Extraction of the relevant delays for temporal modeling

Goutte, Cyril

Published in:
I E E E Transactions on Signal Processing

Link to article, DOI:
10.1109/78.845935

Publication date:
2000

Document Version
Publisher's PDF, also known as Version of record

Extraction of the Relevant Delays for Temporal Modeling

Cyril Goutte

***Abstract*—When modeling temporal processes, just like in pattern recognition, selecting the optimal number of inputs is a central concern. In this paper, we take advantage of specific features of temporal modeling to propose a novel method for extracting the inputs that attempts to yield the best predictive performance. The method relies on the use of estimators of generalization error to assess the predictive performance of the model. This technique is first applied to time series processing, where we perform a number of experiments on synthetic data, as well as a real life dataset, and compare the results to a benchmark physical method. Finally, the method is extended to system identification and illustrated by the estimation of a linear FIR filter on functional magnetic resonance imaging (fMRI) signals.**

***Index Terms*—Delay estimation, functional magnetic resonance imaging, generalization error, identification, modeling, time series.**

I. Overview

In this paper, we will be concerned with modeling the future behavior of a system based on its past plus, possibly, some exogenous control signal with application to time series prediction and system identification. Predictive performance depends, among other things, on the design of a proper input or regression vector. With too few inputs, the model does not have sufficient information and is unable to grasp the inner workings of the system, resulting in a large approximation error. On the other hand, a model with irrelevant inputs is overparameterized, which usually results in poor predictive performance, as suggested by the curse of dimensionality [1], [2].

We will focus mainly on time series prediction and later address the problem of system identification. In this context, potential inputs are past values, or delays, of the time series. Our aim is to select delays that are necessary to model the system while discarding unnecessary delays that could harm the overall performance. To our knowledge, the only provably optimal, general input selection method is exhaustive search, which is NP-complete and computationally unfeasible unless the number of inputs is very limited. Our method is related to iterative feature selection techniques used in traditional statistics [3]. It builds on the specificities of temporal processing to provide an original way of selecting potential inputs. The relevance of a candidate delay is assessed directly by its effect on predictive performance that is measured using estimators of generalization error.

The following sections are organized as follows. First, we give a general presentation of feature selection applied to time series modeling from the statistical and physical points of view (Section II). We suggest the use of estimators of generalization error to evaluate the quality of a subset of features. Our extraction of the relevant delays (ERD) method is described in Section III as a principled alternative. The second part of the paper contains a number of experiments conducted on three different datasets. Time series predictions (Section IV) are addressed using the well-known artificial Hénon map and a real-world time series measuring the mean monthly flow of the Fraser river; ERD is then applied to system identification (Section V) using a functional magnetic resonance imaging (fMRI) dataset. We conclude with a discussion of the method and results.

II. Feature Selection

Let us consider a standard time-series modeling problem; a sequence of measurements $y(t)$, $1 \leq t \leq T$ is collected. We wish to predict $y(t)$ from a set of past values $y(t-d)$, $d > 0$.[1] Extracting the relevant delays consists of finding the set of m delays $(d_1, d_2, \cdots d_m)$ such that the input vector

$$\boldsymbol{x}(t) = [y(t-d_1)y(t-d_2) \cdots y(t-d_m)] \tag{1}$$

yields the best prediction of $y(t)$.

A. Physical Approach

The physical approach relies on estimating the *embedding dimension* of the time series [4]. This is essentially equivalent to finding the set of *primary* delays, i.e., delays with an explicit influence on the observed values. For example, for a time series generated by $x(t) = g(x(t-1))$, 1 is the only primary delay. As $x(t) = g(x(t-2))$, we can also estimate $x(t)$ using $x(t-2)$; however, 2 is a secondary (not a primary) delay.

Several methods have been proposed to estimate the embedding dimension and the embedding space of a time series. Pi and Peterson [5] have introduced the "δ-test" in the neural networks literature. He and Asada [6] proposed the use of "Lipschitz quotients" to identify the order of nonlinear input–output systems. An independent but essentially similar method was applied to time series in the signal processing literature [7]. These methods all rely on the assumption that the underlying mapping g is continuous and reasonably smooth. The existence of a

Manuscript received August 5, 1998; revised November 23, 1999. This work was supported by a Research Fellowship from the Technical University of Denmark and by the Human Brain Project P20 MH57180. The associate editor coordinating the review of this paper and approving it for publication was Dr. Shubha Kadambe.

The author is with the Department of Mathematical Modeling, Technical University of Denmark, Lyngby, Denmark (e-mail: cg@imm.dtu.dk).

Publisher Item Identifier S 1053-587X(00)04065-4.

[1]Note that we do not address the problem of obtaining the best sampling frequency, i.e., length of the basic time delay (difference between $t+1$ and t).

continuous mapping between $\boldsymbol{x}(t)$ and $y(t)$ means that close inputs $\boldsymbol{x}(u)$ and $\boldsymbol{x}(v)$ are mapped to close outputs $y(u)$ and $y(v)$. On the other hand, for insufficient input spaces (i.e., missing delays), close inputs can correspond to arbitrarily distant outputs (see Section IV for a practical example). Quantifying this closeness is done either by estimating the probability that two outputs are close given close inputs (δ-test) or by calculating the ratio between output and input distances (Lipschitz quotients). An important side effect is that all input–output pairs have to be considered, requiring extensive calculation.

Note that the techniques mentioned above are nonparametric, rely on the data alone, and need not specify a given model. This can turn out to be disadvantageous since for a given data set, only one set of relevant delays is selected, regardless of the ability of the model to actually implement the underlying mapping using them. Flexible models such as neural networks should overcome this limitation by their uniform approximation capabilities [8]. However, we will see that in practice, a given model often benefits from the inclusion of *secondary* delays (cf. Section IV).

B. Statistical Approach

From a statistical point of view, the extraction of the relevant delays is a special case of feature selection, which is itself a part of the more general problem of analyzing the structure in the data [3]. The statistical approach relies on specifying a (parametric or nonparametric) model f with which we try to estimate the input–output mapping $\hat{y}(t) = f(\boldsymbol{x}(t))$. In conventional feature selection, an important assumption is the availability of all necessary variables. Provided that our data are sampled correctly, this assumption is usually satisfied in the case of time series.[2] In the following, *delays* are positive integers $d_j > 0, 1 \leq j \leq m$, *variables* or *features* are past values of the time series $y(t-d_j)$, and *inputs* are the vectors $\boldsymbol{x}(t)$ containing these past values. Finding the relevant delays is then equivalent to finding the input that gives optimal predictive performance. Conventional feature selection relies on three different components:

1) a selection method searching through possible subsets of variables;
2) an evaluation criterion assessing the quality of each subset;
3) a stop condition, which decides whether a satisfactory subset was obtained.

To our knowledge, the only general optimal method for selecting the best features among F is to perform an exhaustive search through all $2^F - 1$ possible subsets of variables. This optimal approach becomes unfeasible for moderate values of F. Furthermore, in temporal modeling, the maximum delay and, thus, the total number F of variables, is not known beforehand, but we would typically accept to probe quite far into the past. For monotonous evaluation criteria, the branch and bound algorithm [9] provides an efficient alternative. Unfortunately, the predictive ability is not a monotonous criterion. Common suboptimal alternatives perform an iterative search by regularly increasing or decreasing the number of selected features [3].

Forward Selection: Starting from an empty set (no variables), add variables according to the evaluation criterion until the stop condition is reached.

Backward Elimination: Starting with a full set (all possible variables), delete one variable at a time according to the evaluation criterion until the stop condition is reached.

Stepwise Regression: Alternate between both methods, by, e.g., performing a backward elimination after each inclusion or choosing between adding or deleting variables according to the evolution of the evaluation criterion.

Note that the focus in neural networks research, for example, is almost entirely on backward elimination through various pruning schemes [10]. All these suboptimal methods rely heavily on the evaluation criterion. Typical choices include, for example, the F statistic [3], extensions to nonlinear models, or mutual information [11]. We will now present an alternative to these choices.

C. Generalization Approach

In the context of nonparametric modeling, our goal is to obtain the best prediction. We will thus use a measure of the predictive abilities as our evaluation criterion. For a model f mapping an input vector $\boldsymbol{x}(t) = [y(t-d_1) \cdots y(t-d_m)]$ to output $y(t)$, the risk $\ell(y,\ f(\boldsymbol{x}))$ measures the cost associated with estimating the output y observed in $\boldsymbol{x}$ with the model $f(\boldsymbol{x})$. Assuming Gaussian noise on the output usually leads to the choice of the quadratic risk, $\ell(y,\ f(\boldsymbol{x})) = (y - f(\boldsymbol{x}))^2$. The *generalization error* (or expected risk) associated with model f is defined as the expectation of the risk over the unknown, but fixed, joint input–output distribution

$$G(f) = \int (y - f(\boldsymbol{x}))^2 \, p(\boldsymbol{x},\, y)\, d\boldsymbol{y}\, d\boldsymbol{x}. \tag{2}$$

This is also known as the prediction error or the integrated squared error. Ideally, the evaluation criterion for a given subset should be the generalization error of the model using this subset as input [12]. However, (2) cannot be used directly because the joint input–output probability is unknown. It will be estimated from the available data $(\boldsymbol{x}(t),\, y(t))$, with $t \in \mathcal{D} = \{T_{\min}, \cdots, T_{\max}\}$.

The *split-sample* (SS) estimator is obtained by replacing the joint distribution $p(\boldsymbol{x},\, y)$ by its empirical estimator on the validation set $\mathcal{V} \subset \mathcal{D}$

$$\hat{G}_{\mathrm{SS}} = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \left(y(t) - f^{-\mathcal{V}}(\boldsymbol{x}(t))\right)^2 \tag{3}$$

where $|\mathcal{V}|$ is the cardinality of $\mathcal{V}$, and $f^{-\mathcal{V}}$ is the model estimated on the training set $\mathcal{T} = \mathcal{D} \backslash \mathcal{V}$.[3] Note that $\mathcal{T}$ and $\mathcal{V}$ must be disjoint in order for $\hat{G}_{\mathrm{SS}}$ to be an unbiased estimate of $G(f^{-\mathcal{V}})$.

The *cross-validation*[4] (CV) estimator [13], [14] resamples the validation and training sets from the available data in order

[2]It breaks down when a long-term delay is needed, ranging further in the past than the data itself. However, the relevance of such long-term delays is questionable, as there would be no data to identify the associated parameter(s).

[3]Where $\mathcal{T} = \mathcal{D} \backslash \mathcal{V}$ means $\mathcal{T} = \{t \in \mathcal{D},\, t \notin \mathcal{V}\}$.

[4]The split-sample method is sometimes referred to as "cross-validation" in the neural networks literature. This is inconsistent with the definition of [13], and we will here reserve the term exclusively for the averaging method (4).

to increase the reliability of the resulting estimator. In L-fold CV, $\mathcal{D}$ is split into L disjoint subsets $(\mathcal{S}_j)_{j=1,\dots,L}$ of roughly equivalent size $\bigcup_{j=1}^{L} \mathcal{S}_j = \mathcal{D}$. The split-sample estimators (3) calculated using each $\mathcal{S}_j$ in turn as a validation set are averaged over the subsets

$$\hat{G}_{\mathrm{CV}} = \frac{1}{N} \sum_{j=1}^{L} \sum_{t \in \mathcal{S}_j} \left(y(t) - f^{-\mathcal{S}_j}(\boldsymbol{x}(t))\right)^2 = \frac{1}{N} \sum_{t \in \mathcal{D}} \left(y(t) - f^{-t}(\boldsymbol{x}(t))\right)^2 \quad (4)$$

where for notational convenience, we introduce f^{-t}, which is the model estimated excluding the subset $\mathcal{S}_j$ containing t. This means that $\forall j,\ \forall (u, v) \in (\mathcal{S}_j)^2$, $f^{-u} = f^{-v}$. Note that (3) does not estimate the generalization error (2), but it does estimate its average over all possible datasets of the same size sampled from $p(\boldsymbol{x}, y)$.

Finally, a number of analytical asymptotic estimators of the average generalization error have been proposed in the literature, e.g., final prediction error (FPE) [15], generalized prediction error (GPE) [16], final prediction error for regularized problems (FPER) [17], or network information criterion (NIC) [18]. Without loss of generality,[5] we will settle for a GPE-like expression, i.e., an FPE with an *effective number of parameters* $\hat{P}$

$$\hat{G}_{\mathrm{FPE}} = \left(\frac{N + \hat{P}}{N - \hat{P}}\right) R(f) \quad (5)$$

where $R(f)$ is the training error (or empirical risk) on dataset $\mathcal{D}$: $R(f) = \sum_{t \in \mathcal{D}} (y(t) - f(\boldsymbol{x}(t)))^2$. For unregularized linear models, $\hat{P} = P$, which is the number of parameters. For regularized and/or nonlinear models, we generally have $\hat{P} < P$, and the exact expression for $\hat{P}$ depends on the estimator and the regularization method (see, e.g., [17] and [19]).

Estimators of generalization error, whether they are based on cross-validation or asymptotics, will now be used as evaluation criterion in order to derive an original delay extraction scheme.

III. Extraction of the Relevant Delays

The above presentation of statistical feature selection applies to general regression problems. Time series prediction, and, to some extent, system identification, have a number of distinct characteristics. On the one hand, all potential features are available, but there is no upper bound on the maximum delay. This makes the extraction of relevant delays a rather bad candidate for backward elimination schemes. On the other hand, the chronological order yields a natural ordering of variables that our method uses as a natural selection criterion. The rationale for this scheme is that primary delays will always be tested for inclusion before secondary delays. As a consequence, secondary delays will never be included unless 1) the model is unable to represent the underlying mapping using primary delays alone, or 2) the secondary delay is also a primary delay.[6]

[5]Other estimators lead to similar expressions, and their subsequent use is straightforward.

[6]If $y(t) = g(y(t-1),\ y(t-2))$, 2 is a primary delay but is also a secondary delay, through the primary delay 1.

The ERD method takes all delays in their natural order and adds a candidate variable if and only if it yields a *significant* decrease in generalization error. The algorithm can be described as follows.

1) Initialize: $d = 0$; no input selected; $G_{\min} = \sigma_y^2$ (time series variance).
2) Model: d = d + 1; add delay $t-d$ to selected inputs; Estimate generalization error $\hat{G}$ for resulting model.
3) Test: If $\hat{G}$ is *significantly* smaller than $G_{\min}$, then keep delay $t - d$ and cut $G_{\min} = \hat{G}$; else, discard delay $t - d$.
4) Iterate: Go to step 2 until stop condition is reached.

A. Optimality and Suboptimality

In the ideal case of a complete model (i.e., contains the target mapping) and sufficient amount of noise-free data (i.e., fairly larger than the number of parameters), the above method is optimal. To see this, consider the maximum primary delay d_{MAX}. Clearly, all models obtained for $d < d_{\mathrm{MAX}}$ have strictly positive generalization errors $\hat{G} > 0$. Due to the completeness of the model class, there is a model using delays up to d_{MAX}, which yields $\hat{G} = 0$, and due to the noise-free assumption, this model can be estimated. As models using $d > d_{\mathrm{MAX}}$ cannot yield any decrease in generalization error, no further delay will be selected.

For more general situations, we are not aware of any proof that the model is optimal. Indeed, the experiments presented below suggest that it is not, although results appear to be close to optimum. The correctness and near optimality of the extracted delays will depend on several factors, the approximation capabilities of the model and the noise level among them.

Note that the traditional caveat against sequential selection is its inability to handle variables that are combinations of other variables. This arises naturally in temporal modeling as each observation is a (possibly nonlinear) mapping of previous values. However, chronological selection ensures that each variable is tested for inclusion before the variables on which it depends, guaranteeing a parsimonious selection.

B. Significant Decrease in Error

Step 3 of the above algorithm requires that we assess the significance of an observed decrease in generalization error. This requirement avoids the inclusion of a delay leading to negligible decrease in estimated generalization error, which could happen by chance alone. We take advantage of the fact that the estimators presented in (3)–(5) are based on averaging over the data. The CV and asymptotic estimators can be put in the general expression

$$\hat{G} = \frac{1}{N} \sum_{t \in \mathcal{D}} \epsilon(t) \quad (6)$$

where $\epsilon(t) = (y(t) - f^{-t}(\boldsymbol{x}(t)))^2$ for CV and $\epsilon(t) = ((N + \hat{P})/(N - \hat{P}))(y(t) - f(\boldsymbol{x}(t)))^2$ for GPE [16].[7] For the split-sample estimator (3), $\epsilon(t) = (y(t) - f(\boldsymbol{x}(t)))^2$, and the average runs over $\mathcal{V}$ instead of $\mathcal{D}$.

[7]Other estimators like FPER or NIC will lead to similar forms with different expressions for the penalty term of $\hat{P}$.

We wish to test whether G_m, which is the generalization error for the model with m delays (current best in the algorithm), is significantly different from G_{m+1}, which is the generalization error for the model with $m+1$ delays (augmented with the candidate delay), using the residuals $\epsilon_m(t)$ and $\epsilon_{m+1}(t)$ and their average $\hat{G}_m$ and $\hat{G}_{m+1}$, which are unbiased estimators of G_m and G_{m+1}, respectively. Assuming that the difference between the residuals $\epsilon_m(t)$ and $\epsilon_{m+1}(t)$ is approximately normally distributed and using the fact that all residuals are calculated on the same data, we assess the significance of an observed decrease in generalization error by using a one-tailed paired t test [20] between the residuals. Note that by construction, the candidate model has an additional delay, meaning that fewer input–output pairs $(\boldsymbol{x}(t), y(t))$ can be formed. In order to apply the paired t-test, we will discard the additional data points used by the smaller model. This waste of a couple of examples is counterbalanced by the superior power of the paired t-test, compared with the nonpaired version.

The choice of the t-test is typical for assessing the significance of a difference in means when we are willing to make a normal assumption about the individual differences $(\epsilon_m(t) - \epsilon_{m+1}(t))$. An efficient nonparametric alternative is the Wilcoxon matched-pairs signed-ranks test [21], which has an asymptotic relative efficiency (compared with the t-test) of 0.95 in the normal case. This means that the price for relaxing some of the parametric assumptions in the t-test is that it needs asymptotically only 5% more data to assess a given difference with the same significance.

As in standard iterative subset selection, the significance level α has an influence on the result. When $\alpha \to 0$, no delays are selected. However, contrary to the standard case, $\alpha \to 1$ does not necessarily select all delays. This is because the estimators of generalization take the increase in model complexity into account and do not always decrease for larger models (this is also observed with, e.g., Mallow's C_p [22]). Note that it has been reported that the traditional choice of $\alpha = 0.05$ (the 5% significance level) tends to be overly conservative. In agreement with standard practice in subset selection [23], we will choose $0.15 < \alpha < 0.25$, a range in which experimental results seem to be stable. Finally, an additional level of generalization estimation can be invoked to tune the value of α (e.g., CV in [24]), but the overall process becomes cumbersome and has not been pursued here.

C. Stop Condition

The stop condition is motivated by practical or by problem-specific considerations. As mentioned above, models with larger delays in the input will have less input–output pairs available; for a sequence of T measurements and inputs with maximum delay d_m (1), at most $(T - d_m)$ training examples can be formed. This means that with increasing delays, 1) the estimation becomes more difficult (less data for more parameters), and 2) the degrees of freedom in the statistical test decreases, such that the difference in estimated generalization error must be larger in order to become significant. The available data provides a natural upper bound for the maximum possible delay; the selection is stopped when the data becomes insufficient for proper estimation. Most parametric models, such as linear models or neural networks, require at least as many examples as parameters. Statistical rules-of-thumb suggest a ratio of at least 10 examples per parameter. For the Hénon map example below (Section IV-A) and a linear model, the maximum possible delay becomes 500 (data points) minus six (parameters in the candidate model), i.e., $d_{\max} = 494$ (and $d_{\max} = 440$ if one requires 10 data per parameter).

On some problems, and typically when modeling a physical system, it may make sense to introduce stronger constraints to reflect additional knowledge on the phenomenon or requirements on the model. In the fMRI experiments presented below (Section V), the resulting filter tentatively models the haemodynamic response to neuronal activation. As the experiment is performed as a series of consecutive runs and we are interested in the response to a typical activation pattern, it is sensible to limit the delay extraction to one run, i.e., 48 delays in that case.

Because the selection criterion is independent of the stop criterion, the influence of the latter is in how conservative the resulting model will be. Any criterion that stops the algorithm earlier will necessarily yield a model with fewer (or as many) delays selected, i.e., it will be more conservative. It is, thus, fair to say that the maximum delay gives the least conservative model (everything being otherwise equal). Adding stronger (e.g., physical) constraints on the maximum delay will potentially trade a more conservative model for an increase in interpretability. Note finally that due to the decreasing degrees of freedom when less data are available, the last included delay is usually (in the examples we have processed) much smaller than $d_{\max}$.

IV. Time Series Prediction

A. Hénon Map

Let us first consider a time series generated by the well-known Hénon map [25]

$$y(t) = 1 - 1.4(y(t-1))^2 + 0.3y(t-2). \tag{7}$$

We apply several methods on a dataset containing 500 points. An independent set of 10 000 elements is sampled from (7) to assess the resulting generalization abilities. The noisy map, with additive Gaussian noise ($\sigma^2 = 0.1$) is also investigated. We consider two models:

- linear model (obviously a bad choice to model the nonlinear Hénon map) using FPE [15] as an estimator of generalization error;
- kernel smoother [26], [27] using the *leave-one-out* (i.e., N-fold) CV estimator;

and four delay selection methods

1) the δ-test [5], estimating the embedding dimension;
2) the ERD method (Section III);
3) a "divine guidance" or large validation set (LVS) method;
4) the traditional F-inclusion scheme.

The "divine guidance" method selects delays on the basis of a large independent dataset providing a reliable estimator of generalization. This method is, of course, impractical on real data and is used in order to check the behavior of the estimators used in ERD.

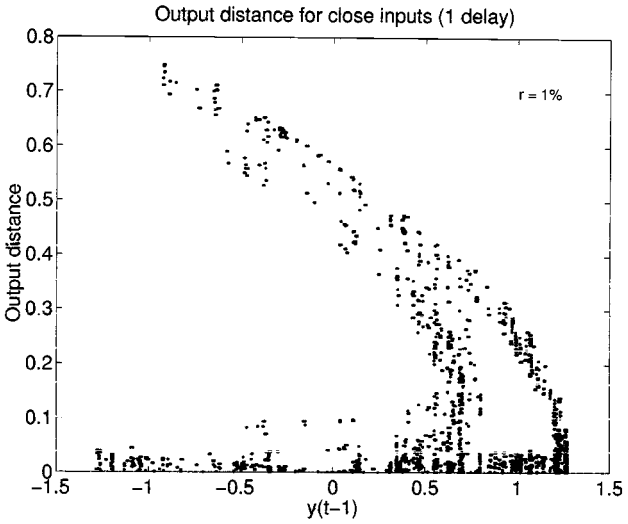

Fig. 1. Output distance versus input when input contains the first delay. Each point represents a pair of data, and only points with the 1% smallest input distances have been included. Many close inputs lead to distant outputs, indicating an insufficient input space.

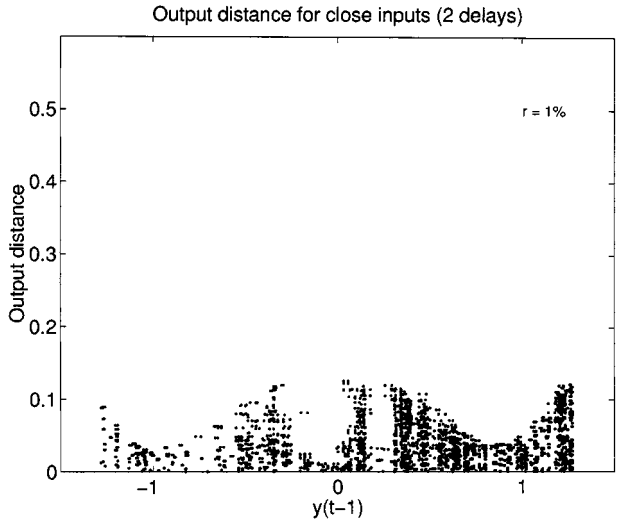

Fig. 2. Output distance versus input when input contains the first two delay. Each point represents a pair of data, and only the points with the 1% smallest input distances have been included. The output distance is always small, indicating a sufficient input space.

In F-inclusion, a candidate delay is included iff

$$F = \left(\frac{R_m(f) - R_{m+1}(f)}{N R_{m+1}(f)/(N-m-1)} \right) > F_\alpha \qquad (8)$$

where $R_m(f)$ and $R_{m+1}(f)$ are the empirical risks, or training errors, for the models with m delays and $m+1$ delays (respectively), calculated on the same data, and F_α is the F-distribution threshold for an α confidence level 1 and $(N-m-1)$ degrees of freedom. A candidate delay is therefore included when it yields a significant improvement in *observed* performance.

Figs. 1 and 2 show a plot proposed by Aleksic [28] to investigate the embedding dimension. All points correspond to a pair of data (t_1, t_2) with $\|\boldsymbol{x}(t_1) - \boldsymbol{x}(t_2)\|$ small and give the output distance $\|y(t_1) - y(t_2)\|$ against $y(t_1 - 1)$. Clearly, with only one delay in the input (Fig. 1), close inputs do not guarantee close outputs. Therefore, there is no continuous mapping from $\boldsymbol{x}(t) = y(t-1)$ to $y(t)$. On the other hand, with the first two delays (Fig. 2), all pairs with close inputs also have close outputs. This suggests that a continuous mapping between the first two delays and the time series value can be implemented. Accordingly, the δ-test selects the first two delays from the dataset [5]. For noise-free data and the kernel smoother model, all other selection methods also select these two delays and are able to implement the Hénon map perfectly from the 500 available observations. In the remaining situations, they always select at least one additional delay, depending on the model-noise combination (see Table I).

TABLE I
DELAYS SELECTED ON THE HÉNON MAP DATASET BY FOUR METHODS: LARGE VALIDATION SET (LVS), EXTRACTION OF THE RELEVANT DELAYS (ERD), F-INCLUSION, AND δ-TEST

Hénon map:	No noise		Additive noise	
	Linear	Kernel	Linear	Kernel
LVS	1–7	1–2	1–7	1–3
ERD	1,3–6	1–2	1,3,4	1–3
F-inclusion	1–6	1–2	1–6,10	1–8,11–13, 16,17,19,20
δ-test	1–2		1–2	

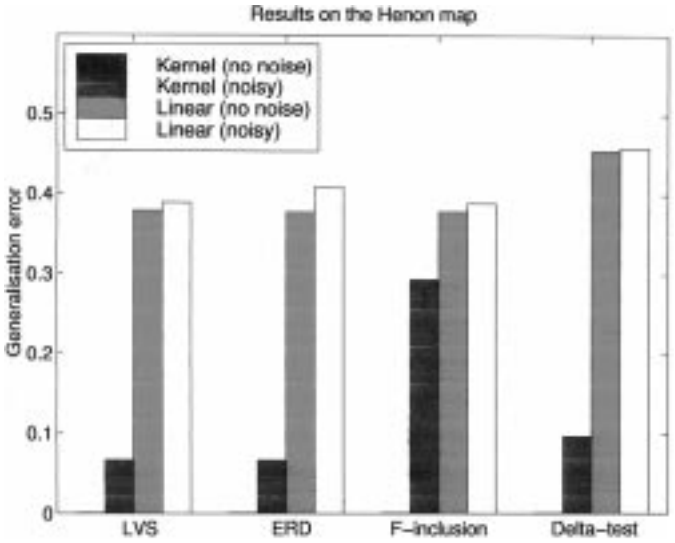

Fig. 3. Results on the Hénon map for both models and both noise conditions with four selection schemes: Large validation set (LVS), extraction of the relevant delays (ERD), F-inclusion, and δ-test.

All resulting models are tested on the large noise-free test set in order to check their generalization abilities (Fig. 3). Predictably, the kernel smoother provides much better performance than the linear model. Note that the δ-test outperforms a statistical method (F-inclusion) only once: for the "kernel + noisy data" combination. This is due to its emphasis on extracting the primary delays; as efficient as the δ-test may be at extracting the "true" delays, it suffers from not addressing the relevant goal in our context, namely, prediction of future time series values. The relative edge of the δ-test for that particular model-noise combination is rather due to the inferior performance of F-inclusion.

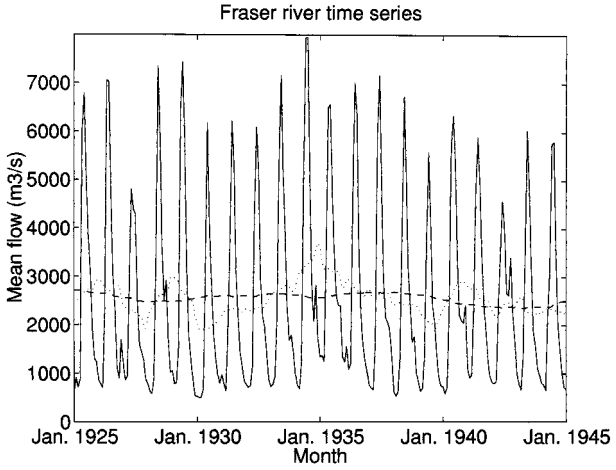

Fig. 4. Mean monthly flow of the Fraser river recorded in Hope, B.C., Canada, between January 1925 and January 1945 (solid line). The one- and ten-year moving averages are shown as dotted and dashed lines, respectively.

This suggests that F-inclusion is ill-suited to flexible nonlinear models. This is because (8) does not penalize overfitting; on the contrary, when $R(f)$ becomes overly small, the F statistic will be artificially increased, leading to an exceedingly optimistic selection (15 delays, instead of two or three). Notice the close agreement between LVS and ERD, suggesting that the estimators of generalization error (FPE and CV) perform well. ERD tends to be more parsimonious, yielding five and three delays for the linear model, instead of seven and seven, for similar performance. Finally, on noisy data, even with the flexible kernel smoother, ERD and LVS select an additional, secondary delay. This is because we try to learn the underlying relationship from a limited amount of data. This additional delay, although not theoretically necessary, yields a 30% improvement in performance.

B. Fraser River

Let us now turn our attention to a real dataset, which is analyzed in [29] and publicly available. The data records mean monthly measurements of the flow of the Fraser river in Hope, B.C., Canada, from March 1913 to December 1990. It contains 946 values with a rough periodicity and maxima every 11–13 months (see Fig. 4). The dataset is split in half: The first 473 observations are used for model estimation and delay extraction, and the last 473 are used for testing the generalization abilities of the resulting model. In the following experiments, the data have been log-transformed and centered. The selection method using the large validation set is, of course, not sensible in this context, as no extra data is available. Accordingly, we will consider only two selection schemes (the δ-test and the proposed ERD) applied to three models:

- linear model;
- kernel smoother;
- nonlinear neural network model (multilayer perceptron with one hidden layer).

The parameters of the linear and neural network models are estimated by minimizing the mean squared error on the transformed data. The estimators of generalization error are the FPE [15] and GPE [16], respectively, for these models and the leave-one-out estimator, as above, for the kernel smoother. The nonparametric

TABLE II
DELAYS EXTRACTED, FOR EACH MODEL, BY BOTH METHODS: THE PROPOSED EXTRACTION OF THE RELEVANT DELAYS (ERD) AND THE δ-TEST

Fraser river	Linear	Kernel	Neural net
ERD	1,2,4–7,10,11, 23,26,35,48	1,2,4,7,11,13	1,2,4,7,11,23
δ-test	1,2,4,7,8,11		

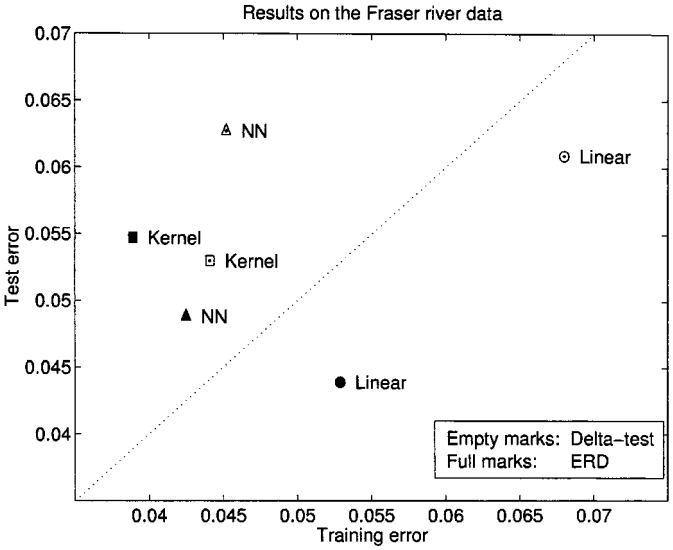

Fig. 5. Result for three different models of both extraction schemes: the proposed extraction of the relevant delays (ERD, full marks) and the δ-test (empty marks). The three models are linear, kernel smoother, and nonlinear neural network (NN).

δ-test selects six delays, whereas the ERD extracts six to 12 delays, depending on the model (see Table II). Not surprisingly, the more flexible models, i.e., kernel smoother and neural network, use fewer delays, whereas the linear model uses twice as many.

Notice that although the δ-test and ERD extract the same number of delays for two out of three models, these delays do not coincide. Typically, ERD probes further into the past: across two periods for the neural network and even four for the linear model. This is an interesting artifact of the δ-test method; as the size of the input space grows larger, fewer and fewer pairs of points have close inputs. As a result, the variance of the estimated probabilities increases, and the test is unable to reliably select additional delays.

The generalization abilities of the resulting models are assessed using the test set. On Fig. 5, points far above the dotted line indicate probable overfitting. Not surprisingly, the kernel smoother overfits the data quite severely. This model is known to suffer acutely from the curse of dimensionality and has difficulties handling high-dimensional inputs when only one smoothing parameter is used [30]. Surprisingly, the best overall performance is achieved by the combination of ERD and linear model. The flexible neural network model does marginally worse and seems to overfit slightly. Note that with a good training algorithm, the neural network should be able to outperform the linear model *using the same inputs*. However, the ERD scheme has

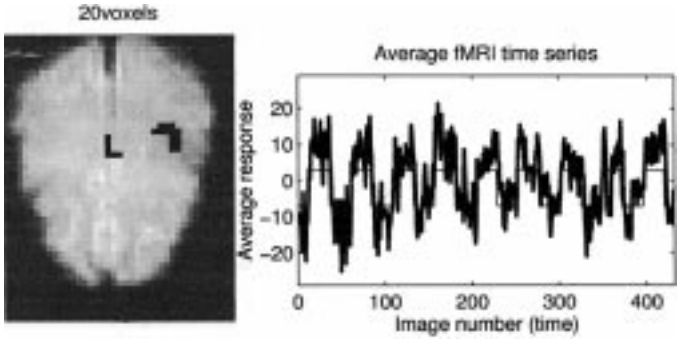

Fig. 6. Right: Brain map indicating the 21 activated voxels (black) analyzed with ERD. The background is an anatomical reference obtained from the average activation. Left: average time series for the activated voxels (thick line) and reference stimulus (thin black).

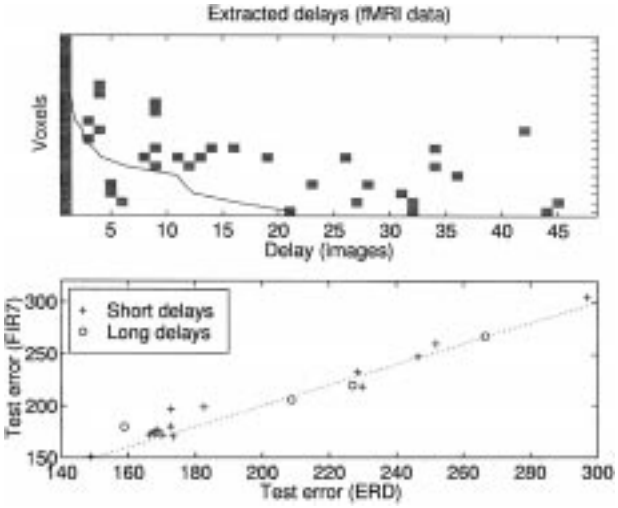

Fig. 7. Top: Delays extracted for each voxel (grey) and effective filter delay (solid black). Bottom: Test error for the seven-parameter FIR filter versus the ERD filter. Crosses indicate voxels for which the effective delay is less than 10 seconds; circles indicate voxels with longer delays (more than 15 s). One voxel with very large test error in both cases has been discarded to improve clarity.

predictably extracted different sets of delays for both models. In particular, the linear model uses three additional long-term delays (26, 35, and 48).

Again, these results illustrate the limitations of the physical approach when prediction performance is the target. The statistical approach based on model performance gathers more information in the past and yields an improvement in performance that is, here, most impressive for the linear model.

V. System Identification

The ERD scheme can be extended to some simple system identification problems, such as the identification of linear or nonlinear finite input response (FIR) filters [31]. The data is acquired during a functional magnetic resonance imaging (fMRI) experiment. While the subject lies in the scanner, he is asked to either lie motionless (rest) or perform a simple motor task (activation), namely, left-handed finger-to-thumb opposition. Two 64×64 slices of whole brain echo-planar fMRI images are acquired every 2.5 s. The dimension of each voxel is $3.1 \times 3.1 \times 8$ mm. Nine runs of 48 images are acquired in the following sequence: 12 images (30 s) of rest, 24 images (60 s) during activation, and 12 images of rest again. The complete time series consists of 432 measurements (18 min) in each voxel. In the following experiments, we focus on the 39×49 area containing the brain and only one slice.

The fMRI signal measures the haemodynamic response to focal neuronal activation. It is widely believed that the response in the brain can be characterized as a convolution of the reference stimulus signal representing the activation with a linear filter. In our experiment, the reference stimulus $h(t)$ is a square wave with positive values for the scans acquired while the subject was performing the motor task and negative values during rest (Fig. 6). Previous approaches tried to characterize the haemodynamic response using several convolution filters: Poisson [32], Gamma [33], or FIR filter with a fixed length [34]. We will try to extend the latter to general FIR filters by extracting the delays that yield the best modeling performance. In other words, we are looking for the set of delays $d_1, \cdots d_m$ such that

$$\hat{y}(t) = \sum_{i=1}^{m} \beta_i h(t - d_i) \tag{9}$$

models the observed fMRI signal $y(t)$ as well as possible. The extension of the ERD algorithm presented in Section III is straightforward. In step 2 of the algorithm, the delays added to the inputs should be the delays of the stimulus reference $h(t)$ rather than of the time series $y(t)$. The modeling and generalization estimations are independent of the particular choice of inputs, and the rest of the algorithm is not affected by this extension.

Based on the cross-correlation between the time series $y(t)$ and the reference signal $h(t)$, we isolate 20 voxels that display a significant activation (see Fig. 6) distributed in two groups that cover the primary motor area and the supplementary motor area. As expected, the left-handed opposition produces a contralateral activation in the right hemisphere. The ERD algorithm is run on each voxel using five out of the nine runs, corresponding to 240 data points, for delay extraction and modeling, whereas the remaining four runs (192 points) are used to test the generalization abilities of the resulting model. The stop condition is set by limiting the extraction to 48 delays, i.e., one run. In agreement with standard statistical practice in subset selection [23], the significance level of the paired t-test is set to 20%. The selected delays range from 1–45 (see the top of Fig. 7). The "effective delay" of each filter is estimated by computing the average of the delay indices d_i, weighted by the absolute filter coefficients $|\beta_i|$. As a result, the activated voxels can be roughly divided in two categories: 14 out of the 20 filters have an average delay between 3 and 10 s, which is roughly within the accepted range of 5–10 s. The remaining voxels have average delays above 15 s, suggesting that the extracted delays do not make sense from a biological point of view, although they efficiently model the relationship between the reference signal $h(t)$ and the fMRI $y(t)$.

The performance of the resulting filters is compared with a FIR filter of fixed size. As the average haemodynamic delay has been observed between 5–10 s [35], i.e., two to four images, we will use an FIR filter with seven delays as reference. In Fig. 7, points above the dotted diagonal mean higher test error, i.e.,

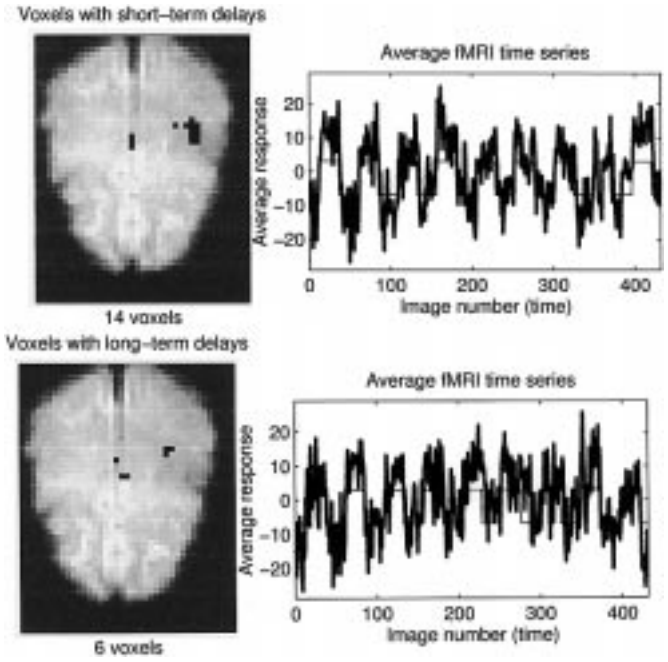

Fig. 8. Voxels with short-term delays (top) between 3–10 s and voxels with long-term delays (bottom), of more than 15 seconds with no biological explanation. The brain maps on the left show the location of the analyzed voxels (black dots); the right plot shows the average fMRI signal in these voxels (thick) and the reference signal (thin).

worse modeling, for the fixed-length filter. Most voxels benefit from using extracted delays, whereas a few display a small decrease in performance. Overall, the ERD filters yield better or comparable performance while using far fewer parameters (2.5 on average). Finally, Fig. 8 displays the location and average fMRI activation signal of both groups of voxels. The voxels with short-term delays (top) and long-term delays (bottom) seem to have slightly different patterns of activation, especially in the last four runs. However, the small sample sizes (14 and 6, respectively) makes a strict comparison difficult.

VI. Discussion and Conclusions

The physical approach adopted by the δ-test investigates the "true" delays with a physical meaning, i.e., the primary delays. As a consequence, the method displays the following salient features.

1) It is data intensive.
2) It extracts one set of delays independently of the model.
3) The selected delays have a natural physical interpretation.

On the other hand, the statistical approach of the ERD or F-inclusion focuses on model performance. It selects the delays for which the model yields the best estimated generalization error by directly optimizing a relevant performance criterion. The resulting delays might not, however, be easily interpreted. This is exemplified by the fMRI example (see Section V), where a quarter of the resulting filters have no apparent biological justification. A second difference is that the variables selected by the statistical methods are model dependent. Note, however, that the overhead is usually limited, as the computational requirements depend on the model estimation; for example, the use of ERD on linear models is extremely fast. ERD improves on traditional statistical selection methods by focusing on the relevant performance criterion, namely, the generalization error. An interesting side effect of the estimators of generalization error is that they naturally take into account the limited size of the available dataset. A similar effect is known from the AIC [36] and BIC [37] information criteria: BIC selects the asymptotically optimal model size, which AIC tends to overestimate, while effectively minimizing the average generalization error for finite datasets. Similarly, ERD selects additional delays that help prediction. In the Hénon map example, the selection of $d = 3$ yields better performance according to both the large validation set method and the independent test set. It should also be noted that it is the *minimum* generalization, rather than its value, that matters, such that the method could very well accommodate an estimator with a consistent bias.

The ERD method is essentially a forward selection method (see Section II-B). It is straightforward to extend it to backward elimination or stepwise regression using the same evaluation criterion. As we argued in Section III, backward elimination is ill suited to temporal modeling. However, stepwise regression can be applied by alternating forward and backward steps. With many models, this can be addressed during the model estimation, e.g., using an adaptive metric approach for kernel smoothing [30], Gaussian processes [38], or pruning and regularization for linear models or neural networks [19], [39], [40]. These methods automatically discard irrelevant variables in the input vector. An additional interesting prospect for future research is the extension of the ERD scheme to more general system identification architecture, e.g., auto-regressive filter with exogenous input (ARX). Note that, in that case, the filter input consists of the system output $y(t)$ and the control signal $u(t)$, and the precise ordering of the candidate inputs has to be determined.

As a conclusion, we have presented a generalization-based method for the extraction of the relevant delays in temporal modeling with a focus on time series prediction and system identification. This method is related to traditional forward selection methods and is therefore well founded. It is easy to implement and requires little time in addition to model estimation. Furthermore, it accommodates efficiently the tradeoff between number of selected delays (model size) and model flexibility (approximation error). The method provides interesting results on a number of experiments performed in different contexts.

Acknowledgment

The Fraser river dataset (Section IV-B) is publicly available on statlib: http://lib.stat.cmu.edu/datasets/. The fMRI dataset (Section V) was acquired by R. Savoy at Massachusetts General Hospital, Boston. The author would like to thank P. Gallinari, J. Larsen, and the learning group at IMM, DTU, for valuable discussions on previous versions of this work.

References

[1] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton Univ. Press, 1961.
[2] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[3] R. R. Hocking, "The analysis and selection of variables in linear regression," *Biometrics*, vol. 32, pp. 1–49, Mar. 1976.
[4] R. Savit and M. Green, "Time series and dependent variables," *Physica D*, vol. 50, no. 1, pp. 95–116, 1991.
[5] H. Pi and C. Peterson, "Finding the embedding dimension and variable dependences in time series," *Neural Comput.*, vol. 6, no. 3, pp. 509–520, May 1994.
[6] X. He and H. Asada, "A new method for identifying orders of input–output models for nonlinear dynamic systems," in *Proc. Amer. Conf. Contr.*, San Francisco, CA, 1993.
[7] C. Molina, N. Sampson, W. J. Fitzgerald, and M. Niranjan, "Geometrical techniques for finding the embedding dimension of time series," in *Proc. IEEE Workshop Neural Networks Signal Process. VI*, 1996, pp. 161–169.
[8] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–368, 1989.
[9] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.*, vol. 26, pp. 917–922, Sept. 1977.
[10] P. Leray and P. Gallinari, "Feature selection with neural networks," Tech. Rep. LIP6 1998/012, 1998.
[11] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Networks*, vol. 5, pp. 537–550, Apr. 1994.
[12] C. Goutte, "Lag space estimation in time series modeling," in *Proc. ICASSP*, 1997, pp. 3313–3316.
[13] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. R. Stat. Soc. B*, vol. 36, pp. 111–147, 1974.
[14] G. Toussaint, "Bibliography on estimation of misclassification," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 472–479, July 1974.
[15] H. Akaike, "Fitting autoregressive models for prediction," *Ann. Inst. Stat. Math.*, vol. 21, pp. 243–247, 1969.
[16] J. Moody, "Note on generalization, regularization and architecture selection in nonlinear learning systems," in *Proc. 1st IEEE Workshop Neural Networks Signal Process.*, B. H. Juang, S. Y. Kung, and C. A. Kamm, Eds. Piscataway, NJ, 1991, pp. 1–10.
[17] J. Larsen and L. K. Hansen, "Generalized performance of regularized neural networks models," in *Proc. IEEE Workshop Neural Networks Signal Processing IV*, J. Vlontzos, J. N. Hwang, and E. Wilson, Eds. Piscataway, NJ, 1994, pp. 42–51.
[18] N. Murata, S. Yoshizawa, and S. Amari, "Network information criterion—Determining the number of hidden units for an artificial neural network model," *IEEE Trans. Neural Networks*, vol. 5, pp. 865–872, June 1994.
[19] C. Goutte, "On the use of a pruning prior for neural networks," in *Proc. IEEE Workshop Neural Networks Signal Processing VI*, S. Usui, Y. Tohkura, S. Katagiri, and E. Wilson, Eds. Piscataway, NJ, 1996, pp. 52–61.
[20] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1992.
[21] S. Siegel, *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 1956.
[22] C. Mallows, "Some comments on C_p," *Technometrics*, vol. 15, pp. 661–675, 1973.
[23] R. B. Bendel and A. A. Afifi, "Comparison of stopping rules in forward "stepwise" regression," *J. Amer. Stat. Assoc.*, vol. 72, no. 357, pp. 46–53, Mar. 1977.
[24] Salahuddin and A. G. Hawkes, "Cross-validation in stepwise regression," *Commun. Stat. Theory Methods*, vol. 20, no. 4, pp. 1163–1182, 1991.
[25] M. Hénon, "A two-dimensional mapping with a strange attractor," *Commun. Math. Phys.*, vol. 50, no. 1, pp. 69–77, 1976.
[26] W. Härdle, "Applied nonparametric regression," in *Econometric Society Monographs*. Cambridge, U.K.: Cambridge Univ. Press, 1990, vol. 19.
[27] M. P. Wand and M.C. Jones, "Kernel Smoothing," in *Monographs on Statistics and Applied Probability*. London, U.K.: Chapman & Hall, 1995, vol. 60.
[28] Z. Aleksić, "Estimating the embedding dimension," *Physica D*, vol. 52, pp. 362–368, 1991.
[29] A. I. McLeod, "Diagnostic checking of periodic autoregression models with application," *J. Time Series Anal.*, vol. 15, no. 2, pp. 221–233, 1994.
[30] C. Goutte and J. Larsen, "Adaptive metric kernel regression," in *Proc. IEEE Workshop Neural Networks Signal Process. VIII*, T. Constantinides, S.-Y. Kung, M. Niranjan, and E. Wilson, Eds. Piscataway, NJ, 1998, pp. 184–193.
[31] C. Goutte, L. K. Hansen, R. Savoy, and S. C. Strother, "Delay analysis of fMRI time series," in *Proc. 4th Int. Conf. Functional Mapping Human Brain*, T. Paus, A. Gjedde, and A. Evans, Eds., 1998, pt. 2 of Neuro Image, vol. 4, p. S611.
[32] K. J. Friston, P. Jezzard, and R. Turner, "Analysis of functional MRI time series," *Human Brain Mapping*, vol. 1, pp. 153–171, 1994.
[33] N. Lange and S. L. Zeger, "Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging," *J. R. Stat. Soc. C*, 1997.
[34] F. Årup Nielsen, L. K. Hansen, P. Toft, C. Goutte, N. Mørch, C. Svarer, R. Savoy, B. Rosen, E. Rostrup, and P. Born, "Comparison of two convolution models for fMRI time series," in *Proc. 3rd Int. Conf. Functional Mapping Human Brain*, L. Friberg, A. Gjedde, S. Holm, N. A. Lassen, and M. Nowak, Eds., May 1997, pt. 2, vol. 3, p. S473.
[35] P. A. Bandettini, A. Jesmanowicz, E. C. Wong, and J. S. Hyde, "Processing strategies for time-course data sets in functional MRI of the human brain," *Magn. Reson. Med.*, vol. 30, no. 2, pp. 161–173, Aug. 1993.
[36] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, Dec. 1974.
[37] G. Schwartz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.
[38] C. K. I. Williams, "Prediction with Gaussian processes: From linear regression to linear prediction and beyond," Neural Comput. Res. Group, Aston Univ., UK, Tech. Rep. NCRG/97/012, 1997.
[39] Y. Le Cun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in Neural Information Processing Systems*, D. S. Touretzky, Ed. San Francisco, CA: Morgan Kaufmann, 1990, pp. 598–605.
[40] C. Goutte and L. K. Hansen, "Regularization with a pruning prior," *Neural Networks*, vol. 10, no. 6, pp. 1053–1059, 1997.

Cyril Goutte received the M.E. degree from ENSTA, Paris, France, in 1992 and the M.Sc. degree in computer science from ENSTA and the University of Paris 11 the same year. He received the Ph.D. degree from the University of Paris 6 in 1997 for the dissertation "Statistical learning and regularistion for regression."

From 1996 to 1999, he was a Post-Doctoral Reseacher and an Assistant Research Professor with the Department of Mathematical Modeling, Technical University of Denmark, Lyngby. In 1999, he joined Nokia Mobile Phones R&D, Copenhagen, Denmark. His main interests lie in statistical learning, pattern recognition, and cryptography