2003s-24

Metric-Based Model Selection For Time-Series Forecasting

Yoshua Bengio, Nicolas Chapados

Série Scientifique
Scientific Series

Montréal
Mai 2003

CIRANO
Centre interuniversitaire de recherche
en analyse des organisations

Les cahiers de la série scientifique (CS) visent à rendre accessibles des résultats de recherche effectuée au CIRANO afin de susciter échanges et commentaires. Ces cahiers sont écrits dans le style des publications scientifiques. Les idées et les opinions émises sont sous l’unique responsabilité des auteurs et ne représentent pas nécessairement les positions du CIRANO ou de ses partenaires.

This paper presents research carried out at CIRANO and aims at encouraging discussion and comment. The observations and viewpoints expressed are the sole responsibility of the authors. They do not necessarily represent positions of CIRANO or its partners.

ISSN 1198-8177

Metric-Based Model Selection For Time-Series Forecasting

Yoshua Bengio[*], *Nicolas Chapados*[†]

Résumé / *Abstract*

Les méthodes métriques, et qui utilisent des données non-étiquetées pour détecter les différences brutes pour les comportements loin des pointes d'entrainement, ont été récemment introduites pour la sélection de modèles, apportant une amélioration dans beaucoup de cas (incluant la validation croisée). Nous présentons des prolongements à ces méthodes qui prennent avantage du cas particulier des séries temporelles pour lesquelles la tâche consiste en une prédiction avec un horizon "h". Les idées sont (i) d'utiliser au temps "t" les "h" exemples non-étiquetés qui précèdent "t", et (ii) profiter des différentes distributions d'erreur de validation croisée et de méthodes métriques. Des résultats expérimentaux établissent l'efficacité de ces prolongements dans le contexte de la sélection d'un sous-ensemble de caractéristiques.

Mots clés : Données non-étiquetées, sélection de modèles, séries temporelles.

Metric-based methods, which use unlabeled data to detect gross differences in behavior away from the training points, have recently been introduced for model selection, often yielding very significant improvements over alternatives (including cross-validation). We introduce extensions that take advantage of the particular case of time-series data in which the task involves prediction with a horizon "h". The ideas are (i) to use at "t" the "h" unlabeled examples that precede "t" for model selection, and (ii) take advantage of the different error distributions of cross-validation and the metric methods. Experimental results establish the effectiveness of these extensions in the context of feature subset selection.

Keywords: *Unlabeled data, model selection, time-series.*

[*] CIRANO et Département d’informatique et recherche opérationnelle, Université de Montréal, Montréal, Québec, Canada, H3C 3J7, tél.: (514) 343-6804. Courriel: bengioy@iro.umontreal.ca.

[†] Département d’informatique et recherche opérationnelle, Université de Montréal, Montréal, Québec, Canada, H3C 3J7. Courriel : chapados@iro.umontreal.ca.

MODEL SELECTION AND REGULARIZATION

Supervised learning algorithms take input/output training pairs $\{(x_1, y_1) \cdots (x_l, y_l)\}$ sampled (usually independently) from an unknown joint distribution $P(X, Y)$ and attempt to infer a function $f \in \mathcal{F}$ that minimizes the expected value of the loss $L(f(X), Y)$ (also called the *generalization error*). In many cases one faces the dilemma that if $\mathcal{F}$ is too "rich" then the average training set loss (*training error*) will be low but the expected out-of-sample loss may be large (*overfitting*), and vice-versa if $\mathcal{F}$ is not "rich" enough (*underfitting*).

In many cases one can define a collection of increasingly complex function classes $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$ (although some methods studied here work as well with a partial order). *Model selection* methods attempt to choose one of these function classes to avoid both overfitting and underfitting. For example, in the case of variable subset selection, these subsets may correspond to the number of input variables that are allowed (e.g. $\mathcal{F}_i$ is the set of linear regressions with i input variables). One approach to model selection is based on *complexity penalization* [5, 3]. Another approach to model selection is based on *held-out data*: one selects the model with the lowest generalization error, estimated by repeatedly training on a subset of the data and testing on the rest, e.g. using the bootstrap, leave-one-out or K-fold cross-validation (XVT). The *metric-based* methods introduced by Schuurmans [6, 7] are somewhat in between in that they take advantage of *unlabeled* data not used for training (but only the input part) in order to introduce a complexity penalty. These methods take advantage of unlabeled data: the behavior of functions corresponding to different choices of complexity are compared on the training data and on the unlabeled data, and differences in behavior that would indicate overfitting are exploited to perform model selection. An overview of advances in model selection and feature selection methods can be found in a recent Machine Learning special issue [1].

After a review of metric-based model selection methods, we introduce the extensions proposed in this paper that deal specifically with time-series data.

METRIC-BASED MODEL SELECTION

Metric-based methods for model selection are based on the idea that solutions that overfit are likely to behave very differently on the training points and on other points sampled from the input density $P_X(x)$. This occurs because the learning algorithm tries to reduce the loss at the training points (but not necessarily elsewhere since no data is available there), whereas we want the solution to work well not only on the training points but in general where $P_X(x)$ is not small. These metric-based methods are all based on the definition of a *metric* (or pseudo-metric) on the space of functions, which allows to judge how far two functions are from each other:

$$d(f, g) = \psi(E[L(f(X), g(X))])$$

where the expectation $E[\cdot]$ is over $P_X(x)$ and ψ is a normalization function. For example with the quadratic loss $L(u, v) = (u - v)^2$, the proper normalization function is $\psi(z) = z^{1/2}$. Although $P_X(x)$ is unknown, Schuurmans (1997) proposed to estimate $d(f, g)$ using an average $d_U(f, g)$ computed an *unlabeled set* U (i.e. points x_i sampled from $P_X(x)$ but for which no associated y_i is given). In what follows we shall use $d_U(f, g)$ to denote the distance estimated on the unlabeled set U:

$$d_U(f, g) = \psi(\frac{1}{|U|} \sum_{i \in U} L(f(x_i), g(x_i))) \tag{1}$$

The metric-based methods proposed in [6, 7] are based on comparing $d_U(f, g)$ with the corresponding average distance $d_T(f, g)$ measured on the *training set* T.

Schuurmans (1997) first introduced the idea of a metric-based model selection by taking advantage of possible violations of the *triangle inequality*. Improved results were described in [7] with a new penalization model selection method, based on similar ideas, called **ADJ**, which chooses the hypothesis function f_l which minimizes the *adj*usted loss

$$d_T(f_l, P_{Y|X}) \max_{k<l} \frac{d_U(f_k, f_l)}{d_T(f_k, f_l)},$$

where $d_T(f_l, P_{Y|X})$ denotes the training error. See [7] for more detailed justification, including proofs of bounds on the maximum overfitting and underfitting, and experiments showing that these methods outperform classical model selection procedures (including XVT) on some small artificial data sets (with between 10 and 30 training examples) on which overfitting can be severe.

EXTENSION TO TIME-SERIES FORECASTING

We now turn to the case of applying statistical learning algorithms to time-series data, such as economic or financial data, that may be non-stationary. At time t, we have an information set $\mathcal{I}_t$ which includes all measurable observations at and prior to time t. We want to forecast some aspect $y_{t+h} = y(\mathcal{I}_{t+h})$ of this information set at a future time $t + h$, using some aspect of the information available at t, $x_t = y(\mathcal{I}_t)$. Because of the possible non-stationarity of the data (dependence on t of $P_t(y_{t+h}|x_t)$), estimating generalization error is often done with the **sequential validation** technique, rather than with leave-one-out or K-fold XVT. Sequential validation is based on the analysis of the sequence of losses obtained by sliding a learning algorithm A over the time sequence, as shown in Algorithm 1.

The most important result of the sequential validation algorithm is the average loss, which can be compared across several algorithms. The individual losses are useful to estimate confidence intervals around the average loss or around differences in average loss.

In the sequential validation algorithm we would in general prefer to choose $\Delta t = 1$ but larger values allow to save computations (in proportion to the value of Δt). The choice of the training window size w_t depends on the degree of non-stationarity expected (or estimated) from the particular data sequences. The most

Algorithm 1 Sequential Validation

Input: data sequences $\{x_t\}, \{y_t\}$ (t ranging from 1 to T), learning algorithm A, loss functional L, forecast horizon h, step Δt, training window size w_t (often fixed to a constant), and first test point t_0.

```
For $t$ ranging from $t_0$ to $T - h$ by steps $\Delta t$
    Training set:  $\mathcal{D}_t = \{(x_s, y_{s+h})\}$,  $s \in [t - w_t, t - h)$
    Solution at $t$ is:  $f_t = A(\mathcal{D}_t)$
    Test set:  $\mathcal{T}_t = \{(x_s, y_{s+h}\}$,  $s \in [t, t + \Delta t)$
        Forecast at $s$:  $f_t(x_s)$
        Loss at $s$:  $l_s = L(f_t, (x_s, y_{s+h}))$
```

Output: the sequence of losses $\{l_t\}$, for $t \in [t_0, T - h]$

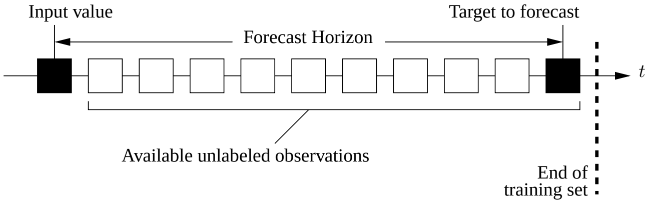

Figure 1. *A natural source of unlabeled data for time series forecasting with an horizon arises at the end of the training set.*

common choices are $w_t = min(w_0, t)$ (a fixed value) and $w_t = t$ (use all the available data). Using sequences shorter than t may be justified when the conditional distribution of the data changes so much with t as for old training pairs to hurt generalization to new cases. Note that unless w_t is constant the amount of training data may change as t increases, thus usually requiring an adaptive model selection algorithm.

Natural Source of Unlabeled Data

Inspection of the sequential validation algorithm quickly reveals that at time t there are h input vectors x_s ($s \in [t - h + 1, t]$) which cannot be associated with a corresponding target output y_{s+h}. The idea of the proposed extension is to use these **unlabeled points** to form the unlabeled set U required in the metric methods (to compute d_U, as in equation 1). This phenomenon is illustrated in Figure 1.

It is also interesting to note that this unlabeled set includes in particular the input for the next test point, x_t. This suggests that a method that uses x_t for model selection is actually doing a form of **transduction**. Vapnik introduced in 1982 the principle of **transductive inference**, which differs from the usual **inductive inference** principle in that the learner chooses a solution based not only the training set but also on the input values of the test point(s). Here, this is particularly true when the sequential validation step Δt is chosen equal to 1 (which is however more computationally costly). Otherwise, only one out of Δt of the test points would be in U. Why would it be useful to use the metric model selection methods with the future test points as unlabeled data?— the intuition is simply that these are the data points that we care about: this is where we want to reject functions that “misbehave”.

Time-Series Transduction Experiments

Experimental Setup.
To verify the potential of metric model selection methods in time-series forecasting applications, we performed feature-selection experiments using artificially-generated data in a controlled setting. We wish to compare model selection algorithms (in this case the metric method ADJ against XVT) on the set of progressively more complex models that arise in forward (stepwise) feature selection.

Data Generation. The artificial data series are generated from the class of linear autoregressive $AR(K)$ models, where given a fixed coefficients vector $\alpha \equiv (\alpha_0, \ldots, \alpha_K)'$ and initial conditions $y_{-1}, y_{-2}, \ldots, y_{-K}$, we have the process

$$y_t = \alpha_0 + \sum_{k=1}^{K} \alpha_k y_{t-k} + \epsilon_t, \quad t \geq 0. \tag{2}$$

with $\epsilon_t \sim N(0, \sigma^2)$ i.i.d. gaussian noise. To simplify matters and ease analysis, we restrict the generating models to the specific form $y_t = \alpha + \alpha y_{t-K} + \epsilon_t$, where in our experiments $K = 1, 2, 3$.

Task Description. We seek to forecast the series $\{y_t\}$ at horizon h, given the realizations of the past $\tilde{K}$ series values (we do not impose that $\tilde{K}$ be equal to the order K of the generating process). One typically considers a *point forecast*, or in other words, at a given time t and given the values of $\{y_t, y_{t-1}, \ldots, y_{t-\tilde{K}+1}\}$, one seeks an estimator of $E[y_{t+h}|\mathcal{I}_t]$. However, in our experiments, we shall consider an "integrated" forecast, consisting of the *sum of the series values* over the horizon. We shall then seek an estimator of $E[y_{t+1} + y_{t+2} + \cdots + y_{t+h}|\mathcal{I}_t]$. In many applications this type of forecast can be interpreted more naturally in terms of the underlying problem variables; for instance, given a financial series of (log) returns, the integrated forecast corresponds to the estimated total portfolio (log) return over the horizon. Obviously, at horizon $h = 1$, the integrated forecast is equivalent to the point forecast.

We shall consider the class of $AR(\tilde{K})$ models. This is equivalent to estimating the coefficients $\hat{\beta} \equiv (\hat{\beta}_0, \ldots, \hat{\beta}_{\tilde{K}})'$ corresponding to the model

$$\sum_{j=1}^{h} y_{t+j} = \beta_0 + \sum_{k=1}^{\tilde{K}} \beta_k y_{t-k+1} + \epsilon_t,$$

where ϵ_t is i.i.d. gaussian noise. The estimation of $\hat{\beta}$, for a fixed $\tilde{K}$, is easily performed analytically using the ridge estimator, $\hat{\beta}^* = (X'X + \lambda I)^{-1} X'Y$, where X is the matrix of regressors, Y is the (column-) vector of targets, λ is a weight-decay hyperparameter, and I is the identity matrix.[1] This estimator implicitly uses a squared-error loss function, which is appropriate for our task of estimating a conditional expectation.

[1] In our procedure, we do not penalize the mean estimator $\hat{\beta}_0$; hence, the mean is estimated without bias.

Feature Selection. The role of feature selection here is to decide which α_k are significant and should be included in the regression. To this end, we use a standard forward stepwise selection algorithm, in which the individual features are the lagged series values, $y_{t-k}, k = 0, \ldots, \tilde{K} - 1$. Forward selection proceeds incrementally, starting from the mean (the lowest-complexity model that we are willing to consider), and at each step adds the feature that minimizes the training error. At a given time step t, we have the following sequence of models produced by the algorithm,

$$\{\hat{f}_t^{(0)}, \hat{f}_t^{(1)}, \ldots, \hat{f}_t^{(\tilde{K})}\},$$

where $\hat{f}_t^{(k)}$ is the estimated regression model containing the k "best" features according to forward selection (which are not necessarily the first k lagged series values). The model $\hat{f}_t^{(0)}$ is simply the mean on the current training set (obtained from $\mathcal{I}_t$). We observe that this sequence of models forms a total order with respect to complexity, and is thence *amenable to selection by metric methods*. We exploit this crucial property, which arises naturally from the nature of the forward selection algorithm, in the experiments.

Experimental Plan. The experiments measure the relative ability of 10-fold XVT versus metric model selection (in this case, ADJ) to select among the sequence of models produced by stepwise selection. We compare the methods across a whole spectrum of parameters, i.e. all permutations of (i) Forecasting horizon $h = \{1, 2, 5, 10, 15\}$, (ii) Generating model AR order $K = \{1, 2, 3\}$, (iii) Generating model coefficient magnitude $\alpha = \{0.1, 0.3, 0.5, 0.7, 0.9\}$. This last coefficient controls the series *signal-to-noise ratio*; $\alpha = 0.1$ yields series very close to white noise, whereas series with $\alpha = 0.9$ exhibit much more structure. Each triplet $\langle \text{horizon}, \text{AR order}, \text{magnitude} \rangle$ is henceforth called an *experiment*.

We fix the maximum model order $\tilde{K} = 10$, and a constant training window size $w_t = 75 = t_0$, making this a challenging task. The sequential validation increment is $\Delta_t = 10$, and the total length of each generated series is 1000 observations. In addition, each "basis" model $f_t^{(k)}$ is estimated with a small ridge penalty $\lambda = 10^{-\frac{1}{4}}$. (This hyperparameter was not tuned extensively, but empirically produced quite reasonable results.)

Statistical Methodology.

We compare the performance of two models by a usual paired t-test on their mean-squared error difference. However, the results of individual experiments (e.g. across different horizons) cannot be pooled randomly, since the expected error distribution is quite different across experiments. For instance, we *expect a priori* the MSE to be higher when forecasting across a longer horizon, given a stationary underlying generating process. To perform a valid statistical test of the performance difference between methods *across experiments*, it is necessary to normalize the distribution of paired differences *within each experiment* to have unit standard deviation, before pooling the observations across experiments, and then performing the statistical test.

More specifically, suppose we perform M experiments, each one with N_m test points. Let $e_i^m, m = 1, \ldots, M, i = 1, \ldots, N_m$ be the squared error differences between two methods we wish to compare (e.g. XVT against ADJ in our case).

The first step is to normalize the distribution of error differences to unit standard deviation,

$$\tilde{e}_i^m = \frac{e_i^m}{\sqrt{\hat{\sigma}^2(e^m)}}, \tag{3}$$

where the variance estimator $\hat{\sigma}^2(e^m)$ is described below. Then we compute the overall mean difference $\bar{e}$ and standard error $\hat{\sigma}_{\bar{e}}$ as

$$\bar{e} = \frac{\sum_{m=1}^{M} \sum_{i=1}^{N_m} \tilde{e}_i^m}{\sum_{m=1}^{M} N_m}, \qquad \hat{\sigma}_{\bar{e}} = \frac{1}{\sqrt{\sum_{m=1}^{M} N_m}}. \tag{4}$$

Throughout this section, the so-obtained mean difference $\bar{e}$ is termed *normalized MSE difference*.

Estimation of $\sigma(e^m)$. The question left open is the estimation of the standard deviation of the error-difference distribution within a single experiment. The usual estimator cannot be used here for it rests upon an i.i.d. assumption, whereas the series we consider exhibit mild to strong autocorrelation patterns. This autocorrelation is induced, on the one hand, by the problem structure, and on the other hand by the sequential validation testing procedure.[2]

To properly estimate the variance, we use the Newey–West estimator well-known to econometricians [4, 2], which in addition to being consistent, has the desirable property of being robust at small sample sizes,[3]

$$\hat{\sigma}^2(e^m) = \hat{\gamma}_0^m + 2\sum_{j=1}^{q} \frac{q-j}{q} \hat{\gamma}_j^m, \tag{5}$$

where q is the maximum lag length to be considered,[4] and $\hat{\gamma}_j^m$ is the empirical lag-j autocovariance,

$$\hat{\gamma}_j^m = \frac{1}{N_m - j} \sum_{i=1}^{N_m - j} (e_i^m - \bar{e}^m)(e_{i+j}^m - \bar{e}^m), \tag{6}$$

with $\bar{e}^m$ the sample mean.

Experimental Results.

Figure 2 presents a summary of the experiments described above, comparing XVT against ADJ. The left plot outlines the effect of the series signal-to-noise (SNR) ratio (for which the generating model AR coefficient magnitude are a proxy) on performance. At very low SNR, the series being essentially white noise, both methods perform about equally poorly (worse, in fact, than a naïve constant model (not shown on the figure)). At the other end of the spectrum, at high coefficient values,

[2] Since successive training sets in sequential validation tend to highly overlap, the trained models are generally very correlated—especially at small step sizes Δ_t—thence inducing correlation in the error structure.

[3] It guarantees the positive-definiteness of the estimated covariance matrix in the multivariate case.

[4] This must scale with the sample size for the estimator to be consistent, but not too rapidly.

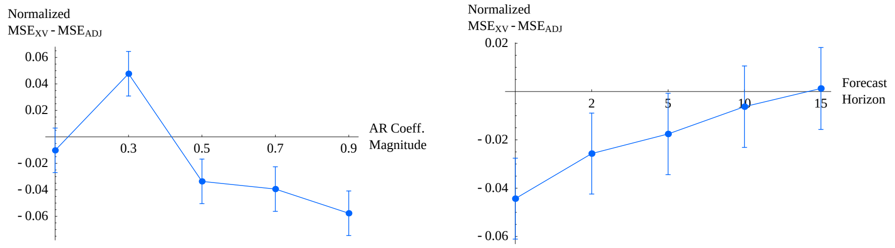

Figure 2: ***Left:*** *Normalized MSE difference between the models chosen by XVT and ADJ, as a function of the magnitude of the AR coefficients (across all forecast horizons and generating model order). The error bars represent 95% confidence intervals on the mean difference (normalized as explained in the text).* ***Right:*** *Same measure, as a function of the forecast horizon; we note that even with* ***extremely few*** *unlabeled observations (one or two), ADJ does not lose catastrophically against XVT, which is very surprising; the two methods become essentially equivalent for longer horizons.*

XVT performs, overall, significantly better than ADJ. However, the opposite picture emerges at small but significant coefficient values, where ADJ significantly beats XVT. We conjecture that at these moderate SNR levels, the intrinsic variance of the choice made by XVT causes costly mistakes, whereas a less-variable (albeit biased) method such as ADJ can pick out important structures without being swamped by the noise level.

The right plot in Figure 2 is, in some ways, more surprising: first, the expected outcome shows a steady improvement in the performance of ADJ with respect to XVT as the forecast horizon increases, as a result of the increase in the number of unlabeled observations that ADJ can use to make its choice. But the unexpected outcome is, relatively speaking, **how well ADJ performs given extremely few unlabeled observations** (one or two); recall that these observations are used to form a Monte Carlo approximator of an expectation (*c.f.* eq. 1), and that so few observations are sufficient to make a reasonable model selection choice in this context strikes us as a surprise.

Moreover, we can count the number of experiments for which each method statistically significantly beats the other; a kind of model selection tournament (we shall take $p \leq 0.05$ as the significance level). The results comparing ADJ to cross validation are shown in Table 1. The hypothesis about the behavior of each method

	XV Wins	ADJ Wins	Total Exp.
Overall	17	7	75
AR Coeff = 0.01			15
AR Coeff = 0.03		7	15
AR Coeff = 0.05	1		15
AR Coeff = 0.07	6		15
AR Coeff = 0.09	10		15
Horizon = 1	7	1	15
Horizon = 2	6	1	15
Horizon = 5	3	3	15
Horizon = 10	1	2	15
Horizon = 15			15

Table 1. *"Tournament" results comparing XVT against ADJ for individual experiments. A "win" indicates that the corresponding method beats the other statistically significantly ($p \leq 0.05$) on the MSE criterion. A blank stands for zero. The results corroborate those of Figure 2.*

Algorithm 2 Logistic Hybrid Model Selection

```
Input at $t$: the sequence of solutions $f_s^{xv}$ and $f_s^p$, respectively for XVT and complexity penalization selected models, and the data sequence $\{(x_s, y_s)\}$ for $s \leq t$.
    1. Let $d_s = f_s^{xv}(x_s) - f_s^p(x_s)$
    2. Let $w_s(\beta) = 1/(1 + exp(-(\beta_0 + \beta_1 d_s + \beta_2 d_s^2)))$
    3. Let $C(\beta) = \sum_{s \leq t-h} (w_s(\beta) f_s^{xv} + (1 - w_s(\beta)) f_s^p - y_{s+h})^2$
    4. Let $\beta^* = \text{argmin}_\beta C(\beta)$
Output at $t$: the solution $f_t = w_t(\beta) f_t^{xv} + (1 - w_t(\beta)) f_t^p$.
```

at a given series SNR finds more confirmation; a further surprise emerges from the horizon data, where we find that ADJ sometimes **significantly beats** XVT even at very small forecast horizon (i.e. using extremely few unlabeled points). The two methods become indistinguishable at longer horizons.

HYBRID MODEL SELECTION

The motivation for this final extension to metric model selection follows from several years of working with various model selection methods and frustratingly comparing them against XVT. XVT does not always work but it almost always performs quite well. However, it tends to have higher variance (in the sense of larger variations in error) than complexity penalization methods. We also know that it is almost unbiased (it is unbiased for training with a bit less examples than what is actually available).[5] Since it is usually almost as good (and often better) than these complexity penalization methods (including the metric methods), it must mean that these other methods must have smaller variance (and none of them is guaranteed to be unbiased, so they are likely to be biased). Can we take advantage of this situation, whereby one method is more biased but has less variance than the other. In this paper we have just begun to explore this opportunity. Let us call f^{xv} the solution obtained by XVT and f^p the solution obtained by some form of complexity penalization, for a particular training set. A simple-minded combination algorithm is the following: if, for a given test point x, the absolute difference $|f^{xv}(x) - f^p(x)|$ is “large”, then trust f^p, else trust f^{xv}. The intuition for this heuristic rule is that a large difference in function value more likely indicates that the cross-validatory choice is wrong, owing to its large variance. This leaves open the question of choosing the proper threshold. A more sophisticated (and better grounded) algorithm for the squared loss, which we have tested in the experiments is shown in Algorithm 2.

This algorithm is based on the idea of the logistic regression: it assigns a weight $w_s(\beta)$ to the XVT model (and $1 - w_s(\beta)$ to the ADJ model) based a quadratic function of the difference $f^{xv}(x) - f^p(x)$; the use of the sigmoid ensures that the weights are always between 0 and 1. Contrarily to traditional logistic regression, the coefficient vector β for the weights is obtained by directly optimizing a squared-loss criterion, estimated from the *past test observations* (i.e. those for $s \leq t$, available

[5] We are talking about the bias of an estimator of generalization error. However, for most model selection methods, the only bias we care about is not in the value of the estimator but only of how it ranks different hypotheses.

TABLE 2: *"Tournament" results comparing the logistic combination against, respectively XVT alone and ADJ alone. A "win" indicates that the corresponding method beats the other statistically significantly ($p \leq 0.05$) on the MSE criterion. A blank stands for zero.*

	Logis Wins	XV Wins	Logis Wins	ADJ Wins	Total Exp.
Overall	6	1	15	1	75
AR Coeff = 0.01			1		15
AR Coeff = 0.03	6				15
AR Coeff = 0.05		1	3	1	15
AR Coeff = 0.07			3		15
AR Coeff = 0.09			8		15
Horizon = 1	1		7		15
Horizon = 2	1		6		15
Horizon = 5	3	1	2		15
Horizon = 10	1				15
Horizon = 15				1	15

without cheating from the sequential validation procedure).

Hybrid Model Selection Experiments

The same experimental setup as described previously was used to compare the logistic combination rule against XVT alone and ADJ alone. "Tournament" results are shown in Table 2. The most significant observation is that the logistic combination **almost always performs better** than either method taken alone. As the table shows, across all experiments, the logistic combination loses only once to XVT and ADJ, whereas it wins quite more frequently against them. Finally, Figure 3 illustrates typical cases of the weight attributed to the XVT model by the logistic combination (as a function of $f^{xv} - f^p$). (The ADJ model, as always, gets the opposite weight). It confirms the intuition outlined above: in case of "small" differences (but with a bias empirically estimated from the data) between f^{xv} and f^p, choose XVT, otherwise choose ADJ.

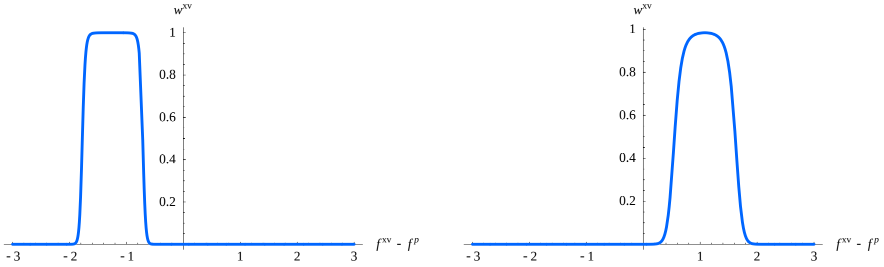

Figure 3: *Examples of the weight $w_s(\beta)$ given to the XVT model (*c.f.* Algorithm 2), obtained by the logistic regression, as a function of the difference $f^{xv} - f^p$ evaluated at the test point.*

CONCLUSION

We have proposed extensions to metric-based model selection (ADJ in particular) to take advantage of (i) the particular structure of time-series data using a transductive inference procedure, and (ii) the difference in error profiles between ADJ and XVT. It is a surprising result that ADJ can work so well with time-series data, using only as few as 1 to 15 unlabeled examples, as one would have expected a very unreliable complexity correction with so few points. Moreover, we have opened a very exciting new avenue to combine model selection methods that exhibit very different error profiles (e.g. one has large variance, the other has bias). The experiments show that the hybrid method is almost never beaten by XVT or by ADJ, and often beats one or the other. Probably more questions have been raised than answered in this work: Why is ADJ working well with so few unlabeled data when these include the next test point?— this is probably related to the transductive effect, but a true theory is lacking. Finally, can we push further the last extension to other sets of model selection methods, or with better combination algorithms?

REFERENCES

[1] Y. Bengio and D. Schuurmans, "Special Issue on New methods for model selection and model combination," 2002, *Machine Learning*, 48(1).

[2] J. Campbell, A. W. Lo and A. MacKinlay, **The Econometrics of Financial Markets**, Princeton: Princeton University Press, 1997.

[3] D. Foster and E. George, "The risk inflation criterion for multiple regression," **Annals of Statistics**, vol. 22, pp. 1947–1975, 1994.

[4] W. Newey and K. West, "A Simple, Positive Semi-Definite, Heteroscedasticity and Autocorrelation Consistent Covariance Matrix," **Econometrica**, vol. 55, pp. 703–708, 1987.

[5] J. Rissanen, "Stochastic complexity and modeling," **Annals of Statistics**, vol. 14, pp. 1080–1100, 1986.

[6] D. Schuurmans, "A new metric-based approach to model selection," in **Proceedings of the National Conference on Artificial Intelligence (AAAI-97)**, 1997, pp. 552–558.

[7] D. Schuurmans and F. Southey, "Metric-based methods for adaptive model selection and regularization," **Machine Learning**, vol. 48, no. 1, pp. 51–84, 2002.

[8] V. Vapnik, **Estimation of Dependences Based on Empirical Data**, Berlin: Springer-Verlag, 1982.