Unexpected opportunities in misspecified predictive regressions

Guillaume Coqueret, Romain Deguest

European Journal of Operational Research 318 (2024) 686–700

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Interfaces with other disciplines

Unexpected opportunities in misspecified predictive regressions ☆

Guillaume Coqueret [a,*], Romain Deguest [b]

[a] EMLYON Business School, 23 avenue Guy de Collongue, 69130 Ecully, France
[b] World Bank, 1818 H Street, NW Washington, DC 20433, United States

ARTICLE INFO

JEL classification:
C13
C22
C53
G11
G12

Keywords:
Predictive regression
Model misspecification
Spurious accuracy
Short samples

ABSTRACT

This article documents surprising learning patterns that can occur under model misspecification. An agent resorts to predictive regressions and fails to take into account autocorrelation in the dependent variable. Remarkably, when the dependent and independent variables are uncorrelated, we find cases for which the resulting out-of-sample R^2 is well above zero, which benefits the agent, in spite of the erroneous model. We refer to them as instances of unexpected opportunity. When both variables exhibit high levels of persistence, we reveal the existence of counter-intuitive configurations for which the R^2 *increases* when the absolute correlation between the series decreases. Our theoretical results are confirmed by extensive simulations and complemented by an empirical exercise of equity premium prediction for which we use 15 predictors commonly referenced in the economic literature.

1. Introduction

Predictive regressions (PRs) are the most elementary forecasting models. Nevertheless, the apparent simplicity of their formulation masks the complexity of the underlying statistical machinery. Numerous studies have warned scholars and practitioners to be careful when relying on simple estimates stemming from PRs.[1] Since the seminal paper of Stambaugh (1999), it is for instance known that when regressors are autocorrelated, the OLS estimators of PRs are biased and should be corrected.

From an inferential standpoint, other issues arise when the dependent variable is also autocorrelated, which occurs for instance in finance when returns are computed over long horizons (Campbell, 2001; Lanne, 2002).[2] In this case, test statistics are artificially larger *by construction* (Boudoukh et al., 2008; Valkanov, 2003), which gives rise to false positive results. This is partly linked to the notion of *spurious correlation* (Granger & Newbold, 1974). Indeed, even if the two series are completely independent random walks, it is possible to uncover significant coefficients, where there should be none.

In this paper, we are interested in the risk that an agent faces whilst using PRs for forecasting purposes whilst the actual data generating process does not correspond to a typical PR model. Our analysis does not focus on inference but on predictive accuracy. In this setting, the major hurdle is not spurious correlation, but model misspecification. The latter is ubiquitous in the operations and management science literature.[3]

Our contributions are threefold. First, we derive analytical identities and properties for the mean quadratic error of strongly misspecified PRs, i.e., when PRs are used despite the absence of a direct link between the two variables. Importantly, our formulae hold when estimates are based on finite samples, whereas most results in statistical inference are obtained asymptotically. They are therefore of particular relevance for decision-makers who resort to small samples ($T < 100$, approximately), either because they have to (when data is scarce), or because

☆ **Disclaimer**: the findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

* Corresponding author.
E-mail addresses: coqueret@em-lyon.com (G. Coqueret), rdeguest@worldbank.org (R. Deguest).

[1] A selective list is: Deng (2014), Ferson et al. (2003) and Phillips (2015).

[2] Indeed, since Fama and French (1988a, 1988b), it has been widely documented that statistical significance is empirically more pronounced when return horizons are large (e.g. Bandi and Perron (2008) and Sizova (2013)) though this phenomenon vanishes beyond 12 to 15 years (Bandi et al., 2019). Two notable exceptions are Ang and Bekaert (2007) and Torous et al. (2004), where the authors argue that predictability can be a short-term phenomenon.

[3] Indeed, this topic is abundantly covered, whether in product pricing in supply/demand models (Cooper et al., 2015; Nambiar et al., 2019), derivative pricing (Coqueret & Tavin, 2016; Lazar & Qi, 2022), decision science (Cerreia-Vioglio et al., 2020), financial risk management (Barrieu & Scandolo, 2015; Wu & Olson, 2010), or expected utility evaluation (Blanchet & Murthy, 2019).

https://doi.org/10.1016/j.ejor.2024.05.044
Received 25 November 2023; Accepted 28 May 2024
Available online 29 May 2024

they choose to. The latter case can occur when data becomes rapidly obsolete so that only the most recent samples are used for estimation and prediction purposes, e.g., in rapidly changing environments.

As is often the case when analyzing regression coefficients, our results rely heavily on moments of ratios of quadratic forms of multivariate Gaussian distributions. This particular topic is reviewed in Appendix B of Bao and Kan (2013) and in Paolella (2018). In a similar vein, Kan and Wang (2010) use an elegant characterization of eigenvalues to derive moments for the sample estimator of the autocorrelation, while Kan and Pan (2021) derive finite sample properties of estimators in predictive regressions, but from an inferential standpoint. These recent results, as well as ours, build on the work of Magnus (1990) who proposes expressions based on simple integral formulas for nontrivial exponents. This contribution generalizes previous work on the subject (Magnus, 1986; Sawa, 1972 notably).

Second, from the theoretical formula obtained for the mean quadratic error, we derive a key relationship between the out-of-sample R^2 of the regression and the correlation ρ between the innovations of the two autoregressive processes. This result shows that the R^2 is a symmetric function of ρ and allows us to introduce the notion of *unexpected gain*, which comes in two varieties. More precisely, we call *unexpected gain of Type I* the case when the out-of-sample R^2 is strictly positive, despite a zero correlation between innovations (and processes). On the other hand, *unexpected gains of Type II* occur counter-intuitively when the R^2 is increasing when $|\rho|$, the absolute value of the correlation between innovations, decreases. These two types of forecasting anomalies are confirmed via simulations which reveal that they occur for high levels of persistence in the dependent variable. The first type also requires small sample sizes while the second type a high persistence in the predictors as well. The fact that predictive accuracy may be obtained in spite of model misspecification has already been documented in a different context in Nelson (1992) and Nelson and Foster (1995).

Finally, our third contribution contextualizes our findings in an empirical study of return predictability for which the dependent variable is the S&P500 index returns, and the predictors are taken from the studies from Novy-Marx (2014) and Welch and Goyal (2008). It confirms the stylized results obtained in the numerical section. Notably, our conclusions corroborate the impact of the persistence of both the dependent variable and the predictor on the R^2 together with the influence of the sample size. Our last study pertains to the short-term predictability of the S&P500 forward volatility (VIX) and concludes that for persistent dependent variables, short training samples yield large positive R^2. Because this stylized fact is hard to rationalize, we attribute it to unexpected gains of Type I.

These results are undoubtedly linked to the abundant empirical literature on the predictability of asset returns. While some contributions argue in favor of predictability (Cochrane, 2008; Lewellen, 2004), or against (Bossaerts & Hillion, 1999; Ferson et al., 2003; Goyal & Welch, 2003; Welch & Goyal, 2008), some studies take a more unifying stance by concluding that the relationship is likely time-varying (Dangl & Halling, 2012; Farmer et al., 2021; Zhu, 2015). Similarly, our conclusions are not clear-cut, because we do not arbitrate for or against predictability. Rather, we detail the econometric configurations under which predictive regressions are likely to deliver unintended out-of-sample accuracy. It is important to underline that while our empirical study is focused on asset pricing, the theoretical results of the paper can be valuable to other disciplines.

Lastly, the present paper contributes to the literature on unexpected patterns in statistical learning. Recent results in the machine learning field document such occurrences for arbitrarily large samples, e.g., with *benign overfitting* and the *double descent* effects (see Bartlett et al. (2020) and Hastie et al. (2022) among many others). In contrast, the results we obtain are most surprising for *small* samples, as if imperfect learning from few observations was able to partly mitigate the handicap of misspecification. We also point to Berk et al. (2014), who propose to learn from imperfect regression models, and who hence recommend to abandon the search for the correct one.

The remainder of the paper is structured as follows. In Section 2, we formulate our research problem, provide all the required notations, lay out our analytical results, and propose our definitions of unexpected gains. Section 3 is dedicated to numerical analyses and illustrates the phenomenon of spurious accuracy over simulated data. In Section 4, we confirm our theoretical conclusions on empirical financial data, while Section 5 concludes. All the technical details and proofs are located in Appendix.

2. Problem and theoretical results

2.1. Data generating process (DGP)

We assume that the data $\boldsymbol{z}_t^* = \begin{bmatrix} \boldsymbol{x}_t^* \\ \boldsymbol{y}_t^* \end{bmatrix}$ is generated by a bivariate first order and stationary vector-autoregression (VAR(1)), with

$$\boldsymbol{z}_{t+1}^* = \boldsymbol{a}_z + \boldsymbol{B}_z \boldsymbol{z}_t^* + \boldsymbol{\epsilon}_{t+1}, \tag{1}$$

where $\boldsymbol{a}_z$ and $\boldsymbol{B}_z$ are respectively a 2×1 vector and a 2×2 matrix. The distributional properties of errors $\boldsymbol{\epsilon}_{t+1}$ will be discussed subsequently. Our notations are such that processes written with stars, e.g. $\boldsymbol{z}_t^*$, have nonzero means, while those written without stars, e.g. $\boldsymbol{z}_t$, have zero-mean. In most of our analysis, means will play no role, which is why we will omit the stars later on, without any loss of generality.

Autoregressive models are ubiquitous in several fields, including economics (Hsiao, 1981; Stock & Watson, 2001) and finance (Campbell et al., 1997; Hsu et al., 2022; Piatti & Trojani, 2020), and are also sometimes used in logistics (Eroglu & Hofer, 2011; Levi et al., 2008; Luong, 2007; Sobel & Babich, 2012) and even politics (Freeman et al., 1989). In finance, for example, modeling independent predictors as autocorrelated processes is commonplace, see, e.g., Campbell and Yogo (2006), Stambaugh (1999), and Van Binsbergen and Koijen (2010), to cite but a few. For instance, at the aggregate level, dividend yields, stock variance, and book-to-market ratios are all persistent. Therefore, autocorrelation has become the norm in papers that propose solutions to biased estimators (e.g., Hjalmarsson (2011), Stambaugh (1999) and Xu (2020)).

When the off-diagonal terms in $\boldsymbol{B}_z$ are nonzero, we can rewrite the dynamics of y^* as[4]

$$y_{t+1}^* = [\boldsymbol{a}_z]_2 + [\boldsymbol{B}_z]_{2,1}\, x_t^* + [\boldsymbol{B}_z]_{2,2}\, y_t^* + \text{error term},$$

and see that both x_t^* and y_t^* have a direct effect on the future value y_{t+1}^*. On the other hand, if the off-diagonal terms in $\boldsymbol{B}_z$ are zero, then the model boils down to two strictly stationary first-order auto-regressive (AR(1)) processes x^* and y^*:

$$x_{t+1}^* = \alpha_x + \rho_x x_t^* + e_{x^*,t+1}, \tag{2}$$

$$y_{t+1}^* = \alpha_y + \rho_y y_t^* + e_{y^*,t+1}, \tag{3}$$

with constants α_x and α_y, autocorrelations ρ_x and ρ_y satisfying $|\rho_x| < 1$ and $|\rho_y| < 1$, and correlated Gaussian white noise processes e_{x^*} and e_{y^*} with variances σ_x^2 and σ_y^2 and correlation ρ satisfying $|\rho| < 1$. As we will show, when $\boldsymbol{B}_z$ is diagonal, x^* and y^* are nevertheless correlated, because of ρ, which characterizes the dependence in the error terms.

Henceforth, we will impose that $\boldsymbol{B}_z$ be diagonal for two reasons:

1. The first reason is analytical tractability. Our theoretical results and their proofs when $\boldsymbol{B}_z$ is diagonal are cumbersome. The general case when $\boldsymbol{B}_z$ has non-zero off-diagonal terms is prohibitively complex and does not allow for closed-form formulae.

[4] We write $[\boldsymbol{M}]_{i,j}$ for the element located at the ith row and jth column for matrix $\boldsymbol{M}$ with similar notation for vectors.

2. The second reason is that we want to assume that the agent makes a modeling error. If the DGP is such that the link between y^* and x^* is direct, then working with a predictive regression (see Section 2.2 below) implies that the choice of the model is correct ex-ante, which will likely yield deceptively favorable results. But in practice, outstanding forecasting performance rarely occurs. As White (2014) puts it: "*Owing to the complexity of economic phenomena, it is perhaps more realistic that the relationship between X_t and Y_t is unknown*". Imposing a diagonal $\boldsymbol{B}_z$ introduces a disconnect between the agents' beliefs and the actual realizations of the world.

Nevertheless, even if the DGP assumes no direct link between x^* and y^*, the agent will still be able to (sometimes) benefit from the non-zero correlation between the two processes.

2.2. Agent model

In the present paper, the agent (e.g., economist, asset manager, logistics analyst) seeks to forecast the future value of y^*, but ignores the true DGP and decides instead to rely on a simple predictive regression using x^* in order to build such a forecast. The agent therefore assumes the following relationship:

$$y^*_{t+k} = a + b\, x^*_t + e_{t+k}, \tag{4}$$

where the current level of the predictor x^*_t aims to predict the k-step ahead value of the dependent variable y^*_{t+k}. In multiple disciplines, examples for the predictor include industrial output, stock levels, credit and term spreads, and dividend yield, while instances of the dependent variable encompass GDP growth, inventory levels, lot sizes, and future market returns. In Eq. (4), we consider predictive regression models in a *strict sense*, meaning that the past of y^* is not included in the prediction. Therefore, this implies that the agent makes an error in her assumptions which is at the root of the model misspecification.

Plainly, in Eq. (4), a is the intercept and b the slope. In practice, these two parameters must be estimated from a dataset, which we will call the training sample, as is customary in the machine learning (ML) jargon. If we denote with T the size of this training sample $D^*_t = \{(x^*_{t-k-T+1}, y^*_{t-T+1}), \dots, (x^*_{t-k}, y^*_t)\}$, then the canonical expressions for the Ordinary Least Squares (OLS) estimator are

$$\hat{b}(D^*_t) = \frac{\sum_{s=0}^{T-1}(x^*_{t-k-s} - \bar{x}^*)(y^*_{t-s} - \bar{y}^*)}{\sum_{s=0}^{T-1}(x^*_{t-k-s} - \bar{x}^*)^2} \tag{5}$$

$$\hat{a}(D^*_t) = \bar{y}^* - \hat{b}(D^*_t)\bar{x}^*, \tag{6}$$

with sample means $\bar{x}^* = \frac{1}{T}\sum_{s=0}^{T-1} x^*_{t-k-s}$ and $\bar{y}^* = \frac{1}{T}\sum_{s=0}^{T-1} y^*_{t-s}$. One important feature of the training sample is naturally the time shift between x^* and y^* due to the forecasting objective. The variable x^*_t is exploited to forecast the variable y^*_{t+k} that will only be measured k periods in the future. Moreover, in the traditional statistics literature, it is customary to make assumptions on the errors e_{t+k} in the model, especially for inference purposes. In the present paper, we are interested in the ex-post consequences of the modeling choice of the agent, that is, the potential cost of model misspecification.

In the present paper, we work under this hypothesis of model misspecification in order to provide a realistic evaluation of loss and risk from the agent's perspective. This corresponds to a scenario where the agent models a direct link between x^* and y^* and ignores the autoregressive component on y^* whereas the underlying DGP is only autoregressive and contains no explicit link between x^* and y^* beyond correlation.

More generally, this model misspecification falls under the umbrella of what Chambers et al. (2018) call "*automatic or blind use of regression models*",[5] which is made possible by the fact that regression routines are readily accessible in all data analysis tools, from Excel, SPSS or STATA to R, Python, Julia and Matlab. This has long been documented in various fields, including economics (Angrist & Pischke, 2010) and policy decision making (Porter et al., 1981), but also in the medical sciences (Porter, 1999), in physics and chemistry (Badertscher & Pretsch, 2006; Exner, 1997), and in the agricultural sciences (Mitchell, 1997).

Overlooking the properties of predictors and dependent variables is common in forecasting practice also because accuracy matters more than inference. Consequently, it is the out-of-sample strength of the link between y and x that is sought by forecasters.[6] In fact, in recent machine learning contributions, large linear models are considered without necessarily focusing on the properties of predictors. We point to Bartlett et al. (2020) and Hastie et al. (2022) for theoretical results on (high-dimensional) linear models.

We can thus summarize the framework as follows:

1. The agent is not sophisticated and opts for standard predictive regression modeling.
2. But, by nature, the underlying data generating process is autoregressive. Our specification precludes *direct* links between x^* and y^* while at the same time allowing for correlations via the innovation terms.

We use the following notations in the paper. Vectors $\boldsymbol{v}$ and matrices $\boldsymbol{M}$ are written with bold fonts. For square matrices, we simplify the notation and use only one subscript: $\boldsymbol{I}_T$, $\boldsymbol{0}_T$ and $\boldsymbol{1}_T$ denote the T dimensional identity, zero (filled with 0) and unit (filled with 1) matrices, respectively, while $\boldsymbol{0}_{N,M}$ and $\boldsymbol{1}_{N,M}$ are $(N \times M)$ matrices filled with zeros and ones. $\boldsymbol{v}'$ and $\boldsymbol{M}'$ are the transpose of the corresponding vector and matrix. Finally, $\mathrm{tr}(\cdot)$ is the trace that operates on the set of square matrices and yields the sum of diagonal elements.

2.3. A first look at squared errors

One focal quantity in the present article is the mean squared error of the prediction, also called quadratic loss in ML parlance and defined by $\mathbb{E}[\hat{e}^2_{t+k}]$. First, we derive the conditional loss for the agent, based on all the information available at time t, namely D^*_t and x^*_t. It can be derived as follows, with $\mathbb{E}_t[\cdot]$ being the expectation, conditional on the data at time t:

$$\begin{aligned} L_t &:= L(k, \alpha_y, \rho_y, \sigma^2_y, D^*_t, x^*_t) \\ &= \mathbb{E}_t\left[\left(y^*_{t+k} - \hat{a}(D^*_t) - \hat{b}(D^*_t)x^*_t\right)^2\right] \\ &= \underbrace{\mathbb{E}_t\left[\left(y^*_{t+k} - \mathbb{E}_t[y^*_{t+k}]\right)^2\right]}_{\mathrm{var}_t(y^*_{t+k})} + \underbrace{\left(\mathbb{E}_t[y^*_{t+k}] - \hat{a}(D^*_t) - \hat{b}(D^*_t)x^*_t\right)^2}_{\mathrm{bias}^2_t} \\ &= \mathbb{E}_t\left[\left(\sum_{s=0}^{k-1}\rho^s_y e_{y^*,t+k-s}\right)^2\right] + \left(\alpha_y\sum_{s=0}^{k-1}\rho^s_y + \rho^k_y y^*_t - \hat{a}(D^*_t) - \hat{b}(D^*_t)x^*_t\right)^2 \\ &= \sigma^2_y\sum_{s=0}^{k-1}\rho^{2s}_y + \left(\alpha_y\sum_{s=0}^{k-1}\rho^s_y + \rho^k_y y^*_t - \bar{y}^* + \hat{b}(D^*_t)(\bar{x}^* - x^*_t)\right)^2. \end{aligned} \tag{7, 8}$$

All items in the last equation are either model parameters or elements of the training sample. More interestingly, this paper is focused on the unconditional loss, which depends only on model parameters and is equal to

$$\begin{aligned} L &:= L(T, k, \alpha_x, \alpha_y, \rho_x, \rho_y, \rho, \sigma^2_x, \sigma^2_y) = \mathbb{E}\left[L_t\right] \\ &= \mathbb{E}\left[\left(y^*_{t+k} - \hat{a}(D^*_t) - \hat{b}(D^*_t)x^*_t\right)^2\right]. \end{aligned} \tag{9}$$

[5] "*Misuse*" and "*abuse*" (see Box (1966)) are others terms that denote improper use of regression models.

[6] Modern tools such as neural networks and tree ensembles are known to be effective forecasting engines. However, when only one predictor is considered, as in the present paper, sophisticated tools have less forecasting edge.

Our first result is the computation of the loss L when the slope b and intercept a parameters of the linear prediction model (4) are known by the agent. We also provide the expression of the minimal value achieved by the loss in that specific framework.

Lemma 1. *Consider two AR(1) processes with correlated innovations given by (2) and (3), and the predictive regression model (4). The quadratic loss L of the predictive model satisfies the following properties:*

1. *If the intercept a and slope b are known, then*

$$L_{ab} = \frac{\sigma_y^2}{1-\rho_y^2} - 2b\rho_y^k \frac{\rho\sigma_x\sigma_y}{1-\rho_x\rho_y} + b^2\frac{\sigma_x^2}{1-\rho_x^2} \tag{10}$$

2. *The minimal value of L_{ab} is achieved for*

$$b_o = \frac{cov(x_t, y_{t+k})}{var(x_t)} = \rho_y^k \frac{\rho\sigma_x\sigma_y}{1-\rho_x\rho_y}\frac{1-\rho_x^2}{\sigma_x^2}, \tag{11}$$

and is equal to

$$L_o = \frac{\sigma_y^2}{1-\rho_y^2}\left(1 - \rho_y^{2k}\frac{\rho^2}{(1-\rho_x\rho_y)^2}(1-\rho_x^2)(1-\rho_y^2)\right). \tag{12}$$

In practice, the slope b is not known but estimated using the training sample D_t^* which makes the computation of the loss much more cumbersome. Nevertheless, elementary results such as the ones above allow us to capture salient properties of the model. For instance, from Eq. (11), we infer that actual predictability of y_{t+k} using x_t (i.e. $b_o \neq 0$) requires that both ρ and ρ_y be nonzero, i.e., both correlations in innovations *and* persistence in y_{t+k} should be non-null.

The aim of the following sections is to provide tractable and analytical formulae for L when a and b are unknown and to characterize its sensitivity to key parameters. As will be proven later on, some parameters of the data generating processes (2) and (3) have little interest because either they have no impact on the loss or their impact is straightforward. In order to focus on the most relevant parameters of the underlying DGP, we will work with an alternative formulation which we specify below.

2.4. Reformulation of the problem

Replacing $\hat{a}(D_t^*)$ with its expression in Eq. (9) leads to

$$L = \mathbb{E}\left[\left(y_{t+k}^* - \bar{y}^* - \hat{b}(D_t^*)(x_t^* - \bar{x}^*)\right)^2\right]. \tag{13}$$

The differences $y_{t+k}^* - \bar{y}^*$ and $x_t^* - \bar{x}^*$ in the two processes seem to indicate that the drift terms α_x and α_y will cancel out and that L will not be affected by these terms. Also, the slope term $\hat{b}(D_t^*)$ is linear in the y process and inversely proportional to the x process, loosely speaking. Hence we expect that L will be linearly dependent on σ_y^2, but insensitive to σ_x^2. These claims are formally proven in our first theorem below.

Taking into account these observations, Eq. (9) can be simplified to the equivalent definition

$$L := L(T, k, \rho_x, \rho_y, \rho, \sigma_y^2) = \sigma_y^2\,\mathbb{E}\left[\left(y_{t+k} - \hat{a}(D_t) - \hat{b}(D_t)x_t\right)^2\right], \tag{14}$$

where (x_t, y_{t+k}) and the training sample $D_t = \{(x_{t-k-T+1}, y_{t-T+1}), \dots, (x_{t-k}, y_t)\}$ are obtained from the two simplified auto-regressive processes x and y defined by

$$x_{t+1} = \rho_x x_t + e_{x,t+1}, \tag{15}$$

$$y_{t+1} = \rho_y y_t + e_{y,t+1}, \tag{16}$$

with autocorrelations ρ_x and ρ_y and correlated Gaussian white noise processes e_x and e_y with variances equal to 1 and correlation ρ.

The definitions of the estimators $\hat{a}$ and $\hat{b}$ remain the same as in (5) and (6), but applied to the new training sample D_t. This means that, even though the constants of the two processes x and y are now equal to 0, the agent does not know that the constant are equal to 0 and therefore still uses the sample means $\bar{x}$ and $\bar{y}$ of processes x and y in the computation of $\hat{a}(D_t)$ and $\hat{b}(D_t)$.

The new formulation (14) leads to

$$L = \sigma_y^2\,\mathbb{E}\left[\left(y_{t+k} - \bar{y} - \hat{b}(D_t)(x_t - \bar{x})\right)^2\right] \tag{17}$$

$$= \sigma_y^2\,\mathbb{E}\left[(y_{t+k} - \bar{y})^2\right] - 2\mathbb{E}\left[(y_{t+k} - \bar{y})(x_t - \bar{x})\hat{b}(D_t)\right] + \mathbb{E}\left[(x_t - \bar{x})^2\hat{b}(D_t)^2\right]$$

$$= \sigma_y^2\,\mathbb{E}\left[\bar{\boldsymbol{z}}'\boldsymbol{E}\bar{\boldsymbol{z}} - 2(\bar{\boldsymbol{z}}'\boldsymbol{C}\bar{\boldsymbol{z}})\left(\frac{\bar{\boldsymbol{z}}'\boldsymbol{A}\bar{\boldsymbol{z}}}{\bar{\boldsymbol{z}}'\boldsymbol{B}\bar{\boldsymbol{z}}}\right) + (\bar{\boldsymbol{z}}'\boldsymbol{D}\bar{\boldsymbol{z}})\left(\frac{\bar{\boldsymbol{z}}'\boldsymbol{A}\bar{\boldsymbol{z}}}{\bar{\boldsymbol{z}}'\boldsymbol{B}\bar{\boldsymbol{z}}}\right)^2\right], \tag{18}$$

$$= \sigma_y^2\Bigg(\underbrace{\mathbb{E}\left[\bar{\boldsymbol{z}}'\boldsymbol{E}\bar{\boldsymbol{z}}\right]}_{\text{variance of } y_{t+k}-\bar{y}} \underbrace{- 2\mathbb{E}\left[(\bar{\boldsymbol{z}}'\boldsymbol{C}\bar{\boldsymbol{z}})\left(\frac{\bar{\boldsymbol{z}}'\boldsymbol{A}\bar{\boldsymbol{z}}}{\bar{\boldsymbol{z}}'\boldsymbol{B}\bar{\boldsymbol{z}}}\right)\right]}_{\text{cross term}}$$

$$\underbrace{+\,\mathbb{E}\left[(\bar{\boldsymbol{z}}'\boldsymbol{D}\bar{\boldsymbol{z}})\left(\frac{\bar{\boldsymbol{z}}'\boldsymbol{A}\bar{\boldsymbol{z}}}{\bar{\boldsymbol{z}}'\boldsymbol{B}\bar{\boldsymbol{z}}}\right)^2\right]}_{\text{quadratic term}}\Bigg), \tag{19}$$

where $\bar{\boldsymbol{z}} = \boldsymbol{N}\boldsymbol{z} = [x_{t-k-T+1} - \bar{x}, \dots, x_{t-k} - \bar{x}, x_t - \bar{x}, y_{t-T+1} - \bar{y}, \dots, y_t - \bar{y}, y_{t+k} - \bar{y}]'$ is a Gaussian vector built from linear combinations of all the random variables of the system stacked in vector $\boldsymbol{z} = [x_{t-k-T+1}, \dots, x_{t-k}, x_t, y_{t-T+1}, \dots, y_t, y_{t+k}]'$ and where $\boldsymbol{N}$ is defined by

$$\boldsymbol{N} = \begin{bmatrix} \boldsymbol{N}_{T+1} & \boldsymbol{0}_{T+1} \\ \boldsymbol{0}_{T+1} & \boldsymbol{N}_{T+1} \end{bmatrix} \quad \text{with} \quad \boldsymbol{N}_{T+1} = \begin{bmatrix} \boldsymbol{M}_T & \boldsymbol{0}_{T,1} \\ -\frac{1}{T}\boldsymbol{1}_{1,T} & 1 \end{bmatrix} \quad \text{and}$$

$$\boldsymbol{M}_T = \boldsymbol{I}_T - \frac{1}{T}\boldsymbol{1}_T.$$

We note that $\boldsymbol{M}_T$ is a demeaning (i.e. centering) operator. Matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ and $\boldsymbol{E}$ in Eq. (19) are given below:

$$\boldsymbol{A} = \frac{1}{2}\begin{bmatrix} \boldsymbol{0}_{T+1} & \boldsymbol{J}_{T+1} \\ \boldsymbol{J}_{T+1} & \boldsymbol{0}_{T+1} \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{J}_{T+1} & \boldsymbol{0}_{T+1} \\ \boldsymbol{0}_{T+1} & \boldsymbol{0}_{T+1} \end{bmatrix} \quad \boldsymbol{C} = \frac{1}{2}\begin{bmatrix} \boldsymbol{0}_{T+1} & \boldsymbol{K}_{T+1} \\ \boldsymbol{K}_{T+1} & \boldsymbol{0}_{T+1} \end{bmatrix}$$

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{K}_{T+1} & \boldsymbol{0}_{T+1} \\ \boldsymbol{0}_{T+1} & \boldsymbol{0}_{T+1} \end{bmatrix}, \quad \boldsymbol{E} = \begin{bmatrix} \boldsymbol{0}_{T+1} & \boldsymbol{0}_{T+1} \\ \boldsymbol{0}_{T+1} & \boldsymbol{K}_{T+1} \end{bmatrix}$$

where

$$\boldsymbol{J}_{T+1} = \begin{bmatrix} \boldsymbol{I}_T & \boldsymbol{0}_{T,1} \\ \boldsymbol{0}_{1,T} & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{K}_{T+1} = \boldsymbol{I}_{T+1} - \boldsymbol{J}_{T+1} = \begin{bmatrix} \boldsymbol{0}_T & \boldsymbol{0}_{T,1} \\ \boldsymbol{0}_{1,T} & 1 \end{bmatrix}.$$

$\boldsymbol{z}$ is a stationary Gaussian vector with mean $\boldsymbol{0}_{2(T+1)}$ and covariance matrix $\boldsymbol{\Sigma}$ given by the following block form

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{T+1}^{\boldsymbol{x}} & \left(\boldsymbol{\Sigma}_{T+1}^{\boldsymbol{xy}}\right)' \\ \boldsymbol{\Sigma}_{T+1}^{\boldsymbol{xy}} & \boldsymbol{\Sigma}_{T+1}^{\boldsymbol{y}} \end{bmatrix}, \tag{20}$$

where $\boldsymbol{\Sigma}_{T+1}^{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{T+1}^{\boldsymbol{y}}$ are the two autocorrelation matrices of $\boldsymbol{x} = [x_{t-k-T+1}, \dots, x_{t-k}, x_t]'$ and $\boldsymbol{y} = [y_{t-T+1}, \dots, y_t, y_{t+k}]'$ and $\boldsymbol{\Sigma}_{T+1}^{\boldsymbol{xy}}$ the covariance matrix between $\boldsymbol{x}$ and $\boldsymbol{y}$. In the following, we will use subscript T and denote with $\boldsymbol{\Sigma}_T^{\boldsymbol{x}}$ (resp. $\boldsymbol{\Sigma}_T^{\boldsymbol{y}}$) the covariance matrix of the first T elements $[x_{t-k-T+1}, \dots, x_{t-k}]'$ of vectors $\boldsymbol{x}$ (resp. the first T elements $[y_{t-T+1}, \dots, y_t]'$ of vector $\boldsymbol{y}$). Moreover, we will denote with $\boldsymbol{\epsilon}_{\boldsymbol{x}}$ (resp. $\boldsymbol{\epsilon}_{\boldsymbol{y}}$) the covariance vector of x_t with the first T elements $[x_{t-k-T+1}, \dots, x_{t-k}]'$ (resp. the covariance vector of y_{t+k} with the first T elements $[y_{t-T+1}, \dots, y_t]'$). Therefore we can write[7]

$$\boldsymbol{\Sigma}_{T+1}^{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{\Sigma}_T^{\boldsymbol{x}} & \boldsymbol{\epsilon}_{\boldsymbol{x}} \\ \boldsymbol{\epsilon}_{\boldsymbol{x}}' & (1-\rho_x^2)^{-1} \end{bmatrix} \quad \text{with} \quad \boldsymbol{\epsilon}_{\boldsymbol{x}} = (1-\rho_x^2)^{-1}\left[\rho_x^{k+T-1}, \dots, \rho_x^k\right]',$$

and where $\boldsymbol{\Sigma}_T^{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_T^{\boldsymbol{y}}$ are Toeplitz symmetric definite positive matrices whose elements on row i and column j are equal to $(1-\rho_x^2)^{-1} \times \rho_x^{|i-j|}$ and $(1-\rho_y^2)^{-1} \times \rho_y^{|i-j|}$ respectively.

[7] The same notations will be used for y.

Matrix $\boldsymbol{\Sigma}^{\boldsymbol{xy}}_{T+1}$ is a bit more complicated but can be written as

$$\boldsymbol{\Sigma}^{\boldsymbol{xy}}_{T+1} = \frac{\rho}{1-\rho_x\rho_y}\begin{bmatrix}\boldsymbol{\Xi}^{\boldsymbol{xy}}_T & \boldsymbol{\xi_1}\\ \boldsymbol{\xi}'_2 & \rho_y^k\end{bmatrix} \quad \text{with} \quad \boldsymbol{\xi_1} = \left[\rho_x^{T-1},\dots,\rho_x,1\right]' \quad \text{and}$$
$$\boldsymbol{\xi_2} = \left[\rho_y^{2k+T-1},\dots,\rho_y^{2k}\right]', \tag{21}$$

and the element on row i and column j of matrix $\boldsymbol{\Xi}^{\boldsymbol{xy}}_T$ corresponds to the covariance between $x_{t-k-T+j}$ and y_{t-T+i} and is given by

$$\left[\boldsymbol{\Xi}^{\boldsymbol{xy}}_T\right]_{i,j} = \begin{cases}\rho_y^{i+k-j} & \text{for } i+k\geq j\\ \rho_x^{j-i-k} & \text{for } j\geq i+k\end{cases}.$$

Finally, we denote with $\bar{\boldsymbol{\Sigma}}$ the covariance matrix $\boldsymbol{N\Sigma N}'$ of the Gaussian vector $\bar{\boldsymbol{z}} = \boldsymbol{Nz}$ built from the linear combination $\boldsymbol{N}$ applied to all the random variables of the system stacked in vector form such that $\boldsymbol{z} = [x_{t-k-T+1},\dots,x_{t-k},x_t,y_{t-T+1},\dots,y_t,y_{t+k}]'$. We write Λ_i the eigenvalues of $\boldsymbol{B}\bar{\boldsymbol{\Sigma}}\boldsymbol{B}$ and $\boldsymbol{\Lambda}$ is the diagonal matrix that contains these values.

With all of these notations, we are equipped to proceed to our theoretical results.

2.5. Main results

Our first result holds in all generality. The first three points are simple yet useful properties of the quadratic loss, while the last item is its core decomposition, which we obtain by using the main theorem of Magnus (1990), which is recalled in Appendix I. Except for the last proposition and lemma, our theoretical results are mostly hard to interpret at first, but they can be confirmed and illustrated with numerical exercises. This will be the purpose of the next section.

Theorem 2. *Consider two AR(1) processes with correlated innovations given by (2) and (3), and the predictive regression model (4). The bias (average error) is equal to zero and the quadratic loss L, defined in (9), satisfies the following properties:*

1. *L is independent of the two constants α_x and α_y.*
2. *L is independent of σ_x^2, the variance of innovations in the predictor process x^*.*
3. *L is proportional to σ_y^2, the variance of innovations in the predicted process y^*.*
4. *L is equal to*

$$L(T,k,\rho_x,\rho_y,\rho,\sigma_y^2) = \sigma_y^2\left(\text{tr}(\boldsymbol{E}\bar{\boldsymbol{\Sigma}}) + \int_0^\infty |\boldsymbol{\Delta}|\,\left(t\,f_2(t) - 2f_1(t)\right)dt\right), \tag{22}$$

where $|\boldsymbol{\Delta}|$ denotes the determinant of matrix $\boldsymbol{\Delta} = (\boldsymbol{I}_{2(T+1)}+2t\boldsymbol{A})^{-1/2}$, and the two functions f_1 and f_2 are given by

$$f_1(t) = \text{tr}(\boldsymbol{A}^*)\text{tr}(\boldsymbol{C}^*) + 2\text{tr}(\boldsymbol{A}^*\boldsymbol{C}^*), \tag{23}$$
$$f_2(t) = 2\text{tr}(\boldsymbol{A}^*\boldsymbol{A}^*)\text{tr}(\boldsymbol{D}^*) + 8\text{tr}(\boldsymbol{A}^*\boldsymbol{A}^*\boldsymbol{D}^*) + \text{tr}(\boldsymbol{A}^*)^2\text{tr}(\boldsymbol{D}^*) + 4\text{tr}(\boldsymbol{A}^*)\text{tr}(\boldsymbol{A}^*\boldsymbol{D}^*), \tag{24}$$

with $\boldsymbol{A}^* = \boldsymbol{A}\bar{\boldsymbol{\Sigma}}\boldsymbol{W}$, $\boldsymbol{C}^* = \boldsymbol{C}\bar{\boldsymbol{\Sigma}}\boldsymbol{W}$ and $\boldsymbol{D}^* = \boldsymbol{D}\bar{\boldsymbol{\Sigma}}\boldsymbol{W}$ where $\boldsymbol{W} = (\boldsymbol{I}_{2(T+1)} + 2t\boldsymbol{B}\bar{\boldsymbol{\Sigma}})^{-1}$.

The three terms we obtain in Eq. (22) pertain to the decomposition obtained in (19), i.e. $\text{tr}(\boldsymbol{E}\bar{\boldsymbol{\Sigma}})$ is the matrix representation of the variance of $y_{t+k} - \bar{y}$, while f_1 relates to the cross term, and f_2 stems from the quadratic term.[8]

In the literature on predictive models, a common yardstick for accuracy assessment is the out-of-sample R^2 which, in our setting, is equal to $R^2 := 1 - L/\text{var}(y^*_{t+k})$. Replacing the variance of y^*_{t+k} with its expression, item 4 of Theorem 2 leads to

$$R^2(T,k,\rho_x,\rho_y,\rho) = 1 - (1-\rho_y^2)\left(\text{tr}(\boldsymbol{E}\bar{\boldsymbol{\Sigma}}) + \int_0^\infty |\boldsymbol{\Delta}|\,\left(t\,f_2(t) - 2f_1(t)\right)dt\right), \tag{25}$$

so that the R^2 does not depend on σ_y^2. In the sequel of the paper, we will focus on the R^2 as it is independent of σ_y^2, but it can easily be translated into the loss L. Typically, if both the intercept a and the slope b of the linear model are known, the optimal R^2 is equal to

$$R^2_o := R^2_o(\rho_x,\rho_y,\rho,k) = \rho_y^{2k}\frac{\rho^2}{(1-\rho_x\rho_y)^2}(1-\rho_x^2)(1-\rho_y^2), \tag{26}$$

which is simply an affine transform of Eq. (12). There are two natural takeaways from this formula. First, R^2_o is never negative for stationary processes and $R^2_o = 0$ when $\rho = 0$. Intuitively, if both processes are uncorrelated, there is not much to learn from one another. The second important property is that R^2_o increases with $|\rho|$: when the link between the two processes intensifies, the precision of predictions improves. Motivated by these two observations (along with our empirical results below), we introduce two definitions for "*unexpected gains*" for the agent which we will henceforth use in the remainder of the paper.

Definition 3. With the out-of-sample $R^2(T,k,\rho_x,\rho_y,\rho)$ defined in Eq. (25), we say that the agent benefits from "unexpected gains" of:

Type I: if $R^2(T,k,\rho_x,\rho_y,0) > 0$ when $\rho = 0$;
Type II: if $R^2(T,k,\rho_x,\rho_y,\rho)$ increases when $|\rho|$ decreases.

Our definitions of unexpected gains differ from those of spurious regressions studied in the econometrics literature when regressing an independent random walk (RW) on another uncorrelated one yields a statistically significant linear relationship. In our study, we focus on stationary processes although the autocorrelation parameters can be close to 1 to recover RW-type behaviors.

Contrary to Eq. (26) where the expression of the R^2 is simple when the parameters a and b of the predictive regression are known, in the present paper, these parameters have to be estimated with a finite sample of size T, leading to a far more complex expression for the R^2. The theorem below sheds light on the behavior of the R^2 with respect to the correlation ρ between the two innovations.

Before we can formulate the result, we need to specify a matrix decomposition at the heart of some simplifications. If we denote with $\bar{\boldsymbol{\Sigma}}^{\boldsymbol{x}}_{\boldsymbol{T+1}}$ the covariance matrix $\boldsymbol{N}_{T+1}\boldsymbol{\Sigma}^{\boldsymbol{x}}_{\boldsymbol{T+1}}\boldsymbol{N}'_{T+1}$ of the first $T+1$ elements of the Gaussian vector $\bar{\boldsymbol{z}}$, then a direct computation shows that

$$\bar{\boldsymbol{\Sigma}}^{\boldsymbol{x}}_{\boldsymbol{T+1}} = \begin{bmatrix}\bar{\boldsymbol{\Sigma}}^{\boldsymbol{x}}_T & \bar{\boldsymbol{\epsilon}}_{\boldsymbol{x}}\\ \bar{\boldsymbol{\epsilon}}'_{\boldsymbol{x}} & \bar{s}^2_x\end{bmatrix},$$

where the components are given by

$$\bar{\boldsymbol{\Sigma}}^{\boldsymbol{x}}_T = \boldsymbol{M}_T\boldsymbol{\Sigma}^{\boldsymbol{x}}_T\boldsymbol{M}_T,$$
$$\bar{\boldsymbol{\epsilon}}_x = \boldsymbol{M}_T\left(\boldsymbol{\epsilon}_{\boldsymbol{x}} - \frac{1}{T}\boldsymbol{\Sigma}^{\boldsymbol{x}}_T\boldsymbol{1}_{T,1}\right),$$
$$\bar{s}^2_x = \frac{1}{T^2}\boldsymbol{1}_{1,T}\boldsymbol{\Sigma}^{\boldsymbol{x}}_T\boldsymbol{1}_{T,1} - \frac{2}{T}\boldsymbol{1}_{1,T}\boldsymbol{\epsilon}_{\boldsymbol{x}} + (1-\rho_x^2)^{-1}.$$

With these notations at hand, we can proceed to our second major result.

Theorem 4. *Under the same assumptions as those of Theorem 2, the accuracy of the predictive model measured by the R^2, defined in (25), is a symmetric quartic function of the correlation ρ between the innovations of the predictor x^* and the predicted y^* processes and takes the following form:*

$$R^2(T,k,\rho_x,\rho_y,\rho) = 1 - (1-\rho_y^2)\left(\bar{s}^2_y + \int_0^\infty |\boldsymbol{\Delta}_T|\,\left(g_4(t)\rho^4 + g_2(t)\rho^2 + g_0(t)\right)dt\right) \tag{27}$$

[8] When the means of processes x^* and y^* are known, we can replace the sample means $\bar{x}^*$ and $\bar{y}^*$ in Eq. (13) by the true means or equivalently replace the sample means $\bar{x}$ and $\bar{y}$ in Eq. (17) by 0. Therefore, Theorem 2 remains true but with the simplification $\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$, which is equivalent to taking $\boldsymbol{N} = \boldsymbol{I}_{2(T+1)}$.

where $|\boldsymbol{\Delta}_T|$ denotes the determinant of matrix $\boldsymbol{\Delta}_T = (\boldsymbol{I}_T + 2t\boldsymbol{\Lambda}_T)^{-1/2}$ built from the diagonal matrix $\boldsymbol{\Lambda}_T$ containing the eigenvalues of $\bar{\boldsymbol{\Sigma}}^{\boldsymbol{x}}_T$, and the functions g_4, g_2 and g_0 only depend on the sample size T, the lag k and the two auto-correlations ρ_x and ρ_y.

Similar to Eq. (25) where a and b are known, this result confirms that, when these parameters are unknown and estimated with a finite sample size, the R^2 is also a symmetric function of ρ and more precisely a second order polynomial of ρ^2. It also justifies our choice to focus on positive values for ρ in the numerical section below.

Our definition of unexpected gains of Type I focuses on positive levels of R^2 in regressions when the innovations between the two processes is equal to 0, implying that the processes are also uncorrelated. Therefore, the proposition below underlines the simplifications that occur in the absence of correlation. In particular, we are able to derive simpler explicit expressions for $|\boldsymbol{\Delta}_T|$ and for the functions f_1 and f_2.

Proposition 5. *Under the same assumptions as those of Theorems 2 and 4 but with the additional assumption that the two processes x^* and y^* have uncorrelated innovations, i.e., $\rho = 0$, the accuracy of the predictive model measured by the R^2, defined in (25), satisfies the following properties:*

1. *R^2 is given by*

$$R^2(T,k,\rho_x,\rho_y) = 1 - (1-\rho_y^2)\left(\bar{s}_y^2 + \int_0^\infty |\boldsymbol{\Delta}_T|\,\left(t\,f_2(t) - 2f_1(t)\right)dt\right), \tag{28}$$

where the two functions f_1 and f_2 simplify into

$$f_1(t) = \bar{\boldsymbol{\epsilon}}'_{\boldsymbol{x}}\boldsymbol{\Omega}\bar{\boldsymbol{\epsilon}}_{\boldsymbol{y}}, \tag{29}$$

$$f_2(t) = \mathrm{tr}(\bar{\boldsymbol{\Sigma}}^{\boldsymbol{x}}_T\boldsymbol{\Omega}\bar{\boldsymbol{\Sigma}}^{\boldsymbol{y}}_T)\left(\bar{s}_x^2 - 2t\,\bar{\boldsymbol{\epsilon}}'_{\boldsymbol{x}}\boldsymbol{\Omega}\bar{\boldsymbol{\epsilon}}_{\boldsymbol{x}}\right) + 2\bar{\boldsymbol{\epsilon}}'_{\boldsymbol{x}}\boldsymbol{\Omega}\bar{\boldsymbol{\Sigma}}^{\boldsymbol{y}}_T\boldsymbol{\Omega}\bar{\boldsymbol{\epsilon}}_{\boldsymbol{x}}, \tag{30}$$

where $\boldsymbol{\Omega} = (\boldsymbol{I}_T + 2t\bar{\boldsymbol{\Sigma}}^{\boldsymbol{x}}_T)^{-1}$.

2. *If we denote with $\boldsymbol{P}_T$ the orthogonal matrix such that $\bar{\boldsymbol{\Sigma}}^{\boldsymbol{x}}_T = \boldsymbol{P}_T\boldsymbol{\Lambda}_T\boldsymbol{P}'_T$ and with $\lambda_1,\dots,\lambda_T$ the eigenvalues of $\bar{\boldsymbol{\Sigma}}^{\boldsymbol{x}}_T$ stacked on the diagonal of $\boldsymbol{\Lambda}_T$, then the determinant of $\boldsymbol{\Delta}_T$ and the functions f_1 and f_2 defined in Eqs. (29) and (30) are equal to*

$$|\boldsymbol{\Delta}_T| = \prod_{j=1}^{T}\left(1+2t\lambda_j\right)^{-\frac{1}{2}} \tag{31}$$

$$f_1(t) = \sum_{j=1}^{T}\frac{p_j q_j}{1+2t\lambda_j}, \tag{32}$$

$$f_2(t) = \sum_{j=1}^{T}\frac{Q_{jj}\lambda_j}{1+2t\lambda_j}\left(\bar{s}_x^2 - 2t\sum_{j=1}^{T}\frac{p_j^2}{1+2t\lambda_j}\right) + 2\sum_{i=1}^{T}\sum_{j=1}^{T}\frac{Q_{ij}p_i p_j}{\left(1+2t\lambda_i\right)\left(1+2t\lambda_j\right)}, \tag{33}$$

where the two vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ denote the quantities $\boldsymbol{P}'_T\bar{\boldsymbol{\epsilon}}_{\boldsymbol{x}}$ and $\boldsymbol{P}'_T\bar{\boldsymbol{\epsilon}}_{\boldsymbol{y}}$ respectively and matrix $\boldsymbol{Q}$ denotes the product $\boldsymbol{P}'_T\bar{\boldsymbol{\Sigma}}^{\boldsymbol{y}}_T\boldsymbol{P}_T$.

The results[9] of Proposition 5 and especially Eqs. (32) and (33) are of particular interest. Indeed, we see that Eq. (28) involves quantities of the form $I = \int_0^\infty h(t)dt$ and, splitting the integral into two terms $I = \int_0^u h(t)dt + \int_u^\infty h(t)dt$ allows us to compute the first part via Riemann trapezoidal sums and, if u is large enough, to approximate the second term as the integral of a polynomial function of t where we infer from both Eqs. (32) and (33) the behavior of the two integrands when t converges to infinity.[10]

[9] When the means of processes x^* and y^* are known, Proposition 5 still holds but with the simplifications $\bar{\boldsymbol{\Sigma}}^{\boldsymbol{x}}_{T+1} = \boldsymbol{\Sigma}^{\boldsymbol{x}}_{T+1}$ and $\bar{\boldsymbol{\Sigma}}^{\boldsymbol{y}}{}_{T+1} = \boldsymbol{\Sigma}^{\boldsymbol{y}}_{T+1}$ or equivalently $\bar{\boldsymbol{\Sigma}}^{\boldsymbol{x}}_T = \boldsymbol{\Sigma}^{\boldsymbol{x}}_T$, $\bar{\boldsymbol{\Sigma}}^{\boldsymbol{y}}{}_T = \boldsymbol{\Sigma}^{\boldsymbol{y}}_T$, $\bar{\boldsymbol{\epsilon}}_x = \boldsymbol{\epsilon}_x$, $\bar{\boldsymbol{\epsilon}}_y = \boldsymbol{\epsilon}_y$, $\bar{s}_x^2 = (1-\rho_x^2)^{-1}$ and $\bar{s}_y^2 = (1-\rho_y^2)^{-1}$.

[10] More details about the behavior of the two integrands when t converges to infinity are given in Appendix G.

Our previous theoretical results involve integrals which are hard to interpret. It is consequently difficult to understand the impact of the different parameters on the accuracy measure, R^2. In the following proposition, we propose to assume that the innovations of the two processes are uncorrelated, but also that the predictor is a white noise process, i.e. $\rho_x = 0$. In that case, we will see that the computations are tractable since we are able to derive a closed-form expression for the R^2 but also there are a set of parameter values, i.e. T, k and ρ_y for which we observe strong unexpected gains of Type I.

Proposition 6. *Consider two AR(1) processes given by (2) and (3) with uncorrelated innovations ($\rho = 0$), and the predictive regression model (4). If we assume that the predictor is a white noise process, i.e. $\rho_x = 0$, then the accuracy of the predictive model measured by the R^2, satisfies the following properties:*

1. *R^2 is defined for a sample size $T > 3$ and given by*

$$R^2(T,k,\rho_x=0,\rho_y) = \frac{2\rho_y^k}{T}\frac{1-\rho_y^T}{1-\rho_y} - \left(1 - \frac{T+1}{(T-1)(T-3)}\right) \times \frac{1}{T^2}\left(2\frac{T - T\rho_y + \rho_y^{T+1} - \rho_y}{(1-\rho_y)^2} - T\right) - \frac{T+1}{(T-1)(T-3)}. \tag{34}$$

2. *Asymptotically, using the Landau notation,*

$$R^2(T,k,0,\rho_y) = O(T^{-1}), \qquad T\uparrow\infty, \tag{35}$$

$$R^2(T,k,0,\rho_y) = 1 + O(1-\rho_y), \qquad \rho_y\uparrow 1, \tag{36}$$

$$R^2(T,k,0,\rho_y) = -2\frac{T-1}{T(T-3)} + O(\rho_y), \qquad \rho_y\downarrow 0. \tag{37}$$

From Eq. (34), we see that the dependence in k is simple: as k increases, the R^2 decreases (assuming $\rho_y \in (0,1)$, which is often the case empirically). However, the impacts of the sample size T and the dependent variable autocorrelation ρ_y are non-trivial. We also observe that the R^2 can take both negative and positive values where the latter automatically translates into unexpected gains of Type I. A surprising result is the convergence of the R^2 to 1 when the autocorrelation in the dependent variable converges towards 1. This result is independent of the sample size and of the horizon k and illustrates potentially strong unexpected gains of Type I for highly persistent dependent variables. Nevertheless, as will be shown subsequently in our empirical results, the convergence of R^2 to one is both slow and depends on T. For reasonable sample sizes ($T > 10$) and for nontrivial horizons ($k > 1$), the R^2 will in fact never be in the vicinity of 1, even when $\rho_y = 0.96$ or $\rho_y = 0.99$,[11] and sometimes it will even be well below zero.

In Fig. 1, we plot the R^2 as a function of T for several values of k and ρ_y. It clearly shows the parametric regions that give rise to unexpected gains of Type I. Simply put, the combination of a high autocorrelation for y with a small sample size will most likely lead to unexpected gains for the agent even when x and y are uncorrelated. These gains tend to vanish when k increases.

3. Numerical experiments

This section illustrates the sensitivity of the R^2 to the model parameters and infers the configurations of the underlying DGP for which the agent's misspecified predictive regression leads to unexpected gains using simulated data. First, we validate the analytical formulae from Section 2.5 by running alternative Monte Carlo simulations and compare the computational efficiency of our closed-form expressions to the simulations.

[11] Results available upon request.

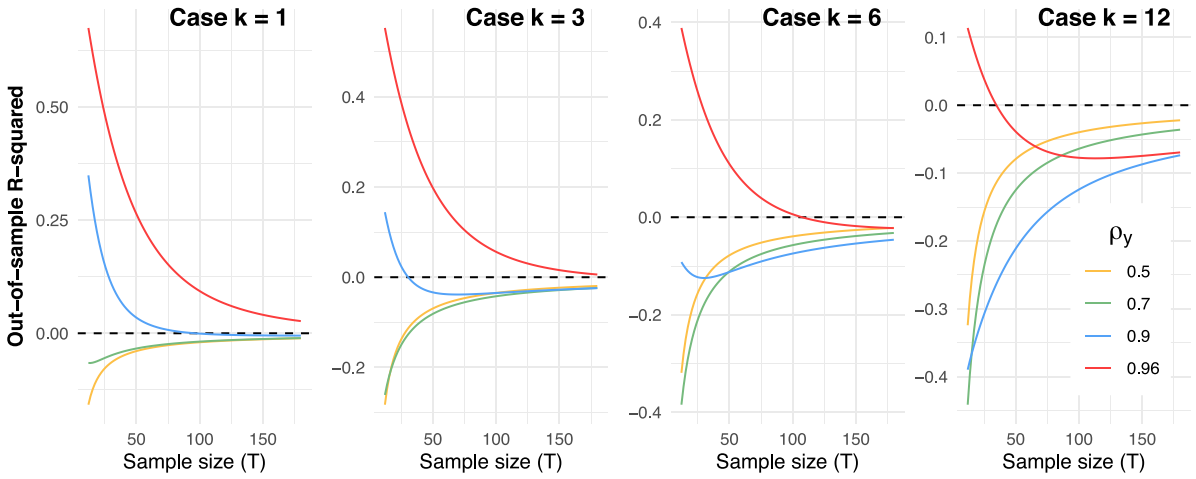

Fig. 1. Out-of-sample R^2 when $\rho = \rho_x = 0$. We plot the theoretical out-of-sample R^2 as a function of three variables: k (for each column of subplots), sample size T on the x-axis and the persistence of labels, ρ_y, is shown with colors. Unexpected gains occur when the curves are above the zero threshold, marked with a black dotted horizontal line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Methods and computational efficiency

We test three numerical routes to calculate the loss defined in Eq. (9). For simplicity, we will refer to them with one word:

- **process**: we run N simulations of the processes x^* and y^*, based on the horizon k and the sample size T. The total number of points is $T+2k$ and the "present" date is $T+k$ (required to avoid any forward-looking bias). The in-sample predictors correspond to the first T points. The in-sample predicted variable is indexed between $k+1$ and $T+k$. The estimates (6) and (5) are computed from these samples. Then, based on these estimated values and on the present predictor value $x^{(n)}_{T+k}$, we predict the k-period ahead value $\hat{y}^{(n)}_{T+2k}$, which we compare to $y^{(n)}_{T+2k}$, which is the actual simulated value. For each simulation, we compute the OLS estimates and the related squared error. Then, we average these errors, which corresponds to the Monte-Carlo estimation of Eq. (9):
$$\widehat{MSE}(N,T,k) = N^{-1}\sum_{n=1}^{N}(\hat{y}^{(n)}_{T+2k} - y^{(n)}_{T+2k})^2;$$
- **multivariate**: we run N simulations of the Gaussian variate $\boldsymbol{z}$ according to the covariance matrix (20). This allows to sample the term which is within the expectation of Eq. (18) and which we call $\hat{\Lambda}^{(n)}$ here for the nth simulation. We then have
$$\widehat{MSE}(N,T,k) = \sigma_y^2 N^{-1}\sum_{n=1}^{N}\hat{\Lambda}^{(n)};$$
- **integral**: we use the pseudo closed-form solution (22) with numerical evaluation of the integral - via Riemann trapezoidal sums. In the case where the two processes have uncorrelated innovations (i.e. $\rho = 0$), we can resort to the simpler form of Eq. (28).

The first two methods are straightforward to implement and both heavily rely on Monte Carlo simulations. The last one requires further clarification which we provide in Appendix G. One of the purposes of the **integral** method is *verification*. The formulae in the above propositions and theorems need to be confirmed numerically. As they rely on one integral, we check if the integral values match the simulation ones.

In Fig. 2, we plot the loss values obtained for different levels of discretizations (for the **integral** method) and Monte Carlo (MC) scenarios (for both the **process** and **multivariate** methods). The sample size is equal to $T = 36$ and the lag equal to $k = 12$ in the left panel.[12] Given the large number of discretization points, the integral method yields a constant result. In contrast, the other methods generate oscillating patterns, which is a clear indication of the superiority of the integral formulae. In the right panel, we plot the computation times (on CPU) obtained for different training sample sizes T. The **multivariate** approach is not shown because it is too slow and not competitive.

Because the speed of convergence of MC averaging is $N^{-1/2}$, the magnitude of the error is expected to have an order of 10^{-3} for a number of points equal to 10^6 — which is what we observe to the right of the left panel. To further confirm such speed we proceed to a complementary exercise in which we repeat the simulation procedures M times, yielding M mean squared errors $\widehat{MSE}_M(N,T,k)$. For each simulation length N, sample size T, and horizon k, we can then evaluate the quantiles of the MSE, across M. In the left panel of Figure S1 in Appendix A.1, we depict the 2.5% and 97.5% quantiles which we obtained from $M = 200$ batches of such MSE values. The parameters are $T = 36$ and $k = 12$. These quantiles form the 95% confidence intervals and therefore, the plot confirms both the patterns and orders of magnitudes from the confidence levels from Fig. 2. Moreover, for the sake of completeness, we also report in the right panel of Figure S1 the smoothed distributions of the $\widehat{MSE}_m(N,T,k)$ across $m = 1,\dots,200$, for five values of N. As N increases, the densities clearly converge to a Dirac distribution centered on the true (asymptotic) value of the MSE.

Computation times (CPU) are exceedingly large for the **multivariate** method which is the reason why we discard this method from the right panel of Fig. 2 and from the remaining of the paper. One reason for this is that it involves lengthy matrix operations, which is not the case for the **process** technique. In terms of computation resources, the **integral** and **process** are hardly comparable. The fastest one is the integral evaluation, by far, which, as expected, highlights the gains stemming from closed-form expressions.

3.2. Unexpected gains

In Theorem 4, we show that the R^2 is linked to ρ via a simple polynomial form with even degrees, namely

$$R^2(\rho) = c_4\rho^4 + c_2\rho^2 + c_0.$$

According to Definition 3, a case of unexpected gains of Type I occurs when $R^2(0) = c_0 > 0$ and a case of Type II happens if $(R^2)'(\rho) =$

[12] Results for other sample sizes are not particularly enlightening. We invite the interested reader to check the online *comparison* notebook on the paper's website: http://www.gcoqueret.com/PRs.html for further details.

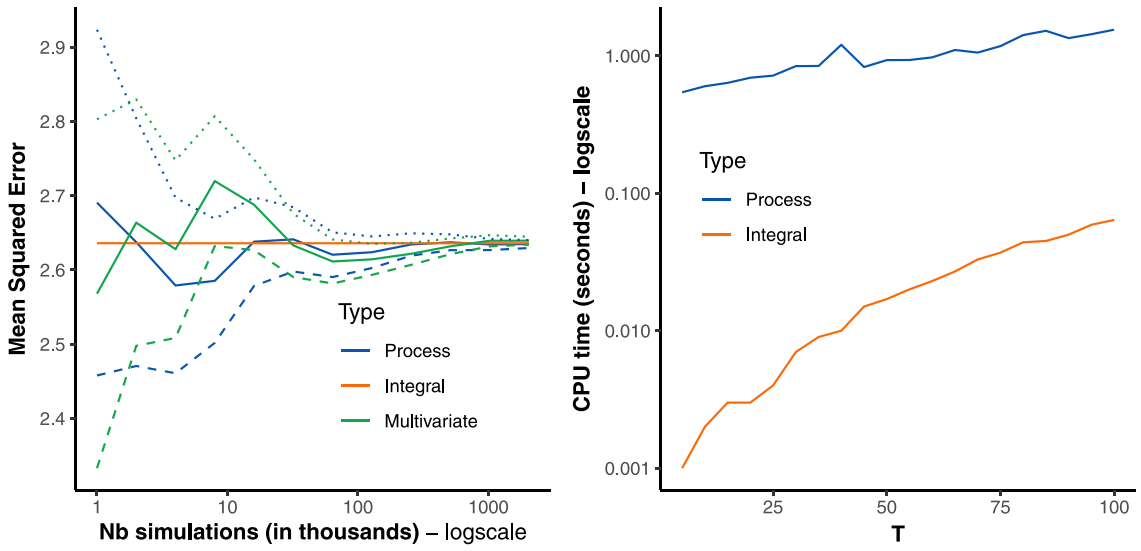

Fig. 2. Comparison of methods — speed and accuracy. In the left panel, we plot the values obtained for the MSE as a function of the computational cost N expressed by the number of Monte Carlo simulations (for both the **process** and **multivariate** methods) or the number of discretization points in the Riemann integral (for the **integral** method). The cost ranges from $N = 10^3$ up to $N = 10^3 \times 2^{11} \approx 2 \times 10^6$, with increments of powers of 2 — which explains the logarithmic scale for the x-axis. We use a training sample size $T = 36$ — equivalent to 3 years of monthly data. The dotted and dashed lines show the upper and lower bounds of the 95% confidence intervals, defined as ± 1.96 times the standard deviations of simulated squared errors, divided by $\sqrt{N}$. In the right graph, we show the CPU time as a function of the training sample size T using a number of points equal to 10^3 for the discretization of the **integral** method and a number of simulations equal to 10^5 for the **process** method. The other parameters are fixed and equal to $\rho_x = 0.9$, $\rho_y = 0.7$, $\rho = 0$ and $k = 12$. For the Integral method, the upper bound for the trapezoidal approximation is taken equal to $u = 150\lfloor \log(N) \rfloor$ — see Appendix G for more details.

$2\rho(2c_4\rho^2 + c_2) < 0$ for $\rho \in (0, 1)$. Unfortunately, the expressions for the coefficients c_2 and c_4 are cumbersome to say the least and there is no simple algebraic way to determine under which conditions they may be positive or negative.

In Fig. 3, we plot the R^2 as a function of sample size T, horizon k, and autocorrelations ρ_x and ρ_y. We restrict the results to the case $\rho = 0$ to see if and when the R^2 is nonzero. This is expected to reveal cases of unexpected gains of Type I. As it turns out, the latter do often occur, and there is a clear sample size effect. When T is large, the magnitude of R^2 is small, which is what we would expect from Eq. (26). When $\rho = 0$, the model captures only noise, so the R^2 should be negligible. However, when T is small, the R^2 can be surprisingly large, even above 0.5 in some cases when the dependent variable is very persistent. Nevertheless, we do also observe negative values for the R^2, especially when ρ_y is small. As the horizon k increases, positive R^2 become more scarce. Our results hold for various levels of persistence ρ_x in the predictor which shows that it has more marginal importance. Finally, when $\rho_x = 0.1$ (leftmost column of graphs), we recover the patterns of Fig. 1 in which ρ_x was equal to zero.

We propose another angle in Fig. 4 by plotting the R^2 as a function of ρ in order to investigate Type II unexpected gains. While the curves with the steepest slopes match the expected quadratic or quartic shape derived in Theorem 4, the surprising feature is that some of them are indeed *decreasing* and this only happens with small samples. This is somewhat counter-intuitive because these cases correspond to situations when decreasing the correlation between the innovations (and hence increasing the misspecification in the predicted regression) improves the out-of-sample fit. One common feature between the two types is that they both require a high level of persistence in the dependent variable: unexpected gains only occur when ρ_y is sufficiently high. However unexpected gains of Type II also require a level of high persistence in the predictor (see both lower panels), which means that both the dependent and independent variables should behave like random walks as for the Grange–Newbold regressions of two independent random walks. This shows our definition of unexpected gains of Type II is similar to the classical spurious regression of independent random walks studied in the econometrics literature.

We underline that both types of unexpected gains are not incompatible. When ρ_y and ρ_x are high and the sample size is small ($T = 6$, $T = 12$), some curves in Fig. 4 corresponding to $k = 3$ and $k = 6$ start slightly above zero for $\rho = 0$ and subsequently decrease when ρ increases.

3.3. Convergence towards theoretical optimality

Our final analysis pertains to the convergence towards asymptotic values defined in Eq. (26). This phenomenon is shown in Fig. 3, i.e., as the sample size increases, the R^2 seems to stabilize towards some constant, which in this configuration, is zero (because $\rho = 0$). In the general case, when k increases to $+\infty$, or when ρ or ρ_y decrease to 0, the asymptotic R^2_o in Eq. (26) is exactly equal to zero. When the training sample increases, we would expect that, despite the absence of independence in the training data set, the estimated coefficient of the models is consistent with the classical OLS form defined by Eq. (11) for $\hat{b}$. The lemma below states a result in this direction.

Lemma 7. *It holds that*

$$\hat{b}(\mathcal{D}_t^*) \xrightarrow[T\to+\infty]{P} b_o,$$

where $\xrightarrow{P}$ *stands for convergence in probability and* b_o *is defined in* (11).

Under an additional integrability condition, the R^2 is thus expected to converge to its corresponding value R^2_o (see Eq. (26)). This is an important subject, because most R^2 we report in the previous subsections are negative, whereas the R^2_o are not, especially if ρ is far enough from zero.

Overall, the lines in Fig. 4 reveal a wide array of situations. The $R^2 := R^2(\rho, T)$ can be locally (depending on k, ρ_x and ρ_y):

- increasing in ρ and T;
- decreasing in ρ and T;
- increasing in ρ and decreasing in T or vice-versa.

For a given set of parameters k, ρ_x, ρ_y and ρ, the value of the R^2 for the smallest sample size (say $T = 3$) can be either *above* or *below* R^2_o:

$$\underbrace{R^2_{\rho_x,\rho_y,\rho,k}(3)}_{\text{small sample}} \lessgtr \underbrace{R^2_o(\rho_x, \rho_y, \rho, k)}_{\text{infinite sample}}$$

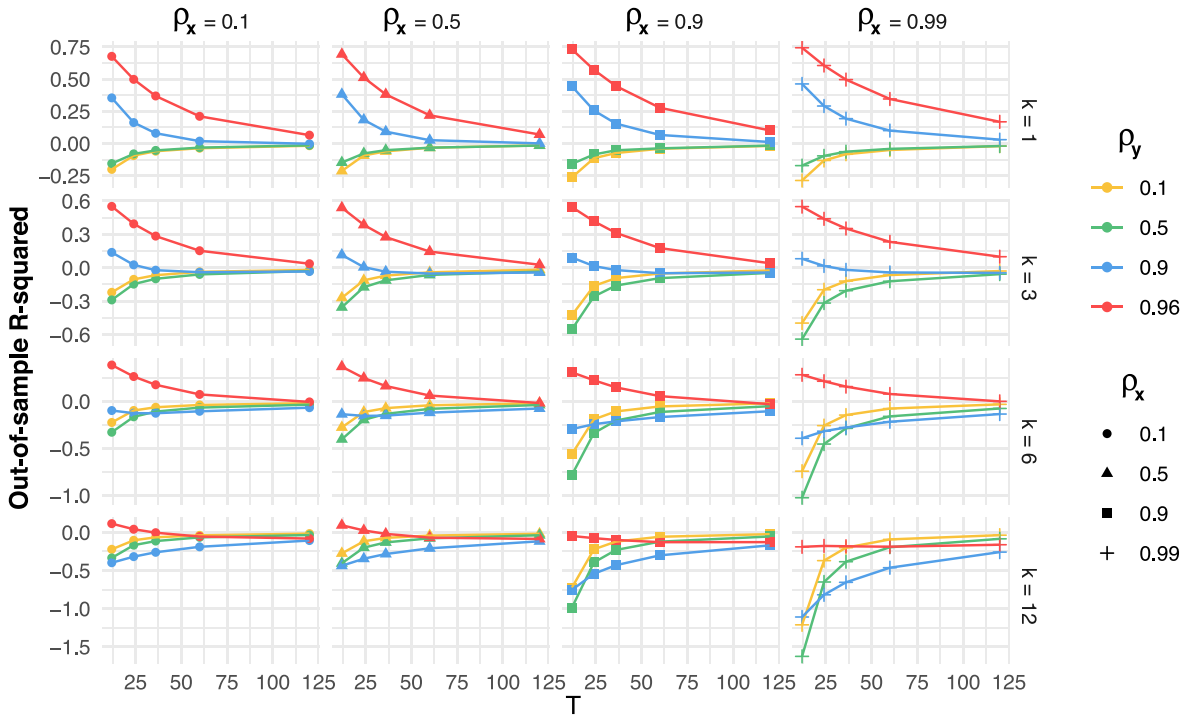

Fig. 3. Unexpected gains of Type I: the case $\rho = 0$. We plot the out-of-sample R^2 as a function of four variables: k (for each row of subplots), sample size T on the x-axis, and persistence of the predictor ρ_x (columns of subplots). The persistence of labels, ρ_y, is shown with colors. he points were obtained via the discretization of the integral of Eq. (28). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Depending on the case, smaller or larger samples will be beneficial because the R^2 will either increase or decrease to the asymptotic values R^2_o. In the situations when $R^2_o > 0$, the estimates are expected to push the R^2 above zero when the agent is endowed with enough data T is large enough.

The interesting question is then: when will that be the case? While we cannot provide a definite answer, a reorganization of the results of Fig. 4 with T on the x-axis gives interesting cues (see Fig. 5). It appears that the most important parameter is ρ_y: when it is high, small samples seem preferable, but when it is small, then long samples yield better results.

4. Empirical evidence

4.1. Data

One of the academic standards in return predictability is the study of Welch and Goyal (2008), which was recently updated in Goyal et al. (2021). Consequently, we resort to the updated version of their dataset for our empirical analysis.[13]

Since the main purpose of the paper is to study the impact of predictor persistence, we remove a few variables because many are highly autocorrelated. This would generate unnecessary redundancies in the results. We also complement our dataset with one of the mildly persistent predictors tested in Novy-Marx (2014): **temp**, the index of global temperature anomalies.[14] The list of all variables is provided in Table 1.

Table 1
List of predictors. We list the predictors used in the empirical study, along with their sources. We keep the same compact notation as in Welch and Goyal (2008), from which most predictors originate. **temp** is documented in Novy-Marx (2014).

Short name	Brief description	Academic source
bm	Aggregate book-to-market ratio	Welch and Goyal (2008)
de	Log of dividend-earnings ratio	Welch and Goyal (2008)
dfr	Default return spread	Welch and Goyal (2008)
dfy	Default yield spread (BAA minus AAA)	Welch and Goyal (2008)
dp	Log of dividend-price ratio	Welch and Goyal (2008)
ep	Log of earnings-price ratio	Welch and Goyal (2008)
ltr	Long term rate of bond returns	Welch and Goyal (2008)
lty	Long term government bond yield	Welch and Goyal (2008)
svar	Stock variance	Welch and Goyal (2008)
temp	Global temperature	Novy-Marx (2014)
tbl	Treasury bill rate	Welch and Goyal (2008)
tms	Bond term spread	Welch and Goyal (2008)

To produce more diversity in autocorrelation, we introduce the difference in the first three of these variables (dp, ep and de). The new variables are simply defined by $\Delta x_t = x_t - x_{t-1}$. The dependent variables will be returns on the S&P500 index minus returns on US Treasury bills. We choose four horizons (k) for these returns: 3 months, 6 months, 12 months, and 24 months. We exclude monthly returns because their autocorrelation is equal to 12%, and as shown in Section 3.2, unexpected gains only occur for high levels of persistence.

The descriptive statistics of all variables used in the study can be found in Table 2. In addition, we provide the histograms of estimated innovations in AR(1) models for all the variables in the study in Figure S2 of Appendix A.2. The correlation between the innovations of y and x being an important feature of our study, we plot their (bimodal) distribution for each horizon of returns in Figure S3 in Appendix A.3. Finally, we show the link between the correlation of innovations and the correlation of processes in the right panel of the same figure and observe that they are both strongly positively linked.

[13] Available on Amit Goyal's website: https://sites.google.com/view/agoyal145.

[14] The other climate-related variables in Novy-Marx (2014) are highly persistent and do not add value, we thus omit them. The Combined Land-Surface Air and Sea-Surface Water Temperature Anomalies can be accessed at https://data.giss.nasa.gov/gistemp/tabledata_v4/GLB.Ts+dSST.csv.

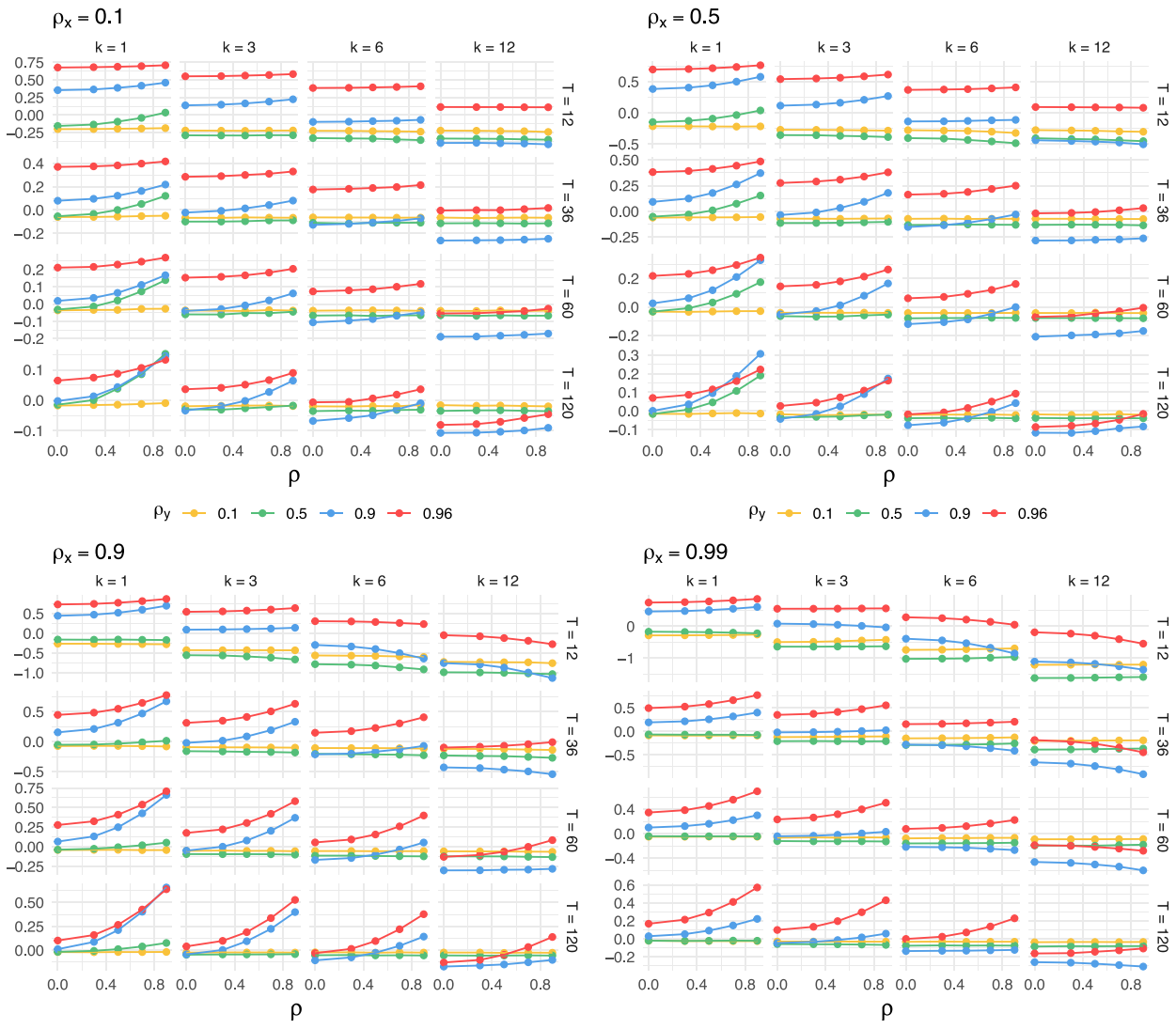

Fig. 4. Unexpected gains of Type II: R^2 as a function of ρ. We plot the out-of-sample R^2 as a function of five variables: the innovations correlation ρ on the x-axis, the sample size T and horizon k (rows and columns of subplots), and ρ_y is shown with colors and ρ_x varies in each panel. Because the R^2 is symmetric in ρ, we only show the right part of the support of ρ. The points were obtained from the **process** simulation method described in Section 3.1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Baseline results

We briefly describe the protocol we follow. For each predictor x, label y (equivalently, horizon k), sample size T, and point in time t where the training sample $\mathcal{D}_t$ is well defined (i.e., with no missing point), we:

1. estimate Eq. (2) for x and (3) for y over a past sample of 120 months (10 years), this yields $\hat{\rho}_{x,t}$, $\hat{\rho}_{y,t}$, $\hat{\rho}_t$, $\hat{\sigma}^2_{x,t}$ and $\hat{\sigma}^2_{y,t}$, which are local estimates for the underlying processes[15];
2. estimate Eq. (4) based on the sample $\mathcal{D}_t$ of size T (and store the residuals);
3. make a prediction for y_{t+k} based on $\hat{a}$ and $\hat{b}$;
4. store the error with respect to the realized y_{t+k}: $e_t(x, y, T)$.

From the above quantities, we are able to compute the out-of-sample R^2, defined as

$$R^2_{\text{oos}}(x, y, t) = 1 - \frac{\sum_{s=1}^{S} e_{t-s}(x, y, T)^2}{\sum_{s=1}^{S} (y_{t-s+k} - \bar{y})^2}, \tag{38}$$

where $s = 1, \dots, S$ are the indices of the out-of-sample dates pertaining to the sample. In order to have sufficient granularity, we compute R^2_{oos} for each decade between 1900 and 2019 so that S is equal to 120.

We report results for six choices of T: 12, 24, 36, 60, 84 and 120 months. Over all combinations of predictors, labels, sample sizes, and dates, we obtain 591,840 predictions. In the spirit of Campbell and Thompson (2008) and Pettenuzzo et al. (2014), we remove the predictions that do not make any economic sense (i.e., that are smaller than −100% or larger than 300%). In Fig. 6, we plot the out-of-sample R^2 as a function of sample size. However, only the instances that correspond to negligible $\hat{\rho}$ are displayed. Each point relates to an estimate for $\hat{\rho}$ between −0.05 and +0.05. Therefore, the points above zero (dashed black line) can be considered as cases of unexpected gains of Type I (see Definition 3). Our results show that while a very large

[15] These estimates do not depend on T: very small samples yields estimates that are too noisy, which is why we resort to ten-year samples by default.

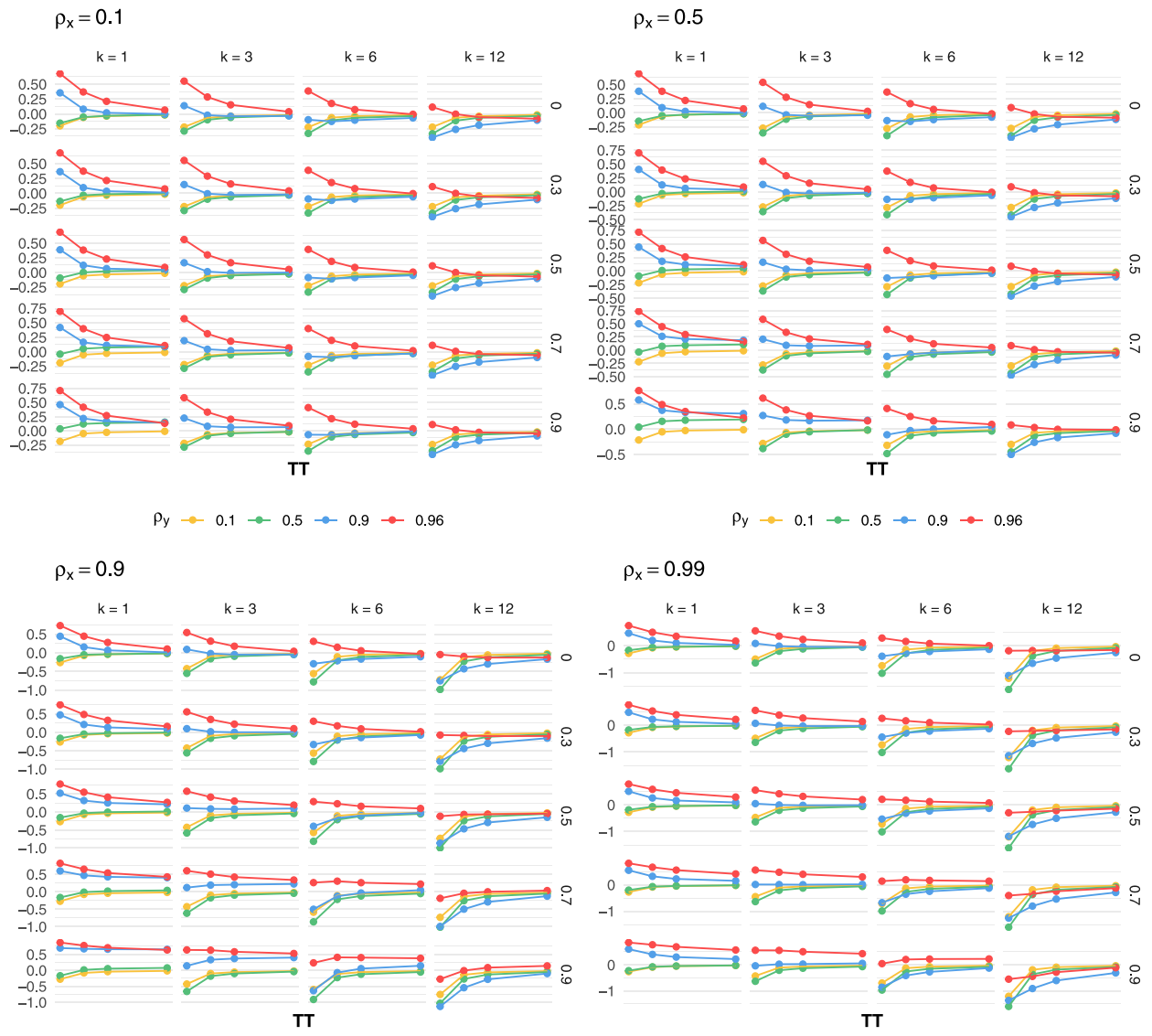

Fig. 5. R^2 as a function of T. We plot the out-of-sample R^2 as a function of five variables: the sample size T on the x-axis, the correlation in innovations ρ and horizon k (rows and columns of subplots), and ρ_y is shown with colors and ρ_x varies in each panel. The points were obtained from the **process** simulation method described in Section 3.1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

majority of points lie below the zero threshold (see the proportions, in black), a few cases do point towards the presence of unexpected gains.

In order to compare them with the results of Fig. 3, we recall that the horizon of returns corresponds to the time lag k therein. The low proportions of positive R^2_{oos} for large horizons (e.g., $k = 12$ months) are in line with the bottom panels in Fig. 3.

In Fig. 7, we turn to a study focused on unexpected gains of Type II. The theoretical patterns shown in Fig. 4 suggest that this type of gain will be more pronounced for persistent predictors, thus we split the analysis in two. In the top plots, we focus on predictors with low to moderate autocorrelation, while in the bottom ones, we provide the graphs for highly persistent independent variables (we refer to Table 2 for the clusters of predictors).

Clearly, there is a marked difference between the upper and lower panels for the slopes of the fitted linear relationships. While the latter are all positive in the upper plots (moderately persistent predictors), several of them are negative for highly persistent independent variables, especially for short-term returns and small sample sizes. For the sake of completeness, we provide in Table 3 the t-statistics associated with the slopes of the linear models fitted in Fig. 7. The statistics obtained are not always significant, meaning that the relationship between $\hat{\rho}$ and the out-of-sample R^2 is not clear cut, but the values change significantly from the class of predictors with low to moderate autocorrelation in Panel A to the class of predictors with high persistence in Panel B. Indeed, Panel A shows large positive t-statistics (among which 7 are significant) while Panel B shows 8 negative t-statistics for short horizons, e.g. when k is equal to 3 and 6 months.

4.3. VIX prediction

Our results imply that the presence of unexpected gains of Type I requires persistence, which is why in the baseline study, monthly returns are omitted. This prevents us from considering an interesting case, which combines a short horizon $k = 1$ to a high degree of persistence in the label. In Fig. 6, it is clear that this combination is likely to lead to unexpected gains of Type I. In Fig. 7, this is less clear and would require persistent predictors.

Unlike returns, proxies for turbulence such as realized volatility or forward-looking measures of risk are easier to predict because they are autocorrelated by construction (we refer to Paye (2012) on this topic). When switching to volatility prediction, we thus expect to obtain much higher R^2 on average, compared to those of Fig. 4. To test this conjecture, we run PRs in which the one month *ahead* VIX value is the

Table 2
Descriptive statistics of the dataset. In the last columns, $\hat{\rho}_v$ stands for the sample autoregressive parameter of variable v, and $\hat{\sigma}_v^2$ refers to the variance of the residuals of the corresponding estimation. The increment variables are defined as $\Delta v_t = v_t - v_{t-1}$ for each variable v.

Varable	Begins	Ends	Min	Max	Mean	Median	std. dev.	$\hat{\rho}_v$	$\hat{\sigma}_v^2$
PANEL A: Predictors (features)									
Panel A1: High persistence ($\hat{\rho}_x > 0.9$)									
bm	1921-03	2020-12	0.463	7.791	2.135	2.038	1.000	0.987	0.028
de12	1871-01	2020-12	−3.965	4.397	−1.752	−1.774	1.000	0.993	0.014
dfy	1919-01	2020-12	0.462	8.139	1.700	1.371	1.000	0.976	0.046
dp12	1871-01	2020-12	−10.183	−4.217	−7.278	−7.108	1.000	0.995	0.012
ep12	1871-01	2020-12	−12.700	−4.385	−7.047	−7.081	1.000	0.990	0.023
lty	1919-01	2020-12	0.229	5.484	1.842	1.560	1.000	0.997	0.007
tbl	1920-01	2020-12	0.003	5.473	1.133	1.012	1.000	0.993	0.014
tms	1920-01	2020-12	−2.819	3.515	1.240	1.244	1.000	0.963	0.072
Panel A2: Moderate persistence									
Δde	1871-02	2020-12	−19.684	14.195	−0.001	0.000	1.000	0.772	0.405
svar	1885-02	2020-12	0.000	14.122	0.491	0.238	1.000	0.570	0.676
temp	1881-01	2020-12	−0.749	9.483	0.127	−0.037	1.000	0.733	0.490
Panel A3: Low persistence ($\hat{\rho}_x < 0.3$)									
Δdp	1871-02	2020-12	−7.470	7.015	−0.015	−0.055	1.000	0.144	0.980
Δep	1871-02	2020-12	−9.315	12.578	−0.012	−0.018	1.000	0.290	0.916
dfr	1926-01	2020-12	−6.970	5.263	0.026	0.039	1.000	−0.102	0.990
ltr	1926-01	2020-12	−4.583	6.210	0.199	0.130	1.000	0.043	0.999
PANEL B: Dependent variables (labels)									
r03m	1871-04	2020-09	−0.467	0.874	0.006	0.008	0.090	0.683	0.004
r06m	1871-07	2020-06	−0.534	0.959	0.012	0.016	0.125	0.849	0.004
r12m	1872-01	2019-12	−0.716	1.449	0.026	0.031	0.190	0.928	0.005
r24m	1873-01	2018-12	−0.900	1.186	0.054	0.040	0.281	0.965	0.005

Fig. 6. R^2_{oos} for low innovation correlation. We plot the out-of-sample R^2 as a function of sample sizes. The instances were filtered to keep only those corresponding to an estimated $\hat{\rho} \in [-0.05, 0.05]$, which is the interval we choose as a baseline for an absence of correlation between innovations. Before averaging, we remove all outliers in predictions, i.e., those values that lie outside the $[-100\%, +300\%]$ interval. Each point corresponds to one predictor and one calendar decade (over which the R^2 is computed). Each vertical panel pertains to one forecasting horizon, which, in turn, corresponds to one persistence level for the dependent variable. The black numbers at the top indicate the proportion of points above zero.

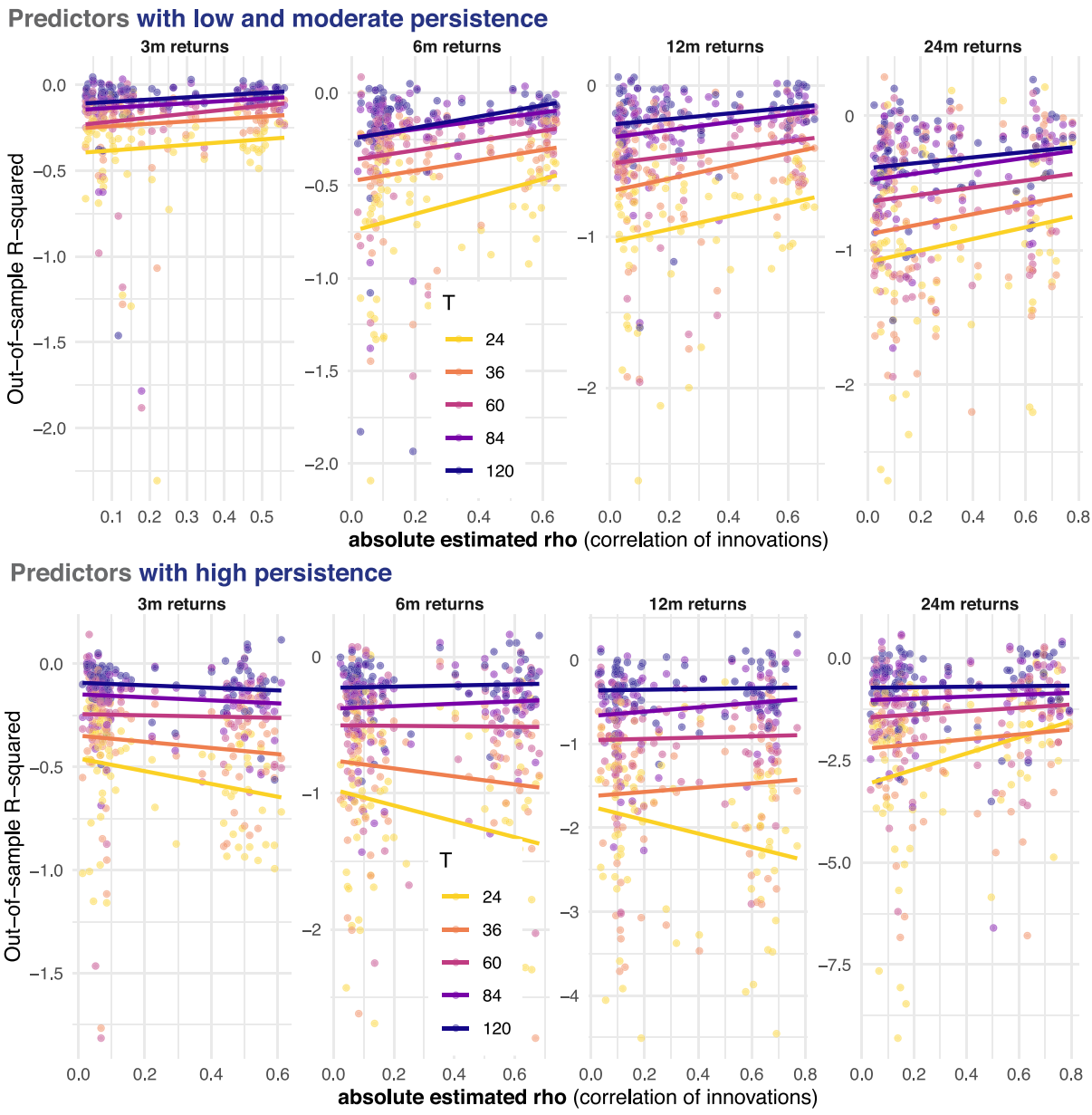

Fig. 7. R^2_{oos} as a function of $\hat{\rho}$. We plot the out-of-sample R^2 as a function of $\hat{\rho}$, the estimated correlation between innovations. Before averaging the squared errors, we remove all outliers in predictions, i.e., those values that lie outside the $[-100\%, +300\%]$ interval. The top panel relates to predictors with low or moderate persistence, as defined in Table 2, while the bottom one shows the results for the predictors with high auto-correlation. Each point corresponds to one predictor and one calendar decade (over which the R^2 is computed). Each vertical panel pertains to one forecasting horizon, which, in turn, corresponds to one persistence level for the dependent variable. For each sample size (T, shown with colors), we fit a linear model on the points to determine the impact of ρ on the R^2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

dependent variable and is regressed against all other predictors, one by one:

$$\text{VIX}_{t+1} = a + bx_t + e_{t+1}. \tag{39}$$

Unlike the svar variable in our study whose estimated autocorrelation is only 0.57 and therefore limits the presence of unexpected gains, the VIX indicator, downloaded from the Federal Reserve of Saint Louis data center, has an autocorrelation of 0.82 when sampled at a monthly frequency (from January 1990 onwards).

Predictions are then made accordingly for the next period value and the out-of-sample R^2 is derived from January 2000 onwards. Predictions (and the evaluation of their out-of-sample performance) only start in 2000 because the 1990–2000 decade is used to estimate the first coefficient (it serves as an initial training sample, which, given the sample size $T = 120$ months requires a 10-year buffer).

In the left panel of Fig. 8, we represent the distribution of estimated correlations between innovations, $\hat{\rho}$. It is interesting to notice the levels of estimated correlations remain low, i.e. between −15% and +12%, which is adequate to illustrate the presence of unexpected gains of Type I but not of Type II.

In the right panel, we show the OOS R^2 clustered by sample size T (x-axis). For short samples ($T = 12$), the proportion of positive R^2 is 96%, but this figure shrinks to 43% for longer samples ($T = 120$). In any case, there is a substantial fraction of cases that correspond to unexpected gains of Type I. In addition, there is a clear pattern of decreasing R^2 with T, which is a theoretical prediction from Figs. 1 and 3 (with high ρ_y).

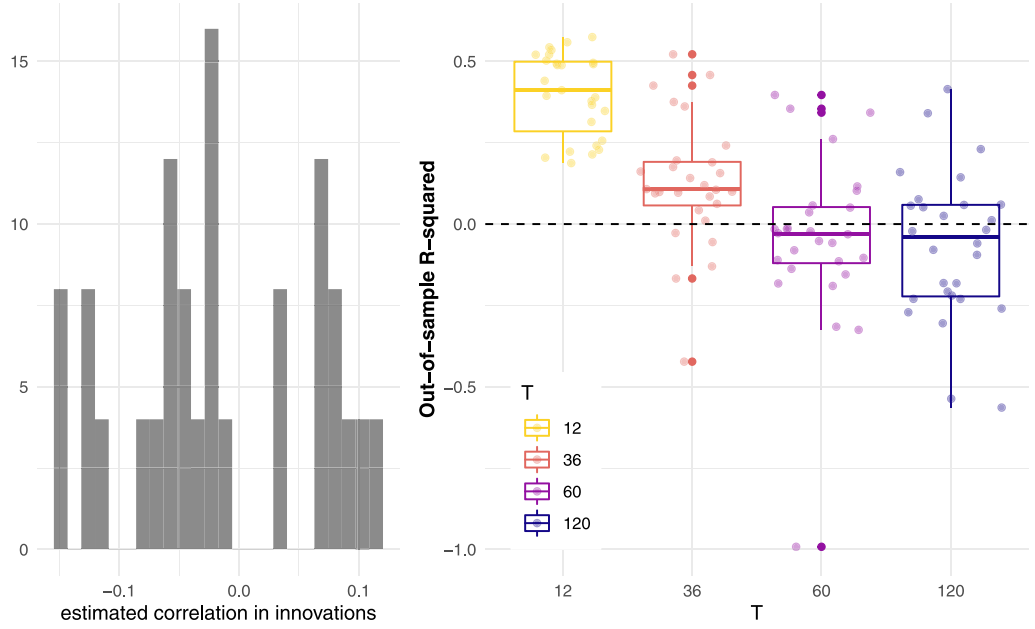

Fig. 8. VIX prediction. In the left panel, we show the distribution of estimated correlations between innovations of the VIX and innovations for all other predictors. In the right panel, we plot the out-of-sample R^2 stemming from PRs in which the VIX is the dependent variable. All other predictors are used, one at a time, except the svar. Before averaging, we remove all outliers in predictions, i.e, those values that lie outside the $[0\%, +500\%]$ interval. Each point corresponds to one predictor and one calendar decade (over which the R^2 is computed).

Table 3
t-statistics of slopes: we provide the *t*-statistics of the fitted linear relationships shown in Fig. 7.

Sample size	Return horizon			
(months)	3 months	6 months	12 months	24 months
PANEL A: predictors with low or moderate persistence				
24	1.084	2.632	2.126	1.990
36	1.286	2.143	2.071	1.714
60	1.365	1.834	1.480	1.187
84	1.040	2.060	2.183	1.753
120	0.901	1.807	1.582	1.364
PANEL B: predictors with high persistence				
24	−0.901	−1.337	−0.253	2.699
36	−0.953	−1.107	1.110	1.578
60	−0.298	0.363	0.279	0.836
84	−1.012	0.606	1.226	0.512
120	−1.014	0.460	0.355	0.193

5. Conclusion

In this article, we evaluate the out-of-sample loss that an agent faces when using a predictive regression subject to a strong model misspecification. En route, we introduce the concept of unexpected gains in predictive regressions. We provide two definitions thereof, based on the level and sensitivity of the out-of-sample R^2. We present closed-form expressions for the out-of-sample mean squared error (and R^2) when regression coefficients are given by the sample OLS estimates.

Our results reveal the parametric configurations in which unintended opportunities may arise and they all involve small sample sizes. First, as in the documented spurious correlation effect, there are cases for which a zero correlation between the variables is associated to unexpectedly positive R^2. Second, and again surprisingly, we find combinations of parameters for which this out-of-sample R^2 increases when the correlation between the processes (and their innovations) decreases in absolute value. This is eminently counter-intuitive because we would presume that a lower correlation be associated with weaker information linkages between the processes and thus with a smaller R^2.

Our theoretical findings are confirmed by many simulation exercises and illustrated via an empirical study of return predictability where the dependent variable is the S&P500 index and the predictive variables are taken from the studies from Novy-Marx (2014) and Welch and Goyal (2008). Focusing on the prediction of the VIX, we observe that short training samples yield large R^2 values despite the very low levels of estimated correlations between innovations, which clearly illustrates an unexpected opportunity of Type I.

The generalization of our results to higher dimensions is left for future work. In particular, three extensions are of interest. Predictive regressions with many predictors are harder to handle analytically because of the inverse matrix in the sample coefficients, but they seem to be a promising direction for research. Panel approaches are another suggestion, as they would allow us to explain dependent variables for a cross-section of assets. Finally and more generally, the generalization to nonlinear models (e.g., tree methods and neural networks) is of interest, though likely out of reach analytically.[16] The joint impact of correlations, variable persistence, and sample size on the performance of these technical tools remains an open question.

CRediT authorship contribution statement

Guillaume Coqueret: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Romain Deguest:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejor.2024.05.044.

[16] A few studies, e.g., Alwosheel et al. (2018) and Turmon and Fine (1994), assess the effect of training sample length on the predictive quality of neural networks. More recently, the double descent effect with non-linear models also touches this topic, as in Hastie et al. (2022).

References

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. (2018). Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*, *28*, 167–182.

Ang, A., & Bekaert, G. (2007). Stock return predictability: Is it there? *The Review of Financial Studies*, *20*(3), 651–707.

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30.

Badertscher, M., & Pretsch, E. (2006). Bad results from good data. *TRAC Trends in Analytical Chemistry*, *25*(11), 1131–1138.

Bandi, F. M., & Perron, B. (2008). Long-run risk-return trade-offs. *Journal of Econometrics*, *143*(2), 349–374.

Bandi, F. M., Perron, B., Tamoni, A., & Tebaldi, C. (2019). The scale of predictability. *Journal of Econometrics*, *208*(1), 120–140.

Bao, Y., & Kan, R. (2013). On the moments of ratios of quadratic forms in normal random variables. *Journal of Multivariate Analysis*, *117*, 229–245.

Barrieu, P., & Scandolo, G. (2015). Assessing financial model risk. *European Journal of Operational Research*, *242*(2), 546–556.

Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, *117*(48), 30063–30070.

Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., & Zhao, L. (2014). Misspecified mean function regression: Making good use of regression models that are wrong. *Sociological Methods & Research*, *43*(3), 422–451.

Blanchet, J., & Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, *44*(2), 565–600.

Bossaerts, P., & Hillion, P. (1999). Implementing statistical criteria to select return forecasting models: what do we learn? *The Review of Financial Studies*, *12*(2), 405–428.

Boudoukh, J., Richardson, M., & Whitelaw, R. F. (2008). The myth of long-horizon predictability. *The Review of Financial Studies*, *21*(4), 1577–1605.

Box, G. E. (1966). Use and abuse of regression. *Technometrics*, *8*(4), 625–629.

Campbell, J. Y. (2001). Why long horizons? A study of power against persistent alternatives. *Journal of Empirical Finance*, *8*(5), 459–491.

Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The econometrics of financial markets*. Princeton University Press.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, *21*(4), 1509–1531.

Campbell, J. Y., & Yogo, M. (2006). Efficient tests of stock return predictability. *Journal of Financial Economics*, *81*(1), 27–60.

Cerreia-Vioglio, S., Hansen, L. P., Maccheroni, F., & Marinacci, M. (2020). Making decisions under model misspecification. arXiv Preprint (2008.01071).

Chambers, J. M., Cleveland, W., Kleiner, B., & Tuckey, P. (2018). *Graphical methods for data analysis*. CRC Press.

Cochrane, J. H. (2008). The dog that did not bark: A defense of return predictability. *The Review of Financial Studies*, *21*(4), 1533–1575.

Cooper, W. L., Homem-de Mello, T., & Kleywegt, A. J. (2015). Learning and pricing with models that do not explicitly incorporate competition. *Operations Research*, *63*(1), 86–103.

Coqueret, G., & Tavin, B. (2016). An investigation of model risk in a market with jumps and stochastic volatility. *European Journal of Operational Research*, *253*(3), 648–658.

Dangl, T., & Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, *106*(1), 157–181.

Deng, A. (2014). Understanding spurious regression in financial economics. *Journal of Financial Econometrics*, *12*(1), 122–150.

Eroglu, C., & Hofer, C. (2011). Inventory types and firm performance: Vector autoregressive and vector error correction models. *Journal of Business Logistics*, *32*(3), 227–239.

Exner, O. (1997). How to get wrong results from good experimental data: A survey of incorrect applications of regression. *Journal of Physical Organic Chemistry*, *10*(11), 797–813.

Fama, E. F., & French, K. R. (1988a). Dividend yields and expected stock returns. *Journal of Financial Economics*, *22*(1), 3–25.

Fama, E. F., & French, K. R. (1988b). Permanent and temporary components of stock prices. *Journal of Political Economy*, *96*(2), 246–273.

Farmer, L., Schmidt, L., & Timmermann, A. (2021). Pockets of predictability. *The Journal of Finance*, *Forthcoming*.

Ferson, W. E., Sarkissian, S., & Simin, T. T. (2003). Spurious regressions in financial economics? *The Journal of Finance*, *58*(4), 1393–1413.

Freeman, J. R., Williams, J. T., & Lin, T.-m. (1989). Vector autoregression and the study of politics. *American Journal of Political Science*, *33*(4), 842–877.

Goyal, A., & Welch, I. (2003). Predicting the equity premium with dividend ratios. *Management Science*, *49*(5), 639–654.

Goyal, A., Welch, I., & Zafirov, A. (2021). *A comprehensive look at the empirical performance of equity premium prediction II*: *SSRN Working Paper 3929119*.

Granger, C., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, *2*(2), 111–120.

Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, *50*(2), 949–986.

Hjalmarsson, E. (2011). New methods for inference in long-horizon regressions. *Journal of Financial and Quantitative Analysis*, *46*(3), 815–839.

Hsiao, C. (1981). Autoregressive modelling and money-income causality detection. *Journal of Monetary Economics*, *7*(1), 85–106.

Hsu, A., Palomino, F., & Qian, L. (2022). Gone with the vol: A decline in asset return predictability during the great moderation. *Management Science*, *Forthcoming*.

Kan, R., & Pan, J. (2021). *Finite sample analysis of predictive regressions with long-horizon returns*: *SSRN Working Paper 3790052*.

Kan, R., & Wang, X. (2010). On the distribution of the sample autocorrelation coefficients. *Journal of Econometrics*, *154*(2), 101–121.

Lanne, M. (2002). Testing the predictability of stock returns. *The Review of Economics and Statistics*, *84*(3), 407–415.

Lazar, E., & Qi, S. (2022). Model risk in the over-the-counter market. *European Journal of Operational Research*, *298*(2), 769–784.

Levi, R., Janakiraman, G., & Nagarajan, M. (2008). A 2-approximation algorithm for stochastic inventory control models with lost sales. *Mathematics of Operations Research*, *33*(2), 351–374.

Lewellen, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics*, *74*(2), 209–235.

Luong, H. T. (2007). Measure of bullwhip effect in supply chains with autoregressive demand process. *European Journal of Operational Research*, *180*(3), 1086–1097.

Magnus, J. R. (1986). The exact moments of a ratio of quadratic forms in normal variables. *Annales d'Economie et de Statistique*, (4), 95–109.

Magnus, J. R. (1990). On certain moments relating to ratios of quadratic forms in normal variables: Further results. *Sankhyā. The Indian Journal of Statistics*, *52*, 1–13.

Mitchell, P. (1997). Misuse of regression for empirical validation of models. *Agricultural Systems*, *54*(3), 313–326.

Nambiar, M., Simchi-Levi, D., & Wang, H. (2019). Dynamic learning and pricing with model misspecification. *Management Science*, *65*(11), 4980–5000.

Nelson, D. B. (1992). Filtering and forecasting with misspecified ARCH models I: Getting the right variance with the wrong model. *Journal of Econometrics*, *52*(1–2), 61–90.

Nelson, D. B., & Foster, D. P. (1995). Filtering and forecasting with misspecified ARCH models II: Making the right forecast with the wrong model. *Journal of Econometrics*, *67*(2), 303–335.

Novy-Marx, R. (2014). Predicting anomaly performance with politics, the weather, global warming, sunspots, and the stars. *Journal of Financial Economics*, *112*(2), 137–146.

Paolella, M. S. (2018). *Linear models and time-series analysis: Regression, ANOVA, ARMA and GARCH*. John Wiley & Sons.

Paye, B. S. (2012). 'Déjà vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics*, *106*(3), 527–546.

Pettenuzzo, D., Timmermann, A., & Valkanov, R. (2014). Forecasting stock returns under economic constraints. *Journal of Financial Economics*, *114*(3), 517–553.

Phillips, P. C. (2015). Halbert white jr. memorial JFEC lecture: Pitfalls and possibilities in predictive regression. *Journal of Financial Econometrics*, *13*(3), 521–555.

Piatti, I., & Trojani, F. (2020). Dividend growth predictability and the price–dividend ratio. *Management Science*, *66*(1), 130–158.

Porter, A. M. (1999). Misuse of correlation and regression in three medical journals. *Journal of the Royal Society of Medicine*, *92*(3), 123–128.

Porter, A. L., Connolly, T., Heikes, R. G., & Park, C. Y. (1981). Misleading indicators: The limitations of multiple linear regression in formulation of policy recommendations. *Policy Sciences*, *13*, 397–418.

Sawa, T. (1972). Finite-sample properties of the k-class estimators. *Econometrica*, *40*(4), 653–680.

Sizova, N. (2013). Long-horizon return regressions with historical volatility and other long-memory variables. *Journal of Business & Economic Statistics*, *31*(4), 546–559.

Sobel, M. J., & Babich, V. (2012). Optimality of myopic policies for dynamic lot-sizing problems in serial production lines with random yields and autoregressive demand. *Operations Research*, *60*(6), 1520–1536.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, *54*(3), 375–421.

Stock, J. H., & Watson, M. W. (2001). Vector autoregressions. *Journal of Economic Perspectives*, *15*(4), 101–115.

Torous, W., Valkanov, R., & Yan, S. (2004). On predicting stock returns with nearly integrated explanatory variables. *Journal of Business*, *77*(4), 937–966.

Turmon, M. J., & Fine, T. L. (1994). Sample size requirements for feedforward neural networks. *Vol. 7*, In *Advances in neural information processing systems* (pp. 327–334).

Valkanov, R. (2003). Long-horizon regressions: Theoretical results and applications. *Journal of Financial Economics*, *68*(2), 201–232.

Van Binsbergen, J. H., & Koijen, R. S. (2010). Predictive regressions: A present-value approach. *The Journal of Finance*, *65*(4), 1439–1471.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508.

White, H. (2014). *Asymptotic theory for econometricians - Revised edition*. Academic Press.

Wu, D., & Olson, D. L. (2010). Enterprise risk management: coping with model risk in a large bank. *Journal of the Operational Research Society*, *61*(2), 179–190.

Xu, K.-L. (2020). Testing for multiple-horizon predictability: Direct regression based versus implication based. *The Review of Financial Studies*, *33*(9), 4403–4443.

Zhu, X. (2015). Tug-of-war: Time-varying predictability of stock returns and dividend growth. *Review of Finance*, *19*(6), 2317–2358.