2002s-51

On Out-of-Sample Statistics for Time-Series

François Gingras, Yoshua Bengio,
Claude Nadeau

Série Scientifique
Scientific Series

CIRANO
Centre interuniversitaire de recherche
en analyse des organisations

Montréal
Mai 2002

CIRANO

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du ministère de la Recherche, de la Science et de la Technologie, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO is a private non-profit organization incorporated under the Québec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the Ministère de la Recherche, de la Science et de la Technologie, and grants and research mandates obtained by its research teams.

Les organisations-partenaires / The Partner Organizations

- •École des Hautes Études Commerciales
- •École Polytechnique de Montréal
- •Université Concordia
- •Université de Montréal
- •Université du Québec à Montréal
- •Université Laval
- •Université McGill
- •Ministère des Finances du Québec
- •MRST
- •Alcan inc.
- •AXA Canada
- •Banque du Canada
- •Banque Laurentienne du Canada
- •Banque Nationale du Canada
- •Banque Royale du Canada
- •Bell Canada
- •Bombardier
- •Bourse de Montréal
- •Développement des ressources humaines Canada (DRHC)
- •Fédération des caisses Desjardins du Québec
- •Hydro-Québec
- •Industrie Canada
- •Pratt & Whitney Canada Inc.
- •Raymond Chabot Grant Thornton
- •Ville de Montréal

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

On Out-of-Sample Statistics for Time-Series[*]

François Gingras[†], Yoshua Bengio[‡], and Claude Nadeau[§]

Résumé / *Abstract*

Cet article étudie une statistique hors-échantillon pour la prédiction de séries temporelles qui est analogue à la très utilisée statistique R^2 de l'ensemble d'entraînement (in-sample). Nous proposons et étudions une méthode qui estime la variance de cette statistique hors-échantillon. Nous suggérons que la statistique hors-échantillon est plus robuste aux hypothèses distributionnelles et asymptotiques pour plusieurs tests faits pour les statistiques sur l'ensemble d'entraînement (in-sample). De plus, nous affirmons qu'il peut être plus important, dans certains cas, de choisir un modèle qui généralise le mieux possible plutôt que de choisir les paramètres qui sont le plus proches des vrais paramètres. Des expériences comparatives furent réalisées sur des séries financières (rendements journaliers et mensuels de l'indice du TSE300). Les expériences réalisées pour plusieurs horizons de prédictions, et nous étudions la relation entre la prédictibilité (hors-échantillon), la variabilité de la statistique R^2 hors-échantillon, et l'horizon de prédiction.

This paper studies an out-of-sample statistic for time-series prediction that is analogous to the widely used R^2 in-sample statistic. We propose and study methods to estimate the variance of this out-of-sample statistic. We suggest that the out-of-sample statistic is more robust to distributional and asymptotic assumptions behind many tests for in-sample statistics. Furthermore we argue that it may be more important in some cases to choose a model that generalizes as well as possible rather than choose the parameters that are closest to the true parameters. Comparative experiments are performed on a financial time-series (daily and monthly returns of the TSE300 index). The experiments are performed for varying prediction horizons and we study the relation between predictibility (out-of-sample R^2), variability of the out-of-sample R^2 statistic, and the prediction horizon.

Keywords*: Out-of-sample statistic, time series, TSE300*

Mots-clés : Statistique hors-échantillon, séries financières, TSE300

[*] We would like to thank John W. Galbraith for his useful comments, and the NSERC funding agency for support.
[†] Hydro-Québec. Email: gingras.francois@hydro.qc.ca
[‡] CIRANO and Département d'informatique et recherche opérationnelle, Université de Montréal, Montréal, Québec, Canada, H3C 3J7. Tel: +1 (514) 343-6804, email: bengioy@iro.umontreal.ca
[§] Statistics Canada. Email: claude.nadeau@statcan.ca

1 Introduction

The purpose of the analysis of time-series such as financial time-series is often to take decisions based on data $D_T = \{z_1, \ldots, z_T\}$, with $Z_t = (X_t, Y_t)$. In this paper, we will focus on decisions which take the form of a prediction $\hat{y}_{T+h}$ of the future value of some variable [1], say Y_{T+h}. The quality of the prediction will be judged a posteriori according to some loss function, such as the squared difference between the prediction $\hat{y}_{T+h}$ and the realization Y_{T+h} of the predicted variable $(\hat{y}_{T+h} - Y_{T+h})^2$. A common approach is to use the historical data D_T to infer a function f that takes as input the value of some summarizing information X_t and produces as output $\hat{y}_t = f(X_t)$, which in the case of the above quadratic loss function would be an estimate of the conditional expectation $E[Y_t|X_t]$. The hope is that if this function worked well on observed past pairs (x_t, y_t), it should work well on (X_{T+h}, Y_{T+h}) [2].

How should we choose the function f? A classical approach is to assume a parametrized class of functions, like affine functions, estimate the value of these parameters by maximum likelihood or least squares. Then the model is accessed via goodness-of-fit tests and statistical tests to verify if these parameters differ significantly from the value that would be consistent with a null hypothesis (e.g., the parameters of the regression are significantly different from zero, so that there is really a linear dependency between the X's and the Y's). In particular, these tests are important to know whether one should use the proposed model at all, or to decide among several models.

In this paper we will consider alternative approaches to address the last question, i.e., how a model should be validated and how several models should be compared. It is very satisfying to obtain a result on the **true value** of the parameters (e.g., to use an efficient estimator, which converges as fast as possible to the true value of the parameters). But in many applications of time-series analysis, the end-user of the analysis may be more interested in knowing whether the model is going to work well, i.e., to generalize well to the future cases. In fact, we will argue that sometimes (especially when data is scarce), the two objectives (estimating the true parameters or choosing the model that generalizes better) may yield very different results. Another fundamental justification for the approach that we are putting forward is that we may not be sure that the true distribution of the data has the form (e.g. linear, Gaussian, etc...) that has been assumed. Therefore it may not be meaningful to talk about the true value of the parameters, in this case. What may be more appropriate is the question of generalization performance: will the model

[1] In this paper we will normally use upper case for random variables and lower case for their value.

[2] Obviously $\hat{y}_{T+h} = f(X_{T+h})$ is computable only if X_{T+h} is available. We will typically consider lagged variables so that X_{T+h} is available at "time" T if X is lagged by an amount greater or equal to h.

yield good predictions in the future? where the notion of "good" can be used to compare two models. To obtain answers to such questions, we will consider statistics that measure **out-of-sample performance**, i.e., measured on data that was not used to form the prediction function. This contrast with the in-sample R^2 used in predictability tests [Campbell et al., 1997, Kaul, 1996].

Using a measure of performance based on out-of-sample errors is an approach gaining in popularity. In econometrics, Diebold and Mariano [Diebold and Mariano, 1995] are using out-of-sample errors (or predictive performances) to build tests on accuracy measures. In the machine learning community, it is common to use such measures of performance. Splitting the data into a training subset and a test subset is a popular option. For smaller data sets, K-fold cross-validation is preferred [Efron and Tibshirani, 1993]. The above methods may not be applicable to sequential data, and in the non-stationary case may yield optimistic estimates.A more honest estimate can be obtained with a **sequential cross-validation procedure**, described in this paper. This estimate essentially attempts to measure the predictability of the time-series when a particular class of models is used. In this context, what the analyst will try to choose is not just a **function** $f(x_t)$, but a **functional** F that maps historical data D_t into such a function (and will be applied to many consecutive time steps, as more historical data is gathered).

The objective of this paper is three-fold. First, establish a distinction between two apparently close null hypotheses: (1) no relationship between the inputs and the outputs and (2) no better predictive power of a given model with respect to a naive model. Second, we propose methods to test the second null hypothesis. Third, we show that these two types of tests yield very different results on commonly studied financial returns data.

In section 2, we present the classical notion of generalization error, empirical risk minimization, and cross-validation, and we extend these notions to sequential data. We also present the notion of a "naive model" used to establish a comparison benchmark (and null hypotheses).

In section 3, we introduce a measure of forecastibility, R_o, that is related to the one defined by Granger and Newbold [Granger and Newbold, 1976]. Its estimator, $\hat{R}_o$, is presented.

Section 4 describes the financial time-series data and presents some preliminary results.

In section 5, we test the hypothesis of non-relation between the inputs and the outputs. Although this hypothesis is not really what we want to test, it allows us to nicely introduce some difficult issues with the data at hand, such as dependency induced by overlapping, and the type of methodologies used later on, including the bootstrap. Furthermore, we will compare the results obtained on that test and the

test concerning generalization error. Concerning the no-dependency test, we perform a simulation study to compare the power of in-sample and out-of-sample statistics.

Section 6 aims at assessing whether inputs may be used to produce forecasts that would outperform a naive forecast. Following section 3, we test if $R_o = 0$ against the alternative that it is positive. We do so for different prediction horizons, using the statistic $\hat{R}_0$ and various bootstrap schemes. The results are compared to those obtained when trying to reject the null hypothesis of no dependency, allowing us to show a notable distinction between the absence of relationship between inputs and outputs and the inability of inputs to forecast outputs.

We conclude the paper with a discussion of the results in section 7.

2 Expected Risk and Sequential Validation

This section reviews notions from the generalization theory of Vapnik [Vapnik, 1995], and it presents an extension to sequential data of the concepts of generalization error and cross-validation. We also define a "naive" model that will be used as a reference for the R_o statistic.

First let us consider the usual i.i.d. case [Vapnik, 1995]. Let $Z = (X, Y)$ be a random variable with an unknown density $P(Z)$, and let the *training set* D_l be a set of l examples $z_1, \ldots, z_l$ drawn independently from this distribution. In our case, we will suppose that $X \in \mathcal{R}^n$ and $Y \in \mathcal{R}$. Let $\mathcal{F}$ be a set of functions from $\mathcal{R}^n$ to $\mathcal{R}$. A measure of loss is defined which specifies how well a particular function $f \in \mathcal{F}$ performs the generalization task for a particular Z: $Q(f, Z)$ is a functional from $\mathcal{F} \times \mathcal{R}^{n+1}$ to $\mathcal{R}$. For example, in this paper we will use the quadratic error $Q(f, Z) = (Y - f(X))^2$. The objective is to find a function $f \in \mathcal{F}$ that minimizes the expectation of the loss $Q(f, Z)$, that is the **generalization error** of f:

$$G(f) = E[Q(f, Z)] = \int Q(f, z)P(z)dz \tag{1}$$

Since the density $P(z)$ is unknown, we can't measure or even less minimize $G(f)$, but we can minimize the corresponding **empirical error**:

$$G_{emp}(f, D_l) = \frac{1}{l} \sum_{z_i \in D_l} Q(f, z_i) = \frac{1}{l} \sum_{i=1}^{l} Q(f, z_i). \tag{2}$$

When f is chosen independently of D_l, this is an unbiased estimator of G(f), since $E[G_{emp}(f, D_l)] = G(f)$. **Empirical risk minimization** [Vapnik, 1982, Vapnik, 1995] simply chooses

$$f = F(D_l) = \text{argmin}_{f \in \mathcal{F}} G_{emp}(f, D_l)$$

where $F(D_l)$ is the functional that maps a data set into a decision function.

An empirical estimate of $G(F(D))$, the generalization error of a functional F, can be obtained by partitioning the data in two subsets: a *training subset* D_1 to pick $f = F(D_1) \in \mathcal{F}$ which minimizes the empirical error in D_1, and a *held-out or test subset* D_2 which gives an unbiased estimate of $G(F(D_1))$. The latter is a slightly pessimistic estimate of $G(F(D))$, the generalization error associated to a functional F when applied to $D = D_1 \cup D_2$, and may be poor for small data sets. When there is not much data, it is preferable but computationally more expensive to use the K-fold cross-validation procedure [Bishop, 1995, Efron and Tibshirani, 1993].

However, in the case where the data are not i.i.d., the results of learning theory are not directly applicable, nor are the procedures for estimating generalization error.

Consider a sequence of points z_1, z_2, ..., with $z_t \in \mathcal{R}^{n+1}$, generated by an unknown process such that the z_t's may be dependent and have different distributions. Nevertheless, at each time step t, in order to make a prediction, we are allowed to choose a function f_t from a set of functions $\mathcal{F}$ using the past observations $z_1^t = (z_1, z_2, \ldots, z_t)$, i.e., we choose $f_t = F(z_1^t)$. In our applications z_t is a pair (x_t, y_t) and the functions $f \in \mathcal{F}$ take an x as input to take a decision that will be evaluated against a y through the loss function $Q(f, z)$, with $z = (x, y)$. In this paper, we consider the quadratic loss

$$Q(f, Z_t) = Q(f, (X_t, Y_t)) = (Y_t - f(X_t))^2.$$

We then define the expected generalization error G_t for the decision at time t as

$$G_t(f) = E[Q(f, Z_{t+h})|Z_1^t] = \int Q(f, z_{t+h}) P_{t+h}(z_{t+h}|Z_1^t) dz_{t+h}. \tag{3}$$

Here we call h the **horizon** because it corresponds to the prediction horizon in the case of prediction problems. More generally it is the number of time steps from a decision to the time when the quality of this decision can be evaluated. The objective of learning is to find, on the basis of empirical data z_1^t, the function $f \in \mathcal{F}$ which has the lowest expected generalization error $G_t(f)$.

The process Z_t may be non-stationary, but as long as the generalization errors made by a good model are rather stable in time, we believe that one can use the data z_1^t to pick a function which has worked well in the past and hope it will work well in the future.

We will extend the above empirical and generalization error (equations 2 and 1). However we consider not the error of a single function f but the error associated with a functional F which maps a data set $D_t = z_1^t$ into a function $f \in \mathcal{F}$.

Now let us first consider the empirical error which is the analogue for non *i.i.d.* data of the K-fold cross-validation procedure. We call it the **sequential cross-validation**

procedure and it measures the out-of-sample error of the functional F as follows:

$$C_T(F, z_1^T, h, M) = C_T(F, z_1^T) = \frac{1}{T-M-h+1} \sum_{t=M}^{T-h} Q(F(z_1^t), z_{t+h}) \tag{4}$$

where $f_t = F(z_1^t)$ is the choice of the training algorithm using data z_1^t (see equation 7 below), and $M > 0$ is the minimum number of training examples required for $F(z_1^M)$ to provide meaningful results.

We define the generalization error associated to a functional F for decisions or predictions with a horizon h as follows:

$$\begin{aligned} E_{Gen}(F) &= E[C_T(F, z_1^T)] = \int \frac{1}{T-M-h+1} \sum_{t=M}^{T-h} Q(F(z_1^t), z_{t+h}) P(z_1^T) dz_1^T \\ &= \frac{1}{T-M-h+1} \sum_{t=M}^{T-h} E[G_t(F(Z_1^t))] \end{aligned} \tag{5}$$

where $P(z_1^T)$ is the probability of the sequence Z_1^T under the generating process. In that case, we readily see that (4) is the empirical version of (5), that is (4) estimates (5) by definition. In the case of the quadratic loss, we have

$$E_{Gen}(F) = \frac{\sum_{t=M}^{T-h} E\Big[Var[F(Z_1^t)(X_{t+h}) - Y_{t+h} | X_1^T] + E^2[F(Z_1^t)(X_{t+h}) - Y_{t+h} | X_1^T]\Big]}{T-M-h+1} \tag{6}$$

To complete the picture, let us simply mention that the functional F may be chosen as

$$F(z_1^t) = \text{argmin}_{f \in \mathcal{F}} R(f) + \sum_{s=1}^{t} Q(f, z_s) \tag{7}$$

where $R(f)$ might be used as a regularizer, to define a preference among the functions of $\mathcal{F}$, e.g., those that are smoother.

For example, consider a sequence of observations $z_t = (x_t, y_t)$. A simple class of functions $\mathcal{F}$ is the class of "constant" functions, which do not depend on the argument x, i.e., $f(x) = \mu$. Applying the principle of empirical risk minimization to this class of function with the quadratic loss $Q(f, (x_t, y_t)) = (y_t - f(x_t))^2$ yields

$$f_t^{const} = F^{const}(z_1^t) = \text{argmin}_\mu \sum_{s=1}^{t} (y_s - \mu)^2 = \bar{y}_t = \frac{1}{t} \sum_{s=1}^{t} y_s, \tag{8}$$

the historical average of the y's up to the current time t. We call this "unconditional" predictor the *naive model*, and its average out-of-sample error is $C_T(F^{const}, z_1^T) = \frac{1}{T-M-h+1} \sum_{t=M}^{T-h} (\bar{y}_t - y_{t+h})^2$.

3 Comparison of generalization abilities

To compare the generalization ability of two functionals F_1 and F_2, let us introduce two measures of performance [3]

$$D_o = D_o(F_1, F_2) = E_{Gen}(F_2) - E_{Gen}(F_1) = E[C_T(F_2, z_1^T)] - E[C_T(F_1, z_1^T)], \quad (9)$$

$$R_o = R_o(F_1, F_2) = 1 - \frac{E_{Gen}(F_1)}{E_{Gen}(F_2)} = 1 - \frac{E[C_T(F_1, z_1^T)]}{E[C_T(F_2, z_1^T)]} = \frac{D_o(F_1, F_2)}{E_{Gen}(F_2)}, \quad (10)$$

where $E_{Gen}(.)$, $C_T(.,.)$ were discussed in the previous section. Typically, we will consider cases where $F_2 \subset F_1$. For example, $F_2 = F^{const}$ could serve as benchmark to a more complexe functional F_1. R_o and D_o will be negative, null or positive according to whether the functional F_1 generalizes worse, as well or better than F_2. Related definitions of measure of forecast accuracy have been proposed by many authors. See [Diebold and Lopez, 1996] for a review and [Diebold and Kilian, 1997] for a general discussion. Note that, unlike D_o, R_o is unitless and therefore easier to interpret.

Broadly speaking, for an arbitrary F, **when $R_o(F, F^{const})$ or $D_o(F, F^{const})$ is positive it means that there is a dependency between the inputs and the outputs**. In other words, when there is no dependency and we use a model (F) with more capacity (e.g., degrees of freedom, $F \supset F^{const}$) than the naive model, then R_o **will be negative**. The converse is not true, i.e. $R_o < 0$ does not imply no dependency but suggests that the dependency (signal) is small relative to overall random variation (noise). So in cases where the "signal-to-noise-ratio" is small, it may be preferable not to try to capture the signal to make predictions.

The empirical versions or estimators of R_o and D_o are the statistics

$$\hat{D}_o = \hat{D}_o(F_1, F_2) = C_T(F_2, z_1^T) - C_T(F_1, z_1^T) = \frac{\sum_{t=M}^{T-h}(e_t^{F_2})^2 - \sum_{t=M}^{T-h}(e_t^{F_1})^2}{T - M - h + 1} \quad (11)$$

and

$$\hat{R}_o = \hat{R}_o(F_1, F_2) = 1 - \frac{C_T(F_1, z_1^T)}{C_T(F_2, z_1^T)} = 1 - \frac{\sum_{t=M}^{T-h}(e_t^{F_1})^2}{\sum_{t=M}^{T-h}(e_t^{F_2})^2}, = \frac{\hat{D}_o(F_1)}{C_T(F_2, z_1^T)} \quad (12)$$

where

$$e_t^F = y_{t+h} - F(z_1^t)(x_{t+h})$$

denotes the prediction error made on y_{t+h} by the functional F. This empirical $\hat{R}_o$ ($\hat{D}_o$) is a "noisy" estimate (due to the finite sample), and thus might be positive even when R_o (D_o) is negative (or vice-versa). While $E[\hat{D}_o] = D_o$, $E[\hat{R}_o] \neq R_o$ because the

[3] Arguments of R_o and D_o will often be omitted to ease notation.

expection of a ratio is not equal to the ratio of expectations. In fact we should expect $\hat{R}_o$ to underestimate R_o. This means that $\hat{R}_o$ tends to be a conservative estimate of R_o, which is not undesirable. It is therefore important to analyze how "noisy" this estimate is in order to conclude on the dependency between the inputs and the outputs. This matter will be addressed in a later section.

An example may clarify all of the above. Take $n = 1$ and let $\mathcal{F}^{lin}$ be the set of affine functions, i.e. linear models $f(x) = \alpha + \beta x$. Sticking with the quadratic loss with no regularization, we have that

$$f_t^{lin}(x) = F^{lin}(z_1^t)(x) = \hat{\alpha}_t + \hat{\beta}_t x,$$

where $(\hat{\alpha}_t, \hat{\beta}_t)$, minimizing

$$\sum_{s=1}^{t}(y_s - \alpha - \beta x_s)^2,$$

are the least square estimates of the linear regression of y_s on x_s, $s = 1, \ldots, t$, and rely only on data known up to time t, i.e. z_1^t. We thus have

$$e_t^{F^{const}} = y_{t+h} - F^{const}(z_1^t)(x_{t+h}) = y_{t+h} - \bar{y}_t,$$

$$e_t^{F^{lin}} = y_{t+h} - F^{lin}(z_1^t)(x_{t+h}) = y_{t+h} - \hat{\alpha}_t - \hat{\beta}_t x_{t+h}.$$

If we assume that the Z_t's are independent with expectation $E[Y_t|x_t] = \alpha + \beta x_t$ and variance $Var[Y_t|x_t] = \sigma^2$, then (6) yields

$$(T - M - h + 1)E_{Gen}(F^{const}) = \sigma^2 \sum_{t=M}^{T-h}\left[1 + \frac{1}{t}\right] + \beta^2 \sum_{t=M}^{T-h} E[(X_{t+h} - \bar{X}_t)^2]$$

and

$$(T - M - h + 1)E_{Gen}(F^{lin}) = \sigma^2 \sum_{t=M}^{T-h}\left[1 + \frac{1}{t}\right] + \sigma^2 \sum_{t=M}^{T-h} E\left[\frac{(X_{t+h} - \bar{X}_t)^2}{\sum_{s=1}^{t}(X_s - \bar{X}_t)^2}\right],$$

where $\bar{X}_t = t^{-1}\sum_{s=1}^{t} X_s$ is the mean of the X's up to X_t. We then see that $R_o(F^{lin}, F^{const})$ is negative, null or positive according to whether $\frac{\beta^2}{\sigma^2}$ is smaller, equal or greater than

$$\theta = \frac{\sum_{t=M}^{T-h} E\left[\frac{(X_{t+h} - \bar{X}_t)^2}{\sum_{s=1}^{t}(X_s - \bar{X}_t)^2}\right]}{\sum_{t=M}^{T-h} E[(X_{t+h} - \bar{X}_t)^2]}. \tag{13}$$

This illustrates the comment made earlier regarding the fact that $R_o < 0$ means that the "signal-to-noise-ratio" ($\frac{\beta^2}{\sigma^2}$ here) is too small for F^{lin} to outperform F^{const}. Thus if the true generating model has $\frac{\beta^2}{\sigma^2} < \theta$, a model trained from a class of models with

$\beta = 0$ (the naive model) should be chosen for its better generalization, rather than a model from a class of models that allows $\beta \neq 0$. It also illustrates the point made in the introduction that when the amount of data is finite, choosing a model according to its expected generalization error may yield a different answer than choosing a model that is closest to the true generating model. See also [Vapnik, 1982] (section 8.6) for an example of the difference in out-of-sample generalization performance between the model obtained when looking for the true generating model versus choosing the model which has a better chance to generalize (in this case using bounds on generalization error, for polynomial regression).

Let us now consider a more complex case where the distribution is closer to the kind of data studied in this paper. If we assume that $E[Y_t|x_1^T] = \alpha + \beta x_t$ and $Var[Y_t|x_1^T] = \sigma^2$ with $Cov[Y_t, Y_{t+k}|x_1^T] = 0$ whenever $|k| \geq h$, then (6) yields

$$(T - M - h + 1)E_{Gen}(F^{const}) = \sum_{t=M}^{T-h}(\sigma^2 + E[Var[\bar{Y}_t|X_1^T]]) + \beta^2 \sum_{t=M}^{T-h} E[(X_{t+h} - \bar{X}_t)^2]$$

and

$$(T - M - h + 1)E_{Gen}(F^{lin}) = \sum_{t=M}^{T-h}(\sigma^2 + E[Var[\bar{Y}_t + \hat{\beta}_t(X_{t+h} - \bar{X}_t)|X_1^T]]).$$

We then see that R_o is negative, null or positive according to whether $\frac{\beta^2}{\sigma^2}$ is smaller , equal or greater than

$$\theta = \frac{\sigma^{-2} \sum_{t=M}^{T-h} E\left[Var[\bar{Y}_t + \hat{\beta}_t(X_{t+h} - \bar{X}_t)|X_1^T] - Var[\bar{Y}_t|X_1^T]\right]}{\sum_{t=M}^{T-h} E[(X_{t+h} - \bar{X}_t)^2]}. \tag{14}$$

Note that it can be shown that the above numerator is free of σ as it involves only expectations of expressions in X_t's (like the denominator).

4 The financial data and preliminary results

Experiments on the out-of-sample statistics and related in-sample statistics where performed on a financial time-series. The data is based on the daily total return, including capital gain as well as dividends, for the Toronto Stock Exchange TSE300 index, starting in January 1982 up to July 1998. The total return series $TR_t, t = 0, 1, \ldots, 4178$, can be described as the result at day t of an initial investment of 1 dollar and the reinvestment of all dividends received.

We construct, for different values of h, the log-return series on a horizon h

$$r_t(h) = \log\left(\frac{TR_t}{TR_{t-h}}\right) = \log(TR_t) - log(TR_{t-h}). \tag{15}$$

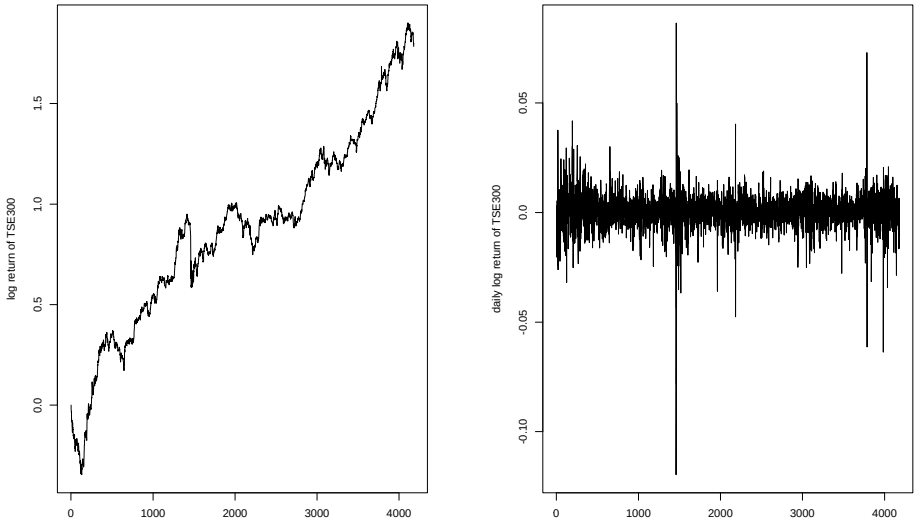

Figure 1:
Left: daily logarithm of TSE300 index from January 1982 to end of July 1998.
Right: daily log returns of TSE300 for the same period

Thus $r_t(h)$ represents the logarithm of the total return at day t of the past h day(s).

There are 4179 trading days in the sample. We consider that there are twenty-one trading days per "month" or 252 trading days per year. The real number of trading days, where the trading activities can occur, can vary slightly from month to month, depending on holidays or exceptional events, but 21 is a good approximation if we want to work with a fixed number of trading days per month. A horizon of H months will mean $h = H \times 21$ days.

Using and predicting returns on a horizon greater than the sampling period creates an overlapping effect. Indeed, upon defining the **daily log-returns**

$$r_t = r_t(1), t = 1, \ldots, 4178,$$

we can write

$$\begin{aligned} r_t(h) &= \log(TR_t) - \log(TR_{t-h}) = \sum_{s=t-h+1}^{t} (\log(TR_s) - \log(TR_{s-1})) \\ &= \sum_{s=t-h+1}^{t} r_s \end{aligned} \tag{16}$$

as a moving sum of the r_t's.

log-returns	skewness	kurtosis
daily	-1.22 (0.04)	33.17 (0.08)
monthly	-1.13 (0.17)	10.63 (0.35)
quarterly	-0.40 (0.30)	3.93 (0.60)

Table 1:
Sample skewness and sample kurtosis of TSE300 daily, monthly and quarterly log-returns. The statistics and their standard deviations (shown in parenthesis) have been computed according to formulas described in [Campbell et al., 1997].

We will work on monthly returns as it has been suggested from empirical evidence [Campbell et al., 1997, Fama and French, 1988] that they can be useful for forecasting, while such results are not documented for daily returns. So our horizons will be multiples of 21 days. Data are slightly better behaved when we take monthly returns instead of daily ones. For instance, the daily return series is far from being normally distributed. It is known that stock indices return distributions present more mass in their tails than the normal distribution [Campbell et al., 1997]. But returns over longer horizons get closer to normality, thanks to equation 16 and the central limit theorem. For example, table 1 shows the sample skewness and kurtosis for the daily, monthly and quarterly returns. We readily notice that these higher moments are more in line with those of the normal distribution (skewness=0, kurtosis=3) when we consider longer term returns instead of daily returns.

Table 1 is the first illustration of the touchy problem of the overlapping effect. For instance, you will notice that the standard deviation are not the same for daily and monthly returns. This is because the daily returns statistics are based on $r_1, \ldots, r_{4178}$, whereas their monthly counterparts are based on $r_{21}(21), r_{42}(21), \ldots r_{21\times 198}(21)$, that is approximatively 21 times fewer points than in the daily case. The reason for this is that we want independent monthly returns. If we assumed that the daily returns were independent, then monthly returns would have to be at least one month apart to be also independent. For instance, $r_{21}(21)$ and $r_{40}(21)$ would not be independent as they share r_{20} and r_{21}. Therefore, if we want to access independence of successive monthly returns, we have to compute the correlation coefficient between $r_{t+21}(21)$ and $r_t(21)$, or between $r_{t+h}(h)$ and $r_t(h)$ for more general h's.

Figure 2 left shows the square of the correlation coefficient obtained on the TSE data for $H = 1, 2, \ldots, 24$. Figure 2 right depicts the values of $\hat{R}_o$ with $z_t = (r_{t+h-1}(h), r_{t+2h-1}(h))$ obtained on the same data. It measures the ability of the past H month return to forecast the future H month return. According to the first plot there appears to be little relationship between past and future returns except, perhaps, when we aggregate the returns on a period of about one year ($H = 12$). Figure 2 right tells a similar story: at best, predictability of future returns seems possible only for yearly returns or so. But how can we decide (formally) if there is

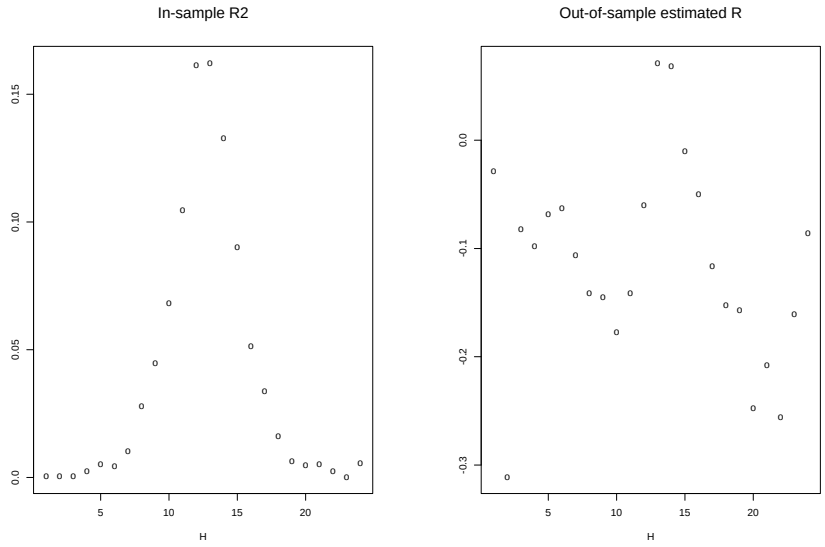

Figure 2:
Left: The evolution of squared correlation with the aggregation period suggests that the stronger input/output relation is for the aggregation period of around a year.
Right: Evolution of $\hat{R}_o$ with the aggregation period.

a relationship between past and future returns, and if such a relationship might be useful for forecasting? This will be the goal of the next section.

5 Testing the hypothesis of no relation between Y and X

Consider testing the hypothesis that there is no relationship between successive returns of horizon h, i.e. $H_0 : E[r_t(h)|r_{t-h}(h)] = \mu$. Note that $r_t(h)$ and $r_{t-h}(h)$ do not overlap but are contiguous h day returns. To put it in section 3's notation, we have $x_t = r_{t+h-1}(h)$ and $y_t = r_{t+2h-1}(h)$, so that, for instance, $x_1 = r_h(h)$ is the first observable x. We wish to test $E[Y_t|x_t] = \mu$.

As mentioned in the introduction, this hypothesis is not what we are actually interested in, but what we do in this section proves to be useful in section 6 as it allows us to introduce the bootstrap, among other things.

To perform a test of hypothesis, one needs a statistic with a behavior that depends on whether H_0 is true or false. We will mainly consider two statistics here. First we have R_o that will take smaller values under H_0 than otherwise. The other approach to testing H_0 is to notice that if $E[r_t(h)|r_{t-h}(h)]$ does not depend on $r_{t-h}(h)$ then

the correlation between $r_{t-h}(h)$ and $r_t(h)$ is null, $\rho(r_t(h), r_{t-h}(h)) = 0$. Thus we will use $\hat{\rho}(r_t(h), r_{t-h}(h))$, an estimator of $\rho(r_t(h), r_{t-h}(h))$, to test H_0 as it will tend to be closer to 0 under H_0 than otherwise.

The second thing needed in a test of hypothesis is the distribution of the chosen statistic under H_0. This may be obtained from theoretical results or approximated from a bootstrap as explained later. In the case of $\hat{\rho}(r_t(h), r_{t-h}(h))$, we do have such a theoretical result [Bartlett, 1946, Anderson, 1984, Box and Jenkins, 1970]. First let us formally define

$$\hat{\rho}(r_t(h), r_{t-h}(h)) = \frac{\sum_{2h}^{T}(r_t(h) - \bar{r}(h))(r_{t-h}(h) - \bar{r}(h))}{\sum_{h}^{T}(r_t(h) - \bar{r}(h))^2}, \tag{17}$$

with $\bar{r}(h)$ being the sample mean of $r_h(h), \ldots, r_T(h)$. Assuming that the r_t's are independent and identically distributed with finite variance then

$$\sqrt{T-h+1}(\hat{\rho}(r_t(h), r_{t-h}(h)) - \rho(r_t(h), r_{t-h}(h))) \longrightarrow N(0, W)$$

with

$$W = \sum_{v=1}^{\infty}(\rho_{v+h} + \rho_{v-h} - 2\rho_h\rho_v)^2, \tag{18}$$

where ρ_k stands for $\rho(r_{t+k}(h), r_t(h))$. If the r_t's are uncorrelated with constant variance and the $r_t(h)$ are running sums of r_t's as shown in equation 16, then

$$\rho_k = \frac{(h - |k|)_+}{h} \tag{19}$$

where $u_+ = \max(u, 0)$. Therefore we have

$$W = \sum_{v=1}^{2h-1}\rho_{v-h}^2 = \sum_{v=1-h}^{h-1}\rho_v^2 = 1 + 2h^{-2}\sum_{v=1}^{h-1}(h-v)^2 = 1 + \frac{(h-1)(2h-1)}{3h}$$

where the identity $1^2 + 2^2 + 3^2 + \ldots + N^2 = \frac{N(N+1)(2N+1)}{6}$ was used in the last equality. Large values of $\sqrt{\frac{T-h+1}{W}}|\hat{\rho}(r_t(h), r_{t-h}(h))|$ are unfavorable to H_0 and their significance are obtained from a $N(0,1)$ table.

The distributions of our out-of-sample statistics are unknown. However we may find an approximation by simulation (bootstrap). So we have to generate data from the hypothesis $H_0 : E[Y_t|x_t] = \mu$ (i.e. do not depend on x_t). This can be done in at least four ways.

1. Generate a set of independent r_t's and compute the $Y_t = r_{t+2h+1}(h)$'s and the $x_t = r_{t+h+1}(h)$'s in the usual way.

2. Keep the Y_t obtained from the actual data, but compute the x_t's as suggested in 1.

3. Keep the x_t's obtained from the actual data, but compute the Y_t as suggested in 1.

4. Generate a set of independent r_t's and compute the Y_t's. Then generate another set of r_t's independently of the first set and compute the x_t's.

The generation of the r_t's may come from the empirical distribution of the actual r_t's (i.e. re-sampling with replacement) or another distribution deemed appropriate. We have considered both the empirical distribution and the $N(0,1)$ distribution [4].

We believe that the generation scheme 1 is the most appropriate here since it looks more like the way the original data was treated: Y_t and x_t obtained from a single set of r_t's.

Once we have chosen a simulation scheme, we may obtain as many (B, say) samples as we want and thus get B independent realizations of the statistic $\hat{R}_o$. We then check if the out-of-sample statistic will take values that are large even in this case, compared to the value observed on the original data series. Formally, compute p-value= $\frac{A}{B}$ where A is the number of simulated $\hat{R}_o$ greater or equal to the $\hat{R}_o$ computed on the actual data. This measures the plausibility of H_0; small values of p-value indicate that H_0 is not plausible in the light of the actual data observed. Another way to use the bootstrap values of $\hat{R}_o$ is to assume that the distribution of $\hat{R}_o$ under H_0 is $N(\hat{E}[\hat{R}_o], \hat{V}[\hat{R}_o])$ where $\hat{E}[\hat{R}_o]$ and $\hat{V}[\hat{R}_o]$ are the sample mean and the sample variance of the B bootstrap values of $\hat{R}_o$. Comparing the actual $\hat{R}_o$ to this distribution yields the normalized bootstrap p-value.

For the scheme 1 method we simply compute the p-value of the observed $\hat{R}_o$ under the null hypothesis of no relationship between the inputs and the outputs, using the empirical histogram of this statistic over the bootstrap replications. When the p-value is very small, a more meaningful quantity might be the mean and the standard deviation of the statistic over the bootstrap replications to provide a z-statistic.

Of course, this bootstrap approach may be used even in the case where the (asymptotic) distribution of a statistic is known. Therefore, we will compute bootstrap p-values for the statistic $\hat{\rho}(r_t(h), r_{t-h}(h))$ as well as its theoretical p-value for comparison purposes.

[4] Since $\hat{R}_o$, just like the usual in-sample R^2, is location-scale invariant, we don't have to bother about matching the mean and variance of the actual series.

T	ρ	R_o	D_o
250	[-0.106,0.101]	$(-\infty$, 0.0012]	$(-\infty$, 0.017]
500	[-0.075,0.072]	$(-\infty$, 0.0009]	$(-\infty$,-0.016]
1000	[-0.052,0.050]	$(-\infty$,-0.0011]	$(-\infty$,-0.043]
2000	[-0.038,0.035]	$(-\infty$,-0.0009]	$(-\infty$,-0.069]
4000	[-0.026,0.025]	$(-\infty$,-0.0006]	$(-\infty$,-0.096]
8000	[-0.019,0.018]	$(-\infty$,-0.0004]	$(-\infty$,-0.119]

Table 2:
Empirical critical points of three statistics estimated on series of different length.

5.1 Results on artificial data

In order to study different properties of in-sample and out-of-sample statistics, we have generated artificial data and tested the null hypothesis of no relationship on them. In this way, we can compare the power of the statistics on the same data set, where the hypothesis behind the use of the autocorrelation statistic is verified.

We chose an autoregressive process of order 1,

$$y_t = \beta y_{t-1} + \epsilon_t$$

for which we vary the coefficient β of auto-regression from a range of values between 0 and 0.1 and where ϵ_t is drawn from a normal distribution $N(0, \frac{1}{5})$. We conduct the tests on the null hypothesis for series of lengths in the set $\mathcal{T} = \{250, 500, 1000, 2000, 4000, 8000\}$.

We first generated, for each value of T in $\mathcal{T}$, five thousand series for which $\beta = 0$. For each of these series we construct the empirical distribution of 3 statistics, namely the autocorrelation $\hat{\rho}$ (equation 17), the out-of-sample $\hat{R}_o$ and $\hat{D}_o$.

From these empirical distributions, we estimated the "acceptance" region at significance level 10%, say $[L_{5\%}, H_{5\%}]$ for $\hat{\rho}$ and $(-\infty, H_{10\%}]$ for the out-of-sample statistics $\hat{D}_o$ and $\hat{R}_o$. For the out-of-sample statistics, we chose $M = 50$ for the minimum number of training examples (see equation 4). The values of these critical points are presented in table 2.

Having established the critical points at 10%, we now want to study the power of these tests, i.e. how each statistic is useful to reject the null hypothesis when the null hypothesis is false. For this goal, we generated two thousand series for different value of β, ranging from 0 to 0.1. We estimated on these series the value of the three statistics considered in table 2, and computed for the different values of β the number of times each of these statistics are outside the interval delimited by the critical values. The results are presented in table 3. Note that proportions in table 3 have standard deviations of at most $\frac{1}{2\sqrt{2000}} = 1.1\%$. We can observe from this table that the out-

T	$\hat{s}$	β					
		0	0.02	0.04	0.06	0.08	0.1
250	ρ	0.10	0.14	0.17	0.24	0.34	0.42
250	R_o	0.10	0.14	0.16	0.23	0.32	0.39
250	D_o	0.10	0.14	0.16	0.23	0.32	0.39
500	ρ	0.09	0.15	0.22	0.38	0.54	0.73
500	R_o	0.11	0.14	0.21	0.36	0.51	0.68
500	D_o	0.11	0.14	0.21	0.36	0.51	0.69
1000	ρ	0.11	0.17	0.36	0.61	0.82	0.94
1000	R_o	0.11	0.16	0.32	0.57	0.77	0.90
1000	D_o	0.11	0.16	0.33	0.57	0.77	0.90
2000	ρ	0.10	0.25	0.58	0.87	0.98	1.00
2000	R_o	0.11	0.22	0.54	0.82	0.96	0.99
2000	D_o	0.11	0.23	0.53	0.82	0.96	0.99
4000	ρ	0.11	0.37	0.81	0.98	1.00	1.00
4000	R_o	0.11	0.36	0.78	0.97	1.00	1.00
4000	D_o	0.11	0.36	0.78	0.98	1.00	1.00
8000	ρ	0.10	0.53	0.98	1.00	1.00	1.00
8000	R_o	0.10	0.49	0.96	1.00	1.00	1.00
8000	D_o	0.10	0.49	0.96	1.00	1.00	1.00

Table 3:
Power of 3 statistics for the hypothesis $H_0 : \beta = 0$ as a function of T and β.

of-sample statistics $\hat{R}_o$ and $\hat{D}_o$ are less powerful than in-sample statistics for the test of $H_o : \beta = 0$. For $\beta = 0$, powers are around 10% as they should be. It would appear from these results that when we want to test against the null hypothesis of no dependency, the classical in-sample tests provide more power. But we must underline again that this is not the null hypothesis of interest here.

5.2 Discussion of the results on financial data

In all cases $B = 1000$ bootstrap replications were generated and the out-of-sample statistic was computed on each of them with $M = 50$, yielding distributions of $\hat{R}_o$ for the null hypothesis of no relationship between input and output.

For $\hat{\rho}(r_t(h), r_{t-h}(h))$, the pure bootstrap p-values and normalized bootstrap p-values agree well, as shown in table 4, suggesting that the distribution of $\hat{\rho}$ is approximatively normal. However, we note a discrepancy between the theoretical p-values and the bootstrap p-values, suggesting that the asymptotic mean and/or variance is not a

proper approximation for small samples. In fact, the mean of the $\hat{\rho}$ are supposed to be 0, which is not the case for finite sample. When using the value of mean obtained by bootstrap to the p-value obtained by the three method are much closer. Regarding $\hat{R}_0$, we see (table 4) that a similar pattern is observed for the positive $\hat{R}_0$. The pure bootstrap p-values seem to indicate a possible dependence of the near one year return on the past year return. Also, in this case, the empirical distributions of the $\hat{R}_0$ are not normal, the observed skewness on these distribution are systematically negative with values around -4, hence the normalized p-values should not be trusted. The theoretical p-values for the out-of-sample statistics are not known.

The table 4 also presents the results of the test conducted on the null hypothesis *no relationship between inputs and outputs* using the statistic D_o. This test statistics **rejects even more strongly the null hypothesis of no linear dependency** than the test based on $\hat{R}_o$.

6 Test of $H_0 : R_o = 0$

Here we attack the problem we are actually interested in: assessing whether generalizations based on past returns are better than the generalizations of an alternative model, here the *naive* (constant) model. We consider linear forecasts, so that we want to know if F^{lin} generalizes better than F^{naive}.

Its distribution not being known, we will have to turn to the bootstrap method and simulate values of $\hat{R}_o$ computed on samples generated under $H_0 : R_o = 0$ (which is equivalent to $D_o = 0$). Strictly speaking, we want to solve the equation $D_o(\beta) = 0$. We can proceed analytically, as we will do when the Y_t and the X_t are different, or numerically when the Y_t are autoregressive. We are using the statistic D_o because its estimator $\hat{D}_o$ is without bias.

Consider first the case where

$$E[Y_t|X_t] = \alpha + \beta X_t, \tag{20}$$

with the X_t is an external covariate series (generated independently from the Y_t series, while the Y_t's are generated conditional on the X_t series).

We saw earlier that this amounts to $\frac{\beta^2}{\sigma^2}$ being equal to the ratio shown in (14). If we let the Y_t's (given x_1^T) have the correlation structure shown in (19), we have

$$\begin{aligned} E[Var[\bar{Y}_t|X_1^T]] &= \frac{\sigma^2}{t^2 h} \sum_{s=1-h}^{h-1} (h - |s|)(t - |s|) \\ &= \frac{\sigma^2}{t^2 h} \left[ht + 2 \sum_{s=1}^{h-1} (h - s)(t - s) \right] \end{aligned}$$

H	$\hat{R}_o$	pbpv	nbpv	$\hat{D}_o$	pbpv	nbpv	$\hat{\rho}$	tpv	pbpv	nbpv
1	-0.03	0.83	0.67	-0.23	0.95	0.93	0.02	0.74	0.60	0.60
2	-0.31	0.99	0.99	-5.10	0.99	1.00	0.02	0.81	0.67	0.67
3	-0.08	0.68	0.51	-1.86	0.84	0.69	-0.02	0.84	0.95	0.98
4	-0.10	0.64	0.46	-3.19	0.83	0.66	-0.05	0.65	0.87	0.87
5	-0.07	0.30	0.31	-2.97	0.62	0.46	-0.07	0.60	0.72	0.72
6	-0.06	0.24	0.29	-3.37	0.52	0.41	-0.06	0.68	0.87	0.89
7	-0.11	0.42	0.39	-6.31	0.70	0.51	-0.09	0.56	0.78	0.75
8	-0.14	0.49	0.39	-8.79	0.69	0.50	-0.15	0.37	0.55	0.51
9	-0.15	0.47	0.39	-8.51	0.61	0.45	-0.18	0.31	0.48	0.46
10	-0.18	0.52	0.43	-9.65	0.57	0.45	-0.22	0.24	0.38	0.38
11	-0.14	0.38	0.35	-7.63	0.39	0.35	-0.26	0.18	0.27	0.26
12	-0.06	0.15	0.25	-3.48	0.17	0.27	-0.32	0.12	0.21	0.20
13	0.07	0.02	0.14	4.57	0.01	0.22	-0.32	0.14	0.20	0.21
14	0.07	0.04	0.14	4.54	0.03	0.17	-0.28	0.21	0.31	0.31
15	-0.01	0.10	0.19	-0.70	0.10	0.23	-0.23	0.33	0.58	0.55
16	-0.05	0.13	0.24	-3.43	0.14	0.25	-0.17	0.48	0.75	0.71
17	-0.11	0.24	0.29	-8.04	0.27	0.29	-0.13	0.58	0.99	0.96
18	-0.15	0.30	0.31	-11.02	0.31	0.34	-0.09	0.73	0.82	0.86
19	-0.16	0.28	0.32	-12.15	0.30	0.32	-0.05	0.85	0.74	0.75
20	-0.25	0.44	0.39	-20.01	0.44	0.39	-0.04	0.88	0.70	0.69
21	-0.21	0.37	0.37	-17.17	0.37	0.35	-0.04	0.89	0.71	0.70
22	-0.26	0.45	0.38	-20.67	0.45	0.36	-0.02	0.94	0.54	0.55
23	-0.16	0.31	0.32	-13.26	0.33	0.32	0.02	0.92	0.37	0.38
24	-0.09	0.19	0.28	-7.41	0.22	0.28	0.08	0.79	0.28	0.28

Table 4:
Test of the hypothesis of no relationship between inputs and outputs. Three statistics are used, and for each, pure bootstrap (pbpv) and normalized (nbpv) p-values are computed. For tests based on $\hat{\rho}$, we also present the theoretical (tpv) p-values computed by Bartlett's formula. The test based on the D_o statistic also give a strong evidence against H_o: no relation between inputs and outputs. The empirical version used to estimate D_o does not suffer of a bias like the empirical version of R_o.

$$
\begin{aligned}
&= \frac{\sigma^2}{t^2 h}\left[ht + 2\sum_{s=1}^{h-1} s(t-h+s)\right] \\
&= \frac{\sigma^2}{t^2}\left[ht - \frac{(h^2-1)}{3}\right] \qquad (21)
\end{aligned}
$$

and

$$E[Var[\bar{Y}_t + \hat{\beta}_t(X_{t+h} - \bar{X}_t)|X_1^T]] = \sigma^2 E[c'Vc],$$

where V is a $t \times t$ matrix with $V_{ij} = \frac{(h-|i-j|)_+}{h}$, and c is a $t \times 1$ vector with

$$c_i = \frac{1}{t} + \frac{(X_{t+h} - \bar{X}_t)(X_i - \bar{X}_t)}{\sum_{j=1}^{t}(X_j - \bar{X}_t)^2}, i = 1, \ldots, t.$$

If we let L be a $(t + h - 1) \times t$ matrix with $L_{ij} = I[0 \leq i - j < h]/\sqrt{h}$, then we may write $c'Vc$ as $W'W$ where $W = Lc$. This representation is useful if we need to compute $Var[F^{lin}(Z_1^t)(X_{t+h})|X_1^T] = \sigma^2 c'Vc$ for various values of t as recursive relations may be worked out in W.

Due the location-scale invariance of the $\hat{R}_o$ mentioned earlier, σ^2 and α may be chosen as one pleases (1 and 0, say). The expectations then depend obviously on the process generating the X_t's. The simplest thing to do is to assume that $X_1^T \sim \delta_{x_1^T}$, that is X_1^T can only take the value observed. This makes the expectation easy to work out. Otherwise, these expectations can be worked out via simulations.

Once X_1^T's process, α, β, σ^2 have been chosen, we generate $Z_1^T = (X_1^T, Y_1^T)$ as follows.

1. Generate X_1^T.
2. Generate $\epsilon_1, \ldots, \epsilon_T$ so that the ϵ_t's are independent of X_1^T with $Var[\epsilon_t] = \sigma^2$ and the covariance structure shown in (19). This may be done by generating independent variates with variance equal to $\frac{\sigma^2}{h}$ and take their moving sums with a window of size h.
3. Put $Y_t = \alpha + \beta x_{t-h} + \epsilon_t$.

The bootstrap test of $H_0 : R_o = 0$ could be performed by generating B samples in the way explained above, yielding B bootstrap values of $\hat{R}_o$. These would be used to compute either a pure bootstrap p-value or a normalized bootstrap p-value.

Needless to say that generating data under $H_{01} : R_o = 0$ is more tedious than generating data under H_{02} : no relationship between inputs and outputs. Furthermore the above approach relies heavily on the distributional assumptions of linearity and the given form of covariance, and we would like to devise a procedure that can be extended to non-linear relationships, for example.

To get the distribution of $\hat{R}_o$ under H_{01}, we can consider an approximation saying that the distribution of $\hat{R}_o - R_o$ is the same under H_{01} and H_{02}. We will call this hypothesis the "shifted distribution" hypothesis (note that this hypothesis can only be approximately true since the domain of $\hat{R}_o$ is $(-\infty, 1]$). This means that we are assuming that the distribution of $\hat{R}_o$ under $R_o = 0$ has the same shape as its distribution under $\beta = 0$ but is shifted to the right, since $R_o < 0$ under $H_0 : \beta = 0$. If that was the case, and we are going the test the validity of this approximation later, generating $\hat{R}_o - 0$ under H_{01} would be the same as simulating $\hat{R}_o - R_o$ under H_{02}, which we have done previously without subtracting off R_o. This R_o can be obtained either analytically or estimated from the bootstrap as

$$1 - \frac{\sum_{b=1}^{B} C_T(F^{lin}, Z_1^T(b))}{\sum_{b=1}^{B} C_T(F^{naive}, Z_1^T(b))}.$$

Note, to make the notation clear, that the bootstrap $\hat{R}_o$'s are simply $1 - \frac{C_T(F^{lin}, Z_1^T(b))}{C_T(F^{naive}, Z_1^T(b))}, b = 1, \ldots, B$. From these $\hat{R}_o - R_o$'s, we obtain the bootstrap p-values and the normalized bootstrap p-values as usual. Note that the bootstrap p-values for H_{01} and H_{02} are the proportion of the $\hat{R}_o$'s (generated under H_{02}) that are greater than $\hat{R}_o(observed) + R_o$ and $\hat{R}_o(observed)$ respectively. Since $R_o < 0$ under H_{02}, we see that p-value$(H_{02}) \leq$ p-value(H_{01}).

6.1 Discussion of the results on artificial data

We present in this section results obtained by two approach. One consist to approximate the distribution of $\hat{R}_o$ under the null $\hat{R}_o = 0$ by shifting the distribution of $\hat{R}_o$ already obtained with $\beta = 0$. The second consist to generate the distribution of $\hat{R}_o$ with and approximation of the value of β associate with $\hat{R}_o = 0$.

In table 5 we show the power of the $\hat{R}_o$ statistic to reject the null hypothesis $R_o = 0$ with a 10% level, for various values of β. The simulations are conducted on the artificial data described in section 5.1. Critical values are those of table 2 except that they must be shifted to the right by 0.0084, 0.0052, 0.0032, 0.0019, 0.0011 and 0.0006 for $T = 250, 500, 1000, 2000, 4000, 8000$ respectively. Comparing to table 3 of section 5.1 we see that, as expected, the hypothesis $R_o = 0$ is more difficult to reject than the hypothesis of *no relation between inputs and outputs*.

We also estimate graphically the value of β for which the autoregressive model generate data with $\hat{R}_o = 0$, for a given length of the series. To do this, we plot the values of $\hat{D}_o$ as a function of the values of β, for each length of the series considered previously. We found the value of β for which the autoregressive model will give a $\hat{R}_o = 0$, called "critical *beta*" (β_c), are shown in table 6.

After having obtained the critical value of β, we simulated, for each length in $\mathcal{T}$,

$T \setminus \beta$	0	0.02	0.04	0.06	0.08	0.1
250	0.03	0.05	0.07	0.11	0.18	**0.22**
500	0.02	0.04	0.08	0.16	**0.29**	**0.45**
1000	0.02	0.04	0.13	**0.28**	**0.51**	**0.74**
2000	0.01	0.04	0.20	**0.53**	**0.81**	**0.95**
4000	0.01	0.07	**0.42**	**0.86**	**0.98**	**1.00**
8000	0.01	0.15	**0.74**	**0.99**	**1.00**	**1.00**

Table 5:
With the "shifted distribution" hypothesis, this table show the power of the $\hat{R}_o$ statistic to reject the null hypothesis $R_o = 0$ with a 10% level, for various alternative hypotheses corresponding to different values of β. The values in bold correspond to the conditions where we know that the null hypothesis is false. Observe than we are doing an error in rejecting the null hypothesis when it is true (e.g. T=250 and β=0.6 and 0.8).

T	250	500	1000	2000	4000	8000
β_c	0.093	0.073	0.058	0.043	0.033	0.025

Table 6:
Empirical value of β_c associate with $R_o = 0$ on artificial data.

5000 series according to the autoregressive model. We compare the empirical distributions of $\hat{R}_o$ obtained here to the empirical distributions obtained by shifting the distributions of $\hat{R}_o$ obtained under $\beta = 0$. Under the approximation saying that the distribution of $\hat{R}_o - R_o$ is the same under H_{01} and H_{02}, the distributions must coincide.

We can observe from figure 3 that the empirical distributions of $\hat{R}_o$ obtained by shifting the distribution generated with $\beta = 0$ and the distribution of the $\hat{R}_o$ generated with $\beta = \beta_c$ do not coincide. However, we note that in the right-hand tail, the shifted approximation has less mass than the true distribution. Therefore, *if we are not able to reject H_{01} using the shifted approximation, it would not have been rejected using the true distribution.*

Table 7 also shows that using the shifted approximation the critical points for rejecting the hypothesis $\hat{R}_o = 0$ are underestimated.

In table 8, we see that power of the $\hat{R}_o$ statistic to reject the null hypothesis $R_o = 0$ is lower when we use the empirical distributions generated under $R_o = 0$ than when we use the shifted distributions generated under $\beta = 0$. A test based on the shifted approximation is therefore liberal (yields a lower critical value, and rejects more often than it should), but less than a test based on H_{02} (null hypothesis of no relation). This is clear by looking at the three curves in the area that is in the right-hand side

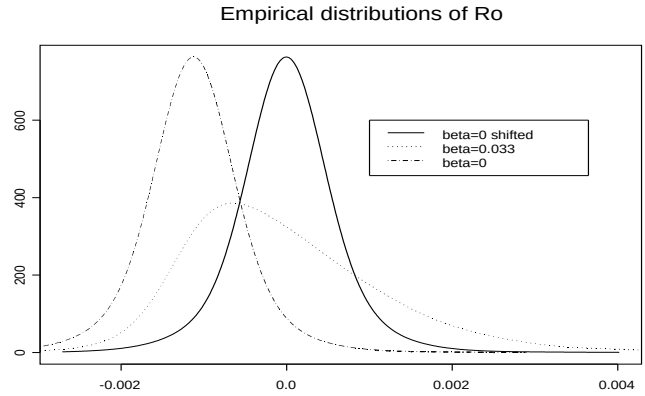

Figure 3: Empirical distributions of $\hat{R}_o$ for $\beta = 0$ shifted and β_c
The length of the series was 4000. *We can observe the discrepancy between the two distributions. However, note that in the right-hand tail, the shifted approximation (full line) has less mass than the true distribution (dotted line).*

T	250	500	1000	2000	4000	8000
$H_{10\%}$ (shifted)	0.0096	0.0061	0.0021	0.0010	0.0005	0.0002
$H_{10\%}$ (empirical)	0.0206	0.0106	0.0058	0.0029	0.0016	0.0008

Table 7:
The critical points corresponding to a coverage surface of 90% for the null hypothesis $\hat{R}_o = 0$ obtained by the "shifted distribution" and by the usage of β_c obtained numerically.

$T \backslash \beta$	0	0.02	0.04	0.06	0.08	0.1
250	0.01	0.01	0.02	0.04	0.07	**0.11**
500	0.01	0.01	0.02	0.05	**0.14**	**0.24**
1000	0.00	0.01	0.04	**0.10**	**0.27**	**0.49**
2000	0.00	0.01	0.06	**0.30**	**0.61**	**0.88**
4000	0.00	0.02	**0.19**	**0.65**	**0.94**	**1.00**
8000	0.00	0.04	**0.52**	**0.97**	**1.00**	**1.00**

Table 8:
Power of the $\hat{R}_o$ statistic to reject the null hypothesis $R_o = 0$ with a 10% level, for various alternative hypotheses corresponding to different values of β. The values in bold correspond to the conditions where we know that the null hypothesis is false. In this case, we observe that we are rejecting the null when it is false.

H	pbpv	H	pbpv	H	pbpv	H	pbpv
1	0.95	7	0.85	13	0.57	19	0.81
2	0.99	8	0.87	14	0.54	20	0.84
3	0.92	9	0.86	15	0.70	21	0.83
4	0.90	10	0.89	16	0.72	22	0.83
5	0.83	11	0.86	17	0.80	23	0.80
6	0.82	12	0.76	18	0.81	24	0.76

Table 9:
P-values for the hypothesis $H_{01} : R_o = 0$ on the financial data. We may suppose this test to be optimistic because we use the "shifted distribution" approximation.

tails, in Figure 3.

With the linear model described in equation 20 and using the analytical method to compute the value of β_c, we obtained the same conclusions on the critical points and the distribution of $\hat{R}_o$.

6.2 Discussion of the results on financial data

For different horizons, we compare the predictive ability of the linear forecast and the naive forecast, i.e. F^{lin} vs F^{naive}. The set of horizons used in the experiments was $H = 1, 3, 6, 9, \ldots, 21, 24$, in number of "months" (21 days), i.e., $h = 21H$. Table 9 gives the p-value of the H_{01} hypothesis $R_o = 0$ using the method based on the "shifted distribution" hypothesis described previously. The p-values are pure histogram counts.

According to this table, there is more than 50% probability to observe values $\hat{R}_o \geq 0.07$ for horizons 13 and 14 under the null hypothesis of $R_o = 0$. Since the true critical values are likely to be larger than those computed using the shifted approximation, we conclude that **we are very far from being able to say that the inputs (past returns) are useful to make a linear forecast, even though we are able to reject the hypothesis** $\beta = 0$ (as shown previously in table 4).

This is a very significant result, since it shows the importance of testing a hypothesis that reflects what we really care about (e.g., out-of-sample error): testing an apparently close hypothesis (no dependency) could yield a very different conclusion!

7 Conclusion

In this paper we have introduced an extension of the notion of generalization error to non-iid data. We also gave a definition of two out-of-sample statistics, R_o and D_o, to compare the forecasting ability of two models in this generalized framework, and we have presented the notion of a naive model used to establish null hypotheses.

The statistic R_o allowed us to establish a link between the signal-to-noise ratio and the particular value $R_o = 0$. We have shown that $R_o \leq 0$ means that the signal-to-noise ratio is too small for the linear functional to outperform the naive model. This does not imply no dependency but indicates that whenever the "signal-to-noise-ratio" is small, it is preferable not to try to capture the signal to make predictions.

We have made a distinction between tests for dependency and for generalization ability and we have described a method, based on bootstrap, to perform these tests.

We have used the proposed bootstrap methods to test the two null hypotheses on simulated data and on real financial data. We have used simulations to better understand the behavior of the in-sample and out-of-sample statistics in a controlled environment. We have observed that the tests based on out-of-sample statistics R_o and D_o had less power to test against the null hypothesis of no dependency than the tests based on an in-sample statistic.

On real financial data, we have observed that we were very far from being able to say that the past returns are useful to make a linear forecast, even though we are able to reject the hypothesis of no relation. This result shows the importance of testing a hypothesis that reflects what we really care about, in that case the out-of-sample error.

In future work, we must find a way to generate the distribution of $\hat{R}_o$ under the hypothesis of no-predictability. That seems for now not trivial when the generating model is more complex than an autoregressive one or a linear model with only two parameters. But this is an important question if we want to avoid making assumptions on the distribution of the errors.

We also wish to investigate the estimation of confidence intervals for R_o. We therefore need the distribution (or at least the standard deviation) of the statistic $\hat{R}_o$ under the process generating the observed data. To do so, we would generate different series of data that preserve the dependency between the inputs and the outputs of the regression. We would use the values of $\hat{R}_o$ on these to test the null hypothesis that R_o is not positive (via confidence intervals). This method, using either parametric models or adequate non-parametric are more difficult to apply (because of the requirement that the proper input/output dependency must be preserved). Models of price returns, based on ARMA(1,1) or equivalent forms proposed in literature were studied, but it seemed in our experimentations that they did not reflect some of the statistical

properties that we observed in the real price return data.

References

[Anderson, 1984] Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis*. John Wiley and Sons, New York.

[Bartlett, 1946] Bartlett, M. (1946). On the theoritical specification of sampling properties of autoc orrelated time series. *J. Royal Stat. Soc. B*, 8:27–41.

[Bishop, 1995] Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, London, UK.

[Box and Jenkins, 1970] Box, G. and Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

[Campbell et al., 1997] Campbell, J., Lo, A. W., and MacKinlay, A. (1997). *The Econometrics of Financial Markets*. Princeton University Press, Princeton.

[Diebold and Kilian, 1997] Diebold, F. X. and Kilian, L. (1997). Measuring predictability:theory and macroeconomics applications. *NBER technical working paper*, 213.

[Diebold and Lopez, 1996] Diebold, F. X. and Lopez, J. A. (1996). Forecast evaluation and combination. In Maddala, G. and Rao, C., editors, *Handbook of Statistics, Vol. 14*, pages 241–268. Elsevier Science.

[Diebold and Mariano, 1995] Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263.

[Efron and Tibshirani, 1993] Efron, B. and Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. Chapman and Hall, New-York.

[Fama and French, 1988] Fama, E. and French, K. (1988). Permanent and temporary components of stock prices. *Journal of Political Economy*, 96(2):246–273.

[Granger and Newbold, 1976] Granger, C. W. J. and Newbold, P. (1976). Forecasting transformed series. *J. Roy. Statist. Soc. B*, 38:189–203.

[Kaul, 1996] Kaul, G. (1996). Predictable components in stock returns. In Maddala, G. and Rao, C., editors, *Handbook of Statistics, Vol. 14*, pages 269–296. Elsevier Science.

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

[Vapnik, 1995] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New-York.

Liste des publications au CIRANO*

Série Scientifique / *Scientific Series* (ISSN 1198-8177)

2002s-51 On Out-of-Sample Statistics for Time-Series / F. Gingras, Y. Bengio et C. Nadeau

2002s-50 Forecasting Non-Stationary Volatility with Hyper-Parameters / Y. Bengio et C. Dugas

2002s-49 Cost Functions and Model Combination for VaR-based Asset Allocation using Neural Networks / N. Chapados et Y. Bengio

2002s-48 Experiments on the Application of IOHMMs to Model Financial Returns Series / Y. Bengio, V.-P. Lauzon et R. Ducharme

2002s-47 Valorisation d'Options par Optimisation du Sharpe Ratio / Y. Bengio, R. Ducharme, O. Bardou et N. Chapados

2002s-46 Incorporating Second-Order Functional Knowledge for Better Option Pricing / C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau et R. Garcia

2002s-45 Étude du Biais dans le Prix des Options / C. Dugas et Y. Bengio

2002s-44 Régularisation du Prix des Options : Stacking / O. Bardou et Y. Bengio

2002s-43 Monotonicity and Bounds for Cost Shares under the Path Serial Rule / Michel Truchon et Cyril Téjédo

2002s-42 Maximal Decompositions of Cost Games into Specific and Joint Costs / Michel Moreaux et Michel Truchon

2002s-41 Maximum Likelihood and the Bootstrap for Nonlinear Dynamic Models / Sílvia Gonçalves, Halbert White

2002s-40 Selective Penalization Of Polluters: An Inf-Convolution Approach / Ngo Van Long et Antoine Soubeyran

2002s-39 On the Mediational Role of Feelings of Self-Determination in the Workplace: Further Evidence and Generalization / Marc R. Blais et Nathalie M. Brière

2002s-38 The Interaction Between Global Task Motivation and the Motivational Function of Events on Self-Regulation: Is Sauce for the Goose, Sauce for the Gander? / Marc R. Blais et Ursula Hess

2002s-37 Static Versus Dynamic Structural Models of Depression: The Case of the CES-D / Andrea S. Riddle, Marc R. Blais et Ursula Hess

2002s-36 A Multi-Group Investigation of the CES-D's Measurement Structure Across Adolescents, Young Adults and Middle-Aged Adults / Andrea S. Riddle, Marc R. Blais et Ursula Hess

2002s-35 Comparative Advantage, Learning, and Sectoral Wage Determination / Robert Gibbons, Lawrence F. Katz, Thomas Lemieux et Daniel Parent

2002s-34 European Economic Integration and the Labour Compact, 1850-1913 / Michael Huberman et Wayne Lewchuk

---

* Consultez la liste complète des publications du CIRANO et les publications elles-mêmes sur notre site Internet :

http://www.cirano.qc.ca

2002s-33 Which Volatility Model for Option Valuation? / Peter Christoffersen et Kris Jacobs

2002s-32 Production Technology, Information Technology, and Vertical Integration under Asymmetric Information / Gamal Atallah

2002s-31 Dynamique Motivationnelle de l'Épuisement et du Bien-être chez des Enseignants Africains / Manon Levesque, Marc R. Blais, Ursula Hess

2002s-30 Motivation, Comportements Organisationnels Discrétionnaires et Bien-être en Milieu Africain : Quand le Devoir Oblige / Manon Levesque, Marc R. Blais et Ursula Hess

2002s-29 Tax Incentives and Fertility in Canada: Permanent vs. Transitory Effects / Daniel Parent et Ling Wang

2002s-28 The Causal Effect of High School Employment on Educational Attainment in Canada / Daniel Parent

2002s-27 Employer-Supported Training in Canada and Its Impact on Mobility and Wages / Daniel Parent

2002s-26 Restructuring and Economic Performance: The Experience of the Tunisian Economy / Sofiane Ghali and Pierre Mohnen

2002s-25 What Type of Enterprise Forges Close Links With Universities and Government Labs? Evidence From CIS 2 / Pierre Mohnen et Cathy Hoareau

2002s-24 Environmental Performance of Canadian Pulp and Paper Plants : Why Some Do Well and Others Do Not ? / Julie Doonan, Paul Lanoie et Benoit Laplante

2002s-23 A Rule-driven Approach for Defining the Behavior of Negotiating Software Agents / Morad Benyoucef, Hakim Alj, Kim Levy et Rudolf K. Keller

2002s-22 Occupational Gender Segregation and Women's Wages in Canada: An Historical Perspective / Nicole M. Fortin et Michael Huberman

2002s-21 Information Content of Volatility Forecasts at Medium-term Horizons / John W. Galbraith et Turgut Kisinbay

2002s-20 Earnings Dispersion, Risk Aversion and Education / Christian Belzil et Jörgen Hansen

2002s-19 Unobserved Ability and the Return to Schooling / Christian Belzil et Jörgen Hansen

2002s-18 Auditing Policies and Information Systems in Principal-Agent Analysis / Marie-Cécile Fagart et Bernard Sinclair-Desgagné

2002s-17 The Choice of Instruments for Environmental Policy: Liability or Regulation? / Marcel Boyer, Donatella Porrini

2002s-16 Asymmetric Information and Product Differentiation / Marcel Boyer, Philippe Mahenc et Michel Moreaux

2002s-15 Entry Preventing Locations Under Incomplete Information / Marcel Boyer, Philippe Mahenc et Michel Moreaux

2002s-14 On the Relationship Between Financial Status and Investment in Technological Flexibility / Marcel Boyer, Armel Jacques et Michel Moreaux

2002s-13 Modeling the Choice Between Regulation and Liability in Terms of Social Welfare / Marcel Boyer et Donatella Porrini

2002s-12 Observation, Flexibilité et Structures Technologiques des Industries / Marcel Boyer, Armel Jacques et Michel Moreaux

2002s-11 Idiosyncratic Consumption Risk and the Cross-Section of Asset Returns / Kris Jacobs et Kevin Q. Wang

2002s-10 The Demand for the Arts / Louis Lévy-Garboua et Claude Montmarquette

2002s-09 Relative Wealth, Status Seeking, and Catching Up / Ngo Van Long, Koji Shimomura

2002s-08 The Rate of Risk Aversion May Be Lower Than You Think / Kris Jacobs

2002s-07 A Structural Analysis of the Correlated Random Coefficient Wage Regression Model / Christian Belzil et Jörgen Hansen

2002s-06 Information Asymmetry, Insurance, and the Decision to Hospitalize / Åke Blomqvist et Pierre Thomas Léger

2002s-05 Coping with Stressful Decisions: Individual Differences, Appraisals and Choice / Ann-Renée Blais

2002s-04 A New Proof Of The Maximum Principle / Ngo Van Long et Koji Shimomura

2002s-03 Macro Surprises And Short-Term Behaviour In Bond Futures / Eugene Durenard et David Veredas

2002s-02 Financial Asset Returns, Market Timing, and Volatility Dynamics / Peter F. Christoffersen et Francis X. Diebold

2002s-01 An Empirical Analysis of Water Supply Contracts / Serge Garcia et Alban Thomas

2001s-71 A Theoretical Comparison Between Integrated and Realized Volatilities Modeling / Nour Meddahi

2001s-70 An Eigenfunction Approach for Volatility Modeling / Nour Meddahi

2001s-69 Dynamic Prevention in Short Term Insurance Contracts / M. Martin Boyer et Karine Gobert

2001s-68 Serial Cost Sharing in Multidimensional Contexts / Cyril Téjédo et Michel Truchon

2001s-67 Learning from Strike / Fabienne Tournadre et Marie-Claire Villeval

2001s-66 Incentives in Common Agency / Bernard Sinclair-Desgagné

2001s-65 Detecting Mutiple Breaks in Financial Market Volatility Dynamics / Elena Andreou et Eric Ghysels

2001s-64 Real Options, Preemption, and the Dynamics of Industry Investments / Marcel Boyer, Pierre Lasserre, Thomas Mariotti et Michel Moreaux

2001s-63 Dropout, School Performance and Working while in School: An Econometric Model with Heterogeneous Groups / Marcel Dagenais, Claude Montmarquette et Nathalie Viennot-Briot

2001s-62 Derivatives Do Affect Mutual Funds Returns : How and When? / Charles Cao, Eric Ghysels et Frank Hatheway

2001s-61 Conditional Quantiles of Volatility in Equity Index and Foreign Exchange Data / John W. Galbraith, Serguei Zernov and Victoria Zinde-Walsh

2001s-60 The Public-Private Sector Risk-Sharing in the French Insurance "Cat. Nat. System" / Nathalie de Marcellis-Warin et Erwann Michel-Kerjan

2001s-59 Compensation and Auditing with Correlated Information / M. Martin Boyer et Patrick González

2001s-58 Resistance is Futile: An Essay in Crime and Commitment / M. Martin Boyer