Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Mathematical Psychology 50 (2006) 149–166

Journal of Mathematical Psychology

www.elsevier.com/locate/jmp

Accumulative prediction error and the selection of time series models

Eric-Jan Wagenmakers[a,*], Peter Grünwald[b], Mark Steyvers[c]

[a]Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands
[b]CWI, Amsterdam, National Research Institute for Mathematics and Computer Science, The Netherlands
[c]Department of Cognitive Sciences, University of California, Irvine, USA

Received 26 March 2005; received in revised form 1 December 2005

Abstract

This article reviews the rationale for using accumulative one-step-ahead prediction error (APE) as a data-driven method for model selection. Theoretically, APE is closely related to Bayesian model selection and the method of minimum description length (MDL). The sole requirement for using APE is that the models under consideration are capable of generating a prediction for the next, unseen data point. This means that APE may be readily applied to selection problems involving very complex models. APE automatically takes the functional form of parameters into account, and the 'plug-in' version of APE does not require the specification of priors. APE is particularly easy to compute for data that have a natural ordering, such as time series. Here, we explore the possibility of using APE to discriminate the short-range ARMA(1,1) model from the long-range ARFIMA$(0, d, 0)$ model. We also illustrate how APE may be used for *model meta-selection*, allowing one to choose between different model selection methods.

Keywords: Long-range dependence; Time series analysis; Model selection; Prediction error; Prequential methods; Predictive MDL

The main purpose of this article is to review the theoretical advantages and practical feasibility of accumulative one-step-ahead prediction error (APE) as a criterion for model selection. To briefly illustrate the motivation that underlies the APE method, consider two time series models, a simple model M_s and a complex model M_c. Both models are fitted to the same data set $x^n = (x_1, x_2, \ldots, x_n)$ of length n, and the aim is to select either M_s or M_c for the purpose of statistical inference. It is likely that the complex model M_c fits the data x^n better than does model M_s. This is not very informative, however, since the better fit may merely reflect the capability of M_c to describe noise. In model selection, what is important is model generalizability, or the ability of a model to predict unseen data from the same process (cf. Myung & Pitt, 1997; Pitt, Myung, & Zhang, 2002). The APE method implements this philosophy quite directly.

According to the APE method, the litmus test for models is not how well they fit the current data x^n, but how well they are able to predict the next unseen data point, x_{n+1}. In other words, according to the APE method the most useful model is the model with the smallest out-of-sample one-step-ahead prediction error.[1] The problem is that the prediction error cannot be calculated because x_{n+1} has not been observed. What can be calculated, however, are the prediction errors for x_{i+1} based on the previous x^i, $0<i<n$. For a given model, the APE method estimates the prediction error for the unavailable x_{n+1} by the sum of the previous prediction errors for data that are available. Note that this technique is similar in spirit to that of leave-one-out cross-validation (e.g., Browne, 2000; Stone, 1974): both methods assess generalizability by computing the predictive performance for a single data point, pretending it had never been observed. The methods differ, however, in that the size of the data set used for fitting the model gradually increases for the APE method but is constant for

*Corresponding author. Fax: +31 20 639 0279.
E-mail address: ewagenmakers@fmg.uva.nl (E.-J. Wagenmakers).

[1] One can posit alternative prediction criteria such as the smallest ten-step-ahead prediction error, and these alternative criteria may yield different results than the one-step-ahead APE discussed here (Bhansali, 1999). As will be apparent later, one of the attractions of the one-step-ahead APE is its close relation to Bayesian model selection and the method of minimum description length.

0022-2496/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2006.01.004

cross-validation. While leave-one-out cross-validation can be statistically inconsistent for a variety of models (Shao, 1993; Stone, 1977a), the APE method is typically consistent.

Indeed, the APE method is closely related to two methods of model selection that are consistent and satisfy certain optimality criteria:

(1) *Relation to Bayes*: The APE method is related to Bayesian model selection via the 'prequential' (i.e., sequential prediction) principle introduced by Dawid (1984, 1991). Two main instantiations of the APE method can be distinguished, depending on how the prediction of the next data points is carried out. The first version is exactly equivalent to Bayesian model selection. The second version may be interpreted as an approximation to Bayesian model selection. In contrast to standard Bayesian methods, however, this second form of the APE criterion can be calculated quite easily for a wide range of different time series models.

(2) *Relation to minimum description length* (*MDL*): The APE implements the principle of 'predictive minimum description length' (Rissanen, 1986b), one of a variety of MDL methods. Again, the advantage is that, in contrast to other MDL methods, the APE/predictive MDL method can be calculated relatively easily, particularly for time series models.[2]

Thus, in terms of computational ease the APE method is more similar to popular methods such as Akaike's information criterion (AIC; Akaike, 1974; Burnham & Anderson, 2002) and the Bayesian information criterion (BIC; Raftery, 1995; Schwarz, 1978) than it is to Bayesian model selection and MDL. The advantage that APE has over AIC and BIC is that APE is sensitive not only to the number of free parameters, but also to their functional form.

Despite its theoretical and practical advantages, the APE method is not used on a regular basis—barring a few recent exceptions in which the method has been applied with great success (Kontkanen, Myllymäki, & Tirri, 2001; Modha & Masry, 1998a, 1998b). The comparatively small amount of recent applications may be the main reason why APE is unfamiliar to most psychologists. Nevertheless, the list of inference problems to which the APE criterion has been applied does include ARMA model selection (Dawid, 1991; Gerencsér, 1994; Rissanen, 1989), density estimation using histograms (Dawid, 1991; Rissanen, Speed, & Yu, 1992), linear least-squares regression (Rissanen, 1986a), and generalized linear models (Qian, Gabor, & Gupta, 1996). Here, we illustrate the APE method by applying it to the problem of selecting between two popular non-nested time series models: the *long-range* autoregressive fractionally integrated moving average model (i.e., ARFIMA$(0, d, 0)$) and the *short-range* autoregressive moving average model (i.e., ARMA(1,1); cf. Basak, Chan, & Palma, 2001; Crato & Ray, 1996; Hosking, 1984; Wagenmakers, Farrell, & Ratcliff, 2004). To the best of our knowledge, no study yet has used the APE criterion to address this model selection problem.[3]

The outline of this paper is as follows. The next section describes the APE criterion and its relation to leave-one-out cross-validation, Bayesian model selection, and MDL. After a discussion of the APE's pros and cons we turn to a time series application involving models with short-range and long-range dependence. We briefly define the ARMA and ARFIMA time series models, outline the setup of a Monte Carlo simulation, and then report the results. Finally, the APE criterion is applied to a real-world time series that involves the estimation of one-second time intervals.

1. Model selection through accumulative one-step-ahead prediction error

Quantitative models seek to separate replicable structure from noise. Underfitting occurs when a model captures too little replicable structure, whereas overfitting occurs when a model captures too much noise. In both cases, predictive performance will suffer. A model that is able to optimally separate replicable structure from noise will have optimal predictive performance. Hence, the widely agreed upon criterion for model selection is the maximization of generalizability (Myung, 2000; Pitt et al., 2002), or equivalently the minimization of prediction error for future data coming from the same source. Several methods have been developed to implement this general idea. In what follows we will confine ourselves to the case of two competing, possibly non-nested models M_1 and M_2, as the extension to more than two competing models is self-evident.[4] The APE method advocated here is based on the intuition that a good indicator for prediction error to as yet unseen data is the sum (or equivalently, the average) of the previous prediction errors. Consider a time series of n observations, $x^n = (x_1, x_2, ..., x_n)$. The APE method proceeds by calculating sequential one-step-ahead forecasts based on a gradually increasing part of the data set. That is, the APE for model M_j is calculated as follows:

(1) Determine the smallest number s of observations that make the model identifiable. Set $i := s + 1$ (so that $i - 1 = s$).

[2] Application of the APE criterion is more involved when the data are not ordered sequentially in time. This issue is discussed in more detail later. The focus of the present paper is on time series data.

[3] This is somewhat surprising given the substantial interest in the forecasting performance of short-range versus long-range models; see for instance the 2002 special issue in the *International Journal of Forecasting* (Baillie, Crato, & Ray, 2002).

[4] This holds as long as the list of candidate models is finite. Certain complications arise in case the list is infinite (Grünwald, 2005).

(2) Based on the first $i-1$ observations, calculate a prediction $\hat{p}_i$ for the next observation i.
(3) Calculate the prediction error for observation i, for instance, take the squared difference between the predicted value $\hat{p}_i$ and the observed value x_i.
(4) Increase i by 1 and repeat steps 2 and 3 until $i=n$.
(5) Sum all of the one-step-ahead prediction errors as calculated in step 3. The result is the APE.

This procedure is completed separately for model M_1 and model M_2. One then proceeds to select the model for which the associated APE is smallest. Thus, for model M_j the accumulative prediction error is given by

$$APE(M_j) = \sum_{i=s+1}^{n} d[x_i, (\hat{p}_i|x^{i-1})], \tag{1}$$

where d indicates the specific function that quantifies the discrepancy between what is observed and what is predicted.

The APE recipe outlined above leaves open three important choices. We consider these in turn:

(1) *Form of predictions*: The first choice is what form the predictions should take: whether to predict using a single value (cf. Skouras & Dawid, 1998) or a probability distribution (cf. Aitchison & Dunsmore, 1975). In the latter case, $\hat{p}_i$ is a distribution on the set of possible outcomes x_i. In the experiments in this paper, we choose the first option: the predictions $\hat{p}_i$ are predictions for the mean value of ith outcome x_i.

(2) *Loss function*: The second choice is how to quantify the discrepancy between the predicted values and the observed values. We can measure the quality of predictions in a variety of different ways. Ideally, as explained in the next section, this measure should reflect the specific application we have in mind, but in practice this may be unknown. In that case, one can use a standard well-behaved loss function. For single-value predictions, one typically uses the squared error $(y_i - \hat{p}_i)^2$; this is also the approach we adopt in this paper. Another choice would be to compute the absolute value loss, $|y_i - \hat{p}_i|$, or, more generally, an α-loss function, $|y_i - \hat{p}_i|^\alpha$, $\alpha \in [1,2]$ (Rissanen, 2003). For probabilistic predictions, one typically uses the logarithmic loss function—ln $\hat{p}_i(y_i)$: thus, the loss depends on the probability mass or density that $\hat{p}_i$ assigns to the actually observed outcome y_i. The larger the probability, the smaller the loss. There are several strong reasons for choosing the logarithmic loss function. Perhaps the most important one is the fact that, as we explain below, it makes the APE method compatible with maximum likelihood (Example 1), Bayesian inference, and MDL inference.

(3) *Prediction algorithm*: Based on the initial $i-1$ outcomes and model M_j, we must output some prediction $\hat{p}_i$ for the ith outcome. The third choice is what prediction algorithm to use for this task. Consider first the case that we decided to predict using a distribution rather than a single value (see item 1, 'Form of Predictions' above). In that case, there exist two standard methods for determining the distribution $\hat{p}_i$:

(i) Set $\hat{p}_i$ equal to the distribution indexed by $\hat{\theta} = \hat{\theta}(x^{i-1})$, the maximum likelihood distribution within M_j for data $x^{i-1} = (x_1, \ldots, x_{i-1})$. This is the so-called *maximum likelihood* (*ML*) *plug-in method* for determining the APE. It is also the approach we adopt in this paper.
(ii) Compute the *Bayes predictive distribution* for x_i based on data x^{i-1} (Eq. (5)). This would be the preferred prediction method according to a Bayesian statistician. Note that the Bayes predictive distribution is a *mixture* of distributions in M_j, rather than a single element of it. A potential difficulty with the Bayesian method is that it requires the specification of a prior distribution $p(\theta \mid M_j)$ over the parameters θ. An advantage is that the first step in the APE algorithm outlined above can be omitted, because one can use the parameter priors to make predictions for the very first data point.

Now consider the case where we make predictions using a single value rather than a distribution. The standard way to determine a single value prediction $\hat{p}_i$ is as follows: one *first* infers a distribution $\hat{q}_i$ based on data x^{i-1}, using either the ML plug-in or the Bayesian method that we just described. One then predicts using the *value* $\hat{p}_i$ that would be the optimal prediction if data were distributed according to the inferred *distribution* $\hat{q}_i$. If the squared error is used, this is the mean of x_i under $\hat{q}_i$.

The next sections outline the theoretical justification of the APE procedure by discussing the relation between APE and three sophisticated model selection procedures: cross-validation, Bayesian model selection and model selection via the principle of MDL.

1.1. APE and leave-one-out cross-validation

Obviously, the APE method is similar in spirit to cross-validation, a model selection technique that has been successfully applied in a wide variety of contexts, including applications in psychology (Browne, 2000). The similarity is particularly striking if we compare APE to *leave-one-out* cross-validation. This frequently employed form of cross-validation is identical to the APE method except for steps 1 and 2 from the APE recipe outlined in the previous section. In cross-validation, *all* $n-1$ observations $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ are used to predict x_i, typically using the ML plug-in predictor with logarithmic or squared loss (Stone, 1974). This difference between APE and leave-one-out may seem small, but it has important repercussions: when the list of models under consideration is infinite, APE is statistically consistent whereas leave-one-out cross-validation can be statistically inconsistent. This means that the method of cross-validation is not guaranteed to select the 'true' underlying model as sample size grows large (Stone, 1977a). Cross-validation is for instance inconsistent for linear models (Shao, 1993).

On a related point, under regularity conditions on the models and the data generating distribution, and under the logarithmic loss function, leave-one-out cross-validation asymptotically behaves like AIC (Stone, 1977b; Zhang, 1993), whereas the APE method asymptotically behaves like BIC (Barron, Rissanen, & Yu, 1998), given appropriate constrains on the rate with which the information in the data increases (Wei, 1992).

1.2. Bayesian model selection and the Bayes factor

Recall that one way to implement the APE is to make subsequent predictions using the Bayesian predictive distribution and measure error by the logarithmic loss. It turns out that this procedure is equivalent (in the sense that it always gives the same results) to standard Bayesian model selection based on Bayes factors. To explain this, it is best to first point out that every probability distribution may be interpreted as a prediction strategy:

Example 1 (*Distributions as prediction strategies*). Let p be an arbitrary probability density or mass function on sequences $x^n = (x_1, \ldots, x_n)$ of some given length n. By the 'chain rule' of probability theory, p can be rewritten as a product of conditional probabilities, so that for all $x_1, \ldots, x_n$,

$$p(x_1, \ldots, x_n) = p(x_n \mid x^{n-1})p(x_{n-1} \mid x^{n-2}) \ldots p(x_2 \mid x_1)p(x_1). \tag{2}$$

To see this, simply note that by the definition of conditional probability, $p(x_i \mid x^{i-1}) = p(x^i)/p(x^{i-1})$. Using this to rewrite all factors on the right-hand side of (2), we see that every denominator cancels with the subsequent numerator, and (2) follows. Now, suppose you have evidence that $x_1, \ldots, x_n$ are distributed according to p. You observe the first $i-1$ outcomes, and you are asked to make a probabilistic prediction of the ith outcome. Then you would of course predict according to the distribution $p(x_i \mid x^{i-1})$. We may therefore think of the individual probabilities $p(x_i \mid x^{i-1})$ as *probabilistic predictions* of the ith outcome conditioned on the first $i-1$ outcomes. Assuming prediction error is measured by logarithmic loss, the total error incurred when sequentially predicting all outcomes, based on all previous outcomes, is $\sum_{i=1}^{n}[-\ln p(x_i \mid x^{i-1})]$. But now note that, since $\sum_{i=1}^{n}[-\ln p(x_i \mid x^{i-1})] = -\ln \prod_{i=1}^{n} p(x_i \mid x^{i-1})$, by (2), we have

$$\sum_{i=1}^{n} -\ln p(x_i \mid x^{i-1}) = -\ln p(x_1, \ldots, x_n). \tag{3}$$

Eq. (3) shows that *every distribution on n outcomes may be interpreted as a sequential prediction strategy. The accumulated logarithmic loss incurred by this strategy on data x^n is equal to minus the log likelihood of data x^n*. This crucial insight connects the APE method to maximum likelihood: suppose we compare two degenerate models M_1 and M_2, each model containing just one distribution (parameter instantiation), say, $p(\cdot \mid \theta_1)$ and $p(\cdot \mid \theta_2)$, respectively. Then, for all i, the ML predictor according to model M_j is simply the distribution $p(x_i \mid x^{i-1}, \theta_j)$, and the APE for model M_j is minus the log likelihood $-\ln p(x^n|\theta_j)$ of x^n. Thus, selecting the model with minimum APE now amounts to selecting the distribution maximizing the likelihood of the data: if the models under consideration are degenerate, then APE coincides with maximum likelihood. We shall now see that if a model contains more than one distribution, then APE mimics Bayesian rather than maximum likelihood inference.

Bayesian model selection procedures prefer the model M_i that has the highest probability $p(M_i \mid x^n)$ given the observed data. Assuming uniform priors on the models M_1 and M_2, by Bayes' rule, this is equal to the model maximizing the marginal probability of the observed data, $p(x^n|M_i)$. Even if the prior on M_1 and M_2 is not uniform, its influence on $p(M_i \mid x^n)$ is typically so small that in practice, one can safely assume that Bayesian model selection picks the model maximizing $p(x^n|M_i)$. The ratio of the marginal probabilities, $B_{12} = p(x^n|M_1)/p(x^n|M_2)$, is called the *Bayes factor* (e.g., Edwards, Lindman, & Savage, 1963; Jeffreys, 1961; Kass & Raftery, 1995), and the logarithm of the Bayes factor, $\log B_{12}$, is the weight of evidence provided by the data for model M_1 versus model M_2 (e.g., Good, 1985). The marginal probability of the data can be calculated by integrating out the model parameters θ:

$$p(x^n|M_j) = \int p(x^n|\theta, M_j)p(\theta \mid M_j)\, d\theta, \tag{4}$$

where $p(\theta \mid M_j)$ denotes the prior distribution of the parameter θ within M_j. Thus, in Bayesian model selection the marginal probability of the data is calculated by integrating or summing the probabilities of the observed data across the entire range of parameter values, weighted by their prior plausibility. A complex model with many parameters will generally have a low marginal probability of the data, because for certain regions of the parameter space the probability of the observed data is likely to be quite small (Myung & Pitt, 1997).

Now let us compare this to the APE with logarithmic loss and Bayesian predictions. The Bayesian predictive distribution of x_i based on model M_j and data x^{i-1} is just the conditional probability of x_i given x^{i-1}, according to $p(\cdot \mid M_j)$, which can be rewritten as follows:

$$p(x_i \mid x^{i-1}, M_j) = \int p(x_i \mid x^{i-1}, \theta, M_j)p(\theta \mid x^{i-1}, M_j)\, d\theta, \tag{5}$$

where $p(\theta \mid x^{i-1}, M_j)$ is the Bayesian *posterior probability* of parameter θ conditioned on data x^{i-1} and model M_j.

Applying (3) to the Bayesian marginal (4) and conditional (5), we see that the APE with logarithmic loss and Bayesian predictions (5) satisfy

$$-\ln p(x^n \mid M_j) = \sum_{i=1}^{n} -\ln p(x_i \mid x^{i-1}, M_j). \tag{6}$$

Since Bayesian model selection picks the model M_j minimizing the left-hand side, whereas the APE method picks the model minimizing the right-hand side, the two procedures are equivalent.

Under mild regularity conditions on the model M_j, the Bayesian posterior will resemble a Gaussian distribution with width of order $1/\sqrt{n}$ and mean equal to the maximum likelihood distribution $\hat{\theta}(x^{i-1})$ within M_j (Bernardo & Smith, 1994). Thus, with increasing n, it will become very sharply concentrated around $\hat{\theta}(x^{i-1})$, so that the predictive distribution (5) will very closely resemble the ML distribution $p(x_i \mid x^{i-1}, \hat{\theta}(x^{i-1}))$. This suggests that we can approximate (6) by adding prediction errors made with the ML distribution rather than the Bayesian posterior, that is:

$$-\ln p(x^n \mid M_j) \approx \sum_{i=1}^{n} -\ln p(x_i \mid x^{i-1}, \hat{\theta}(x^{i-1})). \tag{7}$$

The right-hand side is just the APE calculated using the ML rather than the Bayesian predictions. One can show theoretically that, under regularity conditions on M_j and the data generating distribution, (7) is indeed the case, the approximation holding to within a constant independent of n, and depending on the prior distribution. This makes the ML-based APE (which does not involve priors) an approximation of the log Bayes marginal likelihood that is often much easier to compute than the Bayes marginal likelihood itself. Let us illustrate all this with a simple example.

Example 2 (*APE with ML plug-in versus Bayesian prediction*). Suppose x^n is a sequence of n zeros and ones. Let $n_{[1]}$ denote the number of ones in x^n. The *Bernoulli* or *biased-coin model* M_0 is the set of all distributions $p(\cdot \mid \theta)$ such that the x^n are independent outcomes of a coin with bias θ, that is, $p(x^n \mid \theta) = \theta^{n_{[1]}}(1-\theta)^{n-n_{[1]}}$. The ML estimator $\hat{\theta}(x^n)$ is given by $n_{[1]}/n$ (Grünwald, 2005). Let us calculate the APE based on a Bernoulli model. The ith prediction is given by

$$p(X_i = 1 \mid \hat{\theta}(x^{i-1})) = \hat{\theta}(x^{i-1});$$
$$p(X_i = 0 \mid \hat{\theta}(x^{i-1})) = 1 - \hat{\theta}(x^{i-1}).$$

Note that if the sequence x^{i-1} contains only zeros, whereas X_i turns out be 1, then the logarithmic loss is $-\ln 0 = \infty$, leading to an infinite APE. This cannot be the intention. Indeed, in practice, we cannot start accumulating the APE at $i = 1$, but have to start at i^*, defined as the smallest i such that both a 0 and 1 have been observed in x^{i-1}. We discuss this 'startup problem' in more detail later. To compare the APE to the Bayesian marginal likelihood, let us consider the modified ML estimator

$$\hat{\theta}_{\lambda}(x^n) := \frac{n_{[1]} + \lambda}{n + 2\lambda}. \tag{8}$$

If we take $\lambda = 0$, we get the ordinary ML estimator. If we take $\lambda = 1$, then an exercise involving beta-integrals shows that, for all i, x^i,

$$p(x_i \mid \hat{\theta}_1(x^{i-1})) = p(x_i \mid x^{i-1}, M_0),$$

where $p(x_i \mid x^{i-1}, M_0)$ is the Bayesian predictive distribution (5) relative to the uniform prior $p(\theta \mid M_0) \equiv 1$. Thus, $\hat{\theta}_1(x^{i-1})$ corresponds to the Bayesian predictive distribution for the uniform prior. This prediction rule was advocated by the great probabilist P.S. de Laplace, co-originator of Bayesian statistics. We see that the Bayesian predictions are very closely related to the ML predictions, which strongly suggests that, as long as we start counting at $i = i^*$ rather than $i = 1$,

$$-\ln p((x_{i^*}, \ldots, x_n) \mid x_1, \ldots, x_{i^*-1}, M_j)$$
$$\approx \sum_{i=i^*}^{n} -\ln p(x_i \mid x^{i-1}, \hat{\theta}(x^{i-1})),$$

so that Bayesian model selection and APE based on ML predictions will tend to select the same distributions for all but the smallest n. This can also be shown formally (Grünwald, 2005).

For more general models, such simple modifications of the ML estimator usually do not correspond to a Bayesian predictive distribution; for example, if a model M_j is not convex[5] then a point estimator (i.e., an element of M_j) typically does not correspond to the Bayesian predictive distribution, which is a mixture (weighted average) of elements of M_j. Nevertheless, one can in many cases still show that an appropriate version of (7) holds.

Predictive interpretation of Bayesian inference: An often-voiced objection against the use of Bayes factors is that they supposedly depend on one of the models being true in the sense of being identical to the data generating process (e.g., Gelman, Carlin, Stern, & Rubin, 2004, p. 180; Spiegelhalter, Best, Carlin, & van der Linde, 2002). As Eq. (6) shows, Bayes factors do have a predictive interpretation, however, meaning that the model with the highest marginal probability of the data is guaranteed to also have the smallest accumulative sequential one-step-ahead prediction error—even when none of the models is true (Kass & Raftery, 1995, p. 777). The importance of sequential prediction for statistical inference has been stressed by Dawid and colleagues (Dawid, 1984, 1991, 1992; Dawid & Vovk, 1999; Skouras & Dawid, 1998), who argued that "(...) the purpose of statistical inference is to

[5] A model M is convex if for any two distributions p_1 and p_2 in the model, and every λ with $0 \leqslant \lambda \leqslant 1$, the distribution q defined by $q(x) := \lambda p_1(x) + (1-\lambda)p_2(x)$ is also in the model. This implies that the Bayesian posterior predictive distribution—a weighted average of distributions in M—must itself be a member of M. The Bernoulli model is convex, but most statistical models used in practice are not.

make sequential probability forecasts for future observations, rather than to express information about parameters.'' (Dawid, 1984, p. 278). Dawid, who together with Rissanen may be viewed as the originators of the APE method, termed this method 'prequential', combining 'sequential' and 'prediction'. The importance of prediction has also attracted interest from researchers in the field of mathematical psychology (cf. Busemeyer & Wang, 2000; Forster, 2000, pp. 226–227).

1.3. Model selection and the principle of minimum description length

The minimum description length principle originates from algorithmic coding theory (cf. Gammerman & Vovk, 1999; Li & Vitányi, 1997) and was pioneered as a method for model selection by Rissanen (1986b, 1987, 1996, 1999, 2001).[6] The MDL principle is based on the premise that any regularities in a data set allow it to be compressed (e.g., de Rooij & Grünwald, 2006; Grünwald, 2000, 2005; Grünwald, Myung, & Pitt, 2005; Hansen & Yu, 2001; Pitt et al., 2002). Hence, the best model is the model that minimizes the number of bits required to unambiguously encode both the data and the model itself. A very complex model will take many bits to encode, and this will only be worthwhile when that model is able to greatly reduce the number of bits required to encode the data.

Rissanen (1996, 2001) derived an elegant and very general formulation of MDL as a normalized maximum likelihood. According to this formulation, the MDL criterion may be calculated by first obtaining the maximum likelihood for the observed data, and then dividing this quantity by an integral or sum of maximum likelihoods over all possible other data sets that could have been observed but were not. Hence, the normalized maximum likelihood interpretation of MDL leads to the selection of models that minimize

$$-\ln \frac{p(x^n \mid \hat{\theta}(x^n))}{\sum_{y^n \in \mathscr{X}^n} p(y^n \mid \hat{\theta}(y^n))}, \tag{9}$$

where $\mathscr{X}$ is the set of possible values for x_i, and the sum is to be replaced by an integral if the x_i are real-valued. It turns out that, under conditions on the models M_1 and M_2, for large samples, MDL model selection between M_1 and M_2 *precisely* coincides with Bayesian model selection based on the Bayesian marginal likelihood (4) equipped with *Jeffreys' prior* (Grünwald, 2005). Thus, MDL is closely related to so-called 'objective Bayesian' procedures (Kass & Raftery, 1995). Not surprisingly then—and this is crucial for the present discussion—MDL may also be given a predictive interpretation (Rissanen, 1986a) comparable to the predictive interpretation of Bayesian model selection. This leads to the 'predictive MDL' technique that calculates the accumulative one-step-ahead prediction error using the recipe outlined earlier, and measures error using the logarithmic loss function. Predictions for the next observation are based on plug-in maximum likelihood point estimates for the parameters. The predictive MDL technique may be viewed as a close approximation to the 'ideal' (but hard to compute) version of MDL based on normalized maximum likelihood. This is explained at length in Grünwald (2005). Here, we merely give a suggestive example:

Example 3. Consider once again the modified ML estimator (8) for the Bernoulli model. A similar calculation as the one for $\lambda = 1$, again using beta-integrals, shows that if we take $\lambda = \frac{1}{2}$, the resulting estimator is equal to the predictive distribution (5) defined relative to *Jeffreys' prior* rather than the uniform prior. For the Bernoulli model, Jeffreys' prior is given by $p(\theta|M_0) = 1/(\pi\sqrt{\theta(1-\theta)})$. It follows that the APE calculated with logarithmic loss and a slightly modified ML estimator (using $(n_{[1]} + 0.5)/(n+1)$) is equal to the minus log Bayesian marginal likelihood with Jeffreys' prior, which is asymptotically indistinguishable from the minus log normalized maximum likelihood(9).

In sum, both MDL and Bayesian model selection can be viewed as procedures that aim to minimize one-step-ahead prediction error by making sequential forecasts and accumulating the resulting prediction errors.

1.4. Pros and cons of accumulating prediction errors

The previous sections have shown that the APE procedure is a data-driven method with a strong theoretical foundation. By measuring predictive performance for unseen data, the APE method automatically adjusts for model complexity in a conceptually straightforward manner. If desired, the specification of priors can be avoided by using the plug-in ML estimate for prediction. For larger sample sizes, the plug-in APE method approximates a full Bayesian solution that would have required the specification of priors. Further, the APE method is generally *consistent*, that is, as $n \to \infty$ it will select the data-generating model, if such a model exists, with probability one (cf. Dawid, 1992; De Luna & Skouras, 2003; Hemerly & Davis, 1989).

The APE method also has great practical advantages. In particular, it is easy to apply, as the only requirement is that the model can generate predictions. Hence, the APE method can give informative results almost effortlessly, even for very complicated models. In contrast to the APE method, both Bayesian model selection and MDL require the evaluation of integrals. In Bayesian model selection such integrals refer to the parameter space whereas in the NML interpretation of MDL they refer to the sample space. Even for relatively simple hypothesis testing problems, these integrals can be difficult to evaluate analytically (see also de Rooij & Grünwald, 2006). In these cases, one of the standard solutions is to evaluate the

[6] The similar method of *minimum message length* was proposed earlier by Wallace and Boulton (1968), see also Wallace and Freeman (1987).

integrals numerically, using Markov chain Monte Carlo (MCMC) techniques (e.g., Gilks, Richardson, & Spiegelhalter, 1996; in particular, see Raftery, 1996).[7] The effort needed to understand, construct, program, and check an MCMC simulation is considerable. For certain inference problems, the APE method provides a practical and easy-to-understand alternative to these more involved procedures.

Unfortunately, the APE method also has some drawbacks. Some of these drawbacks, however, are more apparent than real. We will discuss four APE features that may limit its scope of application:

1. *Computational effort* (Hansen & Yu, 2001; Wei, 1992). As a rule, data-driven methods are computationally expensive, and APE is no exception. A model selection problem that involves k competing models and n observations will require at the most $k(n-1)$ model fits, and $k(n-1)$ calculations of prediction error. For most inference problems in the field of psychology, however, the computation burden is anything but prohibitive.

2. *Startup problems* (Qian et al., 1996; Rissanen, 1992; Wei, 1992). For accurate parameter estimation, complex models require a relatively large number of observations. Hence, in the early stages of the APE method a complex model may be heavily punished for making inferior predictions. In fact, predictions may be even undefined or of disastrous quality, as we indicated in Example 2. At the later stages of the APE method, however, the number of observations may be large enough to allow relatively accurate parameter estimation, and as a result the complex model may begin to make superior predictions.[8] With respect to this concern, we agree with Dawid (1992), who states "It seems to me perfectly reasonable that a complex model should be heavily penalized in the initial stages, since the slow rate at which its parameters can be learned means that it may for a long time predict more poorly than a simpler "incorrect" model. In this case I would rather use the simple model until the data are sufficiently extensive as to demand more detailed description" (pp. 124–125).

3. *Choice of loss function* (Dawid, 1992). What loss function should be used to quantify prediction error? For certain problems, the choice of loss function may be guided by substantive knowledge of the problem at hand. Dawid (1992) discusses a case in which the prognosis of a patient can be either 'full recovery', 'partial recovery', or 'death'. When assessing the predictive adequacy of a specific medical examination, one could devise a loss function that formally takes into account the intuitive notion that 'partial recovery' is closer to 'full recovery' than it is to 'death'. In practice, however, a model is selected mainly to obtain insight, or to guide the search for further experiments. In other cases, at the time of model selection, it may be simply be unknown what type of predictions one will want to use the model for. In such cases, one should adopt a generic loss function such as the logarithmic loss (making APE coincide with Bayes/MDL) or the squared loss for point predictions, which, in case the errors are Gaussian, is equivalent to the logarithmic loss (cf. Gelman et al., 2004, p. 180; Rissanen, 1986a). An advantage of the squared loss is computational simplicity, especially when the point prediction is based on the maximum likelihood estimate.

4. *Ordering of data* (Hansen & Yu, 2001). Suppose the data are i.i.d. (independently and identically distributed) according the models M_j under consideration. Then the data is best viewed as unordered, and it seems that in such a case, a reasonable model selection method does not depend on the arbitrary order in which the data are entered. Indeed, in a full Bayesian analysis the order in which the data x^n enter the APE method is inconsequential. This can be easily seen from the factorization of the marginal probability of the data, $\ln p(x^n|M_i) = \sum_{n=1}^{N} \ln p(x_n|x^{n-1}, M_i)$. For non-Bayesian methods such as those using ML point prediction or those using a squared error loss function, however, the sequential ordering of the data does influence the results, particularly for small samples. The dependence of the APE on the ordering of the data is of course undesirable, and several methods have been proposed to remedy the situation. An obvious solution is to calculate the final APE as an average of APEs for many random orderings of the same data set (Kontkanen et al., 2001; Rissanen, 1986a). The drawback of this procedure is that it greatly increases the computational burden. Another solution to the ordering dependence is to derive analytic approximations to APE that are invariant with respect to the ordering of the data (e.g., the Fisher Information Criterion; Wei, 1992). The issue of ordering dependence does not play a role for time series data, since the definition of a time series implies a single natural ordering of the data. The current work applies the APE method to the discrimination of long-range dependence from short-range dependence in time series (for an introduction in time series analysis see Priestley, 1981). These concepts are explained in the next section.

2. Long-range dependence versus short-range dependence

The difference between long-range dependence and short-range dependence is in the rate with which the autocorrelation decays. For long-range dependence, the decay with increasing lag k is so slow that the autocorrelation $C(k)$ sums to infinity, that is,

$$\sum_{k=-\infty}^{\infty} C(k) = \infty. \tag{10}$$

This means that in the frequency domain, long-range dependence is associated with a log–log power spectrum that keeps increasing at the low frequencies.

[7] Another solution is to reply on asymptotic approximations such as those provided by the Laplace method (Raftery, 1996).

[8] In the method of forward validation proposed by Hjorth (1982), sequential predictions are weighted by the number of observations upon which they are based. This reduces the concern that complex models may be overly penalized in the initial stages of the sequential prediction procedure.

In the case of short-range dependence, the rate with which the autocorrelations decay is relatively fast, so that the autocorrelation function sums to a finite number, that is,

$$\sum_{k=-\infty}^{\infty} C(k) = constant < \infty. \tag{11}$$

This means that in the frequency domain, short-range dependence is associated with a log–log power spectrum that flattens off at the low frequencies. The difference between long-range dependence and short-range dependence is quite fundamental. Long-range dependence has special features such as scale-invariance and self-similarity (cf. Baillie, 1996; Beran, 1994; Doukhan, Oppenheim, & Taqqu, 2003; Gisiger, 2001; Mandelbrot, 1977; Rangarajan & Ding, 2003). Short-range dependence, in contrast, is not associated with these special properties.

Long-range dependence is also of scientific interest because it has been reported in a remarkably wide variety of different systems, suggesting the presence of a common underlying principle (cf. Bak, 1996; Bak, Tang, & Wiesenfeld, 1987; Sornette, 2000; Van Orden, Holden, & Turvey, 2003; but see Wagenmakers, Farrell, & Ratcliff, 2005). Examples of long-range dependence include the electric current in transistors, water levels in the river Nile, the size of tree rings, brain activity as recorded by magnetoencephalogram, the stock market, music, and speech (e.g., Handel & Chung, 1993; Hosking, 1984; Hurst, 1951; Novikov, Novikov, Shannahoff-Khalsa, Schwartz, & Wright, 1997; Voss & Clarke, 1975; Wolf, 1978).[9]

In cognitive psychology, evidence for long-range dependence was recently found in a range of tasks such as mental rotation, lexical decision, speeded visual search, estimation of distance, estimation of rotation, estimation of force, estimation of time, simple reaction time, and word naming (Gilden, 1997, 2001; Gilden, Thornton, & Mallon, 1995; Van Orden et al., 2003; but see Wagenmakers, Farrell et al., 2004). Long-range dependence has also been reported in human motor tasks (Chen, Ding, & Kelso, 1997, 2001; Ding, Chen, & Kelso, 2002; Yoshinaga, Miyazima, & Mitake, 2000; Yulmetyev, Emelyanova, Hänggi, Gafarov, & Prokhorov, 2002; but see Pressing & Jolley-Rogers, 1997), in day-to-day fluctuations in self-esteem (Delignières, Fortes, & Ninot, 2004), in the temporal dynamics of tics in Gilles de la Tourette syndrome (Peterson & Leckman, 1998), and in day-to-day fluctuations in self-mood of bipolar patients (Gottschalk, Bauer, & Whybrow, 1995).

Obviously, in order to conclude that a times series is long-range dependent it is crucial that the alternative explanation in terms of a short-range process can safely be excluded. In practical applications, however, the special properties associated with long-range processes can be approximated quite well by certain short-range processes (e.g., Beran, 1994, p. 144; Crato & Ray, 1996; Hosking, 1984; Lawrance & Kottegoda, 1977). These short-range models generally provide competitive fits by mimicking the behavior of long-range models through carefully tuning parameter values. This problem of mimicry is compounded for short series and overall weak serial dependence.

Fig. 1 shows two example time series of length $n = 400$. The time series in panel A originates from a long-range process. In panel B, the autocorrelation of this series decays as a power function. Panel C shows the linear log–log power spectrum that characterizes long-range processes. The thick dotted line indicates a slope of -0.8. The time series in panel D was generated by a short-range process tuned to the time series from panel A. The autocorrelation function of this process decays exponentially, and its log–log spectrum is curved rather than linear. Despite the theoretical differences between long-range dependence and short-range dependence, panels E and F show that in actual practice these types of dependence can be hard to distinguish due to the stochastic nature of the processes involved.

The foregoing illustrates that the problem of deciding whether or not a time series is long-range dependent is ultimately a problem of model selection. This requires the definition of a short-range time series model and a long-range time series model. These models should then be fitted to the same data set and the model that has the highest generalizability should be preferred. A framework that is excellently suited for this particular job is autoregressive fractionally integrated moving average (ARFIMA) time series modeling. The reader not interested in the details may skip to the final paragraph of the next section.

3. The ARMA(1,1) model versus the ARFIMA$(0, d, 0)$ model

Standard Box–Jenkins ARIMA time series models (Box & Jenkins, 1970) exclusively generate short-range dependence. These models consist of three components: the autoregressive ‘AR’ component, the moving-average ‘MA’ component, and the integration ‘I’ component that determines how many times the ARMA series should be integrated. For example, a first-order AR process can be written as AR(1), or ARIMA(1,0,0), or $x_t = \phi x_{t-1} + \varepsilon_t$. In an AR(1) model, the value of a series at time t is a proportion of its value at time $t-1$, plus noise. A second-order AR process would be denoted AR(2), ARIMA(2,0,0), or $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \varepsilon_t$, from which it is evident that the value at time $t-2$ now has an independent effect on the value at time t. Similarly, a first-order MA process is given by MA(1), or ARIMA(0,0,1), or $x_t = \theta \varepsilon_{t-1} + \varepsilon_t$. In an MA(1) model, the current value of a series at time t is determined in part by the value of the noise at time $t-1$. Finally, a white noise series is given by ARIMA(0,0,0), and a random walk series is denoted by ARIMA(0,1,0). Thus, the ARIMA(p, d, q) model consists

[9] A comprehensive bibliography of $1/f$ noise is maintained by Wentian Li at http://www.nslij-genetics.org/wli/1fnoise/.

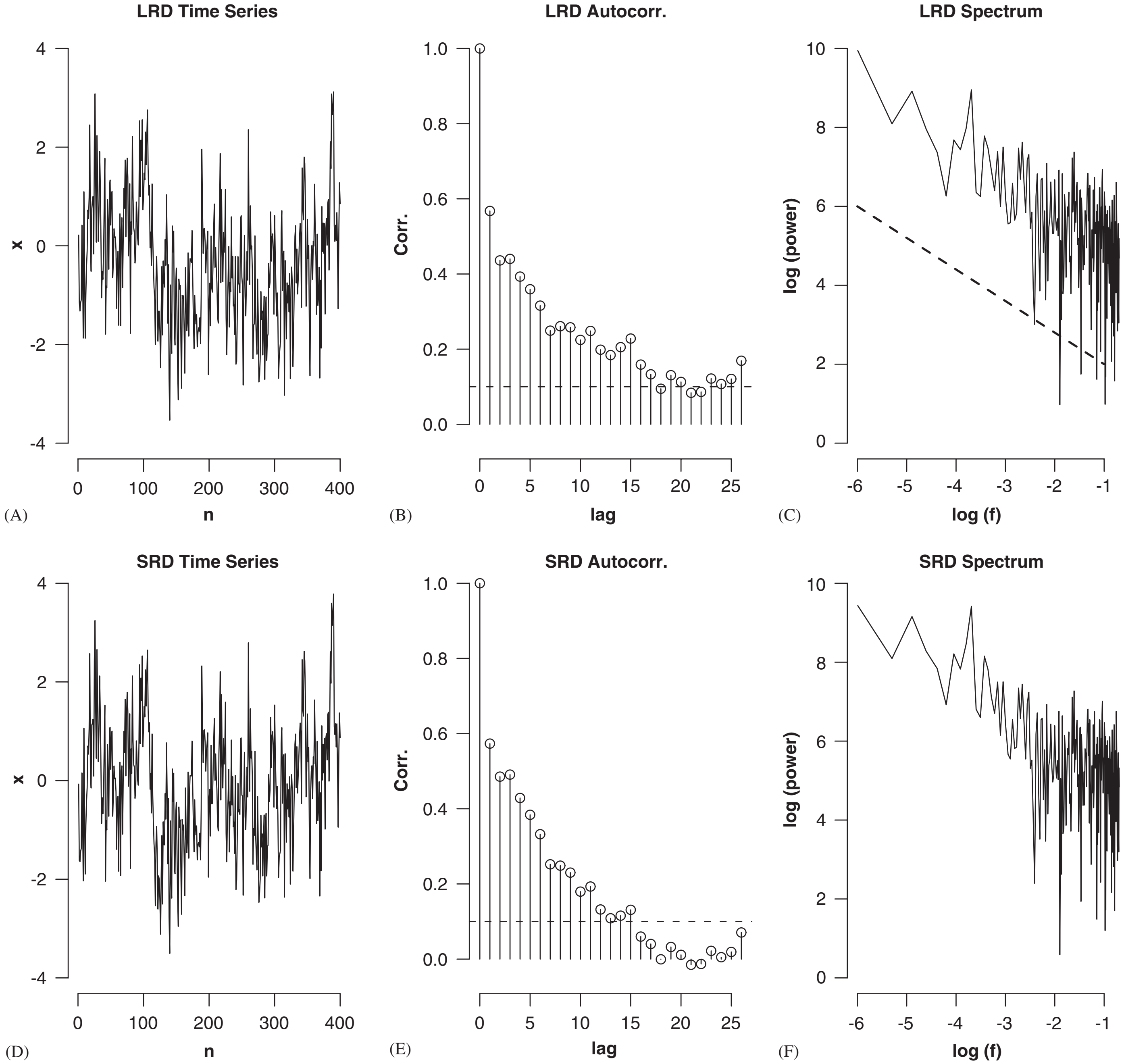

Fig. 1. Long-range or short-range dependence? Panel A: example of a long-range time series (i.e., ARFIMA$(0, d = 0.4, 0)$); panel B, autocorrelation function for the time series in panel A; panel C, log–log power spectrum for the time series in panel A; panel D, example of a short-range time series (i.e., ARMA$(\phi = 0.887, \theta = -0.536)$; parameter values determined by fitting the panel A long-range time series); panel E, autocorrelation function for the time series in panel D; panel F, log–log power spectrum for the time series in panel D.

of a pth order AR process, d times integration, and qth order MA process.

Granger and Joyeux (1980) and Hosking (1981) generalized the ARIMA model to account for long-range correlations. This was accomplished by letting the integration component take on fractional values, resulting in an ARFIMA(p, d, q) model:

$$\Phi(B)(1 - B)^d (X_t - \mu) = \Theta(B)\varepsilon_t, \qquad (12)$$

where B is the backward shift or lag operator defined as $BX_t = X_{t-1}$, $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is the AR polynomial, and $\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ is the MA polynomial (Beran, 1994; Hosking, 1984). For stationarity of the ARMA part of the model, it is required that the roots of $\Phi(z) = 0$ and $\Theta(z) = 0$ lie outside the unit circle (e.g., Priestley, 1981, pp. 132–135). In (12), the mean of the process is given by μ, and ε_t is a purely random process with $E(\varepsilon_t) = 0$, $E(\varepsilon_t \varepsilon_{t+h}) = 0$ $\forall h \neq 0$, and variance σ_ε^2.

In a short-range dependent ARIMA model, d in (12) can take on integer values only. However, if d is allowed to take on real values, the fractional differencing operator ∇^d is given by

$$\nabla^d = (1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k,$$

with binomial coefficients

$$\binom{d}{k} = \frac{d!}{k!(d-k)!} = \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)},$$

where $\Gamma(\cdot)$ is the gamma function. Hence, fractional differentiation of a time series X_t can be described as

$$\begin{aligned}\nabla^d X_t &= (1-B)^d X_t \\ &= \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k X_t \\ &= X_t - dX_{t-1} - \frac{1}{2}d(1-d)X_{t-2} \\ &\quad - \frac{1}{6}d(1-d)(2-d)X_{t-3} - \cdots .\end{aligned}$$

(Hosking, 1984). If $\nabla^d X_t$ is a white noise process, then X_t is an ARFIMA$(0,d,0)$ process or *fractional white noise*. X_t is stationary when $d \in (-\frac{1}{2}, \frac{1}{2})$, and X_t is said to be *persistent* or long-range dependent when $d \in (0, \frac{1}{2})$. When the data generating process is pure fractional white noise, the log–log power spectrum is described by $1/f^{\alpha}$, and $d = \frac{1}{2}\alpha$.

The ARFIMA(p,d,q) model can be thought of as fractional white noise passed through an ARMA(p,q) filter. The low-frequency behavior of the ARFIMA(p,d,q) model is determined by d, as the ARMA(p,q) process is only short-range. One of the advantages of the ARFIMA framework is that it allows simultaneous estimation of a long-range d component and a short-range ARMA(p,q) component. Thus, to determine the presence of long-range dependence one could test whether or not $\hat{d} = 0$. This procedure, however, is spuriously affected by short-range processes. For instance, if the data generating process is a short-range ARMA(1,1) with the AR component slightly higher than the MA component, $\hat{d}$ as estimated from the ARFIMA$(0,d,0)$ model will usually exceed 0.

To avoid this problem, a model selection approach is advisable. The family of ARFIMA models encompasses both short-range and long-range models, and their relative merits may be compared using model selection techniques such as AIC and BIC (e.g., Crato & Ray, 1996; Hosking, 1984; Wagenmakers et al., 2005). In this article we focus on the comparison between one short-range model and one long-range model (cf. Thornton & Gilden, 2005; Wagenmakers, Farrell et al., 2004). The generalization to more than two models is self-evident. The short-range model is the ARFIMA$(1, d = 0, 1)$ or ARMA(1,1) model. The ARMA(1,1) model is known to mimic long-range dependence for specific combinations of its AR and MA parameters (Basak et al., 2001; Beran, 1994; Crato & Ray, 1996; Thornton & Gilden, 2005), and it may be conceptualized as an AR(1) process plus independent white noise (Granger & Morris, 1976; Pagano, 1974).

The long-range model that competes with the short-range ARMA(1,1) model is the long-range ARFIMA$(0,d,0)$ fractional white noise model. An appealing argument that is often advanced in support of the ARFIMA$(0,d,0)$ model is based on the principle of parsimony (Beran, 1994; Thornton & Gilden, 2005). It is argued that although the ARMA(1,1) process may be able to mimic the ARFIMA $(0,d,0)$ process, this mimicry only occurs after the ARMA parameters have been carefully tuned to the data under consideration. Formal model selection methods that focus on generalizability (i.e., predictive performance to unseen data) can be used to assess whether and to what extent this assertion holds.

In sum, the ARFIMA framework allows the comparison between a short-range ARMA(1,1) and a long-range ARFIMA$(0,d,0)$ model. When testing for long-range dependence, self-similarity, or power-law scaling in psychological time series, it is important that the conceptually simply short-range ARMA(1,1) model can be excluded from consideration. This is not always feasible, because for medium size times series the ARMA(1,1) model can mimic the features of the ARFIMA$(0,d,0)$ model quite well. In the remainder of this article, we assess the discriminability and generalizability of the short-range and long-range models using AIC, BIC, and APE.

4. Outline of the Monte Carlo simulations

The primary aim of the Monte Carlo simulations is to illustrate the use of the APE procedure and compare it to other model selection methods such as AIC and BIC. The philosophy behind the simulations is similar to the one advocated by Wagenmakers, Ratcliff, Gomez, and Iverson (2004): representative time series are generated from both the ARMA(1,1) and the ARFIMA$(0,d,0)$ model. These simulated time series are then being fit by both the ARMA(1,1) and ARFIMA$(0,d,0)$ models. As will be apparent later, the result of this exercise is informative with respect to the overall discriminability of the models, and is also informative with respect to the extent to which model complexity should be penalized.

In order to generate representative samples from the ARMA(1,1) model, $x_t = \phi x_{t-1} + \theta \varepsilon_{t-1} + \varepsilon_t$, values for the AR and MA parameters were systematically sampled from a uniform prior distribution. To obtain stationary series, both the AR parameter ϕ and the MA parameter θ were constrained to lie in the interval $(-1,1)$. Further, to obtain series whose autocorrelations and power spectra are in accord with those observed in psychological time series, $\phi \in (0,1)$, $\theta \in (-1,0)$ and $|\theta| < \phi$. These constraints yield stationary series with spectra that decrease in power as frequency increases (Thornton & Gilden, 2005; cf. Fig. 1, panel F). From this joint prior distribution on ϕ and θ, 1225 draws were obtained using systematic sampling (i.e., non-random sampling using equispaced intervals). Each draw of parameter values then served to generate a single time series of length 500. For the ARFIMA$(0, d, 0)$ model, representative time series were generated by taking 1225 systematic draws from a uniform prior distribution for $d \in (0, \frac{1}{2})$. Each value of d served to generate a single time series of length 500. For each of the 2450 time series, the first 100 values were treated as startup values and discarded, leaving $n = 400$ simulated observations. For both the ARMA(1,1) and the ARFIMA$(0, d, 0)$ model, the Gaussian innovations are given by $\varepsilon_t \sim N(0, \sigma^2_\varepsilon = 1)$. Thus, when generating the data the error variance is fixed at 1.

Next, both the ARMA(1,1) model and the ARFIMA $(0, d, 0)$ model were fitted to each simulated time series, interest centering on values for maximum log likelihood, AIC, BIC, and APE. Note that the short-range ARMA(1,1) model and the long-range ARFIMA$(0, d, 0)$ model are not nested and differ in the number of free parameters. That is, the ARMA(1,1) has three free parameters (i.e., ϕ, θ, and the error variance σ^2_ε), and the ARFIMA$(0, d, 0)$ model has two free parameters (i.e., d and σ^2_ε). The fit of both models to data was computed using maximum likelihood (e.g., Beran, 1994; Doornik & Ooms, 2003; Sowell, 1992a, 1992b). Throughout this article, we used the Ox arfima package (Doornik, 2001; Doornik & Ooms, 2003) for generating simulated time series and for the computation of exact Gaussian maximum likelihood.

Thus, the maximum log likelihood ℓ is immediately obtained from the fitting procedure (Doornik & Ooms, 2003). The AIC (Akaike, 1974) is usually calculated as $AIC = -2\ell + 2k$, where k is the number of free parameters. Here, we follow the advise of Burnham and Anderson (2002), and calculate the small sample version of AIC, AICc instead:

$$AICc = -2\ell + 2k\frac{n}{n-k-1}$$

(Hurvich & Tsai, 1989). Thus, the correction factor for small samples is $\frac{n}{n-k-1}$, which means that as $n \to \infty$, AICc $\to$ AIC. The BIC (Schwarz, 1978; Raftery, 1995) is calculated as $BIC = -2\ell + k \ln n$.

Calculation of APE is data-driven, but it is not completely automatic, as several choices need to be made. The APE method used here is very similar to predictive MDL (Rissanen, 1986a), in that it uses the 'plug-in' maximum likelihood parameter estimates for prediction of the next observation (cf. Dawid, 1984, p. 287, and Skouras & Dawid, 1998). The advantages of the plug-in approach are computational ease and the fact that no prior distribution for the parameters needs to be specified. To evaluate the prediction error we used squared error loss (cf. Rissanen, 1986a). Finally, for each time series we did not compute prediction errors until the most complex model had become identifiable (i.e., prediction errors are not calculated for the first four observations). De Luna and Skouras (2003) recently made the same choices for calculating APE.

To assess performance of the different model selection methods, the most often-used measure is the probability of correct model recovery in Monte Carlo simulations. Although very useful, this method presupposes knowledge of the data generating process. In real-world applications, models are only approximately "true". To assess what model selection method is most appropriate for a particular real-world time series, we apply the model meta-selection method introduced by Clarke (2001) and De Luna and Skouras (2003). This method is similar to APE in that it also assesses accumulative prediction errors. The model meta-selection method, however, computes APE for model selection methods instead of models. Thus, the method quantifies the size of the prediction errors if, say, AICc was used to select the model that predicts the next data point.

5. Results of the Monte Carlo simulations

An important research question concerns the extent to which ARMA(1,1) and ARFIMA$(0, d, 0)$ time series can be differentiated in practical applications (i.e., for series with only a modest number of observations). Fig. 2, panel A, shows two distributions of differences in log likelihoods,

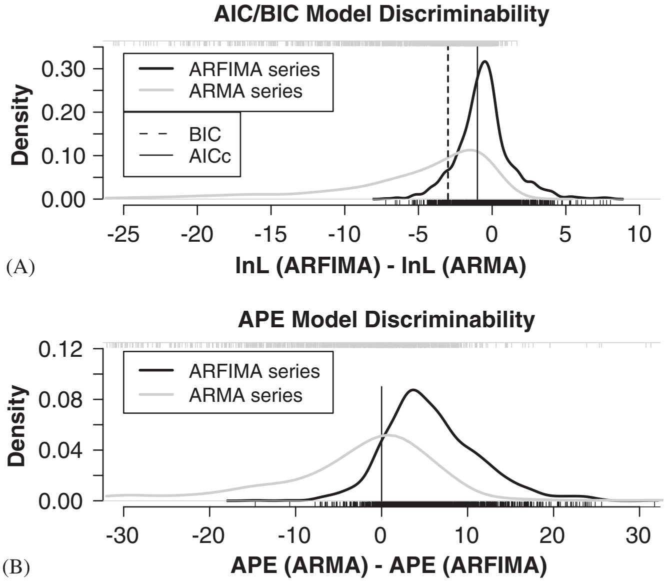

Fig. 2. Model discriminability after 400 simulated observations. Panel A, difference in log likelihood; panel B, difference in APE. Each tick mark inside the figure frame corresponds to a single time series.

Table 1
Probability of correct model recovery using AICc, BIC, and APE for time series generated by ARMA(1,1) and ARFIMA$(0, d, 0)$ models

Index	Model recovery		
	ARMA	ARFIMA	Average
AICc	0.82	0.63	0.73
BIC	0.58	0.92	0.75
APE	0.60	0.89	0.75

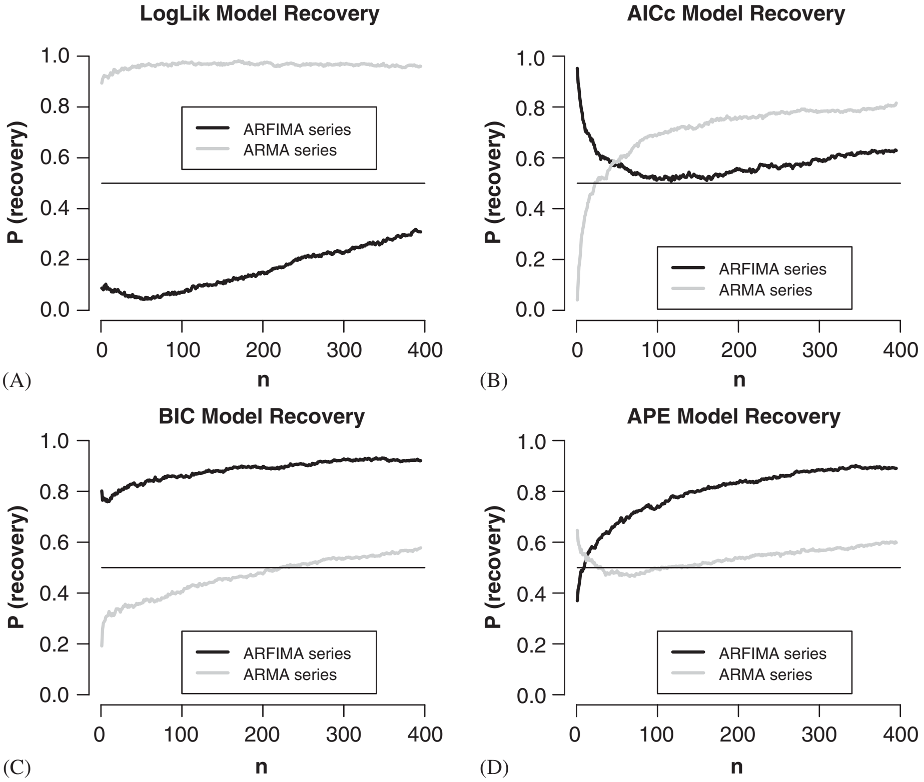

Fig. 3. Probability of correct model recovery as a function of the number of simulated observations. Panel A, recovery based on log likelihood; panel B, recovery based on AICc; panel C, recovery based on BIC; panel D, recovery based on APE.

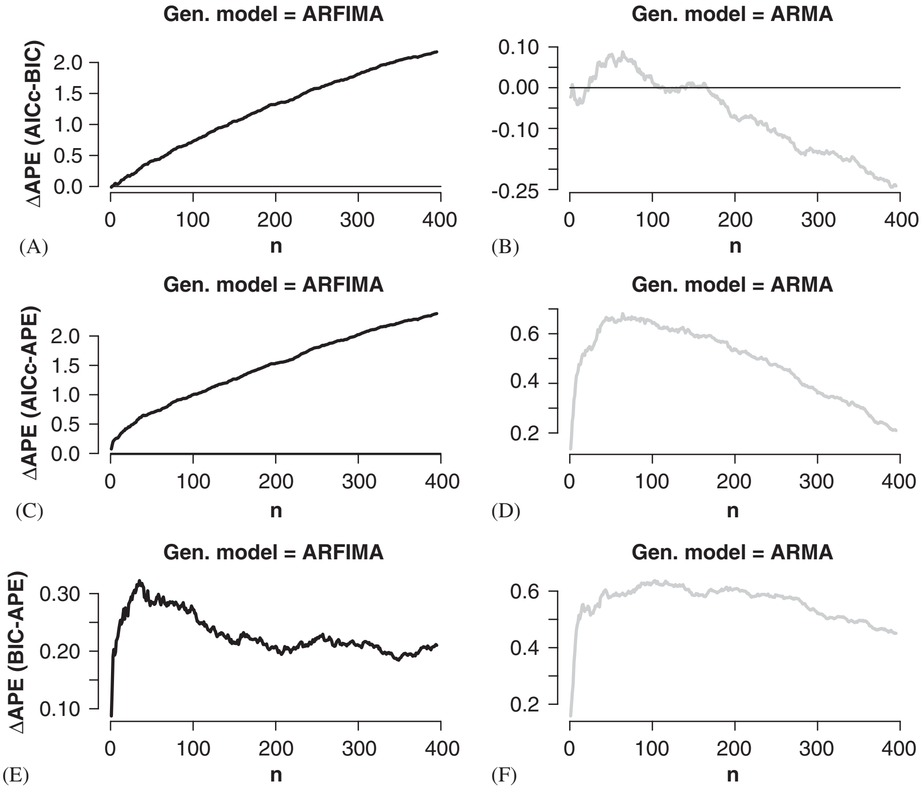

Fig. 4. Model meta-selection as a function of the generating model and the number of simulated observations. Each panel shows the difference in APE for various model selection methods. Panels A and B, $APE_{AICc} - APE_{BIC}$; panels C and D, $APE_{AICc} - APE_{APE}$; panels E and F, $APE_{BIC} - APE_{APE}$.

one for the 1225 series of length 400 generated by the ARMA(1,1) model, and one for the 1225 series generated by the ARFIMA$(0,d,0)$ model.[10] On the x-axis, values greater than zero indicate that the ARFIMA$(0,d,0)$ model provided a better fit than did the ARMA(1,1) model. As may be expected from the fact that the ARMA(1,1) model has an additional parameter compared to the ARFIMA$(0,d,0)$ model, the ARMA(1,1) provides a superior fit not only for most of the ARMA time series, but also for most of the ARFIMA time series. Thus, log likelihood alone forms a poor basis for model selection, as the majority of time series would be classified as ARMA, even when the true generating process is ARFIMA.

In Fig. 2, panel A, vertical lines indicate the AICc and BIC criteria. Both AICc and BIC 'correct' the log likelihood by penalizing the ARMA(1,1) model. The BIC criterion favors the ARFIMA model more than does the AICc criterion, as the BIC punishment for an additional parameter is relatively high (i.e., 5.99 for BIC versus 2.01 for AICc). Note that the total number of correct classifications is maximized when the criterion is placed at the point where the distributions intersect.

Fig. 2, panel B, shows two distributions of differences in APE, one for the 1225 series generated by the ARMA(1,1) model, and one for the 1225 series generated by the ARFIMA$(0,d,0)$ model. Again, values on the x-axis greater than zero indicate that the ARFIMA$(0,d,0)$ model is to be preferred over the ARMA(1,1) model. Panel B shows that model selection by APE is centered correctly, that is, the nominal criterion $\Delta APE = 0$ automatically incorporates a correction for model complexity. Moreover, for the present model selection problem the $\Delta APE = 0$ criterion is approximately optimal, as it is located closely to the point where the distributions intersect. In contrast to APE, log likelihood is not centered correctly, and thus requires a correction using explicit penalty terms.

Table 1 shows the probability that AICc, BIC, and APE recover the model that generated the data. The probability of correct model recovery can also be obtained from Fig. 2. For instance, the probability of correctly identifying an ARFIMA-generated time series using APE is equal to the area under the black curve in panel B that lies to the right of the $\Delta APE = 0$ criterion. Although all three model selection methods show comparable classification performance overall, AICc recovers the ARMA(1,1) model better than the ARFIMA(1,1) model, whereas both BIC and APE show the opposite pattern.

Whereas Table 1 shows model recovery performance after 400 observations, Fig. 3 plots performance as a function of the entire sequence of simulated observations. Thus, Fig. 3 shows, for instance, how the percentage of ARMA(1,1) time series that were classified correctly according to AICc after each of $n \in \{5, 6, \ldots, 400\}$ observations. Panel A confirms that model selection by uncorrected log likelihood is biased toward selection of the ARMA(1,1) model. Panel B shows that the AICc penalty term to a certain extent corrects this bias. After about 50 observations, recovery for both ARMA and ARFIMA series is above chance. Nonetheless, recovery of ARMA series is still consistently higher than that of ARFIMA series. Panel C shows that the BIC penalty term leads to a general preference for the ARFIMA$(0,d,0)$ model. Model recovery using APE shows a qualitatively similar pattern to model recovery using BIC, although the extent of the ARFIMA preference is a little less for APE than it is for BIC, particularly when $n < 200$.

Thus, substantial differences exist between AICc on the one hand and BIC and APE on the other. In particular, AICc tends to prefer the ARMA(1,1) model, whereas BIC and APE tend to prefer the ARFIMA$(0,d,0)$ model. Unfortunately, assessment of performance via model recovery simulations does not provide much guidance on

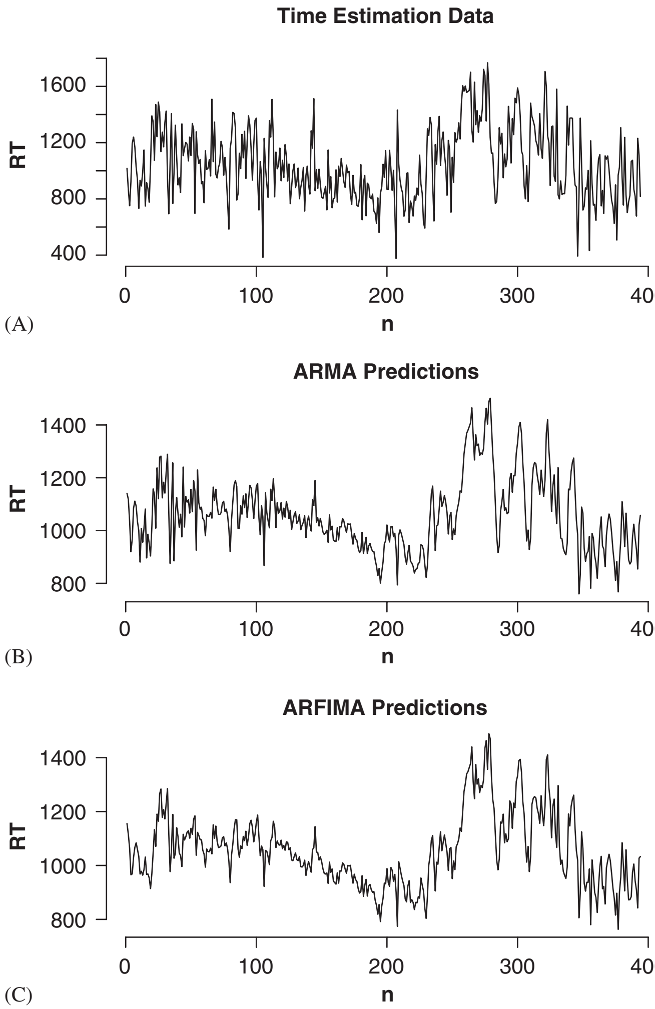

Fig. 5. Estimation of 400 one second time intervals. Panel A, observed response latency; panel B, series of one-step-ahead predictions from the ARMA(1,1) model; panel C, series of one-step-ahead predictions from the ARFIMA$(0,d,0)$ model.

[10] These distributions are based on kernel density estimation using a Gaussian kernel with window width $h = 0.9An^{-\frac{1}{5}}$ (cf. Eq. (3.31) in Silverman, 1986), where $A = \min(\textit{standard deviation}, \textit{interquartilerange}/1.34)$, and n is the number of observations.

which selection method to use, as the overall probability of correct recovery is almost equal (cf. Table 1). Besides, even if the model recovery procedure had shown that BIC and APE outperform AICc, AIC fans could rightly point out that the model that generated the data for fitting is also present among the candidate models. This is a circumstance that is unlikely to hold in scientific practice, where almost all models are abstractions of reality and never precisely "true" (cf. Burnham & Anderson, 2002).

An alternative method to assess performance for model selection methods is to quantify their predictive performance through a procedure known as *model meta-selection* (Clarke, 2001; De Luna & Skouras, 2003). The aim of this procedure is to estimate the predictive value not of the models (i.e., ARMA and ARFIMA), but of the model selection methods (i.e., AICc, BIC, and APE). Just as in the calculation of APE, the meta-selection procedure requires one to fit the ARMA(1,1) and ARFIMA$(0, d, 0)$ models for each of an increasing number of observations. The predictive value of, say, AICc can then be quantified by the accumulative prediction error for the models chosen by AICc. For instance, suppose that for a particular time series, AICc prefers the ARMA model up until the data set has increased to $n = 200$, after which AICc starts to prefer the ARFIMA model. The accumulative prediction error for the AICc model selection procedure is then a sum of the prediction errors made by the ARMA and ARFIMA models (for the first and second half of the time series, respectively). The advantage of using the model meta-selection method is that it allows a concrete, data-driven assessment of the relative value of model selection tools such as AIC and BIC. Also, the model meta-selection method is *consistent* in the sense that it will eventually prefer the model selection tool that provides the best forecasts (cf. De Luna & Skouras, 2003, Theorem 1). Finally, model meta-selection does not require the data generating model to be in the set of candidate models. This advantage will be exploited when we later consider performance of model selection methods for a single real-world time series.

Fig. 4 shows the results of the model meta-selection method applied to the simulated time series, separately for ARMA series and ARFIMA series. The panels of Fig. 4 generally show $\Delta APE > 0$, indicating the following ordering on APE: $APE(AICc) > APE(BIC) > APE(APE)$. That is, the accumulative prediction errors are highest for models selected according to AICc, and are lowest for models selected according to APE. The only exception to this rule concerns panel B, which shows that for ARMA time series the AICc is predictively better than the BIC.

In sum, the above results demonstrate that for time series of length $n = 400$, the ARMA(1,1) model is generally difficult to discriminate from the ARFIMA$(0, d, 0)$ model. AIC behaves very differently from BIC, and BIC is qualitatively similar to APE. The model meta-selection procedure shows that for the simulated time series, the use of AICc leads to larger one-step-ahead prediction errors than the use of BIC or APE.

6. Real-world example: estimation of one second time intervals

In this section, we illustrate the use of both APE and the model meta-selection procedure by application to a real-

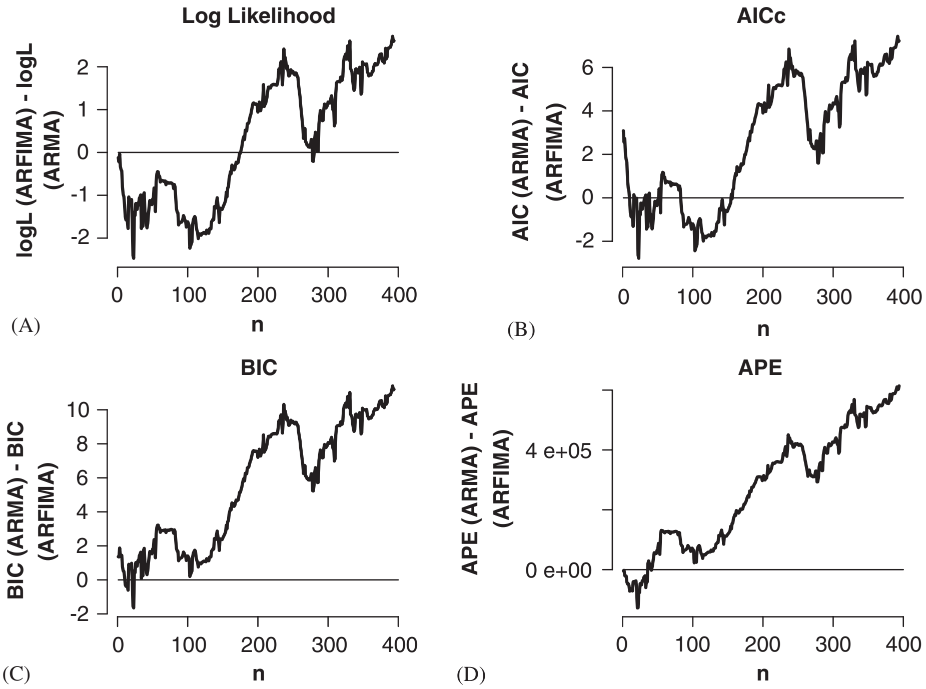

Fig. 6. Difference between ARFIMA model indices and ARMA model indices as a function of the number of observations. Panels A, log likelihood index; panel B, AICc index; panel C, BIC index; panel D, APE index.

world time series. The series under consideration is shown in Fig. 5, panel A, and concerns the repeated estimation of a one second time interval. This particular task is very popular in the research on long-range dependence in psychology (cf. Gilden, 2001; Wagenmakers, Ratcliff et al., 2004). The time estimation task is known to generate relatively sizeable serial correlations, and this greatly facilitates the process of discriminating between models with long-range dependence and models with short-range dependence (Thornton & Gilden, 2005).

A female student completed a practice block of 15 trials with feedback followed by an experimental block of 400 successive trials without feedback. Each trial started with the 1 s presentation of a fixation cross, followed by the presentation of the picture of an orange carrot.[11] The participant was instructed to press a response button one second after the appearance of the carrot stimulus. The stimulus remained visible until the participant had responded.

Prior to model fitting, two data purification techniques were applied. First, RTs exceeding the mean by more than three standard deviations were deemed outliers and were discarded. In the time series under study, one such outlier was identified and removed. Second, the models used here assume stationarity. Quadratic trends such as those resulting from practice effects and effects of fatigue may result in a non-stationary time series. Such time series will generally be described much better by the ARFIMA$(0, d, 0)$ model than by the ARMA(1,1) model (cf. Giraitis, Kokoszka, & Leipus, 2001). Thus, in order to avoid spurious detection of long-range dependence, the time series under study was quadratically detrended (cf. Van Orden et al., 2003; but see Thornton & Gilden, 2005).

Fig. 5, panel A, shows the detrended data.[12] Just as for the simulated data, both the ARMA(1,1) model and the ARFIMA$(0, d, 0)$ model were fit to a gradually increasing part of the time series. Fig. 5, panels B and C, show the one-step-ahead predictions from the ARMA(1,1) and ARFIMA$(0, d, 0)$ models, respectively (cf. Basak et al., 2001).

Fig. 6 shows the difference between log likelihood, AICc, BIC, and APE for the two models as a function of the number of observations in the data set. Fig. 6 clearly demonstrates that as the number of observations increases, so does support for the ARFIMA$(0, d, 0)$ model. This support is particularly strong for BIC and APE, echoing the results from the Monte Carlo simulations. Thus, from Fig. 6 it can be concluded that the most useful model for this time series is more likely to be ARFIMA$(0, d, 0)$ than it is to be ARMA(1,1).

The three panels in Fig. 7 show the result of a model meta-selection procedure. Panels A and B demonstrate that, for this particular time series, use of AICc for model selection results in relatively large one-step-ahead prediction errors, whereas BIC and APE perform about the same. Note that the horizontal stretches in Fig. 7 indicate that the difference in prediction error between two model selection methods does not change. This occurs when two model selection methods prefer the same model. Since after about 180 observations all selection methods prefer the ARFIMA$(0, d, 0)$ model over the ARMA(1,1) model, the x-axis of Fig. 7 only shows the results for the first 200 observations.

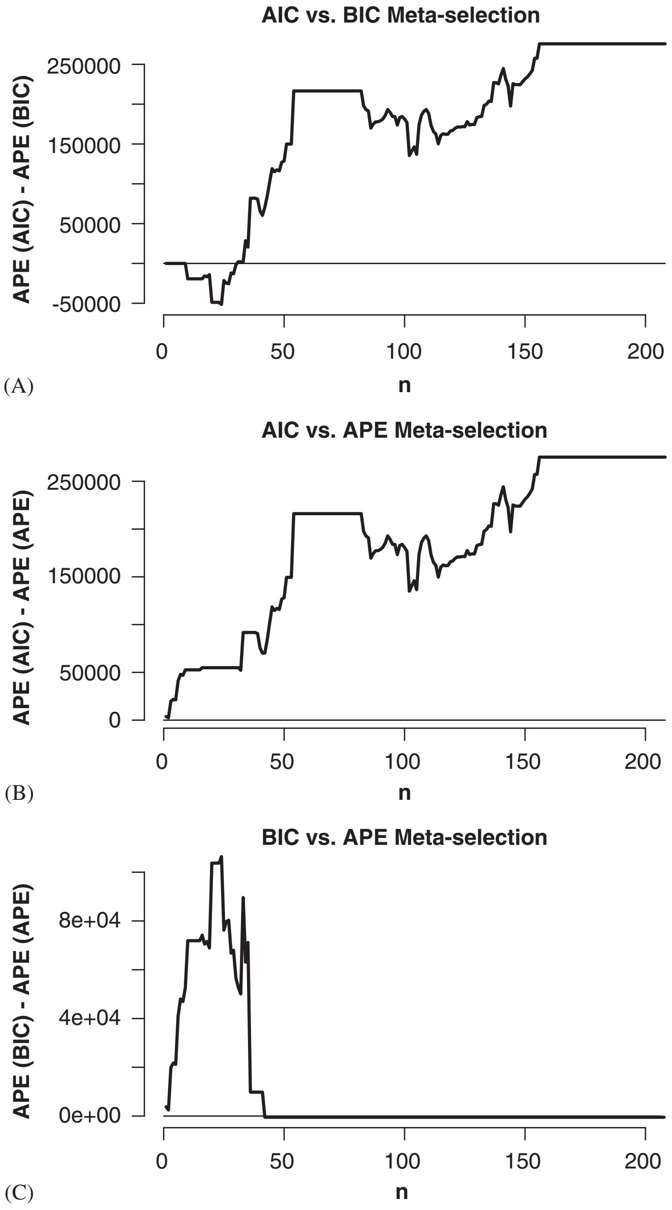

Fig. 7. Model meta-selection as a function of the number of observations. Each panel shows the difference in APE for various model selection methods. Panel A, $APE_{AICc} - APE_{BIC}$; panel B, $APE_{AICc} - APE_{APE}$; panel C, $APE_{BIC} - APE_{APE}$.

[11] Our reasons for using a carrot were twofold. First, a companion two-choice RT task involved classification of rabbits. Second, these two tasks were also administered to children, for whom rabbits and carrots are relatively interesting stimuli.

[12] The raw time series data are available on the first author's internet site, at http://users.fmg.uva.nl/ewagenmakers/2006/time.txt.

7. Concluding remarks

This article reviewed the rationale for using accumulative prediction error as a method of model selection. The APE method was applied to the problem of deciding whether a time series is long-range dependent or short-range dependent, or, more specifically, whether a time series is better described as ARFIMA$(0, d, 0)$ or as ARMA(1,1). As shown in Fig. 2 and Table 1, the APA is able to discriminate these models in an almost optimal fashion (i.e., the $\Delta APE = 0$ point is located very near the intersection of the two distributions, cf. Wagenmakers, Ratcliff et al., 2004) and automatically incorporates a penalty for model complexity.

The use of APE has several advantages. As was discussed in some detail earlier, the APE inherits a firm theoretical foundation from its relation to Bayesian model selection and minimum description length (Dawid, 1984; Rissanen, 1986b). Also, note that the APE interpretation of BMS/MDL provides a new perspective on what is achieved using BMS/MDL (i.e., minimization of accumulative one-step-ahead prediction errors).

One of the most important advantages of the APE method is surely the relative ease with which it is implemented. Further, the APE method can be applied to nested and non-nested models alike, and—in contrast to AIC and BIC—it is sensitive to the functional form of the model parameters (cf. Myung & Pitt, 1997), and not just to their number. The APE method is conceptually straightforward, as it accumulates 'honest' one-step-ahead prediction errors, that is, its predictions always concern unseen data. This distinguishes APE from cross-validation, which is otherwise very similar in spirit. Also, the APE is a data-driven method that does not rely on the accuracy of asymptotic approximations. In particular, use of the APE does not require one of the candidate models to be 'true' in the sense that it should correspond to the data generating process. Finally, the APE method can be used not only for the selection of models, but also for the selection of model selection methods. This method allows model selection methods to be compared for a single real-world time series, and may hence be of considerable practical importance.

Acknowledgment

This research was funded by a VENI Grant from the Dutch Organization for Scientific Research (NWO).

References

Aitchison, J., & Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge: Cambridge University Press.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, *73*, 5–59.

Baillie, R. T., Crato, N., & Ray, B. K. (2002). Long-memory forecasting [Special issue]. *International Journal of Forecasting*, *18*(2).

Bak, P. (1996). *How nature works: The science of self-organized criticality*. New York: Springer.

Bak, P., Tang, C., & Wiesenfeld, K. (1987). Self-organized criticality: An explanation of $1/f$ noise. *Physical Review Letters*, *59*, 381–384.

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, *44*, 2743–2760.

Basak, G. K., Chan, N. H., & Palma, W. (2001). The approximation of long-memory processes by an ARMA model. *Journal of Forecasting*, *20*, 367–389.

Beran, J. (1994). *Statistics for long-memory processes*. New York: Chapman & Hall.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.

Bhansali, R. J. (1999). Autoregressive model selection for multistep prediction. *Journal of Statistical Planning and Inference*, *78*, 295–305.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden Day.

Browne, M. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*, 108–132.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*, 171–189.

Chen, Y., Ding, M., & Kelso, J. A. S. (1997). Long memory processes ($1/f^{\alpha}$ type) in human coordination. *Physical Review Letters*, *79*, 4501–4504.

Chen, Y., Ding, M., & Kelso, J. A. S. (2001). Origin of timing errors in human sensorimotor coordination. *Journal of Motor Behavior*, *33*, 3–8.

Clarke, B. (2001). Combining model selection procedures for online prediction. *Sankhya A*, *63*, 229–249.

Crato, N., & Ray, B. K. (1996). Model selection and forecasting for long-range dependent processes. *Journal of Forecasting*, *15*, 107–125.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, *147*, 278–292.

Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference. *Journal of the Royal Statistical Society B*, *53*, 79–109.

Dawid, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 109–121). Oxford: Oxford University Press.

Dawid, A. P., & Vovk, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli*, *5*, 125–162.

Delignières, D., Fortes, M., & Ninot, G. (2004). The fractal dynamics of self-esteem and physical self. *Nonlinear Dynamics in Psychology and Life Sciences*, *8*, 479–510.

De Luna, X., & Skouras, K. (2003). Choosing a model selection strategy. *Scandinavian Journal of Statistics*, *30*, 113–128.

de Rooij, S., & Grünwald, P. (2006). An empirical study of minimum description length model selection with infinite parametric complexity. *Journal of Mathematical Psychology*, 50(2), 180–192.

Ding, M., Chen, Y., & Kelso, J. A. S. (2002). Statistical analysis of timing errors. *Brain and Cognition*, *48*, 98–106.

Doornik, J. A. (Ed.). (2001). *Ox: An object-oriented matrix language*. London: Timberlake Consultants Press.

Doornik, J. A., & Ooms, M. (2003). Computational aspects of maximum likelihood estimation of autoregressive fractionally integrated moving average models. *Computational Statistics & Data Analysis*, *42*, 333–348.

Doukhan, P., Oppenheim, G., & Taqqu, M. S. (Eds.). (2003). *Theory and applications of long-range dependence*. New York: Springer.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, *44*, 205–231.

Gammerman, A., & Vovk, V. (Eds.). (1999). Kolmogorov complexity [Special issue]. *The Computer Journal*, *42*(4).

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gerencsér, L. (1994). On Rissanen's predictive stochastic complexity for stationary ARMA processes. *Journal of Statistical Planning and Inference*, *41*, 303–325.

Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science*, *8*, 296–301.

Gilden, D. L. (2001). Cognitive emissions of $1/f$ noise. *Psychological Review*, *108*, 33–56.

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). $1/f$ noise in human cognition. *Science*, *267*, 1837–1839.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.) (1996). *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.

Giraitis, L., Kokoszka, P., & Leipus, R. (2001). Testing for long memory in the presence of a general trend. *Journal of Applied Probability*, *38*, 1033–1054.

Gisiger, T. (2001). Scale invariance in biology: Coincidence or footprint of a universal mechanism? *Biological Reviews of the Cambridge Philosophical Society*, *76*, 161–209.

Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 2* (pp. 249–269). New York: Elsevier.

Gottschalk, A., Bauer, M. S., & Whybrow, P. C. (1995). Evidence of chaotic mood variation in bipolar disorder. *Archives of General Psychiatry*, *52*, 947–959.

Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-range time series models and fractional differencing. *Journal of Time Series Analysis*, *1*, 15–30.

Granger, C. W. J., & Morris, M. J. (1976). Time series modelling and interpretation. *Journal of the Royal Statistical Society A*, *139*, 246–257.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*, 133–152.

Grünwald, P. (2005). MDL tutorial. In P. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

Grünwald, P., Myung, I. J., & Pitt, M. A. (Eds.). (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

Handel, P. H., & Chung, A. L. (Eds.). (1993). *Noise in physical systems and $1/f$ fluctuations*. New York: AIP Press.

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, *96*, 746–774.

Hemerly, E. M., & Davis, M. H. A. (1989). Strong consistency of the PLS criterion for order determination of autoregressive processes. *The Annals of Statistics*, *17*, 941–946.

Hjorth, U. (1982). Model selection and forward validation. *Scandinavian Journal of Statistics*, *9*, 95–105.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, *68*, 165–176.

Hosking, J. R. M. (1984). Modeling persistence in hydrological time series using fractional differencing. *Water Resources Research*, *20*, 1898–1908.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, *116*, 770–799.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*, 297–307.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 377–395.

Kontkanen, P., Myllymäki, P., & Tirri, H. (2001). Comparing prequential model selection criteria in supervised learning of mixture models. In T. Jaakkola, & T. Richardson (Eds.), *Proceedings of the eighth international workshop on artificial intelligence and statistics* (pp. 233–238). Los Altos, CA: Morgan Kaufmann Publishers.

Lawrance, A. J., & Kottegoda, N. T. (1977). Stochastic modelling of riverflow time series. *Journal of the Royal Statistical Society A*, *140*, 1–47.

Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications* (2nd ed.). New York: Springer.

Mandelbrot, B. B. (1977). *Fractals: Form, chance, and dimension*. San Francisco, CA: Freeman.

Modha, D. S., & Masry, E. (1998a). Memory-universal prediction of stationary random processes. *IEEE Transactions on Information Theory*, *44*, 117–133.

Modha, D. S., & Masry, E. (1998b). Prequential and cross-validated regression estimation. *Machine Learning*, *33*, 5–39.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Novikov, E., Novikov, A., Shannahoff-Khalsa, D., Schwartz, B., & Wright, J. (1997). Scale-similar activity in the brain. *Physical Review E*, *56*, R2387–R2389.

Pagano, M. (1974). Estimation of models of autoregressive signal plus white noise. *Annals of Statistics*, *2*, 99–108.

Peterson, B. S., & Leckman, J. F. (1998). The temporal dynamics of tics in Gilles de la Tourette syndrome. *Biological Psychiatry*, *44*, 1337–1348.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Pressing, J., & Jolley-Rogers, G. (1997). Spectral properties of human cognition and skill. *Biological Cybernetics*, *76*, 339–347.

Priestley, M. B. (1981). *Spectral analysis and time series*. London: Academic Press.

Qian, G., Gabor, G., & Gupta, R. P. (1996). Generalised linear model selection by the predictive least quasi-deviance criterion. *Biometrika*, *83*, 41–54.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Cambridge, MA: Blackwells.

Raftery, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 163–187). Boca Raton, FL: Chapman & Hall/CRC.

Rangarajan, G., & Ding, M. (Eds.). (2003). *Processes with long-range correlations: Theory and applications*. New York: Springer.

Rissanen, J. (1986a). A predictive least-squares principle. *IMA Journal of Mathematical Control and Information*, *3*, 211–222.

Rissanen, J. (1986b). Stochastic complexity and modeling. *The Annals of Statistics*, *14*, 1080–1100.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society B*, *49*, 223–239.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Teaneck, NJ: World Scientific Publishers.

Rissanen, J. (1992). Discussion of "prequential analysis, stochastic complexity and Bayesian inference" by A. P. Dawid. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 121–122). Oxford: Oxford University Press.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*, 40–47.

Rissanen, J. (1999). Hypothesis selection and testing by the MDL principle. *The Computer Journal*, *42*, 260–269.

Rissanen, J. (2001). Strong optimality of the normalized ml models as universal codes and information in data. *IEEE Transactions on Information Theory*, *47*, 1712–1717.

Rissanen, J. (2003). Complexity of simple nonlogarithmic loss functions. *IEEE Transactions on Information Theory*, *49*, 476–484.

Rissanen, J., Speed, T., & Yu, B. (1992). Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, *38*, 315–323.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 286–292.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Skouras, K., & Dawid, A. P. (1998). On efficient point prediction systems. *Journal of the Royal Statistical Society B*, *60*, 765–780.

Sornette, D. (2000). *Critical phenomena in natural sciences*. Berlin: Springer.

Sowell, F. B. (1992a). Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics*, *53*, 165–188.

Sowell, F. B. (1992b). Modeling long run behavior with the fractional arima model. *Journal of Monetary Economics*, *29*, 277–302.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, *64*, 583–639.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B*, *36*, 111–147.

Stone, M. (1977a). Asymptotics for and against cross-validation. *Biometrika*, *64*, 29–35.

Stone, M. (1977b). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B*, *39*, 44–47.

Thornton, T. L., & Gilden, D. L. (2005). Provenance of correlations in psychological data. *Psychonomic Bulletin & Review*, *12*, 409–441.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, *132*, 331–350.

Voss, R. F., & Clarke, J. (1975). '$1/f$' noise in music and speech. *Nature*, *258*, 317–318.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition. *Psychonomic Bulletin & Review*, *11*, 579–615.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2005). Human cognition and a pile of sand: A discussion on serial correlations and self-organized criticality. *Journal of Experimental Psychology: General*, *134*, 108–116.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50.

Wallace, C. S., & Boulton, D. M. (1968). An information measure for classification. *The Computer Journal*, *11*, 185–194.

Wallace, C. S., & Freeman, P. R. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society B*, *49*, 240–265.

Wei, C. Z. (1992). On predictive least squares principles. *The Annals of Statistics*, *20*, 1–42.

Wolf, D. (1978). *Noise in physical systems*. New York: Springer.

Yoshinaga, H., Miyazima, S., & Mitake, S. (2000). Fluctuation of biological rhythm in finger tapping. *Physica A*, *280*, 582–586.

Yulmetyev, R. M., Emelyanova, N., Hänggi, P., Gafarov, F., & Prokhorov, A. (2002). Long-range memory and non-Markov statistical effects in human sensorimotor coordination. *Physica A*, *316*, 671–687.

Zhang, P. (1993). Model selection via multifold cross-validation. *Annals of Statistics*, *21*, 299–313.