mathematics

MDPI

Article

On Model Improvement Algorithms—Generalised Linear Models and Neural Networks

Manuel L. Esquível [1,2,*,†], Nadezhda P. Krasii [2,3,†] and Raquel M. Gaspar [4,†]

[1] Department of Mathematics, NOVA School of Science and Technology (NOVA FCT), 2829-516 Caparica, Portugal
[2] Center for Mathematics and Applications (NOVA Math), Universidade Nova de Lisboa, 2829-516 Caparica, Portugal; n.krasii@fct.unl.pt
[3] Department of Higher Mathematics, Don State Technical University, Gagarin Square 1, Rostov-on-Don 344000, Russia
[4] ISEG Research, Lisbon School of Economics and Management, Universidade de Lisboa, 1200-781 Lisbon, Portugal; rmgaspar@iseg.ulisboa.pt
* Correspondence: mle@fct.unl.pt; Tel.: +351-96-55-44-623
† These authors contributed equally to this work.

Abstract

We propose a generic approach to stochastic model improvement by first introducing an archetypal algorithm based on error minimisation and establishing two results on the weak convergence of the probability laws associated with the models under improvement. We then present two concrete instances of this approach: Generalised Linear Models and classical multivariate models assessed using a neural network. In both cases, we illustrate the methodology using economic, financial, and social data related to the determination of government bond coupon rates prior to primary market auctions. For each application, we derive weak convergence results that specify conditions under which model improvement occurs, in the sense of convergence in law of the probability distributions associated with successive models. These results ensure the convergence of the proposed archetypal algorithm and provide a probabilistic foundation for systematic model improvement.

Keywords: stochastic models; model improvement algorithms; algorithm convergence; weak convergence of distributions; generalised linear models; neural networks; bond coupon rates prior to auction

MSC: 65Y20; 68W20; 68W40; 62J12; 68T07; 91B26; 91B70

1. Introduction

The main content of this work is a set of instances of results that fall within the generic description of an algorithm for stochastic model improvement. To illustrate this main content, we provide two applications—the first one in the Generalised Linear Model (GLM) context and the second one in the neural networks (NN) context. The algorithm detailed below in schematic form in Algorithm 1 relies on controlling an error that must be defined for each of the particular contexts that the algorithm is applied to. The general form of the conclusion of the results presented is the convergence in law of a sequence of distributions of the successive models to a final optimal model; a model with minimal error that can be either deterministic or stochastic. The two applications that are intended to illustrate the approach for model improvement presented apply stochastic models to data relating economic, financial, and social factors to the coupon rates determined by countries prior to bond auctions in the primary markets.

Check for updates

Academic Editor: Andreas C. Georgiou

Received: 21 November 2025
Revised: 20 January 2026
Accepted: 24 January 2026
Published: 4 February 2026

Algorithm 1 Generic algorithm for model improvement

1: Initial Model Fitting: $\mathcal{M} \leftarrow M$ ▷ An initial model fitting is implemented
2: Error: $\mathcal{E} \leftarrow E$ ▷ The error control E is initialised into $\mathcal{E}$
3: **while** Algorithm can be improved and/or error decreases **do**
4: Determine new Model M and compute new Error E
5: **if** Algorithm can be improved and/or $E < \mathcal{E}$ **then**
6: Error: $\mathcal{E} \leftarrow E$ ▷ The newly computed error control now into $\mathcal{E}$
7: Go to Step (line) 4
8: **else**
9: Error $\mathcal{E}$ control variable does not change
10: Go to Step (line) 4
11: **end if**
12: **end while**

The generic algorithm considered may be thought of as a type of stochastic optimisation algorithm but with changing models, with the optimisation goal being to determine an optimal model. The perspective of considering the possibility of optimizing the models by reducing some error is of great importance for the current practice of model risk management (see, for instance, refs. [1–4]). Under this perspective, the work [5] deserves a special mention, since it deals with periodic recalibration of parameters and initial conditions of pricing and hedging models to eliminate any conflict between model-implied and market prices. The paper shows "under which conditions recalibration improves the hedging errors by limiting the propagation of an initial error".

Financial practitioners, for instance, are well aware that there is a risk inherent to the use of a certain model while searching to describe or predict aspects of financial observable reality and that it is fundamental to take into account the risk of model choice. To the best of our knowledge, there is systematic approach focusing on stochastic model improvement via model error control provided in the literature. The novel idea introduced in this work is to show that by defining a model improvement algorithm with an adequate selection of an error—that should be minimised in order to improve the model—we can control model risk. While the approach appears to align with the paradigm of classical stochastic optimisation, this is not the case; as demonstrated in the application Section The Neural Network as a Tool for an Example of a Model Improvement Algorithm, the methodology applies when models vary inside a family for which a common notion of error can be defined, for instance, via an external computational tool such as a fixed neural network.

Despite the novelty of our approach in this work we can, nevertheless, connect our approach of model improvement to several works that we detail in the following. The review of the selected examples that follows indicates that the idea of algorithms assessing model improvements is latent across varied fields of research, although not explicitly stated as a formal objective.

In the context of a mathematical theory of design proposed by Braha and Maimon containing a "a method for adaptive learning of successful designs that is based on the use of statistical experimental design and a stochastic search algorithm," the authors render explicit an adaptive algorithm that search for an optimal formulation of a design, a search that uses metrics to assess the distance between the current state of the process and an ultimate objective goal for the process (see [6], pp. 374–375). We consider this approach as a precursor example of a model improvement algorithm.

In a more recent work [7], a Kullback–Leibler divergence loss, that is, a measure of how one probability distribution differs from a second probability distribution considered as a reference, is taken as a criterion to assess an improvement of computerised learning.

In the work [8], a new practical methodology was introduced to generate sets in the plane with fractal type patterns. This methodology was compared with classic methodologies to generate fractals such as Barnley's iterated function systems by means of several criteria. One of the criterion used is the precision in structural detail which amounts to an error criterium. In [9], the authors propose, and we quote, an "algorithm for black-box optimisation problems without derivatives in the presence of output constraints".

The title of [10] may suggest an approach similar to the one introduced in this work but it is not so. In fact, the article deals with "a novel denoising optimisation technique specifically designed for time series data that are subject to stochastic processes." The empirical analysis "demonstrates that introducing noise, in conjunction with optimizing latent representations and target variables, effectively reduces overfitting."

In the following, we present a brief summary of the contents of the remaining sections while highlighting the main contributions of this work in each of the sections.

- In Section 2, we describe the algorithm that allows for a sequential improvement of a sequence of stochastic models. We briefly discuss two instances of application of the archetypal algorithm to real data—in particular, in a Generalised Linear Model (GLM) context and using a fixed neural network (NN)—detailing our choice of errors to control model improvement.
- In Section 3, we present the first contributions in this work to stochastic model improvement in which a given dependent variable is a function of a finite set of random factors plus an additive noise that acts as an error associated to the model. In order to clarify this algorithm, we detail, in Theorem 1 and Corollary 1, two results on the convergence of the algorithm; for clarity, we propose, for the reader to suppose that we are in a GLM fitting of data, but the results are general in the sense that they only suppose a sequence of models operating on a finite number of stochastic factors and the corresponding additive random errors. These results show that if the sequence of functions linking the factors to the dependent variable converges in an adequate way and if the errors converge in probability, then the joint laws of the models and errors converge, thus ensuring the algorithm convergence.
- In Sections 4–6, we present real data and two applications to real-word data as a proof of concept of the methodologies proposed. We aim to study the improvement of stochastic models in two different contexts—GLM (in Section 5) and NN (in Section 6)—applied to the problem of determining the initial interest rate for long-term bonds; we stress that the interest rate that we consider is an interest rate fixed by the issuer of the bond prior to the auction in the primary market. We will consider that the interest rate under consideration depends on economic, financial, and social factors of the country issuing the bond. Firstly we investigate—in the GLM context—what the most significant factors are and secondly—in the NN context—we investigate several statistical models for the joint distribution of factors and associated interest rates using only the most significant factors.
- In Section 5.3, we present another contribution of this work—Theorem 4—showing the weak convergence of the joint laws of the sequence of models and correspondent errors in the GLM context that was shown to be supported by—in Theorem 3—also one of the results presented in Section 5.
- In Section 6, we present another contribution of this work Theorem 5, that is, a result showing that we also get the convergence in law of the joint laws of the models; this is done by considering the expected squared prediction error—computed in the application of a neural network fitting—as the error associated to a given model.

2. A Generic Model Improvement Algorithm

In this section, we formally propose an archetypal algorithm that underlies the main results presented in this work. Despite being a simple idea, it is a powerful one since it allows us to describe a general method that has many instances, as is shown in Theorems 1, 4, 5 and Corollary 1. The archetypal algorithm proposed essentially says that if we have a sequence of models together with a coherent notion of error associated with each model, if the error sequence associated with the sequence of models decreases, then model improvement is achieved. As it happens with all algorithms, the main question is proving the convergence of the algorithm in each of its instances of application. This is achieved in Theorems 1, 4, 5 and Corollary 1 for several instances of the archetypal algorithm. The pseudocode is presented next in Algorithm 1.

It is clear in the schematic presentation of Algorithm 1 that it is necessary to define properly what is the error associated to each model. We will consider two instances of this generic model improvement algorithm firstly in the GLM context and secondly in the NN context. The main characteristics of the errors associated with models considered in these two contexts are the following.

1. The Generalised Linear Model (GLM) fitting context. Usually there are at least two ways of controlling the improvement of the algorithm. The first is eliminating variables that are not statistically significant. And the second is to check for a systematic reduction of the Akaike Information Criterion and Bayesian Information Criterion, which are essentially, in this case, goodness-of-fit measures of the model. It is also possible to consider two error measures associated with each fitting.
   (a) The root mean square (RMS) of the standard errors obtained, in the model fitting, for each estimate. The RMS is a summary measure of the uncertainty associated with the coefficient estimates providing a single value that aggregates the variability of the individual coefficient estimates.
   (b) The probability distribution of the fitting errors, that is, the distribution of the sequence of differences between the predicted values and the response of the model. In principle, one would expect a known distribution—specifically, a Gaussian distribution—for the fitting errors. A model choice, between two options, could be justified if the choice falls for the model with the dominated distribution.
2. In the neural network context, we will consider the expected squared prediction error that is provided in each fitting by the R^2 given by the difference to one of the expected squared prediction errors; it is then clear that the convergence to zero of the error will be equivalent to the convergence to one of the correspondent R^2.

In Sections 5 and 6, we explore examples of these two instances of model improvement algorithms.

3. On the Convergence of the Model Improvement Algorithm

In Section 2, we described a two way algorithm—model and error computation—that is applied in Sections 5 and 6 to concrete instances with real data. Concretely, we show how to improve models giving a bond coupon—a nominal interest rate—of bonds issued by a given country, as a function of social, economic, and financial factors of the country. For the reader's benefit, we may suppose to be in this context for the remainder of this section, but the approach is general, as can be seen in the formulation of the main results. In what follows we study—in a general formulation—the convergence of a two step model improvement algorithm in which an initial model may be considered as the first time of application of the first step of the algorithm. For concreteness, the reader can think of the

interest rate $Y(\omega)$—for a country ω—to be initially given by a continuous function f_0 of N observed factors $F_1(\omega), \ldots, F_N(\omega)$ in the form

$$Y(\omega) = f_0(F_1(\omega), \ldots, F_N(\omega)) + E_0(\omega) \, , \tag{1}$$

with $E_0(\omega)$ being an error in the relation between the observed interest rate $Y(\omega)$ and the values observed for the value taken by the model function f_0 on the factors $F_1(\omega), \ldots, F_N(\omega)$.

If we assume that $F_1, \ldots, F_N$ and E_0 are random variables—with the probability space being the set of conditions in a country that may originate the values of the factors for this country—we get a random initial model giving the interest rate Y as a function of the factors in the following formula:

$$Y = f_0(F_1, \ldots, F_N) + E_0 \, . \tag{2}$$

In Theorem 1 and Corollary 1, we exploit the general form of the relation between model f_0, factors $F_1, \ldots, F_N$, and error E_0, in Formula (2), to show that under a set of general hypothesis, there is model improvement as determined by the error reduction. For that purpose, it is possible to consider a sequence of models f_n and errors E_n for Y, depending on factors $F_1, \ldots, F_N$ given by:

$$Y = f_n(F_1, \ldots, F_N) + E_n \, .$$

The convergence of the algorithm just described is the main natural question that we now address. An initial result demonstrates that, under mildly restrictive hypotheses, the sequence of joint laws of the model and the error converge weakly to a probability law on $\mathbb{R}^2$, thus allowing us to establish that the model improvement two-step algorithm converges.

Firstly, for the sake of completeness, we recall some standard definitions and results in the context in which we will be working. These definitions and results are thorough and fully developed in [11–13].

Definition 1. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \longrightarrow \mathbb{R}^N$ be a random variable in this space.*

(a) The probability law of X is the probability measure $\mathcal{L}_X$ defined in the Borel measured space $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$ by:

$$\forall B \in \mathcal{B}(\mathbb{R}^N) \, , \; \mathcal{L}_X(B) := \mathbb{P}[X \in B] \, .$$

(b) A sequence of probability measures $(\mu_n)_{n \geq 1}$, each one of them defined in $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$, converges weakly to a probability measure μ if, for every continuous bounded function defined on $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$, we have that,

$$\lim_{n \to +\infty} \int_{\mathbb{R}^N} f(x) d\mu_n(x) = \int_{\mathbb{R}^N} f(x) d\mu(x) \, .$$

(c) A sequence of random variables $(X_n)_{n \geq 1}$, each one of them taking values in $\mathbb{R}^N$, converges in distribution to a random variable X, if the sequence of their respective probability laws $(\mathcal{L}_{X_n})_{n \geq 1}$ converges weakly to $\mathcal{L}_X$.

(d) A sequence of probability measures $(\mu_n)_{n \geq 1}$, each one of them defined in $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$, is tight if,

$$\forall \epsilon > 0 \, , \; \exists K_\epsilon \Subset \mathbb{R}^N \, , \; \forall n \geq 1 \, , \; \mu_n(\mathbb{R}^N \setminus K_\epsilon) \leq \epsilon \, ,$$

the notation $K_\epsilon \Subset \mathbb{R}^N$ denoting that K_ϵ is a compact set in the topological space $\mathbb{R}^N$.

Related to the notion of tightness, there is a most important result, specifically, Prokhorov's theorem. This theorem asserts that a sequence of probability measures $(\mu_n)_{n\geq 1}$ is tight if and only if every subsequence of $(\mu_n)_{n\geq 1}$ contains a further subsequence converging weakly to a probability measure (for this result and important complementary consequences, see [11] (p. 384), [12] (pp. 59, 60), or [13] (p. 507)).

Theorem 1 (Weak convergence of model errors joint laws—I)**.** *Let $F_1, \ldots, F_N$ be random variables having a known joint law. Consider the sequence of models for Y given by*

$$Y = f_n(F_1, \ldots, F_N) + E_n \,, \tag{3}$$

for $n \geq 1$ and with $f_n : \mathbb{R}^N \longrightarrow \mathbb{R}$ continuous and E_n being an error random variable. Consider the following assumptions.

(i) *The sequence $(f_n)_{n\geq 1}$ converges continuously to some function f, that is*

$$\forall (x_n)_{n\geq 1} \in (\mathbb{R}^N)^{\mathbb{N}_0} \,, \left(\lim_{n\to+\infty} x_n = x\right) \Rightarrow \left(\lim_{n\to+\infty} f_n(x_n) = f(x)\right) .$$

(ii) *The sequence of errors $(E_n)_{n\geq 1}$ converges to zero in probability.*

Let μ_n be the joint law of $f_n(F_1, \ldots, F_N)$ and E_n. Then $(\mu_n)_{n\geq 1}$ converges weakly to the law of $(f(F_1, \ldots, F_N), 0)$.

Proof. We have that the sequence $(f_n(F_1, \ldots, F_N))_{n\geq 1}$ converges in distribution to the random variable $f(F_1, \ldots, F_N)$, by the continuous mapping, Theorem 5.27 in [13] (p. 117), for instance. Since $(E_n)_{n\geq 1}$ converges to zero in probability, it converges to zero in distribution. As a consequence of Theorem 2.7 (v) in [14] (p. 10), we have that $(f_n(F_1, \ldots, F_N), E_n)_{n\geq 1}$ converges in distribution to $(f(F_1, \ldots, F_N), 0)$; that is, the sequence of probability laws $(\mu_n)_{n\geq 1}$ converge weakly to the law of $(f(F_1, \ldots, F_N), 0)$. □

We now deal with a version of the previous Theorem 1 where the behaviour of the sequence of errors $(E_n)_{n\geq 1}$ is more general. We get the tightness of the family of joint laws of the models and the associated errors.

Theorem 2 (Tightness of model errors joint laws)**.** *Let $F_1, \ldots, F_N$ be random variables having a known joint law. Consider the sequence of models for Y given by*

$$Y = f_n(F_1, \ldots, F_N) + E_n \,, \tag{4}$$

for $n \geq 1$ and with $f_n : \mathbb{R}^N \longrightarrow \mathbb{R}$ continuous and E_n being an error random variable. We now consider the following two hypotheses.

(j) *The sequence $(f_n)_{n\geq 1}$ converges—uniformly over compact sets—to some function f.*

(jj) *For every sequence of non-negative constants $(c_n)_{n\geq 1}$ converging to zero, the sequence $(c_n E_n)_{n\geq 1}$ converges to zero in probability.*

Let μ_n be the joint law of $f_n(F_1, \ldots, F_N)$ and E_n. Then the sequence $(\mu_n)_{n\geq 1}$ is tight.

Proof. Let us deal separately with $f_n(F_1, \ldots, F_N)$ and E_n. If hypothesis **(jj)** is verified, we have that $(E_n)_{n\geq 1}$ is tight by reason of Lemma 5.9 in [13] (p. 106). In order to deal with the family of random variables $f_n(F_1, \ldots, F_N)$, for $n \geq 1$, we observe that the multivariate real random variable $(F_1, \ldots, F_N)$ has a probability distribution in $\mathbb{R}^N$ and so this law is tight since $\mathbb{R}^N$ is a Polish space, that is, separable and complete (see [11]) (p. 387) or [12] (p.8)). Let us consider some arbitrary $\epsilon > 0$; since $(F_1, \ldots, F_N)$ is tight, there exists a compact set $K_\epsilon \Subset \mathbb{R}^N$ such that

$$\mathbb{P}[(F_1,\ldots,F_N)\notin K_\epsilon]\leq\epsilon\,. \tag{5}$$

We now take the function $f:\mathbb{R}^N\longrightarrow\mathbb{R}$ in the hypothesis **(j)** above and observe that with f being continuous, we have that $f(K_\epsilon)$ is compact. Since, we have that

$$\{(F_1(\omega),\ldots,F_N(\omega)):\omega\in\Omega\}\subset K_\epsilon\Rightarrow\{f(F_1(\omega),\ldots,F_N(\omega)):\omega\in\Omega\}\subset f(K_\epsilon)\,,$$

we then have that

$$\begin{aligned}\mathbb{P}[\{f(F_1(\omega),\ldots,F_N(\omega)):\omega\in\Omega\}\subset\mathbb{R}\setminus f(K_\epsilon)\,]=\mathbb{P}[f(F_1,\ldots,F_N)\notin f(K_\epsilon)]\leq\\ \leq\mathbb{P}\Big[\{(F_1(\omega),\ldots,F_N(\omega)):\omega\in\Omega\}\subset\mathbb{R}^N\setminus K_\epsilon\Big]=\mathbb{P}[(F_1,\ldots,F_N)\notin K_\epsilon]\leq\epsilon\,,\end{aligned} \tag{6}$$

thus showing that $f(F_1,\ldots,F_N)$ is a tight random variable, as is the bound in Formula (6) that we are now going to take advantage of, together with the hypothesis that the sequence $(f_n)_{n\geq1}$ converges uniformly over compact sets to the function f. We take an arbitrary $\delta>0$ and we have that

$$\exists N_\delta\,,\ (\forall n\geq N_\delta\,,\ \forall y\in f(K_\epsilon))\Rightarrow|f_n(y)-f(y)|\leq\delta\,. \tag{7}$$

Now, since

$$|f_n(y)-f(y)|\leq\delta\Leftrightarrow|f(y)|-\delta\leq|f_n(y)|\leq|f(y)|+\delta\,,$$

we have that $f_n(K_\epsilon)\subset[f(K_\epsilon)]^\delta$, with $[f(K_\epsilon)]^\delta$ being the δ compact neighbourhood of $f(K_\epsilon)$. As a consequence, we now have that

$$\forall n\geq N_\delta\,,\ \mathbb{P}\Big[f_n(F_1,\ldots,F_N)\notin[f(K_\epsilon)]^\delta\Big]\leq\mathbb{P}[f(F_1,\ldots,F_N)\notin f(K_\epsilon)]\leq\epsilon\,,$$

and so the sequence $(f_n(F_1,\ldots,F_N))_{n\geq N_\delta}$ is tight, which implies that $(f_n(F_1,\ldots,F_N))_{n\geq1}$ is tight.

We are now going to show that the sequence $(\mu_n)_{n\geq1}$ of the joint laws μ_n of $f_n(F_1,\ldots,F_N)$ and E_n is a tight sequence Although this is, in general a well-known fact, see exercise 5.9 in [12] (pp. 65) where we give a proof, in this case, for completeness and commodity of the reader. Consider an arbitrary $\epsilon>0$. We have that for a compact set $K^1_\epsilon\Subset\mathbb{R}$

$$\forall n\geq1\,,\ \mathbb{P}\Big[f_n(F_1,\ldots,F_N)\notin K^1_\epsilon\Big]\leq\frac{\epsilon}{2}\,,$$

for another compact set $K^2_\epsilon\Subset\mathbb{R}$

$$\forall n\geq1\,,\ \mathbb{P}\Big[E_n\notin K^2_\epsilon\Big]\leq\frac{\epsilon}{2}\,.$$

We have that $K^1_\epsilon\times K^2_\epsilon\Subset\mathbb{R}^2$; that is $K^1_\epsilon\times K^2_\epsilon$ is a compact set of $\mathbb{R}^2$ and, moreover

$$\begin{aligned}&\forall n\geq1\,,\ \mathbb{P}\Big[\mu_n\Big(\mathbb{R}^2\setminus K^1_\epsilon\times K^2_\epsilon\Big)\Big]=\mathbb{P}\Big[(f_n(F_1,\ldots,F_N),E_n)\notin K^1_\epsilon\times K^2_\epsilon\Big]=\\&=\mathbb{P}\Big[(f_n(F_1,\ldots,F_N),E_n)\in\Big(\mathbb{R}\setminus K^1_\epsilon\times\mathbb{R}\Big)\cup\Big(\mathbb{R}\times\mathbb{R}\setminus K^2_\epsilon\Big)\Big]=\\&=\mathbb{P}\Big[(f_n(F_1,\ldots,F_N),E_n)\in\mathbb{R}\setminus K^1_\epsilon\times\mathbb{R}\Big]+\mathbb{P}\Big[(f_n(F_1,\ldots,F_N),E_n)\in\mathbb{R}\times\mathbb{R}\setminus K^2_\epsilon\Big]=\\&=\mathbb{P}\Big[f_n(F_1,\ldots,F_N)\notin K^1_\epsilon\Big]+\mathbb{P}\Big[E_n\notin K^2_\epsilon\Big]\leq\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon\,,\end{aligned}$$

thus showing that the sequence $(\mu_n)_{n\geq1}$ is tight (see [12] (pp. 57, 59) or [11] (p. 384) for concurrent results). □

Remark 1 (On the hypothesis on the errors of the model)**.** *We observe that if $(E_n)_{n\geq 1}$ converges to zero in probability—as in hypothesis **(ii)** of Theorem 1—we have by definition that:*

$$\forall \epsilon > 0\ ,\ \exists c_\epsilon > 0\ ,\ \mathbb{P}[|E_n| \geq c_\epsilon] \leq \epsilon\ .$$

As a consequence, by definition also, the sequence $(E_n)_{n\geq 1}$ is tight. Also if additionally the sequence $(|E_n|)_{n\geq 1}$ is bounded by M, then by Tchebychev's inequality we have that,

$$\forall n \geq 1\ ,\ \mathbb{P}[|c_n E_n| \geq \epsilon] \leq \frac{\mathbb{E}[|c_n E_n|]}{\epsilon} \leq \frac{c_n M}{\epsilon}\ ,$$

that is, the sequence $(c_n E_n)_{n\geq 1}$ converges to zero in probability and, as a consequence, we have that $(E_n)_{n\geq 1}$ is tight, thus showing that convergence to zero in probability can be replaced by the hypothesis that the sequence $(|E_n|)_{n\geq 1}$ is bounded.

The following result shows that under a more general hypothesis than those of Theorem 1, we can still deduce the weak convergence of the sequence of models. Due to the hypotheses on the sequence of errors, we can no longer obtain an exact model, that is, a final model with zero error. Nevertheless, the approach in the following result is not only more general but also more realistic.

Corollary 1 (Weak convergence of model errors joint laws—II)**.** *With the assumptions and notations of Theorem 2, we have the following conclusions.*

1. *The sequence of the first marginal laws of the sequence $(\mu_n)_{n\geq 1}$—that is, the sequence of the laws of $f_n(F_1, \ldots, F_N)$—converges weakly to the law of $f(F_1, \ldots, F_N)$.*
2. *The sequence of the second marginal laws of the sequence $(\mu_n)_{n\geq 1}$—that is, the sequence of the laws of E_n—converge weakly to the law of $Y - f(F_1, \ldots, F_N)$.*
3. *The sequence of joint laws $(\mu_n)_{n\geq 1}$ converges weakly to the law of the random variable $(f(F_1, \ldots, F_N), Y - f(F_1, \ldots, F_N))$.*

Proof. We proceed in several steps, carefully justifying each convergence claim.

1. On the convergence of $f_n(F)$ to $f(F)$ in distribution. Under hypothesis **(j)**, f_n converges to f uniformly on compact sets. Since f is continuous (as the uniform limit of continuous functions on compact sets), for any sequence $(x_n) \to x$ in $\mathbb{R}^N$, we have,

$$\begin{aligned} |f_n(x_n) - f(x)| &\leq |f_n(x_n) - f(x_n)| + |f(x_n) - f(x)| \leq \\ &\leq \sup_{y\in K} |f_n(y) - f(y)| + |f(x_n) - f(x)| \to 0\ , \end{aligned}$$

where $K = \{x\} \cup \{x_n : n \geq 1\}$ is compact. Thus $f_n(x_n) \to f(x)$ whenever $x_n \to x$.

Now, since F_i are random variables with a known joint law, the random vector $\boldsymbol{F} = (F_1, \ldots, F_N)$ takes values in $\mathbb{R}^N$. By the continuous mapping theorem for convergence in distribution (applied to the function $\boldsymbol{F} \mapsto f_n(\boldsymbol{F})$ and using the pointwise convergence $f_n \to f$ along with the uniform convergence on compacts to handle measurability and continuity issues), we obtain:

$$f_n(F_1, \ldots, F_N) \xrightarrow{d} f(F_1, \ldots, F_N) \quad \text{as } n \to \infty\ .$$

Denote $X_n := f_n(\boldsymbol{F})$ and $X := f(\boldsymbol{F})$.

2. On the tightness of the sequence of joint laws. By Theorem 2, under hypotheses **(j)** and **(jj)**, the sequence of joint laws $\mu_n = \mathcal{L}_{(X_n, E_n)}$ is tight.

3. On the convergence of the first marginal. The first marginal of μ_n is $\mu_n^1 = \mathcal{L}_{X_n}$. From Step 1, we have:

$$\mu_n^1 \xrightarrow{w} \mathcal{L}_X \, .$$

This establishes conclusion (1) of the corollary.

4. On the joint convergence using the deterministic relationship. Recall that by the model Formula (4),

$$E_n = Y - X_n \quad \text{almost surely for each } n \, .$$

Define the function $\Phi : \mathbb{R} \to \mathbb{R}^2$ by $\Phi(x) = (x, Y - x)$. Note that Φ is deterministic and affine, hence continuous. Moreover, $(X_n, E_n) = \Phi(X_n)$ almost surely.

Since $X_n \xrightarrow{d} X$ (Step 1), the continuous mapping theorem applied to Φ yields,

$$(X_n, E_n) = \Phi(X_n) \xrightarrow{d} \Phi(X) = (X, Y - X) \, .$$

This gives joint convergence of the sequence (X_n, E_n) to $(X, Y - X)$ in distribution.

5. On the convergence of the second marginal. The second marginal of μ_n is $\mu_n^2 = \mathcal{L}_{E_n}$. From the joint convergence in Step **4**, by taking the continuous projection $\pi_2(x, y) = y$, we obtain,

$$E_n \xrightarrow{d} Y - X = Y - f(\boldsymbol{F}) \, .$$

This establishes conclusion (2).

6. The identification of the limit joint law. From Step **4**, we have shown directly that $\mu_n \xrightarrow{w} \mathcal{L}_{(X, Y-X)}$. This is conclusion (3).

7. Justification via Prokhorov's theorem in an alternative argument. For completeness, we can also argue using the tightness and uniqueness of the limit point. From Step **2**, $\{\mu_n\}$ is tight. By Prokhorov's theorem, every subsequence $\{\mu_{n_k}\}$ has a further subsequence $\{\mu_{n_{k_\ell}}\}$ that converges weakly to some probability measure μ on $\mathbb{R}^2$. From Step **4**, we know that any such limit μ must satisfy that its first marginal is $\mathcal{L}_X$ and that μ is supported on the line $\{(x, y) : x + y = Y\}$ (since (X_n, E_n) satisfies $X_n + E_n = Y$ almost surely). Moreover, from the joint convergence already established, μ must be $\mathcal{L}_{(X, Y-X)}$. Since all convergent subsequences have the same limit, the entire sequence μ_n converges weakly to $\mathcal{L}_{(X, Y-X)}$.

8. A clarification on the nature of Y. In this proof, Y is treated as a fixed random variable on the same probability space as $\boldsymbol{F}$. The key is that the relationship $E_n = Y - X_n$ holds almost surely for each n, which allows us to express the joint vector (X_n, E_n) as a deterministic function of X_n alone.

This completes the proof of all three conclusions. □

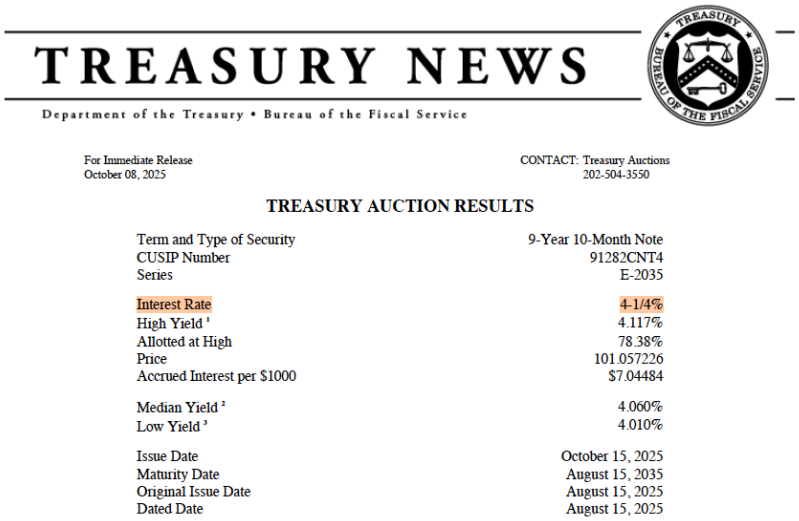
4. Real Data for Two Applications of Model Improvement Algorithms

In this section and also in the following sections, as a proof of concept, we provide applications—of the algorithm for model improvement—to the determination of the coupon rate, that is, the fixed annual interest rate paid by a bond issuer to a bondholder, expressed as a percentage of the bond's face value (par value), specified by the issuer of the bond prior to the auction at the primary market.

Our goal is to use the modelling of the coupon interest rate as a function of socio-economic and financial factors to illustrate the model improvement algorithm.

A bond issued by a national government has two important characteristics at the emission date. The first is the maturity date and the second is the interest rate at the issue date, that is, the coupon interest rate or the coupon, paid by the issuer of the bond to the holders of the bond. An example of an announcement of a bond auction is presented in

Figure 1; the coupon value for that auction is highlighted. This coupon interest rate may be considered as a risk premium. Being so, it matters to the issuer of the bond, to have a way to determine the most adequate coupon interest rate in such a way the agents of the primary market at the auction take the bonds at par, that is, as close as possible of the nominal price of the bonds bought.



TREASURY NEWS

Department of the Treasury • Bureau of the Fiscal Service

For Immediate Release
October 08, 2025

CONTACT: Treasury Auctions
202-504-3550

TREASURY AUCTION RESULTS

Term and Type of Security	9-Year 10-Month Note
CUSIP Number	91282CNT4
Series	E-2035
Interest Rate	4-1/4%
High Yield [1]	4.117%
Allotted at High	78.38%
Price	101.057226
Accrued Interest per $1000	$7.04484
Median Yield [2]	4.060%
Low Yield [3]	4.010%
Issue Date	October 15, 2025
Maturity Date	August 15, 2035
Original Issue Date	August 15, 2025
Dated Date	August 15, 2025

Figure 1. An announcement of a bond auction with the coupon (nominal) rate highlighted.

The determination of the coupon is partially dependent on the values of the yields in the secondary market for a collection of bonds with similar characteristics to the one that is being auctioned.

In the following, we postulate that the coupon—viewed as a risk premium—may be modelled by a function of a collection of social, economic, and financial factors characterising the risk profile of the country issuing the bond; for that purpose, we will use a sequence of GLMs. Having selected a GLM for the coupon dependent on a subset of the original factors, we will use an adequate neural network to select the best joint distribution of the coupon and factor values among an adequate set of probability distributions. In both the GLM and the neural network approaches to model selection, we present a hierarchy of the models studied by means of adequate quantities used as errors, thus providing two examples of the application of the model improvement algorithm detailed in Section 2.

We stress that the yield evolution modelling of the bond is a totally different problem since, by reason of a possible irregular variation of the bond yield with time, it is more of a dynamic type problem: a problem for which a stochastic process type model would be more adequate.

We use the real data of 42 countries. The list of countries considered is in Table 1. We first detail the characteristics of the data used. Next we detail the results of the fitting of the Generalised Linear Model (GLM) to the data giving the interest rate at the issue date as a function of the factors.

Table 1. A selection of countries including G20 countries.

Argentina	Australia	Austria	Belgium	Brazil	Bulgaria
Canada	China	Cyprus	Czech Republic	Denmark	Estonia
Finland	France	Germany	Greece	Hungary	India
Indonesia	Ireland	Italy	Japan	Latvia	Lithuania
Luxembourg	Malta	Mexico	Netherlands	Poland	Portugal
Romania	Russia	Saudi Arabia	Slovakia	Slovenia	South Africa
South Korea	Spain	Sweden	Turkey	United Kingdom	United States

Several studies point to the influence of socio-economic and financial factors on the determination by the market of the bond yields and concomitantly of the coupon—the nominal interest rate—(see [15–18]). In the following we look at a wider set of 11 socio-economic variables that may contribute, or not, to a lower or to a higher country risk, leading to lower or higher coupon interest rates. We list below how each variable may hypothetically affect the country risk and, thus, coupon interest rates.

x_1 Population: It is a factor in the category of Demographics, so it is a social indicator. Given two countries with comparable levels of development, it is to be expected that the country with a significantly larger population will have more capacity to absorb economic shocks and so will present less risk.

x_2 Population Growth: It is a factor in the category of Demographics, so it is a social indicator. Moderate growth can imply less risk since, in principle conjugated with economic growth, it can induce a greater fiscal revenue. But also, very high or low population growth may increase economic instability and subsequently increase the default risk, requiring a higher yield and coupon interest rate.

x_3 GDP Per Capita: It is a factor in the category of Macroeconomic Health. Higher GDP per capita implies a stronger economy, which in turn implies a lower risk of default and, consequently, a lower required yield and coupon interest rate.

x_4 GDP Real Growth: It is a factor in the category of Macroeconomic Health. A higher GDP growth implies better fiscal capacity, that is, further possibilities of tax collection and so lower default risks on bonds.

x_5 Government Debt: It may also be classified as an indicator of Macroeconomic Health at least when considered in the form of the ratio Government Debt to GDP or normalised by population. In this form, a higher ratio implies a higher risk, which, in turn, implies a higher yield and coupon interest rate required by the potential investors.

x_6 Inflation Rate: It is a factor that describes Macroeconomic Health. Higher inflation implies a higher coupon rate needed to compensate investors and also implies a higher default risk since it may be a sign of a stressed economy and failure of central bank control measures. Nevertheless, if the bond is issued in a more robust currency, the risk of a higher yield due to inflation may be reduced.

x_7 Current Account Balance: It may be considered a Financial factor. Deficits are imbalances that may imply weakness of the productive system and may suggest the need for extra reliance on national savings or foreign capital, increasing the default risk.

x_8 Foreign Exchange Reserves: It may be considered a Financial factor. More reserves imply a stronger capacity to repay and so, also a lower default risk.

x_9 External Debt: It is an indicator of Macroeconomic Health at least when considered in the form of the ratio to GDP or normalised by the population. A higher ratio to GDP has a significant default risk for current yields above a certain threshold.

x_{10} Oil Production: It is an Economic factor. In principle, a country with significant oil production will benefit from greater energy self-sufficiency and, in principle also, from a better fiscal strength.

x_{11} Natural Gas Production: It is an Economic factor. In principle, a country with significant natural gas production will benefit from better energy self-sufficiency and, in principle also, from a better fiscal strength.

The characteristics of the factors just presented may, possibly, be further discussed following the practical application results analysis in Remark 2. For each of the countries considered, we choose to look at the socio-economic and financial factors in Table 2. The values were obtained using Wolfram's Mathematica 14.3 ™. We observe that another site, possibly with different information, is the International Monetary Fund (https://data.imf.org/?sk=4BE0C9CB-272A-4667-8892-34B582B21BA6, accessed on 20 November 2025). The

values for these factors—with correspondent variable names presented in Table 2—were statistically standardised in the usual way in order to have values of comparable orders of magnitude.

Table 2. Socio-economic and financial factors with their identifier x_i, $i = 1, \ldots, 11$ and the dependent variable Y.

x_1-Population	x_2-Population Growth	x_3-GDP Per Capita	x_4-GDP Real Growth
x_5-Government Debt	x_6-Inflation Rate	x_7-Current Account Balance	x_8-Foreign Exchange Reserves
x_9-External Debt	x_{10}-Oil Production	x_{11}-Natural Gas Production	y-Coupon Interest Rate

The most reliable data for the 10 year coupon interest rate—in 2021—was extracted from the following site: World government bonds (https://www.worldgovernmentbonds.com/, accessed on 20 November 2025). Other sites that allow for comparing 10 year coupon interest rates are the following: Central Bank Rates (https://www.cbrates.com/, accessed on 20 November 2025), OECD (https://www.oecd.org/en/data/indicators/long-term-interest-rates.html, accessed on 20 November 2025), and Trading Economics (https://tradingeconomics.com/country-list/interest-rate?continent=g20, accessed on 20 November 2025).

5. A Generalised Linear Model Fitting

We fitted GLMs to the data—after an usual statistical standardisation—three times; for the first fitting, we considered all the variables $x_1, \ldots, x_{11}$; for the second fitting, we removed the least statistically significant variables x_{10} and x_{11}; for the third fitting—that produced the final model—we removed again the least statistically significant variables x_5 and x_7. A summary of the results of the third and final fitting is provided in Table 3.

Table 3. Results of the final—and third—GLM fitting.

	Estimate	**Standard Error**	**z-Statistic**	***p*-Value**
1	36.0964	3.40475	10.6018	$2.92411(1)\cdot 10^{-26}$
x_1 — Population	7.90189	1.0599(0)	7.45534	$8.9637(2)\cdot 10^{-14}$
x_2 — Population Growth	−2.14603	0.9748(5)	−2.2014(0)	0.0277(1)
x_3 — GDP Per Capita	33.5276(0)	4.29512	7.80597	$5.9044(2)\cdot 10^{-15}$
x_4 — GDP Real Growth	−6.30668	0.6233(1)	−10.118(00)	$4.5972(3)\cdot 10^{-24}$
x_6 — Inflation Rate	1.35317	0.1815(2)	7.45484	$8.9978(3)\cdot 10^{-14}$
x_8 — Foreign Exchange Reserves	−14.3309(0)	2.14247	−6.68897	$2.2474(4)\cdot 10^{-11}$
x_9 — External Debt	−8.89317	2.64468	−3.36266	0.0007(7)

The fitting was conducted with quasi-likelihood estimation and reciprocal link function, constraining the response to the unit interval $]0, 1[$. Variable selection was performed via backward elimination, retaining only predictors with statistical significance at $\alpha = 0.05$, except for x_8, which was retained due to borderline significance and theoretical relevance. Parameter estimation was performed using Fisher scoring/iteratively re-weighted least squares, as implemented in Mathematica's *GeneralizedLinearModelFit* function. The reciprocal link $g(\mu) = 1/\mu$ selected was the one that provided the best fitting to the data, with us knowing that this inverse relationship is appropriate when the expected response is inversely proportional to linear predictors.

The equation for the third and final model giving the coupon interest rate random variable Y as a function of the random variables factors X_i, $i = 1, 2, 3, 4, 6, 8, 9$ is given in Formula (8).

$$Y = \frac{1}{36.0964 + 7.90189X_1 - 2.14603X_2 + 33.5276X_3 - 6.30668X_4 + 1.35317X_6 - 14.3309X_8 - 8.89317X_9}. \quad (8)$$

Remark 2 (An analysis of the results of the GLM in Formula (8) and Table 3)**.** *In the final model obtained, the factors x_5-Government Debt, x_7-Current Account Balance, x_{10}-Oil Production, and x_{11}-Natural Gas Production are not present since these factors were not found to be statistically significant in the sequence of models tested.*

According to the model—see Formula (8)—we have that an increase in the modulus of a coefficient with a negative coefficient estimate—see also Table 3—induces an increase in the coupon interest rate Y; consequently, we have the following coefficients interpretation with a reciprocal link.

1. *Positive coefficients $\beta_1, \beta_3, \beta_6$—respectively, Population, GDP per capita and Inflation rate—indicate that as these predictors increase, the reciprocal of the response decreases, meaning the response itself increases; this result seems contradictory in what concerns the GDP per capita but this is what the model applied to the data shows. We do not propose any explanation to this fact.*
2. *Negative coefficients $\beta_2, \beta_4, \beta_8, \beta_9$—respectively, Population growth, GDP real growth, Foreign exchange reserves and External debt—indicate that as these predictors increase, the reciprocal of the response increases, meaning the response decreases; this result seems contradictory to what concerns the GDP real growth but this is what the model applied to the data shows. We do not propose any explanation to this fact.*
3. *Let us detail an example. A one-unit increase in X_1—that is Population—is associated with a reciprocal change of $+7.90189$ in the linear predictor, which translates to a decrease of $1/Y$ or equivalently an increase of Y.*

Also due to the functional form of the model in Formula (8), the stronger effects in the coupon, caused by the factors, are verified for x_1-GDP Per Capita and x_8-Foreign Exchange Reserves. The weaker effects are from the factors x_2-Population Growth and x_6-Inflation Rate; in this last case, it may help to explain the previous observation about this factor. These practical application findings only partially agree with the risk-inducing characteristics of the factors discussed just below Table 1.

5.1. Assessing the Quality of the Three Model Fittings

In order to assess the model quality and the goodness-of-fit of the three models, we have the usual indicators presented in Table 4.

Table 4. Quantitative common goodness-of-fit indicators for the three GLMs.

Models	AIC	BIC	Null Deviance	Residual Deviance	Log Likelihood	Pearson Chi Square
First Model	54l.	74.852	$1.13916 \cdot 10^{-7}$	0.0070(1)	−15.	0.0070(1)
Second Model	52.	69.3767	$1.18021 \cdot 10^{-7}$	0.0077(5)	−16.	0.0077(5)
Final Model	50.	63.9014	$1.35452 \cdot 10^{-7}$	0.0094(5)	−17.	0.0094(5)

Remark 3 (Comments on the values of Table 4)**.** *The most important criteria for model comparison are the AIC and the BIC. A lower AIC suggests a better trade-off between fit and parsimony and so there is an improvement from the first to the final model. A lower BIC indicates a better model for Bayesian model selection and so there is, also, an improvement from the first to the final third model. The response variable—interest rates—exhibited limited empirical variation across the sample period (range: 1.36% to 28.10%; interquartile range: 2.6% to 4.4%). Consequently, goodness-of-fit statistics such as deviance and Pearson χ^2 assume numerically small values. Model selection was therefore based primarily on information criteria (AIC, BIC) and predictor significance rather than absolute fit measures.*

5.2. Assessing Model Errors

In order to justify the presentation of this application in this work, we now study the error evolution across the three successive fits. We will consider two methods of assessing

the errors. The first method relies on the comparison of the errors defined for each model by the differences between the response of the model and the predicted responses for that model. The fitting for the errors gave three normal distributions with parameters presented in Table 5.

Table 5. Distributions for the model errors and associated random variables X_1, X_2 and $X_3 = X_f$.

First Model	Second Model	Final Model
$X_1 \frown \mathcal{N}(0.0034(8), 0.0137(7)^2)$	$X_2 \frown \mathcal{N}(0.0043(6), 0.014(2)^2)$	$X_f \frown \mathcal{N}(0.0044(0), 0.0163(1)^2)$

Remark 4 (Comments on the probability distributions of Table 5 and alternative comparison criterium)**.** *It is clear that the means and standard deviations are increasing from the first fitted model to the third and final model. Due to the results of Section 3, it is important to compare for each of the random variables X_1, X_2 and X_f some quantities related to the survival functions $G_1(u) := \mathbb{P}[X_1 > u]$, $G_2(u) := \mathbb{P}[X_2 > u]$ and $G_f(u) := \mathbb{P}[X_f > u]$ for a set of values of $u \in \mathbb{R}$. We observe that there is a well-known stochastic order $\leq_{st}$ between random variables—named the usual stochastic order (see [19] (p. 3))—defined by:*

$$X \leq_{st} Y \Leftrightarrow \forall u \in \mathbb{R}\,,\ \mathbb{P}[X > u] \leq \mathbb{P}[Y > u]\,.$$

This order is not a total order and in the present case, the survival functions $G_1(u)m$ $G_2(u)$ and $G_f(u)$ are not comparable for this stochastic order. Nevertheless, it so happens that for lowest values that the variables X_1, X_2 and X_f can take—say in an interval $[u_n, u_M]$—we can compare the survival functions in such a way that:

$$\forall u \in [u_n, u_M]\ ,\ G_f(u) < \min(G_1(u), G_2(u))\ . \tag{9}$$

Figure 2 shows that Formula (9) is verified, for instance, for $[-0.040, -0.015]$, an approximate interval which corresponds to probabilities in the approximate interval $[0.85, 1]$. This ordering penalizes variability and conveys the idea that the final model presents less variability. Let us observe that, in Figure 2 the result for the survival functions has a counterpart in the quantile functions.

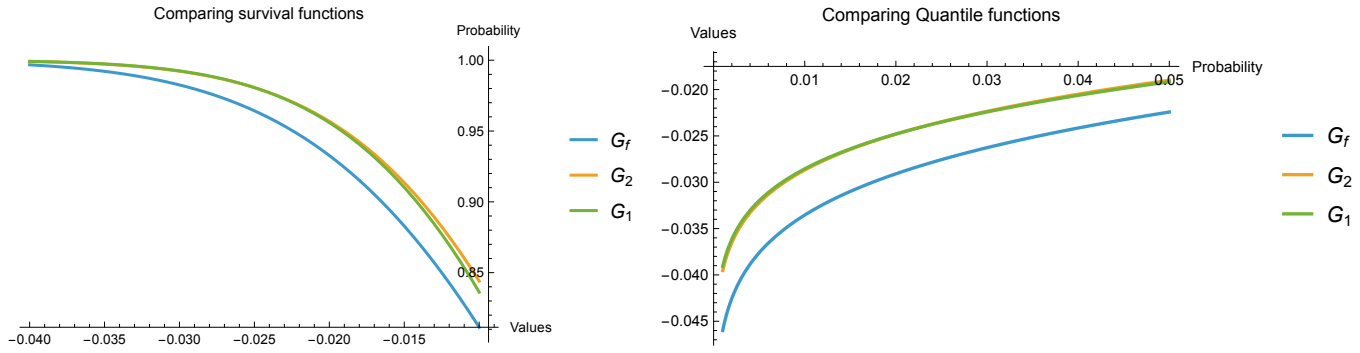


Figure 2. Comparing survival functions $G_1(u)$, $G_2(u)$ and $G_f(u)$ (**left**) and correspondent quantile functions (**right**) for the three models.

The second method consists of computing the RMS—that is, the root mean square—of the standard errors given in Table 3 for the final model. We recall that if the standard errors are given by σ_i, for $i = 0, 1, \ldots, N$ then:

$$RMS = \sqrt{\frac{1}{N+1}\sum_{i=0}^{N}\sigma_i^2}\ . \tag{10}$$

For the first and second models we considered all the standard errors in the respective fittings, that is, for 12 variables for the first model and for 10 variables for the second variable. The results of the RMS for the three models are in Table 6.

Table 6. Root mean square for the three models.

First Model	Second Model	Final Model
$\text{RMS}_1 = 6.91843$	$\text{RMS}_2 = 3.35731$	$\text{RMS}_f = 2.34841$

Remark 5 (Comments on the values of RMS_1, RMS_2, and RMS_f in Table 6)**.** *It is clear that the sequence of values of the RMS of the standard errors decreases from the first to the final model. This observation together with the evolution of the survival functions, discussed in Remark 4 and presented is Figure 2, is an indication of model improvement. However, as a caveat for our methodological illustration, we track RMS reduction across nested models: RMS decreased from 6.92 (12 variables) to 2.35 (8 variables). While RMS quantifies estimation precision gains, we stress that variable elimination can introduce bias when correlated predictors are removed, as evidenced by the* 43% *change in the external debt coefficient when government debt is omitted. In practice, and whenever one is willing to extract sound economic conclusions from real data application, RMS should therefore complement, not replace, theoretical justification and predictive validation.*

5.3. GLM as an Example of a Model Improvement Algorithm

A standard error of a coefficient is the estimated standard deviation of the sampling distribution of the coefficient; it is a statistic obtained from the data. The computed RMS, that is, the square root of the average squared standard errors, may be considered as a descriptive measure of the average uncertainty across the coefficient estimates.

We now show that the GLM falls under the scope of the results in Section 3. In a GLM fitting of a quantity Y—such as an interest rate in the application developed above—depending on a vector of factors $\boldsymbol{F} = (F_1, \ldots, F_N)$ and on a vector of parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_N)$ by means of a function f, we have that,

$$Y = f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}) + \boldsymbol{e}\,, \tag{11}$$

with $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha_0}, \widehat{\alpha_1}, \ldots, \widehat{\alpha_N})$ the vector of estimates of $\boldsymbol{\alpha}$ and the vector $\boldsymbol{e} = (e_0, e_1, \ldots, e_N)$ being the vector of estimation errors. Both $\widehat{\boldsymbol{\alpha}}$ and $\boldsymbol{e}$ are statistics.

We have the following result that shows that the GLM in Formula (11) can be expressed in the form studied in Section 3, that is, essentially, Formula (2).

Theorem 3 (Error propagation in GLM estimation)**.** *Consider the GLM $Y = f(\boldsymbol{\alpha}; \boldsymbol{F})$ with estimates $\widehat{\boldsymbol{\alpha}}$ satisfying:*

(a) The denominator in Formula (8) is bounded away from zero.
(b) The estimation error $\boldsymbol{e} = \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}$ satisfies $\mathbb{E}[\boldsymbol{e}] = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma}$ with $\Sigma_{ii} = \sigma_i^2 = \text{Var}(e_i)$.
(c) There exists $M > 0$ such that $\mathbb{P}(\|\boldsymbol{e}\| > t) \leq \exp(-t^2/(2M^2))$ for all $t > 0$.
(d) $f \in C^2(\mathbb{R}^{N+1} \times \mathbb{R}^N)$ with bounded second derivatives on compact sets.
(e) $\boldsymbol{F}$ takes values in a compact set.

Then, for RMS defined in (10), we have

$$Y = f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}) + E + R\,,$$

where we have the following properties:

1. $|R| \leq K \cdot RMS^2$ *almost surely for some* $K > 0$
2. $\mathbb{E}[E] = 0$,

3. $\mathbb{P}(|E| > u) \leq \dfrac{C(N+1)}{u^2} \cdot RMS^2$ *for all* $u > 0$, *with RMS denoting the realised value from the GLM fit and* $C = \sup \|\nabla f\|^2$ *over the compact set.*

Proof. We provide a step-by-step proof combining Taylor expansion, probabilistic bounds, and the compactness assumptions.

1. On the Taylor expansion with remainder. By hypothesis (d), $f \in C^2$. Applying Taylor's theorem with Lagrange remainder to the function $g(t) = f(\widehat{\boldsymbol{\alpha}} + t\boldsymbol{e}; \boldsymbol{F})$ at $t = 1$ with expansion point $t = 0$, there exists $\theta \in [0, 1]$ such that,

$$Y = f(\widehat{\boldsymbol{\alpha}} + \boldsymbol{e}; \boldsymbol{F}) = f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}) + \nabla f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}) \cdot \boldsymbol{e} + \frac{1}{2}\boldsymbol{e}^\top H_f(\widehat{\boldsymbol{\alpha}} + \theta\boldsymbol{e}; \boldsymbol{F})\boldsymbol{e}\,, \tag{12}$$

where,

- $\nabla f(\boldsymbol{\alpha}; \boldsymbol{F}) = \left(\frac{\partial f}{\partial \alpha_0}, \ldots, \frac{\partial f}{\partial \alpha_N}\right)^\top$ is the gradient;
- $H_f(\boldsymbol{\alpha}; \boldsymbol{F}) = \left(\frac{\partial^2 f}{\partial \alpha_i \partial \alpha_j}\right)_{i,j=0}^{N}$ is the Hessian matrix.

Define,

$$E := \nabla f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}) \cdot \boldsymbol{e},$$
$$R := \frac{1}{2}\boldsymbol{e}^\top H_f(\widehat{\boldsymbol{\alpha}} + \theta\boldsymbol{e}; \boldsymbol{F})\boldsymbol{e}\,.$$

Thus, $Y = f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}) + E + R$.

2. On the properties of the linear term E. From hypothesis (b), $\mathbb{E}[\boldsymbol{e}] = \boldsymbol{0}$. Since $\widehat{\boldsymbol{\alpha}}$ and $\boldsymbol{e}$ are dependent in general, we need to condition. However, note that,

$$\mathbb{E}[E] = \mathbb{E}[\nabla f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}) \cdot \boldsymbol{e}]\,.$$

By the law of total expectation,

$$\mathbb{E}[E] = \mathbb{E}[\mathbb{E}[\nabla f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}) \cdot \boldsymbol{e} \mid \widehat{\boldsymbol{\alpha}}, \boldsymbol{F}]] = \mathbb{E}[\nabla f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}) \cdot \mathbb{E}[\boldsymbol{e} \mid \widehat{\boldsymbol{\alpha}}, \boldsymbol{F}]]\,.$$

Under standard GLM asymptotics, $\boldsymbol{e}$ is asymptotically independent of $\widehat{\boldsymbol{\alpha}}$ (or at least $\mathbb{E}[\boldsymbol{e} \mid \widehat{\boldsymbol{\alpha}}] \to 0$). Assuming either exact independence or that the conditional expectation vanishes, we have $\mathbb{E}[E] = 0$, establishing property (2).

3. Bounding the remainder R. By hypotheses (d) and (e), the second derivatives of f are bounded on compact sets. Specifically, there exists $L > 0$ such that for all $\boldsymbol{\alpha}$ in the compact set containing $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\alpha}} + \theta\boldsymbol{e}$, and all $\boldsymbol{F}$ in its compact set

$$\|H_f(\boldsymbol{\alpha}; \boldsymbol{F})\|_{\text{op}} \leq L\,,$$

where $\|\cdot\|_{\text{op}}$ denotes the operator norm, that is, the largest singular value.

Thus, for the quadratic form we have,

$$|R| = \frac{1}{2}\left|\boldsymbol{e}^\top H_f(\widehat{\boldsymbol{\alpha}} + \theta\boldsymbol{e}; \boldsymbol{F})\boldsymbol{e}\right| \leq \frac{1}{2}\|H_f(\widehat{\boldsymbol{\alpha}} + \theta\boldsymbol{e}; \boldsymbol{F})\|_{\text{op}} \cdot \|\boldsymbol{e}\|^2 \leq \frac{L}{2}\|\boldsymbol{e}\|^2.$$

Now, by definition of RMS,

$$\|\boldsymbol{e}\|^2 = \sum_{i=0}^{N} e_i^2 = (N+1) \cdot \frac{1}{N+1}\sum_{i=0}^{N} e_i^2\,.$$

However, we need to relate this to $\text{RMS}^2 = \frac{1}{N+1}\sum_{i=0}^{N}\sigma_i^2$. From hypothesis (b), $\mathbb{E}[e_i^2] = \sigma_i^2$. While e_i^2 is random and not equal to σ_i^2, we can use the sub-Gaussian assumption (c) to bound it. Alternatively, for an almost sure bound, note that by the compactness in (e) and continuity of the estimation procedure, $\|\boldsymbol{e}\|$ is bounded. Let $B > 0$ be such that $\|\boldsymbol{e}\| \leq B$ almost surely. Then,

$$|R| \leq \frac{L}{2}B^2.$$

Let us get a bound in terms of RMS. Since σ_i^2 are the theoretical variances, and in practice we use their estimates, we need a different approach. Let us state a key observation: in the GLM context, the standard errors σ_i are typically of order $O(n^{-1/2})$, where n is the sample size. For the theorem statement, we can use the following: by the definition of RMS and the fact that σ_i^2 are consistent estimators of $\text{Var}(e_i)$, for large enough samples, we have with high probability,

$$\|\boldsymbol{e}\|^2 \leq C' \cdot (N+1) \cdot \text{RMS}^2 \,,$$

for some $C' > 0$. This follows from Markov's inequality,

$$\mathbb{P}\Big(\|\boldsymbol{e}\|^2 > t \cdot (N+1)\text{RMS}^2\Big) \leq \frac{\mathbb{E}[\|\boldsymbol{e}\|^2]}{t(N+1)\text{RMS}^2} = \frac{1}{t}\,,$$

since $\mathbb{E}[\|\boldsymbol{e}\|^2] = \sum_{i=0}^{N}\sigma_i^2 = (N+1)\text{RMS}^2$. Thus, taking $t = 1/\delta$, with probability at least $1-\delta$,

$$\|\boldsymbol{e}\|^2 \leq \frac{(N+1)\text{RMS}^2}{\delta}\,.$$

Observe that for an almost sure bound in property (1) we need a stronger assumption. If we assume $\|\boldsymbol{e}\|$ is almost surely bounded by some multiple of RMS, then $|R| \leq K \cdot \text{RMS}^2$ almost surely for some $K > 0$.

4. The probability bound for E (property 3). We now prove that $\mathbb{P}(|E| > u)$. By Cauchy–Schwarz,

$$|E| = |\nabla f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}) \cdot \boldsymbol{e}| \leq \|\nabla f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F})\| \cdot \|\boldsymbol{e}\|\,.$$

Let $C := \sup \|\nabla f(\boldsymbol{\alpha}; \boldsymbol{F})\|^2$, where the supremum is taken over the compact sets containing $\widehat{\boldsymbol{\alpha}}$ and $\boldsymbol{F}$ (hypothesis e). Such C exists and is finite by continuity (from d) and compactness. Then:

$$|E|^2 \leq C \cdot \|\boldsymbol{e}\|^2 = C \cdot \sum_{i=0}^{N} e_i^2\,.$$

Now apply Chebyshev's inequality:

$$\mathbb{P}(|E| > u) = \mathbb{P}(|E|^2 > u^2) \leq \frac{\mathbb{E}[|E|^2]}{u^2}\,.$$

Notice that we have,

$$\begin{aligned}
\mathbb{E}[|E|^2] &\leq \mathbb{E}\left[C \cdot \sum_{i=0}^{N} e_i^2\right] \\
&= C \cdot \sum_{i=0}^{N} \mathbb{E}[e_i^2] \quad \text{(by linearity of expectation)} \\
&= C \cdot \sum_{i=0}^{N} \sigma_i^2 \quad \text{(by hypothesis b)} \\
&= C \cdot (N+1) \cdot \text{RMS}^2\,.
\end{aligned}$$

We stress that here RMS^2 is the realized value computed from the estimated variances σ_i^2, not a random variable. In practice, we plug in the estimated RMS from the GLM output. Therefore

$$\mathbb{P}(|E| > u) \leq \frac{C(N+1) \cdot \text{RMS}^2}{u^2},$$

which establishes property (3).

5. The completion of the proof. We have shown the following:

1. From Step 3: $|R| \leq \frac{L}{2}\|e\|^2$. Under the additional assumption that $\|e\|^2 \leq K'(N+1)\text{RMS}^2$ almost surely (which holds if the parameter space is compact and f is Lipschitz), we get $|R| \leq K \cdot \text{RMS}^2$ with $K = \frac{L}{2}K'(N+1)$.
2. From Step 2: $\mathbb{E}[E] = 0$.
3. From Step 4: $\mathbb{P}(|E| > u) \leq \frac{C(N+1)}{u^2} \cdot \text{RMS}^2$.

And this completes the proof of the theorem. □

Remark 6 (On the meaning and consequences of hypothesis **(a)** in Theorem 3)**.** *Since Y in Formula (8) is the interest rate, the hypothesis is natural. To establish minimax prediction error bounds for a Generalised Linear Model (GLM) with an inverse link function $\mu = (a^T X)^{-1}$, we need an analysis of the Fisher Information Matrix (FIM). For the model $Y = (a^T X)^{-1}$, the FIM is*

$$I(a) = \mathbb{E}\left[\frac{(a^T X)^{-4}}{V((a^T X)^{-1})} XX^T\right].$$

The bounded away from zero condition ($|a^T X| \geq \delta > 0$) ensures that the FIM is non-degenerate. Aiming at full rigour, a tight analysis of the minimum eigenvalue $\lambda_{min}(I(a))$ provides the lower bound on the variance of any estimator, which, via Fano's Inequality, yields a minimax prediction error bound.

Next we show that with a sequence of GLM fittings, we have an instance of an algorithm for model improvement. Without loss of generality, we will consider that the number of factors N is constant. In fact, as seen in the application developed above, the number of variables $x_1, \ldots, x_M$ considered to build factors M is definite and bounded. From this finite set of variables it is possible to build an infinite number of factors by considering functions of subsets of variables $F_{i,j,\ldots,k} := g_{i,j,\ldots,k}(x_i, x_j, \ldots, x_k)$ and functions of those functions and so on and so forth. Nevertheless, in practice and in a top-down or backwards approach, we will start with a certain number of factors and, at each step, we will reduce the number of factors in the model; we observe that in this case, we can consider the initial number of factors as a constant by assuming that, if a certain factor is no longer considered in a further model, then both the corresponding coefficient and standard error will be put equal to zero in the further model. We observe that as long as we keep control over the model choice, some quantitative indicator that penalizes a large number of factors—such and the AIC or the BIC—even in the case of a down-up or forward approach, we will always end up with a maximal and finite number of factors.

Theorem 4 (Weak convergence of model errors joint laws–III–the GLM case)**.** *Consider a sequence of GLMs indexed by $n \in \mathbb{N}$, each yielding a decomposition,*

$$Y = f_n(\widehat{\boldsymbol{\alpha}}_n; \boldsymbol{F}) + E_n + R_n \,,$$

where for each n, the pair (E_n, R_n) satisfies the conclusions of Theorem 3 with associated constants K_n, C_n, and root mean square error RMS_n.

Assume the following conditions hold:

(A1) (Function Convergence) The sequences $(f_n)_{n\geq 1}$ *and* $(\nabla f_n)_{n\geq 1}$ *converge* sequentially *to limits* f *and* ∇f*, respectively. That is, for any sequence* (x_n) *in* $\mathbb{R}^{N+1}$ *with* $x_n \to x$*, we have,*

$$\lim_{n\to\infty} f_n(x_n) = f(x) \quad \textit{and} \quad \lim_{n\to\infty} \nabla f_n(x_n) = \nabla f(x)\,.$$

(A2) (Estimate Convergence) The sequence of estimated parameter vectors $(\widehat{\boldsymbol{\alpha}}_n)_{n\geq 1}$ *converges almost surely to a limit vector* $\widehat{\boldsymbol{\alpha}}$*, and there exists a constant* $M_1 > 0$ *such that* $\|\widehat{\boldsymbol{\alpha}}_n\| \leq M_1$ *almost surely for all* n*.*

(A3) (Precision Convergence) The sequence of root mean square errors converges to zero in probability: $RMS_n \xrightarrow{P} 0$ *as* $n \to \infty$*. Furthermore, the constants from Theorem 3 are uniformly bounded:*

$$\sup_n K_n < \infty \quad \textit{and} \quad \sup_n C_n < \infty\,.$$

(A4) (Compact Support) The limit estimate $\widehat{\boldsymbol{\alpha}}$ *and the factor vector* $\boldsymbol{F}$ *take values in compact sets* $\mathcal{K}_{\boldsymbol{\alpha}} \subset \mathbb{R}^{N+1}$ *and* $\mathcal{K}_{\boldsymbol{F}} \subset \mathbb{R}^N$*, respectively.*

Let ν_n *denote the joint law (probability distribution) of the pair* $\big(f_n(\widehat{\boldsymbol{\alpha}}_n; \boldsymbol{F}), E_n\big)$*. Then*

$$\nu_n \xrightarrow{d} \nu_\infty\,,$$

where ν_∞ *is the law of* $\big(f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}), 0\big)$*. That is, the sequence* (ν_n) *converges weakly to the distribution with the first marginal given by the law of* $f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F})$ *and second marginal as the point mass, that is, a Dirac measure, at zero.*

Proof. The proof proceeds in four main steps: establishing the convergence of the model term, showing the error terms vanish in probability, proving tightness of the sequence of joint laws, and finally identifying the unique weak limit.

1. On the convergence of the model term $f_n(\widehat{\boldsymbol{\alpha}}_n; \boldsymbol{F})$. From assumption (A2), we have $\widehat{\boldsymbol{\alpha}}_n \xrightarrow{\text{a.s.}} \widehat{\boldsymbol{\alpha}}$. For almost every elementary outcome ω in the sample space, the sequence $\widehat{\boldsymbol{\alpha}}_n(\omega)$ converges to $\widehat{\boldsymbol{\alpha}}(\omega)$. By the sequential convergence condition (A1), it follows that for almost every ω,

$$\lim_{n\to\infty} f_n(\widehat{\boldsymbol{\alpha}}_n(\omega); \boldsymbol{F}(\omega)) = f(\widehat{\boldsymbol{\alpha}}(\omega); \boldsymbol{F}(\omega))\,.$$

Hence, we have almost sure convergence, which implies convergence in probability,

$$f_n(\widehat{\boldsymbol{\alpha}}_n; \boldsymbol{F}) \xrightarrow{\text{P}} f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F})\,. \tag{13}$$

Since convergence in probability implies convergence in distribution, we also have:

$$f_n(\widehat{\boldsymbol{\alpha}}_n; \boldsymbol{F}) \xrightarrow{\text{d}} f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F})\,. \tag{14}$$

2. On the vanishing of the error terms R_n and E_n. Recall from Theorem 3 that the remainder term satisfies the almost sure bound $|R_n| \leq K_n \cdot \text{RMS}_n^2$. By assumption (A3), $\text{RMS}_n \xrightarrow{\text{P}} 0$ and $\sup_n K_n < \infty$. A standard argument using Markov's inequality shows that for any $\epsilon > 0$,

$$\mathbb{P}(|R_n| > \epsilon) \leq \frac{\mathbb{E}[|R_n|]}{\epsilon} \leq \frac{K_n \mathbb{E}[\text{RMS}_n^2]}{\epsilon}\,.$$

Given $\text{RMS}_n \xrightarrow{\text{P}} 0$ and the uniform bound on K_n, it follows that $\mathbb{P}(|R_n| > \epsilon) \to 0$. Therefore,

$$R_n \xrightarrow{\text{P}} 0\,. \tag{15}$$

For the linear error term E_n, Theorem 3 provides the Chebyshev-type bound,

$$\mathbb{P}(|E_n| > u) \leq \frac{C_n(N+1) \cdot \mathrm{RMS}_n^2}{u^2} \quad \text{for all } u > 0 \,.$$

By assumption (A3), $\mathrm{RMS}_n \xrightarrow{\mathrm{P}} 0$ and $\sup_n C_n < \infty$. Fixing any $\epsilon > 0$ and setting $u = \epsilon$ in the bound above yields,

$$\mathbb{P}(|E_n| > \epsilon) \leq \frac{C_n(N+1) \cdot \mathrm{RMS}_n^2}{\epsilon^2} \to 0 \quad \text{as } n \to \infty \,.$$

Thus, we have convergence in probability to zero:

$$E_n \xrightarrow{\mathrm{P}} 0 \,. \tag{16}$$

3. On the tightness of the sequence of joint laws (ν_n). A sequence of probability measures is tight if for every $\epsilon > 0$, there exists a compact set K_ϵ such that $\nu_n(K_\epsilon) > 1 - \epsilon$ for all n.

- From (14), the sequence of first marginals $\{f_n(\widehat{\boldsymbol{\alpha}}_n; \boldsymbol{F})\}_n$ converges in distribution. A well-known result in probability theory states that a sequence converging in distribution is tight.
- The sequence $\{E_n\}_n$ converges in probability to the constant 0 by (16). A sequence converging to a constant is also tight (the mass concentrates around that constant).

Let $\epsilon > 0$ be given. By the tightness of the first marginal, there exists a compact set $K_1 \subset \mathbb{R}$ such that $\mathbb{P}(f_n(\widehat{\boldsymbol{\alpha}}_n; \boldsymbol{F}) \in K_1) > 1 - \epsilon/2$ for all n. By the tightness of $\{E_n\}$, there exists a compact interval $K_2 = [-\delta, \delta]$ (for some $\delta > 0$) such that $\mathbb{P}(E_n \in K_2) > 1 - \epsilon/2$ for all n. The Cartesian product $K_\epsilon = K_1 \times K_2$ is a compact subset of $\mathbb{R}^2$. By the union bound,

$$\nu_n(K_\epsilon^c) = \mathbb{P}\big(\big(f_n(\widehat{\boldsymbol{\alpha}}_n; \boldsymbol{F}), E_n\big) \notin K_1 \times K_2\big) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \,.$$

Hence, $\nu_n(K_\epsilon) > 1 - \epsilon$ for all n, proving the sequence (ν_n) is tight.

4. On the identification of the weak limit. By Prokhorov's theorem, because (ν_n) is tight, every subsequence (ν_{n_k}) has a further subsequence $(\nu_{n_{k_\ell}})$ that converges weakly to some probability measure ν_* on $\mathbb{R}^2$. We will now show that any such subsequential limit ν_* must be the product measure $\nu_\infty = \mathcal{L}_{f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F})} \otimes \delta_0$, where δ_0 is the Dirac measure at zero.

Let $(\nu_{n_{k_\ell}})$ be a weakly convergent subsequence with limit ν_*. Consider an arbitrary bounded continuous function $g : \mathbb{R}^2 \to \mathbb{R}$. We analyse its expectation.

$$\int_{\mathbb{R}^2} g(x,y)\, d\nu_{n_{k_\ell}}(x,y) = \mathbb{E}\Big[g\big(f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}), E_{n_{k_\ell}}\big)\Big] = \mathbb{E}\Big[g\big(f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}), 0\big)\Big] + \Delta_\ell \,,$$

where the error term Δ_ℓ is defined as,

$$\Delta_\ell = \mathbb{E}\Big[g\big(f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}), E_{n_{k_\ell}}\big) - g\big(f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}), 0\big)\Big] \,.$$

- For the first term, note that the function $h(x) = g(x, 0)$ is bounded and continuous. Since $f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}) \xrightarrow{\mathrm{d}} f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F})$ by (14) (and this property holds for any subsequence), we have by the definition of weak convergence,

$$\lim_{\ell \to \infty} \mathbb{E}\Big[g\big(f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}), 0\big)\Big] = \mathbb{E}\big[g\big(f(\widehat{\boldsymbol{\alpha}}; \boldsymbol{F}), 0\big)\big] = \int_{\mathbb{R}^2} g(x,y)\, d\nu_\infty(x,y) \,.$$

- We now show $\Delta_\ell \to 0$. Since g is continuous, it is uniformly continuous on any compact set in $\mathbb{R}^2$. Let $\epsilon > 0$. There exists $\eta > 0$ such that $|g(x, y_1) - g(x, y_2)| < \epsilon$ whenever $|y_1 - y_2| < \eta$. Now,

$$\begin{aligned}
|\Delta_\ell| &\leq \mathbb{E}\Big[\Big|g\big(f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}), E_{n_{k_\ell}}\big) - g\big(f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}), 0\big)\Big|\Big] \\
&= \mathbb{E}\left[\Big|g\big(f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}), E_{n_{k_\ell}}\big) - g\big(f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}), 0\big)\Big| \cdot \mathbf{1}_{\left\{\left|E_{n_{k_\ell}}\right| < \eta\right\}}\right] \\
&\quad + \mathbb{E}\left[\Big|g\big(f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}), E_{n_{k_\ell}}\big) - g\big(f_{n_{k_\ell}}(\widehat{\boldsymbol{\alpha}}_{n_{k_\ell}}; \boldsymbol{F}), 0\big)\Big| \cdot \mathbf{1}_{\left\{\left|E_{n_{k_\ell}}\right| \geq \eta\right\}}\right] \\
&\leq \epsilon \cdot \mathbb{P}\Big(\Big|E_{n_{k_\ell}}\Big| < \eta\Big) + 2\|g\|_\infty \cdot \mathbb{P}\Big(\Big|E_{n_{k_\ell}}\Big| \geq \eta\Big),
\end{aligned}$$

  where $\|g\|_\infty = \sup_{(x,y)\in\mathbb{R}^2} |g(x, y)| < \infty$. By (16), $\mathbb{P}\Big(\Big|E_{n_{k_\ell}}\Big| \geq \eta\Big) \to 0$ as $\ell \to \infty$. Therefore,

$$\limsup_{\ell\to\infty} |\Delta_\ell| \leq \epsilon \cdot 1 + 2\|g\|_\infty \cdot 0 = \epsilon.$$

  Since $\epsilon > 0$ was arbitrary, we conclude $\lim_{\ell\to\infty} \Delta_\ell = 0$.

Combining these two results, we have for our arbitrary bounded continuous g and any convergent subsequence

$$\lim_{\ell\to\infty} \int_{\mathbb{R}^2} g \, d\nu_{n_{k_\ell}} = \int_{\mathbb{R}^2} g \, d\nu_\infty.$$

This means every weakly convergent subsequence of (ν_n) has the same limit ν_∞. By a standard corollary of Prokhorov's theorem, if a tight sequence of measures has the property that every convergent subsequence converges to the same limit, then the entire sequence converges to that limit. Therefore, $\nu_n \xrightarrow{\mathrm{d}} \nu_\infty$, which completes the proof. □

Conclusion 1 (Improvement of GLM algorithm application)**.** *Theorem 4 was not applied to obtain the final GLM fitting above; as usual, we were guided by the principle of eliminating the less statistically significant factors. The usefulness of this theorem in the GLM context is that it shows that as long as the RMS of successive models decreases and the model functions f_n keep some coherence with one another—such as being of the same functional form family—we are sure to guarantee model improvement.*

Theorem 4 provides a precise mathematical justification for the core idea in this work. Whenever we have a sequence of improving GLMs—the f_n, f functions with $f_n \to f$—meaning that their parameter estimates become more precise as $RMS_n \to 0$, that is $\widehat{\boldsymbol{\alpha}}_n \to \widehat{\boldsymbol{\alpha}}$, we then have that the fitted model stabilises to the limit model f, the associated linearised prediction error E_n vanishes, and the joint behaviour of the model and its error term converges to a scenario where all uncertainty is captured by the limit model itself, f, with no residual random error. This is exactly the model improvement that the archetypal algorithm seeks.

6. Feeding Models Simulated Data to a Neural Network

In this section, we present a second instance of the archetypal algorithm for model improvement. This is a paradigmatic example since model improvement, in this case, refers to a hierarchy of different statistical models for the data presented in the beginning of Section 4. Namely, we consider four statistical models, the normal multivariate distribution, the Student t multivariate distribution, a multivariate kernel distribution, and the empirical distribution associated with the data. Let us stress that, in the data considered in this section, we only used the factors that integrate the final—and best—GLM determined in Section 5.1.

The hierarchy of the four statistical models is established with some quantitative error measures provided by the training of a fixed neural network fed by Monte Carlo samples of the models. The important point to be noticed is that the different statistical models for the data are compared by means of their error performances, which are computed from the neural network trainings with samples from each of the four models. Again, this application of the archetypal algorithm for model improvement is a proof of concept for the archetypal algorithm and there was no attempt to optimize the neural network used as a classifying tool of the statistical models of the data, beyond the reasonable heuristic initial considerations.

The motivation for the use of the neural network is the following. The dataset is moderate in size—seven factors for 42 countries—and so it is expectable that training a neural network with the data considered will not allow us to recover all the real world different possibilities for the production of an interest rate as a function of the different factors. By considering a large sample—size 30,000—of the different statistic models fitted to the data, it is to be expected that a much larger set of possibilities, relating the interest rate to the factors, will occur. Using a fixed NN and Monte Carlo simulated data for each statistical model, to compute an error associated with each model, a model will be considered an improvement if the neural network training, with its Monte Carlo-generated sample, gives rise to a smaller error. Our interpretation of this smaller error is that, under the perspective of the neural network, fewer extreme and non-real possibilities will arise with that given model.

The procedure implemented is as follows.

(i) We consider a neural network, described below in detail, that will be trained firstly by the data and subsequently by samples of the same dimension—specifically, with dimension 30,000—generated by Monte Carlo simulation from four probability distributions previously fitted to the raw data; the probability distributions used are the normal multivariate distribution, the Student t multivariate distribution, a multivariate kernel distribution, and the empirical distribution associated to the data.

(ii) The quality of the training of the neural network for each random sample of the raw data generated by one of the four fitted statistical distributions is assessed by three numerical measures, denominated training progress measurements: the mean deviation of the residuals, the mean square deviation of the residuals, and the RSquare.

(iii) The three numerical measures are computed at each round of the verification of the trained neural network and the evolution of these values for each of the measures is presented in Figure 3, showing the stability of these measures along the whole set of rounds; being so, the numerical results corresponding to the last round are presented in Table 7.

Let us first describe in more detail the neural network that was implemented with Wolfram's Mathematica™ 14.3 running on a platform Mac OS X ARM (64 bits). The neural network has three hidden linear layers with the activation function tanh, the first layer with 25 neurons, the second layer with 20 neurons, and the third layer with 15 neurons. There is an output layer with one neuron to produce the interest rate. See Figure 4 for two schematic representations of the neural network described; notice that the upper representation gives the data flows through the layers and the lower representation also represents the loss net that allows for the determination of the training progress measurements. As usual, the sample dataset was split into training (70%), validation ([70%, 85%]), and test sets ([85%, 100%]).

The choice of this configuration was guided by heuristic considerations and by trial and error of different configurations trained with the raw data. Let us stress that other choices of neural networks are possible.

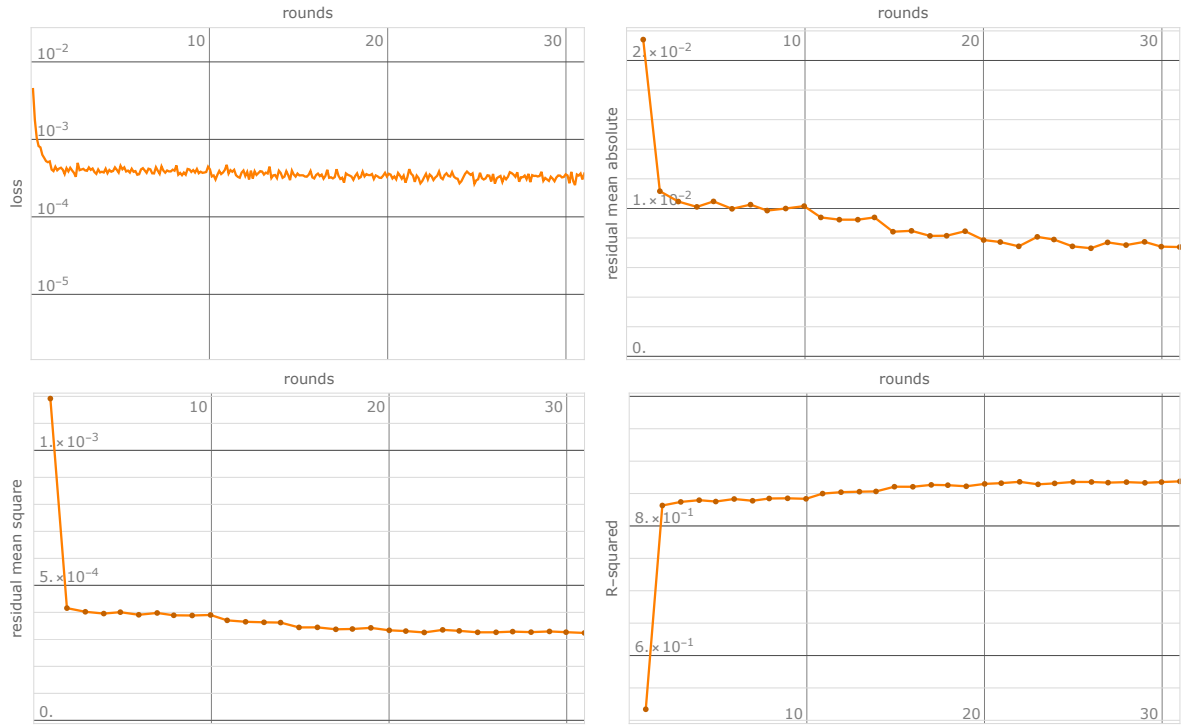


Figure 3. The evolution of the three training progress measurements, mean deviation, mean square deviation, and R Square, all along the whole set of rounds for the empirical distribution model.

The results obtained are given in Table 7.

Table 7. Results of data and different models feeding the neural network.

Data and Models	Mean Deviation	Mean Square Deviation	R Square	s-Wasserstein Distance
Raw Data	$5.13981 \cdot 10^{-3}$	$3.1427(1) \cdot 10^{-4}$	0.8741(5)	–
Empirical Distribution	$9.43129 \cdot 10^{-3}$	$3.90412 \cdot 10^{-4}$	0.84243	442.574
Student *t*	$2.50449 \cdot 10^{-2}$	$1.09434 \cdot 10^{-3}$	0.5613(9)	588.997
Multinormal Distribution	$3.1773 \cdot 10^{-2}$	$1.59688 \cdot 10^{-3}$	0.3495(4)	618.168
Kernel fitting	$4.10254 \cdot 10^{-2}$	$2.88931 \cdot 10^{-3}$	0.23317	993.980

Remark 7 (Comments on the results presented in Table 7)**.** *According to all three quantitative measures, mean deviation, mean square deviation, and R Square, there is a clear improvement of models starting in the reverse order presented in the table. The worse model is the kernel fitting then the multivariate normal distribution, then the multivariate Student t distribution, and then the better model is the empirical distribution. As naturally expected, the model given by the empirical distribution is quite close—relatively to the quantitative measures referred—to the training of the neural network with the raw data. The column on the right contains the values of the sliced-Wasserstein distances between the four fitted distribution samples and the distribution induced by the raw data; it is to be interpreted as an indicator of weak convergence of the successive models towards the true distribution of the raw data (see Remark 8 and Appendix A).*

Each neural network was trained for 31 rounds on 30,000 Monte Carlo samples from the fitted distribution, with a batch size of 64, ensuring convergence and stable error metrics for model comparison. We stress that the neural network in this experiment functions as a consistent diagnostic instrument, not as an optimisable predictive model in its own right. Its architecture was fixed after initial heuristic calibration on the raw data to ensure a stable platform for comparing the generative quality of different statistical models. The R^2 values reported in Table 7 are not expected to reach 1, nor to surpass the value obtained from the raw data. Instead, they quantify how well

data simulated from each approximating model preserves the complex relationships learned by the network from the original dataset. The clear, consistent hierarchy of performance across three distinct error metrics—and its correlation with the independent sliced-Wasserstein distances—validates this approach for comparative model assessment.

The global behaviour of the three training progress measurements, mean deviation, mean square deviation, and R Square, in Table 7 is described in Figure 3. The figure in the bottom right hand side shows the early stabilisation of the R^2 behaviour all along the 30 rounds of the training of the neural network with a sample of dimension 30,000 of the empirical distribution model; for the other models, the figures are similar although with slightly different values.

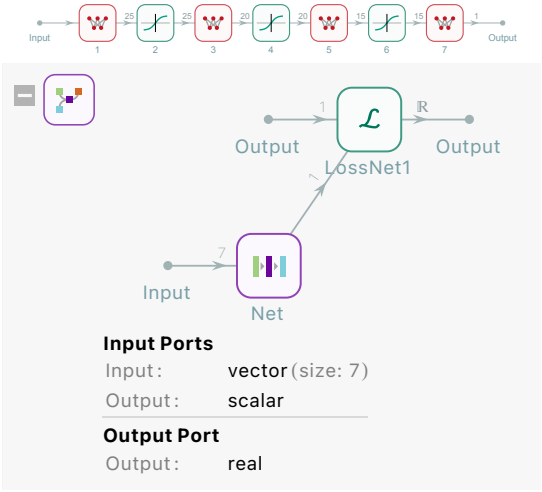


Figure 4. Schematic representations of the configuration of the neural network considered.

The Neural Network as a Tool for an Example of a Model Improvement Algorithm

We present next a result that justifies the claim that by considering the R^2 increasing to one, we achieve model improvement. We first present formal definitions of the quantities needed to formulate the result, that is, the coefficient of determination and a R^2-convergence condition.

Let (X, Y) denote a random vector in $\mathbb{R}^d \times \mathbb{R}$. Consider a sequence of models for the joint law of (X, Y) given by $\mathcal{L}^n_{(X,Y)}$, for $n \geq 1$. For each n, let $f_n : \mathbb{R}^d \to \mathbb{R}$ be the regression function obtained (for instance and in our case, by training a neural network) to minimise the expected squared prediction error under $\mathcal{L}^n_{(X,Y)}$, that is,

$$f_n = \arg\min_{f \in \mathcal{F}} \mathbb{E}^{\mathcal{L}^n_{(X,Y)}} \left[(Y - f(X))^2 \right],$$

where $\mathcal{F}$ denotes the admissible class of predictors.

Definition 2 (Coefficient of determination R^2)**.** *The coefficient of determination R^2 of Y with respect to X under $\mathcal{L}^n_{(X,Y)}$ is defined by:*

$$R^2_n = 1 - \frac{\mathbb{E}^{\mathcal{L}^n_{(X,Y)}} \left[(Y - f_n(X))^2 \right]}{\mathrm{Var}^{\mathcal{L}^n_{(X,Y)}}(Y)},$$

whenever $\mathrm{Var}^{\mathcal{L}^n_{(X,Y)}}(Y) > 0$.

Definition 3 (R^2-convergence condition)**.** *We say that the sequence of fitted distributions* $(\mathcal{L}^n_{(X,Y)})_{n\geq 1}$ *satisfies the* R^2*-convergence condition if,*

$$\lim_{n\to+\infty} R^2_n = 1 \quad \Leftrightarrow \quad \lim_{n\to+\infty} \frac{\mathbb{E}^{\mathcal{L}^n_{(X,Y)}}\left[(Y - f_n(X))^2\right]}{\mathrm{Var}^{\mathcal{L}^n_{(X,Y)}}(Y)} = 0\,.$$

We observe that when the variance $\mathrm{Var}^{\mathcal{L}^n_{(X,Y)}}(Y)$ remains bounded away from zero, this condition is equivalent to:

$$\lim_{n\to+\infty} \mathbb{E}^{\mathcal{L}^n_{(X,Y)}}\left[(Y - f_n(X))^2\right] = 0\,,$$

implying that the conditional law $\mathcal{L}^n_{(X,Y)|X=x}$ becomes increasingly concentrated around a deterministic limit $g(x)$.

We now formulate and prove the main result of this section, that is, the weak convergence of the input marginals—the convergence being denoted by $\mathcal{L}^n_{\boldsymbol{X}} \Rightarrow \mathcal{L}_X$—in turn, entails the weak convergence of the joint distributions $\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}} \Rightarrow \mathcal{L}_{(X,Y)} \equiv \mathcal{L}_{(X,g(X))}$, where $\mathcal{L}_{(X,g(X))}$ denotes the law of $(X, g(X))$.

Theorem 5 (Weak convergence of model errors joint laws when R^2 is increasing to one– IV)**.** *Let* $(\boldsymbol{X}, d_{\boldsymbol{X}})$ *be a Polish space—in the application it is* $\mathbb{R}^d$*—*$\mathbb{R}$ *denoting the real numbers with the usual distance. Let* $g : \boldsymbol{X} \to \mathbb{R}$ *be continuous and bounded. With* $X \frown \mathcal{L}_X$ *a random variable taking values in* $\boldsymbol{X}$ *with law* $\mathcal{L}_X$*, denote by* $\mathcal{L}_{(X,Y)} := \mathcal{L}_{(X,g(X))}$ *the law of* $(X, Y) := (X, g(X))$.

Let $(\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}})_{n\geq 1}$ *be a sequence of probability measures on* $\boldsymbol{X}\times\mathbb{R}$ *with marginals* $\mathcal{L}^n_{\boldsymbol{X}}$ *on* $\boldsymbol{X}$*. Suppose that the following three hypotheses hold:*

1. $\mathcal{L}^n_{\boldsymbol{X}} \Rightarrow \mathcal{L}_X$*, that is,* $(\mathcal{L}^n_{\boldsymbol{X}})_{n\geq 1}$ *converges weakly to* $\mathcal{L}_X$ *on* $\boldsymbol{X}$
2. *The coefficient of determination* R^2_n*—defined when predicting* Y *from* X *under* $\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}}$*—satisfies Definition 3. Alternatively, the mean square error of predicting* Y *from* X *under* $\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}}$ *converges to zero, that is*

$$\lim_{n\to+\infty} \mathbb{E}^{\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}}}[(Y - f_n(X))^2] = 0\,,$$

   where $f_n(x) = \mathbb{E}^{\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}}}[Y|X = x] = \mathbb{E}^{\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}|X=x}}[Y]$ *is the conditional mean under* $\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}}$*.*
3. *The sequence* $(f_n)_{n\geq 1}$ *converges to* g *in probability with respect to* $\mathcal{L}^n_{\boldsymbol{X}}$*, that is for every* $\delta > 0$

$$\lim_{n\to\infty} \mathcal{L}^n_{\boldsymbol{X}}(\{x : |f_n(x) - g(x)| \geq \delta\}) = 0\,.$$

Then, we have:

$$\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}} \Rightarrow \mathcal{L}_{(X,Y)} \equiv \mathcal{L}_{(X,g(X))}\,,$$

that is, the sequence $(\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}})_{n\geq 1}$ *converges weakly to* $\mathcal{L}_{(X,g(X))}$ *on* $\boldsymbol{X}\times\mathbb{R}$*.*

Proof. The proof strategy leverages the fact that under the R^2-convergence condition, the conditional distributions concentrate around their means, and the convergence of f_n to g ensures these means approach the deterministic limit.

Let us firstly express the joint integral in terms of marginals and conditionals. Let $\varphi \in C_b(\boldsymbol{X}\times\mathbb{R})$ be an arbitrary bounded continuous test function. Then by the law of iterated expectations

$$\int_{\boldsymbol{X}\times\mathbb{R}} \varphi(x,y)d\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}}(x,y) = \int_{\boldsymbol{X}} \underbrace{\int_{\mathbb{R}} \varphi(x,y)d\mathcal{L}^n_{\boldsymbol{X}\times\mathbb{R}|X=x}(y)}_{=:\widetilde{\varphi}_n(x)} d\mathcal{L}^n_{\boldsymbol{X}}(x)\,.$$

Our goal is to show that:

$$\lim_{n\to+\infty}\int_{\boldsymbol{X}}\widetilde{\varphi}_n(x)d\mathcal{L}^n_{\boldsymbol{X}}(x)=\int_{\boldsymbol{X}}\varphi(x,g(x))d\mathcal{L}_X(x)\ .$$

The outline of the proof may be described in two main steps and the conclusion. Firstly we will show that $\widetilde{\varphi}_n(x)\to\varphi(x,g(x))$ in probability with respect to $\mathcal{L}^n_{\boldsymbol{X}}$ and secondly we will use the weak convergence $\mathcal{L}^n_{\boldsymbol{X}}\Rightarrow\mathcal{L}_X$, the uniform boundedness of $\widetilde{\varphi}_n$, and the convergence in probability to conclude that the integrals converge.

1. On the convergence of $\widetilde{\varphi}_n$. Let $\varphi\in C_b(\boldsymbol{X}\times\mathbb{R})$ be fixed, and let $M=\|\varphi\|_\infty$. Consider some arbitrary $\varepsilon>0$. Since φ is uniformly continuous on compact sets and the sequence $\{\mathcal{L}^n_{\boldsymbol{X}}\}$ is tight (due to weak convergence), there exists $\delta>0$ such that for all $x\in\boldsymbol{X}$ and $y,y'\in\mathbb{R}$ with $|y-y'|<2\delta$, we have,

$$|\varphi(x,y)-\varphi(x,y')|<\varepsilon\ .$$

Now, define the sets:

$$\begin{aligned}A_n^{(1)}&=\{x:P(|Y-f_n(x)|\geq\delta\mid X=x)<\varepsilon\},\\ A_n^{(2)}&=\{x:|f_n(x)-g(x)|<\delta\},\\ A_n&=A_n^{(1)}\cap A_n^{(2)}\ .\end{aligned}$$

We will show that $\lim_{n\to+\infty}\mathcal{L}^n_{\boldsymbol{X}}(A_n^c)=0$. By Tchebyshev's inequality and the R^2-convergence condition,

$$0\leq\limsup_{n\to+\infty}\int_{\boldsymbol{X}}P(|Y-f_n(x)|\geq\delta\mid X=x)d\mathcal{L}^n_{\boldsymbol{X}}(x)\leq\limsup_{n\to+\infty}\frac{1}{\delta^2}\mathbb{E}^{\mathcal{L}^n}[(Y-f_n(X))^2]=0\ ,$$

so $\lim_{n\to+\infty}P(|Y-f_n(X)|\geq\delta\mid X=x)=0$ in $L^1(\mathcal{L}^n_{\boldsymbol{X}})$, hence in probability. Thus, $\lim_{n\to+\infty}\mathcal{L}^n_{\boldsymbol{X}}((A_n^{(1)})^c)=0$. By the additional assumption that $(f_n))_{n\geq1}$ converges to g in probability, we have $\lim_{n\to+\infty}\mathcal{L}^n_{\boldsymbol{X}}((A_n^{(2)})^c)=0$. Therefore, $\lim_{n\to+\infty}\mathcal{L}^n_{\boldsymbol{X}}(A_n^c)=0$. Now, for $x\in A_n$, we have $|f_n(x)-g(x)|<\delta$ and $P(|Y-f_n(x)|\geq\delta\mid X=x)<\varepsilon$. Then:

$$\begin{aligned}|\widetilde{\varphi}_n(x)-\varphi(x,g(x))|&\leq\int_{\mathbb{R}}|\varphi(x,y)-\varphi(x,g(x))|d\mathcal{L}^n_{Y|X=x}(y)\\ &=\int_{|y-f_n(x)|<\delta}|\varphi(x,y)-\varphi(x,g(x))|d\mathcal{L}^n_{Y|X=x}(y)\\ &\quad+\int_{|y-f_n(x)|\geq\delta}|\varphi(x,y)-\varphi(x,g(x))|d\mathcal{L}^n_{Y|X=x}(y)\ .\end{aligned}$$

On the set $\{|y-f_n(x)|<\delta\}$, we have,

$$|y-g(x)|\leq|y-f_n(x)|+|f_n(x)-g(x)|<2\delta\ ,$$

so, by uniform continuity, $|\varphi(x,y)-\varphi(x,g(x))|<\varepsilon$. Hence, the first integral is at most ε. The second integral is over a set of conditional probability less than ε, and the integrand is bounded by $2M$, so the second integral is at most $2M\varepsilon$. Thus, for $x\in A_n$,

$$|\widetilde{\varphi}_n(x)-\varphi(x,g(x))|\leq\varepsilon+2M\varepsilon=\varepsilon(1+2M)\ .$$

Since $\lim_{n\to+\infty}\mathcal{L}^n_{\boldsymbol{X}}(A_n)=1$, it follows that $\lim_{n\to+\infty}\widetilde{\varphi}_n(x)=\varphi(x,g(x))$ in probability with respect to $\mathcal{L}^n_{\boldsymbol{X}}$.

2. We now deal with the convergence of the integrals, observing that

1. $\mathcal{L}_{\boldsymbol{X}}^{n} \Rightarrow \mathcal{L}_{X}$;
2. $\widetilde{\varphi}_n(x) \to \varphi(x, g(x))$ in probability with respect to $\mathcal{L}_{\boldsymbol{X}}^{n}$;
3. $|\widetilde{\varphi}_n(x)| \leq M$ for all n and x.

As a consequence, by a standard result in weak convergence theory—that is, variant of the dominated convergence theorem for varying measures (see Theorem A1)—these conditions imply:

$$\lim_{n\to\infty} \int_{\boldsymbol{X}} \widetilde{\varphi}_n(x) d\mathcal{L}_{\boldsymbol{X}}^{n}(x) = \int_{\boldsymbol{X}} \varphi(x, g(x)) d\mathcal{L}_{X}(x) \, .$$

3. We are now able to conclude that since $\varphi \in C_b(\boldsymbol{X} \times \mathbb{R})$ was arbitrary, we have that $\mathcal{L}_{\boldsymbol{X}\times\mathbb{R}}^{n} \Rightarrow \mathcal{L}_{(X, g(X))}$, as announced. □

Remark 8 (Verifying the weak convergence of laws)**.** *In practice, for a finite but large sequence of models, it may be necessary to verify hypothesis 1 of Theorem 5; that is* $(\mathcal{L}_{\boldsymbol{X}}^{n})_{n\geq 1}$ *converges weakly to* $\mathcal{L}_X$ *on* $\boldsymbol{X}$*. A practical way to perform this verification is to use the sliced-Wassertein distance as shown in Appendix A. Let us stress that in our case with only four models, this verification is not strictly necessary but, nevertheless, was included in Table 7 as an output of the NN trainings and commented in Remark 7.*

7. Conclusions, Discussion, and Further Work

Under an archetypal algorithm for model improvement, we have presented several results of weak convergence of laws associated with either GLM or multivariate models, which were classified according to errors produced using neural networks; these results require hypotheses that are natural to be assumed in the contexts of each type of model.

With illustration purposes, we considered the models applied to economic, financial, and social data of country bond issuers with the purpose of determining the coupon interest determined by the issuer country prior to the bond auction in the primary market. The model improvement instance using GLM allowed us to select a subset of initial variables according to statistical significance to explain the dependent interest rate variable. Next, we used the data with the reduced set of variables—obtained from the GLM fitting—to fit multivariate distributions, specifically, multidimensional Gaussian distribution, multivariate Student t distribution, kernel distribution, and empirical distribution, and considered samples generated from these distributions. With these four samples, we trained a built on purpose neural network in order to obtain fitting metrics relevant to the model improvement algorithm such as the R^2. In both instances, it was possible to observe model improvement according to the decrease to zero of the associated chosen error.

The practical implementations presented may be considered as illustrations of both the results presented and of the archetypal algorithm for model improvement considered. The determination of an optimal algorithm in the context of model improvement, as considered in this work, remains a task for future work.

Author Contributions: Conceptualisation, M.L.E., N.P.K. and R.M.G.; methodology, M.L.E.; software, M.L.E.; validation, M.L.E., N.P.K. and R.M.G.; formal analysis, M.L.E., N.P.K. and R.M.G.; investigation, M.L.E., N.P.K. and R.M.G.; resources, M.L.E.; data curation, M.L.E.; writing—original draft preparation, M.L.E.; writing—review and editing, M.L.E., N.P.K. and R.M.G.; visualisation, M.L.E., N.P.K. and R.M.G.; supervision, M.L.E.; project administration, M.L.E.; funding acquisition, M.L.E. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded–for M.L.E. and N.P.K.—through FCT (the Fundação para a Ciência e a Tecnologia, I.P.) under the scope of projects UID/297/2025 and UID/PRR/297/2025

(Center for Mathematics and Applications—NOVA Math); for R.M.G. under the scope of project UID/06522/2025.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: AI-1: The authors acknowledge the assistance of OpenAI's ChatGPT (GPT-5 model) in the preparation of parts of this work. The system was used interactively to refine the mathematical formulation of the model improvement algorithm, to formalize convergence theorems and their proofs (in particular those relating the coefficient of determination R^2 and weak convergence of fitted distributions), and to produce concise definitions and notation consistent with the theoretical framework of weak convergence of probability measures. The tool also assisted in the conceptual clarification and exposition of the sliced-Wasserstein distance used as a numerical indicator of convergence of multivariate distribution laws. All mathematical ideas, structure, and interpretation were directed and validated by the authors. AI-2: The authors would like to acknowledge the significant contribution of DeepSeek AI, an artificial intelligence system, to the development of some of the theoretical results in this paper. DeepSeek AI acted as a collaborative research assistant, providing instrumental help in formulating Theorem 5, identifying the necessity of the convergence-in-probability assumption, and constructing a detailed proof strategy.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. The Sliced-Wassertein Distance

For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{R}^d$, the p-Wasserstein distance is defined as,

$$W_p(\mathbb{P},\mathbb{Q}) = \left(\inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^p d\Pi(x,y)\right)^{1/p},$$

where $\Pi(\mathbb{P},\mathbb{Q})$ is the set of couplings of $\mathbb{P}$ and $\mathbb{Q}$—that is, all joint distributions whose marginals are $\mathbb{P}$ and $\mathbb{Q}$ (see [20] (p. 93)). Intuitively, $W_p(\mathbb{P},\mathbb{Q})$ measures the minimal cost to transport the mass off $\mathbb{P}$ onto $\mathbb{Q}$.

For measures $\mathbb{P}_\theta$, $\mathbb{Q}_\theta$ on $\mathbb{R}$, with cumulative distribution functions $F_{\mathbb{P}_\theta}$ and $F_{\mathbb{P}_\theta}$, and corresponding quantile functions $F^{-1}_{\mathbb{P}_\theta}(t)$ and $F^{-1}_{\mathbb{P}_\theta}(t)$, we have,

$$W_p^p(\mathbb{P}_\theta,\mathbb{Q}_\theta) = \int_0^1 \left|F^{-1}_{\mathbb{P}_\theta}(t) - F^{-1}_{\mathbb{Q}_\theta}(t)\right|^p dt\,.$$

The sliced-Wasserstein distance $SW_p^p(\mathbb{P},\mathbb{Q})$ between probability measures $\mathbb{P}$ and $\mathbb{Q}$ (see [21]) is defined as the average of the one dimensional Wasserstein distances over random directions:

$$SW_p^p(\mathbb{P},\mathbb{Q}) = \int_{S^{d-1}} W_p^p(\mathbb{P}_\theta,\mathbb{Q}_\theta)d\sigma(\theta)\,,$$

where $d\sigma(\theta)$ denotes is the uniform measure on the unit sphere S^{d-1}. In practice, this integral is approximated by sampling a finite set of random directions $\{\theta_1,\ldots,\theta_M\}$ and computing,

$$\widehat{SW}_p^p(\mathbb{P},\mathbb{Q}) = \frac{1}{M}\sum_{m=1}^{M} W_p^p(\mathbb{P}_{\theta_m}),\mathbb{Q}_{\theta_m}))\,.$$

This yields a Monte Carlo estimate of the sliced-Wasserstein distance. The sliced-Wasserstein distance defines a true metric on the space of probability measures with finite p-moments, and in particular, weak convergence follows from sliced-Wasserstein convergence.

Appendix B. A Generalised Lebesgue Dominated Convergence Theorem

The following result—see [22,23] for other formulations—is a generalisation of Lebesgue dominated convergence theorem for the case of varying measures and it is used in the proof of Theorem 5. We present a simple proof for completeness.

Theorem A1 (Weak Convergence and Integral Convergence)**.** *Let (S, d) be a metric space, and let $\{\mu_n\}$ be a sequence of probability measures on S converging weakly to a probability measure μ. Let $\{h_n\}$ be a sequence of measurable functions on S verifying the following.*

1. *Uniform boundedness: There exists $M > 0$ such that:*

$$|h_n(x)| \leq M \quad \text{for all } n \in \mathbb{N} \text{ and } x \in S\,.$$

2. *Convergence in probability: For every $\epsilon > 0$,*

$$\lim_{n\to\infty} \mu_n(\{x \in S : |h_n(x) - h(x)| > \epsilon\}) = 0\,.$$

3. *Limit function properties: h is bounded and continuous.*

Then we have:

$$\lim_{n\to\infty} \int_S h_n d\mu_n = \int_S h d\mu\,.$$

Proof. Let an arbitrary $\epsilon > 0$ be given. We aim to show that for all sufficiently large n,

$$\left| \int_S h_n d\mu_n - \int_S h d\mu \right| < \epsilon\,.$$

We decompose the difference as follows:

$$\left| \int_S h_n d\mu_n - \int_S h d\mu \right| \leq \left| \int_S h_n d\mu_n - \int_S h d\mu_n \right| + \left| \int_S h d\mu_n - \int_S h d\mu \right|.$$

We start by estimating the second term using weak convergence. Since $\mu_n \Rightarrow \mu$ and h is bounded and continuous, by the definition of weak convergence we have that:

$$\lim_{n\to\infty} \int_S h d\mu_n = \int_S h d\mu.$$

Thus, there exists $N_1 \in \mathbb{N}$ such that for all $n \geq N_1$,

$$\left| \int_S h d\mu_n - \int_S h d\mu \right| < \frac{\epsilon}{2}\,.$$

We now estimate the left-hand side term with the hypothesis of convergence in probability. We now bound the first term:

$$\left| \int_S h_n d\mu_n - \int_S h d\mu_n \right| = \left| \int_S (h_n - h) d\mu_n \right| \leq \int_S |h_n - h| d\mu_n\,.$$

Let $\delta = \frac{\epsilon}{4}$ and define the set:

$$A_n(\delta) = \{x \in S : |h_n(x) - h(x)| > \delta\}\,.$$

By the convergence in probability hypothesis,

$$\lim_{n\to\infty} \mu_n(A_n(\delta)) = 0\,.$$

We now split the integral,

$$\int_S |h_n - h| d\mu_n = \int_{A_n(\delta)} |h_n - h| d\mu_n + \int_{S \setminus A_n(\delta)} |h_n - h| d\mu_n \, .$$

We observe that on $A_n(\delta)$, since $|h_n| \leq M$ and $|h| \leq M$—since h is bounded—we have $|h_n - h| \leq 2M$. Hence,

$$\int_{A_n(\delta)} |h_n - h| d\mu_n \leq 2M \cdot \mu_n(A_n(\delta)) \, .$$

Also on $S \setminus A_n(\delta)$, we have $|h_n - h| \leq \delta$ and so,

$$\int_{S \setminus A_n(\delta)} |h_n - h| d\mu_n \leq \delta \, .$$

Combining these upper estimates we get

$$\int_S |h_n - h| d\mu_n \leq 2M \cdot \mu_n(A_n(\delta)) + \delta \, .$$

Now, since $\mu_n(A_n(\delta)) \to 0$, there exists $N_2 \in \mathbb{N}$ such that for all $n \geq N_2$

$$\mu_n(A_n(\delta)) < \frac{\epsilon}{8M} \, .$$

Then, for $n \geq N_2$

$$\int_S |h_n - h| d\mu_n < 2M \cdot \frac{\epsilon}{8M} + \frac{\epsilon}{4} = \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2} \, .$$

As a consequence, the final estimate is as follows. Let $N = \max(N_1, N_2)$. Then, for all $n \geq N$,

$$\left| \int_S h_n d\mu_n - \int_S h d\mu \right| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \, .$$

Since $\epsilon > 0$ was arbitrary, the result follows,

$$\lim_{n \to \infty} \int_S h_n d\mu_n = \int_S h d\mu \, ,$$

as claimed. □

References

1. Marketing and Communication Department. *Model Risk Management; Quantitative and Qualitative Aspects*; Technical Report; Management Solutions : Madrid, Spain, 2014.
2. Wallace, T.; Raggl, A.; Tejada, M.; Agarwal, R. *Model Risk Management: Global Update. Latest Insights into the Evolution of Model Governance Practices Across North America, Europe and Asia*; Technical Report; McKinsey & Company: New York, NY, USA, 2019.
3. Kumar, P.; Laurent, M.P.; Rougeaux, C.; Tejada, M. *Model Risk Management 2.0 Evolves to Address Continued Uncertainty of Risk-Related Events*; Technical Report; McKinsey & Company: New York, NY, USA, 2022.
4. Gupta, R.; Sengupta, S.; Mukherjee, A.; Banerjee, R.; Akhilesh, T. *Model Risk Management: Key Considerations in Effective Management of Models*; Technical Report 012_THL1024_KP; KPMG Assurance and Consulting Services LLP., Lodha Excelus: Mumbai, India, 2024.
5. Buraschi, A.; Corielli, F. Risk management implications of time-inconsistency: Model updating and recalibration of no-arbitrage models. *J. Bank. Financ.* **2005**, *29*, 2883–2907. [CrossRef]
6. Braha, D.; Maimon, O. *A Mathematical Theory of Design: Foundations, Algorithms and Applications*, 1st ed.; Series Applied Optimization; Springer: New York, NY, USA, 1998; p. XXII+682. [CrossRef]
7. Han, Y.; Liu, C.; Gao, S. Computer Mathematical Modeling Based on Improved Genetic Algorithm. In *Proceedings of the 2nd International Conference on Cognitive Based Information Processing and Applications (CIPA 2022)*; Jansen, B.J., Zhou, Q., Ye, J., Eds.; Springer: Singapore, 2023; pp. 597–603.

8. Buriboev, A.S.; Sultanov, D.; Ibrohimova, Z.; Jeon, H.S. Mathematical Modeling and Recursive Algorithms for Constructing Complex Fractal Patterns. *Mathematics* **2025**, *13*, 646. [CrossRef]
9. Hannanu, M.I.; Camponogara, E.; Silva, T.L.; Hovd, M. A modified derivative-free SQP-filter trust-region method for uncertainty handling: Application in gas-lift optimization. *Optim. Eng.* **2025**, *26*, 401–429. [CrossRef]
10. Jelenčič, J.; Besher Massri, M.; Todorovski, L.; Grobelnik, M.; Mladenić, D. Improving stochastic models by smart denoising and latent representation optimization. *Inf. Sci.* **2025**, *692*, 121672. [CrossRef]
11. Shiryaev, A.N. *Probability. 1*, 3rd ed.; Graduate Texts in Mathematics; Springer: New York, NY, USA, 2016; Volume 95, p. xvii+486.
12. Billingsley, P. *Convergence of Probability Measures*, 2nd ed.; Wiley Series in Probability and Statistics: Probability and Statistics; John Wiley & Sons, Inc.: New York, NY, USA, 1999; p. x+277. [CrossRef]
13. Kallenberg, O. *Foundations of Modern Probability*, 3rd ed.; Probability Theory and Stochastic Modelling; Springer: Cham, Switzerland, 2021; Volume 99, p. xii+946. [CrossRef]
14. van der Vaart, A.W. *Asymptotic Statistics*; Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: Cambridge, UK, 1998; Volume 3. [CrossRef]
15. Afonso, A.; Arghyrou, M.G.; Bagdatoglou, G.; Kontonikas, A. On the time-varying relationship between EMU sovereign spreads and their determinants. *Econ. Model.* **2015**, *44*, 363–371. [CrossRef]
16. Afonso, A.; Arghyrou, M.G.; Bagdatoglou, G.; Kontonikas, A. *The Determinants of Sovereign Bond Yield Spreads in the EMU*; Working paper of the European Central Bank; European Central Bank: Frankfurt am Main, Germany, 2015.
17. Goel, R.; Malik, S. What Is Driving the Rise in Advanced Economy Bond Yields? *Glob. Financ. Stab. Notes* **2021**, *2021*, A001. [CrossRef]
18. Nose, M.; Menkulasi, J. Fiscal Determinants of Domestic Sovereign Bond Yields in Emerging Market and Developing Economies. *IMF Work. Pap.* **2025**, *2025*. [CrossRef]
19. Shaked, M.; Shanthikumar, J.G. *Stochastic Orders*; Springer Series in Statistics; Springer: New York, NY, USA, 2007; p. xvi+473. [CrossRef]
20. Villani, C. *Optimal Transport, Old and New*; Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]; Springer: Berlin/Heidelberg, Germany, 2009; Volume 338, p. xxii+973. [CrossRef]
21. Bonneel, N.; Rabin, J.; Peyré, G.; Pfister, H. Sliced and Radon Wasserstein Barycenters of Measures. *J. Math. Imaging Vis.* **2015**, *51*, 22–45. [CrossRef]
22. Serfozo, R. Convergence of Lebesgue integrals with varying measures. *Sankhyā Indian J. Stat. Ser. A* **1982**, *44*, 380–402.
23. Di Piazza, L.; Marraffa, V.; Musiał, K.; Sambucini, A.R. Convergence for varying measures. *J. Math. Anal. Appl.* **2023**, *518*, 126782. [CrossRef]