Selecting Models with Judgment

Simone Manganelli*

January, 2026

Abstract

A fundamental challenge in econometrics is how to use misspecified models for critical decisions. We propose a frequentist decision-theoretic approach that evaluates performance conditional on the realized sample, rather than in terms of unconditional risk. This shift resolves the inferential pathologies and non-uniform convergence problems highlighted in the post-model-selection literature. We further demonstrate that under misspecification, the objective statistical guarantee (i.e., the ability to bound the probability that a model-based decision underperforms a judgmental decision) becomes unobservable. This *uncertainty about the level of uncertainty* arises because probabilistic guarantees are themselves conditional on an unverifiable approximation to the data-generating process. The framework is applied to a vector autoregressive model for euro area inflation. Our results provide a formal decision-theoretic rationale for the continued role of judgment in policy-making, even after model-based analysis has been fully taken into account.

Keywords: Statistical Decision Theory; Model Selection Criteria; Problem of Induction; Inflation Forecasting.

JEL Codes: C1; C11; C12; C13.

*European Central Bank, simone.manganelli@ecb.int. The views expressed are mine and do not necessarily reflect those of the European Central Bank.

1 Introduction

Decision-makers routinely act by informally combining personal experience, private information, and subjective beliefs. We define these as *judgmental decisions*. This paper investigates how econometricians can help improve such decisions, moving beyond the ideal scenario of a perfectly specified econometric model and into the more realistic setting of model misspecification.

A formal frequentist framework for incorporating judgment, introduced by Manganelli (2025), involves adjusting a judgmental decision toward what the data suggest. Under correct specification, this method provides a valuable statistical guarantee: the probability that a model-based decision performs worse than the initial judgment is known and controlled by the chosen level of significance (the probability of a Type I error). In this idealized setting, the econometrician can improve upon expert judgment while maintaining a controlled risk profile.

This paper takes a different perspective on model selection. Rather than viewing model selection as a preliminary step for inference on model-specific parameters, it treats it as a choice among a finite set of data-dependent decisions. The object of interest is therefore not a random post-selection estimator, but a realized action whose performance is evaluated conditionally on the observed data. This shift in perspective is central to both the methodological and conceptual contributions of the paper.

We make two main contributions.

First, we demonstrate that the inferential pathologies emphasized in the post-model-selection literature (e.g. Leeb and Pötscher (2005)) are artifacts of evaluating random estimators under unconditional risk. When the object of interest is a realized decision and performance is evaluated conditionally on the observed data, as advocated in this paper, these pathologies disappear. Section 3 illustrates this point formally using the canonical example of Leeb and Pötscher (2005). This shift allows us to derive a consistent model selection criterion that asymptotically identifies the least misspecified decision within a

given class. Furthermore, we show that while model averaging provides no asymptotic benefit for nested models as decisions converge, it leads to strictly superior outcomes when candidate models are non-nested.

Second, the paper identifies and formalizes a deeper and irreducible source of uncertainty inherent in any statistical procedure under misspecification. We show that when probabilistic guarantees are constructed relative to a misspecified model, the true probability of error generally differs from the nominal significance level. This generates an *uncertainty about the level of uncertainty*, which cannot be eliminated by larger samples or richer model classes. Section 3 provides a precise mathematical characterization of this phenomenon, which we identify as the formal econometric counterpart to Hume's problem of induction.

Following the framework of Wald (1950), a statistical decision problem is defined by an unknown probability distribution, a decision rule, and a loss function. To this classical triad, Manganelli (2025) adds judgment, defined as a combination of a judgmental decision (the choice made absent a formal model) and a level of significance (reflecting the decision-maker's tolerance for statistical error). The resulting rule adjusts the judgmental decision toward the closest boundary of a forecast confidence interval. This acts as a shrinkage mechanism that pulls judgment toward the data-driven optimum. Under the ideal assumption of correct specification, this procedure offers a rigorous statistical guarantee: the probability that the model-based adjustment underperforms the initial judgment is bounded by the nominal significance level.

To address the reality of model misspecification, we extend this framework using the quasi-maximum likelihood (QML) theory of White (1996). While White provides a coherent inferential theory for misspecified models, casting the problem in decision-theoretic terms yields a new insight: under misspecification, the objective probability of a statistical decision underperforming judgment becomes unobservable. It is no longer governed solely by the decision-maker's significance level. This result is not a failure of the method, but a fundamental feature of statistical decision-making under misspecification.

We apply our framework to the problem of forecasting euro area inflation, a critical challenge for central banks. Using a judgmental forecasts ranging from 0% to 4.0% over the medium term, we estimate a vector autoregressive model with five endogenous variables (inflation, core inflation, unemployment, industrial production and interest rate) and 12 lags. Given the exceedingly large number of possible combinations within this class of models, a complete grid search for the model with the lowest in-sample loss is infeasible. We show how this problem can be formulated as an integer optimization program and solved using existing algorithms. Our findings indicate that the best forecasting model for euro area inflation includes all variables with complex non-contiguous lag structures. Specifically, the best inflation forecast at a two-year horizon depends on the judgmental decision used and lies between 2.5% and 3.5% as of May 2025. We also find that while a simple autoregressive model performs well in the short term, it lacks the dynamic complexity to accurately forecast inflation over longer, two- to three-year horizons, highlighting the importance of our approach.

The rest of the paper is structured as follows. Section 2 develops the formal theory of model selection within a decision-theoretic framework. Section 3 connects the paper to the post-model-selection literature, shows how conditional evaluation resolves standard inferential pathologies, and formally derives the irreducible uncertainty (the *Humean gap*) inherent in misspecified decision-making. Section 4 outlines the empirical implementation, including the integer optimization procedure. Section 5 presents the forecasting results for the euro area inflation, and Section 6 concludes by discussing the broader implications for policy-making.

2 Model Selection Criterion

Model selection requires first an estimate of the data-generation process under possible misspecification, second a loss function for the decision-maker and her judgment, third the construction of a decision rule incorporating the judgment, and fourth a criterion to assess

the degree of model misspecification. The structure of this section follows this order.

2.1 The Quasi-Maximum Likelihood Framework

Let $x_t \in \mathbb{R}^g$ denote the vector of observations at time t, and let $x^n = (x_1', \ldots, x_n')'$ represent the history of observations up to time n. We assume that x_t is a realization of a stochastic process with cumulative distribution function (cdf) G, so that $G(x_t^1, \ldots, x_t^g | x^{t-1}) = P(X_t^1 < x_t^1, \ldots, X_t^g < x_t^g | x^{t-1})$, $t = 1, 2, \ldots$

Following White (1996), define the quasi-log likelihood function as:

$$\ell_n(X^n, \theta) \equiv n^{-1} \sum_{t=1}^{n} \log f_t(X^t, \theta), \tag{1}$$

where $f_t(\cdot, \theta) : \mathbb{R}^{gt} \to \mathbb{R}^+$ is a suitably chosen function and $\theta \in \mathbb{R}^p$ is a finite-dimensional parameter vector. Consistent with White (1996) (e.g., page 13), $f_t(X^t, \theta)$ denotes a potential general function. While it can be interpreted as the conditional density $f_t(X_t | X^{t-1}, \theta)$ in a standard time series context, White's *quasi-likelihood* is more general, not requiring f_t to specify a full parametric probability model.

The quasi-maximum likelihood estimator (QMLE) is defined as the parameter vector that maximizes this function:

$$\hat{\theta}(X^n) = \arg\max_{\theta} \ell_n(X^n, \theta). \tag{2}$$

We use the notation $\hat{\theta}(X^n)$ to denote the estimator (a random variable) and $\hat{\theta}(x^n)$ for the estimate (a specific realization of the estimator for a given dataset).

We define the pseudo-true parameter vector θ^* as the unique maximizer of the quasi-loglikelihood function:

$$\theta^* = \arg\max_{\theta \in \Theta} E_G(\log f_t(X^t, \theta)).$$

For simplicity, we impose a stationarity assumption on the expectation of the log-likelihood function, specifically $E_G(\log f_t(X^t, \theta))$ does not depend on t. The subscript

G under the expectation indicates that the expectation is taken under the true cdf G. This assumption rules out heterogeneity in the expectations of the log-likelihood function, assuring the convergence of the QMLE to a fixed vector θ^* that does not depend on n. It can be relaxed following White (1996), who provides a more general treatment. We maintain this assumption because it simplifies the exposition and allows us to focus on the core aspects of our contribution.

Assumption 2.1 (**data-generating Process**)**.**

(a) X_t is strictly stationary and ergodic with unconditional distribution G.

(b) The process satisfies the regularity conditions in Theorems 3.5, 6.4 and Corollary 8.28 of White (1996) ensuring QMLE consistency, asymptotic normality, and variance consistency.

Theorem 2.1 (**QMLE Consistency and Asymptotic Normality**)**.** *Suppose Assumption 2.1 holds. Then:*

(i) The quasi-maximum likelihood estimator

$$\hat{\theta}(X^n) = \arg\max_{\theta \in \Theta} \ell_n(X^n, \theta)$$

is consistent for the pseudo-true parameter θ^*, i.e. $\hat{\theta}(X^n) \xrightarrow{p} \theta^*$ as $n \to \infty$.

(ii) Furthermore,

$$B^{*-1/2} D^* \sqrt{n}\,(\hat{\theta}(X^n) - \theta^*) \xrightarrow{d} N(0, I_p),$$

where $D^* \equiv E_G(\nabla^2 \ell_n(X^n, \theta^*))$ and $B^* \equiv Var_G\Big(\sqrt{n}\,\nabla \ell_n(X^n, \theta^*)\Big)$.

The operator ∇ denotes the gradient with respect to θ, i.e., $\nabla g(\theta) = (\partial g/\partial\theta_1, \ldots, \partial g/\partial\theta_p)'$, and ∇^2 denotes the Hessian matrix. The term I_p represents the identity matrix of dimension p. The asymptotic covariance matrix of $\hat{\theta}(X^n)$, given by $D^{*-1}B^*D^{*-1}$, can be estimated

by $\hat{D}_n^{-1}\hat{B}_n\hat{D}_n^{-1}$, where:

$$\hat{D}_n \equiv n^{-1}\sum_{t=1}^{n}\nabla^2 \log f_t(X^t, \hat{\theta}(X^n)), \tag{3}$$

$$\hat{B}_n \equiv n^{-1}\sum_{t=1}^{n}\nabla \log f_t(X^t, \hat{\theta}(X^n))\nabla' \log f_t(X^t, \hat{\theta}(X^n)). \tag{4}$$

2.2 Loss Function and Judgment

Consider a decision-maker whose objective is to minimize an expected loss. Let us formally define it as follows.

Definition 2.1 (**Loss Function**)**.** Fix a forecast horizon $h \geq 1$. Define the vector of future observations

$$X_{t,h} \equiv (X'_{t+1}, \ldots, X'_{t+h})' \in \mathbb{R}^{gh},$$

and let $\kappa : \mathbb{R}^{gh} \to \mathbb{R}^q$ be a mapping from future outcomes into the action space. For an action $a_t \in \mathbb{R}^q$, the **loss function** is defined by

$$L_t^0(a_t) \;\equiv\; E_G\Big[\mathscr{L}(\kappa(X_{t,h}), a_t)\,\Big|\,X^t\Big], \tag{5}$$

where $\mathscr{L} : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ is a measurable function.

Following the terminology in Mas-Colell et al. (1995), $\mathscr{L}(\kappa(X_{t,h}), \cdot)$ can be interpreted as the negative of a Bernoulli utility function, and, consequently, $L_t^0(a_t)$ is the negative of the von-Neumann–Morgenstern expected utility function. This $L_t^0(a_t)$ represents the true expected loss, based on the unknown true data-generating process.

We impose the following standard regularity conditions, to make the problem analytically tractable.

Assumption 2.2 (**Loss Function**)**.** The function $\mathscr{L}(\cdot, \cdot)$ satisfies the following conditions:

(a) The mapping $\kappa : \mathbb{R}^{gh} \to \mathbb{R}^q$ is measurable.

(b) For every $y \in \mathbb{R}^q$, the function $\mathscr{L}(y, a)$ is convex and continuously differentiable on $\mathbb{R}^q$.

(c) For all $a \in \mathbb{R}^q$, $E_G\Big|\mathscr{L}(\kappa(X_{t,h}), a)\Big| < \infty$.

Given the loss function, the optimal action can be defined as:

$$a_t^0 = \arg\min_{a_t} L_t^0(a_t). \tag{6}$$

Since the true data-generating process (represented by its cdf G) is unknown, the decision-maker must instead work with an approximation. They can minimize the expected loss, for the selected parametric specification $f_t(X^t, \theta)$. This gives:

$$\begin{aligned} L_t(\theta^*, a_t) &\equiv E_{\theta^*}(\mathscr{L}(\kappa(X_{t,h}), a_t)|X^t) \\ &= \int \mathscr{L}(y, a_t) f_{t,h}(y, \theta^*) dy, \end{aligned} \tag{7}$$

where $f_{t,h}$ is a suitable h-step ahead conditional density of $\kappa(X_{t,h})$, derived from the chosen $f_t(X^t, \theta^*)$ and evaluated at θ^*, the fixed vector to which the QMLE $\hat{\theta}(X^n)$ converges asymptotically (as defined in the previous subsection). The subscript θ^* under the expectation indicates that the expectation is taken with respect to the probability law implied by the maintained parametric model, $f_t(X^t, \theta^*)$.

This distinction between $L_t^0(a_t)$ and $L_t(\theta^*, a_t)$ highlights that, under misspecification, the decision-maker evaluates actions using a surrogate model, not the true DGP. This is an important distinction that plays a crucial role in the discussion of Section 3.

Following White (1996), we formalize correct model specification in this context:

Definition 2.2 (**Correct Specification**)**.** A parametric specification $f_t(X^t, \theta)$ is **correctly specified** for the decision-maker's purpose if there exists a parameter θ^0 such that $L_t^0(a_t) = L_t(\theta^0, a_t)$ for all a_t and $t = 1, 2, \ldots$ It is **misspecified** otherwise.

We next introduce the concept of judgment in this framework. Following Manganelli (2025), we adopt the following definition.

Definition 2.3 (**Judgment**)**. Judgment** is defined by the pair $A \equiv \{\tilde{a}_n, \alpha\}$, where $\tilde{a}_n \in \mathbb{R}^q$ is referred to as **judgmental decision** and $\alpha \in [0,1]$ is the **level of significance**.

The judgmental decision $\tilde{a}_n$ could represent any action based on domain expertise, historical precedents, or some other non-statistical reasoning. The level of significance α indicates the maximum probability that, by integrating this judgment into the statistical decision rule, the chosen action will lead to a higher loss compared to simply using the judgmental decision $\tilde{a}_n$.

2.3 The Decision Incorporating Judgment

Given the decision-maker's judgment $A = \{\tilde{a}_n, \alpha\}$, the key question is whether the judgmental decision $\tilde{a}_n$ is statistically optimal, and, if not, how it can be improved. This can be addressed by testing the null hypothesis that the judgmental decision $\tilde{a}_n$ sets the gradient of the loss function to zero. Non-rejection of the null hypothesis suggests that the available statistical evidence is insufficient to warrant any deviation from $\tilde{a}_n$. Conversely, rejection at level α justifies moving towards the QMLE-implied action. The probability of an unwarranted deviation (Type I error) is therefore bounded by α.

Here are the main building blocks needed to construct the decision with judgment: 1. The QMLE-decision, 2. The shrinkage decision rule and 3. The test function.

Definition 2.4 (**QMLE-implied and Shrinkage Decisions**)**.** For the observed sample x^n, the **QMLE-implied decision** is

$$\hat{a}_n(x^n) \equiv \arg\min_{a \in \mathbb{R}^q} L_t(\hat{\theta}(x^n), a),$$

where $\hat{\theta}(x^n)$ is the quasi-maximum likelihood estimate.

For $\lambda \in [0,1]$, the **shrinkage decision rule** is the convex combination

$$a_n(\lambda; x^n) \equiv \lambda \hat{a}_n(x^n) + (1-\lambda)\tilde{a}_n. \tag{8}$$

Clearly, $\tilde{a}_n = a_n(0; x^n)$ and $\hat{a}_n(x^n) = a_n(1; x^n)$. It is crucial to note that, conditional on the observed sample realization x^n, both $\hat{a}_n(x^n)$ and $\tilde{a}_n$ are fixed (non-random) values. The importance of conditioning is thoroughly discussed in Section 3.2.

Given the continuity, differentiability, and convexity of $\mathscr{L}$, a necessary and sufficient condition for the optimality of the judgmental decision $\tilde{a}_n$ with respect to the quasi-maximum likelihood loss function at the end of the sample, $L_n(\theta^*, \tilde{a}_n)$ (from Equation (7)) is that the gradient of the loss with respect to λ, evaluated at $\lambda = 0$, is zero :

$$\begin{aligned} H_0 : \nabla_\lambda L_n(\theta^*, a_n(0; x^n)) &= E_{\theta^*}(\nabla_\lambda \mathscr{L}(\kappa(X_{n,h}), a_n(0; x^n))|X^n) \\ &= 0. \end{aligned} \tag{9}$$

To construct a test statistic for this hypothesis, we replace θ^* with its QMLE estimator. To maintain clarity between random variables and observed realizations, we introduce the auxiliary random vector Y^n:

Definition 2.5 (**Auxiliary Random Vector**)**.** Let Y^n denote an **auxiliary random vector** that is i.i.d. and distributed as X^n.

This analytical device is necessary to study the asymptotic sampling distribution of our test statistics, which depends on the decision rule $a_n(\lambda; x^n)$ that is held fixed by the observed sample x^n, and on the estimator $\hat{\theta}(Y^n)$ that is random as it depends on the auxiliary random variable Y^n.

To obtain a test statistic, consider first a mean value expansion:

$$\nabla_\lambda L_n(\hat{\theta}(Y^n), a_n(\lambda; x^n)) = \nabla'_{\lambda,\theta} L_n(\bar{\theta}(Y^n), a_n(\lambda; x^n))(\hat{\theta}(Y^n) - \theta^*), \tag{10}$$

where $\bar{\theta}(Y^n)$ lies between $\hat{\theta}(Y^n)$ and θ^*. Exploiting the asymptotic results of Section 2.1 and following Le Cam (1986), for fixed, large n, the test based on the unknown distribution

is replaced by an approximating test based on the normal distribution:

$$\sigma^{-1}\sqrt{n}\nabla_\lambda L_n(\hat{\theta}(Y^n), a_n(\lambda; x^n)) \sim N(0,1). \tag{11}$$

An estimate of the asymptotic variance σ^2 can be obtained by replacing the random variable Y^n with the sample realization x^n, because Y^n and X^n are assumed to be i.i.d.:

$$\hat{\sigma}^2 \equiv \nabla'_{\lambda,\theta} L_n(\hat{\theta}(x^n), a_n(\lambda; x^n))\hat{D}_n^{-1}\hat{B}_n\hat{D}_n^{-1}\nabla_{\lambda,\theta} L_n(\hat{\theta}(x^n), a_n(\lambda; x^n)) \tag{12}$$

and $\hat{D}_n$ and $\hat{B}_n$ have been defined in equations (3)-(4).

Finally, define the following test function:

Definition 2.6 (**Test Function**)**.** For critical value c_α of the standard normal distribution, the **test function** is

$$\psi^A(y^n, \lambda; x^n) = \begin{cases} 0 & \text{if } z > c_{\alpha/2} \\ \gamma & \text{if } z = c_{\alpha/2} \qquad 0 < \gamma < 1, \\ 1 & \text{if } z < c_{\alpha/2} \end{cases} \tag{13}$$

where $z \equiv \hat{\sigma}^{-1}\sqrt{n}\,\nabla_\lambda L_n(\hat{\theta}(y^n), a_n(\lambda; x^n))$ and $\hat{\sigma}$ is defined in (12).

As explained in Manganelli (2025) (see footnote 4), only the lower tail is relevant, since the sign of z is known ex ante.

The decision incorporating judgment can be obtained by applying Theorem 3.2 of Manganelli (2025), leading to the following corollary:

Corollary 2.2 (**Decision with Judgment**)**.** *If Assumptions 2.1 and 2.2 hold, a decision-maker with judgment* $A \equiv \{\tilde{a}_n, \alpha\}$ *and with the QMLE loss function* (7) *selects the action:*

$$a_n(\hat{\lambda}; x^n) \tag{14}$$

from (8)*, where* $\hat{\lambda} = \max\{0, \lambda^*\}$ *and the scalar* λ^* *solves* $\psi^A(x^n, \lambda^*; x^n) = \gamma$ *in* (13)*.*

Proof — See Appendix A.

The decision rule outlined in Corollary 2.2 prescribes that the decision-maker adopts the judgmental decision if there is insufficient statistical evidence to reject it. Otherwise, it advocates shrinking the decision towards the QMLE-implied decision. The degree of shrinkage is determined by the level of significance α. The distinguishing characteristic of this decision rule is that it performs worse than the judgmental decision $\tilde{a}_n$ with a probability less than the level of significance α, according to the quasi-maximum likelihood model as defined by the pdf $f_t(X^t, \theta)$.

2.4 Selecting the Least Misspecified Model

Suppose the decision-maker has access to M alternative statistical models. For each model $m \in \mathscr{M} \equiv \{1, \ldots, M\}$, let $\theta^m \in \mathbb{R}^{p_m}$ denote its parameter and $f_t^m(X^t, \theta^m)$ its functional form. The decision incorporating judgment implied by model m is $a_n(\hat{\lambda}^m; x^n)$.

Ideally, one would compare the true losses $L_n^0(a_n(\hat{\lambda}^m; x^n))$ across models and select the lowest. In practice, it is impossible to compute $L_n^0(a_n(\hat{\lambda}^m; x^n))$, because the data-generating process is unknown. Nevertheless, it is possible to approximate its average value over time, by considering all the implied past actions.

To evaluate a model's in-sample performance, we need a history of decisions $a_t(\hat{\lambda}^m; x^n)$ for $t = 1 \ldots n$. Our original judgment $\tilde{a}_n$ is only defined for time n. To construct a full history of decisions consistent with this final judgment, we define a judgment-implied decision based on a constrained QMLE.

Definition 2.7 (**Judgment-implied Decision**)**.** The **judgment-implied decision** is derived from the constrained QMLE

$$\tilde{\theta}(x^n) \equiv \arg\max_{\theta} \ell_n(x^n, \theta) \quad \text{s.t.} \quad \nabla_a L_n(\theta, \tilde{a}_n) = 0,$$

where $\tilde{a}_n$ is the judgmental decision. The associated action at any time t is

$$\tilde{a}_t(x^n) \equiv \arg\min_{a \in \mathbb{R}^q} L_t(\tilde{\theta}(x^n), a).$$

The convexity of the function $\mathscr{L}(\kappa(x_{t,h}), \cdot)$ ensures the uniqueness of such a decision. By construction, at time n, $\tilde{a}_n(x^n)$ coincides with the original judgmental decision $\tilde{a}_n$. For earlier times $t < n$, however, it can be used to compute the implied judgmental decision given $\tilde{\theta}(x^n)$.

Define the population loss distance between the optimal action a_t^0 and model m as:

$$\Pi_n(a^0 : a^m) \equiv (n-h)^{-1} \sum_{t=1}^{n-h} (L_t^0(a_t(\hat{\lambda}^m; x^n)) - L_t^0(a_t^0)), \tag{15}$$

where a_t^0 is the truly optimal action at time t as defined in Equation (6). Similar to the Kullback-Leibler Information Criterion, $\Pi_n(a^0 : a^m) \geq 0$. It is equal to 0 if and only if $a_t(\hat{\lambda}^m; x^n) = a_t^0$, for all $t = 1, \ldots, n$, implying that the model is correctly specified for the given sample.

The model m^* that minimizes the empirical in-sample loss consistently selects the model that minimizes the population loss distance. This is stated formally in the following theorem, which analyzes the probability limit of the empirical loss over repeated samples.

Theorem 2.3 (**Consistent Model Selection**). *Assume the conditions of Corollary 2.2 are satisfied and that the process* $\{\mathscr{L}(\kappa(Y_{t,h}), \cdot)\}$ *obeys the Uniform Law of Large Numbers. Then, if model* $m^* \in \mathscr{M}$ *minimizes* $\operatorname{plim} \hat{L}_n^m(Y^n)$*, where* $\hat{L}_n^m(Y^n) \equiv (n-h)^{-1} \sum_{t=1}^{n-h} \mathscr{L}(\kappa(Y_{t,h}), a_t(\hat{\lambda}^m; x^n))$*, it also minimizes* $\operatorname{plim}_{n\to\infty} \Pi_n(a^0 : a^m)$*, for all* $m \in \mathscr{M}$*. If model* m^* *is correctly specified,* $\operatorname{plim} \hat{L}_n^{m^*}(Y^n) = \operatorname{plim}_{n\to\infty} (n-h)^{-1} \sum_{t=1}^{n-h} L_t^0(a_t^0)$.

Proof — See Appendix A.

According to Theorem 2.3, choosing the model with lowest empirical in-sample loss ensures that the decision-maker eventually selects the decision rule which is closest to the optimal one in population, as measured by the Kullback-Leibler-type information criterion

$\Pi_n(a^0 : a^m)$.

Here it is crucial that decisions are fixed by the observed x^n, while performance is evaluated through repeated draws of the auxiliary sample Y^n. This separation ensures we evaluate the actual rule implied by x^n, rather than averaging over alternative possible realizations of X^n.

If the level of significance $\alpha = 1$, the forecast horizon $h = 1$ and the loss is quadratic, the criterion reduces to in-sample mean squared error. In nested models, this always favors the largest model. However, when $\alpha < 1$, this preference for larger models no longer necessarily holds. The choice is restricted to decision rules consistent with judgment, and among these, the one with best in-sample fit is selected.

While Theorem 2.3 ensures consistency of the proposed model selection procedure, for any finite sample, the model m^* with the lowest empirical loss is not necessarily the best model. As originally pointed out by Diebold and Mariano (1995), for any particular realization one should take into account the statistical uncertainty associated with the empirical loss functions of the various decision rules.

Under the conditions of Theorem 2.3, we have the following asymptotic result:

$$\sqrt{n}(\hat{L}_n^{m^*}(Y^n) - \mu^{m^*}) \overset{A}{\sim} N(0, (\sigma^{m*})^2), \tag{16}$$

where $\mu^{m^*} = \text{plim}\hat{L}_n^{m^*}(Y^n) \geq \text{plim}\hat{L}_n^0(Y^n) \equiv \mu^0$, and $\hat{L}_n^0(Y^n) \equiv (n-h)^{-1}\sum_{t=1}^{n-h}\mathscr{L}(\kappa(Y_{t,h}), a_t^0)$. The variance term can be estimated by

$$(\hat{\sigma}_n^{m*})^2 \equiv (n-h)^{-1}\sum_{t=1}^{n-h}(\mathscr{L}(\kappa(x_{t,h}), a_t(\hat{\lambda}^{m*}; x^n)) - \hat{L}_n^{m^*}(x^n))^2. \tag{17}$$

For any finite sample size, it is possible that $\hat{L}_n^{m^*}(x^n) < \hat{L}_n^0(x^n)$. Since it is impossible to know whether a given value of $\hat{L}_n^m(x^n)$ is low due to genuine superiority or merely by chance, it seems reasonable to average across the different decision rules. The proposed averaging is based on the *p-value* associated with the null hypothesis $H_0 : \mu^m = \mu^{m^*}$ for all $m \neq m^*$. Other averaging schemes are possible.

Theorem 2.4 (**Consistent Model Averaging**). *Let*

$$\omega_n^m(x^n) \equiv 1 - \Phi(\sqrt{n}(\hat{L}_n^m(x^n) - \hat{L}_n^{m^*}(x^n))/\hat{\sigma}_n^{m*}), \tag{18}$$

where m^* is the model with lowest $\hat{L}_n^m(x^n)$ among $m \in \mathscr{M}$, and $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. Under the conditions of Theorem 2.3, the decision rule

$$\bar{a}(x^n) \equiv (\sum_{m=1}^{M} \omega_n^m(x^n) a_n(\hat{\lambda}^m; x^n))/\sum_{m=1}^{M} \omega_n^m(x^n) \tag{19}$$

converges in probability to the decision rule(s) with lowest $\mu^m \equiv \operatorname{plim} \hat{L}_n^m(Y^n)$, $m \in \mathscr{M}$. Letting $\mathscr{J} = \{1, \ldots, J\}$ denote the set of models associated with lowest μ^m, the asymptotic weight of decision rule $a_n(\hat{\lambda}^i; x^n)$, $i \in \mathscr{M} \setminus \mathscr{J}$, is 0.

Furthermore, $\operatorname{plim}_{n\to\infty} \Pi_n(a^0 : \bar{a}(x^n)) \leq \operatorname{plim}_{n\to\infty} \Pi_n(a^0 : a_n(\hat{\lambda}^m; x^n))$, *for all $m \in \mathscr{M}$.*

Proof — See Appendix A.

The decision rule $\bar{a}(x^n)$ is a weighted average of all decision rules $a_n(\hat{\lambda}^m; x^n)$, with weights $\omega_n^m(x^n)$ normalized by their sum. The weight is proportional to the cdf of the standard normal distribution evaluated at a scaled difference between its empirical loss and the minimum empirical loss. This construction gives higher weights to models whose empirical losses are closest to the minimum. Asymptotically, models that are correctly specified will converge in probability to μ^0 (the true minimum loss) and will be the only ones to receive positive weight. Misspecified models m, for which $\mu^m > \mu^0$, will asymptotically receive zero weight. If all models are misspecified, positive weights are assigned only to those models with lowest μ^m.

The second part of the theorem states that the weighted average of the asymptotically best decision rules produces a combined decision rule which is weakly superior to any single best decision rule in terms of minimizing the Kullback-Leibler-type criterion $\Pi_n(a^0 : a_n(\hat{\lambda}^m; x^n))$. In case the best decision rules are associated with nested models, this averaging procedure offers no asymptotic improvement, as all decision rules converge to the same

point. If the best-performing models are non-nested, the convexity of the loss function implies that averaging provides a strictly lower asymptotic loss than any of the individual constituent models, highlighting the benefits of diversification across different modeling approaches.

3 The Conditional Approach to Model Selection and Radical Uncertainty

There is a large literature on model selection in statistics and econometrics. Several methods have been proposed and range from hypothesis testing to goodness of fit and forecast evaluation (see Dufour (2008)). Hypothesis testing strategies typically follow either a specific-to-general path (MacKinnon (1992)) or a general-to-specific path (Davidson et al. (1978)). Penalized goodness-of-fit criteria are motivated by the idea that model selection should be interpreted as a problem minimizing a given loss function. One problem with this approach is that a typical in-sample loss function always decreases when more variables are added to the model. Solutions have been proposed to penalize the number of regressors (Akaike (1969), Hannan and Quinn (1979), Schwarz (1978)). Forecast evaluation methods evaluate the performance of the model on post-sample data (Diebold (2004)).

The model selection criterion proposed in this paper combines all three elements in a coherent fashion. Hypothesis testing is used to test the optimality of the judgmental decision and, in case of rejection, to arrive at the decision with judgment. Goodness of fit and forecast evaluation are combined in Theorem 2.3, where the in-sample loss function is used to compare the fit of the different models.

A fourth approach is Bayesian model selection, which is methodologically closest to the framework proposed in this paper. In a Bayesian setting, priors are imposed on model parameters and on models themselves, and optimal decisions are obtained by minimizing the posterior expected loss. As shown in Manganelli (2023), Bayesian decision rules can be mapped into a frequentist decision with judgment. The key difference is that in the Bayesian

approach judgment is encoded in the prior distribution over the statistical parameters θ, whereas in the present framework, judgment is explicitly defined as a benchmark decision $\tilde{a}$ and controlled via the significance level α. Bayesian model selection proceeds by comparing posterior model probabilities, which implicitly combine goodness of fit and model complexity. In contrast, the model selection criterion proposed here compares the conditional performance of the decisions associated with alternative models, as characterized in Theorem 2.3, without requiring the specification of a prior distribution over parameters or models.

There is a deeper problem connected with the standard frequentist model selection approaches, which makes inference on post-selection estimators particularly tricky and often unreliable, as clearly illustrated by Leeb and Pötscher (2005). The approach of this paper resolves these pathologies by shifting the object of interest from the random estimator to the realized decision. By evaluating performance conditionally on the observed data, we demonstrate that the non-uniform convergence issues vanish.

In the following, Section 3.1 presents the canonical setting of Leeb and Pötscher (2005) and Section 3.2 recasts it within our conditional decision-theoretic framework. Section 3.3 derives an impossibility result in statistical decision theory: under model misspecification, the true level of significance of any statistical decision is fundamentally unknowable.

3.1 The Setup

Consider the linear regression model, from Section 2 of Leeb and Pötscher (2005)

$$Z_t = \theta_1 w_{t1} + \theta_2 w_{t2} + \epsilon_t \qquad (1 \leq t \leq n), \tag{20}$$

where $\epsilon_t \sim i.i.d.N(0,1)$ and the regressors are assumed to be non-random. Let $z = (z_1, \ldots, z_n)'$, $w_1 = (w_{11}, \ldots, w_{n1})'$, $w_2 = (w_{12}, \ldots, w_{n2})'$, and $x^n = (z, w_1, w_2)$ denote the $(n \times 3)$ matrix containing all observed data. Let $X^n = (Z, w_1, w_2)$ denote the random data array, where randomness arises only through Z. Consistent with our notation, we

denote random variables with capital letters and realized values in lower case letters. Since the regressors are assumed to be non-random, we maintain the lower case notation in the matrix defining X^n.

The parameter of interest is θ_1 and the question is whether to estimate only the restricted model (i.e., imposing $\theta_2 = 0$) or the unrestricted model. The restricted and unrestricted models are defined as

$$R : \theta_2 = 0, \qquad U : \theta_2 \in \mathbb{R}.$$

For notational simplicity, we index models by $m \in \{R, U\}$.

Traditional model selection is performed using a pre-test based on the statistical significance of the least-squares estimator $\hat{\theta}_2(X^n)$:

$$\hat{m}(X^n) = \begin{cases} R & \text{if } |\sqrt{n}\hat{\theta}_2(X^n)/\sigma_2| \leq c, \\ U & \text{otherwise} \end{cases}$$

where $(w'w/n)^{-1} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ and $w = [w_1, w_2]$ is a $(n \times 2)$ matrix collecting all the regressors.

The resulting post-selection estimator is

$$\hat{\hat{\theta}}_1(X^n) = \hat{\theta}_1(R; X^n) I(\hat{m}(X^n) = R) + \hat{\theta}_1(U; X^n) I(\hat{m}(X^n) = U),$$

where $\hat{\theta}_1(R; X^n)$ and $\hat{\theta}_1(U; X^n)$ denote the least-squares estimators of θ_1 under the restricted and unrestricted models, respectively.

Leeb and Pötscher (2005) show that although $\hat{\hat{\theta}}_1(X^n)$ is pointwise consistent, its finite-sample behavior is non-standard and can be highly misleading, because of non-uniform convergence issues.

3.2 Conditional Evaluation of Realized Decisions

The setup of this paper addresses the issue of post-model-selection inference by re-framing the problem. Consider the squared-error loss

$$\mathscr{L}(Z_t, a_t) = (a_t - Z_t)^2,$$

so that the conditional loss given θ^m is

$$L(\theta^m, a_t) = E_{\theta^m}[\mathscr{L}(Z_t, a_t)], \quad \text{for } m \in \{R, U\},$$

where $\theta^R = \theta_1$ and $\theta^U = [\theta_1, \theta_2]'$, and the expectation is taken with respect to the *quasi-likelihood* of Z_t induced by θ^m.

We now reinterpret model selection as a *decision problem*. Denote the decisions associated with the two models as $a_t(\hat{\lambda}^m; x^n)$ for $m \in \{R, U\}$ as defined in Corollary 2.2. Crucially, conditional on the sample x^n, these are *non-random* elements of the action space $\mathbb{R}$. The object of interest is not the random estimator $\hat{\hat{\theta}}_1(X^n)$, but the realized decisions $a_t(\hat{\lambda}^m; x^n)$. According to Theorem 2.3, the choice depends on the best in-sample fit.

Introduce an auxiliary random process ε_t, such that

$$Y_t = \theta_1 w_{t1} + \theta_2 w_{t2} + \varepsilon_t \qquad (1 \leq t \leq n),$$

where $\varepsilon_t \overset{i.i.d.}{\sim} N(0, 1)$ and ε_t is independent of ϵ_t for all t. This auxiliary random variable serves to evaluate the performance of a fixed, data-dependent decision under repeated realizations of the outcome variable, holding the decision itself constant.

The finite-sample approximation to the population risk is given by:

$$\hat{L}_n^m = n^{-1} \sum_{t=1}^{n} \mathscr{L}(Y_t, a_t(\hat{\lambda}^m; x^n)), \text{ for } m \in \{R, U\},$$

and an estimate of it can be obtained by replacing Y_t with the observed realization z_t.

Since $a_t(\hat{\lambda}^m; x^n)$ is fixed conditional on x^n, no averaging over alternative samples that would induce different actions takes place. The decision $a_t(\hat{\lambda}^m; x^n)$ is evaluated solely on the basis of its performance as measured by the finite sample approximation induced by the unknown true distribution G. In contrast, the unconditional risk considered by Leeb and Pötscher (2005), $E_{\theta^U}[(\hat{\hat{\theta}}_1(X^n) - \theta_1)^2]$, averages over samples that would lead to different realized decisions.

The pathologies emphasized in the post-model-selection literature stem from the evaluation of the *random estimator* $\hat{\hat{\theta}}_1(X^n)$ under unconditional frequentist risk. This risk function is discontinuous because it averages over the selection event. However, once the object of interest is a realized decision and performance is evaluated conditionally on the observed data, this source of non-uniformity disappears. These pathologies are therefore not intrinsic to data-dependent procedures, but to the evaluation of random estimators rather than realized decisions. Our analysis does not refute the results of Leeb and Pötscher (2005). It rather demonstrates that those results apply to unconditional parameter inference, a goal that is orthogonal to conditional decision-making.

In our framework, "model selection" is shorthand for selecting among a finite set of data-dependent decisions, rather than selecting a statistical model as an intermediate object for inference.

3.3 Uncertainty About the Level of Uncertainty

The conditional decision framework developed in this paper resolves the inferential pathologies associated with unconditional evaluation of post-model-selection estimators. However, when models are misspecified, a deeper and irreducible source of uncertainty remains. This section formalizes this source of uncertainty and shows that it concerns not the optimality of realized decisions conditional on a maintained model, but the reliability of the probabilistic guarantees attached to them. Importantly, this limitation is not a finite-sample problem, nor an artifact of asymptotic approximations, but persists asymptotically unless one assumes correct specification or a stationary environment in which the maintained model

class is known to approximate the true data-generating process arbitrarily well (Hornik et al. (1989)).

Recall from Corollary 2.2 that the decision with judgment is determined by the test statistic $Z_n \equiv \hat{\sigma}^{-1}\sqrt{n}\nabla_\lambda L_n(\hat{\theta}(Y^n), a_n(0; x^n))$ and the critical value $c_{\alpha/2}$.[1] The decision-maker shrinks away from the judgmental decision $\tilde{a}_n$ if and only if $z_n < c_{\alpha/2}$, where, as usual, lower case z_n denotes the realized value of Z_n. Recall also that the significance level α plays the role of a user-chosen bound on the probability that the statistical decision performs worse than the judgmental decision.

There are two possibilities.

Case 1: Correct Specification Assume the parametric model $f_t(Y^t, \theta)$ is correctly specified. That is, there exists a true parameter θ^0 such that the data x^n is generated by the distribution G, which is perfectly characterized by the function $f_t(\cdot, \theta^0)$. In this case, the null hypothesis that the judgmental decision is optimal can be equivalently rewritten as $\nabla_\lambda L_n(\theta^0, a_n(0; x^n)) = \nabla_\lambda L_n^0(a_n(0; x^n)) = 0$, where $L_n^0 \equiv E_G[\mathscr{L}(\kappa(X_{n,h}), a_n | X^n)]$ is defined in Equation (5).

The probability of Type I error (rejecting the judgmental decision when it is actually optimal) is exactly controlled by the level of significance:

$$P_{\theta^0}(Z_n < c_{\alpha/2} | Z_n < 0) \to \alpha \quad \text{as } n \to \infty. \tag{21}$$

Following Manganelli (2025), the rejection region is one-sided and conditioning on $Z_n < 0$ reflects the direction of improvement relative to the judgmental decision.

As discussed in Section 4 of Manganelli (2025), the level of significance α represents the *level of uncertainty* implicit in any statistical decision problem, which in the ideal case of correct specification is bounded by α.

[1] As in the rest of the paper, estimation is based on the observed sample realizations x^n; auxiliary variables Y^n are introduced only to conceptualize repeated-loss evaluation.

Case 2: Misspecification Now assume, as is realistic in practice, that the model is misspecified. The true DGP is governed by the unknown cumulative distribution function G, which does not belong to the parametric family $f_t(\cdot, \theta)$. The QMLE $\hat{\theta}(Y^n)$ converges to the pseudo-true parameter θ^*. The test statistic Z_n is constructed using the asymptotic variance $\hat{\sigma}^2$ derived from the *assumed* model $f_t(\cdot, \hat{\theta}(Y^n))$.

However, the optimality of the judgmental decision should be evaluated using the distribution of the gradient of the true loss function, which in turn is determined by the distribution G. Let Z_n^G denote the test statistic associated with the gradient of the true loss function, whose distribution is generated by G. Note that Z_n^G is infeasible, as it requires knowledge of G. In case of misspecification, the null hypothesis of optimality of the judgmental decision under G will generally be different from the one under the chosen parametric specification, that is $\nabla_\lambda L_n^0(a_n(0; x^n)) = 0$ is not necessarily the same as $\nabla_\lambda L_n(\theta, a_n(0; x^n)) = 0$. As a consequence, the true probability of rejection becomes:

$$P_G(Z_n^G < c_{\alpha_{true}/2} | Z_n^G < 0) \to \alpha_{true} \neq P_{\theta^*}(Z_n < c_{\alpha/2} | Z_n < 0) \to \alpha \quad \text{as } n \to \infty. \tag{22}$$

The term α_{true} is the actual uncertainty the decision-maker faces. The difference between the perceived uncertainty and the actual uncertainty is:

$$\Delta_{Hume} = |\alpha_{true} - \alpha|. \tag{23}$$

The quantity Δ_{Hume} measures the inductive gap: the distance between nominal and actual frequentist error control induced by model misspecification. Unlike robustness analyses, which study sensitivity across a neighborhood of models, Δ_{Hume} reflects ignorance about which neighborhood is relevant in the first place.

The Problem of Induction as a Statistical Quantity The term Δ_{Hume} represents *uncertainty about the level of uncertainty.*

1. In the case of correct specification, $\Delta_{Hume} = 0$, as the past reliability of the model

guarantees future control over Type I errors.

2. In the case of misspecification, Δ_{Hume} is unknown because G is unknown.

This gap Δ_{Hume} provides a precise mathematical formulation of Hume's problem of induction (Hume (1739)) expressed in terms of econometric model misspecification. The decision-maker selects α hoping to bound the probability of Type I error. However, because the structural link between the model ($f_t(\cdot, \hat{\theta}(Y^n))$ and the reality (G) is incorrect, the actual error rate α_{true} is unobservable. Standard analyses of misspecification focus on bias, efficiency loss, or pseudo-true parameters. In contrast, Δ_{Hume} captures a different object: the gap between nominal and actual error probabilities. This gap persists even when the pseudo-true parameter is well defined and consistently estimated.

Importantly, this form of uncertainty cannot be eliminated by increasing the sample size or by refining the asymptotic approximation. It arises from the impossibility of learning whether the maintained model provides an adequate approximation for the purpose of evaluating losses. It can be mitigated by increasing the number of estimated statistical models as suggested in this paper, but ultimately the decision-maker faces uncertainty about the level of uncertainty, reflecting the fact that probabilistic guarantees are themselves conditional on an unverifiable approximation of the data-generating process. The remaining uncertainty is therefore not a defect of the decision rule, but a logical consequence of model-based reasoning itself, and hence must be addressed through judgment rather than eliminated by statistical refinement.

4 Inflation Forecasting Models

The implementation of the theory developed in Section 2 requires the econometrician to specify three key ingredients of forecasting:

1. The loss function and the judgment of the decision-maker;

2. The data used in the forecasting model;

3. The econometric model specification (functional form and parameters).

This section reviews each of these elements, with the objective of forecasting inflation in the euro area.

4.1 Loss Function and Judgment

The loss function and judgment are both inherently subjective choices of the decision-maker, and thus will vary from one to another. We take here a specific perspective on the problem of inflation forecasting.

We assume a standard quadratic loss function:

$$\mathscr{L}(\kappa(Y_{t,h}), a_t) = 0.5(\kappa(Y_{t,h}) - a_t)^2, \tag{24}$$

where

$$\kappa(Y_{t,h}) = h^{-1} \sum_{j=1}^{h} Y_{t+j}^1 \tag{25}$$

represents the realized average inflation over the forecast horizon h. Recall that $Y_t \equiv (Y_t^1, \ldots, Y_t^g)'$ and we adopt the convention that the term Y_{t+j}^1 denotes inflation at period $t+j$, with t being the current period from which the forecast is made.

A sensible strategy to form judgment about inflation is by appealing to the statutory mandate of the European Central Bank. According to Article 127 of the Treaty on the Functioning of the European Union, *'the primary objective of the European System of Central Banks [...] shall be to maintain price stability.'* Following its Monetary Policy Strategy Review, concluded in July 2021, the Governing Council of the European Central Bank has updated its quantitative definition: *'The Governing Council considers that price stability is best maintained by aiming for 2% inflation over the medium term. This target is symmetric, meaning that negative and positive deviations from this target are considered equally undesirable.'* (European Central Bank (2021), paragraph 5). This definition was subsequently reconfirmed by the 2025 ECB Strategy Assessment (European Central Bank

(2025)).

As this definition aims for a symmetric 2% over the medium term, it still requires a precise quantitative interpretation for our specific forecasting exercise. We therefore personally interpret the 2% target as a precise point of 2.0% and the *medium term* as 24 months. Other interpretations of the ECB definition could, of course, be justified.

Trusting that the ECB will deliver on its mandate, the judgmental forecast can be defined as:

$$\tilde{a}_{n,24} = 2\%, \tag{26}$$

which implies that our judgment for the expected value of $\kappa(Y_{n,24})$ is 2%. In other words, a reasonable judgmental forecast is that the average annualized inflation over the next two years is 2%. While this appears to be justifiable when facing a credible central bank with an explicitly defined inflation objective, such information is rarely incorporated in empirical applications of inflation forecasting. One exception is Diron and Mojon (2008). They argue that a major advantage of quantified inflation objectives is to anchor inflation expectations. Since economic agents usually set prices and wages over some horizon, it is reasonable to expect that when taking these decisions they take into account their own expectations of the evolution of inflation. When facing a credible central bank, the official inflation objective may act as an anchor to the expectations of the economic agents. The coordination of all expectations around the official inflation objective should inherently help to deliver realized inflation close to the objective. Setting the judgmental decision as in (26) may therefore prove to be a sound forecasting decision.

Any other judgmental forecast at any other horizon can be constructed using a similar strategy.

4.2 Data

Given the ECB definition of price stability, a necessary variable of the forecasting model is euro area HICP. The application of this paper considers the annual rate of change of the overall index with changing composition, as provided by Eurostat. Given the short term volatility of headline inflation, many central banks monitor other measures of underlying inflation, which may give a better sense of the trends in inflation and its likely evolution in the medium term. Many different proxies for underlying inflation exist. We use here the euro area HICP excluding energy and unprocessed food (HICPX). We also consider real economic variables, such as unemployment (UN) and industrial production (IP), because as suggested by the Phillips curve there may be an inverse relationship between inflation and the level of economic activity, and the 3-month euro area interest rate (IR). All data are monthly, ranging from January 1999 (when the euro was introduced) to May 2025.

Key summary statistics of the five variables and the time series are reported in Table 1 and in Figure 6 of the Online Appendix C.

4.3 Models

The last ingredient of the forecasting process is the choice of the functional form of f_t, the building block of the quasi-likelihood function introduced in Section 2.1. For simplicity, we consider only the class of VAR(p) models:

$$Y_t = c + \sum_{i=1}^{p} C_i Y_{t-i} + U_t,$$

where $Y_t = (HICP, HICPX, UN, IP, I)'$, c is a 5×1 vector, C_i are 5×5 matrices of coefficients, and U_t is a 5×1 vector of residuals. For a given p, we allow for selection of any possible subset of individual lagged regressors. Unlike in standard VAR estimation where all lags up to lag p are included, a key contribution of this paper is to treat any

lag of the VAR as a regressor and select only a subset of those lags. Since the algebra is conceptually straightforward, but rather cumbersome, we refer to the Appendix B for detailed derivations, which build on Lütkepohl (2005). In this context, the quasi-likelihood contribution f_t takes the following functional form:

$$f_t(X^t, \theta) = \exp(-0.5 u_t' u_t), \tag{27}$$

where u_t is a vector of residuals defined in Equation (29) of Appendix B.

Estimation of asymptotic variance-covariance matrices follows the procedure detailed in Section 2.1.

4.4 Further Modeling Choices

The actual implementation of the forecasting process requires the econometrician to specify additional modeling choices. In theory, each of these choices could give rise to potentially different models and therefore be optimized using the model selection procedure described in this paper. In practice, however, limitations regarding time and computing power typically necessitate making certain pragmatic decisions.

The primary modeling choice concerns the functional form of the econometric model. In the case of a VAR(p) model, we have specified the length of the vector Y_t to be 5, comprising headline inflation (HICP), core inflation (HICPX), the unemployment rate (UN), industrial production (IP), and the interest rate (I). Other options for the functional form could include factor models, Dynamic Stochastic General Equilibrium (DSGE) models, or deep neural networks from the machine learning literature (see, for instance, Faust and Wright (2013) for a comprehensive overview of common methods used for inflation forecasting). As long as the assumptions about f_t are satisfied, any of these alternative models can be evaluated within the same conceptual framework. In fact, the framework developed in this paper can be employed to weigh these alternative modeling strategies.

The second specification choice concerns the maximum lag order, p. We set this to 12,

which corresponds to one year of monthly lags.

The third specification choice pertains to the number of regressors included in the model. This is particularly relevant given our approach of selecting only a subset of lags, which can still lead to a substantial number of parameters relative to the available data. Recall that the arguments for the asymptotic validity of our estimators, as detailed in Section 2.1, typically rely on the time series length (n) going to infinity while the number of variables (cross-sectional dimension) is fixed. In such standard settings, to ensure reliable estimation and valid inference, the ratio between the effective number of observations and the total number of parameters to be estimated cannot be too small.

Pragmatically, we arbitrarily choose this ratio to be 50. This means that, on average, each model parameter is estimated with at least 50 observations. While this rule of thumb is helpful for managing finite-sample properties, it is important to note that the asymptotic behavior of estimators in scenarios where the number of parameters grows with the sample size (e.g., if this ratio were strictly held constant as $n \to \infty$) typically requires specialized high-dimensional asymptotic theory. The specific validity of our chosen ratio for our models could be further tested through a Monte Carlo experiment.

The fourth specification choice concerns the initialization of the VAR model. We handle the pre-sample values by setting them equal to their respective historical average of the first year. Other approaches are possible, such as treating these initial values as parameters to be estimated within a full maximum likelihood framework. Furthermore, in line with the stationarity assumption imposed in Section 2.1, we discard any quasi-maximum likelihood estimates that imply non-stationary VAR processes.

Even with these simplifications, a VAR model with $p = 12$ lags, $g = 5$ dependent variables, and a constraint of including at most $MR = 32$ maximum number of regressors can result in a vast number of possible model specifications. The total pool of potential regressors (including g constant terms and g^2p lagged terms) for a full VAR(p) is $g + g^2p = 5 + 5^2 \times 12 = 305$. The number of ways to select up to MR regressors from this pool is given by $\sum_{i=1}^{MR} \binom{g+g^2p}{i}$. For our parameters, this sum is approximately $2.5 \cdot 10^{42}$. Even the

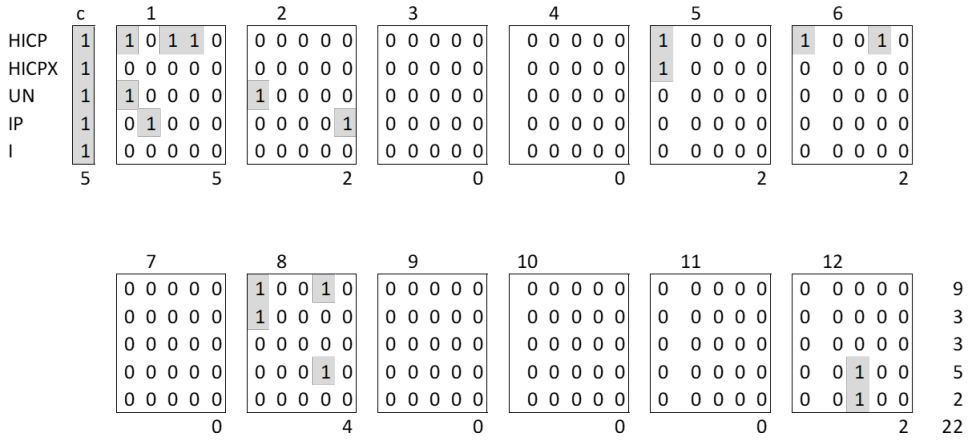


Figure 1: Example of Model Specification

Note: The figure reports an example of a possible VAR model specification, with 12 lags, 5 variables and 22 regressors. The entries with 1s represent the parameters to be estimated. The entries with zeros represent variables which are excluded from the model. The number under each autoregressive block is the sum of parameters estimated for each block. The numbers in the last column represent how many variables from each of the VAR equations are included in the model.

fastest computer cannot cope with a complete grid search.

We use, instead, an integer optimization algorithm to navigate this vast model space. The logic is the following. Let ϑ be a $(g+g^2p)$-vector of binary indicators (0s and 1s), where a 1 signifies the inclusion of a specific regressor. Let θ^m denote the vector of parameters to be estimated corresponding to the regressors indicated by 1s in ϑ. Each unique combination of ϑ corresponds to exactly one model, identified by θ^m, for which the associated loss $\hat{L}_n^m(x^n)$ can be computed. Figure 1 illustrates an example of such a model specification. Model selection can now be framed as the following integer optimization program:

$$\min_{\vartheta} \hat{L}_n^m(x^n) = (n-h)^{-1} \sum_{t=1}^{n-h} 0.5 \left(\kappa(x_{t,h}) - a_t(\hat{\lambda}^m; x^n)\right)^2 \tag{28}$$

$$s.t. \quad ||\vartheta||_0 \leq MR,$$

where $||\vartheta||_0$ is the L_0 operator that counts non-zero elements.

While Matlab's genetic algorithm offers options for solving integer-valued optimization problems, we found it to be rather inefficient and time-consuming for our problem. A more effective alternative proved to be the *patternsearch* algorithm. Although not specifically designed for integer optimization, it can be adapted to solve this type of problem under

suitable choices of its options. While integer optimization techniques for subset selection are increasingly recognized in the broader econometric and statistical literature (e.g., for linear regression models), our specific application of the *patternsearch* algorithm to VAR model selection in this integer-valued context represents a less common, yet computationally pragmatic, approach. For detailed implementation, we refer to the associated codes.[2] In our experience, the algorithm appears quite sensitive to the user-supplied initial conditions, suggesting that finding the global optimum is challenging. However, finding the global optimum is not strictly necessary to implement the procedure, it simply implies that the performance of the selected model could be further improved by a more accurate optimization routine. We provided the following initial conditions: models containing only constants, models including the first MR regressors, and three models selected using LASSO techniques.

5 Inflation Forecasts for the Euro Area

Figure 2 presents the inflation forecasts derived from the best 5-variables VAR model, selected according to the methodology as described in Theorem 2.3. The best model for forecasting average 24-month inflation as of May 2025 is defined as the one with the lowest in-sample loss. The horizontal axis reports the forecasting horizon in months, while the vertical axis measures the forecast of the average annualized inflation.

The top line in Figure 2 shows the forecast from the 5-variables VAR model with a judgmental forecast of 4.0% for the 24-month horizon. The specific 24-month inflation forecast from this model is around 3.5%, and the three-year forecast is 4.6%. It is worth noting that the null hypothesis that our judgmental forecast of 4.0% is correct is rejected at a 10% level of significance.

A key feature of our approach is that a judgmental forecast, even when expressed at a single horizon (e.g., 24 months), is consistently mapped across all other forecasting horizons.

[2]Codes and data can be downloaded from www.simonemanganelli.org.

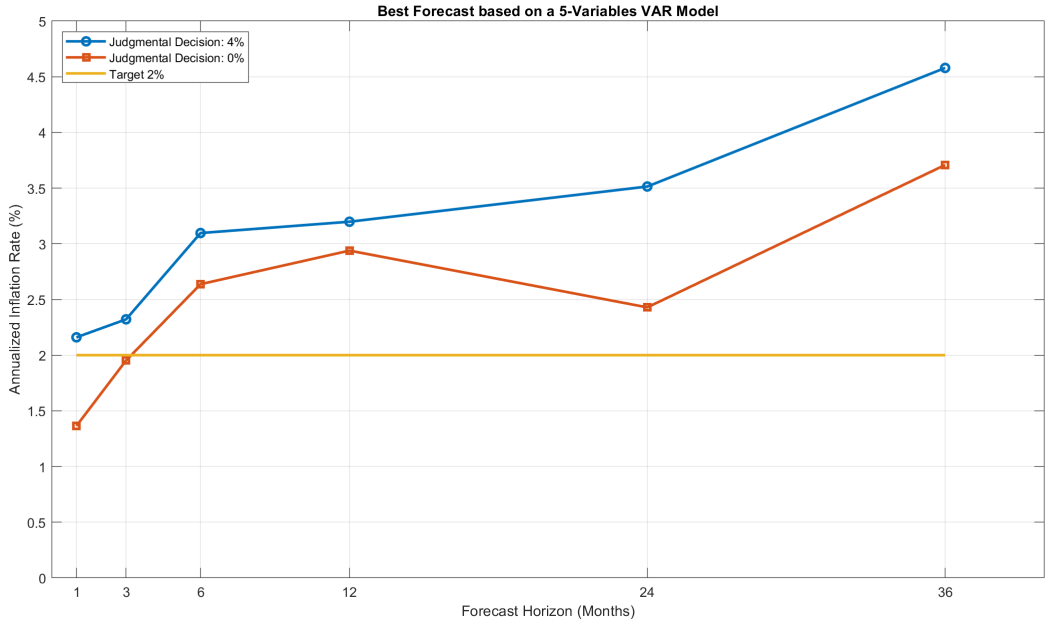


Figure 2: Forecasts of Euro Area Inflation

Note: The figure exhibits the best euro area inflation forecast at different horizons, with information available as of May 2025. The horizontal axis reports the forecasting horizon measured in months. The vertical axis measures the forecast of the average annualized inflation, where the average is taken over the months indicated on the horizontal axis. The top line is the forecast associated with a 24-month judgmental forecast of 4.0%. The bottom line is the forecast associated with a 24-month judgmental forecast of 0%. The horizontal line at 2% is reported as reference point.

This means that the inflation forecast at the 1-month horizon, for instance, incorporates the judgment expressed at the 24-month horizon. This is accomplished by using the constrained parameter estimates $\tilde{\theta}(x^n)$ obtained from the QMLE solving $\max_\theta \ell_n(x^n, \theta)$ *s.t.* $E(\kappa(Y_{n,24}(\theta))|x^n) = \tilde{a}_{n,24}$, to compute the judgmental forecast at any horizon of interest.

For comparison, the bottom line of the Figure 2 shows the forecast associated with an alternative 24-month judgmental forecast of 0%. While these two forecasts would converge with an infinite sample size, they differ substantially in finite samples. As shown in the figure, forecasts based on a 0% judgment are significantly lower, particularly for longer horizons.

For further comparison, Figure 3 presents the best forecast derived from a model selection based only on an autoregressive process for euro area inflation. While the very short term forecasts are similar, the long-term forecasts diverge. The AR model forecasts re-

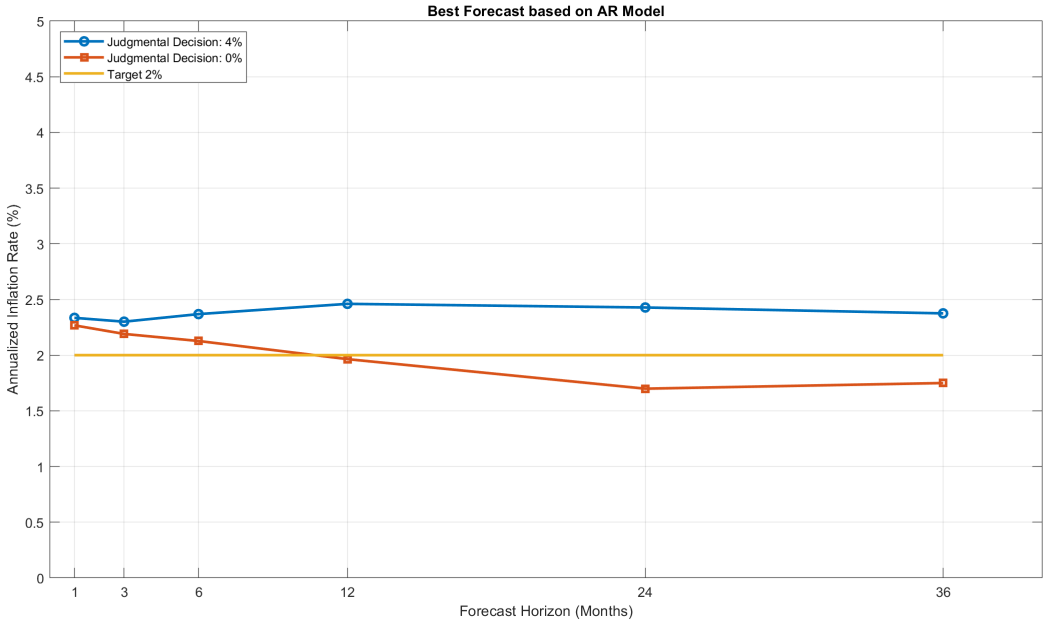


Figure 3: Forecasts of Euro Area Inflation Based on AR Model

Note: Best euro area inflation forecast based on model selection from a simple autoregressive model. See the notes of Figure 2 for explanatory details.

main close to 2% at longer horizons, whereas the 5-variables VAR model forecasts increase towards 4%.

Figure 4 highlights the differences in performance between the models by reporting the average in-sample loss at different horizons. The figure compares the best models selected from the univariate AR and the 5-variable VAR models. At short horizons, the performance is similar, with the univariate AR model having a slightly lower average loss. This is possible because, although the 5-variables VAR model can replicate a univariate AR model, the testing procedure used for constructing judgmental decisions can allow the simpler model to outperform the more complex one.

Significant differences start to emerge as the forecasting horizon increases, with 5-variables VAR model showing a stronger outperformance at 24 and 36 months. Figure 5 illustrates the in-sample performance of the best models selected from the sets of AR and 5-variables VAR models. The continuous blue line is the backward h-month moving average of HICP. The red dashed line represents the 5-variables VAR best model forecast

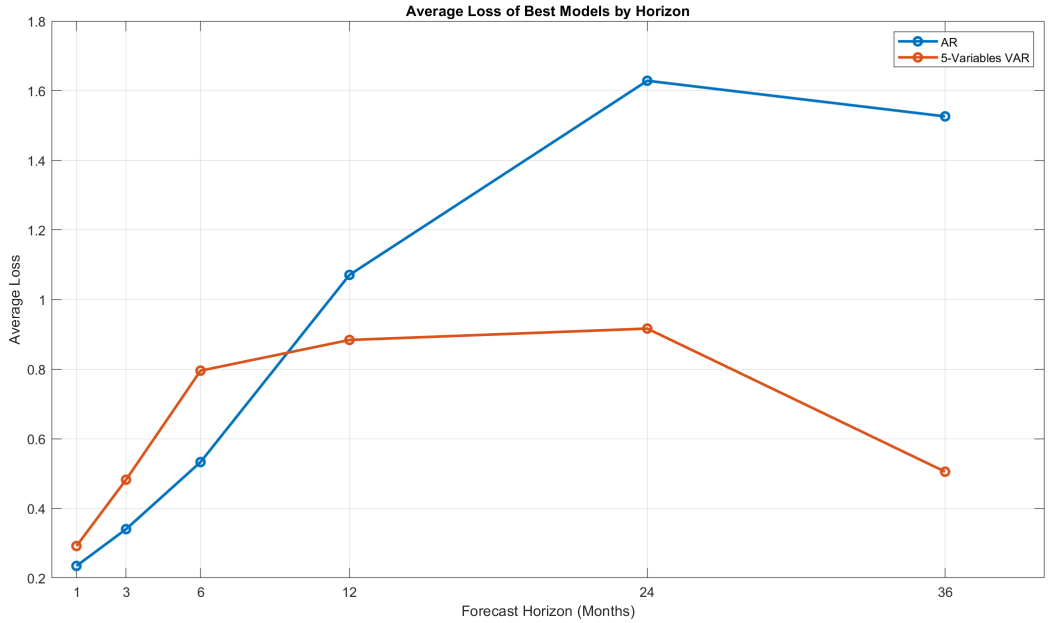
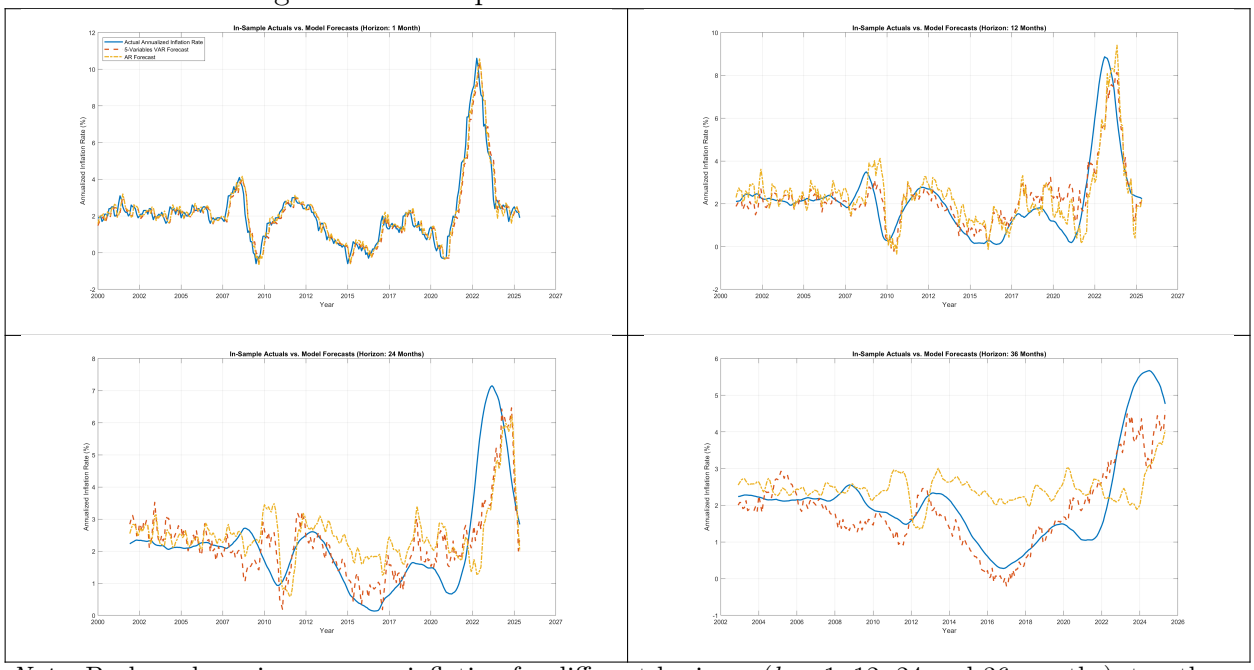


Figure 4: Performance of the Best Models

Note: Average in-sample loss when the best forecast model is chosen from different sets.

Figure 5: In-sample Performance at Different Horizons



Note: Backward moving averages inflation for different horizons ($h = 1$, 12, 24 and 36 months), together with the best in-sample forecasts from the set of all possible AR or five-variables VAR models.

at the beginning of the h-month forecasting window (but using the parameter estimates from the full sample). The yellow dashdot line is the best forecast associated with the AR

model. The average in-sample loss of Figure 4 is simply the average of the square difference between these forecasts lines and the actual inflation line. It is clear that, unlike the 5-variables VAR, the simple AR model lacks the dynamic complexity to accurately describe the stochastic process of inflation over two- and three-year periods.

6 Conclusion: Hume and the Problem of Induction

This paper studies how expert judgment can be integrated into econometric decision-making in the presence of model uncertainty and misspecification. It makes two main contributions. First, by evaluating realized decisions conditionally on the observed data, it resolves the inferential pathologies associated with unconditional evaluation of post-model-selection estimators. Second, it shows that under model misspecification a deeper and irreducible uncertainty remains, concerning the reliability of probabilistic guarantees themselves.

The distinction between conditional and unconditional evaluation is central to our findings. While the literature on post-model-selection inference has rightly highlighted the dangers of unconditional frequentist guarantees (e.g., Leeb and Pötscher (2005)), we demonstrate that conditioning on the realized sample resolves these pathologies for decision-making purposes. This suggests that the solution to the fragility of econometric inference lies not in refining the asymptotic approximations of random estimators, but in recognizing that statistical decisions are fixed actions taken in light of specific evidence.

Our framework represents a fundamental departure from Manganelli (2025), by explicitly embracing the premise of model misspecification. The test function is formulated to assess the optimality of the judgmental decision relative to the quasi-maximum likelihood loss function, rather than against the true, unobservable loss function. When the model is misspecified, the optimal action under the maintained model and the optimal action under the true data-generating process generally diverge. Consequently, the probability of Type I error is no longer consistent with the nominal significance level.

As discussed in Manganelli (2025), the significance level embodies the decision-maker's aversion to uncertainty: it reflects a willingness to adopt a statistical procedure only if its probability of yielding an action inferior to the judgmental input remains below a specified threshold. Under misspecification, this guarantee is lost. Even selecting the least misspecified model, the foundational premise that any model is an imperfect representation of reality implies an inherent uncertainty concerning the true Type I error. This constitutes an *uncertainty about the level of uncertainty*.

This profound epistemological challenge directly connects to *Hume's problem of induction*, which we formalized in Section 3 as the non-vanishing quantity Δ_{Hume}. Hume (1739), in his philosophical treatise, questioned the purely logical justification for drawing conclusions about unobserved phenomena based solely on past experience. We show that this philosophical problem has a precise statistical counterpart: the irreducible gap between the nominal significance level chosen by the decision-maker and the true probability of error under the unknown data-generating process, when probabilistic guarantees are evaluated relative to a maintained model. The irreducibility of Δ_{Hume} stems from the impossibility of validating probabilistic guarantees without assuming correct specification or structural stability. Econometric modeling relies on an inductive leap, assuming that the chosen parametric specification accurately captures the persistence of past relationships. In reality, the decision-maker faces an *uncertainty about the level of uncertainty*.

The paper does not claim to eliminate this uncertainty, nor to restore exact probabilistic guarantees under misspecification. Rather, it clarifies where and why such guarantees fail. As a consequence, judgment necessarily enters both at the beginning and at the end of the decision process. Our framework places judgment as the central starting point of the statistical decision process. The first step is to test whether the initial judgmental decision is optimal when confronted with the statistical evidence from the data and the class of estimated models. However, the preceding discussion on inductive uncertainty and the pervasive possibility of model misspecification provides a critical justification for judgment to play a further role at the end of the statistical decision process. Policy makers are

often wary of taking a purely model-driven outcome at face value and they may still wish to apply their own judgment to the final result, even after they have already influenced the process with their initial judgmental decision. While this ultimate deviation from a purely statistically-driven outcome might appear to lack a strictly scientific basis, it is defensible given the pervasive possibility of model misspecification and the decision-maker's ultimate accountability for their choices. The application of human judgment both at the beginning and at the end of the statistical process reflects how policy decisions are truly made (Kay and King (2020)). By systematically blending qualitative information, model-based evidence, and the informed judgment of decision-makers, our framework is consistent with real-world decision-making, particularly in fields like central banking.

Our findings have immediate implications for the growing reliance on Artificial Intelligence (AI) and machine learning in policy domains. AI systems are, at their core, sophisticated statistical models that identify patterns in historical data by optimizing an objective function. Consequently, they are subject to the same *Humean* limitations derived in this paper: they can minimize empirical losses, but they cannot quantify the *uncertainty about their own uncertainty* (Δ_{Hume}) in the presence of structural breaks or regime changes.

It is important, however, to distinguish between environments in which modern machine-learning methods excel and those typical of policy decision-making. Large language models and related architectures are trained on extraordinarily large cross-sectional datasets generated in relatively stationary environments, where the data-generating process is sufficiently rich and stable for universal approximation results (Hornik et al. (1989)) to be empirically relevant. In contrast, policy decisions are typically based on macroeconomic and financial time series that are short, evolving, and subject to structural change. In such settings, no increase in model complexity or sample size can guarantee that the maintained model class spans the true data-generating process.

As a result, while AI can be extremely effective at extracting regularities from abundant and stable data, it cannot eliminate the inductive gap that arises in decision problems characterized by scarcity of observations and instability of economic regimes. Unlike hu-

man experts, algorithms lack the capacity for *extra-statistical judgment*, that is, the ability to assess when the inductive link between past data and future outcomes has plausibly weakened. In the language of Knight (1921), AI operates well under *risk* (where probabilities are known or learnable), but is fragile under *uncertainty* (where the structure itself is in doubt). The gap Δ_{Hume} ensures that the human component in decision-making is a permanent statistical necessity.

The contribution of this paper is ultimately not to limit the use of statistical or machine-learning tools for decision-making, but rather to clarify the irreducible role of the decision-maker's judgment and accountability whenever probabilistic guarantees depend on unverifiable inductive assumptions.

A Appendix — Proofs

Proof of Corollary 2.2 — The result follows directly from Theorem 3.2 of Manganelli (2025). □

Proof of Theorem 2.3 — If model m^* minimizes

$$\operatorname{plim} n^{-1} \sum_{t=1}^{n} \mathscr{L}(\kappa(Y_{t,h}), a_t(\hat{\lambda}^m; x^n)),$$

it will also minimize:

$$\begin{aligned}
&\operatorname{plim} n^{-1} \sum_{t=1}^{n} \Big(\mathscr{L}(\kappa(Y_{t,h}), a_t(\hat{\lambda}^m; x^n)) - L_t^0(a_t^0) \Big) \\
&\qquad = \operatorname{plim} n^{-1} \sum_{t=1}^{n} \Bigg(\Big[\mathscr{L}(\kappa(Y_{t,h}), a_t(\hat{\lambda}^m; x^n)) - E_G(\mathscr{L}(\kappa(X_{t,h}), a_t(\hat{\lambda}^m; x^n) | X^t)) \Big] \\
&\qquad\qquad + \Big[E_G(\mathscr{L}(\kappa(X_{t,h}), a_t(\hat{\lambda}^m; x^n) | X^t)) - L_t^0(a_t^0) \Big] \Bigg).
\end{aligned}$$

By the Law of Large Numbers for martingale difference sequences (given the Uniform Law of Large Numbers assumption in the theorem), the first bracketed term converges in

probability to zero:

$$\text{plim}\, n^{-1} \sum_{t=1}^{n} \left(\mathscr{L}(\kappa(Y_{t,h}), a_t(\hat{\lambda}^m; x^n)) - E_G(\mathscr{L}(\kappa(X_{t,h}), a_t(\hat{\lambda}^m; x^n)|X^t)) \right) = 0.$$

Therefore, the quantity being minimized simplifies to:

$$\text{plim}\, n^{-1} \sum_{t=1}^{n} \left(E_G(\mathscr{L}(\kappa(X_{t,h}), a_t(\hat{\lambda}^m; x^n)|X^t)) - L_t^0(a_t^0) \right),$$

which, by definition of $L_t^0(\cdot)$ in Equation (5), is precisely

$$\text{plim}\, n^{-1} \sum_{t=1}^{n} \left(L_t^0(a_t(\hat{\lambda}^m; x^n)) - L_t^0(a_t^0) \right) \equiv \text{plim}_{n\to\infty} \Pi_n(a^0 : a^m).$$

For the second part of the theorem, we want to show that if model m^* is correctly specified, $\text{plim}\hat{L}_n^{m^*}(Y^n) = \lim_{n\to\infty} n^{-1} \sum_{t=1}^{n} L_t^0(a_t^0)$. By definition,

$$\text{plim}\, \hat{L}_n^{m^*}(Y^n) = \text{plim}\, n^{-1} \sum_{t=1}^{n} \mathscr{L}(\kappa(Y_{t,h}), a_t(\hat{\lambda}^{m^*}; x^n)).$$

Given the consistency of $\hat{\lambda}^{m^*}$ and the assumed continuity, this converges to

$$E_G[\mathscr{L}(\kappa(Y_{t,h}), a_t^{*m^*})],$$

where $a_t^{*m^*} \equiv \text{plim}\, a_t(\hat{\lambda}^{m^*}; x^n)$.

Since model m^* is correctly specified for the decision-maker's purpose, Definition 2.2 implies that $L_t(\theta^{*m^*}, a_t) = L_t^0(a_t)$ for all a_t and t. By the definition of optimal actions, $a_t^{*m^*} = \arg\min_a L_t(\theta^{*m^*}, a_t)$ and $a_t^0 = \arg\min_a L_t^0(a_t)$. Therefore, if $L_t(\theta^{*m^*}, a_t) = L_t^0(a_t)$, it must follow that $a_t^{*m^*} = a_t^0$. Substituting $a_t^{*m^*} = a_t^0$ and using the definition $L_t^0(a_t) \equiv E_G(\mathscr{L}(\kappa(X_{t,h}), a_t)|X^t)$, we get:

$$\text{plim}\, \hat{L}_n^{m^*} = E_G[\mathscr{L}(\kappa(Y_{t,h}), a_t^{*m^*})] = E_G[\mathscr{L}(\kappa(Y_{t,h}), a_t^0)] = \lim_{n\to\infty} n^{-1} \sum_{t=1}^{n} L_t^0(a_t^0).$$

□

Proof of Theorem 2.4 — The weight associated with decision $a_n(\hat{\lambda}^m; x^n)$ is given by $\bar{\omega}_n^m(x^n) \equiv \omega_n^m(x^n)/\sum_j \omega_n^j(x^n)$. Since plim $(\hat{L}_n^m(x^n) - \hat{L}_n^{m^*}(x^n))/\hat{\sigma}_n^{m*} = (\mu^m - \mu^{m^*})/\sigma^{m*}$, it follows from Equation (18) that $\bar{\omega}_n^m(x^n) \xrightarrow{p} 0$ if $\mu^m - \mu^{m^*} > 0$.

For the second part of the theorem, the convexity of $\mathscr{L}(\kappa(Y_{t,h}), \cdot)$ implies that:

$$\begin{aligned}
&\lim_{n\to\infty} (\Pi_n(a^0 : \bar{a}(x^n)) - \Pi_n(a^0 : a_n(\hat{\lambda}^m; x^n))) \\
&= \lim_{n\to\infty} (n-h)^{-1} \sum_{t=1}^{n-h} (L_t^0(\bar{a}_t(x^n)) - L_t^0(a_n(\hat{\lambda}^m; x^n))) \\
&\leq \lim_{n\to\infty} (n-h)^{-1} \sum_{t=1}^{n-h} \left(\sum_j \bar{\omega}_n^j(x^n) L_t^0(a_n(\hat{\lambda}^j; x^n)) - L_t^0(a_n(\hat{\lambda}^m; x^n)) \right).
\end{aligned}$$

Let $\mathscr{J} = \{j \in \mathscr{M} \mid \mu^j = \mu^{m^*}\}$ be the set of models with the lowest asymptotic loss. From the first part of the theorem, for any $j \notin \mathscr{J}$, we have $\bar{\omega}_n^j(x^n) \xrightarrow{p} 0$. For $j \in \mathscr{J}$, under standard assumptions on the weighting scheme, plim$\bar{\omega}_n^j(x^n) = \frac{1}{|\mathscr{J}|}$.

Since the summation over j is finite, assuming $(n-h)^{-1} \sum_{t=1}^{n-h} L_t^0(a_n(\hat{\lambda}^j; x^n))$ converges to μ^j in probability, we can interchange the limit and summation, and apply the product rule for probability limits:

$$\begin{aligned}
\lim_{n\to\infty} (n-h)^{-1} \sum_{t=1}^{n-h} \left(\sum_j \bar{\omega}_n^j L_t^0(a_n(\hat{\lambda}^j; x^n)) \right) &= \sum_j \left(\text{plim}_{n\to\infty} \bar{\omega}_n^j(x^n) \right) \\
&\qquad \left(\lim_{n\to\infty} (n-h)^{-1} \sum_{t=1}^{n-h} L_t^0(a_n(\hat{\lambda}^j; x^n)) \right) \\
&= \sum_j \bar{\omega}^j(x^n) \mu^j,
\end{aligned}$$

where $\bar{\omega}^j(x^n) = 1/|\mathscr{J}|$ for $j \in \mathscr{J}$ and $\bar{\omega}^j(x^n) = 0$ for $j \notin \mathscr{J}$.

Thus, the expression simplifies to:

$$\frac{1}{|\mathscr{J}|} \sum_{j \in \mathscr{J}} \mu^j - \mu^m.$$

Since $\mu^j = \mu^{m^*}$ for all $j \in \mathscr{J}$, this further simplifies to:

$$\mu^{m^*} - \mu^m.$$

Since m^* minimizes μ^m, it implies $\mu^{m^*} \leq \mu^m$ for all $m \in \mathscr{M}$. Therefore, $\mu^{m^*} - \mu^m \leq 0$, which proves $\lim_{n\to\infty}(\Pi_n(a^0 : \bar{a}(x^n)) - \Pi_n(a^0 : a_n(\hat{\lambda}^m; x^n))) \leq 0$, or $\lim_{n\to\infty} \Pi_n(a^0 : \bar{a}(x^n)) \leq \lim_{n\to\infty} \Pi_n(a^0 : a_n(\hat{\lambda}^m; x^n))$. □

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics 21*, 243–247.

Davidson, J. E. H., D. F. Hendry, F. Srba, and S. Yeo (1978). Econometric modelling of the aggregate time-serie relationship between consumers' expenditure and income in the united kingdom. *Economic Journal 88*, 661–692.

Diebold, F. and R. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics 13*, 253–263.

Diebold, F. X. (2004). *Elements of Forecasting* (3 ed.). Thomson South-Western, Mason, Ohio.

Diron, M. and B. Mojon (2008). Are inflation targets good inflation forecasts? *Economic Perspectives, Federal Reserve Bank of Chicago Q II*, 33–45.

Dufour, J.-M. (2008). Model selection. *The New Palgrave Dictionary of Economics (Online Edition)*, 1–7.

European Central Bank (2021, July). The ECB's Monetary Policy Strategy Statement. `https://www.ecb.europa.eu/mopo/strategy/strategy-review/ecb.strategyreview202506_strategy_statement.en.html`.

European Central Bank (2025, June). ECB's Governing Council updates its monetary policy strategy. https://www.ecb.europa.eu/press/pr/date/2025/html/ecb.pr250630~aabf988af8.en.html.

Faust, J. and J. Wright (2013). Forecasting inflation. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Number 2A. Elsevier.

Hannan, E. J. and B. Quinn (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B 41*, 190–191.

Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural Networks 2*(5), 359–366.

Hume, D. (1739). *A Treatise of Human Nature*. London: John Noon.

Kay, J. A. and M. A. King (2020). *Radical Uncertainty: Decision-making for an Unknowable Future*. London: Bridge Street Press.

Knight, F. H. (1921). *Risk, Uncertainty, and Profit*. Boston: Houghton Mifflin.

Lütkepohl, H. (2005). *The New Introduction to Multiple Time Series Analysis*. Springer-Verlag Berlin.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag.

Leeb, H. and B. Pötscher (2005). Model selection and inference: facts and fiction. *Econometric Theory 75*(21), 29–59.

MacKinnon, J. G. (1992). Model specification tests and artificial regressions. *Journal of Economic Literature 30*, 102–146.

Manganelli, S. (2023). Double conditioning: the hidden connection between Bayesian and classical statistics. *European Central Bank 2786*, 1–24.

Manganelli, S. (2025). Statistical decision functions with judgment. *Journal of Economic Theory 223*, 105940.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995). *Microeconomic Theory*. Oxford University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Wald, A. (1950). *Statistical Decision Functions*. New York, John Wiley & Sons.

White, H. (1996). *Estimation, Inference and Specification Analysis*. Cambridge Books, Cambridge University Press.

B Online Appendix — Technical Derivations for the VAR Model

NOTATION: For vector and matrix differentiation we follow the conventions of Lütkepohl (2005), except that to simplify notation we use the symbol ∇ to indicate derivatives.

In particular, if $f(\theta)$ is a scalar function that depends on the $(m \times 1)$ vector $\theta = (\theta_1, \cdots, \theta_m)'$, we define the following derivatives:

$$\nabla f \equiv \frac{\partial f}{\partial \theta} \equiv \begin{bmatrix} \frac{\partial f}{\partial \theta_1} \\ \vdots \\ \frac{\partial f}{\partial \theta_m} \end{bmatrix} \qquad \nabla' f \equiv \frac{\partial f}{\partial \theta'} \equiv \begin{bmatrix} \frac{\partial f}{\partial \theta_1}, \cdots, \frac{\partial f}{\partial \theta_m} \end{bmatrix}$$

and

$$\nabla^2 f \equiv \frac{\partial^2 f}{\partial \theta \partial \theta'} \equiv \left[\frac{\partial^2 f}{\partial \theta_i \partial \theta_j} \right] = \begin{bmatrix} \frac{\partial^2 f}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \theta_m \partial \theta_1} & \cdots & \frac{\partial^2 f}{\partial \theta_m \partial \theta_m} \end{bmatrix}.$$

Application of the rules for matrix differentiation follows Appendix A.13 of Lütkepohl (2005).

We follow closely the setup of chapter 5 of Lütkepohl (2005). Consider the following general specification of a VAR(p) model:

$$\underset{(g\times 1)}{y_t} = \underset{(g\times 1)}{\nu} + \underset{(g\times g)}{C_1} y_{t-1} + ... + C_p y_{t-p} + \underset{(g\times 1)}{u_t} \qquad u_t \sim i.i.d.(0, \underset{(g\times g)}{\Sigma_u}), \tag{29}$$

for $t = 1, ..., n$. The model can be rewritten as:

$$\underset{(g\times n)}{Y} = \underset{(g\times 1+gp)}{C} \underset{(1+gp\times n)}{Z} + \underset{(g\times n)}{U}, \tag{30}$$

where

$$Y \equiv [y_1, ..., y_n] \qquad C \equiv [\nu, C_1, ..., C_p]$$
$$\underset{(1+gp\times 1)}{Z_t} \equiv [1, y'_{t-1}, ..., y'_{t-p}]' \qquad Z \equiv [Z_1, ..., Z_n]$$
$$U \equiv [u_1, ..., u_n].$$

Any subset VAR model of the general model above can be obtained by imposing appropriate zero constraints on the coefficients:

$$\beta \equiv \text{vec}(C) = \underset{(g(1+gp)\times q)(q\times 1)}{R}\ \theta, \tag{31}$$

where q is the number of regressors (including constants) in the subset VAR model under consideration.

The matrix R can be easily constructed as follows. Let $W \in \mathbb{N}^q$ denote a vector indicating the position of θ relative to the full vector of coefficients β. Then the matrix R can be constructed from a $(g(1 + gp) \times q)$ matrix of zeros with 1 in the $(W(i), i)$ position, for $i = 1, ..., q$.

The model can be further rewritten as:

$$\begin{aligned} \underset{(gn\times 1)}{y} &\equiv \text{vec}(Y) \\ &= (Z' \otimes I_g)\text{vec}(C) + \text{vec}(U) \\ &\equiv (Z' \otimes I_g)R\theta + u. \end{aligned} \tag{32}$$

B.1 The Quasi-Likelihood Function and Its First and Second Derivatives

Define

$$f_t(X^t, \theta) \equiv \exp(-0.5 u_t' u_t), \tag{33}$$

where $u_t = y_t - (Z_t' \otimes I_g) R\theta$, so that the quasi log-likelihood function is

$$\ell_n(X^n, \theta) = -0.5 n^{-1} u' u. \tag{34}$$

Solving the first order conditions with respect to θ gives the QMLE:

$$\hat{\theta} = (R'(ZZ' \otimes I_g)R)^{-1} R'(Z \otimes I_g) y. \tag{35}$$

Chapter 5 of Lütkepohl (2005) discusses alternative strategies to estimate θ, including full maximum likelihood estimation.

To obtain the QMLE variance-covariance matrix estimate, the first and second derivatives of the elements of the log-likelihood are needed.

The first derivative is:

$$\nabla \log f_t(X^t, \theta) = R'(Z_t \otimes I_g) u_t \tag{36}$$

and its transpose is

$$\nabla' \log f_t(X^t, \theta) = u_t'(Z_t' \otimes I_g) R. \tag{37}$$

The second derivative is

$$\nabla^2 \log f_t(X^t, \theta) = -R'(Z_t Z_t' \otimes I_g) R. \tag{38}$$

Therefore

$$\begin{aligned}\hat{D}_n &= -n^{-1}\sum_{t=1}^{n} R'(Z_t Z_t' \otimes I_g)R \\ &= -n^{-1}R'((\sum_{t=1}^{n} Z_t Z_t') \otimes I_g)R \\ &= -n^{-1}R'(ZZ' \otimes I_g)R \end{aligned} \tag{39}$$

and

$$\hat{B}_n = n^{-1}\sum_{t=1}^{n} R'(Z_t \otimes I_g)u_t u_t'(Z_t' \otimes I_g)R.$$

B.2 Forecasting

The objective is to forecast the average inflation h periods ahead:

$$\hat{y}_{n,h} = h^{-1}\sum_{i=1}^{h} \hat{y}_{n+i}. \tag{40}$$

Following closely Lütkepohl (2005), page 96:

$$\hat{y}_{n+i} = J\bar{C}^i Z_n, \tag{41}$$

where

$$\underset{(1+gp\times 1+gp)}{\bar{C}} \equiv \begin{bmatrix} 1 & \mathbf{0}_{1,g} & \mathbf{0}_{1,g} & \cdots & \mathbf{0}_{1,g} & \mathbf{0}_{1,g} \\ \nu & C_1 & C_2 & \cdots & C_{p-1} & C_p \\ \mathbf{0}_{g,1} & I_g & \mathbf{0}_{g,g} & \cdots & \mathbf{0}_{g,g} & \mathbf{0}_{g,g} \\ \mathbf{0}_{g,1} & \mathbf{0}_{g,g} & I_g & \cdots & \mathbf{0}_{g,g} & \mathbf{0}_{g,g} \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \mathbf{0}_{g,1} & \mathbf{0}_{g,g} & \mathbf{0}_{g,g} & \cdots & I_g & \mathbf{0}_{g,g} \end{bmatrix}, \tag{42}$$

which can be more compactly rewritten as:

$$\underset{(1+gp\times 1+gp)}{\bar{C}} = \left[\begin{array}{c:c:c} 1 & \mathbf{0}_{1,g(p-1)} & \mathbf{0}_{1,g} \\ \hdashline & C & \\ \hdashline \mathbf{0}_{g(p-1),1} & I_{g(p-1)} & \mathbf{0}_{g(p-1),g} \end{array}\right],$$

and

$$\underset{(g\times 1+gp)}{J} \equiv [\underset{(g\times 1)}{\mathbf{0}}, I_g, \underset{(g\times g(p-1))}{\mathbf{0}}]. \tag{43}$$

Consider the following mean value expansion:

$$\begin{aligned} \hat{y}_{n,h} &\equiv y_{n,h}(\hat{\beta}) \\ &= y_{n,h}(\beta) + \nabla_\beta y_{n,h}(\bar{\beta})(\hat{\beta}-\beta) \\ &= y_{n,h}(R\theta) + \nabla_\beta y_{n,h}(R\bar{\theta})R(\hat{\theta}-\theta), \end{aligned}$$

where $\bar{\theta}$ is a mean value which lies between θ and $\hat{\theta}$ and we have imposed the constraint (31). It follows that:

$$\sqrt{n}(\hat{y}_{n,h} - y_{n,h}) \sim N(0, \nabla_\beta y_{n,h} R D^{*-1} B^* D^{*-1} R' \nabla'_\beta y_{n,h}). \tag{44}$$

It remains to compute $\nabla_\beta y_{n,h} = h^{-1}\sum_{i=1}^{h} \nabla_\beta y_{n+i}$:

$$\nabla_\beta y_{n+i} = \sum_{j=0}^{i-1} Z'_n (\bar{C}')^{i-1-j} \otimes J\bar{C}^j J'. \tag{45}$$

B.3 Constructing the Decision with Judgment

The decision with judgment is constructed in two steps. First, one tests whether the gradient of the loss function evaluated at the judgmental decision is statistically different from zero. Second, if the test does not reject, the judgmental decision is retained as optimal

one, and if the test rejects, the optimal decision is at the closest boundary of the confidence interval.

To construct the test, notice first that the empirical gradient for the quadratic loss function is:

$$\nabla_a L_n(\theta(X^n), \tilde{a}_n) = -e_1' y_{n,h}(\theta(X^n)) + \tilde{a}_n, \tag{46}$$

where $e_1 \equiv [1, 0, \ldots, 0]'$ is a g vector of zeros with 1 in the first position, $y_{n,h}(\theta(X^n)) \equiv \hat{y}_{n,h}$, and X^n represents the information set over which the parameter θ is estimated, following the convention of Section 2. As usual, random variables depend on X^n and their realizations on x^n.

Exploiting the result (44) of the previous subsection, under the null $H_0 : -e_1 y_{n,h}(\theta^*) + \tilde{a}_n = 0$:

$$\sqrt{n}\sigma^{-1}(-e_1 y_{n,h}(\theta(X^n)) + \tilde{a}_n) \sim N(0, 1), \tag{47}$$

where $\sigma^2 \equiv e_1 \nabla_\beta y_{n,h} R D^{*-1} B^* D^{*-1} R' \nabla_\beta' y_{n,h} e_1'$ can be consistently estimated with standard plug-in estimators.

Denoting with c_α the α percentile of the standard normal and defining $T(x^n) \equiv \sqrt{n}\sigma^{-1}$ $(-e_1 y_{n,h}(\theta(x^n)) + \tilde{a}_n)$, the optimal decision of Equation (8) can be rewritten as:

$$\begin{aligned} a_n(\hat{\lambda}^m; x^n) =& \tilde{a}_n I(c_{\alpha/2} \leq T(x^n) \leq c_{1-\alpha/2}) + \\ &+ (e_1 y_{n,h}(\theta(x^n)) + \hat{\sigma} c_{\alpha/2}/\sqrt{n}) I(T(x^n) < c_{\alpha/2}) + \\ &+ (e_1 y_{n,h}(\theta(x^n)) + \hat{\sigma} c_{1-\alpha/2}/\sqrt{n}) I(T(x^n) > c_{1-\alpha/2}). \end{aligned} \tag{48}$$

B.4 Model Selection

Implementation of the model selection procedure requires the construction of the in-sample average loss for each model m:

$$\hat{L}_n^m(x^n) = 0.5(n-h)^{-1}\sum_{t=1}^{n-h}(h^{-1}\sum_{i=1}^{h} e_1 x_{t+i} - a_t(\hat{\lambda}^m; x^n))^2. \tag{49}$$

This in turn requires the construction of the in-sample optimal decision $a_t(\hat{\lambda}^m; x^n)$. The only missing element to construct $a_t(\hat{\lambda}^m; x^n)$ from (48) is $\tilde{a}_t$, that is the equivalent at time t of the judgmental decision expressed at time n. This can be reconstructed using the estimated coefficients from the following constrained optimization problem:

$$\begin{aligned} &\max_{\theta} \ell_n(x^n, \theta) \\ &s.t. \quad y_{n,h}(\theta(x^n)) = \tilde{a}_n, \end{aligned} \tag{50}$$

where $\ell_n(x^n, \theta)$ was defined in (34). Denoting with $\tilde{\theta}(x^n)$ the constrained parameter vector:

$$\tilde{a}_t = y_{t,h}(\tilde{\theta}(x^n)). \tag{51}$$

Denote with $\mathscr{M}$ the set of models considered in the selection procedure and with $m^* \in \mathscr{M}$ the model with lowest in-sample loss $\hat{L}_n^m(x^n)$.

The last and final step is the construction of the weights for model averaging. The asymptotic variance of $\hat{L}_n^{m^*}(x^n)$ can be estimated as:

$$(\hat{\sigma}^{m^*})^2 = (n-h)^{-1}\sum_{t=1}^{n-h}\left((h^{-1}\sum_{i=1}^{h} e_1 x_{t+i} - a_t(\hat{\lambda}^{m^*}; x^n))^2 - \hat{L}_n^{m^*}(x^n)\right)^2. \tag{52}$$

It is now possible to construct the model weights $\omega_n^m(x^n)$ given in (18) and the model averaging decision rule $\bar{a}(x^n)$ given in (19).

Table 1: Summary Statistics

	HICP	HICPX	UN	IP	IR
Mean	2.1	1.6	8.9	99.9	1.7
Min	-0.6	0.2	6.2	70.9	-0.6
Max	10.6	5.7	12.2	112.8	5.1
n	317				

Note: Summary statistics for the time series (January 1999 - May 2025) used in the forecasting models. HICP is the headline inflation, HICPX is core inflation, UN is the unemployment rate, IP is the industrial production, and IR is the 3-month euro area interest rate. n is the length of the time series. See the notes of the figures for more details.

Source: Eurostat, http://sdw.ecb.europa.eu

C Online Appendix — Data

The five variables inlcuded in the VAR estimation are: euro area inflation (HICP), euro area HICP excluding energy and unprocessed food (HICPX), unemployment (U), industrial production (IP), and the 3-month euro area interest rate (IR). All data are monthly, ranging from January 1999 (when the euro was introduced) to May 2025.

Key summary statistics of the five variables are reported in Table 1. The time series are reported in Figure 6.

Figure 6: Data

Note: Time series plots of inflation (HICP), core inflation (HICPX), unemployment (UN), industrial production (IP) and 3-month interest rate (I) in the euro area, from January 1999 until May 2025.
HICP: Overall index, annual rate of change, euro area (changing composition), neither seasonally nor working day adjusted.
Core Inflation: All-items excluding energy and unprocessed food, annual rate of change, euro area (changing composition), neither seasonally nor working day adjusted.
Unemployment: Standardized unemployment, rate, euro area (changing composition), total (all ages), total (male and female), seasonally adjusted, not working day adjusted, percentage of civilian workforce.
Industrial Production: Industrial production index, total industry, euro area 20 (fixed composition), working day and seasonally adjusted.
Interest Rate: 3-month interest rates for the euro area (19 countries).

Source: Eurostat, http://sdw.ecb.europa.eu.