

Received February 15, 2019, accepted April 24, 2019, date of publication May 7, 2019, date of current version May 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2915109

Data Consistency Approach to Model Validation

ANDREAS LINDHOLM¹, DAVE ZACHARIAH¹, PETRE STOICA¹, (Fellow, IEEE),
AND THOMAS B. SCHÖN¹, (Senior Member, IEEE)

Department of Information Technology, Uppsala University, 751 05 Uppsala, Sweden

Corresponding author: Andreas Lindholm (andreas.lindholm@it.uu.se)

This work was supported in part by the Swedish Foundation for Strategic Research (SSF) through the project ASSEMBLE under Contract RIT15-0012, and in part by the Swedish Research Council through the project NewLEADS - New Directions in Learning Dynamical Systems under Contract 621-2016-06079.

ABSTRACT In scientific inference problems, the underlying statistical modeling assumptions have a crucial impact on the end results. There exist, however, only a few automatic means for validating these fundamental modeling assumptions. The contribution in this paper is a general criterion to evaluate the consistency of a set of statistical models with respect to observed data. This is achieved by automatically gauging the models' ability to generate data that is similar to the observed data. Importantly, the criterion follows from the model class itself and is therefore directly applicable to a broad range of inference problems with varying data types, ranging from independent univariate data to high-dimensional time-series. The proposed data consistency criterion is illustrated, evaluated, and compared with several well-established methods using three synthetic and two real data sets.

INDEX TERMS Statistical models, model validation, Monte Carlo method.

I. INTRODUCTION

In many scientific applications, statistical models provide a basis for inferences about real world phenomena. These inferences are typically dependent on the model assumptions being correct. However, there are few automatic means of evaluating these assumptions. In this paper, we address the problem of model validation by developing a method that automatically assesses the consistency of a set of models with respect to the observed data.

Let the observed data set be denoted as

$$\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}, \quad (1)$$

which consists of n data blocks of equal dimension, referred to as data points. We describe the mechanism that gave rise to the data by a probability density or mass function $p_0(\mathbf{y})$, which is unknown to us. In many applications, the objective of statistical inference is to determine certain properties of the unknown function $p_0(\mathbf{y})$. Statistical methods typically specify a family or class of probability distributions that aim to model $p_0(\mathbf{y})$. We denote this *model class* as

$$\mathcal{P}_\Theta \triangleq \{p(\mathbf{y} | \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\},$$

where each model $p(\mathbf{y} | \boldsymbol{\theta})$ is indexed by the parameter vector $\boldsymbol{\theta}$. Our aim in this paper is to assess whether the models

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenliang Zhang.

in \mathcal{P}_Θ that best approximate $p_0(\mathbf{y})$ are consistent with the observed data \mathbf{y} or not. The idea behind the proposed data consistency criterion (see below) is that if the best models in \mathcal{P}_Θ fail to generate data sets $\tilde{\mathbf{y}}$ that are 'similar' to \mathbf{y} , then \mathcal{P}_Θ can hardly be a valid modelling choice for \mathbf{y} . This notion will be made precise in the subsequent sections.

Past research efforts have mostly focused on *comparing* model classes, let us say \mathcal{P}'_Θ and \mathcal{P}''_Θ , using tools such as the Akaike or Bayesian information criteria and Bayes factors [1]–[4]. These criteria typically assume that some model class is well specified, meaning that one of the classes contains the unknown $p_0(\mathbf{y})$. While the consistency criterion proposed in this paper also can be used for model comparison, our focus here is rather on *validation* of *one* specified model class \mathcal{P}_Θ , which may or may not contain $p_0(\mathbf{y})$.

An established approach to validation is to use a residual-based criterion which assesses whether there is any "information" left in the data after fitting a model. In the restricted context based on the assumption of linear dynamical systems, such validation criteria are capable of rejecting model classes that are inconsistent with the data, cf. [5] and [6, Ch. 11]. For nonlinear dynamical systems, there also exists a non-statistical validation method to assess whether a deterministic model is capable of matching the observed data [7]. For statistical models based on the assumption of independent and identical distribution of the data points, classical tests such

as Cramér-von Mises, Anderson-Darling and Kolmogorov-Smirnov tests are applicable [8], [9]. Those tests are, however, constructed for very specific model classes. In the context of Bayesian modelling, validation of a model class can also be performed using posterior predictive checks which require the user to specify a discrepancy measure, cf. [10]–[12].

Our proposed data consistency criterion (DCC) evaluates the ability of a model to generate data similar to the observed one. In contrast to posterior predictive checks, DCC is automatic and does not require the user to specify any quantities except the model class \mathcal{P}_Θ itself. Furthermore, it applies directly to a broad range of model classes with various data types, e.g., linear regression models, count models, hidden Markov models, autoregressive models, etc. In general, $p(\mathbf{y} | \boldsymbol{\theta})$ can be factored as $p(\mathbf{y} | \boldsymbol{\theta}) = \prod_{i=1}^n p(\mathbf{y}_i | \mathbf{y}_1, \dots, \mathbf{y}_{i-1}, \boldsymbol{\theta})$. DCC is applicable whenever it is possible to point-wise evaluate $p(\mathbf{y}_i | \mathbf{y}_1, \dots, \mathbf{y}_{i-1}, \boldsymbol{\theta})$ for all i , and simulate new data $\tilde{\mathbf{y}}$ from the model, for any given $\boldsymbol{\theta} \in \Theta$.

In summary, the proposed DCC has the following features:

- DCC is a method for validating a given model class,¹
- DCC returns an interpretable test value,
- DCC follows automatically from the specified model class,
- DCC is readily applicable to a wide range of model classes, from models of univariate data points to high-dimensional time-series models.

The paper proceeds as follows: As an introductory application of our approach, we consider a modeling problem in seismology. Thereafter, in II, we explain the principles behind DCC for a single model $p(\mathbf{y} | \boldsymbol{\theta}_*)$ and, subsequently in III, for an entire model class \mathcal{P}_Θ . DCC is then applied to the seismological problem as an illustration in IV. We also discuss its implementation and compare it to classical methods, and illustrate how it can be applied to some interesting model classes, including regression, autoregressive and latent variables models in IV. The source code for all experiments is available online.²

A. MOTIVATING EXAMPLE: EARTHQUAKE COUNTS

A standard assumption in earthquake analysis is that earthquakes occur independently as described by a Poisson point process. That is, the number of earthquakes in any given region during any given time interval is Poisson distributed. However, it is also well-known that earthquakes tend to be clustered (both in time and space), where each cluster typically has several ‘foreshocks’ and ‘aftershocks’ and one larger ‘mainshock’. By modeling the earthquakes within a cluster as a branching process, the negative binomial distribution has been suggested for earthquake counts [13]. We consider both model classes, $\mathcal{P}_\Theta =$ Poisson distribution and $\mathcal{P}_\Theta =$ negative binomial distribution, which have one and two free parameters, respectively. We will use our proposed

¹Unlike model selection methods, that performs a relative comparison between classes, e.g. Akaike and Bayesian information criteria.

²<https://github.com/saerdna-se/consistency-criterion>

Magnitude	2012	2013	2014	2015	2016	2017
≥ 8	2	2	1	1	0	1
≥ 7	16	19	12	19	16	7
≥ 6	133	142	155	146	146	111
≥ 5	1680	1596	1729	1558	1696	1560

FIGURE 1. A snippet of the global earthquake count data, for different magnitudes and years. Each row (magnitude class) is a different data set \mathbf{y} . We would like to assess the consistency between each of these data sets and the two model classes, the Poisson and negative binomial distributions, respectively.

method to assess whether these model classes are consistent with the data \mathbf{y} in the United States Geological Survey earthquake catalog³ (partly shown in Fig. 1; the full data is found in Fig. 9). We will return to this example after developing the DCC.

II. DATA CONSISTENCY CHECK FOR A SINGLE MODEL

We begin by considering a model class consisting of only a single model, i.e., $\mathcal{P}_\Theta = \{p(\mathbf{y} | \boldsymbol{\theta}_*)\}$ where $\boldsymbol{\theta}_*$ is a specified parameter. Let $\tilde{\mathbf{y}} \sim p(\mathbf{y} | \boldsymbol{\theta}_*)$ denote a sample generated from the model and $\mathbb{P}_{\tilde{\mathbf{y}} | \boldsymbol{\theta}_*}(\cdot)$ the probability of an event under the same model.

Initially, consider the simpler case of models in which the data points $i = 1, \dots, n$ in (1) are assumed to be independent. Let $\tilde{z}_i \triangleq \ln p(\tilde{\mathbf{y}}_i | \boldsymbol{\theta}_*)$ denote the log-likelihood for the i th generated data point, and let its mean be denoted as $\mathbb{E}[\tilde{z}_i]$. For the i th observed data point, let $z_i \triangleq \ln p(\mathbf{y}_i | \boldsymbol{\theta}_*)$. The observed and generated log-likelihoods, z_i and \tilde{z}_i form the basis of our criterion. Intuitively, if the deviation of z_i from $\mathbb{E}[\tilde{z}_i]$ is much larger or much smaller than the deviation of \tilde{z}_i from $\mathbb{E}[\tilde{z}_i]$, we consider the observed data \mathbf{y} to be *atypical* for the given model $p(\mathbf{y} | \boldsymbol{\theta}_*)$. More formally, we define the following statistic

$$T(\mathbf{y}; \boldsymbol{\theta}_*) = \frac{1}{n} \sum_{i=1}^n \frac{(z_i - \mathbb{E}[\tilde{z}_i])^2}{\text{Var}[\tilde{z}_i]}, \quad (2)$$

where $\text{Var}[\tilde{z}_i]$ is variance of \tilde{z}_i . Similarly, we define the statistic $T(\tilde{\mathbf{y}}; \boldsymbol{\theta}_*)$ for generated data by replacing \mathbf{y} with $\tilde{\mathbf{y}}$. Let us now define the random event $S(\tilde{\mathbf{y}}, \mathbf{y})$ of generating a larger statistic than the observed one:

$$S(\tilde{\mathbf{y}}, \mathbf{y}) : T(\tilde{\mathbf{y}}; \boldsymbol{\theta}_*) > T(\mathbf{y}; \boldsymbol{\theta}_*). \quad (3)$$

When the probability of this event $\mathbb{P}_{\tilde{\mathbf{y}} | \boldsymbol{\theta}_*}(S(\tilde{\mathbf{y}}, \mathbf{y}))$ is close to 0, it is highly improbable that \mathbf{y} could have been generated by $p(\mathbf{y} | \boldsymbol{\theta}_*)$ and we deem the model to be inconsistent with the observed data. See Fig. 2a for an illustration. This type of inconsistency is due to under-dispersion of the generated log likelihoods, compared to the observed ones. The probability of the complementary event $\mathbb{P}_{\tilde{\mathbf{y}} | \boldsymbol{\theta}_*}(S^c(\tilde{\mathbf{y}}, \mathbf{y})) = 1 - \mathbb{P}_{\tilde{\mathbf{y}} | \boldsymbol{\theta}_*}(S(\tilde{\mathbf{y}}, \mathbf{y}))$ indicates inconsistency as well, see Fig. 2b. Also in this case it is improbable that \mathbf{y} could have been generated by $p(\mathbf{y} | \boldsymbol{\theta}_*)$. By contrast, if both aforementioned

³<https://earthquake.usgs.gov/earthquakes/search/>

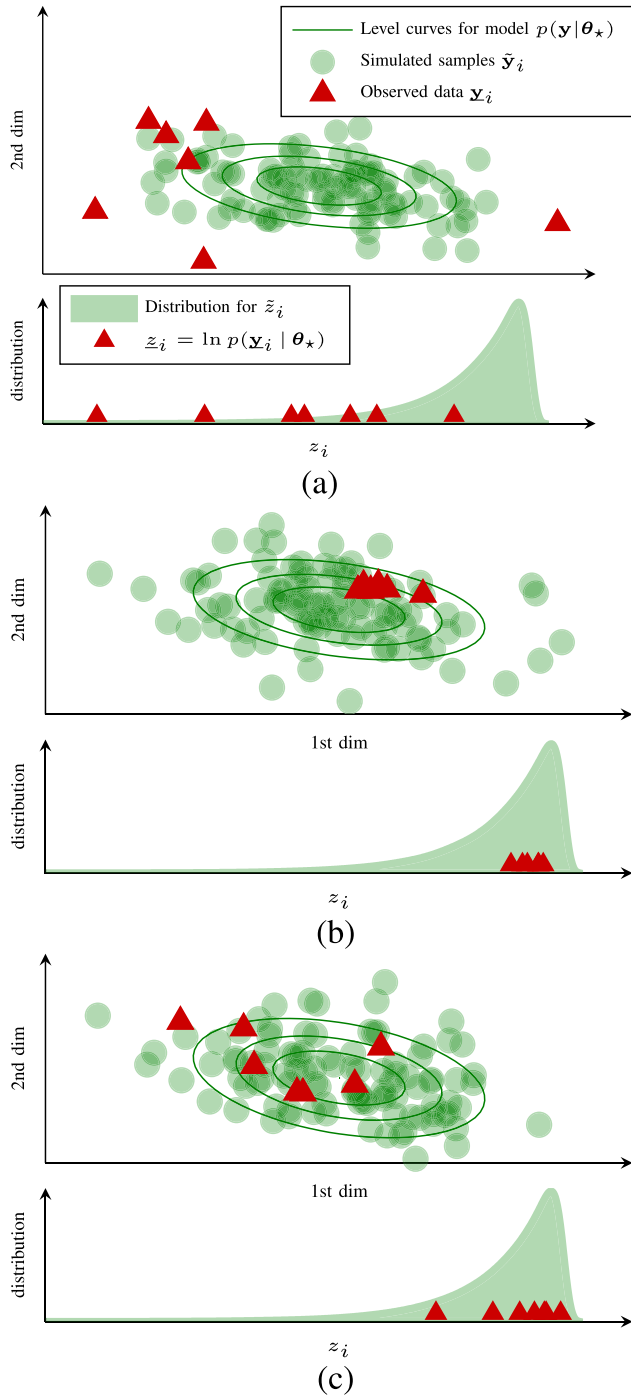


FIGURE 2. Consider a data set $\mathbf{y} = \{\mathbf{y}_i\}$ containing $n = 7$ two-dimensional data points (red triangles, upper panels) and assume a Gaussian i.i.d. model $p(\mathbf{y} | \theta_\star) = \prod_i p(\mathbf{y}_i | \theta_\star)$ (green level curves). The log-likelihoods $z_i = \ln p(\mathbf{y}_i | \theta_\star)$ for each data point are shown as red triangles in the lower panels. The generated log-likelihoods \tilde{z}_i follow the distribution illustrated in green in the same panels. When the deviation of z_i from $\mathbb{E}[\tilde{z}_i]$ is significantly different from that of \tilde{z}_i , it is unlikely that the model could have generated the observed sample. The deviation is quantified by the statistic $T(\mathbf{y}; \theta_\star)$. The figure illustrates three cases: two cases of inconsistency (a, b) and a balanced case (c). (a) The observed data points appear as atypical with respect to the model, since they fall into regions of low probability. We obtain $T(\mathbf{y}; \theta_\star) = 14$, and the probability of generating a higher statistic is $\mathbb{P}_{\tilde{\mathbf{y}}|\theta_\star}(S(\tilde{\mathbf{y}}, \mathbf{y})) = 0.00$. Thus, the dispersion of \tilde{z}_i is significantly lower than that of z_i and we reject the model as inconsistent with the observed data. (b) The observed data points

FIGURE 2. (Continued.) appear as atypical with respect to the model, since they are concentrated asymmetrically. We obtain $T(\mathbf{y}; \theta_\star) = 0.24$, and the probability of generating a lower statistic is $\mathbb{P}_{\tilde{\mathbf{y}}|\theta_\star}(S^c(\tilde{\mathbf{y}}, \mathbf{y})) = 0.04$. Thus, the dispersion of \tilde{z}_i is significantly higher than that of z_i and we reject the model as inconsistent with the observed data. (c) The observed data points appear to be typical samples from the model. We obtain $T(\mathbf{y}; \theta_\star) = 0.51$, and the probability of generating a lower statistic is $\mathbb{P}_{\tilde{\mathbf{y}}|\theta_\star}(S^c(\tilde{\mathbf{y}}, \mathbf{y})) = 0.39$. We do therefore not reject the model as inconsistent.

probabilities are significantly different from 0, we do not reject the model as inconsistent, see Fig. 2c.

The criterion above is readily generalized to models in which the data points in (1) are dependent, by extending the definition of z_i to

$$z_i \triangleq \ln p(\mathbf{y}_i | \mathbf{y}_1, \dots, \mathbf{y}_{i-1}, \theta_\star). \quad (4)$$

As above, the same symbols $\mathbb{E}[\tilde{z}_i]$ and $\text{Var}[\tilde{z}_i]$ are used to define the mean and variance of \tilde{z}_i .

If the model were a match of the unknown data-generating distribution, i.e., $p(\mathbf{y} | \theta_\star) = p_0(\mathbf{y})$, then the quantity $\mathbb{P}_{\tilde{\mathbf{y}}|\theta_\star}(S(\tilde{\mathbf{y}}, \mathbf{y}))$ would be uniformly distributed between 0 and 1 with respect to the observed data \mathbf{y} , see the proof below. In this situation, the probability of *falsely* rejecting the model due to the generated log likelihoods being under-dispersed would be

$$\text{PFA}_u(\theta_\star) \triangleq \mathbb{P}_{\tilde{\mathbf{y}}|\theta_\star}(S(\tilde{\mathbf{y}}, \mathbf{y})). \quad (5)$$

Thus, when $p(\mathbf{y} | \theta) = p_0(\mathbf{y})$, the probability of $\text{PFA}_u(\theta_\star)$ to be less than ρ , is equal to ρ . Symmetrically, the false alarm probability due to over-dispersion is $1 - \text{PFA}_u(\theta_\star)$. If neither $\text{PFA}_u(\theta_\star)$ nor $1 - \text{PFA}_u(\theta_\star)$ are small, we cannot reject the model $p(\mathbf{y} | \theta_\star)$ on the ground that the observed data \mathbf{y} is atypical, as discussed above.

This criterion, that neither $\text{PFA}_u(\theta_\star)$ nor $1 - \text{PFA}_u(\theta_\star)$ is close to 0, follows automatically from the specified model class and does not require user choices specific to the application scenario. The false alarm probabilities PFA above can be approximated numerically using Monte Carlo methods, as we will detail later.

A. PROOF OF UNIFORM DISTRIBUTION OF PFA_u

Let $\xi \triangleq T(\tilde{\mathbf{y}}; \theta_\star)$, whose distribution is characterized by a cumulative density function denoted $F_\xi(x)$. Further, let $\xi \triangleq T(\mathbf{y}; \theta_\star)$, which allows us to write

$$\mathbb{P}_{\tilde{\mathbf{y}}|\theta_\star}(S(\tilde{\mathbf{y}}, \mathbf{y})) = \mathbb{P}_{\tilde{\mathbf{y}}|\theta_\star}(\xi > \xi) = F_\xi(\xi). \quad (6)$$

Now, if $\mathbf{y} \sim p(\mathbf{y} | \theta_\star)$, then also the distribution of ξ is characterized by F_ξ , implying that $F_\xi(\xi) \sim \mathcal{U}[0, 1]$ according to the probability integral transform [14, Thm. 2.1.10].

III. DATA CONSISTENCY CHECK FOR THE BEST MODELS IN A CLASS

In most applications, θ_\star is not given. Instead \mathcal{P}_Θ may consist of a large number of models, possibly an uncountable number

when $\theta \in \Theta$ is continuous. In such a case, we will aim to evaluate the false alarm probabilities $PFA_u(\theta)$ and $1 - PFA_u(\theta)$ with respect to the models $p(\mathbf{y} | \theta)$ that best approximate $p_0(\mathbf{y})$. A natural measure for quantifying the accuracy of the approximation is the Kullback-Leibler divergence [15]. The best models are then defined as those that minimize the divergence:

$$\theta_* \in \arg \min_{\theta} \underbrace{\mathbb{E}_0 [\ln p_0(\mathbf{y}) - \ln p(\mathbf{y} | \theta)]}_{\text{model divergence}}, \quad (7)$$

where the expectation \mathbb{E}_0 is with respect to $\mathbf{y} \sim p_0(\mathbf{y})$. In particular, if θ_* exists such that the model divergence attains the minimum value 0, it follows that $p(\mathbf{y} | \theta_*) = p_0(\mathbf{y})$.

Since $p_0(\mathbf{y})$ is unknown, (7) cannot be used to identify the best models. Consequently we resort to an alternative approach. We assign weights to each model in \mathcal{P}_{Θ} so as to average the false alarm probability $PFA_u(\theta)$ across those models that are likely to be the best approximations of $p_0(\mathbf{y})$. The averaged false alarm probability due to under-dispersion is given by

$$PFA_u^* = \int_{\Theta} PFA_u(\theta) w(\theta | \mathbf{y}) d\theta \quad (8)$$

where the weights $w(\theta | \mathbf{y}) \geq 0$ are high for models in the neighborhood of θ_* and integrate to unity. By defining

$$w(\theta | \mathbf{y}) \triangleq \frac{w_0(\theta) p(\mathbf{y} | \theta)}{\int_{\Theta} w_0(\theta) p(\mathbf{y} | \theta) d\theta}, \quad (9)$$

the weights reflect the uncertainty about the location of θ_* in the parameter space Θ [14], [16]. The default choice of the initial weights is $w_0(\theta) \equiv 1$. In certain applications, however, we may have prior information about the location of θ_* in Θ . Then the initial weights $w_0(\theta)$ can be chosen to describe these prior beliefs. Under certain regularity conditions, the weights in (9) concentrate at θ_* as $n \rightarrow \infty$ if θ_* is unique, cf. [17] and [18].

We denote our final criterion

$$PFA^* = \min(PFA_u^*, 1 - PFA_u^*) \quad (10)$$

In summary, the proposed data consistency criterion (DCC) for the model class \mathcal{P}_{Θ} is the minimum of PFA_u^* and $1 - PFA_u^*$ (8). A value close to 0 indicates that the observed data \mathbf{y} is atypical for the best models in \mathcal{P}_{Θ} , and thereby we consider the model class \mathcal{P}_{Θ} to be inconsistent with \mathbf{y} .

A. IMPLEMENTATION

The averaged PFA^* is available in closed-form only in very few special cases, since there are non-trivial integrals in (5), (8), and (9). However, the integrals can be efficiently approximated by Monte Carlo integration techniques, provided that data $\tilde{\mathbf{y}}$ can be generated from $p(\mathbf{y} | \theta)$ and that $p(\mathbf{y}_i | \mathbf{y}_1, \dots, \mathbf{y}_{i-1}, \theta)$ can be evaluated point-wise. We outline such a generic Monte Carlo-based implementation⁴ in

⁴Where $\mathbb{I}[\cdot]$ denotes the indicator function.

Construct $w(\theta | \mathbf{y})$;

Draw N samples $\theta^{(j)} \sim w(\theta | \mathbf{y}), j = 1, \dots, N$;

for $j = 1, \dots, N$ **do**

Simulate M' data sets $\tilde{\mathbf{y}}^{(k)} \sim p(\mathbf{y} | \theta^{(j)})$,

$k = 1, \dots, M'$;

Compute $\tilde{z}_i^{(k)}$ as (4) for all generated data points $\tilde{\mathbf{y}}_i^{(k)}$;

Compute sample mean \hat{m}_i and variance \hat{v}_i of \tilde{z}_i' , $i = 1, \dots, n$;

Simulate M data sets $\tilde{\mathbf{y}}^{(\ell)} \sim p(\mathbf{y} | \theta^{(j)})$,

$\ell = 1, \dots, M$;

Compute $\tilde{z}_i^{(\ell)}$ as (4) for all generated data points $\tilde{\mathbf{y}}_i^{(\ell)}$;

Compute \underline{z}_i as (4) for all observed data points \mathbf{y}_i ;

Compute $\underline{T}^{(j)} = \frac{1}{n} \sum_{i=1}^n \frac{(\underline{z}_i - \hat{m}_i)^2}{\hat{v}_i}$;

Compute $\tilde{T}^{(j,\ell)} = \frac{1}{n} \sum_{i=1}^n \frac{(\tilde{z}_i^{(\ell)} - \hat{m}_i)^2}{\hat{v}_i}$;

Set $\hat{PFA}_u^{(j)} = \frac{1}{M} \sum_{\ell=1}^M \mathbb{I}[\tilde{T}^{(j,\ell)} > \underline{T}^{(j)}]$;

end

Set $\hat{PFA}_u^* = \frac{1}{N} \sum_{j=1}^N \hat{PFA}_u^{(j)}$ and

$\hat{PFA}^* = \min(\hat{PFA}_u^*, 1 - \hat{PFA}_u^*)$;

Algorithm 1 Monte Carlo Implementation of DCC (8)

Algorithm 1, where N parameters are first drawn using $w(\theta | \mathbf{y})$, and then (for each such draw) $M + M'$ samples of $\tilde{\mathbf{y}}$ are generated from $p(\mathbf{y} | \theta)$, giving a computational complexity on the order of $N(M + M')$.

The operations needed to execute Algorithm 1 are common in most statistical software packages. The weights $w(\theta | \mathbf{y})$ can be computed (at least approximately) by methods that estimate or learn θ . Numerical evaluation of the (incremental) likelihoods $p(\mathbf{y}_i | \mathbf{y}_1, \dots, \mathbf{y}_{i-1}, \theta)$, which in Algorithm 1 has to be performed both for generated data $\tilde{\mathbf{y}}$ (line 5) and observed data \mathbf{y} (line 6), is possible for many statistical models. If n is large compared to the number of parameters, it can be justified to use the following weights

$$w(\theta | \mathbf{y}) = \begin{cases} 1, & \theta = \hat{\theta} \\ 0, & \theta \neq \hat{\theta} \end{cases},$$

where $\hat{\theta}$ is the maximum likelihood/maximum a posteriori point estimate (in this case $N = 1$).

IV. EXAMPLES

A. EARTHQUAKE COUNTS (CONT'D)

To assess whether the Poisson or the negative binomial distribution is best suited for the data (partly) presented in Fig. 1, we are now ready to apply the DCC. We use the Monte Carlo approach in Algorithm 1 with $N = 200$ and $M = M' = 200$, and obtain the results in Fig. 3. From this table we draw the conclusion that the Poisson distribution and the negative binomial distribution are both consistent with the data for earthquakes with magnitude ≥ 7 . However, only the negative binomial distribution is consistent with the data for smaller

Magnitude	Poisson distribution	Negative binomial distribution
≥ 8	$\hat{PFA}^* = 0.40$	$\hat{PFA}^* = 0.39$
≥ 7	$\hat{PFA}^* = 0.29$	$\hat{PFA}^* = 0.38$
≥ 6	$\hat{PFA}^* = 0.00$	$\hat{PFA}^* = 0.30$
≥ 5	$\hat{PFA}^* = 0.00$	$\hat{PFA}^* = 0.13$

FIGURE 3. Assessment of the two earthquake count models (Poisson and negative binomial) for earthquakes of different magnitudes. The low average false alarm probabilities \hat{PFA}^* for the Poisson distribution suggests that this model class is not consistent with the observed data for magnitudes ≤ 6 .

magnitudes. This result supports the qualitative reasoning in the literature [13]: the Poisson distribution does not take the clustering effects into account, but since each cluster typically does not contain more than one major earthquake, the number of really big earthquakes can still follow the Poisson distribution.

B. SYNTHETIC DATA: GAUSSIAN MODELS

To further illustrate the behavior of PFA^* , we conduct a simulation study with two Gaussian model classes $\mathcal{P}_\Theta = \{\mathcal{N}(0, 1)\}$ and $\mathcal{P}'_\Theta = \{\mathcal{N}(\mu, \sigma^2) : \sigma^2 > 0\}$ with $\theta = \{\mu, \sigma^2\}$. Note that \mathcal{P}_Θ contains only a single model. We consider data sets $\underline{y} \sim p_0(\underline{y})$ of different sizes n generated from a standard Gaussian distribution $p_0 = \mathcal{N}(0, 1)$ and a standard uniform distribution $p_0 = \mathcal{U}[0, 1]$, respectively. We use $N = 50$ and $M = M' = 100$, and evaluate DCC in 1 000 experiments. The results are summarized as histograms in Fig. 4.

In the case when data comes from $p_0 = \mathcal{N}(0, 1)$ (Fig. 4a), both \mathcal{P}_Θ and \mathcal{P}'_Θ contain p_0 . Furthermore, the only model in \mathcal{P}_Θ is p_0 . Thus for \mathcal{P}_Θ the averaged PFA^* is uniformly distributed across experiments (cf. the proof of uniform distribution for PFA). In contrast to this, the weights $w(\theta | \underline{y})$ for \mathcal{P}'_Θ concentrate at the best model p_0 only as $n \rightarrow \infty$. Thus the averaged PFA^* approaches a uniform distribution only asymptotically. Neither model class is falsely rejected more frequently than the PFA^* indicates.

In the case when data comes from $p_0 = \mathcal{U}[0, 1]$ (Fig. 4b), neither \mathcal{P}_Θ nor \mathcal{P}'_Θ contains p_0 . Unlike \mathcal{P}_Θ , however, the best model in \mathcal{P}'_Θ matches the mean and variance of p_0 . Therefore, the inconsistency of the best model is discernible only for data points generated from the distribution tails. Consequently, it takes more samples n to reject the larger model class \mathcal{P}'_Θ than it takes to reject \mathcal{P}_Θ . As n increases, both model classes are clearly rejected.

Using the same example, we next make a comparison to the classical Massey [19], Lilliefors [20], Anderson and Darling [8], and Jarque and Bera [21] tests of normality. We consider $n = 100$, and select the respective thresholds such that the rate of falsely rejecting the model is either 5% or 10%. The Kolmogorov-Smirnov test is by construction a test against $\mathcal{P}_\Theta = \mathcal{N}(0, 1)$, whereas the others are tests against $\mathcal{P}'_\Theta = \mathcal{N}(\mu, \sigma^2)$. To achieve 5% or 10% false rejection probability in the classical tests, tabulated threshold values are available. To select the DCC threshold value for \mathcal{P}_Θ (no unknown

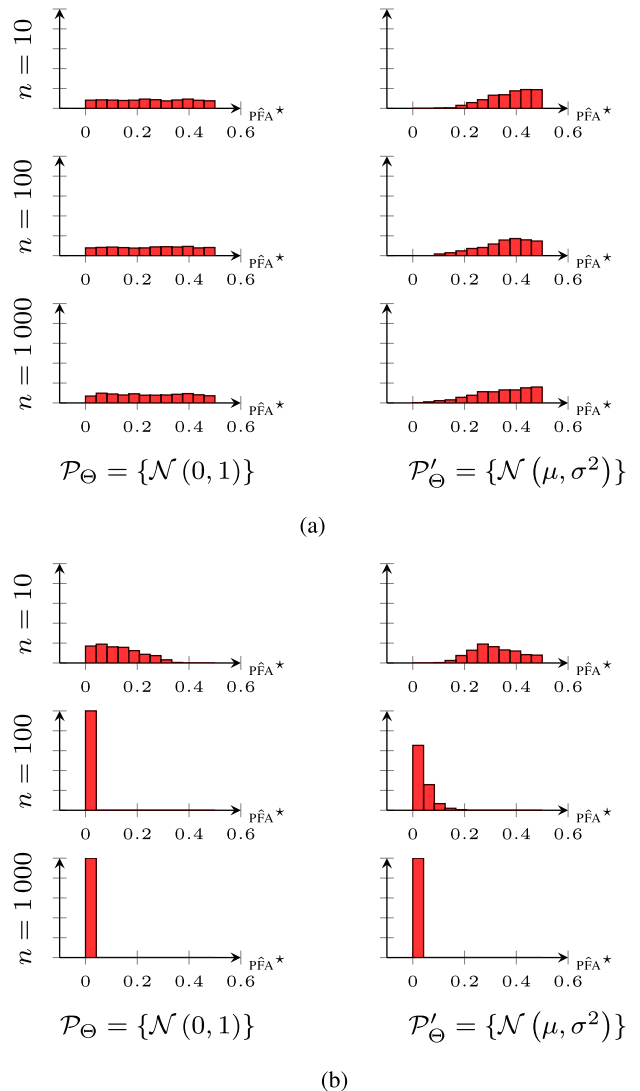


FIGURE 4. Histograms of DCC obtained from 1000 experiments, when p_0 is the standard Gaussian distribution (a) and the standard uniform distribution (b). Two different model classes are considered in the left and right columns, respectively. DCC is able to correctly identify the cases where the model class is consistent/inconsistent with the data. (a) Data generating process $p_0 = \mathcal{N}(0, 1)$. Both model classes contain p_0 . As n increases, PFA^* approaches a uniform distribution also for \mathcal{P}'_Θ . (b) Data generating process $p_0 = \mathcal{U}[0, 1]$. Neither model class contains p_0 . As n increases, PFA^* concentrates at 0 and both model classes are rejected.

parameters), the uniform distribution result can be directly applied, meaning that a false rejection rate of ρ is achieved by selecting the threshold⁵ as $\rho/2$. In the case of \mathcal{P}'_Θ with unknown parameters, however, the same procedure would be overly conservative (as indicated by Fig. 4a), since the uniform property is only asymptotic as $n \rightarrow \infty$. Instead, the threshold is set using simulations: Data with $n = 100$ is repeatedly simulated from $\mathcal{N}(0, 1)$ (which lies in \mathcal{P}'_Θ , hence a consistent case), and a threshold is selected such that the desired false rejection rate is achieved (cf. second column

⁵The factor 1/2 is because PFA_u is uniformly distributed on $U[0, 1]$, which implies that $\min(PFA_u, 1 - PFA_u)$ is uniformly distributed on half the interval, $[0, 0.5]$.

Method	$\mathcal{P}_\Theta = \mathcal{N}(0, 1)$ $p_0 = \mathcal{N}(0, 1)$	$\mathcal{P}'_\Theta = \mathcal{N}(\mu, \sigma^2)$ $p_0 = \mathcal{N}(0, 1)$	$\mathcal{P}_\Theta = \mathcal{N}(0, 1)$ $p_0 = \mathcal{U}[0, 1]$	$\mathcal{P}'_\Theta = \mathcal{N}(\mu, \sigma^2)$ $p_0 = \mathcal{U}[0, 1]$
Oracle	0	0	1	1
DCC	0.10	0.10	1.00	1.00
Kolmogorov-Smirnov	0.10	-	1.00	-
Anderson-Darling	-	0.09	-	1.00
Lilliefors	-	0.09	-	0.77
Jarque-Bera	-	0.10	-	0.99

(a)

Method	$\mathcal{P}_\Theta = \mathcal{N}(0, 1)$ $p_0 = \mathcal{N}(0, 1)$	$\mathcal{P}'_\Theta = \mathcal{N}(\mu, \sigma^2)$ $p_0 = \mathcal{N}(0, 1)$	$\mathcal{P}_\Theta = \mathcal{N}(0, 1)$ $p_0 = \mathcal{U}[0, 1]$	$\mathcal{P}'_\Theta = \mathcal{N}(\mu, \sigma^2)$ $p_0 = \mathcal{U}[0, 1]$
Oracle	0	0	1	1
DCC	0.06	0.05	1.00	1.00
Kolmogorov-Smirnov	0.05	-	1.00	-
Anderson-Darling	-	0.04	-	0.96
Lilliefors	-	0.04	-	0.60
Jarque-Bera	-	0.05	-	0.75

(b)

FIGURE 5. Probability of rejection, estimated using 1000 simulations. For each method, the rejection thresholds are set to achieve a certain probability of falsely rejecting the model class (i.e., the first and second columns). The proposed DCC clearly performs on par with the classical methods, even though DCC is a much more generally applicable method. The “oracle” is a fictitious test which knows the ground truth. (a) 10% false rejections. (b) 5% false rejections.

in Fig. 5). In fact, a threshold selected in this fashion will be independent of the parameters used in the simulations, since different parameter values essentially only change a constant in both \tilde{z}_i and \tilde{z}_i . This simulation-based procedure for selecting DCC thresholds for model classes with unknown parameters is hence useful also for practical purposes.

We simulate 1000 experiments, and report the estimated probability of rejections in Fig. 5. The results suggest that DCC has a performance comparable to the performance of the considered classical tests. In fact, DCC even outperforms the Lilliefors and Jarque-Bera tests for this particular example. It should, moreover, be remembered that whereas these classical tests are essentially tailor-made for testing normality of scalar random variables, the proposed DCC is generally applicable also to much more complex models, such as multivariate non-linear time-series models.

SYNTHETIC DATA: REGRESSION MODELS

We illustrate the capability of DCC to reject model classes with an inappropriate model order, using the example of polynomial regression. If the observed data is well described by a high-order polynomial, the best model in \mathcal{P}_Θ , which contains models of lower-order polynomials plus noise, will yield a good fit in the likelihood sense because the noise variance is scaled to match the residuals. Such an example is shown in Fig. 6, where \mathcal{P}_Θ contains 1st order polynomial models

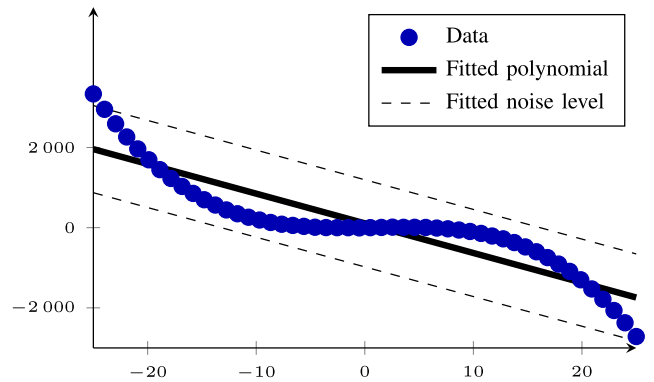


FIGURE 6. Data points from p_0 (3rd order polynomial, blue dots) and model class $\mathcal{P}_\Theta = \{1\text{st order polynomial} + \text{Gaussian noise}\}$ where θ contains the polynomial coefficients and noise variance. A fitted model is shown with θ estimated using the maximum likelihood method (black solid line: Estimated polynomial, dashed lines: 2 estimated noise standard deviations). While the estimated noise variance produces a good fit in terms of likelihoods, this model class should ideally be rejected. Using DCC for \mathcal{P}_Θ , we obtain $\text{PFA}^* = 0.01$ and can reject \mathcal{P}_Θ . Similarly, for a model class \mathcal{P}'_Θ containing 2nd order polynomials we obtain $\text{PFA}^* = 0.01$. By contrast, for 3rd order polynomials, \mathcal{P}''_Θ , we have $\text{PFA}^* = 0.37$ and this class is correctly found not to be inconsistent with the data.

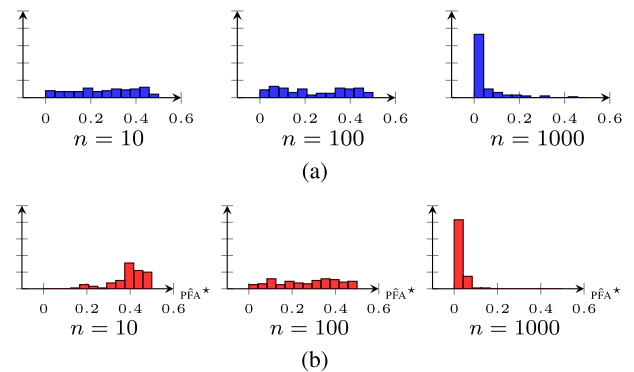


FIGURE 7. Histograms for the Ljung-Box p-value (top) and $\hat{\text{PFA}}^*$ (bottom). For both methods, small values indicate an inconsistency. The data y is generated by a saturated (non-linear) autoregressive model, but the assumed model class \mathcal{P}_Θ consists of linear autoregressive models. In this example, the amount of data required by the Ljung-Box method to detect the mismatch is slightly larger than for the proposed criterion. (a) Results for the Ljung-Box method. (b) Results for the DCC.

with independent Gaussian noise. When assessing \mathcal{P}_Θ using DCC, we obtain $\text{PFA}^* = 0.01$ (with $N = M = M' = 100$), which clearly indicates an inconsistency. On the other hand, for the model class containing 3rd order polynomials we obtain $\text{PFA}^* = 0.37$, which (correctly) indicates no inconsistency.

SYNTHETIC DATA: TIME-SERIES MODELS

For the case of time-series models driven by white noise, whiteness tests such as the Ljung-Box test [5], [22] are common validation techniques. The Ljung-Box test constructs a p-value from the fact that the statistic $n(n + 2) \sum_{k=1}^h \frac{\hat{r}_k^2}{n-k}$ will follow a χ_{h-d}^2 distribution if the model is correct, with \hat{r}_k being the lag k sample correlation of the prediction residuals, and d the dimension of θ . We set the upper lag limit h to $\log n$ (rounded to nearest integer).

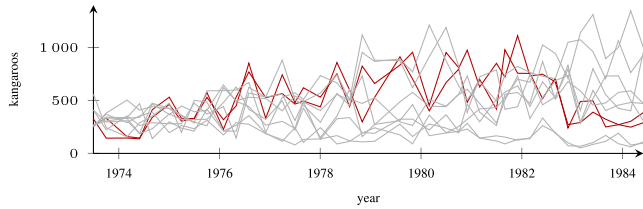


FIGURE 8. The kangaroo time series data (\mathbf{y} , red), together with a few time series (grey) that are generated from the model in (12) (parameters sampled from $w(\theta | \mathbf{y})$; cf. Step 7 of Algorithm 1). DCC indicates no inconsistency between the data and the model ($P\hat{F}A^* = 0.28$), which resonates with the intuition since the generated data behaves ‘similarly’ to the observed data.

We conduct a simulation study with data from the saturated first order autoregressive model $y_i = \max(0.7 y_{i-1} + e_i, -0.3)$, where $e_i \sim \mathcal{N}(0, 1)$. We assume a misspecified model class which consists of first-order linear autoregressive models $\mathcal{P}_\Theta = \{p(\mathbf{y}|\theta) : y_i = ay_{i-1} + e_i, e_i \sim \mathcal{N}(0, \sigma^2) : \sigma^2 > 0\}$, with unknown parameters $\theta = \{a, \sigma^2\}$. We consider cases where \mathbf{y} contains different amount of data samples n , and use $N = 200$ and $M = M' = 200$ in Algorithm 1.

The results are shown in Fig. 7. As the amount of data n grows, both methods correctly reject the misspecified model, with the proposed DCC needing slightly less data to do so than the much more specialized Ljung-Box method.

C. LATENT-VARIABLE MODELS: EVOLUTION OF A KANGAROO POPULATION

Certain models are too complex to be described using closed-form expressions. Instead these models are often parameter-

ized using latent variables η , with a prior distribution $p(\eta | \theta)$. This includes, e.g., hidden Markov or state-space models, mixed-effect models, latent topic models, etc. Then the data distribution can be written as the integral

$$p(\mathbf{y} | \theta) = \int p(\mathbf{y} | \eta, \theta)p(\eta | \theta)d\eta. \tag{11}$$

If the (incremental) likelihoods $p(y_i | \mathbf{y}_1, \dots, y_{i-1}, \theta)$ can be evaluated or well approximated, DCC can be computed also for this model class, as we will illustrate in the following.

We consider the dynamics of a population of red kangaroos (*Macropus rufus*) in New South Wales, Australia. The data \mathbf{y} , from [23, Appendix 8.2; available in Fig. 9], is a time series of $n = 41$ bi-variate observations from double transect counts at irregular time intervals between 1973 and 1984. In [24] the authors propose three different models for this data. These models are then compared in a pairwise fashion using the Bayes factor. The comparison has recently been repeated in [25] using the Hyvärinen score. Both [24] and [25] conclude that among the three different models, the preferred model is the following continuous-time stochastic differential equation

$$x_1 = \mathcal{LN}(0, 5), \tag{12a}$$

$$\frac{dx_t}{x_t} = \frac{\sigma^2}{2}dt + \sigma dW_t, \tag{12b}$$

$$y_{1,t}, y_{2,t} | x_t \stackrel{i.i.d.}{\sim} \mathcal{NB}(x_t, x_t + \tau x_t^2), \tag{12c}$$

where \mathcal{LN} is the log-normal distribution, W_t is the standard Brownian motion, \mathcal{NB} is the negative binomial distribution,

Magnitude	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
≥ 8	0	0	0	0	0	2	1	0	0	1	0	0	0	0	2	2	1	0	1
≥ 7	6	10	7	14	14	15	11	13	11	8	18	17	13	12	13	20	15	16	12
≥ 6	96	88	90	139	141	162	140	173	126	139	154	137	179	148	159	201	164	136	121
≥ 5	1408	1265	1505	1802	1683	1806	1765	1572	1598	1561	1765	1583	1675	1564	1693	1501	1373	1234	1074

Magnitude	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
≥ 8	0	1	1	0	1	2	1	2	4	0	1	1	1	2	2	1	1	0	1
≥ 7	18	15	16	13	15	16	11	11	18	12	17	24	20	16	19	12	19	16	7
≥ 6	136	160	137	139	156	157	151	153	196	179	161	175	207	133	142	155	146	146	111
≥ 5	1192	1495	1352	1309	1364	1672	1843	1877	2283	1965	2075	2395	2692	1680	1596	1729	1558	1696	1560

(a)

Date	1973 Jul	1973 Oct	1974 Mar	1974 Jun	1974 Sep	1975 Jan	1975 Apr	1975 Jul	1975 Oct	1976 Feb	1976 May	1976 Aug	1976 Dec	1977 Apr
Counts	267	333	159	145	340	463	305	329	575	227	532	769	526	565
	326	144	145	138	413	531	331	329	529	318	449	852	332	742

Date	1977 Jul	1977 Sep	1978 Jan	1978 May	1978 Aug	1978 Nov	1979 Feb	1979 Aug	1979 Nov	1980 Mar	1980 Jul	1980 Oct	1980 Dec	1981 Mar
Counts	466	494	440	858	599	298	529	912	703	402	669	796	483	700
	479	620	531	751	442	824	660	834	955	453	953	808	975	627

Date	1981 Jul	1981 Sep	1981 Dec	1982 Mar	1982 Jun	1982 Sep	1982 Dec	1983 Mar	1983 Jun	1983 Sep	1983 Dec	1984 Mar	1984 Jun
Counts	418	979	757	755	517	710	240	490	497	250	271	303	386
	851	721	1112	731	748	675	272	292	389	323	272	248	290

(b)

FIGURE 9. Complete data sets for the earthquake and kangaroo counting examples. (a) Data for the earthquake count example: The number of earthquakes above a certain magnitude (left column) in the entire world for 1980-2017, retrieved from the U. S. Geological Survey earthquake catalog. (b) Data for the kangaroo population example, adopted from [23, Appendix 8.2].

and $\theta = \{\sigma, \tau\}$ are unknown parameters. The latent variables are $\eta = \{x_t\}_{t \geq 1}$. Note that this model describes a bi-variate time-series $\{y_i\}$.

While [24] and [25] favor the model given in (12) over other alternatives, we consider the different question whether the model in (12) is consistent with the observed data \underline{y} or not. We first solve (12) analytically to obtain a discrete-time nonlinear state-space model, and use the particle marginal Metropolis-Hastings method [26] to sample the unknown parameters. A standard particle filter is used to approximate $p(y_i | y_1, \dots, y_{i-1}, \theta)$. We use $N = 1000$ and $M' = M = 200$ and obtain $\text{PFA}^* = 0.28$. Thus the model in (12) is deemed to be consistent with the observed kangaroo population data (also see Fig. 8 for intuitive support of this conclusion).

V. DISCUSSION

We have proposed a data consistency criterion (DCC) to assess the consistency of a model class \mathcal{P}_Θ with respect to the observed data \underline{y} . By comparing the observed (incremental) likelihoods $p(y_i | y_1, \dots, y_{i-1}, \theta)$ to the ones of generated data \tilde{y} , DCC rejects a model class for which \underline{y} is atypical for the best models in the class. The criterion follows automatically from the specification of \mathcal{P}_Θ and does not require additional application-specific choices. It yields an (approximate) false alarm probability PFA^* of erroneously declaring the best models to be inconsistent. When PFA^* falls below some set threshold, there is a sound ground for ruling out the model class \mathcal{P}_Θ . In a sense, the criterion is quantifying the lack of inconsistency, see [27] for a discussion about model validation on such grounds.

We have compared DCC to some well-established methods for consistency checks⁶, such as the Kolmogorov-Smirnov and Ljung-Box test. We concluded that DCC performs, at least, on par with these methods. Moreover, whereas these methods are designed specifically for a certain, rather restricted, model type, DCC is much more general; indeed DCC is applicable to a much broader range of classes, ranging from univariate data to high-dimensional time-series models.

By exploiting properties of the Fisher information matrix of \mathcal{P}_Θ , it is possible to construct misspecification tests, cf. [28]. That is, decide whether $p(\underline{y} | \theta_*) = p_0(\underline{y})$ is true or not. Since all practical models are incomplete in some respect, such tests may not always be relevant. Instead, what matters in many applications is whether the model is accurate or not, and the proposed DCC provides a practically useful criterion in this respect.

Using a Bayesian interpretation of (8), DCC can be understood as a certain posterior predictive check [10]–[12]. Posterior predictive checks are, however, not available for plug-and-play since they require the user to specify a ‘discrepancy variable’, in contrast to the fully automatic DCC. The DCC

⁶Note that model selection methods, such as the Akaike and the Bayesian information criteria, serve a different purpose than DCC, namely that of relative comparisons between models.

also readily admits a frequentist interpretation in terms of false alarm probability as a sampling property of $p_0(\underline{y})$.

The computational cost of the criterion increases indeed with the dimension of θ and n . The operations required are, however, available for many models and well developed in most statistical software packages. DCC could therefore be used as a routine check in existing statistical modeling methods, in order to better guide the end user to well-grounded scientific conclusions.

REFERENCES

- [1] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.
- [2] G. Schwarz, “Estimating the dimension of a model,” *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
- [3] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Upper Saddle River, NJ, USA: Prentice-Hall, 2005.
- [4] P. Stoica and Y. Selén, “Model-order selection: A review of information criterion rules,” *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, 2004.
- [5] L. 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.
- [6] T. Söderström and P. Stoica, *System Identification*. Upper Saddle River, NJ, USA: Prentice-Hall, 1989.
- [7] S. Prajna, “Barrier certificates for nonlinear model validation,” *Automatica*, vol. 42, no. 1, pp. 117–126, 2005.
- [8] T. W. Anderson and D. A. Darling, “Asymptotic theory of certain goodness of fit criteria based on stochastic processes,” *Ann. Math. Statist.*, vol. 23, no. 2, pp. 193–212, 1952.
- [9] E. L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco, CA, USA: McGraw-Hill, 1975.
- [10] D. B. Rubin, “Bayesianly justifiable and relevant frequency calculations for the applied statistician,” *Ann. Statist.*, vol. 12, no. 4, pp. 1151–1172, 1984.
- [11] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. Boca Raton, FL, USA: CRC Press, 2014.
- [12] G. E. P. Box, “Sampling and Bayes’ inference in scientific modelling and robustness,” *J. Roy. Stat. Soc. A, Gen.*, vol. 143, no. 4, pp. 383–430, 1980.
- [13] Y. Y. Kagan, *Earthquakes: Models, Statistics, Testable Forecasts*. Washington, DC, USA: American Geophysical Union, 2013.
- [14] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove, CA, USA: Brooks/Cole, 2002.
- [15] S. Kullback, *Information Theory and Statistics*. New York, NY, USA: Dover, 1959.
- [16] P. G. Bissiri and S. G. Walker, “Converting information into probability measures with the Kullback–Leibler divergence,” *Ann. Inst. Stat. Math.*, vol. 64, no. 6, pp. 1139–1160, 2012.
- [17] L. Lennart and P. E. Caines, “Asymptotic normality of prediction error estimators for approximate system models,” *Stochastics*, vol. 3, nos. 1–4, pp. 29–46, 1979.
- [18] R. H. Berk, “Limiting behavior of posterior distributions when the model is incorrect,” *Ann. Math. Statist.*, vol. 37, no. 1, pp. 51–58, 1966.
- [19] F. J. Massey, Jr., “The Kolmogorov-Smirnov test for goodness of fit,” *J. Amer. Statist. Assoc.*, vol. 46, no. 253, pp. 68–78, 1951.
- [20] H. W. Lilliefors, “On the Kolmogorov–Smirnov test for normality with mean and variance unknown,” *J. Amer. Stat. Assoc.*, vol. 62, no. 318, pp. 399–402, 1967.
- [21] C. M. Jarque and A. K. Bera, “A test for normality of observations and regression residuals,” *Int. Stat. Rev.*, vol. 55, no. 2, pp. 136–172, 1987.
- [22] P. Stoica, “A test for whiteness,” *IEEE Trans. Autom. Control*, vol. 22, no. 6, pp. 992–993, Dec. 1977.
- [23] P. Bayliss, “Kangaroo dynamics,” in *Kangaroos: Their Ecology and Management in the Sheep Rangelands of Australia*. Cambridge, U.K.: Cambridge Univ. Press, 1987.
- [24] J. Knape and P. D. Valpine, “Fitting complex population models by combining particle filters with Markov chain Monte Carlo,” *Ecology*, vol. 93, no. 2, pp. 256–263, 2012.

- [25] S. Shao, P. E. Jacob, J. Ding, and V. Tarokh, "Bayesian model comparison with the Hyvärinen score: Computation and consistency," *J. Amer. Stat. Assoc.*, to be published.
- [26] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 72, no. 3, pp. 269–342, 2010.
- [27] A. P. Robinson and R. E. Froese, "Model validation using equivalence tests," *Ecol. Model.*, vol. 176, nos. 3–4, pp. 349–358, 2004.
- [28] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, no. 1, pp. 1–25, 1982.



ANDREAS LINDHOLM (né Svensson) received the M.Sc. degree in applied physics and electrical engineering from Linköping University, Sweden, in 2001, and the Ph.D. degree in automatic control from Uppsala University, Sweden, in 2018. He is currently a Postdoctoral Fellow with the Department of Information Technology, Uppsala University. His research interests include machine learning and computational statistics, particularly within the context of dynamical systems.



DAVE ZACHARIAH received the M.S. degree in electrical engineering and the Tech. Lic. and Ph.D. degrees in signal processing from the Royal Institute of Technology, Stockholm, Sweden, in 2007, 2011, and 2013, respectively. From 2007 to 2008, he was a Research Engineer at Global IP Solutions, Stockholm. He is currently an Associate Senior Lecturer with Uppsala University, Uppsala, Sweden. His research interests include statistical signal processing, machine learning, sensor fusion, and localization.



U.K. He is also a Fellow of the EURASIP and the Royal Statistical Society.

PETRE STOICA (F'94) is currently a Professor in signal and system modeling with Uppsala University, Uppsala, Sweden. He is a member of the Royal Swedish Academy of Engineering Sciences, Stockholm, Sweden, the United States National Academy of Engineering (foreign member), Washington, DC, USA, the Romanian Academy, Bucuresti, Romania (Honorary), the European Academy of Sciences, Liège, Belgium, and the Royal Society of Sciences, London,



THOMAS B. SCHÖN received the B.Sc. degree in business administration and economics, the M.Sc. degree in applied physics and electrical engineering, and the Ph.D. degree in automatic control from Linköping University, in 2001, 2001, and 2006, respectively. He has held visiting positions with the University of Cambridge, U.K., the University of Newcastle, Australia, and the Universidad Técnica Federico Santa María, Valparaíso, Chile. He is currently a Professor of the Chair of Automatic Control with the Department of Information Technology, Uppsala University. His main research interest is probabilistic machine learning, with close links to signal processing and automatic control. He is a member of The Royal Swedish Academy of Engineering Sciences (IVA) and The Royal Society of Sciences at Uppsala. He received the Arnberg Prize and the Tage Erlander Prize for Natural Sciences and Technology from the Royal Swedish Academy of Sciences (KVA), in 2016 and 2017, respectively, the Automatica Best Paper Prize, in 2014, and the Best Ph.D. Thesis Award from The European Association for Signal Processing (EURASIP), in 2013. He received the Best Teacher Award at the Institute of Technology, Linköping University, in 2009.

...