Selecting Models with Judgment

Simone Manganelli*
European Central Bank
simone.manganelli@ecb.int

June, 2018

Abstract

A statistical decision rule incorporating judgment does not perform worse than a judgmental decision with a given probability. Under model misspecification, this probability is unknown. The best model is the least misspecified, as it is the one whose probability of underperforming the judgmental decision is closest to the chosen probability. It is identified by the statistical decision rule incorporating judgment with lowest in sample loss. Averaging decision rules according to their asymptotic performance results in decisions which are weakly better than the best decision rule. The model selection criterion is applied to a vector autoregression model for euro area inflation.

Keywords: Statistical Decision Theory; Model Selection Criteria; Inflation Forecasting.
JEL Codes: C1; C11; C12; C13.

*I would like to thank for useful comments and suggestions Igor Custodio Joao and Gabriela Miyazato Szini. The views expressed are mine and do not necessarily reflect those of the European Central Bank.

1 Introduction

It is often the case that people take decisions by informally combining personal experience, limited amount of information and beliefs about the likelihood of uncertain events. I refer to these decisions as *judgmental decisions*. This paper is concerned with the question of how an econometrician can help decision makers take better decisions. In Manganelli (2018) I have addressed this question under the assumption that the econometrician can estimate a correctly specified econometric model. This paper explores the implications of relaxing the assumption of correct specification.

According to the framework developed by Wald (1950), for a given sequence of random variables, a statistical decision problem is formally defined by three essential ingredients: a probability distribution, a decision rule and a loss function. The probability distribution fully describes the stochastic process generating the observed realizations of the random variables. It is unknown and the econometrician tries to approximate it using statistical parameters. The decision rule prescribes the decision to take, given the observed realization. The loss function expresses the loss suffered by the decision maker when a decision is taken and the true probability distribution is revealed. The concept of judgment is not explicitly incorporated, although it is implicitly introduced via the prior distribution used to compute Bayes risk. In Manganelli (2018) I show that once judgment is explicitly introduced, it provides a unifying framework for statistics of which the Bayesian and classical approaches become special cases.

I define *judgment* as a pair formed by a *judgmental decision* and a *confidence level*. The judgmental decision should be the starting point of any statistical analysis, as it represents what the decision maker would choose without statistical support. The fundamental goal of statistics is to provide a decision rule which improves upon the judgmental decision, in the sense that it provides a lower loss. Given the lack of knowledge of the true probability distribution, the decision rule can improve upon the judgmental decision only in a statistical sense. Under correct model specification, the

statistical decision rule provided by Manganelli (2018) performs worse than the judgmental decision with a probability which is bounded above by the confidence level. This statistical decision rule prescribes to take the decision identified by the closest bound of the confidence interval, which amounts to a shrinkage estimation. The confidence level, which I also refer to as *statistical risk aversion*, reflects the attitude of the decision maker towards uncertainty. An extreme statistical risk averse person (characterized by a confidence level of 0%) would never engage in statistical decision making. At the other extreme, a statistical risk loving person (characterized by a confidence level of 100%) would ignore any risk that the statistical decision rule might underperform the judgmental decision. This is what happens with standard classical plug-in estimators, which ignore any estimation risk. Bayesian priors are associated with a time varying, sample dependent statistical risk aversion, a fact which also seems to be at odds with actual behavior.

This paper starts from here and pushes forward in several directions. First, it draws from the quasi-maximum likelihood theory developed by White (1996) to approximate the distribution of the estimators, which in turn is needed to construct the confidence interval. White's theory is quite general and applies to the most commonly used models in econometrics. The main advantage of White's framework is that it provides a coherent inference theory even in the presence of model misspecification. Casting the statistical problem in terms of decision theory reveals that in the presence of misspecification the probability that the statistical decision rule performs worse than the judgmental decision becomes unknown and it is therefore no longer determined by the confidence level supplied by the decision maker. If one takes the view that any model is inherently misspecified, this analysis reveals that there is an irreducible level of uncertainty about the probability with which any given statistical model performs worse than the judgmental decision. Decision makers engaging in statistical decision making need to accept this additional layer of *uncertainty about the level of uncertainty*.

The second major contribution of this paper is to provide a consistent

model selection criterion which asymptotically identifies the least misspecified model. The best model asymptotically minimizes a Kullback-Leibler-type of distance between the in sample loss associated with the model decision rule and the optimal, but unknown decision. If the confidence level is equal to 100% and the loss function is quadratic, this selection criterion is equivalent to minimizing in sample mean squared error. When the models to choose from are nested, it will by construction select the largest model. This is due to the fact that a decision maker with confidence level equal to 100% does not care about the probability of committing statistical error, that is of selecting decisions which may perform worse than the judgmental decision. When the confidence level is less than 100%, however, this is no longer necessarily the case, as the largest model may be characterized by larger confidence intervals and be associated with suboptimal decision rules. I also show how averaging decision rules according to their asymptotic performance results in decisions which are asymptotically weakly better than the single best decision rule.

I apply the model selection criterion to the problem of forecasting inflation in the euro area. I use as judgmental forecast the European Central Bank's definition of price stability, that is an average inflation rate of close but below 2% over the medium term, a confidence level of 10% and a quadratic loss function. The models are selected from a vector autoregression with four endogenous variables (inflation, core inflation, unemployment and industrial production) and 12 lags. Given the extremely large number of possible combinations associated with this class of models, a complete grid search of the model with lowest in sample loss is not feasible. I show, instead, how the problem can be cast in terms of an integer optimization program, which can be solved using available optimization algorithms. I find that the best forecasting model for the euro area inflation includes all variables with different and long lags. In particular, the best inflation forecast at a two year horizon is 1.9%. Forecasts based on a simple autoregressive process reveal significant evidence of model misspecification, especially for forecasts at longer horizons. I also find that core inflation does not help much forecasting at longer hori-

zons, unlike unemployment, suggesting that Phillip's curve type mechanisms may be at work within the euro area.

The rest of the paper is structured as follows. Section 2 develops the theory of model selection. Section 3 describes in detail the various elements needed to implement the forecasting process. Section 4 reports the results for the euro area inflation forecast. Section 5 concludes.

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 Asymptotic approximation to the data generation process

The observations at time t are $x_t \in \mathbb{R}^v$, $v \in \mathbb{N}$, where $x_t = (w_t', y_t')'$, y_t is a $l \times 1$ vector of dependent variables, and w_t is a $v - l \times 1$ vector of potential explanatory variables. The history of observations available at time n is $x^n = (x_1', \ldots, x_n')'$. The observed data x_t are assumed to be a realization of a stochastic process with c.d.f. F_t, so that $F_t(x_t^1, \ldots, x_t^v) = P(X_t^1 < x_t^1, \ldots, X_t^v < x_t^v | x^{t-1})$, $t = 1, 2, \ldots$

Define the quasi-log-likelihood function (in the sense of White, 1996) $\ell_n(X^n, \theta) \equiv n^{-1} \sum_{t=1}^n \log f_t(X^t, \theta)$, where $f_t(\cdot, \theta) : \mathbb{R}^{vt} \to \mathbb{R}^+$ is a suitably chosen function and $\theta \in \mathbb{R}^p$, $p \in \mathbb{N}$, is a finite dimensional parameter vector. $\theta(X^n) = \arg\max_\theta \ell_n(X^n, \theta)$ is the quasi-maximum likelihood estimator (QMLE).[1] For simplicity, I impose the stationarity assumption $E(\log f_t(X^t, \theta)) = c$, for $c \in \mathbb{R}$ constant, where the expectation is taken

[1] I use the notation $\theta(X^n)$ to denote the estimator and $\theta(x^n)$ for the estimate.

with respect to the true data generation process. This assumption rules out heterogeneity in the expectations of the log-likelihood function, so that the QMLE converges to a fixed vector θ^* which does not depend on n.[2] Assuming also that the conditions for consistency, asymptotic normality and consistent asymptotic variance estimation are satisfied (see Theorems 3.5, 6.4 and Corollary 8.28 of White, 1996) gives:

$$B^{*-1/2}A^*\sqrt{n}(\theta(X^n) - \theta^*) \overset{A}{\sim} N(0, I_p)$$

where $A^* \equiv E(\nabla^2 \ell_n(X^n, \theta^*))$, $B^* \equiv \text{var}(\sqrt{n}\nabla \ell_n(X^n, \theta^*))$ and I_p is the identity matrix of dimension p. The asymptotic covariance matrix $A^{*-1}B^*A^{*-1}$ is consistently estimated by $\hat{A}_n^{-1}\hat{B}_n\hat{A}_n^{-1}$, where:

$$\hat{A}_n \equiv n^{-1}\sum_{t=1}^{n} \nabla^2 \log f_t(X^t, \theta(X^n))$$
$$\hat{B}_n \equiv n^{-1}\sum_{t=1}^{n} \nabla \log f_t(X^t, \theta(X^n))\nabla' \log f_t(X^t, \theta(X^n))$$

2.2 Loss function and judgment

Consider a decision maker with the following loss function:

$$L_t(a_t|F_t) \equiv E(\mathscr{L}(Y_{t,h}, a_t)|X^t) \tag{1}$$

where $a_t \in \mathbb{R}^q$, $q \in \mathbb{N}$, is the action chosen at time t by the decision maker, $Y_{t,h} \equiv (Y'_{t+1}, \ldots, Y'_{t+h})'$, $h = 1, 2, \ldots$ is the horizon of interest, and $\mathscr{L}(Y_{t,h}, \cdot)$ is a continuously differentiable and strictly convex function on $\mathbb{R}^q$. I assume that the expectation exists and is finite.

Since F_t is unknown, the best the decision maker can do is to minimize the loss function using the selected parametric specification $f_t(X^t, \theta)$:

$$\begin{aligned} L_t(\theta^*, a_t) &\equiv E_\theta(\mathscr{L}(Y_{t,h}, a_t)|X^t) \\ &\equiv \int \mathscr{L}(y, a_t) f_{t,h}(y, \theta) dy \end{aligned} \tag{2}$$

[2] See White (1996) for the more general treatment.

where $f_{t,h}$ is the suitable h-step ahead conditional density of $Y_{t,h}$ constructed from $f_t(X^t, \theta)$, and is such that the expectation exists and is finite. Adapting from White (1996), $f_t(X^t, \theta)$ is said to be correctly specified for the purposes of the decision maker if θ^* is such that $L_t(a_t|F_t) = L_t(\theta^*, a_t)$, $\forall a_t$, $t = 1, 2, \ldots$, and misspecified otherwise.

Remark: Direct estimation — An alternative estimation strategy is to choose a parametric specification for the decision $a_t(X^t, \theta)$ and set $f_t(X^t, \theta) = \exp(\mathscr{L}(Y_{t,h}, a_t(X^t, \theta)))$.[3] When f_t is chosen as the p.d.f. of a normal distribution and $\mathscr{L}$ is quadratic the two estimation strategies coincide. In general, they generate different QMLE and different decision rules. They should therefore be treated as alternative model specifications, to be selected according to the criterion presented later in this section. □

Continuity, differentiability and convexity of the function $\mathscr{L}$ guarantee that a necessary and sufficient condition for optimality according to (2) is:

$$\begin{aligned} \nabla_a L_t(\theta^*, a_t) &= E_\theta(\nabla_a \mathscr{L}(Y_{t,h}, a_t)|X^t) \\ &= 0 \end{aligned} \tag{3}$$

assuming again that the expectation exists and is finite.

The decision maker deciding at time n has judgment $\{\tilde{a}_n, \alpha\}$ as defined in Manganelli (2018), where $\tilde{a}_n \in \mathbb{R}^q$ is referred to as *judgmental decision* and $\alpha \in [0, 1]$ as the *confidence level*. The judgmental decision $\tilde{a}_n$ implies a specific constraint on the model parameters θ. The null hypothesis that $\tilde{a}_n$ is optimal according to (2) can be expressed by imposing that the first order condition (3) holds:

$$H_0 : \nabla_a L_n(\theta^*, \tilde{a}_n) = 0 \tag{4}$$

This hypothesis can be tested using a Wald, likelihood ratio (LR) or Lagrange

[3]For a general treatment, see section 5.2 of White (1996).

multiplier (LM) test statistic (Theorem 8.10 of White, 1996):

$$\begin{aligned}
\mathcal{W}_n(X^n) &\equiv n\hat{k}_n'(\hat{K}_n\hat{A}_n^{-1}\hat{B}_n\hat{A}_n^{-1}\hat{K}_n')^{-1}\hat{k}_n \overset{A}{\sim} \chi_q^2 \\
\mathcal{LR}_n(X^n) &\equiv -2n(\tilde{\ell}_n - \hat{\ell}_n) \overset{A}{\sim} \chi_q^2 \\
\mathcal{LM}_n(X^n) &\equiv n\nabla'\tilde{\ell}_n(\nabla^2\tilde{\ell}_n)^{-1}\tilde{K}_n'(\tilde{K}_n\hat{A}_n^{-1}\hat{B}_n\hat{A}_n^{-1}\tilde{K}_n')^{-1}\tilde{K}_n(\nabla^2\tilde{\ell}_n)^{-1}\nabla\tilde{\ell}_n \\
&\overset{A}{\sim} \chi_q^2
\end{aligned} \tag{5}$$

where $\hat{k}_n \equiv \nabla_a L_n(\theta(X^n), \tilde{a}_n)$, $\hat{K}_n \equiv \nabla_{a,\theta} L_n(\theta(X^n), \tilde{a}_n)$, $\hat{\ell}_n \equiv \ell_n(X^n, \theta(X^n))$, $\tilde{\ell}_n \equiv \ell_n(X^n, \tilde{\theta}(X^n))$, with $\tilde{\theta}(X^n)$ the constrained QMLE solving the problem $\max_\theta \ell_n(X^n, \theta)$ *s.t.* $\nabla_a L_n(\theta, \tilde{a}_n) = 0$ and $\tilde{K}_n \equiv \nabla_{a,\theta} L_n(\tilde{\theta}(X^n), \tilde{a}_n)$.

2.3 The decision incorporating judgment

The inference apparatus outlined so far is needed to test the optimality of the judgmental decision $\tilde{a}_n$ of the decision maker.

Given the judgment $\{\tilde{a}_n, \alpha\}$, testing the null (4) is equivalent to testing whether the judgmental decision $\tilde{a}_n$ is optimal. If the null is not rejected, statistical evidence is not strong enough to suggest any deviation from $\tilde{a}_n$. Rejection at the confidence level α, however, implies that *marginal* moves away from $\tilde{a}_n$ in the direction of the QMLE decision increase (instead of decreasing) the loss function with probability less than α, the probability of a Type I error (see Manganelli, 2018, for a formal development of this argument). The willingness to take this risk depends on the decision maker's attitude towards uncertainty and is summarized by the confidence level α.

To formalize this reasoning, let's first define the QMLE implied decision:

$$\hat{a}_t = \arg\max_a L_t(\theta(x^n), a) \tag{6}$$

Analogously, it is possible to define the judgmental decision at time t implied by $\tilde{a}_n$ by exploiting the constrained QMLE :

$$\tilde{a}_t = \arg\max_a L_t(\tilde{\theta}(x^n), a) \tag{7}$$

and the associated shrinking action:

$$a_t(\lambda) \equiv \lambda \hat{a}_t + (1-\lambda)\tilde{a}_t, \quad \lambda \in [0,1] \tag{8}$$

Clearly, $\tilde{a}_t = a_t(0)$ and $\hat{a}_t = a_t(1)$, $t = 1, \ldots, n$.

It is important to note that both $\hat{a}_t$ and $\tilde{a}_t$ are not random, as they are defined as a function of the sample realization x^n. $\hat{a}_t$ is the decision that would obtain at time t by minimizing the loss function (2) using standard plug-in estimators. $\tilde{a}_t$ is the equivalent at time t of the original judgment $\tilde{a}_n$.

Defining the Wald statistic (5) in terms of $a_t(\lambda)$ (similar definitions hold for the LR and LM statistics):

$$\mathcal{W}_{t,\lambda}(X^n) \equiv n\hat{k}'_{t,\lambda}(\hat{K}_{t,\lambda}\hat{A}_n^{-1}\hat{B}_n\hat{A}_n^{-1}\hat{K}'_{t,\lambda})^{-1}\hat{k}_{t,\lambda} \tag{9}$$

where $\hat{k}_{t,\lambda} \equiv \nabla_a L_t(\theta(X^n), a_t(\lambda))$ and $\hat{K}_{t,\lambda} \equiv \nabla_{a,\theta} L_t(\theta(X^n), a_t(\lambda))$. Notice that $\mathcal{W}_{n,\lambda}(X^n) \overset{A}{\sim} \chi^2_q$ under the null $H_0 : \nabla_a L_n(\theta^*, a_n(\lambda)) = 0$, for any non random $\lambda \in [0,1]$. Notice also that since the likelihood ratio test $\mathcal{LR}_n(X^n)$ depends on the constraint (4) only via $\tilde{\theta}(X^n)$ and is asymptotically equivalent to $\mathcal{W}_n(X^n)$ and $\mathcal{LM}_n(X^n)$, testing $H_0 : \nabla_a L_n(\theta^*, a_n(0)) = 0$ is equivalent to testing $H_0 : \nabla_a L_t(\theta^*, a_t(0)) = 0$, $t = 1, \ldots, n$.

Letting c_α denote the critical value of χ^2_q, that is $P(Z \leq c_\alpha) = \alpha$ for $Z \sim \chi^2_q$, define the following test function:

$$\psi_\alpha(z) = \begin{cases} 0 & \text{if } z < c_\alpha \\ \gamma & \text{if } z = c_\alpha \qquad 0 \leq \gamma \leq 1 \\ 1 & \text{if } z > c_\alpha \end{cases}$$

The optimal decision can be obtained by applying Theorem 2.1 of Manganelli (2018).

Corollary 2.1 (**Optimal Decision**). *Given the judgment* $\{\tilde{a}_n, \alpha\}$*, if the assumptions of Theorem 8.10 of White (1996) hold, the optimal decision at time* $t = 1, \ldots, n$ *associated with the loss function* (2) *is:*

$$\delta_t(X^n) = a_t(0)[1 - \psi_\alpha(\mathcal{W}_{t,0}(X^n))] + a_t(\hat{\lambda})\psi_\alpha(\mathcal{W}_{t,0}(X^n)) \tag{10}$$

where $\hat{\lambda}$ *is implicitly defined by* $\mathcal{W}_{t,\hat{\lambda}}(x^n) = c_\alpha$.

Proof — See Appendix.

Theorem 2.2 of Manganelli (2018) shows that this decision is admissible. The key element behind the admissibility result is that the decision rule (10) conditions on the observed sample realization x^n via $\hat{\lambda}$.

The decision of Corollary 2.1 coincides with the judgmental decision if there is not enough statistical evidence against it, and shrinks towards the QMLE decision otherwise. The amount of shrinkage is determined by the confidence level α. If $\alpha = 0$, $\psi_0(\mathcal{W}_{t,0}(X^n)) = 0$ for all X^n and $\delta_t(X^n) = \tilde{a}_t$. If $\alpha = 1$, $\psi_1(\mathcal{W}_{t,0}(X^n)) = 1$ for all X^n and $\delta_t(X^n) = \hat{a}_t$. As also pointed out by Manganelli (2009), this decision converges asymptotically to the QMLE decision. However, in finite samples it is characterized by the property that it performs worse than the judgmental decision $\tilde{a}_n$ with probability less than the confidence level α according to the pdf $f_n(X^n, \theta)$ (see Manganelli 2018).

One key difference with respect to the setup of Manganelli (2018) is that the present framework allows for the possibility of misspecification. In particular, the test function $\psi_\alpha(\mathcal{W}_{n,0}(X^n)$ tests the null hypothesis that the judgmental decision $\tilde{a}_n$ is optimal according to (2), rather than (1). Define the optimal action in population according to (1) as:

$$a_t^0 \equiv \arg\min L_t(a_t|F_t) \tag{11}$$

Define also $a_t^* \equiv \arg\min L_t(\theta^*, a_t)$, the optimal action in population according to (2). Under model misspecification, it will generally be the case that $a_n^0 \neq a_n^*$ and therefore the probability of rejecting the null $H_0 : \nabla_a L_n(\tilde{a}_n|F_n) = 0$ when $\tilde{a}_n = a_n^0$ under the chosen parametric specification $f_n(X^n, \theta)$ and critical value c_α is different from α, since $H_0 : \nabla_a L_n(\theta^*, \tilde{a}_n) = c \neq 0$.

The confidence level α reflects the degree of statistical risk aversion of the decision maker, as illustrated by figure 1 of Manganelli (2018): A decision maker is willing to abandon her judgmental decision $\tilde{a}_n$ for a statistical procedure only if the probability that the statistical procedure will result in an action worse than $\tilde{a}_n$ is less or equal than α. Under model misspecification, this guarantee cannot be given. Searching through many model specifications

to find the least misspecified model will help alleviating this problem. However, if one takes the view that any model is inherently misspecified, there will remain an irreducible level of uncertainty about α (which one could refer to as *'uncertainty about the level of uncertainty'*), which any decision maker engaging in statistical decision making has to live with, as it will be made clear in the next subsection.

2.4 Selecting the least misspecified model

Suppose now that the decision maker can choose from M alternative statistical models. Denote with $\theta^m \in \mathbb{R}^{p_m}$, $p_m \in \mathbb{N}$, $m \in \mathscr{M} \equiv \{1, \ldots, M\}$, the parameterization of the alternative models, with $f_t^m(x^t, \theta^m)$ the model m specification and with $\delta_t^m(x^n)$ the related decision incorporating judgment. Define the following distance in the loss space:

$$\Pi_n(a^0 : \delta^m) \equiv n^{-1} \sum_{t=1}^{n} (L_t(\delta_t^m(x^n)|F_t) - L_t(a_t^0|F_t)) \tag{12}$$

where a_t^0 is defined in (11). Like the Kullback-Leibler Information Criterion, $\Pi_n(a^0 : \delta^m) \geq 0$ and is equal to 0 if and only if $\delta_t^m(x^n) = a_t^0$, for all $t = 1, \ldots, n$.

The model with the lowest empirical in sample loss will asymptotically converge to the model minimizing $\Pi_n(a^0 : \delta^m)$, as stated in the following theorem.

Theorem 2.1 (**Consistent Model Selection**). *Assume the conditions of Corollary 2.1 are satisfied and that the process* $\{\mathscr{L}(Y_{t,h}, \cdot)\}$ *obeys the Uniform Law of Large Numbers. Then, if model* $m^* \in \mathscr{M}$ *minimizes* $\text{plim}\hat{L}_n^m$, *where* $\hat{L}_n^m \equiv n^{-1} \sum_{t=1}^n \mathscr{L}(Y_{t,h}, \delta_t^m(x^n))$, *it also minimizes* $\lim_{n\to\infty} \Pi_n(a^0 : \delta^m)$, *for all* $m \in \mathscr{M}$. *If model* m^* *is correctly specified,* $\text{plim}\hat{L}_n^{m^*} = \lim_{n\to\infty} n^{-1} \sum_{t=1}^n L_t(a_t^0|F_t)$.

Proof — See Appendix.

According to theorem 2.1, 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, according to the Kullback-Leibler-type Information Criterion $\Pi_n(a^0 : \delta^m)$.

Here again it is important to pay attention to what is random and what is not in the asymptotic argument. The proof of the theorem makes clear that the loss function $\mathscr{L}(Y_{t,h}, \delta_t^m(x^n))$ depends on the random variable $Y_{t,h}$ for the *given* decision rule $\delta_t^m(x^n)$. The asymptotic thought experiment is to evaluate the given decision rule over repeated draws of $Y_{t,h}$, according to the loss function $\mathscr{L}$. Since we are interested in the performance of $\delta_t^m(x^n)$ for the observed sample realization x^n, it does not make sense to average also other potential decision rules $\delta_t^m(X^n)$ which are not relevant for the current decision problem.

If the confidence level $\alpha = 1$, the forecast horizon $h = 1$ and the loss function is quadratic, this selection criterion is equivalent to the one-step ahead in sample mean squared error. In finite samples, it will by construction select the largest model, in the case all models in $\mathscr{M}$ are nested. This should come as no surprise, since a decision maker with $\alpha = 1$ does not care about the probability that the statistical decision rule might underperform her judgmental decision. However, when $\alpha < 1$, this is not necessarily the case, as the choice is restricted to decision rules which satisfy the condition that $P(L(\theta^{*m}, \delta_n^m(X^n)) > L(\theta^{*m}, a_n^0)) < \alpha$. Among the decision rules satisfying this condition, the one associated with the largest model does not necessarily provide the best in sample fit.

Even though theorem 2.1 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 2.1:

$$\sqrt{n}(\hat{L}_n^{m^*} - \mu^{m^*}) \overset{A}{\sim} N(0, \sigma^2) \tag{13}$$

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

$$\hat{\sigma}_n^2 \equiv n^{-1}\sum_{t=1}^n (\mathscr{L}(Y_{t,h}, \delta_t^{m^*}(x^n)) - \hat{L}_n^{m^*})^2 \tag{14}$$

For any finite sample size, it may therefore happen that $\hat{L}_n^{m^*} < \hat{L}_n^0$ even if $\delta_t^{m^*}(x^n) \neq a_t^0$ for some t. Since for any given finite sample it is impossible to know whether a given value of $\hat{L}_n^m$ is low because of luck or because it represents a draw from the true DGP, it makes sense 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 $m \neq m^*$.

Theorem 2.2 (**Consistent Model Averaging**). *Let*

$$\omega_n^m \equiv \Phi(\sqrt{n}(\hat{L}_n^m - \hat{L}_n^{m^*})/\hat{\sigma}_n) \tag{15}$$

where m^* is the model with lowest $\hat{L}_n^m$, $m \in \mathscr{M}$, and $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. Under the conditions of Theorem 2.1, the decision rule

$$\bar{\delta} \equiv (\sum_{m=1}^M \omega_n^m \delta^m)/\sum_{m=1}^M \omega_n^m \tag{16}$$

converges in probability to the decision rule(s) with lowest μ^m, $m \in \mathscr{M}$. Letting $\mathscr{J} = \{1, \ldots, J\}$ denote the set of models associated with such decision rules, the asymptotic weight of decision rule δ^i, $i \in \mathscr{M} \setminus \mathscr{J}$, is 0.

Furthermore, $\lim_{n\to\infty} \Pi_n(a^0 : \bar{\delta}) \leq \lim_{n\to\infty} \Pi_n(a^0 : \delta^m)$, for all $m \in \mathscr{M}$.

Proof — See Appendix.

The decision rule $\bar{\delta}$ is a weighted average of all decision rules δ^m, with weights given by $\omega_n^m / \sum_{m=1}^M \omega_n^m$. The weight is proportional to the cdf of the

standard normal distribution evaluated at $\sqrt{n}(\hat{L}_n^m - \hat{L}_n^{m^*})/\hat{\sigma}_n$, so that it will give higher weights to realizations closest to zero. Eventually, realizations from correctly specified models will converge in probability to μ^0 and the associated decision rules will be the only ones to receive positive weight. Misspecified models m, on the other hand, will converge in probability to $\mu^m > \mu^0$ and will asymptotically receive zero weights. If all models are misspecified, positive weights are given only to the models with lowest μ^m.

The second part of the theorem says that the weighted average of the asymptotically best decision rules is a decision rule which is weakly better than the single best decision rule. In case the best decision rules come from nested models, averaging does not provide any asymptotic improvement, as all decision rules converge to the same rule. If, however, the best decision rules are derived from non-nested models, the convexity of the loss function implies that averaging among these rules provides a strictly lower asymptotic loss than the one associated with the individual rules.

3 Inflation Forecasting Models

The implementation of the theory developed in the previous section requires the econometrician to take a stance on three key ingredients of the forecasting process:

1. The loss function and the judgment of the decision maker
2. The data used in the forecasting model
3. The functional form of the asymptotic approximation

This section reviews each of these elements.

3.1 Loss Function and Judgment

The loss function and judgment are both personal choices of the decision maker. It will therefore vary from decision maker to decision maker. I take

here my personal perspective on the problem of inflation forecasting.

I assume a standard quadratic loss function:

$$\mathscr{L}(Y_{n,h}, a_n) = 0.5(Y_n^h - a_n)^2 \tag{17}$$

The term Y_n^h is a function of the forecast horizon h and future inflation realizations.

I form my judgment about inflation by appealing to the statutory mandate of the European Central Bank. According to the 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.'* The Governing Council of the European Central Bank has subsequently provided the following quantitative definition: *'Price stability shall be defined as a year-on-year increase in the Harmonised Index of Consumer Prices (HICP) for the euro area of below, but close to, 2% over the medium term'* (ECB, 2011, p. 64). This is still not sufficient to arrive at a precise quantitative definition. I therefore take my personal definition of *'below, but close to, 2%'* as 1.9% and of *'medium term'* as 24 months. Of course other interpretations of the ECB definition could be justified.

Trusting that the ECB will deliver on its mandate, the above definition of price stability implies:

$$Y_n^{24} = 24^{-1} \sum_{i=1}^{24} Y_{n+i} \tag{18}$$

$$\tilde{a}_{n,24} = E(Y_n^{24} | x^n) = 1.9\% \tag{19}$$

that is, my judgmental forecast is that the annualized monthly inflation over the next two years will average 1.9%. Although this looks like a reasonable judgment when facing a credible central bank with an explicitly defined inflation objective, this 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 anchor to the expectations of the economic agents. The coordination of all expectations around the official inflation objective should by itself help to deliver realized inflation close to the objective. Setting the judgmental decision as in (19) may therefore prove quite a good forecasting decision. In fact, the quantitative results of Diron and Mojon (2008) reveal that such a simple rule of thumb yields smaller forecast errors than widely used forecasting models.

3.2 Data

Given the ECB definition of price stability, one 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 in changing composition, as provided by Eurostat.[4] 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. I use here the euro area HICP excluding energy and unprocessed food. I also consider real economic variables, such as unemployment and industrial production, because as suggested by the Phillips curve there may be an inverse relationship between inflation and the level of economic activity. All data are monthly, ranging from January 1999 (when the euro was introduced) to February 2018.

The time series behavior of the four variables is reported in figures 1, 2 and 3. Key summary statistics are reported in table 1.

[4]All data have been downloaded from the ECB database, http://sdw.ecb.europa.eu

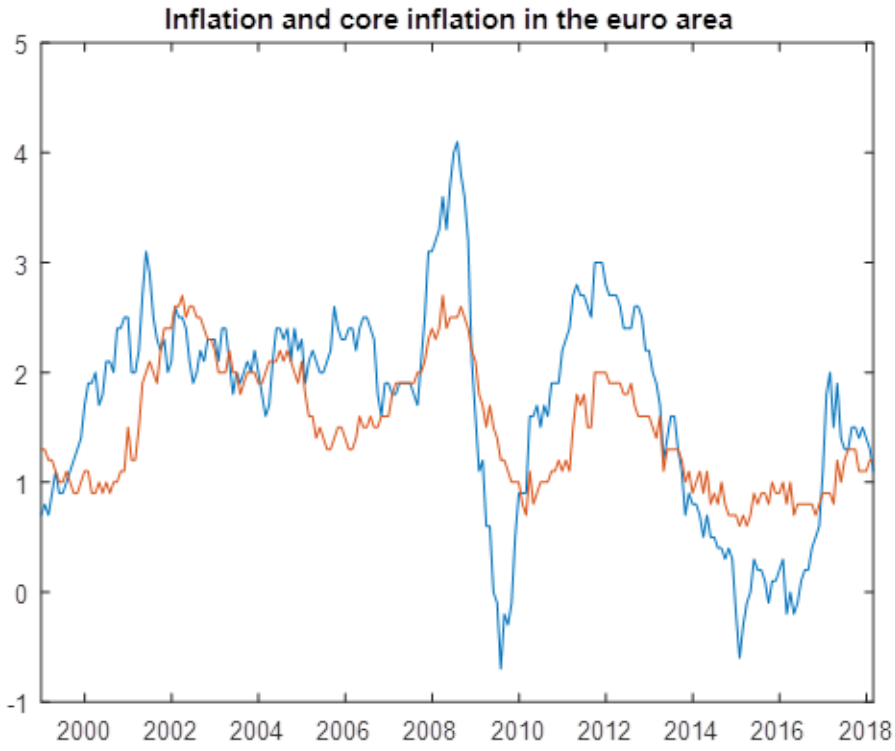

Figure 1: Headline and core inflation in the euro area

Note: HICP - Overall index, annual rate of change, euro area (changing composition), neither seasonally nor working day adjusted.

HICPX - All-items excluding energy and unprocessed food, annual rate of change, euro area (changing composition), neither seasonally nor working day adjusted.

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

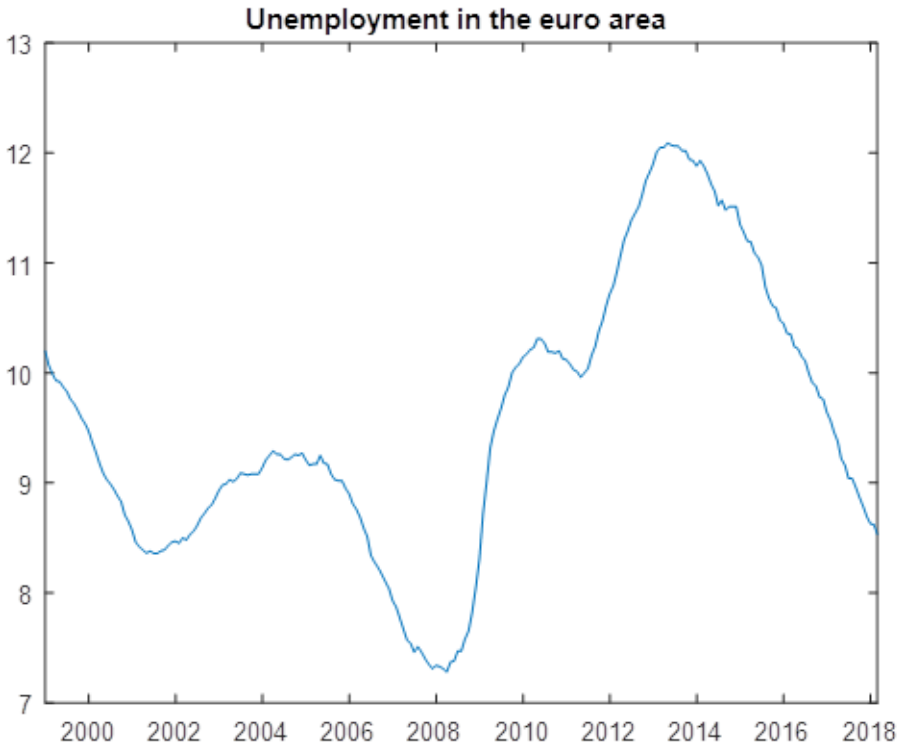

Figure 2: Unemployment in the euro area

Note: UN - Standardized unemployment, rate, euro area 19 (fixed composition), total (all ages), total (male and female), seasonally adjusted, not working day adjusted, percentage of civilian workforce.

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

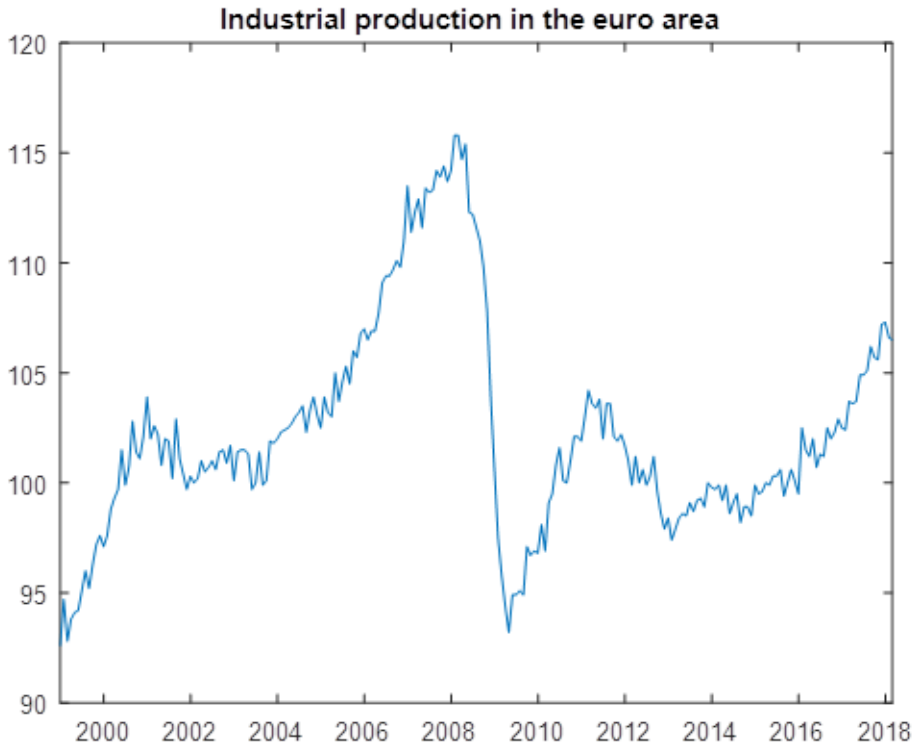

Figure 3: Industrial production in the euro area

Note: IP - Industrial production index, total industry, euro area 19 (fixed composition), working day and seasonally adjusted.

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

Table 1: Summary statistics

	HICP	HICPX	UN	IP
Mean	1.70	1.50	9.55	102
Std Dev	0.96	0.55	1.26	4.83
Min	-0.70	0.60	7.28	92
Max	4.10	2.70	12.09	115
n	231			

Note: Summary statistics for the time series used in the forecasting models. HICP is the headline inflation, HICPX is core inflation, UN is the unemployment rate, IP is the industrial production. n is the length of the time series. See the notes of the figures for more details.

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

3.3 Models

The last ingredient of the forecasting process is to choose the functional form of f_t, the building block of the quasi-log-likelihood function introduced in section 2.1. For simplicity, I limit myself to the class of VAR(p) models:

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

where $Y_t = (HICP, HICPX, UN, IP)'$, c is a 4×1 vector, C_i are 4×4 matrices of coefficients, and U_t is a 4×1 vector of residuals. For a given p, I allow for any possible combination of lags. Since the algebra is conceptually straightforward, but rather cumbersome, I refer to the Appendix 2 for detailed derivations, which build on Lütkepohl (2005). In this case, f_t takes the following functional form:

$$f_t(X^t, \theta) = \exp(-0.5 u_t' \Sigma_u^{-1} u_t) \tag{20}$$

where u_t is a suitably defined vector of residuals and Σ_u the associated variance covariance matrix. This is the VAR equivalent of a Generalised Least Squares estimator.

Estimation of asymptotic variance-covariance matrices follows the procedure detailed in section 2. See again Appendix 2 for technical details.

3.4 A Few More Arbitrary Choices

The actual implementation of the forecasting process still requires the econometrician to specify some more parameters. In theory, each of these parameters would give rise to potentially different models and therefore could be optimized using the model selection procedure described in this paper. In practice, time and computing power limitations usually require the econometrician to take some shortcuts.

The first arbitrary choice is obviously the functional form of the econometric model. In the case of a VAR(p) model, I have arbitrarily chosen the length of the vector Y_t to be 4 and the other variables to be core inflation, unemployment and industrial production. Other options for the functional form could be factor models, DSGE models or deep neural nets from the machine learning literature (see, for instance, Faust and Wright (2013) for an exhaustive list of the most common methods used for forecasting inflation). As long as the assumptions about f_t are satisfied, any of these alternative models fits the same conceptually framework. In fact, the framework of this paper can be used to choose among these alternative modelling strategies.

The second arbitrary choice is the maximum number of possible lags, p. I set this to 12, which corresponds to one year of possible monthly lags.

The third arbitrary choice is how many regressors to include in the model. Recall that the arguments of section 2 rely on the asymptotic approximation being valid. The ratio between number of observations (which is given by the number of time series in the VAR times their length) and parameters to be estimated cannot therefore be too small. I arbitrary choose this ratio to be 50, that is each model parameter is estimated, on average, with at least 50

observations. One could run a Monte Carlo experiment to test the validity of this choice.

The fourth arbitrary choice is the initialization of the VAR model. I simply put the pre-sample values equal to their full sample averages. Other choices are possible, including treating the pre-sample values as parameters to be estimated. I also discard all non stationary QMLE estimates, in line with the stationarity assumption imposed in section 2.1.

Even with these simplifications, a VAR model with $p = 12$ lags, $g = 4$ dependent variables and $MR = 18$ maximum number of regressors can be combined in

$$\sum_{i=1}^{MR} \binom{g + g^2 p}{i} \approx 1.46 \cdot 10^{25}$$

different possible ways. Even the fastest computer cannot cope with a complete grid search. I use, instead, an integer optimization algorithm to choose among all the possible models. The logic is the following. Let ϑ be a $(g+g^2p)$-vector of 0s and 1s and let θ^m be the vector of parameters to be estimated corresponding to the 1s in ϑ. To each combination of ϑ corresponds exactly one model identified by θ^m and for each θ^m it is possible to compute the associated loss $\hat{L}^m$. Figure 4 provides an example of a possible model. Model selection can now be framed as the following integer optimization program:

$$\begin{aligned} \min_{\vartheta} \hat{L}^m &= n^{-1} \sum_{t=1}^{n} 0.5(\delta_t^m(x^n) - Y_n^h)^2 \\ s.t. \quad & ||\vartheta||^2 \leq MR \end{aligned} \tag{21}$$

The genetic algorithm of Matlab has specific options for solving optimization problems for integer-valued variables. I found, however, that it is rather inefficient and time consuming. A better alternative seems to be the algorithm *patternsearch*, which, even though not specifically designed to solve integer optimization problems, under suitable choices of its otpions can be tricked into solving this type of problems. I refer to my codes for details.[5] In my

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

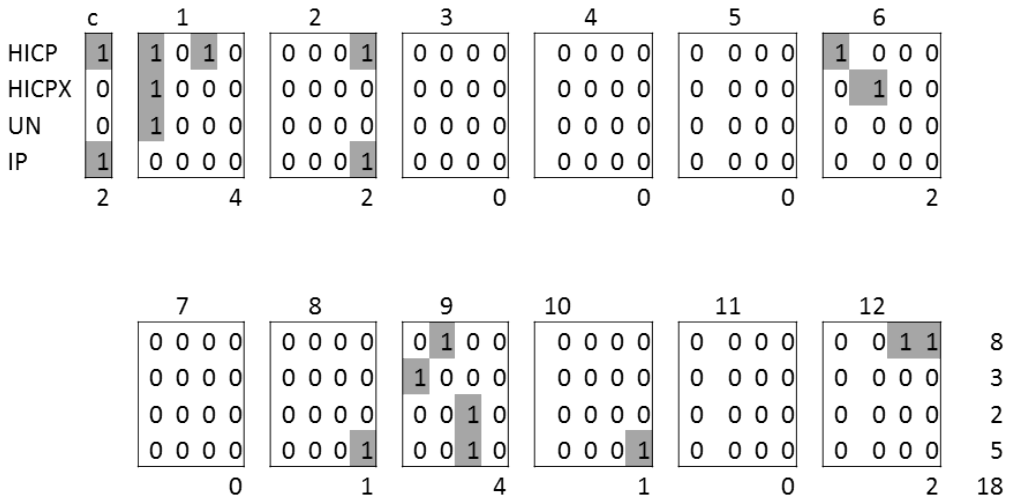

Figure 4: Example of model specification

Note: The figure reports an example of a possible VAR model specification, with 12 lags, 4 variables and 18 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.

experience, the algorithm seems to be quite sensitive to the user supplied initial conditions, suggesting that the global optimum is difficult to find. I have provided as initial conditions the models with only the constants, with the first MR regressors, and three models selected with lasso techniques. I note, in passing, that models selected with lasso are usually very far from the best ones.

4 Inflation Forecasts for the Euro Area

Figure 5 reports the forecasts associated with the best models selected using the implementation details described in the previous section. The single best model for forecasting average 24-month inflation as of February 2018, that is the one delivering the lowest in sample loss as prescribed by Theorem 2.1, is the one reported in figure 4. 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 as in equation (18) over the months indicated on the horizontal axis. The top line is the forecast associated with a 24-month judgmental forecast of 1.9%, and it is the forecast that I would choose if forced to bet on one number. The 24-month inflation forecast is exactly 1.9%. In other words, the null hypothesis that my judgmental forecast of inflation is correct cannot be rejected at a 10% confidence level. The three-year forecast is 1.93%.

One subtle point to notice is that even though I express my judgment in terms of the average inflation over 24 months being equal to 1.9%, I map this judgment over any forecasting horizon. That is, the inflation forecast at the 1-month horizon incorporates my 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.* $Y_n^{24}(\theta) = \tilde{a}_n$, to compute the forecast at any horizon of interest.

The bottom line of the figure is the forecast associated with an alternative 24-month judgmental forecast, in this case set equal to 0%. While asymp-

totically the two lines would coincide (judgment is irrelevant with an infinite sample size, as the model parameters would converge to their QMLE value), with finite samples forecasts associated with different judgments may differ. By construction, the bottom line will never lie above the top line, as the best forecast is identified by the closest bound of the confidence interval. It is interesting to note that in this case inflation forecasts do not seem overly sensitive to drastically different judgments.

For comparison, figure 6 reports the best forecast derived from a model selection based only on an autoregressive process for the euro area inflation. Two interesting observations are in order. First, while the very short term forecasts are all very close to each other, the forecasts at longer horizons are quite different, with the two-year forecast standing at 1.7%. Second, the distance between the forecasts associated with the two alternative judgments is much larger than in the case of a VAR model with four variables.

To highlight the differences in performance of the various models, figure 7 reports the average in sample loss at different horizons for the best models selected from four different sets, the univariate AR model and the VAR models with 2, 3 and 4 endogenous variables. The VAR models with two variables contains HICP and HICPX, while the VAR model with three variables includes also unemployment (but no industrial production). Since the larger set of models from which to choose always contains the smaller set, the corresponding average in sample loss cannot increase: it would always be possible to choose the best model from the smaller set. At short horizons, the performance of all the models is similar. Significant differences start to emerge only as the forecasting horizon increases. Two interesting findings emerge from this figure. The first one is that core inflation does not seem to help much in forecasting inflation at longer horizons, as the difference between the performance of the AR and the 2-variable VAR is small. The short term volatility induced by the items excluded in core inflation is averaged out as the forecasting horizon increases and becomes less relevant. The second interesting finding is that including unemployment significantly

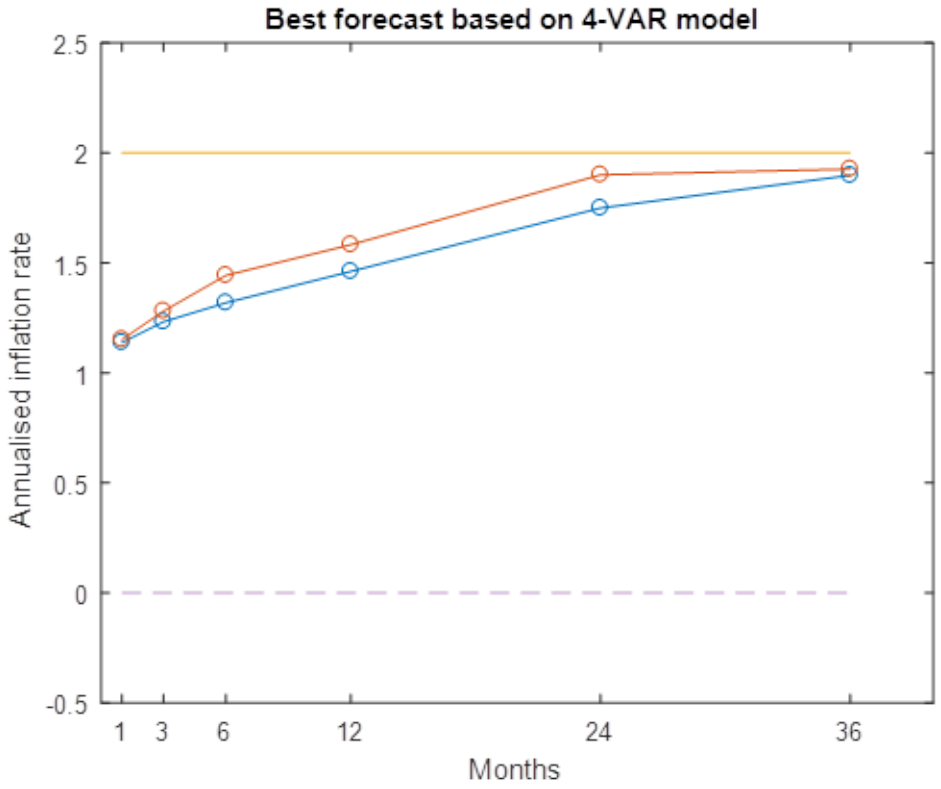

Figure 5: Forecasts of the euro area inflation

Note: The figure exhibits the best euro area inflation forecast at different horizons, with information available as of February 2018. 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 1.9%. The bottom line is the forecast associated with a 24-month judgmental forecast of 0%. The horizontal lines at 0% and 2% are reported as reference points.

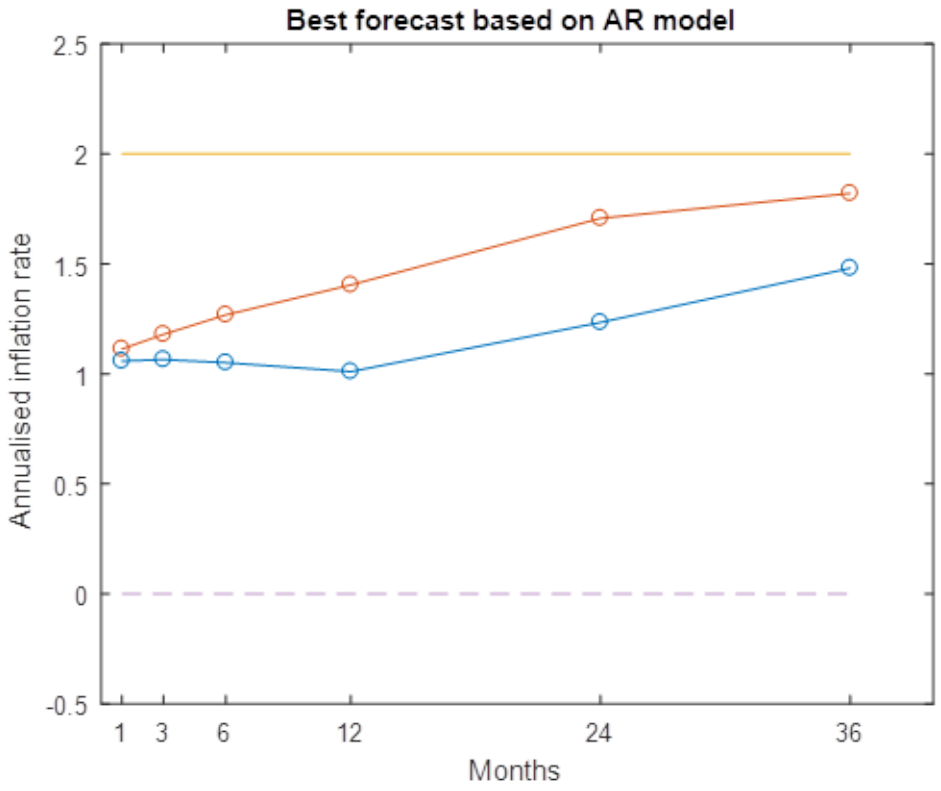

Figure 6: Forecasts of the euro area inflation based on AR model

Note: Best euro area inflation forecast based on model selection from a simple autoregressive model. See the notes of figure 5 for explanatory details.

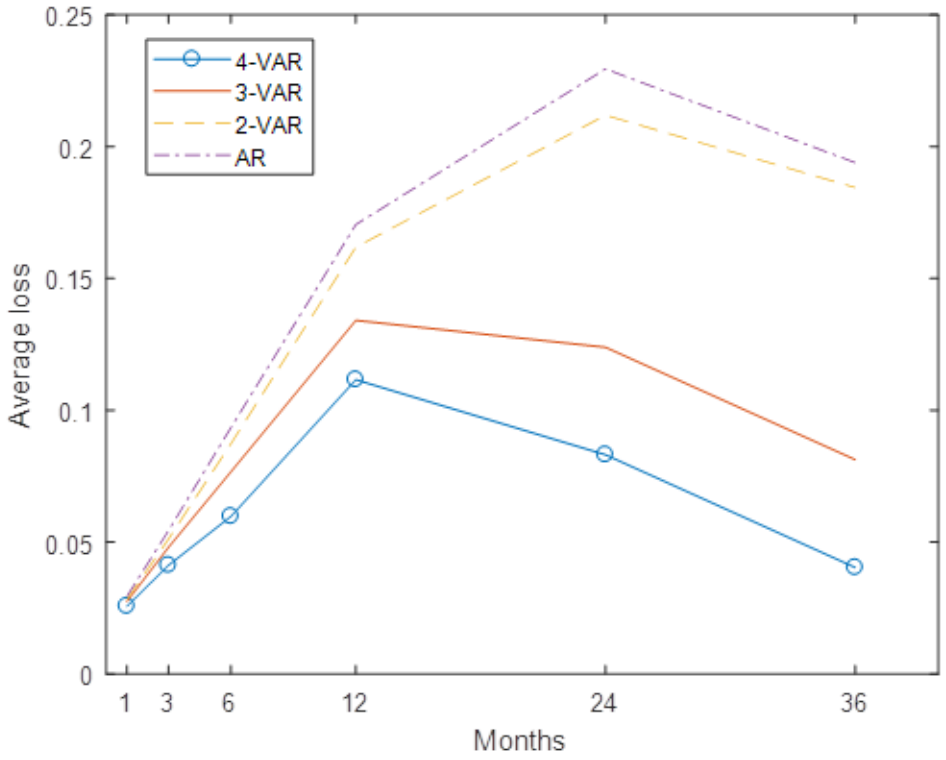

Figure 7: Performance of the best models

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

improves the forecasting performance of the VAR model.[6] Even though this is only evidence that unemployment Granger causes medium term inflation and no structural interpretation can yet be given to this finding, this result may usefully inform the debate about the relevance of the Phillips curve.

Finally, figure 8 reports the in sample, time series performance of the best models selected from the sets of AR and four-variable VAR models. The smooth blue line is the backward 36-month moving average of HICP. The red dashed line is the four-variable VAR best model forecast at the beginning of

[6]Running a similar three-variable VAR with industrial production in place of unemployment produces much lower improvement.

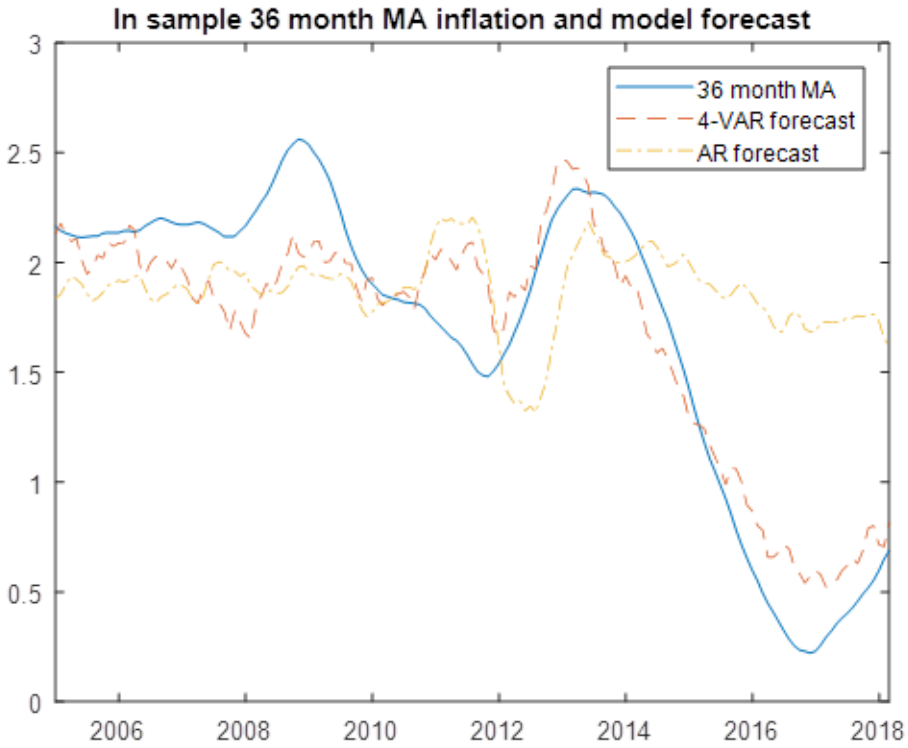

Figure 8: Performance of the best AR model

Note: Backward 36-month moving average inflation, together with the best in sample forecasts from the set of all possible AR or four-variable VAR models.

the 36-month forecasting window (but using the parameter estimates from the full sample). The dashdot line if the best forecast associated with the AR model. The average in sample loss reported in the previous figure is simply the average of the square difference between these two lines and the smooth blue lines. It is clear from the figures that, unlike the four-variable VAR, the simple AR model does not contain sufficient dynamics to describe the stochastic process represented by three-year inflation.

5 Conclusion

An econometrician is asked to provide a statistical decision rule which does not perform worse than a judgmental decision with a given confidence level provided by the decision maker. If the statistical model is correctly specified and the distribution of the estimator is known, the econometrician can construct a confidence interval around the maximum likelihood decision. The best decision is given by the decision associated with the boundary of the confidence interval which is closest to the judgmental decision. In most real world situations, the econometrician does not know the correctly specified model, but must choose from a given class of probability distributions. One important consequence of model misspecification is that the econometrician can no longer guarantee that the proposed statistical decision rule has the chosen bounded probability of underperforming the given judgmental decision. Searching for the least misspecified model alleviates this problem, but since any statistical model is likely to be misspecified, a decision maker engaging in statistical decision making has to live with this additional level of uncertainty. The least misspecified model is the one associated with the decision rule with lowest in sample empirical loss. If the class of models considered by the econometrician contains the true model, this model selection criterion asymptotically selects the true model with probability one. Averaging decision rules according to their asymptotic performance provides a new decision rule which is weakly better than the single best decision rule. If the best decision rules come from nested models, averaging does not provide any asymptotic improvement, as all rules converge asymptotically to the same rule. If, however, there are two or more asymptotically equivalent best decision rules derived from non-nested misspecified models, the convexity of the loss function implies that their average results in a new decision rule which is strictly better than any of the individual rules.

A Appendix — Proofs

Proof of Corollary 2.1 — The derivation of (9) is based on the following mean value expansion:

$$\hat{k}_{n,\lambda} = k^*_{n,\lambda} + \bar{K}_{n,\lambda}(\theta(X^n) - \theta^*)$$

where $k^*_{n,\lambda} \equiv \nabla_a L_{t,h}(\theta^*, a_t(\lambda))$ and $\bar{K}_{n,\lambda} \equiv \nabla_{a,\theta} L_{t,h}(\bar{\theta}(X^n), a_t(\lambda))$, for $\bar{\theta}(X^n)$ denoting a mean value vector between θ^* and $\theta(X_n)$.

Defining $Z_\lambda \equiv \sqrt{n}(\hat{K}_{n,\lambda}\hat{A}_n^{-1}\hat{B}_n\hat{A}_n^{-1}\hat{K}'_{n,\lambda})^{-1/2}\hat{k}_{n,\lambda}$, under the null that $a_t(\lambda)$ is optimal $k^*_{n,\lambda} = 0$ and $Z_\lambda \overset{A}{\sim} N(0,1)$. Since $\mathcal{W}_{n,\lambda}(X^n) = Z'_\lambda Z_\lambda$, to every critical value associated with z_λ corresponds one and only one critical value associated with $\mathcal{W}_{n,\lambda}(x^n)$. We are therefore in the same decision environment of Theorem 2.1 of Manganelli (2018), except for the fact that the loss function is now generically convex and continuously differentiable.

It remains to determine the action to be taken in case the test statistic $\psi(\mathcal{W}_{n,0}(x^n)) = 1$. Following the same logic as the proof of Theorem 2.1 of Manganelli (2018), rejection of the null hypothesis implies that marginal moves away from $a_t(0)$ in the direction of $a_t(1)$ increase the loss function with probability less than α. Given her confidence level α, the decision maker is willing to take this marginal move until one reaches the action $a_t(\hat{\lambda})$ where the null is no longer rejected. This action is implicitly defined by $\mathcal{W}_{t,\hat{\lambda}}(x^n) = c_\alpha$. Given the convexity and continuity assumptions on $\mathcal{L}$, this value exists and is unique. □

Proof of Theorem 2.1 — If model m^* minimizes

$$\text{plim}\, n^{-1} \sum_{t=1}^{n} \mathscr{L}(Y_{t,h}, \delta_t^m(x^n))$$

it will also minimize:

$$\begin{aligned}\operatorname{plim} n^{-1}\sum_{t=1}^{n}\Big(\mathscr{L}(Y_{t,h},\delta_t^m(x^n))-L_t(a_t^0|F_t)\Big)=\\ =\operatorname{plim} n^{-1}\sum_{t=1}^{n}\Bigg(\mathscr{L}(Y_{t,h},\delta_t^m(x^n))-E(\mathscr{L}(Y_{t,h},\delta_t^m(x^n)))+\\ +E(\mathscr{L}(Y_{t,h},\delta_t^m(x^n)))-E(\mathscr{L}(Y_{t,h},\delta_t^m(x^n)|X_t))+\\ +E(\mathscr{L}(Y_{t,h},\delta_t^m(x^n)|X_t))-L_t(a_t^0|F_t)\Bigg)\end{aligned}$$

By the law of large numbers:

$$\operatorname{plim} n^{-1}\sum_{t=1}^{n}\Big(\mathscr{L}(Y_{t,h},\delta_t^m(x^n))-E(\mathscr{L}(Y_{t,h},\delta_t^m(x^n)))\Big)=0$$

and by the law of iterated expectations:

$$\operatorname{plim} n^{-1}\sum_{t=1}^{n}\Big(\mathscr{L}(Y_{t,h},\delta_t^m(x^n))-E(\mathscr{L}(Y_{t,h},\delta_t^m(x^n)|X_t))\Big)=0$$

and the first result follows.

Furthermore, since $\operatorname{plim}\delta_t^m(X^n)=a_t^{*m}$, if model m^* is correctly specified:

$$\begin{aligned}a_t^{*m^*}&=\arg\min_a L_t(\theta^{*m^*},a_t)\\&=\arg\min_a L_t(a_t|F_t)\\&=a_t^0\end{aligned}$$

□

Proof of Theorem 2.2 — The weight associated with decision δ^i is given by $\omega_n^i/\sum_j\omega_n^j$. Clearly $\sum_j\omega_n^j\leq M$. Since $\operatorname{plim}(\hat{L}_n^i-\hat{L}_n^{m^*})^2/(\hat{\sigma}_n^m)^2=(\mu^i-\mu^{m^*})^2/(\sigma^m)^2$, it follows that $\omega_n^i\xrightarrow{p}0$ if $(\mu^i-\mu^{m^*})^2>0$.

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

that:

$$\begin{aligned}
&\lim_{n\to\infty}(\Pi_n(a^0:\bar{\delta})-\Pi_n(a^0:\delta^m))\\
=&\lim_{n\to\infty}n^{-1}\sum_{t=1}^{n}(L_t(\bar{\delta}_t(x^n)|F_t)-L_t(\delta_t^m(x^n)|F_t))\\
\leq&\lim_{n\to\infty}n^{-1}\sum_{t=1}^{n}(\sum_i w^i L_t(\delta_t^i(x^n)|F_t)-L_t(\delta_t^m(x^n)|F_t))
\end{aligned}$$

□

B Appendix 2 — Technical Derivations for the VAR model

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

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

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

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).

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

$$\underset{(g\times 1)}{y_t} = \underset{(g\times 1)}{\nu} + \underset{(g\times g)}{A_1}\, y_{t-1} + ... + A_p y_{t-p} + \underset{(g\times 1)}{u_t} \qquad u_t \sim i.i.d.(0, \underset{(g\times g)}{\Sigma_u}) \tag{22}$$

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

$$\underset{(g\times n)}{Y} = \underset{(g\times 1+gp)}{C}\ \underset{(1+gp\times n)}{Z} + \underset{(g\times n)}{U} \tag{23}$$

where

$$Y \equiv [y_1, ..., y_n] \qquad C \equiv [\nu, A_1, ..., A_p]$$
$$\underset{(1+gp\times 1)}{Z_t} \equiv [1, y_t', ..., y_{t-p+1}']' \qquad Z \equiv [Z_0, ..., Z_{n-1}]$$
$$U \equiv [u_1, ..., u_n]$$

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

$$\beta \equiv vec(C) = \underset{(g(1+gp)\times q)}{R} \underset{(q\times 1)}{\theta} \tag{24}$$

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:

$$\begin{aligned} \underset{(gn\times 1)}{y} &\equiv vec(Y) \\ &= (Z' \otimes I_g)vec(C) + vec(U) \\ &\equiv (Z' \otimes I_g)R\theta + u \end{aligned} \tag{25}$$

B.1 The quasi-likelihood function and its first and second derivatives

Define

$$f_t(X^t, \theta) \equiv \exp(-0.5u_t'\Sigma_u^{-1}u_t) \tag{26}$$

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

$$\begin{aligned} \ell_n(X^n, \theta) &= -0.5n^{-1}\sum_{t=1}^{n} u_t'\Sigma_u^{-1}u_t \\ &= -0.5n^{-1}u'(I_n \otimes \Sigma_u^{-1})u \end{aligned} \tag{27}$$

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

$$\hat{\theta} = (R'(ZZ' \otimes \hat{\Sigma}_u^{-1})R)^{-1}R'(Z \otimes \hat{\Sigma}_u^{-1})y \tag{28}$$

The matrix $\hat{\Sigma}_u$ is unknown. Under the assumption that the residuals u_t are normally distributed, the variance covariance matrix could be efficiently estimated by maximum likelihood. For the purpose of this paper, this could be computationally expensive. I therefore estimate it as:

$$\hat{\Sigma}_u = n^{-1}(Y - \hat{\hat{C}}Z)(Y - \hat{\hat{C}}Z)' \tag{29}$$

where

$$\begin{aligned} vec(\hat{\hat{C}}) &= \hat{\hat{\beta}} \\ \hat{\hat{\beta}} &= R\hat{\hat{\theta}} \\ \hat{\hat{\theta}} &= (R'(ZZ' \otimes I_g)R)^{-1}R'(Z \otimes I_g)y \end{aligned}$$

See the discussion in chapter 5 of Lütkepohl (2005) for alternative strategies to estimate $\hat{\Sigma}_u$.

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:

$$\nabla' \log f_t(X^t, \theta) = u_t'\hat{\Sigma}_u^{-1}(Z_t' \otimes I_g)R \tag{30}$$

and its transpose is

$$\nabla \log f_t(X^t, \theta) = R'(Z_t \otimes I_g)\hat{\Sigma}_u^{-1}u_t \tag{31}$$

The second derivative is

$$\begin{aligned} \nabla^2 \log f_t(X^t, \theta) &= -R'(Z_t \otimes I_g)\hat{\Sigma}_u^{-1}(Z_t' \otimes I_g)R \\ &= -R'(Z_tZ_t' \otimes \hat{\Sigma}_u^{-1})R \end{aligned} \tag{32}$$

Therefore

$$\begin{aligned} \hat{A}_n &= -n^{-1}\sum_{t=1}^{n} R'(Z_tZ_t' \otimes \hat{\Sigma}_u^{-1})R \\ &= -n^{-1}R'((\sum_{t=1}^{n} Z_tZ_t') \otimes \hat{\Sigma}_u^{-1})R \\ &= -n^{-1}R'(ZZ' \otimes \hat{\Sigma}_u^{-1})R \end{aligned} \tag{33}$$

and

$$\begin{aligned}\hat{B}_n &= n^{-1}\sum_{t=1}^{n} R'(Z_t \otimes I_g)\hat{\Sigma}_u^{-1}u_t u_t'\hat{\Sigma}_u^{-1}(Z_t' \otimes I_g)R \\ &= n^{-1}R'(Z \otimes \hat{\Sigma}_u^{-1})diag(U)diag(U')(Z' \otimes \hat{\Sigma}_u^{-1})R \end{aligned} \tag{34}$$

where $diag(U)$ is a block diagonal matrix with its columns along the diagonal.

B.2 Forecasting

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

$$\hat{y}_{t,h} = h^{-1}\sum_{i=1}^{h}\hat{y}_{t+i} \tag{35}$$

Following closely Lütkepohl (2005), page 96:

$$\hat{y}_{t+i} = J\bar{C}^i Z_t \tag{36}$$

where

$$\underset{(1+gp\times 1+gp)}{\bar{C}} \equiv \begin{bmatrix} 1 & \mathbf{0}_{1,g(p-1)} & \cdots & \mathbf{0}_{1,g} \\ C & & & \\ \mathbf{0}_{g(p-1),1} & I_{g(p-1)} & & \mathbf{0}_{g(p-1),g} \end{bmatrix} \tag{37}$$

$$\underset{(g\times 1+gp)}{J} \equiv [\underset{(g\times 1)}{\mathbf{0}}, I_g, \underset{(g\times g(p-1))}{\mathbf{0}, \cdots, \mathbf{0}}] \tag{38}$$

Consider the following mean value expansion:

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

where $\bar{\theta}$ is a mean value which lies between θ and $\hat{\theta}$ and I have imposed the constraint (24). It follows that:

$$\sqrt{n}(\hat{y}_{t,h} - y_{t,h}) \sim N(0, \nabla y_{t,h}RA^{-1}BA^{-1}R'\nabla' y_{t,h}) \tag{39}$$

It remains to compute $\nabla y_{t,h} = h^{-1} \sum_{i=1}^{h} \nabla y_{t+i}$:

$$\nabla y_{t+i} = \sum_{j=0}^{i-1} Z_t'(\bar{C}')^{i-1-j} \otimes J\bar{C}^j J' \tag{40}$$

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 confience interval.

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

$$\nabla_a L_n(\theta(X^n), \tilde{a}_n) = -e_1 y_{n,h}(\theta(X^n)) + \tilde{a}_n \tag{41}$$

where e_1 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 (39) of the previous subsection, under the null H_0 : $-e_1 y_{n,h}(\theta^*) + \tilde{a}_n = 0$:

$$\sqrt{n}\sigma^{-1}(-e_1 y_{n,h}(\theta(X^n)) + \tilde{a}_n) \sim N(0,1) \tag{42}$$

where $\sigma^2 \equiv e_1 \nabla y_{n,h} R A^{-1} B A^{-1} R' \nabla' y_{n,h} e_1'$ can be consisently estimated with standard plug-in estimators.

Denoting wiht 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 (10) can be rewritten as:

$$\begin{aligned} \delta_n(x^n) =& \tilde{a}_n I(c_{\alpha/2} \leq T(x^n) \leq c_{1-\alpha/2}) + \\ &+ (e_1 y_{n,h}(\theta(x^n)) + \hat{\sigma} c_{\alpha/2}/\sqrt{n}) I(T(x^n) < c_{\alpha/2}) + \\ &+ (e_1 y_{n,h}(\theta(x^n)) + \hat{\sigma} c_{1-\alpha/2}/\sqrt{n}) I(T(x^n) > c_{1-\alpha/2}) \end{aligned} \tag{43}$$

B.4 Model selection

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

$$\hat{L}_n^m = 0.5(n-h)^{-1} \sum_{t=1}^{n-h} (h^{-1} \sum_{i=1}^{h} e_1 y_{t+i} - \delta_t^m(x^n))^2 \tag{44}$$

This in turn requires the construction of the in sample optimal decision $\delta_t^m(x^n)$. The only missing element to construct $\delta_t^m(x^n)$ from (43) 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:

$$\begin{aligned} &\max_{\theta} \ell_n(x^n, \theta) \\ &s.t. \quad y_{n,h}(\theta(x^n)) = \tilde{a}_n \end{aligned} \tag{45}$$

where $\ell_n(x^n, \theta)$ was defined in (27). Denoting with $\tilde{\theta}(x^n)$ the constrained parameter vector:

$$\tilde{a}_t = y_{t,h}(\tilde{\theta}(x^n)) \tag{46}$$

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 $\hat{L}_n^m$.

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

$$(\hat{\sigma}^{m^*})^2 = (n-h)^{-1} \sum_{t=1}^{n-h} \left((h^{-1} \sum_{i=1}^{h} e_1 y_{t+i} - \delta_t^{m^*}(x^n))^2 - \hat{L}_n^{m^*} \right)^2 \tag{47}$$

It is now possible to construct the model weights ω_n^m given in (15) and the model averaging decision rule $\bar{\delta}$ given in (16).

References

Diebold, F.X. and R.S. Mariano (1995), Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13, 253–263.

Diron, M. and B. Mojon (2008), Are Inflation Targets Good Inflation Forecasts?, Economic Perspectives, Federal Reserve Bank of Chicago, issue Q II, pages 33-45.

ECB (2011), *The Monetary Policy at the ECB*, European Central Bank, available at http://www.ecb.europa.eu

Faust, J. and J. Wright (2013), Forecasting Inflation, in *Handbook of Economic Forecasting* (G. Elliott and A. Timmermann (eds.)), Volume 2A, Elsevier.

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

Manganelli, S. (2018), Deciding with Judgment, ECB manuscript, available at www.simonemanganelli.org.

Manganelli, S. (2009), Forecasting with Judgment, *Journal of Business and Economic Statistics*, 27 (4), 553-563.

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

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