CCIMI Initial Research Report
Instabilities in statistical modelling
and
a proposal of the adaptive learning method in time series

Parley Ruogu Yang

This version: 5 June 2021
The latest version can be found on my website.

Abstract

We investigate the instabilities in statistical modelling and propose an innovative learning method in time series. In the first half of the report, we start by reviewing the overarching idea of stabilities and statistical learning in a regression setting, followed by their relations to linear models. Then, we present studies towards the gradient of the linear models and neural network models, followed by distributional analysis. The second half concerns specifically with the topics in time series: stationarity and forecasting. By reviewing the contemporary methods with computational experiments, we analyse the stability and instability of some autoregressive models. Lastly, we propose a forecast-centric learning method called adaptive learning.

Acknowledgement

I thank my supervisor Dr Anders Hansen for his supervision, advice, support, and wisdoms which have inspired me to continue the research in the relevant areas. I appreciate the helpful comments and contributions from the members of OPRG: in particular, Raena McElwee and Ryan Lucas for parts of the computing implementations; and Dr Camilla Schelpe for her helpful feedbacks.

Contents

1 General phenomena about stability and instability

1.1 Introduction

1.1.1 Overview

In this subsection, we present a general framework of stability and instability, followed by a review on the setup of regression in subsection 1.2. Then, in subsection 1.3 and subsection 1.4, we present stability and instability in the following manner: we start by stating the relevant assumption on the data generating process / underlying relationship between variables. Then, we review a common training method or the parametric model and state the relevant consequences. At last, we state and prove Lemmas that can describe stability and instability of the model. To highlight the difference between these two subsections: in subsection 1.3, we focus on deterministic models, where we ignore the concept of noises and purely focus on the deterministic relationships between variables; whereas in subsection 1.4, we consider random variables and distributional assumptions, which engages more intimately with the statistical community. This can then be linked with the analysis of time series, as further discussed in section 2.

1.1.2 Notations and Abbreviations

Let n be a positive integer. $[n] := \{1, ..., n\}$, that is, a set of all integers from 1 to n.

When dealing with $\mathbb{R}^d$, we use the metric space $(\mathbb{R}^d, ||\cdot||)$ where $||\cdot|| := ||\cdot||_2$ is the standard Euclidean norm, unless specified otherwise.

The closed ε-ball around a set A is written as

$$B(A;\varepsilon) := \{z \in \mathbb{R}^d \ : \ \exists x \in A \ s.t. \ ||x - z|| \leq \varepsilon\}$$

The collection of explanatory variables in a dataset $\mathcal{D}$ is written as

$$\mathcal{D}_x := \{x \ : \ \exists y \ \ s.t. \ (x, y) \in \mathcal{D}\}$$

The expectation of functions of random variables are written as follows: if $x \in X$ and P is the probability measure of x over X, then for a function $f : X \to \mathbb{R}^d$, we denote

$$\mathbb{E}_x[f(x)] := \int_{x \in X} f(x) dP(x)$$

Cauchy Schwartz inequality (Lemma A.1) is abbreviated as CS; Law of Large Numbers (Theorem A.3) is abbreviated as LLN; and the Central Limit Theorem (Theorem A.4) is abbreviated as CLT.

1.1.3 Overview of stability and instability

Let us consider a regression set-up (further detailed in subsection 1.2) where $f(x)$ denotes the true value of response to a given input x, and $g(x;\hat{\theta})$ denotes the response from our

best fitted model.[1] In the case of deterministic models (elaborated in subsection 1.3), we can make general statements about stability and instability as follows:

- Stability:

$$\forall \varepsilon > 0, \forall v \in B(\mathcal{D}_x, \varepsilon), \quad |g(v;\hat{\theta}) - f(v)| \leq \texttt{UpperBound}(f, \hat{\theta}, \varepsilon)$$

- Instability:

$$\forall \varepsilon > 0, \exists v \in B(\mathcal{D}_x, \varepsilon) \ \texttt{ such that } \ |g(v;\hat{\theta}) - f(v)| \geq \texttt{LowerBound}(f, \hat{\theta}, \varepsilon)$$

The stability statement, as shown above, deals with the global behaviour of our best fitted model $g(\cdot;\hat{\theta})$ compared against the true model $f(\cdot)$. In particular, a ε amount of perturbation may result in an upper-bounded difference in terms of the output of the results as quantified by $\texttt{UpperBound}(f, \hat{\theta}, \varepsilon)$. Likewise, the instability statement details the existence of worst perturbation $v \in B(\mathcal{D}_x, \varepsilon)$, for which the model can yield a significantly different result compared to the truth. Such a difference is lower bounded by $\texttt{LowerBound}(f, \hat{\theta}, \varepsilon)$.

The statements for stability and instability in probabilistic models are distributional-dependent, hence will be introduced under further set-up of the frameworks in subsection 1.4 and section 2.

1.2 On statistical contemporaries

This subsection reviews the set-up of regression, where we adapt the framework of regression estimation (Vapnik 1999).

Let y be the true value of response to a given input x, and $f(x, \alpha)$ be the response provided by the learning machine, with $\alpha \in A$, where A denotes the set of parameters. Then, given some specification of the loss function $L(\cdot, \cdot)$, we aim to minimise the risk functional

$$R(\alpha) = \mathbb{E}_{x,y}[L(y, f(x, \alpha))]$$

by finding over the class of functions $\alpha \in A$. Write $\alpha^* := \arg\min_{\alpha \in A} R(\alpha)$

Theorem 1.1 (Regression estimation (Vapnik 1999))
Let the loss to take the form $L(a, b) = (a - b)^2$. Define the regression function as

$$f(x, \alpha_0) := \mathbb{E}[y|x]$$

Then

$$\alpha^* = \arg\min_{\alpha \in A} \mathbb{E}_x[(f(x, \alpha) - f(x, \alpha_0))^2]$$

In particular, if $\alpha_0 \in A$, then $\alpha^* = \alpha_0$

Proof. Observe that

$$R(\alpha) = \mathbb{E}_x[(f(x, \alpha) - f(x, \alpha_0))^2] + \mathbb{E}_{x,y}[(y - f(x, \alpha_0))^2]$$

[1] In statistical and econometric literature, we refer x as explanatory variables and y as dependent variables.

And that the second term is independent of α. □

The above theorem helps understanding the common statistical framework — it is under the squared loss function $L(a,b) = (a-b)^2$ that we end up in an exercise of finding the conditional expectation $\mathbb{E}[y|x]$, and it also justifies the common set-up of probabilistic model, that we usually focus on the conditional distribution of the dependent variable given the explanatory ones, i.e. $y|x$, rather than the joint distribution (x,y) or marginal distribution of the dependent variable y.

Under this framework, Joseph (2019) presents a Shapley-Taylor decomposition for interpreting statistical machine learning, and practical extensions can be made as a consequence of such a decomposition (Bluwstein et al. 2020). For more contemporary reviews, one can also refer to Efron and Hastie (2016).

1.3 Examples in deterministic models

The plan of this subsection goes as follows: in Lemma 1.2 and Lemma 1.3, we analyse the stability and instability of deterministic linear models respectively. The instability of omitted relevant variable follows immediately in Corollary 1.4. Then, we use a similar method to imply the instability of single layer neural network in Corollary 1.5.

1.3.1 Example: linear model (deterministic case)

Let $f(x) = y$ be the underlying relationship between the explanatory variable x and the dependent variable y, and let $f : \mathbb{R}^N \to \mathbb{R}$ be once continuously differentiable. For training, consider the functional class $\mathcal{G}$ of linear functions on variable x and parameter $\theta \in \Theta$:

$$g(\cdot;\theta) \in \mathcal{G} \qquad \Longleftrightarrow \qquad g(x;\theta) = \langle x, \theta \rangle$$

Here, parametric space $\Theta \subset \mathbb{R}^N$. Let $\mathcal{D} := \{(x_i, y_i) : y_i = f(x_i), i \in [n]\}$ be the training set. The usual Ordinary Least Square (OLS) estimator can be written as a standard squared loss minimisation:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{i \in [n]} (y_i - g(x_i;\theta))^2$$

Lemma 1.2 (Stability of linear model)
If $||\nabla f(x)||_\infty \leq K \ \forall x \in \mathbb{R}^N$, *then* $\forall \varepsilon > 0, \forall v \in B(\mathcal{D}_x, \varepsilon)$,

$$|g(v;\hat{\theta}) - f(v)| \leq (K + ||\hat{\theta}||)\varepsilon + \min_{x \in \mathcal{D}_x \ : \ ||x-v|| \leq \varepsilon} |f(x) - g(x;\hat{\theta})| \tag{1}$$

Proof. $\forall x \in \mathcal{D}_x$ s.t. $||x - v|| \leq \varepsilon$, we have

$$|g(v;\hat{\theta}) - f(v)| \leq |g(v;\hat{\theta}) - g(x;\hat{\theta})| + |g(x;\hat{\theta}) - f(x)| + |f(x) - f(v)|$$

The first term on the right hand side is bounded by $||\hat{\theta}||\varepsilon$ due to CS, and the third term is bounded by $K\varepsilon$ due to the Fundamental Theorem of Calculus (FTC). □

Lemma 1.3 (Instability of linear model in deterministic case)
Let $\varepsilon > 0$. If there exists $M > 0$ such that

$$\max_{x\in\mathcal{D}_x}\max_{i\in[N]}\inf_{v\in B(x;\varepsilon)}|\frac{\partial f}{\partial x_i}(v)-\hat{\theta}_i| \geq M \tag{2}$$

then $\exists v \in B(\mathcal{D}_x;\varepsilon)$ s.t.

$$|g(v;\hat{\theta})-f(v)| \geq M\varepsilon \tag{3}$$

Proof. Denote

$$x^* := \arg\max_{x\in\mathcal{D}_x}\max_{i\in[N]}\inf_{v\in B(x;\varepsilon)}|\frac{\partial f}{\partial x_i}(v)-\hat{\theta}_i| \tag{4}$$

$$j := \arg\max_{i\in[N]}\inf_{v\in B(x^*;\varepsilon)}|\frac{\partial f}{\partial x_i}(v)-\hat{\theta}_i| \tag{5}$$

The key idea is to separate into four cases by the signs of $(f-g)(x^*)$ and the derivative.
Write $\tilde{f}:\mathbb{R}\to\mathbb{R}$, $z\mapsto f(x_1^*,...,x_{j-1}^*,z,x_{j+1}^*,...,x_N^*)$
Case 1a: $\tilde{f}'(x_j^*)-\hat{\theta}_j \geq M$ and $f(x^*)-g(x^*)\geq 0$.
Case 1b: $\tilde{f}'(x_j^*)-\hat{\theta}_j \geq M$ and $f(x^*)-g(x^*)< 0$.
Case 2a: $\tilde{f}'(x_j^*)-\hat{\theta}_j < -M$ and $f(x^*)-g(x^*)\geq 0$
Case 2b: $\tilde{f}'(x_j^*)-\hat{\theta}_j < -M$ and $f(x^*)-g(x^*)< 0$.

In cases 1a and 2b, we construct $v := (x_1^*,...,x_{j-1}^*,x_j^*+\varepsilon,x_{j+1}^*,...,x_N^*)$

Observe, for 1a,

$$f(v)-g(v) = [f(v)-f(x^*)+g(x^*)-g(v)]+[f(x^*)-g(x^*)] \geq (f-g)(v-x^*) \geq M\varepsilon$$

due to FTC; and likewise for 2b,

$$g(v)-f(v) = [g(v)-g(x^*)+f(x^*)-f(v)]+[g(x^*)-f(x^*)] \geq (g-f)(v-x^*) \geq M\varepsilon$$

The same proof goes with cases 1b and 2a, where we construct
$v := (x_1^*,...,x_{j-1}^*,x_j^*-\varepsilon,x_{j+1}^*,...,x_N^*)$ □

Comments: We note the proof here only utilises the first order derivatives of f, g, thus easy extension can be made to non-linear models. See subsubsection 1.3.3 for example.

1.3.2 Example: Omitted relevant variable

Let $\Theta = \mathbb{R}^{N_0}$ with $N_0 < N$. This means we omit some relevant variables, in particular, those with indices $N_0, N_0+1, ..., N$.

Corollary 1.4 (Instability due to omitted relevant variable)
Let $\varepsilon > 0$. If there exists $x \in \mathcal{D}_x, i \in [N]\setminus[N_0]$ such that

$$\inf_{v\in B(x;\varepsilon)}\left|\frac{\partial f}{\partial x_i}(v)\right| \geq M$$

Then there exists $v \in B(\mathcal{D}_x; \varepsilon)$ *such that*

$$|g(v; \hat{\theta}) - f(v)| \geq M\varepsilon$$

The above is a trivial application from Lemma 1.3, as it satisfies Equation 2 due to $\hat{\theta}_i = 0$ for all $i \in [N] \setminus [N_0]$

1.3.3 Example: Single Layer Neural Network (NN)

Let $\rho : \mathbb{R}^{N_1} \to \mathbb{R}^{N_1}$ be an element-wise activation function. Let W_1, W_2 be affine transformations where $W_1 : \mathbb{R}^N \to \mathbb{R}^{N_1}$ and $W_2 : \mathbb{R}^{N_1} \to \mathbb{R}$. We note that affine transformations can be represented by linear coefficients, in particular, we can write matrix $A_1 := (a_{1,i,j})_{i\in[N_1],j\in[N]} \in \mathbb{R}^{N_1 \times N}$ and vectors $A_2 := (a_{2,i})_{i\in[N_1]} \in \mathbb{R}^{N_1}, b_1 \in \mathbb{R}^{N_1}, b_2 \in \mathbb{R}$ such that

$$W_2(\rho(W_1(x))) = A_2\rho(A_1x + b_1) + b_2$$

By element-wise, we can write $\rho(x) = (\tilde{\rho}(x_1), ..., \tilde{\rho}(x_{N_1}))$ and in this example, we focus on $\tilde{\rho}$ which are twice continuously differentiable with bounded derivatives. [2] To put the above set-up into our framework of functional class $\mathcal{G}$, we can say

$$g(\cdot; \theta) \in \mathcal{G} \qquad \Longleftrightarrow \qquad g(x; \theta) = W_2\left(\rho\left(W_1\left(x\right)\right)\right)$$

And here the parameter $\theta = (W_1, W_2)$

Observe that Lemma 1.3 can be generalised to our functional class, as all members of $\mathcal{G}$ are once continuously differentiable. In particular, the left hand side of Equation 2 can be written as

$$\max_{x\in\mathcal{D}_x} \max_{i\in[N]} \inf_{v\in B(x;\varepsilon)} |\frac{\partial f}{\partial x_i}(v) - \frac{\partial g}{\partial x_i}(v; \hat{\theta})|$$

Corollary 1.5 (Instability by perturbation: single layer NN)
Let $\varepsilon > 0$. *For all* $M > 0$, *there exists* $\theta \in \Theta$ *such that* $\exists v \in B(\mathcal{D}_x; \varepsilon)$ *satisfying*

$$|g(v; \theta) - f(v)| \geq M\varepsilon$$

Proof. By chain rule,

$$\frac{\partial g}{\partial x_i}(v; \hat{\theta}) = a_{2,i} \sum_{j\in[N_1]} a_{1,i,j}\tilde{\rho}'((W_1(x))_j)$$

Fix some $i \in [N]$ and perturb $a_{2,i}$ and $a_{1,i,j}$ accordingly so that

$$\max_{x\in\mathcal{D}_x} \max_{i\in[N]} \inf_{v\in B(x;\varepsilon)} |\frac{\partial f}{\partial x_i}(v) - \frac{\partial g}{\partial x_i}(v; \theta)| \geq M \tag{6}$$

This is achievable because $\tilde{\rho}'(\cdot)$ is bounded. The rest follows from Lemma 1.3. □

[2] Examples: Logistic activation function $\tilde{\rho}(x) = (1 + \exp(-x))^{-1}$, and we have $\tilde{\rho}'(x) = \tilde{\rho}(x)(1 - \tilde{\rho}(x)) \in (0, \frac{1}{4}]$.
Tanh activation function $\tilde{\rho}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, and we have $\tilde{\rho}'(x) = 1 - (\tilde{\rho}(x))^2 \in (0, 1)$

1.3.4 Example: Deep Learning for classification

This and the next subsubsection is TBC. I will inherit most of the things done in MT 20 and HT 21 here. The main paper of concern is the work-in-progress paper on deep learning (Bastounis, Hansen, and Vlacic 2021)

1.3.5 An extension to Deep Learning towards a special class of continuous function

TBC

1.4 Examples in probabilistic models

1.4.1 Example: linear model with correct specification (probabilistic case)

Consider a Data Generating Process (DGP)

$$y_i|x_i \sim N(f(x_i), \sigma^2) \quad \text{independently } \forall i \in [n]$$

In this example, we consider $f(x) = \langle \beta, x \rangle$. In linear models, we write $\boldsymbol{X} = [x_1|x_2|...|x_n]^T \in \mathbb{R}^{n \times N}$. We estimate $\hat{\beta}$ by Maximum Likelihood Estimation (MLE), which is equivalent to OLS in the linear model, and that if we assume $\boldsymbol{X^T X}$ to be invertible, we have $\hat{\beta} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}$. The distribution can be computed as a N dimensional normal distribution

$$\hat{\beta}|\boldsymbol{X} \sim N(\beta, \sigma^2 (\boldsymbol{X^T X})^{-1})$$

For any j, write $s_j(\boldsymbol{X}, \sigma^2) := (\sigma^2 (\boldsymbol{X^T X})^{-1})_{j,j}$, then we have a marginal distribution

$$\hat{\beta}_j|\boldsymbol{X} \sim N(\beta_j, s_j(\boldsymbol{X}, \sigma^2)) \tag{7}$$

For the rest of this example, we omit the condition "$|\boldsymbol{X}$" in notation. That is, when probability $\mathbb{P}[\cdot]$ being considered, it is meant by $\mathbb{P}[\cdot|\boldsymbol{X}]$ and are directly or indirectly a consequence of Equation 7.

Now, from Equation 7, as a consequence of the tail bounds of Gaussian distribution (Lemma A.2), for all positive real c,

$$\mathbb{P}[|\hat{\beta}_j - \beta_j| > c] < \sqrt{\frac{2}{\pi}} \frac{\exp(-\frac{c^2}{2 s_j(\boldsymbol{X}, \sigma^2)})}{c} \tag{8}$$

Now consider the following statement.

Lemma 1.6 (Stability of linear model under correct specification)
Pick any positive real constant c. $\forall v \in \mathbb{R}^N$,

$$\mathbb{P}\left[|g(v;\hat{\theta}) - f(v)| < \sqrt{N} c ||v||\right] \geq 1 - \sqrt{\frac{2}{\pi}} c^{-1} \sum_{i \in [N]} \exp\left(-\frac{c^2}{2 s_i(\boldsymbol{X}, \sigma^2)}\right) \tag{9}$$

Proof.

$$\begin{aligned}\mathbb{P}\left[|\langle\hat{\beta}-\beta, v\rangle| \geq \sqrt{N}c||v||\right] &\leq \mathbb{P}\left[||\hat{\beta}-\beta||^2 \geq Nc^2\right] \\ &\leq \sum_{i\in[N]} \mathbb{P}\left[|\hat{\beta}_i-\beta_i|^2 \geq c^2\right] \\ &\leq \sqrt{\frac{2}{\pi}}c^{-1}\sum_{i\in[N]} \exp\left(-\frac{c^2}{2s_i(\boldsymbol{X},\sigma^2)}\right)\end{aligned}$$

The first line follows from CS, and the third line follows from Equation 8. □

1.4.2 Example: linear model with incorrect specification (probabilistic case)

Lemma 1.7 (Instability of linear model under incorrect specification)
TBC

2 Time series and adaptive learning

2.1 Introduction to time series

What is time series? For such a fundamental concept, one finds many potential statements that can serve as a succinct and overarching definition. To illustrate a few:

- Hamilton (1994): Dynamic consequences of events over time.
- Prado and West (2010): A set of observations collected sequentially in time.
- Tsay (2010): Financial time series analysis is concerned with the theory and practice of asset valuation over time.
- Tsay and Chen (2018): Time series analysis is concerned with understanding the dynamic dependence of real-world phenomena.

The emphasis we highlight here, is that time series is ordered and irreversible, thus are commonly autocorrelated ('dynamic dependence' as per Tsay and Chen (2018)). In more formal settings, we denote $y_t \in \mathbb{Y}$ as the observation of a random variable y at time t, with $\mathbb{Y}$ as the domain of the random variable, typically a subset of $\mathbb{R}^d$ with some $d \geq 1$. The ordered and irreversible can be understood as the fact that at time t, we only have access to its past observations, noted $\{y_t, y_{t-1}, ..., y_1\} = \{y_\tau : \tau \in [t]\}$, but we can not reverse the time series in modelling, e.g. at time t, to make prediction about y_{t+1}, we should not be using y_{t+2}, as naturally future observations can not be used to explain the current or the past. This relates to two concerns. Firstly, a modelling concern, that many variables observed currently depend on its past values — hence autocorrelation. Traditional ARMA models address to such an issue (see Definition 2.1 for formal introduction), but that raises model complexity, and raises the associated concern of how to correctly determine model complexity. The other concern is forecasting — it is a naturally unique issue, in the way that it asks for the expectation of future state conditional on the current state, hence distinguishes itself from model-fitting.

To illustrate numerical examples of time series data, in Figure 1, we plot a one-dimensional simulated time series, and in Figure 2, we plot various 10-step-ahead forecasts at time index 100 by different models. Further numerical details of these are specified in Appendix B. An example of empirical time series of higher dimension is illustrated in Figure 3, where two time-slices of a 11-dimensional variable (known as the 'Yield Curve' in Macroeconomic literature) and two time-slices of a 9-dimensional variable (known as the 'VIX Curve' in Financial literature) are plotted on the left and right panels, with a 4-dimensional variable over time being plotted in the centre, representing the SP500 Index over time.

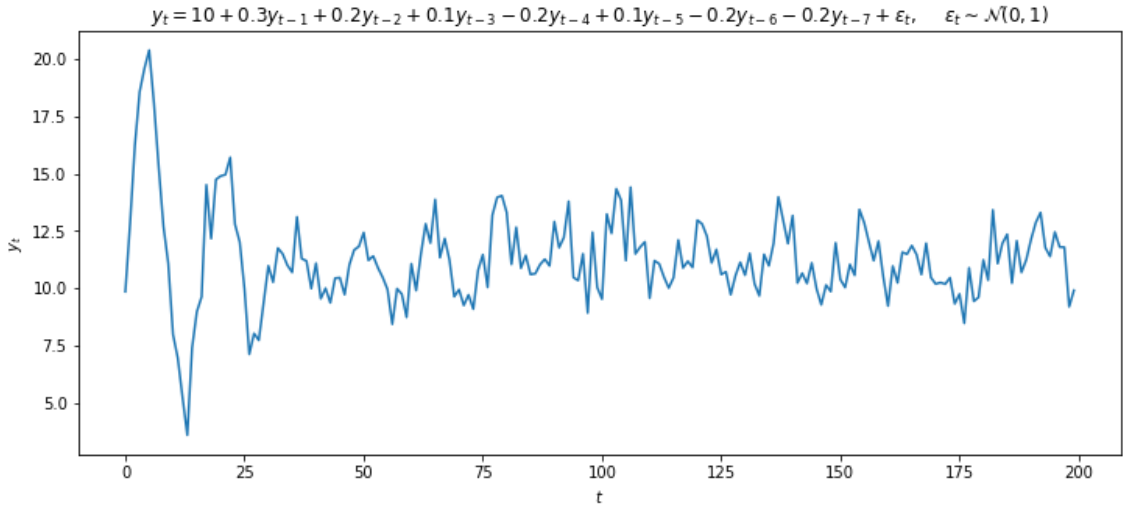

Figure 1: An example of simulated time series

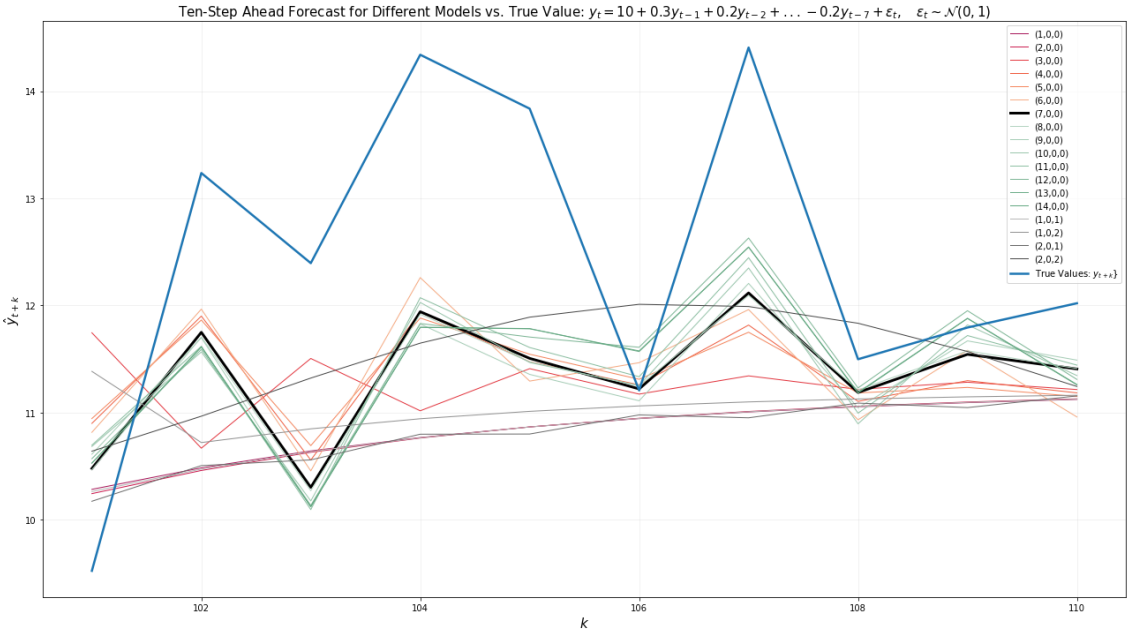

Figure 2: An example of forecasts by different models. The legend notes the models by $ARIMA(p, d, q)$.

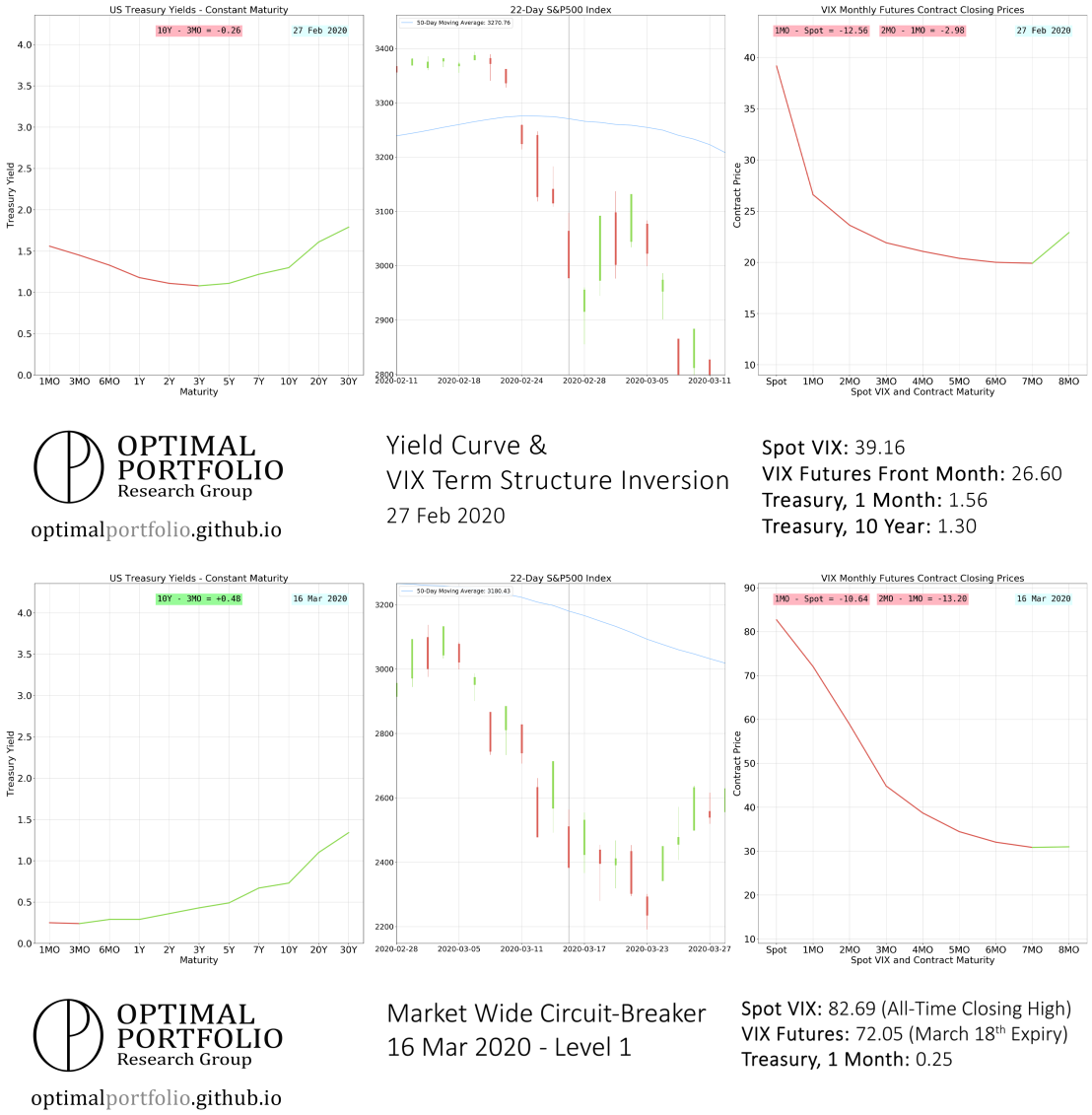

Figure 3: Examples of empirical time series: Yield curve, SP500, and VIX curve (see https://parleyyang.github.io/Res/Notes/poster.pdf for further information)

The problem of forecasting can be formally articulated as follows. At time t, we have access to an information set representing all data available up to t, written as[3]

$$\Phi_t := \{y_\tau : \tau \in [t]\}$$

Given such, we are in the business of producing a conditional forecast of future values of y, that is,

$$y_{t+k|t} = \mathbb{E}[y_{t+k}|\Phi_t]$$

for a step size $k \geq 1$. At this stage, one should notice the expectation can be computed upon knowing which model y_t follows. Below, we introduce a standard ARMA model, followed by an example of forecast computation.

[3] For simplicity, we just consider the dependent variable itself, before drawing the explanatory variables into discussion.

Definition 2.1 (ARMA(p,q) model)
Let L be an operator, defined as $L(x_t) := x_{t-1}$ for any time series.
Let p, q be non-negative integers.
Define the autoregressive polynomial of order p as

$$\Phi_p(L) := 1 - \phi_1 L - ... - \phi_p L^p = 1 - \sum_{i \in [p]} \phi_i L^i$$

and the moving average polynomial of order q as

$$\Psi_q(L) := 1 - \psi_1 L - ... - \psi_q L^q = 1 - \sum_{i \in [q]} \psi_i L^i$$

Then an autoregressive moving average (ARMA) model of order p, q can be written as

$$\Phi_p(L)(y_t - \mu) = \Psi_q(L)\varepsilon_t \quad where \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2) \quad \forall t \tag{10}$$

We denote an ARMA(p,0) model as AR(p), and an ARMA(0,q) model as MA(q).

Example 2.2 (Forecasts in AR(1))
In an AR(1) model, with $|\phi_1| < 1$, [4] *we can write*

$$y_t = c + \phi_1 y_{t-1} + \varepsilon_t$$

where $c := (1 - \phi_1)\mu$. Now, for $k = 1$, we have

$$y_{t+1|t} = c + \phi_1 y_{t|t} + \varepsilon_{t+1|t} = c + \phi_1 y_t$$

And likewise, for $k > 1$,

$$y_{t+k|t} = c + \phi_1 y_{t+k-1|t} = \frac{c(1 - \phi_1^k)}{1 - \phi_1} + \phi_1^k y_t$$

And note that, under the assumption of $|\phi_1| < 1$,

$$y_{t+k|t} \xrightarrow[k \to \infty]{} \frac{c}{1 - \phi_1} = \mu$$

A standard method for evaluating the quality of forecasts is introduced below.

Definition 2.3 (MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) of forecasting)
For any positive integer k, collect the vector of k-step-ahead forecasts by a model h as

$$\underline{\hat{y}_{t+k|t}(h)} := \big(\hat{y}_{t+1|t}(h), ..., \hat{y}_{t+k-1|t}(h), \hat{y}_{t+k|t}(h)\big)$$

Suppose we have access to a vector of data

$$\underline{y_{t:(t+k)}} := (y_{t+1}(h), ..., y_{t+k-1}(h), y_{t+k}(h))$$

[4] This assumption is necessary for stationarity, as being discussed in subsection 2.2.

And write $||\cdot||_2$ *and* $||\cdot||_1$ *as the 2-norm and 1-norm in an Euclidean space respectively, then*

$$RMSE_t(h) := ||\underline{\hat{y}_{t+k|t}(h)} - \underline{y_{t:(t+k)}}||_2 \tag{11}$$

$$MAE_t(h) := ||\underline{\hat{y}_{t+k|t}(h)} - \underline{y_{t:(t+k)}}||_1 \tag{12}$$

To engage with the standard statistical machine learning literature, we plot Figure 4 to make a point about 'training loss' versus 'validation loss' — in time series context, this can be understood as in-sample log likelihood versus out-of-sample forecasts performance evaluated by MAE and RMSE.

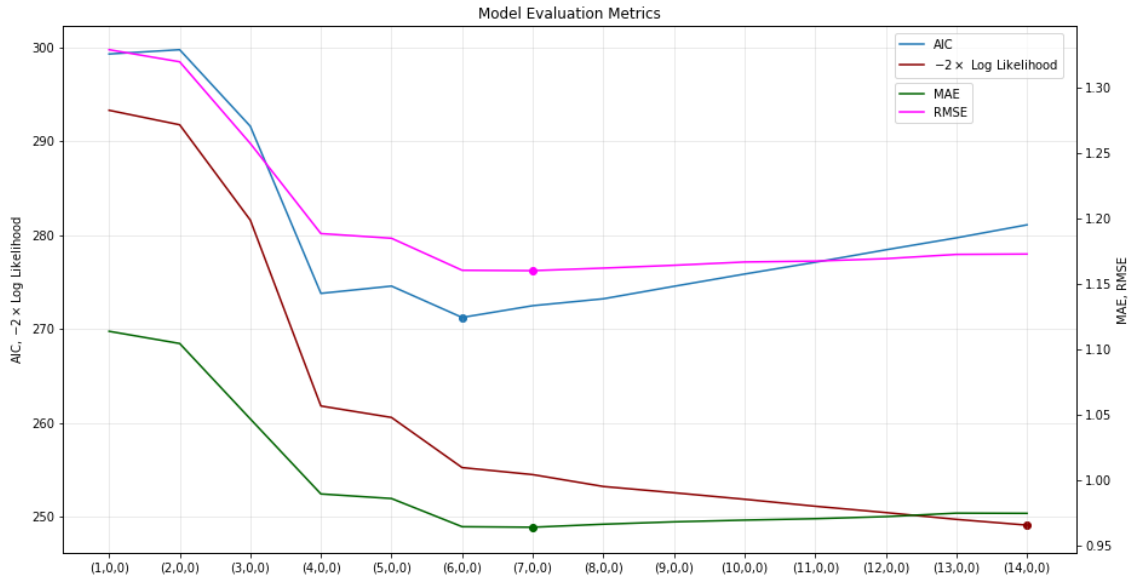

Figure 4: On the ability to detect an $AR(7)$: AIC under-fits and forecasting MAE and RMSE fits correctly.

As can be seen, the negative of log likelihood, as the training loss, monotonically decreases as model complexity increases. Certainly, to avoid overfit, in-sample penalisation was introduced, a common version of which is the Akaike Information Criterion (AIC). In the setting below, AR(7) is the underlying DGP, and all statistics are plotted as the average of 1000 repetitions. It can be seen that the out-of-sample loss is being minimised by the correctly specified modelling — AR(7), whereas the AIC over-penalises models. Furthermore, it seems that higher model complexity (over-specified), at least in this example and from Figure 2, does not increase the loss largely, compared to the ones with lower complexity (under-specified). This motivates further research in the area, proposed by the previous work (Yang 2020; Yang 2021), with the methodologies being summarised as 'adaptive learning'.

The rest of this section proceeds as follows: in subsection 2.2, we discuss the relationship between stationarity and stability by showing the consequence of incorrect specification, and how it can potentially damage the ability to forecast. This analysis is furthered by computational examples in subsection 2.3, where we demonstrate such instabilities. At

last, but not least, we propose a novel learning method, namely 'adaptive learning' in subsection 2.4.

2.2 Stationarity and stability

2.2.1 Definitions of stationarity

A key concept in time series is stationarity. More general definition in stochastic process concerns the probability distribution over a very general space, as illustrated below.

Definition 2.4 (A general definition of (strict) stationarity (Klenke 2020))
Let E be a Polish space, $I \subset \mathbb{R}$ and closed under addition. An E-valued stochastic process $(X_t)_{t\in I}$ is called stationary if, for all $s \in I$, $\mathbb{P}_{(X_t)_{t\in I}} = \mathbb{P}_{(X_t+s)_{t\in I}}$

In more specific settings — natural-number-indexed real-valued time series, we usually concern with a weaker form of stationarity, as defined below.

Definition 2.5 (A more specific definition of (weak) stationarity (Prado and West 2010; Tsay and Chen 2018))
Let $(x_t)_{t\in\mathbb{N}}$ be a real-valued time series. It is weakly stationary and denoted $x_t \sim I(0)$ if the first two moments of $(x_t)_{t\in\mathbb{N}}$ exists and time invariant. That is, for all $t, l \in \mathbb{N}$,

$$\begin{aligned}\mu :=& \mathbb{E}[x_t] < \infty \\ \gamma_l :=& Cov(x_t, x_{t+l}) < \infty\end{aligned}$$

And note that both μ, γ_l are independent of t.

For a positive integer d, we denote $x_t \sim I(d)$ if $\Delta^d x_t := (1-L)^d x_t \sim I(0)$.

Throughout this paper, we take Definition 2.5 as the formal definition of stationarity. As an example, in ARMA models, we can categorise stationarity by its parameters:

Example 2.6 (Stationarity in ARMA(p,q) model (See Hamilton (1994) for further information))
Consider the ARMA(p,q) model as per Equation 10. We can find the inverse of $\Phi_p(\cdot)$ as an infinite sequence $\tilde{\psi}_0 + \tilde{\psi}_1 L + + \tilde{\psi}_2 L^2 + ... =: \Phi_p^{-1}(L)$ Hence denote $\Theta(L) := \Phi_p^{-1}(L)\Psi_q(L)$ and write $\Theta(L) := \sum_{i\in\mathbb{N}} \theta_i L^i$.
Then, such a model is stationary if and only if $\sum_{i\in\mathbb{N}} \theta_i^2$ is finite.

If one were given a time series and were asked to test for stationarity, a common way to do so is via an ADF test, as explained next.

Consider an AR(p) model

$$\Phi_p(L)(y_t - \mu) = \varepsilon_t$$

Suppose $\Phi_p(1) = 0$, also known as unit root, then we note y_t to be non-stationary. Define $\zeta_{p-1}(L) := \Phi_p(L)(1-L)^{-1}$ and we can write $\zeta_{p-1}(L)\Delta(y_t) = \varepsilon_t$. Hence, ADF test aims to find out whether $\Phi_p(1) = 0$ is true.

Definition 2.7 (Augmented Dicky Fuller (ADF) test, random walk version)
Estimate the below model

$$\Delta(y_t) = \gamma y_{t-1} + \theta_1 \Delta(y_{t-1}) + ... + \theta_{p-1}\Delta(y_{t-p+1}) + \varepsilon_t \tag{13}$$

And test against the hypothesis

$$H_0 : \gamma = 0$$
$$H_A : \gamma < 0$$

Should H_0 be rejected, we conclude y_t has no unit root, implying stationarity.

There are two underlying hypotheses that the ADF tests rely on: parametric assumption based on an AR(p) specification, and the assumption of correct model specification itself.

First, in terms of the parametric assumption — we note that, even given the data to be an AR(p) process, it is certainly incomplete to say that as long as the alternative hypothesis holds, we have stationarity. An easy counter-example can be made for $p = 2$. Stationarity in an AR(2) model requires $\phi_2 - \phi_1 < 1$, but that implies, in Equation 13, $-2(\theta_1 - 1) < \gamma$, which is a lower bound for γ, and not necessarily being satisfied under H_A. In more general terms, one should note that having the unit-root avoided does not necessarily avoid non-stationarity. Hence, ultimately ADF only tests for part of the problem of stationarity.

Second, in terms of the models — no one has prior certainty about how the true DGP behaves. Of course, when practitioners deal with the natural frequency of the data, some artificial assumptions can be made, e.g. one may choose $p = 12$ for monthly data. But, when the model specification becomes wrong, there is a high chance for ADF test to fail bizarrely. The following lemma gives a general statement on counter-examples.

Lemma 2.8 (Limitation of ADF test)
Suppose the true DGP is an AR(p) process, so the ADF test as defined in Definition 2.7 is being correctly specified. If the alternative hypothesis of the ADF is satisfied, then the AR(p) process has no unit roots.
However, if we suppose the true DGP is an AR(p+1) process, then, when an ADF test as defined in Definition 2.7 is being considered with an AR(p) model, even if the alternative hypothesis of the ADF is satisfied, there still exists ϕ_{p+1} such that y_t is non-stationary.

Proof. To show the first part of the lemma, we observe the polynomial

$$\Phi_p(L) := 1 - \sum_{i \in [p]} \phi_i L^i$$

can be written as

$$\Phi_p(L) = 1 - \left((\gamma + 1 + \theta_1)L + \sum_{i=2}^{p-1} (\theta_i - \theta_{i-1})L^i - \theta_{p-1}L^p \right)$$

Hence $\Phi_p(1) = -\gamma > 0$. No unit root as desired.

Now, for the second part of the lemma, to show the AR(p+1) non-stationary, it is sufficient to show that the polynomial $\Phi_{p+1}(L)$ has unit root. Now, observe

$$\Phi_{p+1}(L) = \Phi_p(L) + \phi_{p+1}L^{p+1}$$

Therefore, if we set $\phi_{p+1} = \gamma$, then $\Phi_{p+1}(1) = 0$, hence there is a unit root. □

Lemma 2.8 purely works from a theoretical argument — that due to omitted relevant variable, i.e. the ϕ_{p+1}, one has failed to acknowledge the existence of unit root even the ADF test suggests stationarity. In what follows, we observe more closely on the estimation front, how a misspecification can be bad for estimation and therefore forecasting.

2.2.2 Stationarity as a matter of stability

Lemma 2.9 (The consequence of correct specification)
Suppose y_t to have a true stationary AR(1) process

$$y_t = \phi_1 y_{t-1} + \varepsilon_t \tag{14}$$

And that we use an AR(1) model to fit the data $\{y_0, ..., y_T\}$, noted

$$y_t = \rho y_{t-1} + e_t \quad , \; e_t \sim N(0, \sigma^2) \quad \forall t \in \{0, 1, ..., T\} \tag{15}$$

Then, for the OLS estimate $\hat{\rho}$,

$$\hat{\rho} \xrightarrow[T\to\infty]{P} \phi_1 \tag{16}$$

$$T^{\frac{1}{2}}(\hat{\rho} - \phi_1) \xrightarrow[T\to\infty]{D} N(0, 1 - \phi_1^2) \tag{17}$$

Hence

$$\hat{y}_{T+1|T} - y_{T+1|T} \xrightarrow[T\to\infty]{P} 0$$

Comment: note the difference compared to subsubsection 1.4.1 — in time series, we are unable to claim conditional distribution like Equation 7, as the "explanatory variable" is in fact part of the dependent variable. This also makes small-sample inference barely possible.

Proof. Observe that

$$\hat{\rho} - \phi_1 = \frac{\sum_{t\in[T]} y_{t-1}\varepsilon_t}{\sum_{t\in[T]} y_{t-1}^2}$$

And that by CLT,

$$T^{-\frac{1}{2}} \sum_{t\in[T]} y_{t-1}\varepsilon_t \xrightarrow[T\to\infty]{D} N(0, \sigma^4(1-\phi_1^2)^{-1})$$

While by LLN,

$$T^{-1}\sum_{t\in[T]} y_{t-1}\varepsilon_t \xrightarrow[T\to\infty]{P} 0$$
$$T^{-1}\sum_{t\in[T]} y_{t-1}^2 \xrightarrow[T\to\infty]{P} \sigma^2(1-\phi_1^2)^{-1}$$

□

Lemma 2.10 (The consequence of misfit)
Suppose y_t to have a true stationary AR(2) process

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t \tag{18}$$

Suppose further that $v_1 := Var(y_t y_{t-1})$ is finite and time-independent. Denote $v_2 := Var(y_t) = (1-\phi_2)\sigma^2((1+\phi_2)((1-\phi_2)^2-\phi_1^2))^{-1}$.

Suppose that we use an AR(1) model to fit the data $\{y_0, ..., y_T\}$ as noted in Equation 15, then

$$\hat{\rho} - \phi_1 \xrightarrow[T\to\infty]{P} \frac{\phi_1\phi_2}{1-\phi_2} \tag{19}$$
$$T^{\frac{1}{2}}(\hat{\rho} - \phi_1 - \frac{\phi_1\phi_2}{1-\phi_2}) \xrightarrow[T\to\infty]{D} N(0, \frac{v_1\phi_2^2 + v_2\sigma^2}{v_2^2}) \tag{20}$$

Hence

$$\hat{y}_{T+1|T} - \frac{\phi_1}{1-\phi_2} y_T \xrightarrow[T\to\infty]{P} 0$$

Proof. Observe that

$$\hat{\rho}-\phi_1 = \frac{\left(\phi_2\sum_{t\in[T]} y_{t-2}y_{t-1}\right) + \left(\sum_{t\in[T]} y_{t-1}\varepsilon_t\right)}{\sum_{t\in[T]} y_{t-1}^2} = \phi_2\left(\frac{\sum_{t\in[T]} y_{t-2}y_{t-1}}{\sum_{t\in[T]} y_{t-1}^2}\right) + \frac{\sum_{t\in[T]} y_{t-1}\varepsilon_t}{\sum_{t\in[T]} y_{t-1}^2}$$

and that

$$T^{-1}\sum_{t\in[T]} y_{t-2}y_{t-1} \xrightarrow[T\to\infty]{P} \frac{\phi_1}{1-\phi_2} v_2$$
$$T^{-1}\sum_{t\in[T]} y_{t-1}\varepsilon_t \xrightarrow[T\to\infty]{P} 0$$
$$T^{-1}\sum_{t\in[T]} y_{t-1}^2 \xrightarrow[T\to\infty]{P} v_2$$

This leads to Equation 19. Now, observe that

$$T^{\frac{1}{2}}(\hat{\rho}-\phi_1-\frac{\phi_1\phi_2}{1-\phi_2}) = \phi_2\left(\frac{T^{\frac{1}{2}}\left(T^{-1}\sum_{t\in[T]} y_{t-2}y_{t-1}\right)}{T^{-1}\sum_{t\in[T]} y_{t-1}^2}\right) - \left(\sqrt{T}\frac{\phi_1\phi_2}{1-\phi_2}\right) + \frac{T^{-\frac{1}{2}}\sum_{t\in[T]} y_{t-1}\varepsilon_t}{T^{-1}\sum_{t\in[T]} y_{t-1}^2}$$

and that

$$T^{\frac{1}{2}}\left(\left(T^{-1}\sum_{t\in[T]}y_{t-2}y_{t-1}\right)-\left(\frac{\phi_1}{1-\phi_2}T^{-1}\sum_{t\in[T]}y_{t-1}^2\right)\right)\xrightarrow[T\to\infty]{D}N(0,v_1)$$

$$T^{-\frac{1}{2}}\sum_{t\in[T]}y_{t-1}\varepsilon_t\xrightarrow[T\to\infty]{D}N(0,\sigma^2 v_2)$$

□

Corollary 2.11 (Instability under misfit, a consequence of Lemma 2.10)

$$\hat{y}_{T+1|T}-y_{T+1|T}-\left(\phi_2(\phi_1 y_T(1-\phi_2)^{-1}-y_{T-1})\right)\xrightarrow[T\to\infty]{P}0$$

Lemma 2.10 and Corollary 2.11 describes, under a time series environment, how underfit hurts the estimation, and ultimately forecasting. The usage of CLT offers certain level of asymptotic variance analysis. In the next subsection, we demonstrate further the instability issues by examples.

2.3 Examples of instability

2.3.1 ADF test

Here, we simulate 3 short time series $\{\{y_t^p\}_{t=1}^{30}\}_{p\in[3]}$ that follows AR(1), AR(2), and AR(3) respectively. We run an ADF test with $p=1,2,3$ respectively, and hence observe the behaviour under misspecification.

For the specification of the DGP of the AR(1) model, we put $\phi_1=0.9$, for the DGP of the AR(2) model, we put $\phi_1=0.7$ and $\phi_2=-0.2$, and for AR(3), we put $\phi_1=\phi_2=0$ with $\phi_3=0.9$. From theory, we know all three series should be stationary, note as $\{y_t^p\}_{t=1}^{30}\sim I(0)$.

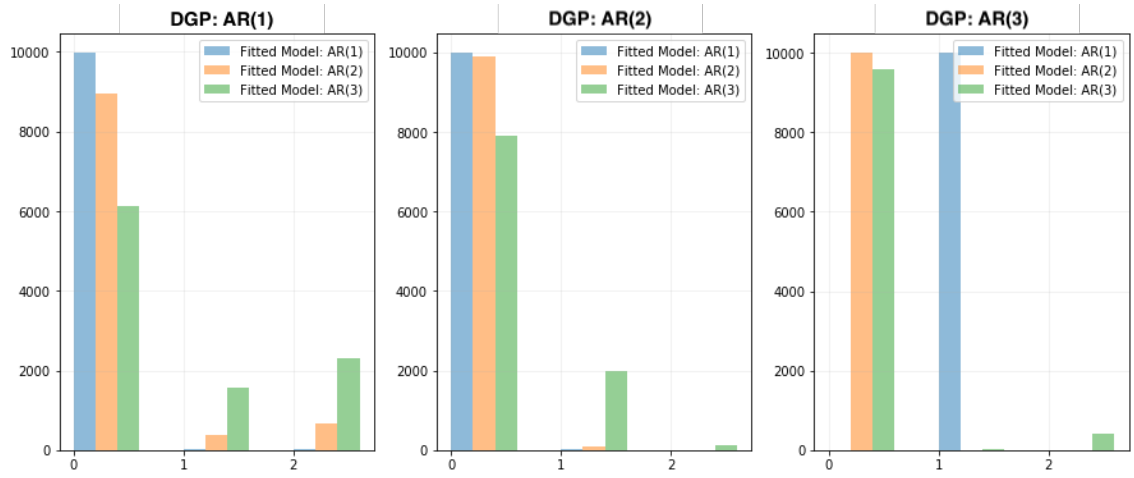

Figure 5: Count of conclusions of the hypothesis testing by ADF tests. 0 and 1 indicates $\{y_t^p\}_{t=1}^{30}\sim I(0)$ and $I(1)$ respectively. 2 indicates we concluded $\{y_t^p\}_{t=1}^{30}\nsim I(d)$ for any $d\leq 2$

We run the ADF tests with specified lags $p = 1, 2, 3$ separately on each of the three data series, and make a 5% significance level conclusion. We repeat such an experiment for 10000 times and report the count of $I(0)$, $I(1)$, or else in Figure 5.

As reported, we see the correct specification does the job of concluding stationarity correctly, whereas incorrect specification fails massively to conclude the right process — over-specification, for instance the green bars on the left panel, demonstrates only around 60% of times when $I(0)$ is concluded, due to over-fitting an AR(3) onto an AR(1) process. Likewise, under-specification, see the blue bars on the right panel, concludes $I(1)$ for almost all experiments, due to under-fitting an AR(1) onto an AR(3) process.

2.3.2 Model behaviour: misfit and forecasting miseries

Here, we further investigate the misfit and its associated evaluations. Based on 1000 replications of the time series as per specified in Figure 1, we produce a count of the number of the best models induced by each of the AIC, MAE, and RMSE model evaluations. This is illustrated in Figure 6. The average of these evaluation statistics has been plotted in Figure 4.

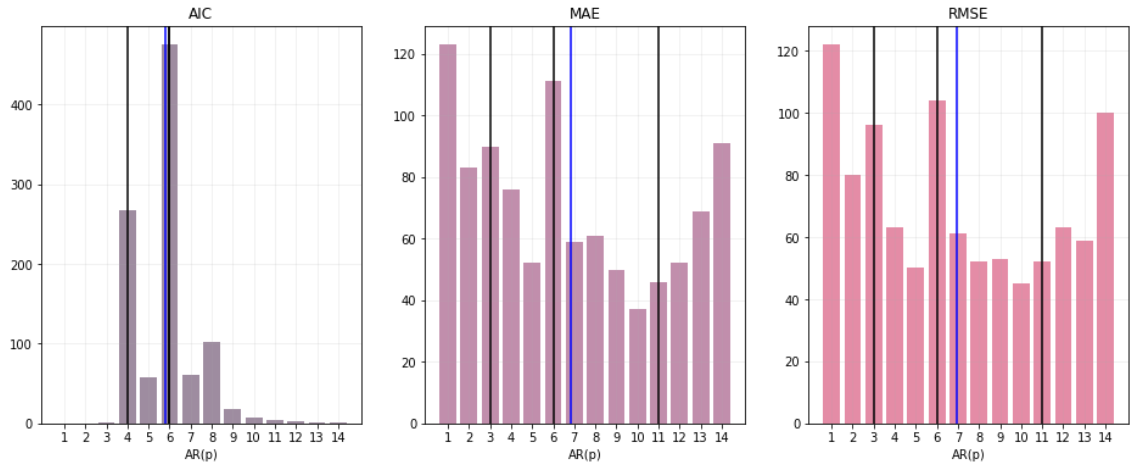

Figure 6: Count of the best models induced by the in-sample AIC and out-of-sample MAE and RMSE. The three black lines indicates 25%, 50%, and 75% quantiles, and the blue line indicates the mean.

A direct observation would be that the MAE and RMSE do not have a huge distaste towards large models. Moreover, unlike in AIC where there exists a dominating mode towards $AR(6)$, the distributions in MAE and RMSE have relatively even shape. This, nonetheless, does not justify AIC as an appropriate model evaluator — AIC has its own problems — it favours a model which under-fits, whereas the average of MAE and RMSE would favour $AR(7)$, which is correctly fitted.

Overall, it comes into further miseries about what is the best modelling — in the above experiment, many models can outperform AR(7) in forecasting, as demonstrated by the centre and right panels of Figure 6. But AR(7) is the correct model specification, though AIC fails to detect it for most of the times, and there are no clear favour towards the

AR(7) in the other two statistics. This contributes to one of the key motivations to the proposal of adaptive learning, as presented in the next subsection.

2.4 Adaptive learning

2.4.1 Introduction

In this subsection, we propose a novel learning method in time series, named as adaptive learning. The key concept is that one should exploit the ordered nature of the time series, that because it is ordered, its past can verify forecasts that were produced based on previous observations, hence dynamically offer information which can be helpful for model determination. This can be similar to some basic statistical learning concepts including cross validation and can overlap with methods including model ensemble, which are detailed in the later part of this section.

Despite scarce study on the asymptotic or theoretical behaviour, the use of adaptive learning has contributed to empirical studies in Macroeconomics and Finance (Yang 2019; Yang 2020; Yang 2021), which are illustrated as examples in the later part of this section.

2.4.2 Notation

We start by summarising the operation of forecasting in a conclusive expression. To do so, we re-introduce the information set representing all data available up to time t, with the inclusion of explanatory variables x_τ:

$$\Phi_t := \{(x_\tau, y_\tau) : \tau \in [t]\}$$

Hence, we can write the k-step-ahead forecast at time t into the following notation

$$\mathbb{E}[y_{t+k}|\Phi_t] =: f_k(\Phi_t; \theta_t; h_t)$$

where the right hand side is a general notation for a forecast — θ_t stands for the contemporary parameter and h_t stands for the contemporary functional form.

Let $\mathcal{D} := \{(x_t, y_t) : t \in T\}$ be a time series dataset where T is the index set. Let the validation and testing index sets be $T^{\texttt{validation}}$ and $T^{\texttt{test}}$ where $\max(T^{\texttt{validation}}) < \min(T^{\texttt{test}})$. We notice that the index set is partitioned such that the dataset is created for its labelled purposes, e.g. the validation dataset can be written as $\{(x_t, y_t) : t \in T^{\texttt{validation}}\}$. Also note that the training takes place at each time $t \in T^{\texttt{validation}} \cup T^{\texttt{test}}$, and that the training set varies by t. For instance, a w windowed training set at time t can be expressed as

$$\{(x_\tau, y_\tau) : \tau \in [t] \setminus [t - w]\}$$

2.4.3 Definition of adaptive learning and examples

Definition 2.12 (Adaptive learning)
An adaptive learning on a time series dataset $\mathcal{D}$ and forecasting horizon k, is a learning method in which there is a learning function l, a set of models H, and estimation techniques

$$\Lambda_h : \Theta(h) \mapsto \theta(h) \quad \forall h \in H$$

where $\Theta(h)$ is the domain of the parameter and $\theta(h)$ is the estimated parameter, such that for all time $t \in T^{validation} \cup T^{test}$, we are able to obtain the followings:

1. *$\forall h \in H$, an estimated statistic $\theta_t(h)$*
2. *$\forall h \in H$, an estimated forecast $y_{t+k|t}(h)$*
3. *an optimal model $h_t^* \in H$ selected by l, with an induced optimal forecast $y_{t+k|t}(h_t^*)$*

The design of $(l, H, \{\Lambda_h\}_{h \in H})$ should be made on the basis of theoretical validity and tuned in the validation data. Their ability to improve forecasts can then be demonstrated in the testing data.

Example 2.13 (Dynamic model selection)
Let H be finite. Each $h \in H$ gives $f_k(\cdot;\cdot;h)$ a functional specification. For instance, it could be $f_k(\Phi_t;\theta_t;h) = \theta_{t,1} + \theta_{t,2}x_t$. $\Theta(h) = \mathbb{R}^2$ in this case, and the training of $\theta(h)$ can be done by MLE over a w windowed training set at time t.

The above specifies H and $\{\Lambda_h\}_{h \in H}$. Now, we introduce an MSE-styled loss function, using the notations as per Equation 11:

$$l_t(h) = \sum_{\tau \in \tilde{T}(t)} RMSE_\tau(h)$$

At time t, $\tilde{T}(t) = \{t - v - k + 1, ..., t - k - 1, t - k\}$ indicating that we equally evaluate the loss over the v windowed history up to time $t - k$.

Then, we can obtain an optimal model $h_t^* \in H$ by executing a finite minimisation operation:

$$h_t^* = \arg\min_{h \in H} l_t(h)$$

For computing implementations, we illustrate one of the previous works in Figure 7. A more detailed sample algorithm for Example 2.13 is supplied in subsection B.2.

Algorithm 2: Algorithm for obtaining the forecasts with a time-varying h_t

Input: Data $\{\Phi_t\}_{t\in T}$, functional sets H, specification of the loss function $\ell(\cdot,\cdot)$, desired forecasting index set T, and validation data $\{y_{t+1}\}_{t\in T}$.

Output: Forecasts $\{y_{t+1|t}(h_t^*)\}_{t\in T}$ with the associated functions $\{h_t^*\}_{t\in T}$, and the performance metric.

1. For $t \in T$, repeat:
   (a) Produce $\tilde{H}_t$ according to Equation 11.
   (b) Evaluate and execute the minimisation given by Equation 12. Then get h_t^* with $y_{t+1|t}(h_t^*) = \mathbb{E}[y_{t+1}|\theta_t(h_t^*), \Phi_t, h_t^*]$.
2. Evaluate the performance metric.

Figure 7: A sample algorithm for dynamic model selection (Yang 2021)

Example 2.14 (Model ensemble)
We inherit the set-up from Example 2.13. Denote $|H| = n$ *and* $H = \{h_1, ..., h_n\}$. *Now, consider an n-dimensional simplex* $P := \{p \in \mathbb{R}^n : ||p||_1 = 1\}$

By model ensemble, we mean

$$y_{t+k|t}(h^{ensemble}) = \langle p_t^*, \big(y_{t+k|t}(h_1), ..., y_{t+k|t}(h_n)\big)\rangle$$

And as an example, a simple way to train $p_t^* \in P$ *would be the empirical distribution of*

$$\left\{\arg\min_{h\in H} RMSE_\tau(h)\right\}_{\tau\in\tilde{T}(t)}$$

Example 2.15 (Extension to penalised MLE)
TBC

Example 2.16 (Portfolio optimisation: a semi-parametric extension)
TBC

3 Additional research remarks

Key contributions:

- Section 1:
  - Provided a framework for analysing stability and instability of statistical modelling.
  - Engaged with the analysis of NN and adversarial attacks from a different dimension.
- Section 2:
  - Extended the framework to time series analysis.
  - Demonstrated the distributional results of autoregressive misfit, and proposed a new learning regime with potential extensions.

Further works:

- Section 1:
  - Further diggings in to more distributions, analysis of NN, and adversarial attacks.
  - Relevant work-in-progress in AFHA: Bastounis, Hansen, and Vlacic (2021).
- Section 2:
  - More inference and computational experiments. Then upload to Arxiv.
  - Combine empirical results (collaboratively with OPRG) to submit to the BoE annual conference and RSSC.

A Mathematical Appendix

A.1 Useful results

Lemma A.1 (Cauchy Schwartz Inequality)
Let $(V, \langle , \rangle)$ be an inner product space and let $||v|| := \langle v, v\rangle$ for all $v \in V$. Then

$$|\langle x, y\rangle| \leq ||x||||y|| \quad \textit{for all } x, y \in V$$

Lemma A.2 (Tail bounds of Gaussian distribution)
Let $X \sim N(0, \sigma^2)$, then $\forall t > 0$,

$$\mathbb{P}[|X| > t] < \sqrt{\frac{2}{\pi}} \frac{\exp(-\frac{t^2}{2\sigma^2})}{t}$$

Proof. Note that by the property of scaling, it is sufficient to show the case under $\sigma = 1$. Observe

$$\begin{aligned}
\mathbb{P}[X > t] &= \frac{1}{\sqrt{2\pi}} \int_t^\infty \exp\left(-\frac{x^2}{2}\right) dx \\
&< \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} \exp\left(-\frac{x^2}{2}\right) dx \\
&= \frac{1}{t\sqrt{2\pi}} \exp(\frac{-t^2}{2})
\end{aligned}$$

where the last line comes from integration by parts. This finishes the proof as $\mathbb{P}[X > t] = \mathbb{P}[X < -t]$ □

Theorem A.3 (Law of Large Numbers)
Let $\{y_\tau\}_{\tau \in \mathbb{N}}$ be stationary as defined in Definition 2.5. Define $\bar{y}_t := t^{-1} \sum_{\tau=1}^t y_\tau$, then

$$\bar{y}_t \xrightarrow[t\to\infty]{P} \mu$$

Theorem A.4 (Central Limit Theorem (White 1984))
Inherit the notations from Theorem A.3. Suppose further that there exists $r > 2$ such that for all τ, $\mathbb{E}[|y_\tau - \mu|^r] < \infty$, then

$$\sqrt{t}(\bar{y}_t - \mu) \xrightarrow[t\to\infty]{D} N(0, \gamma_0)$$

B Computational appendix

B.1 Misfit & forecasting problems

ARIMA	(12,0,0)	(13,0,0)	(14,0,0)	(11,0,0)	(7,0,0)	(8,0,0)	(10,0,0)	(6,0,0)	(2,0,2)	(4,0,0)	(5,0,0)	(9,0,0)	(3,0,0)	(1,0,2)	(1,0,0)	(1,0,1)	(2,0,0)	(2,0,1)
AMAE	1.2905	1.2870	1.2890	1.2828	**1.2737**	1.2846	1.3134	1.3664	1.3650	1.3677	1.3747	1.3294	1.6121	1.7470	1.7510	1.7499	1.7486	1.7495
ARMSE	**1.2402**	1.2414	1.2419	1.2425	1.2505	1.2579	1.2582	1.2619	1.2676	1.2722	1.2737	1.2761	1.4108	1.4478	1.4638	1.4642	1.4646	1.4681

Table 1: MAE and RMSE of the forecasts as per Figure 2

B.2 A sample algorithm for implementing adaptive learning

Input:

- Data: $\{y_t : t \in T\}$.
- Model specifications: H.
- Forecasting horizon k.
- Window size for the model estimations w.
- Specification of loss function: $l(\cdot)$. Default choice can be ARMSE($\tilde{T}$).

$$l(\{e_\tau\}_{\tau\in\tilde{T}}) = \sum_{\tau\in\tilde{T}} ||e_\tau||_2 \tag{21}$$

  At time t, $\tilde{T} = \{t-v-k+1, ..., t-k-1, t-k\}$

Output:

- T^{test} the index of testing forecasts.
- Induced-optimal functional forms $(h_t^*)_{t\in T^{test}}$
- Induced-optimal forecasts for testing purposes: $\{\hat{y}_{t+k|t}(h_t^*)\}_{t\in T^{test}}$
- General evaluation of MAE and MSE at time $t \in T^{test}$ and the overall evaluations.

Computing steps:

1. Call $T^{max} := \max(T)$. Define $T^{test} = \{w+v+k, ..., T^{max}-k-1, T^{max}-k\}$
2. For all $t \in \{w+1, ..., T^{max}-k\}$:
   (a) For all $h \in H$:
      i. Train h over the data $\{y_\tau : \tau \in [t] \setminus [t-w-1]\}$
      ii. Obtain vector $\hat{y}_{t+k|t}(h)$
      iii. Obtain vector $e_t(h)$
3. For all $t \in T^{test}$:
   (a) Declare $\tilde{T}$.
   (b) For all $h \in H$:
      i. Collect the set$\{e_\tau(h)\}_{\tau\in\tilde{T}}$
      ii. Evaluate $\ell(h) = l(\{e_\tau(h)\}_{\tau\in\tilde{T}})$
   (c) Find $h^* := \arg\min\{\ell(h) : h \in H\}$.
   (d) Hence save this h_t^* and make and save the associated forecast $\hat{y}_{t+k|t}(h_t^*)$
4. Produce general evaluation for $\{\hat{y}_{t+k|t}(h_t^*)\}_{t\in T^{test}}$. Return the output.

References

Bastounis, A., A. C. Hansen, and V. Vlacic (2021). "The mathematics of adversarial attacks in AI." In: Work in progress.

Bluwstein, Kristina et al. (2020). "Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach." In: *Bank of England Working Paper* 848.

Efron, Bradley and Trevor Hastie (2016). *Computer age statistical inference : algorithms, evidence, and data science*. CUP.

Hamilton, James D (1994). *Time Series Analysis*. Princeton University Press.

Joseph, Andreas (2019). *Parametric inference with universal function approximators*. eprint: arXiv:1903.04209.

Klenke, Achim (2020). *Probability Theory: A Comprehensive Course*. 3rd ed. Springer.

Prado, Raquel and Mike West (2010). *Time Series: Modeling, Computation, and Inference*. 1st ed. Taylor and Francis Group.

Tsay, R.S. and R. Chen (2018). *Nonlinear Time Series Analysis*. Wiley.

Tsay, Ruey S. (2010). *Analysis of Financial Time Series*. Wiley.

Vapnik, Vladimir (1999). *The Nature of Statistical Learning Theory*. Springer.

White, Halbert (1984). *Asymptotic Theory for Econometricians*. San Diego: Academic Press.

Yang, Parley Ruogu (2019). *Bank of England: Modelling with Big Data and Machine Learning Conference*. URL: https://www.bankofengland.co.uk/events/2019/november/modelling-with-big-data-and-machine-learning.

— (2020). "Using the yield curve to forecast economic growth." In: *Journal of Forecasting* 39.7, pp. 1057–1080. DOI: 10.1002/for.2676.

— (2021). "Forecasting high-frequency financial time series: an adaptive learning approach with the order book data." In: eprint: arXiv:2103.00264.