actually simulate the samples $\tilde{y}_t^{(1)}, \tilde{y}_t^{(2)}, \ldots, \tilde{y}_t^{(K)}$, at every time t, according to the assumed model $\mathcal{M}_0$. To that end, we use particle filtering as follows. Recall that with particle filtering we process the data $y_{1:t}$ sequentially and obtain an approximation of the posterior of $\boldsymbol{x}_t$ using a random measure $\chi_t = \{\boldsymbol{x}_t^{(m)}, w_t^{(m)}\}_{m=1}^M$, where M is the total number of particles, and $\boldsymbol{x}_t^{(m)}$ and $w_t^{(m)}$ are the particles and weights, respectively, that define the random measure [8], [10]. For the predictive distribution $p(y_t|y_{1:t-1}, \mathcal{M}_0)$ we can formally write

$$p(y_t|y_{1:t-1}, \mathcal{M}_0) = \int p(y_t|\boldsymbol{x}_t, y_{1:t-1}, \mathcal{M}_0) p(\boldsymbol{x}_t|y_{1:t-1}, \mathcal{M}_0) d\boldsymbol{x}_t.$$

If the predictive distribution $p(\boldsymbol{x}_t|y_{1:t}, \mathcal{M}_0)$ is approximated by

$$p(\boldsymbol{x}_t|y_{1:t-1}, \mathcal{M}_0) \simeq \sum_{m=1}^{M} w_{t-1}^{(m)} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(m)}, \mathcal{M}_0)$$

one can readily generate samples $\boldsymbol{x}_t^{(j)}$ from it, provided it is easy to draw from $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(m)}, \mathcal{M}_0)$, which we assume here. Thus, we sample according to

$$\boldsymbol{x}_t^{(j)} \sim \sum_{m=1}^{M} w_{t-1}^{(m)} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(m)}, \mathcal{M}_0), \quad j = 1, 2, \ldots, J \quad (6)$$

which then allows us to draw $\tilde{y}_t^{(k)}$, $k = 1, 2, \ldots, K$ according to

$$\tilde{y}_t^{(k)} \sim \frac{1}{J} \sum_{j=1}^{J} p(y_t \,|\, \boldsymbol{x}_t^{(j)}, \mathcal{M}_0). \quad (7)$$

Therefore, the generation of samples of y_t is a two-step procedure where in the first step we obtain J samples of $\boldsymbol{x}_t$ according to (6) and in the second step we draw K samples of y_t using (7), making a total of $J + K$ samples.

In summary, at every time instant t, the method is implemented in four steps, as shown in Table I.

The first two steps represent simulation of $\tilde{y}_t$ based on $\mathcal{M}_0$, the third step is the computation of the KS distances d_t, and the fourth step is testing the validity of the model from the computed d_ts.

IV. DISCUSSION

In this section, we discuss some issues related to the implementation and complexity of the proposed method. In this paper, we refer to the assessment methods based on estimating the probabilities $P(Y_t \leq y_t|y_{1:t-1}, \mathcal{M}_0)$ as standard tests. They need to implement the following integration:

$$\psi_t = \int P(Y_t \leq y_t|\boldsymbol{x}_t, \mathcal{M}_0) p(\boldsymbol{x}_t|y_{1:t-1}) d\boldsymbol{x}_t. \quad (8)$$

This integration can be carried out analytically only in a small number of situations. When a closed form solution in (8) is not available, one has to perform the estimation of ψ_t by Monte

Carlo methods. For example, the authors of [11] propose to a) generate samples of $\boldsymbol{x}_t^{(j)}, j = 1, 2, \ldots, J$, from $p(\boldsymbol{x}_t|y_{1:t-1})$, b) for each $\boldsymbol{x}_t^{(j)}$, draw K samples $y_t^{(k)}$ from $p(y_t|\boldsymbol{x}_t^{(j)})$, $k = 1, \ldots, K$, and c) compute an empirical estimate of $P(Y_t \leq y_t|\boldsymbol{x}_t^{(j)}, \mathcal{M}_0)$, denoted $\hat{P}(Y_t \leq y_t|\boldsymbol{x}_t^{(j)}, \mathcal{M}_0)$, from the obtained $y_t^{(k)}$s. Finally, an estimate of ψ_t is obtained as

$$\hat{\psi}_t = \frac{1}{J}\sum_{j=1}^{J} \hat{P}(Y_t \leq y_t|\boldsymbol{x}_t^{(j)}, \mathcal{M}_0).$$

So, in total, the estimate of $\hat{\psi}_t$ requires drawing of JK samples at each time instant t. This approach besides being computationally intensive, may produce unstable estimates of $\hat{\psi}_t$ [11]. By contrast, our method requires the generation of $J + K$ samples only, of which K are drawn from $p(y_t|y_{1:t-1}, \mathcal{M}_0)$. Moreover, K can be rather small, even as small as two. We point out that the particle filter has to complete the step of computing the random measure that is used for the next time instant in order to generate the needed samples. A simple approach would be to use the SIR filter from [12], where at time instant t, one generates particles $\boldsymbol{x}_t^{(j)}, j = 1, \ldots, J$ from $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)}, \mathcal{M}_0)$, draws samples $\tilde{y}_t^{(k)}$ according to (7), computes d_t, evaluates the weights of $\boldsymbol{x}_t^{(j)}$, and resamples the particles $\boldsymbol{x}_t^{(j)}$. In that case $J = M$.

Second, in the calculation of d_t only the value of y_t is used. However, the model is assessed using $d_{1:t}$. The advantage of the proposed procedure is that one can choose one of many possible tests for assessing the model $\mathcal{M}_0$. For example, since we know the distribution of D_t, we can directly test whether the outcomes $d_{1:t}$ come from that distribution. Alternatively, we can compute a single statistic using $d_{1:t}$ that allows us to decide whether we should reject the model $\mathcal{M}_0$. The latter approach is explored in the next section. Third, in our paper we assumed that the postulated model was nonlinear. If the model is linear and Gaussian, the predictive distribution of Y_t is also Gaussian. More specifically, let our model be

$$\boldsymbol{x}_t = \boldsymbol{G}_t\boldsymbol{x}_{t-1} + \boldsymbol{u}_t \tag{9}$$
$$y_t = \boldsymbol{h}_t^\top \boldsymbol{x}_t + v_t \tag{10}$$

where $\boldsymbol{G}_t$ and $\boldsymbol{h}_t$ are a known matrix and a vector, respectively, and $\boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ and $v_t \sim \mathcal{N}(0, \sigma_v^2)$ are the state and observation noises, where $\boldsymbol{\Sigma}_u$ is the covariance matrix of $\boldsymbol{u}_t$. We assume that the matrix and vectors in (9) and (10) all have compatible dimensions. Clearly, these two equations are a special case of the system described by(2) and (3). It is well known that the predictive distribution of Y_t is given by [30]

$$p(y_t|y_{1:t-1}) = \mathcal{N}(\eta_t, \xi_t^2)$$

where η_t and ξ_t^2 are the estimated mean and variance of the Gaussian distribution obtained by the Kalman filter. If $p(\boldsymbol{x}_{t-1}|y_{1:t-1}) = \mathcal{N}(\boldsymbol{\zeta}_{t-1}, \boldsymbol{C}_{t-1})$, where $\boldsymbol{\zeta}_{t-1}$ and $\boldsymbol{C}_{t-1}$ are the mean and covariance of $\boldsymbol{x}_{t-1}$, then

$$\eta_t = \boldsymbol{h}_t^\top \boldsymbol{G}_t\boldsymbol{\zeta}_{t-1}$$
$$\xi_t^2 = \boldsymbol{h}_t^\top \boldsymbol{R}_t\boldsymbol{h}_t + \sigma_v^2$$

where $\boldsymbol{R}_t = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^\top + \boldsymbol{\Sigma}_u$. For finding the mean and the covariance of $\boldsymbol{x}_t$, we employ

$$\boldsymbol{\zeta}_t = \boldsymbol{G}_t\boldsymbol{\zeta}_{t-1} + \frac{1}{\xi_t^2}\boldsymbol{R}_t\boldsymbol{h}_t\left(y_t - \boldsymbol{h}_t^\top \mathbf{G}_t\boldsymbol{\zeta}_{t-1}\right)$$
$$\boldsymbol{C}_t = \boldsymbol{R}_t - \frac{1}{\xi_t^2}\boldsymbol{R}_t\boldsymbol{h}_t\mathbf{h}_t^\top \boldsymbol{R}_t.$$

Since the parameters of the Gaussian are known, this allows us to obtain d_t without generating fictitious observation samples $\tilde{y}_t^{(k)}$. Moreover, under $\mathcal{H}_0$,

$$d_t \sim \mathcal{U}[0.5,\ 1]$$

which implies that one can use tests for determining whether $d_{1:t}$ are indeed samples from the continuous uniform distribution with support $[0.5,\ 1]$. This approach is illustrated in Section V-A.

V. Examples

We illustrate the application of the proposed method by way of three examples. The first one is on a time-varying channel estimation problem, well known in the communications arena. There we use the Kalman filter to track the channel and obtain the Gaussian predictive densities of the observations analytically. The second example involves a highly nonlinear dynamic system which has been studied in the context on nonlinear filtering by several authors (see, e.g., [16] and the references therein). Finally, in the last example we consider the problem of tracking a moving target using a network of wireless sensors.

A. Linear Channel Tracking

Consider a time-varying communication channel with an impulse response $\boldsymbol{x}_t \in \mathbb{R}^{n_x}$. The channel evolution is modeled by the first-order autoregressive process

$$\boldsymbol{x}_t = 0.95\boldsymbol{x}_{t-1} + \sqrt{10^{-2}}\boldsymbol{u}_t, \quad t = 1, 2, \ldots, T \tag{11}$$

where $\boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n_x})$ is a sequence of i.i.d. n_x- dimensional noise vectors. The parameter T denotes the length of the sequence of channel responses. We also assume that $p(\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n_x})$.

A sequence of known binary symbols, $b_{1:T-n_x+1}$, $b_t \in \{\pm 1\}$, $t = 1, \ldots, T - n_x + 1$, is transmitted over the communication channel. For $t < 1$ and $t > T - n_x + 1$, we assume $b_t = 0$. At the receiving end, the observations have the form

$$y_t = \boldsymbol{b}_t^\top \boldsymbol{x}_t + v_t, \quad t = 1, \ldots, T \tag{12}$$

where $\boldsymbol{b}_t = [b_t, b_{t-1}, \ldots, b_{t-n_x+1}]^\top$ and $v_t \sim \mathcal{N}(0, \sigma_v^2)$ is a sequence of i.i.d. noise samples. The variance σ_v^2 is fixed to yield an average signal-to-noise ratio (SNR) of 10 dB. The state-space system given by (11) and (12), together with the prior $p(\boldsymbol{x}_0)$ and the channel length $n_x = 3$, represents the model $\mathcal{M}_0$.

The model $\mathcal{M}_0$ is linear and Gaussian, hence the predictive distribution $p(y_t|y_{1:t-1}, \mathcal{M}_0)$ is Gaussian and can be computed analytically using a Kalman filter, as shown in Section IV. As a

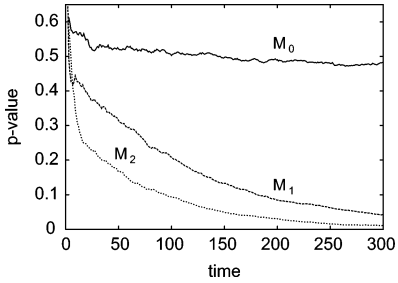


Fig. 1. Sequence of p-values obtained by the standard one-sample KS test for models $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$. The curves are averaged over 500 independent simulation trials.

consequence, if the assumption of $\mathcal{M}_0$ is correct, the KS statistics $D_1, \ldots, D_T$ are i.i.d. uniform random variables on $[1/2,\ 1]$. In order to illustrate the performance of the proposed assessment method, we also consider two more models $\mathcal{M}_1$ and $\mathcal{M}_2$, where the channel length under $\mathcal{M}_1$ is $n_x = 1$ and, under $\mathcal{M}_2$, $n_x = 4$.

Given a sample sequence of statistics $D_{1:T} = d_{1:T}$, we assess the validity of $\mathcal{M}_0$ by testing whether $d_{1:T}$ is a sequence of i.i.d. samples from $\mathcal{U}[0.5, 1]$. Specifically, we apply a standard KS one-sample test (as described in Section II-B) at each time step t, $t = 1, \ldots, T$, to obtain a sequence of p-values.[4]

Fig. 1 shows the sequence of p-values of the tests, averaged over 500 independent simulation runs, when the observations $y_{1:T}$ (with $T = 300$) are generated from $\mathcal{M}_0$, $\mathcal{M}_1$, and $\mathcal{M}_2$, respectively. It is seen that the p-values obtained when the data actually come from $\mathcal{M}_0$ are always clearly higher and converge to a mean value of ≈ 0.5. This means that the data would be recognized to be generated from $\mathcal{M}_0$ by a test with a type I error as high as ≈ 0.5. When the observations are generated from models $\mathcal{M}_1$ and $\mathcal{M}_2$, the p-values of the tests quickly decrease with time. In particular, it is seen that model $\mathcal{M}_2$ (with a longer channel length) is "more different" from the model $\mathcal{M}_0$, as the p-values converge toward zero more quickly.

We also illustrate the mismatch among the distributions of D_t under $\mathcal{M}_0$, $\mathcal{M}_1$, and $\mathcal{M}_2$. Fig. 2 shows the histograms of $d_{1:300}$ obtained in a single simulation with observations generated from each model. It is seen that the histograms for $\mathcal{M}_1$ and, especially, $\mathcal{M}_2$, depart clearly from the uniform distribution.

B. A Nonlinear Dynamical System

Let the nonlinear state-space model $\mathcal{M}_0$ be given by

$$x_t = \frac{1}{2}x_{t-1} + \frac{25x_{t-1}}{1+x_{t+1}^2} + 8\cos(\phi t) + \sigma_x u_t$$
$$y_t = \frac{x_t^2}{20} + \sigma_y v_t$$

where $x_t \in \mathbb{R}$ is the system state, ϕ is a frequency parameter (in rad/s), $u_t, v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and σ_x and σ_y are the standard deviations of the state and observation noises, respectively. Thus, the model has a parameter vector defined by $\boldsymbol{\theta} = [\phi, \sigma_x, \sigma_y]^\top$. Note that, given $\boldsymbol{\theta}$, the state transition distribution, $p(x_t|x_{t-1}, \mathcal{M}_0)$ and the likelihood function, $p(y_t|x_t, \mathcal{M}_0)$, are Gaussian. We complete the model with the assumption of a Gaussian prior, $p(x_0|\mathcal{M}_0) = \mathcal{N}(\cos(0), 1)$. Under $\mathcal{H}_0$, we adopt the model $\mathcal{M}_0$ with parameters $\boldsymbol{\theta}_0 = [1.2,\ 1,\ 1]$.

In order to apply the proposed method, we assume $\mathcal{H}_0$ is true and run a standard particle filter to approximate the predictive distributions $p(y_t|y_{1:t-1}, \mathcal{M}_0)$, $t = 1, 2, \ldots$ As a result, we obtain the point-mass approximations

$$p(y_t|y_{1:t-1}, \mathcal{M}_0) \simeq \hat{p}_K(y_t|y_{1:t-1}, \mathcal{M}_0) = \frac{1}{K}\sum_{k=1}^{K}\delta(y_t - y_t^{(k)}), \quad t \in \mathbb{N}$$

where $\delta(\cdot)$ denotes the Dirac delta function. The corresponding (empirical) predictive CDF of y_t is

$$\hat{P}(Y_t \le y_t|y_{1:t-1}, \mathcal{M}_0) = \frac{1}{K}\sum_{k=1}^{K} I_k(y_t)$$

and $I_k(y_t)$ is the indicator function that yields 1 when $y_t^{(k)} \le y_t$ and 0 otherwise. Given the actual observation y_t, the KS statistic at time t can be easily computed as

$$d_t = \max\left\{\hat{P}(Y_t \le y_t|y_{1:t-1}, \mathcal{M}_0), 1 - \hat{P}(Y_t \le y_t|y_{1:t-1}, \mathcal{M}_0)\right\}.$$

When K is odd, the random variable D_t is discrete uniform, as shown in Section II-B, and its mean and variance can easily be obtained. Respectively, they are given by

$$\mu_t = E(D_t|\mathcal{M}_0) = \frac{3K+1}{4K} \tag{13}$$
$$\sigma_t^2 = Var(D_{K,t}|\mathcal{M}_0) = \frac{K^3 + 3K^2 - K - 3}{48K^2(K+1)}. \tag{14}$$

Although there exist various possibilities for assessment tests, for this example we focus on those that rely on the sample mean $\bar{D}_t = \sum_{\tau=1}^{t} D_\tau/t$, which is itself a random variable with mean $E\left(\bar{D}_t|\mathcal{M}_0\right) = \mu_t$ and variance $Var\left(\bar{D}_t|\mathcal{M}_0\right) = \sigma_t^2/t$. We note that the sample mean is a consistent estimator of μ_t, since $\lim_{t\to\infty} Var\left(\bar{D}_t|\mathcal{M}_0\right) = 0$ irrespective of K. As a consequence, we can design efficient assessment procedures without the need of drawing a large number of samples to approximate $p(y_t|y_{1:t-1}, \mathcal{M}_0)$. This is a clear advantage with respect to the standard methods that rely on the statistic $\hat{w}_t = \hat{P}(Y_t \le y_t|y_{1:t-1}, \mathcal{M}_0)$, because $\hat{w}_t$ approximates a realization of a uniform random variable only when $\hat{P}(Y_t \le y_t|y_{1:t-1}, \mathcal{M}_0)$ is a sufficiently good approximation of the actual probability $P(Y_t \le y_t|y_{1:t-1}, \mathcal{M}_0)$. This, in turn, demands a large K, that is, the generation of a large number of samples from the random variable Y_t.

Besides computational savings, assessing the model with small K has the additional advantage of allowing the exact derivation of the distribution of the sample mean $\bar{D}_t$. Indeed, when $K = 3$, for example, D_t has a uniform distribution on the binary set $\{2/3,\ 1\}$. As a consequence, $\bar{D}_t$ takes values on the set $\{\tau + (t-\tau)2/3\}_{\tau=0,1,\ldots,t}$ with probabilities

$$K = 3: \quad P\left(\bar{D}_t = \tau + \frac{2}{3}(t-\tau)\Big|\mathcal{M}_0\right) = \frac{1}{2^t}C_{t,\tau}$$

[4]The p-value is the maximum value of the type I error probability of the test such that the null hypothesis is still accepted (conditional on $\mathcal{M}_0$ being true). Small p-values indicate that the test easily fails to recognize that the data come from the true model, while large p-values signify good agreement of the data and the model.

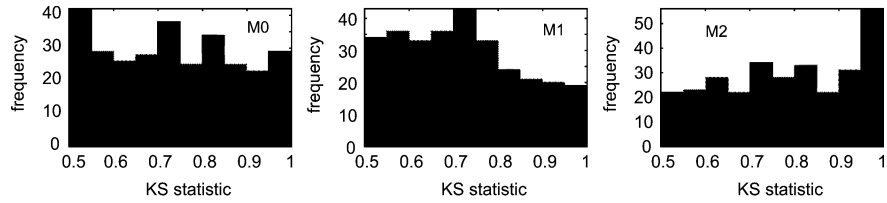


Fig. 2. Histograms of $d_{1:300}$ obtained from a single simulation trial with (a) $\mathcal{M}_0$, (b) $\mathcal{M}_1$, and (c) $\mathcal{M}_2$.

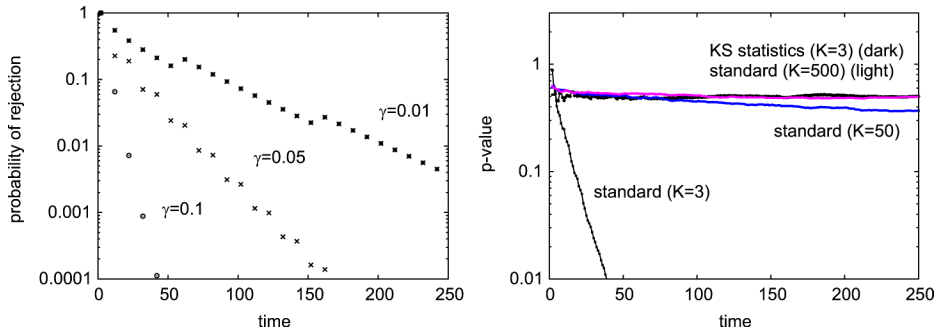


Fig. 3. (Left) Type I error probability, $P\left(Q_t > \gamma|\mathcal{M}_0\right)$, for $K = 3$ and three different thresholds, $\gamma = 0.01, 0.05, 0.1$ as a function of t. (Right): p-values for the standard test (with various K) and the proposed test (with $K = 3$).

where $C_{t,\tau} = t!/[\tau!(t-\tau)!]$ is the number of combinations of t elements taken in subsets of size τ. The CDF of $\bar{D}_t$ can also be evaluated exactly, namely

$$K = 3: \quad P\left(\bar{D}_t \le d|\mathcal{M}_0\right) = \frac{1}{2^t} \sum_{0\le\tau\le t:\tau+2/3(t-\tau)\le d} C_{t,\tau}$$

which turns out useful for the purpose of assessment, as shown below.

In this example, we assessed the validity of model $\mathcal{M}_0$ by evaluating the departure of the sample-mean statistics $\bar{d}_{1:T}$ from the theoretical means $\mu_{1:T}$ for $K = 3$, where T is the length of the observed time series, $y_1, y_2, \ldots, y_T$. Let $Q_t = |\bar{D}_t - \mu_t|$ be the (random) absolute deviation from the mean under the null hypothesis. We considered a test that rejects the null hypothesis at time t (for $0 \le t \le T$) if $q_t > \gamma$, where $\gamma > 0$ is some prescribed threshold. It is apparent that, for fixed $\gamma > 0$,

$$P\left(Q_t > \gamma\,|\mathcal{M}_0\right) = 1 - \left[P\left(\bar{D}_t \le \mu_t + \gamma|\mathcal{M}_0\right) - P\left(\bar{D}_t \le \mu_t - \gamma|\mathcal{M}_0\right)\right]$$

and, under $\mathcal{H}_0$, $\lim_{t\to\infty} P\left(Q_t > \gamma|\mathcal{M}_0\right) = 0$, due to the consistency of $\bar{D}_t$. Therefore, the probability of rejecting $\mathcal{H}_0$ when it is true ("type I error" of the test, in the classical terminology [7]) vanishes with time. Fig. 3 (left) shows how $P\left(Q_t > \gamma|\mathcal{M}_0\right)$ decreased for $t = 1, \ldots, 250$ and three different values of the threshold, namely $\gamma = 0.01, 0.05$ and 0.1.

We compared the proposed assessment procedure with the standard procedure that tests the distribution of the statistics $\hat{\psi}_t = \hat{P}(Y_t \le y_t|y_{1:t-1}, \mathcal{M}_0)$. To that end, using $\mathcal{M}_0$, we generated 500 independent sequences of observations of length $T = 250$. For each of these sequences, we applied the proposed method with $K = 3$ and the conventional approach with $K = 3$, 50 and 500 samples drawn using (6) and (7). For the standard procedure, we chose to validate the null hypothesis "$\hat{\psi}_{1:250}$ are samples from the uniform distribution on [0, 1]" directly, by means of the conventional KS one-sample test. We recall that other test procedures can readily be used as well. Any procedure, however, will be limited by the number of particles, K, available to approximate the CDF $P(Y_t \le y_t|y_{1:t-1}, \mathcal{M}_0)$. Indeed, for small K, the $\hat{\psi}_t$s are far from uniform (correspondingly, the $\Phi^{-1}(\hat{\psi}_t)$ are far from standard normal), which ultimately leads to an incorrect assessment.

This is illustrated by Fig. 3 (right), which depicts the average p-values, versus time, achieved by the proposed method with $K = 3$ (labeled "KS statistic") and the conventional KS test of the $\hat{\psi}_t$s with $K = 3$, 50, and 500 (labeled "standard"). For both types of tests, we computed the p-values exactly for each simulation run and then averaged them for each time instant t. The figure shows that the standard test fails completely to recognize that the data come from $\mathcal{M}_0$ when $K = 3$. The number of samples in the computation of the $\hat{\psi}_t$'s has to be increased up to $K = 50$ to obtain some acceptable results, yet clearly inferior to the proposed method with $K = 3$. In fact, the new KS-based technique attains the same average p-value with $K = 3$ as the standard test with $K = 500$ samples.

The result of this experiment showed that we can already obtain a good performance with only $K = 3$ samples from the predictive distribution. Nevertheless, we wanted to study the gain if we increased the value of K. To that end, we computed the absolute deviations of $\bar{d}_t$ from the theoretical mean of $\bar{D}_t$ for $t = 1, \ldots, 250$, $K = 3$ and $K = 500$, in 500 independent simulated realizations from $\mathcal{M}_0$. The obtained results are depicted in Fig. 4. Let $\bar{d}_t(j)$ denote the sample mean of the KS statistics at time t, in the jth simulation. Fig. 4 (left) shows the histograms of the realizations $\bar{d}_T(1), \ldots, \bar{d}_T(500)$ for $K = 3$ (light color), $K = 500$ (dark color), and $T = 250$. It can be clearly seen that the effect of increasing K is a shift of the mean of the random

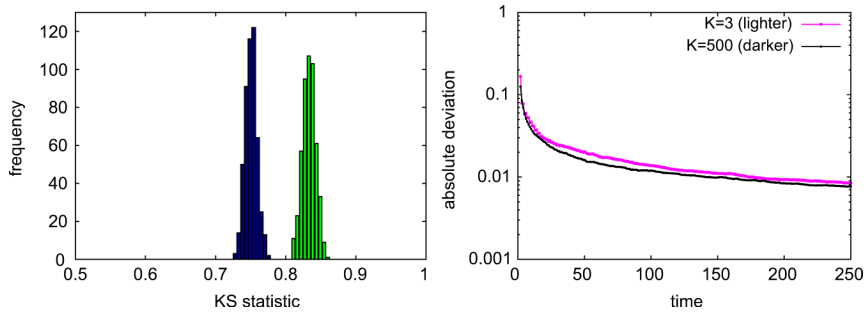


Fig. 4. (Left): Histograms of the sample-mean $\bar{d}_{250}$ for $K = 3$ (light color) and $K = 500$ (dark color). (Right): Average absolute deviations of $\bar{d}_t$ from $\mu_t = E(\bar{D}_t)$ for $K = 3$ and 500 and $t = 1, \ldots, 250$.

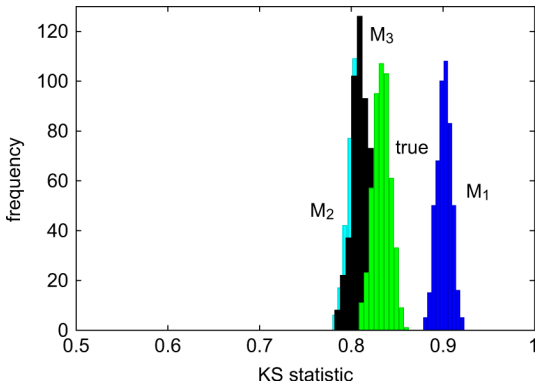


Fig. 5. Histograms of the sample-mean KS statistics $\bar{d}_{250}$ and $K = 3$ for $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_3$. The dark-colored histogram on the right corresponds to $\mathcal{M}_1$, the light-colored histogram in the middle corresponds to $\mathcal{M}_0$ (labeled "true"), and the two overlapping histograms to the left correspond to $\mathcal{M}_2$, (light-colored in the background) and $\mathcal{M}_3$ (dark-colored in the foreground).

variable $\bar{D}_T$ and a more skewed distribution, as predicted by (13) and (14).

Fig. 4 (right) depicts the average value of the absolute deviations $q_t(j) = |\bar{d}_t(j) - \mu_t|$, $j = 1, \ldots, 500$, for $t = 1, \ldots, 250$. The increase in the number of samples, K, led to a marginal performance improvement.

In the sequel, we study the performance of the proposed technique when $y_{1:T}$ are generated from a model $\mathcal{M}' \neq \mathcal{M}_0$. Specifically, we generated observations from different models by varying the parameters in $\boldsymbol{\theta}$ and then computed the KS statistics $\bar{d}_t(j)$ and absolute deviations $q_t(j)$ for $K = 3$ and $t = 1, \ldots, 250$ in 500 independent simulation trials, $j = 1, \ldots, 500$.

Fig. 5 shows the histograms of the sample-mean KS statistics at time $T = 250$, $\bar{d}_T(1), \ldots, \bar{d}_T(500)$ for $K = 3$, when the observations were generated from the following:

- $\mathcal{M}_0$: $\boldsymbol{\theta}_0 = [1.2, 1, 1]$ true model (for comparison);
- $\mathcal{M}_1$: $\boldsymbol{\theta}_1 = [1.21, 1, 1]$ ($\phi = 1.21$ is incorrect);
- $\mathcal{M}_2$: $\boldsymbol{\theta}_2 = [1.2, 2, 1]$ ($\sigma_x = 2$ is incorrect);
- $\mathcal{M}_3$: $\boldsymbol{\theta}_3 = [1.2, 1, 0.5]$ ($\sigma_y = 0.5$ is incorrect).

The shift in the mean of $\bar{D}_T$ is apparent for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ and it shows how these model mismatches can be detected easily. The assessment method is more sensitive to the variation in the frequency parameter ϕ (model $\mathcal{M}_1$), while the histograms obtained for $\mathcal{M}_2$ and $\mathcal{M}_3$ are very similar and closer to the histogram obtained for $\mathcal{M}_0$.

It is also of interest to determine how the histograms of the test statistics evolve with time both for $\mathcal{M}_0$ and $\mathcal{M}_1$. Fig. 6 shows the histograms of the sample means across the 500 simulation trials, $\bar{d}_{t_k}(1), \ldots, \bar{d}_{t_k}(500)$, $k = 0, 1, 2, 3$, at times $t_0 = 20$ [Fig. 6(a)], $t_1 = 40$, [Fig. 6(b)], $t_2 = 80$ [Fig. 6(c)] and $t_3 = 160$ [Fig. 6(d)]. The light-colored plot in the foreground corresponds to $\mathcal{M}_0$, while the dark-colored histogram in the background, to $\mathcal{M}_1$. It can be seen that both histograms overlap considerably up to $t_2 = 80$. At time $t_3 = 160$, the histograms are clearly separated.

Finally, we assessed the sensitivity of the method to variations in these three parameters, ϕ, σ_x and σ_y. Fig. 7 shows the average absolute deviation of $\bar{d}_{250}$, for $K = 3$, from its mean when ϕ is varied from 1.16 to 1.24 ($\phi_0 = 1.20$), when σ_x is varied from 0.25 to 4 ($\sigma_{x,0} = 1$), and when σ_y is varied from 0.25 to 4 ($\sigma_{y,0} = 1$). The results for ϕ are depicted in the left plot, while the results for σ_x and σ_y are jointly shown in the right plot. The proposed assessment method is particularly sensitive to mismatches in the frequency parameter ϕ. The mean of $\bar{D}_{250}$ has a clearcut minimum at the assumed value $\phi_0 = 1.20$. The incorrect values of the standard deviations σ_x and σ_y are also reflected in the respective absolute deviations from the means of the KS statistics.

C. Target Tracking With a Sensor Network

As a third example, we consider the problem of tracking a target that moves along a two-dimensional region using a network of sensors. We assume that the target is equipped with a radio transmitter and the sensors independently measure the power of the transmitted signal. The position and velocity of the target have to be sequentially estimated from the noisy measurements. Similar problems have recently been addressed in, e.g., [13] and [19].

Formally, we represent the system of interest using the state-space model

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{B}\boldsymbol{u}_t$$
$$y_{i,t} = 10\log_{10}\left(\frac{1}{\|\boldsymbol{r}_t - \boldsymbol{s}_i\|^{\alpha}}\right) + \sigma_y v_{i,t}, \quad i = 1, \ldots, N$$

where $\boldsymbol{x}_t = [r_{1,t}, r_{2,t}, \nu_{1,t}, \nu_{2,t}]^{\top} \in \mathbb{R}^4$ is the 4×1 state vector at time t, $\boldsymbol{r}_t = [r_{1,t}, r_{2,t}]^{\top} \in \mathbb{R}^2$ denotes the target position in the plane, $\boldsymbol{\nu}_t = [\nu_{1,t}, \nu_{2,t}]^{\top} \in \mathbb{R}^2$ is the target velocity,

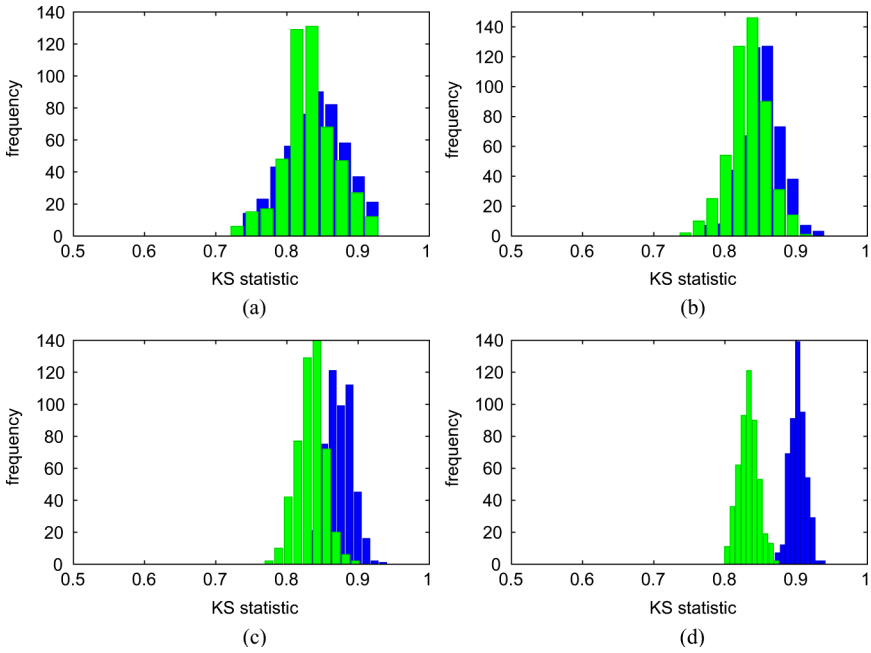


Fig. 6. Histograms of the KS sample-mean statistics at times (a) $t_0 = 20$, (b) $t_1 = 40$, (c) $t_2 = 80$ and (d) $t_3 = 160$. The dark-colored histograms in the background correspond to $\mathcal{M}_1$ while the light-colored histograms in the foreground, to $\mathcal{M}_0$.

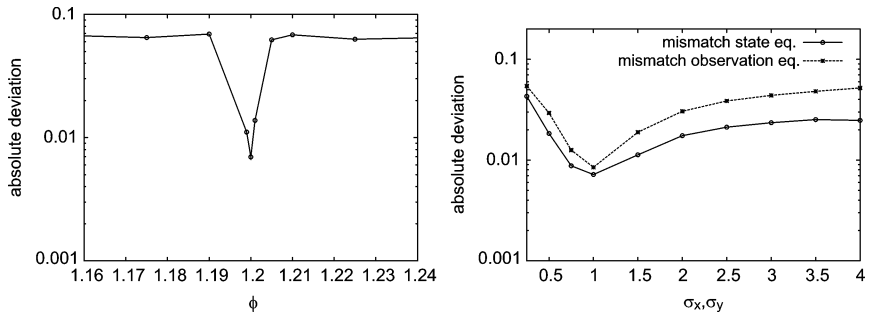


Fig. 7. Absolute deviations from the mean of the KS statistic, $q_{250} = |\bar{d}_{250} - \mu_{250}|$ for $K = 3$, obtained for several values of ϕ (left) and σ_x and σ_y (right).

$\boldsymbol{A} \in \mathbb{R}^{4\times4}$ and $\boldsymbol{B} \in \mathbb{R}^{4\times2}$ are the state transition matrix and the covariance of the state noise, respectively, defined by

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_2 & T_s\boldsymbol{I}_2 \\ \boldsymbol{0} & \boldsymbol{I}_2 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \frac{T_s^2}{2}\boldsymbol{I}_2 \\ T_s\boldsymbol{I}_2 \end{bmatrix}$$

with $\boldsymbol{I}_n$ denoting the $n \times n$ identity matrix, $T_s > 0$ is the sampling interval, and $\boldsymbol{u}_t \sim \mathcal{N}(0, \boldsymbol{I}_2)$. In the observation equation, $y_{i,t}$ is the power measurement, in decibels, collected by the ith sensor (with $i = 1, 2, \ldots, N$), $\mathbf{s}_i \in \mathbb{R}^2$ denotes the ith sensor position, $\|\boldsymbol{r}_t - \boldsymbol{s}_i\|$ is the Euclidean distance between the target and the ith sensor, α is a path-loss parameter that depends on the physical environment, $v_{i,t} \underset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and σ_y is the standard deviation of the observation noise.

Here, we compute the KS statistics for each scalar observation, i.e., we generate samples from the predictive distributions $p(y_{i,t}|\boldsymbol{y}_{1:t-1}, \mathcal{M}_0)$, $i = 1, \ldots, N$, using a particle filter, and then compute the statistics $d_{i,t}$, $i = 1, \ldots, N$, as functions of the actual scalar observations $y_{i,t}$. It can be shown that each time sequence of $D_{i,t}$s represents i.i.d. random variables, and thus all the previous results can be applied safely (per-sensor).

In the experiment, we set $T_s = 1/5$ s and the number of sensors was $N = 8$. In this way, the specific form of the model depended only on two parameters, $\boldsymbol{\theta} = [\alpha, \sigma_y^2]^\top$. We assessed the reference model $\mathcal{M}_0$ defined by $\boldsymbol{\theta}_0 = [2, 1]$. All the computations were based on the predictive distributions $p(y_{i,t}|\boldsymbol{y}_{1:t-1}, \mathcal{M}_0)$, $i = 1, \ldots, N$, and for their approximation we again ran the standard particle filter. As in the previous example, we carried out a test based on the statistics $\bar{d}_{i,t}$. Specifically, for each sensor i, we computed the absolute deviation $q_{i,t} = |\bar{d}_{i,t} - \mu_t|$. The test was designed to reject the model $\mathcal{M}_0$ whenever $q_{i,t} > \gamma$, where $\gamma > 0$ was a predefined threshold.

Fig. 8 (left) shows the absolute deviation $q_{1,t}$, for $t = 1, \ldots, 250$ and $K = 3$, averaged over 400 independent simulation runs. We observe how the deviation decreases with time, since $\bar{D}_{1,t}$ is a consistent estimator of μ_t under $\mathcal{M}_0$. This implies that, for any fixed threshold γ, the type I error probability of the test can be made as small as we wish by simply letting t grow. In the same plot, we show the average value of the deviation over the $N = 8$ sensors, i.e., $\sum_{i=1}^{N} q_{i,t}/N$, $t = 1, \ldots, 250$. We see that the average deviation is almost identical as the deviation for a single sensor.

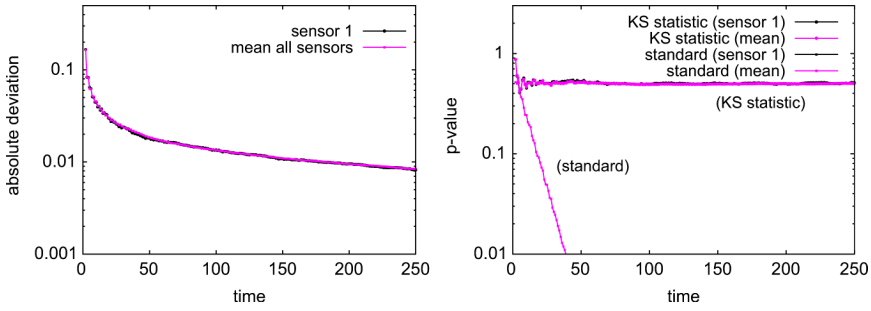


Fig. 8. (Left): Absolute deviation of $\bar{D}_t$ from its mean μ_t for sensor 1 and the mean of the absolute deviations of all the sensors ($N = 8$). (Right): p-values for the proposed test and the standard procedure based on $\hat{\psi}_t$. (See details in the text).

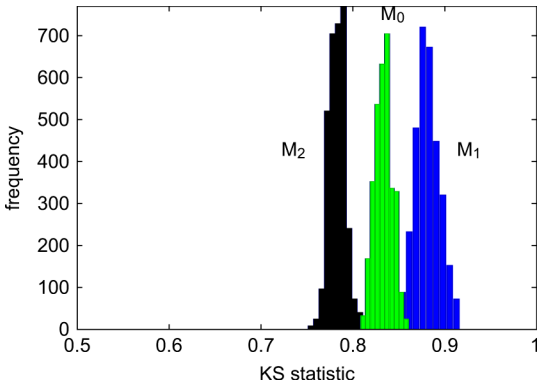


Fig. 9. Histograms of realizations of $D^i_{K,T}$ for $i = 1, \ldots, N$, $K = 3$ and $T = 250$. The results are obtained from realizations based on $\mathcal{M}_0$, $\mathcal{M}_1$ (with $\alpha = 2.04$) and $\mathcal{M}_2$ (with $\sigma_y = 2$).

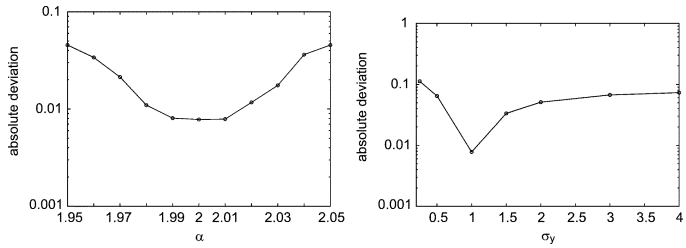


Fig. 10. Average of the absolute deviations q_T, $K = 3$ and $T = 250$, over 400 independent simulations. (Left) For several values of α. (Right) For several values of σ_y.

Fig. 8 (right) shows the evolution of the p-value of the test for sensor 1 with time ($t = 1, \ldots, 250$). Since $K = 3$, the p-value can be computed exactly for each simulation run because the $\bar{D}_{i,t}$s, $i = 1, \ldots, N$ and $t = 1, \ldots, 250$, follow the distribution in (15). The curve shown in the figure is the average of the exact p-values for sensor 1 over 400 independent simulation runs. On the same graph, we displayed the average of the p-values for the $N = 8$ sensors. Again, the curves for a single sensor and the mean are almost identical. We also depicted the p-values for the standard KS one-sample test of the null hypothesis obtained with $K = 3$ (labeled "standard"). We observe that the p-value for sensor 1 and the mean over all sensors quickly go to 0. This is because $K = 3$ samples are not enough to obtain a good approximation of the probability $P(Y_{1,t} \leq y_{1,t} | \boldsymbol{y}_{1:t-1}, \mathcal{M}_0)$ and, as a consequence, the $\hat{\psi}_t$s are far from uniform.

We also studied the changes in the distribution of $\bar{D}_t$ due to mismatches in α and σ_y. In particular, we considered two alternative models, $\mathcal{M}_1$ and $\mathcal{M}_2$, defined by $\boldsymbol{\theta}_1 = [2.04, 1]$ and $\boldsymbol{\theta}_2 = [2, 2]$, respectively. We carried out 400 independent simulations for each model and computed $d_{i,T}(j)$ for $K = 3$, $i = 1, \ldots, N$, $T = 250$ and $j = 1, \ldots, 400$ under the null hypothesis. Fig. 9 shows the resulting histograms. The histogram obtained when the observations are generated from the true model is depicted in light color, in the middle. The histogram of $\mathcal{M}_1$ appears clearly shifted to the right, while the histogram of $\mathcal{M}_2$ is shifted to the left and is slightly more skewed.

Fig. 10 shows the average of 400 independent realizations of $q_T = \sum_{i=1}^{N} q_{i,T}/N$ (where $q_{i,T} = \left|\bar{d}_{i,T} - \mu_T\right|$) for $K = 3$, $T = 250$ and different values of the path-loss exponent α (ranging from 1.95 to 2.05), and the observation-noise standard deviation, σ_y (ranging from 0.5 to 4). It is apparent that the assessment procedure is very sensitive to mismatches in α. Indeed, very small variations in α result in large differences in absolute deviation.

Finally, we studied a scenario in which the model parameters can change with time. Specifically, we performed an experiment where the value of α depended on the multipath propagation characteristics of the radio channel, which is time-varying in many environments. We set up a simulation for which $\alpha_t = \alpha_0 - 0.04 = 1.96$, when $T' < t < T''$, and $\alpha_t = \alpha_0 = 2$, otherwise. We ran 400 independent simulations using $\mathcal{M}_3 = [\alpha_t, 1]^\top$, with $T' = 350$ and $T'' = 550$. In order to account for the variability in the model, we computed a statistic different from the previous experiments. Instead of using the accumulated sample-means $\bar{d}_{i,t}$, we calculated sliding sample-means of the form $\tilde{d}_{i,t} = \sum_{\tau=t-L+1}^{t} d_{i,\tau}/L$ for the different sensors, and then averaged them, $\tilde{d}_t = \sum_{i=1}^{N} \tilde{d}_{i,t}/N$.

Fig. 11 depicts the mean absolute deviation $\tilde{q}_t = |\tilde{d}_t - \mu_t|$, with $K = 3$ and $t = 1, 2, \ldots, 800$, averaged over 400 independent runs. The parameter shift at time $t = T' + 1 = 351$ is clearly observed, as the absolute deviation grows sharply. In the same way, a steady decrease of the deviation starts at $t = 550$, as the parameter α_t falls back to the reference value α_0.

VI. Conclusion

We have proposed a method for the assessment of dynamic models that is based on the Kolmogorov–Smirnov distance. The method tests whether the observations come from the predictive distribution of the data conditioned on the model. To that end, a predictive CDF is constructed from generated data and compared to the CDF obtained from the single observation. The KS

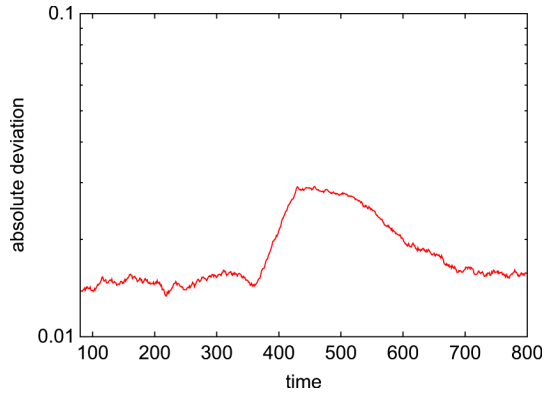


Fig. 11. Average absolute deviation of the KS statistic $\bar{d}_T$, with $K = 3$ and $T = 250$, from its mean for a time-varying path-loss parameter α_t.

distances are i.i.d. random variables with known discrete distributions which allow us to implement various types of tests. We have demonstrated the performance of the proposed approach on three examples and showed its superiority over a standard procedure.

REFERENCES

[1] C. Andrieu, A. Doucet, S. S. Singh, and V. B. Tadić, "Particle methods for change detection, system identification and control," *Proc. IEEE*, vol. 92, pp. 423–438, 2004.

[2] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. New York: Wiley, 1994.

[3] J. P. Bickel and K. Docksum, *Mathematical Statistics*. New York: McGraw-Hill, 2001.

[4] K. O. Bowman and L. R. Shenton, "Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2," *Biometrika*, vol. 62, pp. 243–250, 1975.

[5] G. E. P. Box, "Sampling and Bayes' inference in scientific modelling and robustness," *J. R. Stat. Soc.*, vol. 143, pp. 383–430, 1980.

[6] P. Brockwell and R. Davis, *Introduction to Time Series and Forecasting*. New York: Springer, 2002.

[7] M. H. DeGroot and M. J. Schervish, *Probability and Statistics*, 3rd ed. New York: Addison-Wesley, 2002.

[8] P. M. Djurić, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Míguez, "Particle filtering," *IEEE Signal Process. Mag.*, vol. 20, no. 5, pp. 19–38, 2003.

[9] P. M. Djurić and J. Míguez, "Model assessment with Kolmogorov–Smirnov statistics," presented at the IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Taipei, Taiwan, 2009.

[10] *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds. New York: Springer, 2001.

[11] R. Gerlach, C. Carter, and R. Kohn, "Diagnostics for time series analysis," *J. Time Series Anal.*, vol. 20, no. 3, pp. 309–330, 1999.

[12] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *Proc. Inst. Electr. Eng.—F*, vol. 140, pp. 107–113, 1993.

[13] F. Gustafsson and F. Gunnarsson, "Mobile positioning using wireless networks," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 41–53, Jul. 2005.

[14] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, U.K.: Cambridge Univ. Press, 1989.

[15] S. Kim and N. Shephard, "Stochastic volatility: Likelihood inference and comparison with ARCH models," in *Economics Papers 3*. Oxford, U.K.: Nuffield College, Univ. of Oxford, 1994.

[16] G. Kitagawa and S. Sato, "Monte Carlo smoothing and self-organising state-space model," in *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds. New York: Springer, 2001, ch. 9, pp. 177–196.

[17] G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, 1978.

[18] F. J. Massey, "Distribution table for the deviation between two sample cumulatives," *Ann. Math. Stat.*, vol. 23, pp. 435–441, 1952.

[19] L. Mihaylova, D. Angelova, S. Honary, D. R. Bull, C. N. Canagarajah, and B. Ristić, "Mobility tracking in cellular networks using particle filtering," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3589–3599, Oct. 2007.

[20] A. O'Hagan, *Kendall's Advanced Theory of Statistics: Bayesian Inference*. New York: Halsted, 1994, vol. 2B.

[21] R. L. Plackett, "Karl Pearson and the chi-squared test," *Int. Stat. Review.*, vol. 51, pp. 59–72, 1983.

[22] V. K. Rohatgi and A. K. Md. E. Sales, *An Introduction to Probability and Statistics*. New York: Wiley, 2001.

[23] M. Rosenblatt, "Remarks on multivariate transformation," *Ann. Math. Stat.*, vol. 23, no. 3, pp. 470–472, 1952.

[24] S. S. Shapiro and R. S. Francia, "An approximate analysis of variance test for normality," *J. Amer. Stat. Assoc.*, vol. 67, pp. 215–216, 1972.

[25] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, pp. 591–611, 1965.

[26] N. V. Smirnov, "On the estimation of the discrepancy between empirical curves of distributions for two independent samples," *Bull. Moscow Univ.*, vol. 2, pp. 3–16, 1939.

[27] J. Q. Smith, "Diagnostic check for non-standard time series Models," Dept. of Statistics, Univ. of Warwick, Warwick, U.K., 1985.

[28] M. A. Stephens, "EDF statistics for goodness of fit and some comparisons," *J. Amer. Stat. Assoc.*, vol. 69, pp. 730–737, 1974.

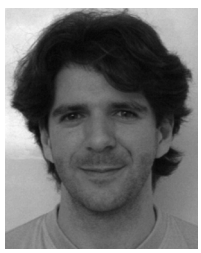
[29] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 2744–2756, 2002.

[30] M. West, *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag, 1997.

Petar M. Djuric (S'86–M'90–SM'99–F'06) received the B.S. and M.S. degrees in electrical engineering from the University of Belgrade in 1981 and 1986, respectively, and the Ph.D. degree in electrical engineering from the University of Rhode Island in 1990.

From 1981 to 1986, he was Research Associate with the Institute of Nuclear Sciences, Vinca, Belgrade. Since 1990, he has been with Stony Brook University, Stony Brook, NY, where he has been Professor in the Department of Electrical and Computer Engineering. He works in the area of statistical signal processing, and his primary interests are in the theory of signal modeling, detection, and estimation and application of the theory to a wide range of disciplines.

Prof. Djuric has been invited to lecture at universities in the United States and overseas and has served on numerous committees for the IEEE. During 2008–2009, he was Distinguished Lecturer of the IEEE Signal Processing Society. He was the Area Editor for special issues of the *IEEE Signal Processing Magazine* and Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He has also been on the Editorial Boards of several IEEE and EURASIP professional journals. In 2007, he received the Best Paper Award for a paper published in the *IEEE Signal Processing Magazine*.



Joaquín Míguez (S'99–AM'00–M'03) received the M.Sc. and Ph.D. degrees in computer engineering from the University of A Coruña, Spain, in 1997 and 2000, respectively.

Late in 2000, he joined the Department of Electronics and Systems of the same university, where he became Associate Professor in July 2003. From April 2001 through December 2001, he was a visiting scholar in the Department of Electrical and Computer Engineering of Stony Brook University, Stony Brook, NY. In September 2004, he moved to the Department of Signal Theory and Communications of University Carlos III, Madrid, Spain, also as Associate Professor. His research interests are in the field of statistical signal processing, with emphasis in the topics of Bayesian analysis, sequential Monte Carlo methods, stochastic filtering and optimization, and their applications to communications, sensor networks, target tracking and vehicle positioning/navigation.

Dr. Míguez is a co-recipient of the *IEEE Signal Processing Magazine* Best Paper Award 2007.