Jueves 20 y Viernes 21 de Noviembre de 2014
International Conference on Regional Science:
FINANCING AND THE ROLE OF THE REGIONS AND TOWNS IN ECONOMIC RECOVERY
Facultad de Economía y Empresa y Paraninfo de la Universidad de Zaragoza

Entropy Econometrics for combining regional economic forecasts: A Data-Weighted Prior Estimator

Esteban Fernández-Vázquez (evazquez@uniovi.es)
Blanca Moreno (morenob@uniovi.es)

REGIOlab- Laboratorio de Análisis Económico Regional,
Departamento de Economía Aplicada
Universidad de Oviedo

Área Temática: Econometría espacial y métodos de análisis regional

Resumen:

Regional economic agents have access to a wide variety of information and they can use different forecasting techniques, thus leading to a considerable degree of heterogeneity between the obtained forecasts. A combination of individual forecasts, each one capturing a different aspect of the available information, would be expected to perform better than the individual forecasts. Forecast combination methods vary in nature and involve different degrees of sophistication.

Although there are several forecasts combination methods, it is usual to use the simple arithmetic mean in order to summarize the individual forecasts. This strategy appears to be justified empirically by the fact that the resulting simple averaging schemes tend to outperform more sophisticated ones. Such phenomenon is usually referred to as the "forecasting combination puzzle". Thus, the simple equal-weight averages constitute a Benchmark. Theoretically, the use of the simple arithmetic mean could be justify when all the forecasters have the same past performance or when we do not have enough information about that.

In this paper we explore the possibility of using Entropy Econometrics as a procedure for combining forecasts that allows to discriminate between bad and good forecasters even in that situation of small information. We provide an explicit framework for combining forecast based on the Data-Weighted Prior (DWP) Estimator proposed by Golan (Journal of Econometrics, 2001) that allows for simultaneous parameter estimation and forecasters selection in linear statistical models. In particular, we examine the ability of the DWP Estimator to effectively select relevant forecasts among all forecasts. We test the validity of the proposed model by a simulation exercise and compare its ex ante forecasting performance with other combining methods. The simulation results suggest that the proposed method dominates other forecast combination strategies which are examined, as equal-weight averages or ordinal least square methods, among others.

Palabras Clave: Regional Economic forecasts; Combined forecast; Entropy Econometrics.

Clasificación JEL: C53; C55

1. Introduction.

Regional economic agents have access to a wide variety of information and they can use different forecasting techniques, thus leading to a considerable degree of heterogeneity between the obtained forecasts.

Therefore, a combination of individual forecasts, each one capturing a different aspect of the available information, would be expected to perform better than the individual forecasts (leads to increased forecast accuracy). In fact, growing amount of literature have empirically demonstrated the superior performance of forecast combination (e.g. Holden and Peel, 1988, Marcellino, 2004; Stock and Watson, 2004; Greer, 2005, Timmerman, 2006, Moreno and López, 2013)

The pioneers in the theoretical study of the combination of forecasts were Bates and Granger (1969). Since their seminal article a large amount of literature has appeared showing a variety of methods to estimate the weights assigned to each individual forecast. A summary of the literature on combination of forecasts can be found in Clemen, 1989 and Bunn 1989; moreover an update of this review from a practical perspective can be found in de Menezes *et al.* 2000 and Timmerman, 2006).

Forecast combination methods vary in nature and involve different degrees of sophistication. Thus, in variance-covariance methods (Bates and Granger, 1969; Newbold and Granger, 1974) weights are calculated to minimize the error variance of the combination, while probabilistic methods (Bunn, 1975; Bordley, 1982) allow the interpretation of each weight as the probability that a specific forecast will perform best on the next occasion. If we assume a random character for both the variable being predicted (y) and the individual forecasts, then the *a priori* distribution of y could be modified with the sample information, obtaining the *a posteriori* distribution whose expected value would be the combined forecast. This approach leads to Bayesian methods which were originally put forward by Winkler (1981) and extended by Lindley (1983), Winkler and Makridakis, (1983), Agnew (1985), Anandalingam and Chen (1989) and Clemen and Winkler (1993) among others.

In 1984 Granger and Ramanathan introduced the regression-based method of combining forecasts allowing the weights to be interpreted as the coefficient vector of a linear projection of the variable being predicted onto the individual forecasts.

The estimation of the individual weights through regression methods is based on the relative past performance of the forecasts to be combined. Therefore, the efficient estimation of the individual weights is reliable when the number of forecasts is relatively small and a sufficiently large number of observations are available.

However, the number of institutions providing forecasts has increased considerably in the last years thus such projection involves the estimation of a large number of parameters. Therefore, when trying to combine forecasts through regression procedures (OLS) a dimensionality problem arises.

Although there are some techniques trying to extract relevant information from a large number of forecasts such as the subset selection, ridge regression (Fang 2003), factor-based methods (Chan *et al.* 1999, Stock and Watson, 2002), Latent Root Regression (Gerard and Clemen, 1989), srinkage methods (Aiolfi y Timmerman 2006), LASSO (de Mol *et al*., 2008, Conflotti et al 2012), it is usual to use the simple arithmetic mean in order to summarize the individual forecasts. This strategy appears to be justified empirically by the fact that the resulting simple averaging schemes tend to outperform more sophisticated ones (Genre *et al*. (2013), Stock and Watson (2004), Makridakis et al (1982), Makridakis and Winkler (1983), and Smith and Wallis (2009) are five notable studies highlighting the empirical success of the equal weighted combination). Such phenomenon is usually referred to as the “forecasting combination puzzle” and has been recently documented by Genre, Kenny, Meyer and Timmermann (2013), who show that the simple equal-weight averages constitute a Benchmark.

Theoretically, the use of the simple arithmetic mean could be justify when all the forecasters have the same past performance, or when we do not have enough information about that.

A natural question is how to combine individual forecasts even in the case of we have small information about individual forecast’s past performance. This drawback of the combination forecast is one of the potential problems which we address in this paper. In that situation, some type of forecast combination can be carried out using all the forecasts but using weights that discriminate “bad” forecasts of “good” forecasts.

In this paper, we provide an explicit framework for combining forecast based on the Data-Weighted Prior (DWP) Estimator proposed by Golan (Journal of Econometrics, 2001) that allows for simultaneous parameter estimation and forecasters selection in linear statistical models. In particular, we examine the ability of the DWP Estimator to effectively select relevant forecasts among all forecasts. In particular, we examine the ability of DWP to effectively discriminate the sequences of forecasters into groups of "good" and "bad" forecasts. We test the validity of the proposed model by a simulation exercise and compare its ex ante forecasting performance with other combining methods.

The simulation results suggest that the proposed method dominates other forecast combination strategies which are examined, as equal-weight averages or ordinal least square methods, among others.

The paper is organized in five more sections. Next section introduces some basic concepts regarding the general framework of forecast combination methods. Section 3 presents the Data-Weighted Prior (DWP) estimator. Section 4 shows an simulation experiment and discusses the main results. Finally, Section 5 concludes the paper.

2. Forecast combination methods based on OLS approach

As noted earlier, the vast amount of prospective sources and methods provides a wide variety of forecasts for any given economic variable. Once the alternative h-step forecasts for an economic variable y_{t+h} are available at time t, the theory suggests the convenience of combining the individual results $\boldsymbol{x} = (x_{1t}, \dots, x_{Kt})$ to obtain an aggregated prediction $\hat{y}_t$ through a vector of weights $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)'$ that calibrates different degrees of experts´ ability. We denote by x_{it} the forecast referred to t+h, provided in period t by an organization i (i=1..., K).

The pioneers in the theoretical study of the combination of forecasts are Bates and Granger (1969) who propose techniques to obtain a combined forecast from linear combinations of two individual forecasts, whose weights are obtained from their forecast error variances. This proposal was extended by Newbold and Granger (1974), leading to a combined forecast given by $\hat{y}_t = \boldsymbol{x\beta}$, where $\boldsymbol{l'\beta} = \boldsymbol{1}$ and $0 \leq \beta_i \leq 1$, $\boldsymbol{l}$ being a vector (Kx1) of ones. The variance of the error of the combined forecast is diminished since

$$\widehat{\boldsymbol{\beta}} = \frac{(\Sigma^{-1} l)}{(l'\Sigma^{-1} l)} \quad \text{where} \sum = E(\boldsymbol{e_t}\boldsymbol{e_t'}) \text{ and } \boldsymbol{e_t} = y\boldsymbol{l'} - \boldsymbol{x} \tag{1}$$

This method essentially ignores any correlation in the errors of the individual forecasts Granger and Ramanathan (1984) show that the weights obtained by conventional methods of combination can be interpreted as a vector of coefficients of the linear projection of y_t from the K forecasts:

$$y_{t+h} = \boldsymbol{x\beta} + e_{t+h} \tag{2}$$

where y_{t+h} is unobservable and the vector of weights is estimated from past observations of the variable $\boldsymbol{y} = (\, y_1, y_2, \dots, y_T)$ and experts' past performances $\boldsymbol{X} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_K})$:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{3}$$

where $\boldsymbol{y}$ is a $(T \times 1)$ vector of observations for y, $\boldsymbol{X}$ is a $(T \times K)$ matrix of experts' past performances, being each $\boldsymbol{x_i}$ a Tx1 vector of individual past forecasts, $\boldsymbol{\beta}$ is the $(K \times 1)$ vector of unknown parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$ to be estimated, and $\boldsymbol{\epsilon}$ is a $(T \times 1)$ vector with the random term of the linear model.

The work of Granger and Ramanathan marked the beginning of the development of more sophisticated econometric methods for combining forecasts. Thus, time varying combining weights proposed in the variance-covariance context by Granger and Newbold (1973) were introduced in the regression context by Diebold and Pauly (1987). Moreover, non linear regression (Deutsch *et al.*, 1994) and serially correlated errors are considered in dynamic combined regressions (Coulson and Robins, 1993), and the problem of non stationarity is considered by Hallman and Kamstra (1989) and Miller *et al.* (1992) among others.

However, the number of institutions providing forecasts has increased considerably in the last years thus (3) projection involves the estimation of a large number of parameters. This implies loss of degrees of freedom and poor forecast ("curse of dimensionality problem").

In such case, it is usual to use the simple arithmetic mean in order to summarize the individual forecasts.

A natural question is how to combine individual forecasts even in the case of we have small information about individual forecast's past performance. Capistrán and Timmermann (2009) show how an affine transformation of the equal weighted forecast performers reasonably well in small samples (entry and exit individual forecasters renders conventional least squares regression approach infeasible); More recently, Moreno and Lopez (2013) cite evidence in support of an alternative that allows the calibration of individual forecast when the small amount of information available does not allow the use of regression procedures.

3. Entropy Econometrics: A data-weighted prior (DWP) estimator

In this section we propose the application of an extension of the Generalized Cross Entropy (GCE) technique in the context of combining individual predictors, given that it has interesting properties when dealing with ill-conditioned datasets (small samples or data sets affected by large collinearity), which are relatively frequent when estimating regional indicators.[1]

Let us suppose a variable y that depends on K explanatory variables x_i:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{4}$$

where $\boldsymbol{y}$ is a $(T \times 1)$ vector of observations for y, $\boldsymbol{X}$ is a $(T \times K)$ matrix of observations for the x_i variables, $\boldsymbol{\beta}$ is the $(K \times 1)$ vector of unknown parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)'$ to be estimated, and $\boldsymbol{\epsilon}$ is a $(T \times 1)$ vector with the random term of the linear model. Each β_i is assumed to be a discrete random variable. We assume that there is some information about its $M \geq 2$ possible realizations. This information is included for the estimation by means of a support vector $\boldsymbol{b}' = (b_1, \dots, b_M)$ with corresponding probabilities $\boldsymbol{p'}_i = (p_{i1}, \dots, p_{iM})$. The vector $\boldsymbol{b}$ is based on the researcher's a priori belief about the likely values of the parameter. For the sake of convenient exposition, it will be assumed that the M values are the same for every parameter, although this assumption can easily be relaxed. Now, vector $\boldsymbol{\beta}$ can be written as:

[1] In Golan *et al.* (1996) or Kapur and Kesavan (1992) extensive descriptions of the entropy estimation approach can be found.

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} = \boldsymbol{BP} = \begin{bmatrix} \boldsymbol{b}' & 0 & \cdots & 0 \\ 0 & \boldsymbol{b}' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{b}' \end{bmatrix} \begin{bmatrix} \boldsymbol{p_1} \\ \boldsymbol{p_2} \\ \vdots \\ \boldsymbol{p_K} \end{bmatrix} \tag{5}$$

where $\boldsymbol{B}$ and $\boldsymbol{P}$ have dimensions $(K \times KM)$ and $(KM \times 1)$ respectively. Now, the value of each parameter β_i is given by the following expression:

$$\beta_i = \boldsymbol{b}' \boldsymbol{p_i} = \sum_{m=1}^{M} b_m \, p_{im}; \; i = 1, \dots, K \tag{6}$$

For the random term, a similar approach is followed. Oppositely to other estimation techniques, GCE does not require rigid assumptions about a specific probability distribution function of the stochastic component, but it still is necessary to make some assumptions. $\boldsymbol{\epsilon}$ is assumed to have mean $E[\boldsymbol{\epsilon}] = 0$ and a finite covariance matrix. Basically, we represent our uncertainty about the realizations of vector $\boldsymbol{\epsilon}$ treating each element ϵ_t as a discrete random variable with $J \geq 2$ possible outcomes contained in a convex set $\boldsymbol{v}' = \{v_1, \dots, v_J\}$, which for the sake of simplicity is assumed as common for all the ϵ_t. We also assume that these possible realizations are symmetric around zero $(-v_1 = v_J)$. The traditional way of fixing the upper and lower limits of this set is to apply the three-sigma rule (see Golan et al. (1996) or Kapur and Kesavan (1992) Pukelsheim, 1994). Under these conditions, vector $\boldsymbol{\epsilon}$ can be defined as:

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_T \end{bmatrix} = \boldsymbol{VW} = \begin{bmatrix} \boldsymbol{v}' & 0 & \cdots & 0 \\ 0 & \boldsymbol{v}' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{v}' \end{bmatrix} \begin{bmatrix} \boldsymbol{w_1} \\ \boldsymbol{w_2} \\ \vdots \\ \boldsymbol{w_T} \end{bmatrix} \tag{7}$$

and the value of the random term for an observation *t* equals:

$$\epsilon_t = \boldsymbol{v}' \boldsymbol{w_t} = \sum_{j=1}^{J} v_j \, w_{tj}; \; t = 1, \dots, T \tag{8}$$

and, consequently, model (7) can be transformed into:

$$\boldsymbol{y} = \boldsymbol{XBp} + \boldsymbol{Vw} \tag{9}$$

In this context, we would need also to estimate the elements of matrix $\boldsymbol{W}$ (denoted by $\widetilde{w}_{tj}$) and the estimation problem for the general linear model is transformed into the estimation of $K + T$ probability distributions. Basing on this idea, Golan (2001) proposed an estimator that allows for simultaneous parameter estimation and variable selection in linear statistical models. Related to the Bayesian Method of Moments (BMOM, see Zellner, 1996, 1997), this data-based method estimator uses both sample and non-sample information in determining a basis for coefficient reduction and extraneous variable identification. In other words, this technique allows for shrinking the coefficient of the explanatory variables that can be classified as irrelevant in the linear model. A recent empirical application of this method can also be found in Bernardini-Papalia (2008) or Fernandez-Vazquez and Rubiera (2103).

The objective of the DWP estimator, in the context of combining forecasters, is to identify which predictors should receive a weight significantly different from the weighting scheme specified in a simple arithmetic mean and simultaneously to forecast the target variable based on a combination of individual predictors. We start by specifying a discrete support space $\boldsymbol{b}$ for each β_i (and the same for $\boldsymbol{v}$) symmetric around the value $1/K$, with large lower and upper bounds for $\boldsymbol{b}$ and the three-sigma rule for $\boldsymbol{v}$, so that each β_i and ϵ_t are contained in the chosen interval with high probability. For the estimation of the β_i parameters, the specification of some a priori distribution $\boldsymbol{q}$ for the values in the supporting vectors is required. Besides fixing an uniform probability distribution that will be used as $\boldsymbol{q}$ in the GCE estimation (i.e.; $q_m = \frac{1}{M}$)., we also specify a "spike" prior for each β_i, with a very high probability at $b_m = 1/K$ (i.e.; $q_m \cong 1$ if $b_m = 1/K$ and $q_m \cong 0$ for the remaining values). Thus, a flexible, data-based prior is specified such that for each β_i coordinate either a spike prior at the $b_m = 1/K$, a uniform prior over the discrete support space $\boldsymbol{b}$, or any convex combination of the two can result. If we denote with $\boldsymbol{q}^{\boldsymbol{u}}$ and $\boldsymbol{q}^{\boldsymbol{s}}$ the uniform and spike a priori distributions respectively, the objective proposed can be achieved by minimizing the following constrained problem:

$$
\begin{aligned}
\underset{\boldsymbol{P},\boldsymbol{P}^{\gamma},\boldsymbol{W}}{\text{Min}} D(\boldsymbol{P},\boldsymbol{P}^{\gamma},\boldsymbol{W} \| \boldsymbol{Q},\boldsymbol{Q}^{\gamma},\boldsymbol{W}^{\mathbf{0}}) = & \sum_{i=1}^{K}(1-\gamma_i)\sum_{m=1}^{M} p_{im}\, ln\left(\frac{p_{im}}{q_{im}^{u}}\right) \\
& + \sum_{i=1}^{H}\gamma_i \sum_{m=1}^{M} p_{im}\, ln\left(\frac{p_{im}}{q_{im}^{s}}\right) \\
& + \sum_{i=1}^{K}\sum_{n=1}^{N} p_{in}^{\gamma}\, ln\left(\frac{p_{in}^{\gamma}}{q_{in}^{\gamma}}\right) \\
& + \sum_{t=1}^{T}\sum_{j=1}^{J} w_{tj}\, ln\left(\frac{w_{tj}}{w_{tj}^{0}}\right)
\end{aligned}
\tag{10}
$$

subject to:

$$y_t = \sum_{i=1}^{K}\sum_{m=1}^{M} b_m p_{im} x_{it} + \sum_{j=1}^{J} v_j\, w_{tj}; \;\; t = 1, \dots, T \tag{11}$$

$$\sum_{m=1}^{M} p_{im} = 1\,; \; i = 1, \dots, K \tag{12}$$

$$\sum_{j=1}^{J} w_{tj} = 1\,; \; t = 1, \dots, T \tag{13}$$

$$\sum_{n=1}^{N} p_{in}^{\gamma} = 1\,; \; i = 1, \dots, K \tag{14}$$

$$\gamma_i = \sum_{n=1}^{N} b_{in}^{\gamma}\, p_{in}^{\gamma} \tag{15}$$

The γ_i parameters are estimated simultaneously with the rest of β_i coefficients of the model in (10). Each γ_i measures the weight given to the spike prior $\boldsymbol{q^s}$ for each parameter β_i and it is defined as $\tilde{\gamma}_{\mathrm{i}} = \sum_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{b}_{\mathrm{in}}^{\gamma}\, \tilde{\mathrm{p}}_{\mathrm{in}}^{\gamma}$, where $b_{i1}^{\gamma} = 0$ and $b_{iN}^{\gamma} = 1$ are respectively the lower and upper bound defined as the support of these parameters. The *a priori* probability distributions fixed for them are always uniform ($q_i^{\gamma} = \frac{1}{N}\ \forall n = 1, .., N$) and the same is applied for the errors (again $w_{tj}^{0} = \frac{1}{J}\ \forall j = 1, .., J$).

To understand the logic of this data-weighted prior (DWP) estimator an explanation on the objective function of the previous minimization program is required. Note that equation (10) is divided in four terms. The last term is exactly the same as in the GCE program and it measures the Kullback divergence between the posterior and the prior probabilities for the noise component of the model. The first term quantifies the divergence between the recovered probabilities and the uniform priors for each β_i coefficient, being this divergence weighted by $(1-\gamma_i)$. On the contrary, the second element of (10) measures the divergence with the spike prior and it s weighted by γ_i. The third element in (10) relates to the Kullback divergence of the weighting parameters γ_i.

From the recovered $\tilde{p}_{im}$ probabilities, the estimated value of each parameter β_i is obtained as:

$$\tilde{\beta}_i = \sum_{m=1}^{M} b_m \, \tilde{p}_{im}; \; i = 1, \dots, K \tag{16}$$

Under some mild assumptions (see Golan 2001, page 177) the consistency and asymptotic normality of the DWP estimates can be ensured. Additionally, these assumptions also guarantee that the approximate variances of the DWP estimator is lower than the approximate variance of the GCE estimator, which in turn is lower than the approximate variance of a ML-LS estimator (see Golan, 2001, page 179).

Simultaneously to the estimation of the parameters of the model, the DWP estimator allows for discriminating between predictors. The proposed estimation strategy provides two indications for this objective. Firstly, estimates of the weighting parameters γ_i, obtained as:

$$\tilde{\gamma}_i = \sum_{n=1}^{N} b_{in}^{\gamma} \, \tilde{p}_{in}^{\gamma}; \; i = 1, \dots, K \tag{17}$$

can be used as a tool for this purpose: as $\tilde{\gamma}_i \to 0$ the prior becomes more uniform and the estimates approach those of the GME estimator. This indicates that the parameter associated to this predictor can take values far from the center of the support vector (i.e., $1/K$). On the contrary, for large values of $\tilde{\gamma}_i$ the part of the objective function with spike prior on $1/K$ takes over. Consequently, the predictors considered in the combination that should receive a weight equal to those in a simple arithmetic mean will be characterized by large values of $\tilde{\gamma}_i$ (Golan considers sufficiently large values when $\tilde{\gamma}_{ih} > 0\ .49$) together with estimates of β_i close to $1/K$.

Moreover, a χ^2 statistic can be constructed in order to test if the estimate for β_i is significantly different from $1/K$. The basic idea is to test if the recovered $\tilde{p}_{im}$ are significantly different from the respective spike prior q_{im}^s. The Kullback-Leibler divergence between our posterior and these a priori probabilities is:

$$D_i(\tilde{\boldsymbol{p}}_i \| \boldsymbol{q}_i^s) = \sum_{m=1}^{M} \tilde{p}_{im}\, ln\left(\frac{\tilde{p}_{im}}{q_{im}^s}\right) \tag{18}$$

And the chi-squared divergence between both distributions is:

$$\chi_{M-1}^2 = M \sum_{m=1}^{M} \frac{(\tilde{p}_{im} - q_{im}^s)^2}{q_{im}^s} \tag{19}$$

A second-order approximation of $D_h(\tilde{\boldsymbol{p}}_h \| \boldsymbol{q}_h^s)$ is the entropy-ratio statistic for evaluating $\tilde{\boldsymbol{p}}_h$ versus $\boldsymbol{q}_h^s$:

$$D_i(\tilde{\boldsymbol{p}}_i \| \boldsymbol{q}_i^s) \cong \frac{1}{2} \sum_{m=1}^{M} \frac{(\tilde{p}_{im} - q_{im}^s)^2}{q_{im}^s} \tag{20}$$

Consequently,

$$2MD_i(\tilde{\boldsymbol{p}}_i \| \boldsymbol{q}_i^s) \to \chi_{M-1}^2 \tag{21}$$

Given this relationship, we can use the measure $2MD_i(\tilde{\boldsymbol{p}}_i \| \boldsymbol{q}_i^s)$ in order to test the hypothesis $H_0 \colon \beta_i = 1/K$. If the null hypothesis is not rejected, a predictor x_i that should be weighted according with a simple mean is identified.[2]

[2] To prevent computational problems that appear when computing log(0), in the empirical application on the next section the spike priors $\boldsymbol{q}_i^u$ have been specified with a point mass at zero equal to 0.999 and 0.0005 respectively for the other points of the support vectors.

4. A simulation study

In this section we try to find some empirical evidences, by means of some numerical simulations, on the comparative performance of the proposed DWP estimator in the context of combining individual forecasts. The point of departure of the experiment is the unknown series y_t $(t = 1, \dots, T)$ that contains the target variable of a given region and a $(T \times K)$ matrix $\boldsymbol{X}$ with K potential unbiased forecasters of this series along the T time periods. The basic idea is that $\boldsymbol{X}$ should contain some imperfect information on the target series. Specifically, in the experiment the elements of $\boldsymbol{X}$ will be generated in the following way:

$$x_{it} = y_t + u_{it};\ t = 1, \dots, T;\ i = 1, \dots, K \tag{22}$$

Where $\boldsymbol{u_i} \sim N(0, \sigma_i)$ is a noise term that reflects the accuracy of $\boldsymbol{x_i}$ as a forecaster of $\boldsymbol{y}$ and σ_i is a scalar that adjusts the variability of this noise. Note that σ_i indicates the degree of information on the target series that is contained in predictor $\boldsymbol{x_i}$, i. e., the higher the value of σ_i, the less informative about $\boldsymbol{y}$ is $\boldsymbol{x_i}$.

Given that in our numerical experiment we would like to replicate situations normally observed in the context of forecasting regional series, instead of numerically generate the values of our target variable $\boldsymbol{y}$, we opted for taking actual values of a regional indicator. More specifically, we have taken the annual GVA rate of change of the region of Catalonia (Spain) from 1980 to 2008. We have extracted this information (at constant prices of 2000) from the BDmores database.[3]

Concerning the configuration of matrix $\boldsymbol{X}$, we consider a different numbers of potential predictors (dimension K) to be combined. Given that in the context of forecasting regional indicators, the number of forecasters is normally smaller than when national or supra-national variables are predicted, we have set three different values for K, being K set to 6, 12 and 24. Moreover, we have considered that the behavior of these predictors can be heterogeneous when aiming at forecasting variable $\boldsymbol{y}$. In particular, we have divided our set of K forecasters in two different subsets, which can be classified as "good" or "bad" predictors. The logic of this idea is that the information that the predictors provide for

[3] This database is elaborated by the Spanish Ministry of Economy. More details can be found in: http://www.sepg.pap.minhap.gob.es/sitios/sepg/en-GB/Presupuestos/Documentacion/paginas/basesdatosestudiosregionales.aspx

forecasting variable $\boldsymbol{y}$ can vary among them, being a "good" predictor preferable to a "bad" one, but still the comparatively "bad" forecaster containing some potentially useful information to be considered in the combination. In order to reflect this idea, the elements of matrix $\boldsymbol{X}$ will be generated differently in the following two subsets:

$$x_{it} = y_t + u_{it}^{g};\ t = 1, \dots, T;\ i = 1, \dots, G \tag{23}$$

$$x_{it} = y_t + u_{it}^{b};\ t = 1, \dots, T;\ i = G + 1, \dots, K \tag{24}$$

Where u_{it}^{g} is the noise term for the subset of G "good" predictors and u_{it}^{b} is the corresponding element for the comparatively "bad" ones. The difference between u_{it}^{g} and u_{it}^{b} is on its variability, since:

$$\boldsymbol{u_i^g} \sim N\left(0, \frac{s}{2}\right) \tag{25}$$

$$\boldsymbol{u_i^b} \sim N(0, s) \tag{26}$$

Being s the standard deviation in the sample 1980-2008 of the target variable $\boldsymbol{y}$. Equations (25) and (26) indicate that the variance of the forecasters classified as "good" present a variance 4 times lower than the classified as "bad".

In the simulation we have set different proportions between these two subsets of predictors. First, a more realistic situation where 5/6 of the total of K forecasters belong to the group of "good" predictors and only 1/6 are classified as "bad". Additionally and for comparative purposes, a situation where they are distributed in equal parts (50%) to each group is considered as well.

In the experiment, all the simulated predictors are combined through the regression-based method of combining forecasts:

$$y_t = \sum_{i=1}^{K} \beta_i x_{it} + e_{it};\ t = 1, \dots, T \tag{27}$$

being the target of the different method for combining these forecasters to determine the best possible values for the $\beta's$ parameters.

The benchmark for comparing the competing methods will be the arithmetic mean of the forecasters, where $\beta_i = 1/K, \forall i$. The simple mean of individual predictors is normally taken as a valid reference in the recent literature on combination of forecasters, being sometimes considered the best way of combining information of individual predictors. For example, Genre *et al*. (2013), Stock and Watson (2004), Makridakis et al (1982), Makridakis and Winkler (1983), and Smith and Wallis (2009) are five notable studies highlighting the empirical success of the equal weighted combination.

Additionally, a restricted Least Squares weight scheme (see Granger and Ramanathan, 1984, for the original unrestricted LS approach; or Timmerman, 2006 for the restricted version) is considered as well, where the $\beta's$ weights (restricted to sum to one) are estimated by minimizing the sum of squared errors e_{it}.

Our comparison is extended in order to include the proposals made in recent forecasting literature, where forecasts based on Bayesian Model Averaging (BMA) has received considerable attention (see Buckland et al. 1997; or Burnham and Anderson, 2002). In this approach, the weights are determined basing on the Bayesian Information Criterion (BIC) as:

$$\beta_i = \frac{exp\left[-\frac{1}{2}BIC_i\right]}{\sum_{i=1}^{K} exp\left[-\frac{1}{2}BIC_i\right]}; \tag{28}$$

And

$$BIC_i = Tln(\hat{\sigma}_i^2) + ln(T) \tag{29}$$

Where $\hat{\sigma}_i^2$ stands for the LS estimation of σ_i^2.

These techniques for combining the individual predictors $\boldsymbol{x_i}$ will be compared with the estimation of the optimal $\beta's$ weights when the DWP estimator is applied. Consequently, specifying some support for the set of parameters to estimate and the errors is required. We have fixed the same vector $\boldsymbol{b}$ for all the $\beta's$ parameters. In particular, the proposed DWP estimator assumes as a prior value for each β_i the solution provided by the simple mean of forecasters, where all are equally weighted as $1/K$. More specifically, we have considered $M = 3$ with vectors $\boldsymbol{b}' = (1/K - 1, 1/K , 1/K + 1)$; in other words, the bouds with the minimum and maximum possible values for the weights are set as the center $1/K \pm 1$.

For the weighting parameters γ_i we fixed supporting vectors composed only by $N = 2$ values $\boldsymbol{b}' = (0,1)$. Finally, the usual three-sigma rule (with the sample standard deviations of the dependent variable $\boldsymbol{y}$) has been applied for specifying the supports of the error terms.

Table 1 and Table 2 summarize the results of comparing the actual target values of our variable of interest (y_t) with the combined forecasts ($\hat{y}_t$) obtained according to the different methods, following two different deviation measures: (i) the mean squared forecast errors (MSFE); and (ii), the mean absolute percentage forecast error (MAPFE), being respectively defined by the two following expressions:

$$MSFE = \sum_{f=1}^{F} \left(y_f - \hat{y}_f\right)^2 \tag{30}$$

$$MAPFE = 100 \sum_{f=1}^{F} \left|y_f - \hat{y}_f\right| \tag{31}$$

The mean values of these deviation measures are computed along 1,000 trials and for a forecast horizon of 4 periods ahead ($f = 1, \ldots ,4$), which means that the last 4 periods in our sample are not included in the estimation of the weights but taken as reference for evaluating the performance of our combination of predictions.

Table 1. Mean Squared Forecasting Error (MSFE); 1,000 trials

Mean Squared Forecasting Error (MSFE)

		Method			
K	**G**	**mean**	**LS**	**BIC**	**DWP**
6	5 good	2.1856	2.0951	2.7123	2.1841
	3 good	2.8403	2.493	3.1027	2.6736
12	10 good	1.478	1.7988	2.5636	1.3896
	6 good	1.9238	2.1185	2.8167	1.8117
24	20 good	1.0749	5.2904	2.5438	0.9863
	12 good	1.365	7.1449	2.7047	1.2655

Table 2. Mean Absolute Percentage Forecasting Error (MAPFE); 1,000 trials

Mean Absolute Percentage Forecasting Error (MAPFE)

		Method			
K	**G**	**mean**	**LS**	**BIC**	**DWP**
6	5 good	2.1856	2.0951	2.7123	2.1841
	3 good	2.8403	2.493	3.1027	2.6736
12	10 good	1.478	1.7988	2.5636	1.3896
	6 good	1.9238	2.1185	2.8167	1.8117
24	20 good	1.0749	5.2904	2.5438	0.9863
	12 good	1.365	7.1449	2.7047	1.2655

Error figures in Table 1 and Table 2 show how the simple mean outperform the combining methods based on some regression analysis (LS or BIC) in situations where the number of potential forecasters is large in relative terms to the available sample size. When the predictors considered are 12 or 24, the combination based on LS and BIC present problems derived from an ill-conditioned dataset (too many parameters to estimate from a relatively small sample size), whereas the arithmetic mean of predictors is not affected by this problem. The proposed DWP estimator seems to beat the competing combination techniques, given that it takes the weighting scheme as the arithmetic mean and only departs from these weights if the information contained in the sample provides a strong empirical evidence on this respect. On the contrary, when the number of predictor is low, a LS-based combination of forecasters performs better than any of the other techniques, given that now the sample size is large enough in relative terms to the number of predictors considered.

We would like to point out, that we have only evaluated the results of the combined forecast under an accuracy criterion (forecast error). When additional criteria are considered (such as error variance, error distribution asymmetry or error correlation) combining forecasts becomes clearly a multi-attribute decision problem and using different criteria leads to distinct preferences (de Menezes *et al.* 2000).

5. Conclusions

The most common alternative for combining individual forecasts consists in computing the simple arithmetic mean, which implies assuming that all the individuals are equally important. Theoretically, the use of the simple arithmetic mean could be justify when all the forecasters have the same past performance or when we do not have enough information about that (which implies not taking advantage of the information available about each forecaster's ability).

The paper proposes the use of an entropy-based technique estimator to obtain an affine transformation of the equal weighted forecast combination: A data-weighted prior (DWP) estimator.

We test the validity of the proposed model by a simulation exercise and compare its ex ante forecasting performance with other combining methods. The benchmark for comparing the competing method are the arithmetic mean of the forecasters, a restricted

Least Squares and weight scheme forecasts based on Bayesian Model Averaging (where the weights are determined basing on the Bayesian Information Criterion).

We set three different values for the number of individual forecasts to be combined (6, 12 and 24) and we have divided our set of forecasters in two different subsets, which can be classified as “good” or “bad” predictors.

The simulation results indicate that the proposed DWP estimator seems to beat the competing combination techniques, given that it takes the weighting scheme as the arithmetic mean and only departs from these weights if the information contained in the sample provides a strong empirical evidence on this respect. The most relevant advantage of this estimator is that, even in situations characterized by a large number of forecasters, the DWP estimator generate a better set of recovered forecasters´ weights than arithmetic mean which is capable to identify groups of forecasters into groups of "good" and "bad" forecasts.

REFERENCES

Agnew, CE. (1985): “Bayesian consensus forecast of macroeconomic variables”, *Journal of Forecasting* 4: 363-376.

Aiolfi, M.; Timmerman, A. (2006) “Persistence in forecasting performance and conditional combination strategies”, *Journal of Econometrics*, 135: 31-53.

Anandalingam. G.; Chen, L. (1983): “Linear combination of forecasts: a general Bayesian model”, *Journal of Forecasting*, 8: 199-214.

Bates, J.M.; Granger, C.W.J. (1969): “The Combination of Forecasts”, *Operational Research Quarterly*, 20: 451-468.

Bernardini-Papalia, R. (2008): “A Composite Generalized Cross Entropy formulation in small samples estimation”, *Econometric Reviews*, 27: 596-609

Bordley, R.F. (1982): “The combination of forecast: a Bayesian approach”, *Journal of Operational Research Society*, 33: 171-174.

Buckland, S. T.; Burnham, K. P.; Augustin, N.H. (1997): “Model selection: an integral part of inference”, *Biometrics*, 53: 603-618.

Bunn, D.W. (19875): “A Bayesian approach to the linear combination of forecasts”, *Operational Research Quarterly*, 26: 325-329.

Bunn, D.W. (1989):”Forecasting with More than One Model”, *Journal of Forecasting,* 8: 161-166.

Burnham, K.P.; Anderson, D.R. (2002): *Model selection and multimodel inference: a practical information-theoretic approach.* Springer: New York

Capistrán, C.; Timmermann, A. (2009): “Forecast combination with entry and exit of expert”, *Journal of Business and Economic Statistics*, 27(4): 428–440.

Chan, Y.; Stock, J.; Watson, M. A. (1999): “Dynamic Factor Model Framework for Forecast Combination”, *Spanish Economic Review*, 1: 91–121.

Clemen, R.T. (1989):”Combining forecasts: A review and annotated bibliography”, *International Journal of Forecasting*, 5: 559–583

Clemen, R.T.; Winkler, R.L. (1999): “Aggregating point estimates: A flexible modelling approach”, *Management Science*, 39: 501-515.

Conflitti, C.; De Mol, C.; Giannone, D. (2012): *Optimal Combination of Survey Forecasts*, working paper CEPR Discussion Papers 9096, C.E.P.R. Discussion Papers.

Coulson, N.E.; Robins, R.P. (1993).”Forecast Combination in a Dynamic Setting”, *Journal of Forecasting,* 12: 63-67.

De Menezes, L.M.; Bunn, D.W.; Taylor, J.W. (2000): “Review of Guidelines for the Use of Combined Forecasts”, *European Journal of Operational Research*, 120: 190-204.

De Mol, C.; Giannone, D.; Reichlin, L (2008): “Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?” *Journal of Econometrics*, 146: 318-328.

Diebold, F.X.; Pauly, P. (1987):”Structural change and the combination of forecast”, *Journal of Forecasting*, 6: 21-40.

Fang, Y. (2003): “Forecasting combination and encompassing tests”, *International Journal of Forecasting*, 19: 87-94.

Fernandez-Vazquez, E.; Rubiera-Morollon, F. (2013). "Estimating Regional Variations of R&D Effects on Productivity Growth by Entropy Econometrics," *Spatial Economic Analysis*, 8(1): 54-70.

Genre, V.; Kenny, G.; Meyler, A.; Timmermann, A. (2013): “Combining expert forecasts: Can anything beat the simple average?” *International Journal of Forecasting*, 29: 108-121.

Golan, A. (2001): "A Simultaneous Estimation and Variable Selection Rule," *Journal of Econometrics,* 101 (1): 165-193.

Golan, A.; Judge, G.; Miller, D. (1996): *Maximum Entropy Econometrics: robust estimation with limited data.* John Wiley & Sons Ltd: London

Granger, C.W.J.; Newbold, P. (1973): “Some comments on the evaluation of economic forecasts”, *Applied Economics*, 5: 35-47.

Granger, C.W.J.; Ramanathan, C. (1984): “Improved Methods of Combining Forecasts”, *Journal of Forecasting*, 3: 197-204.

Greer, M.R. (2005): “Combination forecasting for directional accuracy: An application to survey interest rate forecasts”, *Journal of Applied Statistics*, 32: 607-615.

Guerard, J. B.; Clemen, R. T. (1989): “Collinearity and the use of latent root regression for combining GNP forecasts”, *Journal of Forecasting,* 8: 231-238.

Hallman, J.; Kamstra, M. (1989): “Combining algorithms based on robust estimation techniques and co-integration restrictions”, *Journal of Forecasting*, 8: 189-198.

Holden, K.; Peel, D.A. (1988): “Combining Economic Forecasts”, *Journal of the Operational Research Society*, 39: 1005-1010.

Kapur, J.N.; Kesavan, H.K. (1992): *Entropy Optimization Principles with Applications.* Academic Press: New York.

Makridakis, S. A.; Andersen, R.; Carbone, R.; Fildes, M.; Hibon, R.; Lewandowski, J.; Newton, E.; Parsen, and R. Winkler (1982) “The accuracy of extrapolation (time series) methods: results of a forecasting competition”, *Journal of Forecasting*, 1: 111- 153

Makridakis, S.; Winkler, R.L. (1983): “Averages of Forecasts: some empirical results”, *Management Science*, 29: 987-996.

Marcellino M. Forecast Pooling for European Macroeconomic Variables. *Oxford Bulletin of Economics and Statistics* 2004*;* 66: 91-112.

Miller, C.M.; Clemen, R.T.; Winkler, R.L. (1992): “The effect of nonstationarity on combined forecasts.”, *International Journal of Forecasting*, 7: 515-529.

Moreno, B.; López, A.J. (2013): “Combining economic forecasts by using a Maximum Entropy Econometric”, *Journal of Forecasting*, 32(2): 124-136.

Newbold, P.; Granger, C.W.J. (1974): “Experience with Forecasting Univariate Time Series and the Combination of Forecasts”, *Journal of the Royal Statistical Society,* Serie A 137: 131-165.

Pukelsheim, F. (1994): “The three sigma rule”, *The American Statistician,* 48: 88–91

Smith, J.; Wallis, K.F. (2009) “A Simple Explanation of the Forecast Combination Puzzle”, *Oxford Bulletin of Economics and Statistics*, 71(3): 331-355.

Stock, J. H.; Watson, M. W. (2002): “Forecasting Using Principal Components from a Large Number of Predictors,” *Journal of the American Statistical Association*, 97: 147–162

Stock, J.H.; Watson, M.W. (2004) “Combining Forecasts of Output Growth in a Seven-Country Data Set”, *Journal of Forecasting*, 23: 405-430.

Timmermann, A. (2006): Forecast Combinations. Elliott G, Granger CWJ, Timmermann A. (eds.) *Handbook of Economic Forecasting*, Volume 1, 2006, 135-196: Amsterdam: Noth-Holland.

Winkler, R.L. (1981):”Combining probability distributions from dependent information sources”, *Management Science*, 27: 479-488.

Winkler, R.L.; Makridakis, S. (1983): “The combination of forecasts”, *Journal of the Royal Statistical Society*, Serie A 146: 150-157.

Zellner, A. (1996): “Bayesian Method of Moments/ Instrumental Variable (bmom/iv) Analysis of Mean and Regression Models,” In J.C. Lee, W.C. Johnson, and A. Zellner, (eds.), *Modeling and Prediction: Honoring Seymour Geisser*, Springer-Verlag, 61-75.

Zellner, A. (1997): “The Bayesian Method of Moments (BMOM): Theory and Applications,” in, T. Fomby and R.C. Hill (eds.), *Advances in Econometrics*, Vol. 12: 85-106.