entropy

MDPI

Article

A Data-Weighted Prior Estimator for Forecast Combination

Esteban Fernández-Vázquez [1,*], Blanca Moreno [1] and Geoffrey J.D. Hewings [2]

[1] REGIOlab and Department of Applied Economics, University of Oviedo, Faculty of Economics and Business, Avda. del Cristo, s/n, 33006 Oviedo, Spain; morenob@uniovi.es

[2] Regional Economics Applications Laboratory, University of Illinois at Urbana-Champaign 607 S. Matthew, Urbana, IL 61801-367, USA; hewings@illinois.edu

* Correspondence: evazquez@uniovi.es

Received: 7 February 2019; Accepted: 18 April 2019; Published: 23 April 2019

Abstract: Forecast combination methods reduce the information in a vector of forecasts to a single combined forecast by using a set of combination weights. Although there are several methods, a typical strategy is the use of the simple arithmetic mean to obtain the combined forecast. A priori, the use of this mean could be justified when all the forecasters have had the same performance in the past or when they do not have enough information. 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 the situation of little information. With this purpose, the data-weighted prior (DWP) estimator proposed by Golan (2001) is used for forecaster selection and simultaneous parameter estimation 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 accuracy of the proposed model with a simulation exercise and compare its *ex ante* forecasting performance with other methods used to combine forecasts. The obtained results suggest that the proposed method dominates other combining methods, such as equal-weight averages or ordinal least squares methods, among others.

Keywords: data-weighted prior; generalized maximum entropy method; combined forecast

1. Introduction

Forecasting agents can use an ample variety of forecasting techniques and different information sets, thus leading to a wide variety of obtained forecasts. Hence, as each individual forecast captures a different aspect of the available information, a combination of them would be expected to perform better than the individual forecasts. In fact, a growing volume of literature has demonstrated that a combined forecast increases forecast accuracy in several fields (e.g., [1–7]).

The first study about the forecast combination was carried out by [8]. Since their study, several researchers have shown a variety of modeling procedures to estimate the weights of each individual forecast in the combined forecast (a review of the literature can be found in [5,9,10]).

There are several methods for forecast combination that can be classified as variance–covariance methods, probabilistic methods, Bayesian methods, or regression-based methods, among others. The first kind of method allows the calculation of weights of the combined forecast by minimizing the error variance of the combination ([8,11]); Probabilistic methods ([12,13]) weights are linked to the probability that an individual forecast will perform best on the next occasion; Bayesian methods, which were originally put forward by [14], assume that the variable being predicted (y) and the individual forecasts have a random character and the combined forecast is the expected value of the a posteriori

distribution of y that is modified from its a priori distribution with the sample information of the individual forecasts ([14–18], among others).

The regression-based methods were introduced by [19]. These methods link the weights of the combined forecasts to the coefficient vector of a linear regression, where individual forecasts are explanatory variables of the variable being predicted. The estimation of the coefficient vector is based on the past available information of individual forecasts and realizations of the variable being predicted. However, when the number of agents providing forecasts increases, the combined regression method involves the estimation of a large number of parameters and a dimensionality problem could arise.

In such a situation, in order to take out relevant information from a large number of forecasts, some procedures can be used, such as the subset selection, factor-based methods ([20,21]), ridge regression [22], shrinkage methods [23], latent root regression [24] or least absolute shrinkage, and the selection operator method ([25,26]), among others. Nevertheless, the simple arithmetic mean of the individual forecasts is the most used strategy to obtain the combined forecast. This strategy could be justified, as some researchers have empirically shown that simple averaging procedures dominate other, more complicated schemes ([2,27–29], among others). Such a phenomenon is usually referred to as the "forecasting combination puzzle" which has been documented by [10], who shows that the simple arithmetic mean constitutes a benchmark. From a theoretical point of view, the simple equal-weight average could be justified when all the forecasters have shown the same forecast performance in the past, or there is not available information about individual forecast´s past performance to calibrate them differently.

In such a situation of limited information, the following question arises: Could it be possible to combine individual forecasts differently from the simple average procedure? This drawback of the combination forecast is one of the potential problems which we address in this paper. In fact, under a regression-based combination method framework we propose a procedure that allows for simultaneous parameter estimation and forecast selection in linear statistical models. This procedure is based on the data-weighted prior (DWP) estimator proposed by [30]. This estimator has been previously applied to standard regression analysis, but not specifically to the field of forecast combination. More specifically, we analyze how DWP is able to reduce the number of potential forecasters and estimate a vector of weights different from the simple average in the combined forecast. We use a simulation exercise to compare the ex-ante forecasting performance of the proposed method with other combining methods, such as equal-weight averages or ordinal least square methods, among others. The obtained results indicate that the method based on DWP outperform other examined forecast combination methods.

The paper is organized in five additional sections. Section 2 introduces the framework of the regression-based combination methods. Section 3 presents the data-weighted prior (DWP) estimator. Section 4 shows the simulation experiment and presents the results. Finally, Section 5 summarizes the conclusions of the research.

2. Forecast Combination Methods Based on Regression Methods

There is a large number of individual forecasts to forecast any given variable (y) with forecast horizon h at time t, y_{t+h}. We indicate by x_{it} the forecast referred to $t+h$, given in period t by a forecasting agent or model i ($i = 1, \ldots, K$). The theory of combining forecasts indicates that it could be possible to obtain an aggregated prediction $\hat{y}_t$ that combines the individual forecasts $\boldsymbol{x} = (x_{1t}, \ldots, x_{Kt})$ through a vector of weights $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$.

The first study about forecast combination focused on the combination of two forecasts whose vector of weights was obtained from the error variances of the individual forecast [8]. Afterward, [11] showed a combined forecast obtained by $\hat{y}_t = \boldsymbol{x\beta}$, with the sum of weights is $\boldsymbol{l'\beta} = 1$, $\boldsymbol{l}$ being a vector $(K \times 1)$ of ones and $0 \leq \beta_i \leq 1$. The combined forecast reduces its error variance since:

$$\hat{\boldsymbol{\beta}} = \frac{\left(\Sigma^{-1} l\right)}{\left(l' \Sigma^{-1} l\right)}; \text{ where } \sum = E(\boldsymbol{e_t e_t'}) \text{ and } \boldsymbol{e_t} = y\boldsymbol{l'} - \boldsymbol{x}, \tag{1}$$

where $\boldsymbol{e_t}$ is the vector $(K \times 1)$ containing the forecast error specific to each forecasting agent or model i.

However, the method does not take into account the possible correlation in the errors of the forecasts being combined. [19] showed that weights of the combined forecasts obtained through conventional methods can be interpreted as the coefficient vector of the linear projection of the variable being predicted from the K individual forecasts as:

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

where y_{t+h} is the variable being predicted (unobservable). The estimation of $\boldsymbol{\beta}$ is based on the past observations of the variable $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$ and experts' past performances $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_K})$:

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

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 $T \times 1$ vector of individual past forecasts, $\boldsymbol{\beta}$ is the $(K \times 1)$ vector of unknown parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ to be estimated, and $\boldsymbol{\epsilon}$ is a $(T \times 1)$ vector with the random term of the linear model.

The combining regression-based methods introduced by [19] were extended in several ways. Thus, [31] introduced time varying combining weights and [32] introduced nonlinear specifications in combined regression context. The dynamic combined regressions were introduced by [33] to take into account the serially correlated errors. Moreover, [34,35] considered the problem of non-stationarity.

However, the number of institutions carrying out forecasts has increased considerably in the last few years, thus the projection methodology suggested by Equation (3) would involve the estimation of a large number of weights. Thus a "curse of dimensionality problem" could arise when losing degrees of freedom for the regression estimation. In such cases, it is usual to use the simple mean average of the individual forecasts as a combined forecast.

In this situation of limited information about the past performance of individual forecasts, a question that arises is how to combine individual forecasts differently from the simple mean average. Some authors have shown evidence in support of an alternative that allows the calibration of individual forecasts when the small amount of information available does not allow the use of regression procedures. In a context where entry and exit of individual forecasters makes the regression estimation unfeasible, [36] shows how an affine transformation of the uniform weighted forecast performs reasonably well in small samples. [6] proposes a combination method based on the generalized maximum entropy approach [37]. Through the application of the maximum entropy principle, their method leads the adjustment of a priori weights (which are associated with the simple mean average) into posterior weights by considering a large number of forecasters, for which there is limited available information about their past performances.

3. A Data-Weighted Prior (DWP) Estimator

Generalized cross entropy (GCE) technique has interesting properties when dealing with ill-conditioned datasets (those affected by significant collinearity or small samples) An extensive description of the entropy estimation approach can be found in [37,38]. Thus, in this section we propose the application of an extension of the GCE technique in the context of combining individual predictors.

Let us suppose we are interested in forecasts of 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 the variable being predicted 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 to be estimated $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$, and $\boldsymbol{\epsilon}$ is a $(T \times 1)$ vector containing the random errors. Each unknown parameter β_i is assumed to be a discrete random variable with $M \geq 2$ possible realizations. We suppose that there is some information about those possible realizations based on the researcher's a priori beliefs

about the likely values of β_i. That information is included in a support vector $\boldsymbol{b}' = (b_1, \ldots, b_M)$ with corresponding probabilities $\boldsymbol{p}'_i = (p_{i1}, \ldots, p_{iM})$. Although each parameter could have different M values, it is assumed that the M values are the same for every parameter. Thus, vector $\boldsymbol{\beta}$ can be rewritten as:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} = \boldsymbol{BP} = \begin{bmatrix} \boldsymbol{b}' & 0 & \cdots & 0 \\ 0 & \boldsymbol{b}' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{b}' \end{bmatrix} \begin{bmatrix} \boldsymbol{p}_1 \\ \boldsymbol{p}_2 \\ \vdots \\ \boldsymbol{p}_K \end{bmatrix}, \tag{5}$$

where $\boldsymbol{B}$ and $\boldsymbol{P}$ are matrixes with dimensions $(K \times KM)$ and $(KM \times 1)$ respectively. The following expression gives each parameter β_i as:

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

A similar approach is followed for $\boldsymbol{\epsilon}$. It is highlighted that, although GCE does not require rigid assumptions about the probability distribution function of the random error, as with other traditional estimation methods, some assumptions are still necessary to be made. It is assumed that $\boldsymbol{\epsilon}$ has a mean $E[\boldsymbol{\epsilon}] = 0$ and a finite covariance matrix. Moreover, each element ϵ_t is considered to be a discrete random variable with $J \geq 2$ possible values contained in the vector $\boldsymbol{v}' = \{v_1, \ldots, v_J\}$. Although each ϵ_t could have different J values, it is assumed as common for all of them ϵ_t $(t = 1, \ldots, T)$. We also assume that the random errors are symmetric around zero $(-v_1 = v_J)$. The upper and lower limits (v_1 and v_J, respectively) are fixed by applying the three-sigma rule (see [37–39]). Thus, vector $\boldsymbol{\epsilon}$ can be defined as:

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

and each element ϵ_t has the value equals:

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

Therefore, model (7) can be transformed into:

$$\boldsymbol{y} = \boldsymbol{XBP} + \boldsymbol{VW} \tag{9}$$

In this context, we need to estimate the elements of matrix $\boldsymbol{P}$, but also the elements of matrix $\boldsymbol{W}$ (denoted by $\widetilde{w}_{tj}$). The problem of the estimation of the vector of unknown parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ for the general linear model is transformed into the estimation of $K + T$ probability distributions. Based on this idea, [30] proposed an estimator that simultaneously allows for the estimation of parameters and the selection of variables in linear regression models. In order to have a basis for extraneous variable identification and coefficient reduction, the estimator uses sample but also non-sample information, as it is related to the Bayesian method of moments (BMOM) (see [40,41]). In other words, this technique allows for classifying some the explanatory variables in the linear model as irrelevant by shrinking the coefficients. Recent empirical applications of this method can also be found in [42–44].

Focusing on the context of combination of predictions, the objective of the DWP estimator is to identify which individual forecaster should receive a weight significantly different from the equal weighting scheme (simple arithmetic mean) and simultaneously to forecast the target variable based on a combination of individual predictors. We begin by specifying a discrete support space $\boldsymbol{b}$ for

each β_i symmetric around the value $1/K$ and with large lower and upper limits, so that each β_i is contained in the chosen interval with high probability. The upper and lower bounds for $\boldsymbol{v}$ (v_1 and v_J, respectively) are fixed by applying the three-sigma rule. 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 a 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, where a very high probability $q_m \cong 1$ is associated with the value $1/K$ for b_m (i.e., $q_m \cong 0$ for the remaining values). Thus, data-based prior is specified so flexibly that for each β_i coordinate either a spike prior at the $b_m = 1/K$, a uniform prior over support space $\boldsymbol{b}$, or any convex combination of the two, can result. The weight (a weighted formulation in an entropy optimization problem has been also proposed by [45] who proposed a weighted generalized maximum entropy (W-GME) estimator where different weights are assigned to the two entropies (for coefficient distributions and disturbance distributions) in the objective problem. Moreover, under a linear regression model estimation, [46] proposed a streaming generalized cross entropy (Stre-GCE) method to update the estimation of the parameters β_i by combining prior information and new data) given to the spike prior $\boldsymbol{q^s}$ for each parameter β_i is given by γ_i. For each γ_i, a discrete support space $\boldsymbol{b_i^\gamma}$ is specified with n possible values ($n = 1, \ldots, N$) and corresponding probability distribution $\boldsymbol{p_i^\gamma}$. Thus, γ_i is defined as $\gamma_i = \sum_{n=1}^{N} b_{in}^\gamma p_{in}^\gamma$, where $b_{i1}^\gamma = 0$ and $b_{iN}^\gamma = 1$ are, respectively, the lower and upper bounds defined as the support of these parameters.

If $\boldsymbol{q^u}$ and $\boldsymbol{q^s}$ denote the uniform and spike a priori distributions, respectively, we can achieve the objective proposed by minimizing the following constrained problem:

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

subject to:

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

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

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

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

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

The γ_i parameters and the β_i coefficients of the model in (10) are estimated simultaneously. Please note the symmetry between the terms γ and $1-\gamma$. Permuting the part of the objective function (10) to which they are connected would not change the final result in terms of the weighting scheme estimated.

To understand the logic of the DWP estimator, an explanation regarding the objective function (10) is useful, which is divided into four terms. The first one measures the divergence between the

posterior probabilities and the uniform priors for each β_i parameter, this being part of the divergence weighted by $(1-\gamma_i)$. The second element of (10) measures the divergence between the uniform priors for each β_i with the spike prior and it is weighted by γ_i. The third element in (10) relates to the Kullback divergence of the weighting parameters γ_i. It is highlighted that the a priori probability distribution fixed for each one of those parameters is always uniform ($q_i^{\gamma} = \frac{1}{N}\ \forall n = 1, \ldots, N$). The last term measures the Kullback divergence between the prior and the posterior probabilities for the random error of the model. The prior distribution of the errors is uniform (again $w_{tj}^0 = \frac{1}{J}\ \forall \mathrm{t} = 1, ..., \mathrm{T}$).

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

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

Under some mild assumptions (see [30], page 177), there is a guarantee that DWP estimates are consistent and asymptotically normal. Moreover, it is also ensured that the approximate variance of the DWP estimator is lower than the approximate variance of the GCE estimator, where the variance is lower than the approximate variance of an Maximum Likelihood- Least Squares estimator (see [30], page 179).

As it was highlighted, the DWP estimator allows simultaneously the estimation of parameters and the selection of predictors in linear regression models. The strategy to reach this objective has two steps. First, the estimates of the weighting parameters γ_i are obtained as:

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

which can be used as a tool for this purpose: As $\widetilde{\gamma}_i \to 0$, the prior gets closer to the uniform and the estimated parameters approach those of the GME estimator. This indicates that the parameter associated with this predictor can take values far from the center of the support vector (i.e., $1/K$). On the other hand, for large values of $\widetilde{\gamma}_i$, the part of the objective function with the 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 mean average will be characterized by large values of $\widetilde{\gamma}_i$ ([30] considers sufficiently large values when $\widetilde{\gamma}_{ih} > 0.49$), together with estimates of β_i close to $1/K$.

Moreover, it is possible to test if the estimate for β_i is significantly different from $1/K$ by constructing an χ^2 statistic. In other words, the statistic allows us to test if the estimated $\widetilde{p}_{im}$ is significantly different from the respective spike prior q_{im}^s. The Kullback–Leibler divergence measure between the estimated and the a priori probabilities related to the spike prior is:

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

The χ^2 divergence between both probabilities distributions is:

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

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

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

Consequently:

$$2MD_i\left(\widetilde{\boldsymbol{p}}_i \| \boldsymbol{q}_i^s\right) \rightarrow \chi^2_{M-1} \tag{21}$$

Thus, the measure $2MD_i\left(\widetilde{\boldsymbol{p}}_i \| \boldsymbol{q}_i^s\right)$ allows us to test the null hypothesis $H_0: \beta_i = 1/K$. If H_0 is not rejected, we conclude that a predictor x_i should be weighted as a simple arithmetic. (We would like to point out that, when computing, log(0) presents problems in the computation. In order to overcome this, 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.) In such a case, the vector of weights of the combined forecast estimated by using the DWP estimator is not different from the simple average. It means that the sample does not contain information providing strong empirical evidence to weigh differently than equal.

4. A Numerical Simulation Study

In this section of the paper, we compare the performance of the proposed DWP estimator with other methods used to combine individual forecasts by carrying out a numerical simulation study. Forecast combinations have been successfully applied in several areas of forecasting, such as economy (gross valued added, inflation, or stock returns), meteorology (wind speed, rainfall, see e.g., [47] in *Entropy* journal), or energy fields (wind power), among others. We focus our empirical exercise in the economic area; in fact, we take variable $\boldsymbol{y}$ as the gross value added being forecasted. (It is supposed that $\boldsymbol{y}$ is measured without error. In a situation in which $\boldsymbol{y}$ was measured with error, [48] proposed a method to extend the simple linear measurement error model through the inclusion of a composite indicator by using the GME estimator.)

The starting point of the numerical simulation is the unknown series y_t ($t = 1, \ldots, T$) that contains the target variable 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:

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

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 for the target series that is contained in predictor $\boldsymbol{x_i}$, i.e., the higher the value of σ_i, the less informative $\boldsymbol{x_i}$ is about $\boldsymbol{y}$.

Given that in our numerical experiment we would like to replicate situations normally observed in the context of forecasting economic series, instead of numerically generating the values of our target variable $\boldsymbol{y}$, we opted for taking actual values of an economic indicator. More specifically, we have taken the annual Gross Value Added rate of change in the region of Catalonia (Spain) from 1980 to 2013. We have extracted this information (at constant prices of 2008) from the BDmores database. (This database is generated by the Spanish Ministry of Economy, Industry and Competitiveness. More details can be found in: http://www.sepg.pap.minhap.gob.es/sitios/sepg/en-GB/Presupuestos/Documentacion/paginas/base0sdatosestudiosregionales.aspx).

Concerning the configuration of matrix $\boldsymbol{X}$, we consider 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, with 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 into two different subsets that can be classified as "good" or "bad" predictors. The logic of this idea is that the information that the predictors provide for forecasting variable $\boldsymbol{y}$ can vary among them, with a "good" predictor preferable to a "bad" one, but with the caveat that the comparatively "bad" forecaster may still contain 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:

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

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

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:

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

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

where s is the standard deviation in the sample 1980–2013 of the target variable $\boldsymbol{y}$. Equation (25) and Equation (26) indicate that the variance of the forecasters classified as "good" presents a variance four times lower than for those 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:

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

with the target of the different methods 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 $\beta_i = 1/K,\ \forall i$, which is normally the strategy taken as a valid reference in the literature on combination of forecasters. In fact, it is sometimes considered as the best way of combining information of individual predictors as some studies have pointed out (for example, [2,10,27–29]). Additionally, a restricted least squares weight scheme (see [19], for the original unrestricted Leas Squares approach; or [5] 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 to include the proposals made in recent forecasting literature, where forecasts based on Bayesian model averaging (BMA) has received considerable attention (see [49,50]). In this approach, the weights are determined based 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

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

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 be estimated 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 that each unknown parameter β_i has $M = 3$ possible realizations with

values $\boldsymbol{b}' = (1/K - 1, 1/K, 1/K + 1)$; in other words, the bounds with the minimum and maximum possible values for the weights are set as the center $1/K \pm 1$.

For the weighting parameters, we have considered a support vector with two possible realizations $N = 2$ and values $\boldsymbol{b}' = (0, 1)$. Finally, the supports of the random error terms have been specified by guarantying symmetry around zero and by using the three-sigma rule $(-3s,\ 0,\ 3s)$, with s being the sample standard deviation of the dependent variable.

Tables 1 and 2 summarize the results of comparing the actual target values of our variable of interest (y_t) with the combined individual forecasts ($\hat{y}_t$) obtained according to the different methods, namely; the simple mean (mean), Least Squares (LS), Bayesian Information Criterion (BIC) and the proposed Data Weighted Prior (DWP), and following two different deviation measures: (i) The mean squared forecast errors (MSFE); and (ii), the mean absolute percentage forecast error (MAPFE), respectively, defined by the two following expressions:

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

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

Table 1. Mean squared forecasting error (MSFE); 1000 trials.

Mean Squared Forecasting Error (MSFE)					
		Method			
K	***G***	**mean**	**LS**	**BIC**	**DWP**
6	5 good	0.0160	0.0136	0.0298	0.0156
	3 good	0.0269	0.0180	0.0379	0.0261
12	10 good	0.0077	0.0099	0.0256	0.0076
	6 good	0.0128	0.0141	0.0288	0.0125
24	20 good	0.0040	0.0147	0.0191	0.0039
	12 good	0.0064	0.0205	0.0243	0.0062

Table 2. Mean absolute percentage forecasting error (MAPFE); 1000 trials.

Mean Absolute Percentage Forecasting Error (MAPFE)					
		Method			
K	***G***	**mean**	**LS**	**BIC**	**DWP**
6	5 good	2.0312	1.8454	2.7303	2.0023
	3 good	2.6217	2.1553	3.1300	2.5799
12	10 good	1.4251	1.5797	2.5231	1.4079
	6 good	1.8182	1.8762	2.7280	1.7976
24	20 good	1.0132	1.8501	2.1976	0.99836
	12 good	1.2749	2.2305	2.5106	1.2556

The mean values of these deviation measures are computed from 1000 trials and for a forecast horizon of four periods ahead ($f = 1, \ldots, 4$), which means that the last four 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.

Error figures in Tables 1 and 2 show how the simple mean outperforms the combining methods based on some regression analysis (LS or BIC) in situations where the number of potential forecasters is large relative to the available sample size. When the predictors considered are 12 or 24, the combination

based on LS and BIC presents problems derived from an ill-conditioned dataset (the number of parameters is large relative to the 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 sample contains information providing strong empirical evidence to weigh differently than equal. On the contrary, when the number of predictors is low, an 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. One important aspect to consider, however, is that the performance of the proposed combined forecast methods has only been evaluated under the criterion of accuracy (measured through some forecast error-based indicators). However, other criteria could be considered (such as forecast error variance or asymmetry) leading to a different relative performance of the combining methods [9].

5. Conclusions

One of the most widespread strategies for combining individual forecasts is to take a simple average of the forecasts. Empirically, many studies have shown that the mean outperforms complex combining strategies. Theoretically, the use of the simple arithmetic mean could be justified when all the forecasters have shown the same forecasting ability or when the available information about their ability seems to be not enough to calibrate the forecasters differently. This paper proposes the use of an entropy-based technique estimator to obtain an affine transformation of the equal weighted forecast combination by using the small available information, a data-weighted prior (DWP) estimator.

We tested the validity of the proposed model by a simulation exercise and compared its ex-ante forecasting performance with other combining methods. The benchmarks for comparing the competing method were the arithmetic mean of the forecasters, a restricted least squares, and weight scheme forecasts based on Bayesian model averaging (where the weights are determined on the basis of 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 obtained results of the simulation 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 sample contains information providing strong enough empirical evidence to weigh differently than equal. The most relevant advantage of this estimator is that, even in situations characterized by a large number of forecasters, the DWP estimator generates a better set of recovered forecasters´ weights than the arithmetic mean which is capable to identify groups of forecasters into groups of "good" and "bad" forecasts. Additionally, the empirical application could be extended by comparing the forecasting performance of the proposed method with other combining methods based on an information-theoretic approach [6].

Author Contributions: Conceptualization, E.F.-V., B.M. and G.J.D.H.; Methodology, E.F.-V.; Validation, E.F.-V. and B.M.; Formal Analysis, E.F.-V.; Resources, B.M.; Writing-Original Draft Preparation, E.F.-V., B.M. and G.J.D.H.; Writing-Review & Editing, E.F.-V., B.M. and G.J.D.H.; Funding Acquisition, E.F.-V. and B.M.

Funding: This research was partially funded by the research project "Integrative mechanisms for addressing spatial justice and territorial inequalities in "Europe (IMAJINE)" in the EU Research Framework Programme H2020.

Acknowledgments: The authors acknowledge the support of the guest editors of this special issues and the comments received by two anonymous reviewers.

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

References

1. Holden, K.; Peel, D.A. Combining Economic Forecasts. *J. Oper. Res. Soc.* **1988**, *39*, 1005–1010. [CrossRef]
2. Stock, J.H.; Watson, M.W. Combining Forecasts of Output Growth in a Seven-Country Data Set. *J. Forecast.* **2004**, *23*, 405–430. [CrossRef]

3. Marcellino, M. Forecast Pooling for European Macroeconomic Variables. *Oxf. Bull. Econ. Stat.* **2004**, *66*, 91–112. [CrossRef]
4. Greer, M.R. Combination forecasting for directional accuracy: An application to survey interest rate forecasts. *J. Appl. Stat.* **2005**, *32*, 607–615. [CrossRef]
5. Timmermann, A. Forecast Combinations. In *Handbook of Economic Forecasting*; Elliott, G., Granger, C.W.J., Timmermann, A., Eds.; North-Holland: Amsterdam, The Netherlands, 2006; Volume 1, pp. 135–196.
6. Moreno, B.; López, A.J. Combining economic forecasts by using a Maximum Entropy Econometric. *J. Forecast.* **2013**, *32*, 124–136. [CrossRef]
7. Fernandez-Vazquez, E.; Moreno, B. Entropy econometrics for combining regional economic forecasts: A data-weighted prior estimator. *J. Geogr. Syst.* **2017**, *19*, 349–370. [CrossRef]
8. Bates, J.M.; Granger, C.W.J. The Combination of Forecasts. *Oper. Res. Q.* **1969**, *20*, 451–468. [CrossRef]
9. De Menezes, L.M.; Bunn, D.W.; Taylor, J.W. Review of Guidelines for the Use of Combined Forecasts. *Eur. J. Oper. Res.* **2000**, *120*, 190–204. [CrossRef]
10. Genre, V.; Kenny, G.; Meyler, A.; Timmermann, A. Combining expert forecasts: Can anything beat the simple average? *Int. J. Forecast.* **2013**, *29*, 108–121. [CrossRef]
11. Newbold, P.; Granger, C.W.J. Experience with Forecasting Univariate Time Series and the Combination of Forecasts. *J. Royal Stat. Soc. Ser. A* **1974**, *137*, 131–165. [CrossRef]
12. Bunn, D.W. A Bayesian approach to the linear combination of forecasts. *Oper. Res. Q.* **1975**, *26*, 325–329. [CrossRef]
13. Bordley, R.F. The combination of forecast: A Bayesian approach. *J. Oper. Res. Soc.* **1982**, *33*, 171–174. [CrossRef]
14. Winkler, R.L. Combining probability distributions from dependent information sources. *Manag. Sci.* **1981**, *27*, 479–488. [CrossRef]
15. Winkler, R.L.; Makridakis, S. The combination of forecasts. *J. Royal Stat. Soc. Ser. A* **1983**, *146*, 150–157. [CrossRef]
16. Agnew, C.E. Bayesian consensus forecast of macroeconomic variables. *J. Forecast.* **1985**, *4*, 363–376. [CrossRef]
17. Anandalingam, G.; Chen, L. Linear combination of forecasts: A general Bayesian model. *J. Forecast.* **1983**, *8*, 199–214. [CrossRef]
18. Clemen, R.T.; Winkler, R.L. Aggregating point estimates: A flexible modelling approach. *Manag. Sci.* **1999**, *39*, 501–515. [CrossRef]
19. Granger, C.W.J.; Ramanathan, C. Improved Methods of Combining Forecasts. *J. Forecast.* **1984**, *3*, 197–204. [CrossRef]
20. Chan, Y.; Stock, J.; Watson, M.A. Dynamic Factor Model Framework for Forecast Combination. *Span. Eco. Rev.* **1999**, *1*, 91–121. [CrossRef]
21. Stock, J.H.; Watson, M.W. Forecasting Using Principal Components from a Large Number of Predictors. *J. Am. Stat. Association* **2002**, *97*, 147–162. [CrossRef]
22. Fang, Y. Forecasting combination and encompassing tests. *Int. J. Forecast.* **2003**, *19*, 87–94. [CrossRef]
23. Aiolfi, M.; Timmerman, A. Persistence in forecasting performance and conditional combination strategies. *J. Eco.* **2006**, *135*, 31–53. [CrossRef]
24. Guerard, J.B.; Clemen, R.T. Collinearity and the use of latent root regression for combining GNP forecasts. *J. Forecast.* **1989**, *8*, 231–238. [CrossRef]
25. De Mol, C.; Giannone, D.; Reichlin, L. Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *J. Eco.* **2008**, *146*, 318–328. [CrossRef]
26. Conflitti, C.; De Mol, C.; Giannone, D. Optimal Combination of Survey Forecasts. 2012. Available online: https://ideas.repec.org/p/eca/wpaper/2013-124527.html (accessed on 6 February 2019).
27. Makridakis, S.A.; Andersen, R.; Carbone, R.; Fildes, M.; Hibon, R.; Lewandowski, J.; Newton, E.; Parsen, E.; Winkler, R. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *J. Forecast.* **1982**, *1*, 111–153. [CrossRef]
28. Makridakis, S.; Winkler, R.L. Averages of Forecasts: Some empirical results. *Manag. Sci.* **1983**, *29*, 987–996. [CrossRef]
29. Smith, J.; Wallis, K.F. A Simple Explanation of the Forecast Combination Puzzle. *Oxf. Bull. Eco. Stat.* **2009**, *71*, 331–355. [CrossRef]
30. Golan, A. A Simultaneous Estimation and Variable Selection Rule. *J. Eco.* **2001**, *10*, 165–193. [CrossRef]

31. Diebold, F.X.; Pauly, P. Structural change and the combination of forecast. *J. Forecast.* **1987**, *6*, 21–40. [CrossRef]
32. Deutsch, M.; Granger, C.W.; Teräsvirta, T. The combination of forecasts using changing weights. *Int. J. Forecast.* **1994**, *10*, 47–57. [CrossRef]
33. Coulson, N.E.; Robins, R.P. Forecast Combination in a Dynamic Setting. *J. Forecast.* **1993**, *12*, 63–67. [CrossRef]
34. Hallman, J.; Kamstra, M. Combining algorithms based on robust estimation techniques and co-integration restrictions. *J. Forecast.* **1989**, *8*, 189–198. [CrossRef]
35. Miller, C.M.; Clemen, R.T.; Winkler, R.L. The effect of nonstationarity on combined forecasts. *Int. Forecast.* **1992**, *7*, 515–529. [CrossRef]
36. Capistrán, C.; Timmermann, A. Forecast combination with entry and exit of expert. *J. Bus. Eco. Stat.* **2009**, *27*, 428–440. [CrossRef]
37. Golan, A.; Judge, G.; Miller, D. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*; John Wiley & Sons Ltd.: London, UK, 1996.
38. Kapur, J.N.; Kesavan, H.K. *Entropy Optimization Principles with Applications*; Academic Press: New York, NY, USA, 1992.
39. Pukelsheim, F. The three sigma rule. *Am. Stat.* **1994**, *48*, 88–91.
40. Zellner, A. Bayesian Method of Moments/Instrumental Variable (bmom/iv) Analysis of Mean and Regression Models. In *Modeling and Prediction: Honoring Seymour Geisser*; Lee, J.C., Johnson, W.C., Zellner, A., Eds.; Springer: Berlin, Germany, 1996; pp. 61–75.
41. Zellner, A. The Bayesian Method of Moments (BMOM): Theory and Applications. In *Advances in Econometrics*; Fomby, T., Hill, R.C., Eds.; Emerald Group Publishing Limited: Bingley, UK, 1997; Volume 12, pp. 85–106.
42. Bernardini-Papalia, R. A Composite Generalized Cross Entropy formulation in small samples estimation. *Eco. Rev.* **2008**, *27*, 596–609. [CrossRef]
43. Fernandez-Vazquez, E. Recovering matrices of economic flows from incomplete data and a composite Prior. *Entropy* **2012**, *12*, 516–527. [CrossRef]
44. Fernandez-Vazquez, E.; Rubiera-Morollon, F. Estimating Regional Variations of R&D Effects on Productivity Growth by Entropy Econometrics. *Spat. Eco. Anal.* **2013**, *8*, 54–70.
45. Wu, X. A weighted generalized maximum entropy estimator with a data-driven weight. *Entropy* **2009**, *11*, 917–930. [CrossRef]
46. Angelelli, M.; Ciavolino, E. Streaming Generalized Cross Entropy. *arXiv* **2018**, arXiv:1811.09710.
47. Men, B.; Long, R.; Li, Y.; Liu, H.; Tian, W.; Wu, Z. Combined Forecasting of Rainfall Based on Fuzzy Clustering and Cross Entropy. *Entropy* **2017**, *19*, 694. [CrossRef]
48. Carpita, M.; Ciavolino, E. A Generalized Maximum Entropy Estimator to Simple Linear Measurement Error Model with a Composite Indicator. *Adv. Data Anal. Classif.* **2017**, *11*, 139–158. [CrossRef]
49. Buckland, S.T.; Burnham, K.P.; Augustin, N.H. Model selection: An integral part of inference. *Biometrics* **1997**, *53*, 603–618. [CrossRef]
50. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: New York, NY, USA, 2002.