BANK OF CANADA
BANQUE DU CANADA

Staff Discussion Paper/Document d'analyse du personnel—2022-19

Last updated: September 20, 2022

Calculating Effective Degrees of Freedom for Forecast Combinations and Ensemble Models

by James Younker

Corporate Services Department
Bank of Canada
jyounker@bankofcanada.ca

DOI: https://doi.org/10.34989/sdp-2022-19 | **ISSN** 1914-0568

Acknowledgements

I would like to thank James Chapman, Fuchun Li, Matthew Strathearn and Marcel Voia for the assistance they provided on this paper.

Abstract

Forecast combinations, also known as ensemble models, routinely require practitioners to select a model from a massive number of potential candidates. Ten explanatory variables can be grouped into 2^{1078} forecast combinations, and the number of possibilities increases further to $2^{1078+2^{1078}}$ if we allow for forecast combinations of forecast combinations. This paper derives a calculation for the effective degrees of freedom of a forecast combination under a set of general conditions for linear models. It also supports this calculation with simulations. The result allows users to perform several other computations, including the F-test and various information criteria. These computations are particularly useful when there are too many candidate models to evaluate out of sample. Furthermore, computing effective degrees of freedom shows that the complexity cost of a forecast combination is driven by the parameters in the weighting scheme and the weighted average of parameters in the auxiliary models as opposed to the number of auxiliary models. This identification of complexity cost contributions can help practitioners make informed choices about forecast combination design.

Topics: Econometric and statistical methods
JEL codes: *C, C01, C02, C1, C13, C5, C50, C51, C52, C53*

Résumé

Les combinaisons de prévisions, aussi appelées modèles d'ensemble, obligent régulièrement les praticiens à sélectionner un modèle parmi un grand nombre de modèles potentiels. Dix variables explicatives peuvent être groupées en 2^{1078} combinaisons de prévisions, et le nombre de possibilités atteint même $2^{1078+2^{1078}}$ si l'on tient compte des combinaisons de prévisions des combinaisons de prévisions. Dans cette étude, l'auteur s'attache à calculer les degrés de liberté effectifs d'une combinaison de prévisions en fonction d'un ensemble de conditions générales applicables aux modèles linéaires. Il conforte aussi son calcul par des simulations. Le résultat permet aux utilisateurs d'effectuer plusieurs autres calculs, dont le test de Fisher (test F) et divers critères d'information. Ces calculs sont particulièrement utiles lorsqu'il y a trop de modèles possibles à évaluer hors échantillon. De plus, le calcul des degrés de liberté effectifs montre que le coût de complexité d'une combinaison de prévisions dépend des paramètres du système de pondération et de la moyenne pondérée des paramètres dans les modèles auxiliaires, plutôt que du nombre de modèles auxiliaires. Cette détermination de la contribution du coût de complexité peut aider les praticiens à faire des choix éclairés pour la formation de combinaisons de prévisions.

Sujets : Méthodes économétriques et statistiques
Codes JEL : C, C01, C02, C1, C13, C5, C50, C51, C52, C53

Introduction: Effective degrees of freedom as a tool in model evaluation

Several forecasting methods use an ensemble of multiple auxiliary forecasts combined through a weighting scheme to create a forecast combination. Forecast combination methodologies have a rich history. Some notable papers include Bates and Granger (1969), Breiman (1996), Leblanc and Tibshirani (1996), Stock and Watson (2004) and Hansen (2007). Originally, these methods were used to combine third-party forecasts. Recently, however, forecast combinations have also been considered a shrinkage estimator to mitigate parameter uncertainty (Hansen 2007) and a means of diversifying a forecast to achieve robustness to regime changes (Elliott and Timmermann 2005).

The problem with forecast combinations is that model selection can be challenging when the number of candidate models is massive.[1] Several papers have tried to address this problem by developing an optimal forecast combination design to bypass the search requirement. This has led to a fascinating literature on optimal forecast combinations, featuring papers such as Elliott and Timmermann (2005), Hansen (2007), Hsiao and Wan (2014), Claeskens et al. (2016), Samuels and Sekkel (2017) and Diebold and Shin (2019). However, the model selection problem has still not been resolved.

This paper attempts to mitigate the model selection problem by estimating the degrees of freedom of a forecast combination. This is done in a linear model setting under broad conditions that cover a range of practical applications. The methodology follows the literature on effective degrees of freedom: Efron (2004) discusses the approach in general; Efron et al. (2004) apply it to the LARs model; Zou, Hastie and Tibshirani (2007) and Tibshirani and Taylor (2012) apply it to the LASSO model; Kato (2009) applies it to shrinkage estimation; and Mukherjee et al. (2015) apply it to reduced rank estimators.

Calculating effective degrees of freedom allows for other computations, including the F-test and several information criteria.[2] These measures are useful for selecting models from groups of candidate models that are too large to evaluate exhaustively out of sample. Additionally, this identifies the complexity contributions of the weighting scheme and individual auxiliary models. This allows practitioners to make better decisions about the design of forecast combinations.

The first section of the paper creates a single model representation of a forecast combination, subject to the conditions that the auxiliary models are single equations that are linear in specification and that the auxiliary model weights are non-time varying, linear in specification and estimated on a balanced sample as a function of auxiliary model fitted values. The second section follows the literature and uses Stein's unbiased risk estimate (SURE) to calculate a forecast combination's effective degrees of freedom (Proposition I). We then show this to be a generalization of the result Hansen (2007) derived for fixed-weight forecast combinations. The third section simplifies the vector derivative and identifies the effective

[1] Appendix 2 itemizes the number of forecast combinations that can be produced from a given set of explanatory variables. Equations 21 and 22 in Appendix 2 establish that 10 explanatory variables can be grouped into 2^{1078} forecast combinations—or $2^{1078+2^{1078}}$ when we include forecast combinations of forecast combinations: $2^{\sum_{i=1}^{10}\binom{10}{i}} = 2^{1078}$, and $2^{\sum_{i=1}^{10}\binom{10}{i}+2^{\sum_{i=1}^{10}\binom{10}{i}}} = 2^{1078+2^{1078}}$.

[2] In particular, effective degrees of freedom allow the computation of the Akaike information criterion, the Bayesian information criterion, the generalized cross-validation statistic and Mallow's *Cp* statistic.

degrees of freedom contribution from each of a forecast combination's components (Proposition II). The Proposition II result is then simplified further for practical applications, with the final term being approximated with its limit, resulting in equation (19). The fourth section evaluates the effectiveness of equation (19) using a simulation. Appendix 1 provides detailed derivations of propositions I and II. Appendix 2 provides a formula for the number of ways variables can be combined into forecast combinations. Appendix 3 details the simulation procedure.

1. A single model representation of a forecast combination

As a first step to determining effective degrees of freedom, we represent a forecast combination as a single model by stacking the auxiliary models and expressing the weighting scheme as a matrix. We start with the M underlying auxiliary models indexed by i, where $i = 1, \ldots M$. Each auxiliary model has an n by p_{m_i} matrix of observations X_{m_i} with p_{m_i} explanatory variables and a p_{m_i} by 1 column vector of parameters $\hat{\beta}_{m_i}$. In accordance with the structure of a forecast combination, each auxiliary model is required to have the same n by 1 dependent variable Y. To achieve the single model representation, each auxiliary model must be linear in specification and must be estimated as one equation and not as a system. Across the set of M models, the explanatory variables may repeat, and estimation samples may differ.

The weighting scheme is a diagonal matrix q that weights the contribution of each auxiliary model m by a value q_{m_i}. The weights in q must satisfy the conditions of being non-time varying, linear in specification and estimated on a balanced sample as a function of Y and the n by 1 fitted values from each of the auxiliary models $\hat{Y}_{m_i}$.

The forecast combination fitted values, explanatory variables and parameters are notated as $\hat{Y}$, X and $\hat{\beta}$, respectively, with dimensions n by 1, n by p and p by 1. Throughout the paper, the following indexing convention will be used: variables within an auxiliary model will be indexed with k, where $k = 1, \ldots p_m$; auxiliary models themselves will be indexed with i, where $i = 1, \ldots M$; the forecast combination will include all the variables of the auxiliary models and will be indexed with j, where $j = 1, \ldots p$, such that $p = \sum_{i=1}^{M} p_{m_i}$; and observations will be indexed with l, where $l = 1, \ldots n$.

$$X = [X_{m_1} \quad \cdots \quad X_{m_M}], \text{with dimensions } n \text{ by } p, \text{where } p = \sum_{i=1}^{M} p_{m_i}. \tag{1}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_{m_1} \\ \vdots \\ \hat{\beta}_{m_M} \end{bmatrix}, \text{with dimensions } p \text{ by } 1. \tag{2}$$

$$q = \begin{bmatrix} q_{j=1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & q_{j=p} \end{bmatrix} \text{is a diagonal matrix with dimensions } p \text{ by } p, \text{such that diagonal} \tag{3}$$

values are constant within M blocks that correspond to the blocks in X and $\hat{\beta}$. In particular,

$[q_j]_{j=1}^{j=p_{m_1}} = q_{m_1}, [q_j]_{j=p_{m_1}+1}^{j=p_{m_2}} = q_{m_2}, \ldots, [q_j]_{j=p_{m_M}-1}^{j=p_{m_M}} = q_{m_M}$.

$$\hat{Y} = Xq\hat{\beta}. \tag{4}$$

2. Stein's unbiased risk estimate to calculate effective degrees of freedom and generalize Hanson's result

Estimating degrees of freedom (DF) with effective degrees of freedom (EDF) follows the common practice in the literature of utilizing Stein's unbiased risk estimate (Stein 1981). The resulting EDF is considered interchangeable with DF for practical purposes of computing F-tests and information criteria. Following Efron (2004), Efron et al. (2004), Zou, Hastie and Tibshirani (2007), Kato (2009), Tibshirani and Taylor (2012) and Mukherjee et al. (2015), the starting point is the well-known equations (5) and (6). Equation (5) is a representation of Stein's unbiased risk estimate, where σ^2 is the fitted residual's variance. The expectation relationship in equation (5) then motivates estimating DF with EDF, where EDF is defined by equation (6).

$$DF = \sum_1^n \frac{Cov(\hat{Y},Y)}{\sigma^2} = E\left(Trace\left(\frac{\partial \hat{Y}}{\partial Y}\right)\right). \tag{5}$$

$$EDF = Trace\left(\frac{\partial \hat{Y}}{\partial Y}\right). \tag{6}$$

Branching off from the literature, we substitute the single model representation of a forecast combination in equation (4) into the EDF measure of equation (6), which results in equation (7):

$$EDF = Trace\left(\frac{\partial}{\partial Y} Xq\hat{\beta}\right). \tag{7}$$

Throughout the paper, vec is the vectorization operator, $\otimes$ is the Kronecker product, and I is the identity matrix. Matrix derivatives of matrices will utilize the conventions presented in Magnus (2010), which are provided in equations (8) and (9).[3] The derivative in equation (7) can be calculated using (9), which results in (10), where I_1 is the 1 by 1 identity matrix. Equation (10) can then be simplified into (11). From the properties of block matrix multiplication[4] and the properties of the trace operator,[5] (11) can be rewritten as (12). Further details are in Appendix 1.

[3] The conventions of Magnus are provided below in equations (8) and (9), where F is an r by s matrix, G is a t by u matrix, and Z is a v by w matrix.

$$Derivative\ of\ F(Z)\ with\ respect\ to\ Z = \frac{\partial\ vec\ F(Z)}{\partial (vec\ Z)'}, where\ matrix\ F(Z)\ is\ a\ function\ of\ matrix\ Z \tag{8}$$

$$Product\ rule\text{: } Derivative\ of\ F(Z)G(Z)\ with\ respect\ to\ Z = (G' \otimes I_r)DF(Z) + (I_u \otimes F)DG(Z),$$

$$where\ matrixes\ F(Z)\ and\ G(Z)\ are\ functions\ of\ matrix\ Z \tag{9}$$

[4] See chapter 2 in Harville (2008).

[5] See chapter 5 in Harville (2008).

$$EDF = Trace\left(X\left(\left(\hat{\beta}^T \otimes I_p\right)\frac{\partial q}{\partial Y} + (I_1 \otimes q)\frac{\partial \hat{\beta}}{\partial Y}\right)\right). \quad (10)$$

$$EDF = Trace\left(Xq\frac{\partial \hat{\beta}}{\partial Y}\right) + Trace\left(X\left(\hat{\beta}^T \otimes I_p\right)\frac{\partial q}{\partial Y}\right). \quad (11)$$

$$EDF = \sum_{i=1}^{M} q_{m_i} Trace\left(\frac{\partial \hat{Y}_{m_i}}{\partial Y}\right) + Trace\left(X\left(\hat{\beta}^T \otimes I_p\right)\frac{\partial q}{\partial Y}\right). \quad (12)$$

Equation (6) allows for the substitution $EDF_{m_i} = Trace\left(\frac{\partial \hat{Y}_{m_i}}{\partial Y}\right)$, where EDF_{m_i} is the EDF of auxiliary model m_i. With the substitution, (12) becomes (13).

Proposition I

Given auxiliary models that are single equations and linear in specification, and a weighting scheme that is non-time varying, linear in specification and estimated on a balanced sample as a function of auxiliary model fitted values, the effective degrees of freedom are defined by equation (13).

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace\left(X\left(\hat{\beta}^T \otimes I_p\right)\frac{\partial q}{\partial Y}\right). \quad (13)$$

Proposition I is a solution for the effective degrees of freedom of a forecast combination. Additionally, Proposition I is a generalization of Hansen's Lemma 1 (Hansen 2007), which is the first term in equation (13). Hansen's Lemma 1 $EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i}$ states that the effective degrees of freedom of a forecast combination with fixed weights is the weighted average of the EDFs of the individual auxiliary models. Proposition I generalizes this by adding a second term, $Trace\left(X\left(\hat{\beta}^T \otimes I_p\right)\frac{\partial q}{\partial Y}\right)$, to capture the effective degrees of freedom associated with estimating the weighting scheme. In the case of a forecast combination weighting scheme with fixed weights not estimated as a function of *Y*, $\frac{\partial q}{\partial Y} = 0$ and equation (13) reduces to Hansen's Lemma 1.

3. Attributing effective degrees of freedom to the components of a forecast combination

Proposition I provides a solution for effective degrees of freedom of a forecast combination; however, the term $Trace\left(X\left(\hat{\beta}^T \otimes I_p\right)\frac{\partial q}{\partial Y}\right)$ is awkward for most purposes. This section simplifies the term into a contribution from estimating the weighting scheme in a linear setting and an interaction term arising from estimating the weighting scheme on the fitted values of the auxiliary models that are a function of *Y*.

Using the properties of block matrices,[6] equation (13) can be rewritten as equation (14), where X_{nj} is the $n\,j$ entry of matrix X, and $\hat{\beta}_q$ is a 1 by m vector $\begin{bmatrix} q_{m_1} \\ \vdots \\ q_{m_M} \end{bmatrix}$. Further details are in Appendix 1. Also, through the properties of block matrices,[7] equation (14) can be rewritten as equation (15), where $\hat{Y}_{1i}$ is a fitted value of the i^{th} auxiliary model.

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace\left(\begin{bmatrix} \sum_{j=1}^{P} X_{1j}\hat{\beta}_j \frac{\partial q_j}{\partial Y_1} & \cdots & \sum_{i=1}^{P} X_{1j}\hat{\beta}_j \frac{\partial q_j}{\partial Y_n} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{P} X_{nj}\hat{\beta}_j \frac{\partial q_j}{\partial Y_1} & \cdots & \sum_{i=1}^{P} X_{nj}\hat{\beta}_j \frac{\partial q_j}{\partial Y_n} \end{bmatrix}\right). \quad (14)$$

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace\left(\begin{bmatrix} \sum_{i=1}^{M} \hat{Y}_{1i} \frac{\partial q_{m_i}}{\partial Y_1} & \cdots & \sum_{i=1}^{M} \hat{Y}_{1i} \frac{\partial q_{m_i}}{\partial Y_n} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{M} \hat{Y}_{ni} \frac{\partial q_{m_i}}{\partial Y_1} & \cdots & \sum_{i=1}^{M} \hat{Y}_{ni} \frac{\partial q_{m_i}}{\partial Y_n} \end{bmatrix}\right). \quad (15)$$

Equation (15) can then be rewritten as equation (16), where X_A is an n by m matrix of auxiliary model fitted values $\begin{bmatrix} \hat{Y}_{11} & \cdots & \hat{Y}_{1m} \\ \vdots & \ddots & \vdots \\ \hat{Y}_{n1} & \cdots & \hat{Y}_{nm} \end{bmatrix}$. Using the condition that $\hat{\beta}_q$ is estimated by a single equation linear in specification as a function of X_A results in the form $\hat{\beta}_q = F(X_A)Y$, which is non-linear with respect to Y because X_A is a function of Y:

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace\left(X_A \frac{\partial \hat{\beta}_q}{\partial Y}\right). \quad (16)$$

The derivative in equation (16) is taken using equation (9), resulting in equation (17). The effective degrees of freedom contribution from the weighting scheme is then represented as two terms: $Trace(X_A F(X_A))$ and $Trace\left(X_A\left((Y^T \otimes I_m)\frac{\partial F(X_A)}{\partial Y}\right)\right)$.

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace(X_A F(X_A)) + Trace\left(X_A\left((Y^T \otimes I_m)\frac{\partial F(X_A)}{\partial Y}\right)\right). \quad (17)$$

Using equation (6), $Trace(X_A F(X_A))$ equals EDF_w—the effective degrees of freedom of the weighting scheme if X_A were not dependent on Y.[8] The impact of X_A's dependency on Y is then captured in the final term, $Trace\left(X_A\left((Y^T \otimes I_m)\frac{\partial F(X_A)}{\partial Y}\right)\right)$.

[6] See chapter 2 in Harville (2008).

[7] See chapter 2 in Harville (2008).

[8] For models that can be written in the linear form $\hat{Y} = X\hat{\beta}$, where $\hat{\beta} = F(X)Y$, the effective degrees of freedom is the trace of the hat matrix (Hastie, Tibshirani and Friedman 2001, chapter 7). This is readily derived from equation 6 and the derivative of $\hat{Y}$: $\frac{\partial \hat{Y}}{\partial Y} = XF(X)$.

Equation (17) is simplified by substituting $Trace(X_A F(X_A))$ for EDF_w and rewriting the final term as $Trace\left(X_A\hat{\beta}_{\Delta q}\right)$, where $\hat{\beta}_{\Delta q}$ is an m by n matrix expressing the change in $\hat{\beta}_q$ as a result of X_A being a function of Y. See Appendix 1 for more details.

This results in Proposition II, where the effective degrees of freedom of a forecast combination is decomposed into the Hansen result $\sum_{i=1}^{M} q_{m_i} EDF_{m_i}$ for fixed weight combinations, plus the EDF of the weighting scheme in a deterministic setting EDF_w, plus an interaction term $Trace\left(X_A\hat{\beta}_{\Delta q}\right)$ accounting for the dependency between X_A and Y.

Proposition II

Given auxiliary models that are single equations that are linear in specification, and a weighting scheme that is non-time varying, linear in specification and estimated on a balanced sample as a function of auxiliary model fitted values, the effective degrees of freedom is defined by equation (18):

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + EDF_w + Trace\left(X_A\hat{\beta}_{\Delta q}\right). \quad (18)$$

Therefore, the effective degrees of freedom of a forecast combination is the weighted average of the effective degrees of freedom of the individual auxiliary models, plus the effective degrees of freedom of the weighting scheme ignoring the dependency between X_A and Y, plus an interaction term for X_A and Y.

Given the complexity of computing $Trace\left(X_A\hat{\beta}_{\Delta q}\right)$, it may be advantageous in many practical situations to approximate $Trace\left(X_A\hat{\beta}_{\Delta q}\right)$ with is limit $\lim_{\hat{Y}\prime S \to Y} Trace\left(X_A\hat{\beta}_{\Delta q}\right)$, which is shown in Appendix 1 to be zero for well-behaved cases. As a result, a forecast combination's EDF can be approximated as equation (19):

$$EDF \approx \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + EDF_w. \quad (19)$$

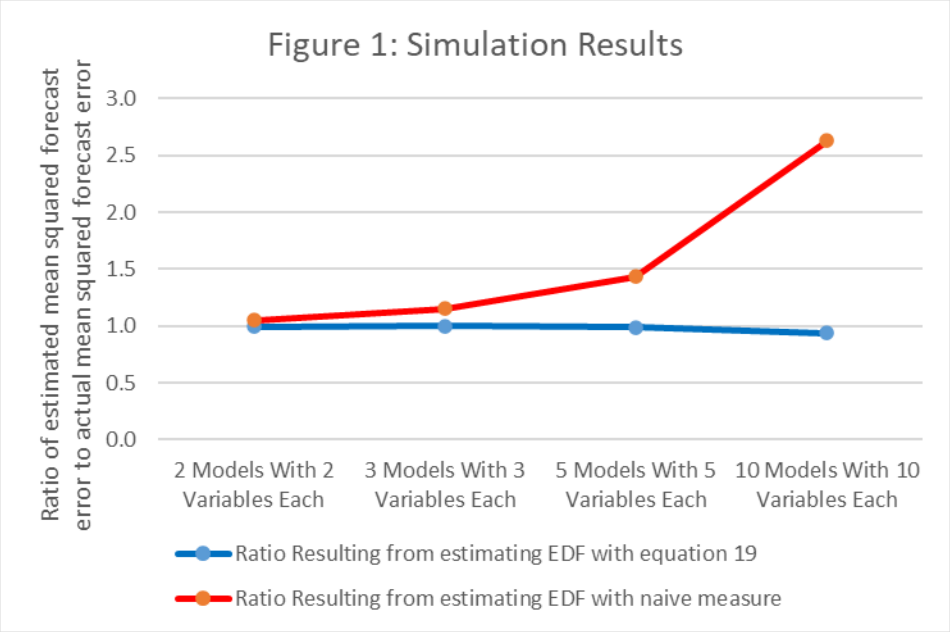
4. Simulation results

We use a Monte Carlo simulation to illustrate the performance of equation (19) as an approximation of the EDF under different model specifications. The simulation procedure is specified in Appendix 3.

The simulation computes the Mallow's *Cp* estimate of squared forecast error and compares it to actual out-of-sample squared forecast errors. Mallow's *Cp* is computed using both the EDF approximation of equation (19) and the naïve EDF measure of counting the number of estimated parameters. The simulation covers four forecast combination specifications: 2 auxiliary models of 2 variables each, 3 auxiliary models of 3 variables each, 5 auxiliary models of 5 variables each, and 10 auxiliary models of 10 variables each. In all cases 100 observations are used, the explanatory variables have a correlation of 0.8, the auxiliary

models and the weighting schemes are linear models estimated with ordinary least squares, the constraint[9] of $\sum_{i=1}^{M} q_{m_i} = 1$ is applied to the weighting scheme, and results are averaged over 100,000 repetitions.

Simulation results are shown in Figure 1. The EDF approximation from equation (19) results in a ratio close to 1, indicating that the Mallow's *Cp* resulting from equation (19) is a relatively accurate estimate of squared forecast error. The small difference from 1 may be the result of approximating $Trace\left(X_A\hat{\beta}_{\Delta q}\right)$ with its limit of zero in equation (19), or the number of observations being insufficiently large for Mallow's *Cp*. By contrast, the naive EDF measure deviates notably from 1, indicating relatively weaker performance. Overall, this simulation would appear to support the use of equation (19) to approximate a forecast combinations EDF as defined by propositions I and II.



Conclusion

An established literature is followed to compute the EDF of a forecast combination using Stein's unbiased risk estimate. This results in Proposition I, which is shown to be a generalization of Hansen's Lemma 1 (Hansen 2007). Proposition I simplifies to Proposition II, which expresses the EDF of a forecast combination as the weighted average of the EDF of the individual auxiliary models, plus the EDF of the weighting scheme plus an interaction term. Given the complexity of the interaction term, which limits to zero, practitioners may find it advantageous to use the approximate EDF from equation (19). Simulation results support the effectiveness of this approach. Equation (19) provides practitioners with a simple calculation for the EDF of a forecast combination—simply the weighted average of the EDF of the auxiliary models plus the EDF of the weighting scheme.

[9] This is a commonly used constraint found in many applied papers, and its use is advocated in chapter 8 of Hastie, Tibshirani and Friedman (2001).

A calculation for the EDF of a forecast combination has two main applications. First, as a result of their design, forecast combinations tend to produce massive numbers of candidate models, as shown in Appendix 2. An EDF allows candidate models to be evaluated with information criteria or F-tests without resorting to out-of-sample forecast evaluations, which are challenging to conduct on a massive scale. Second, the EDF allows practitioners to see what elements of model design drive complexity cost. In particular, equation (19) shows that the complexity cost of a forecast combination is driven by parameters in the weighting scheme and the weighted average of parameters in the auxiliary models, as opposed to the number of auxiliary models.

Appendix 1: Detailed derivations of propositions I and II

Proposition I begins with equation (6) from the literature:

$$EDF = Trace\left(\frac{\partial \hat{Y}}{\partial Y}\right). \tag{6}$$

Equation (4) is substituted into equation (6).

$$EDF = Trace\left(\frac{\partial}{\partial Y} Xq\hat{\beta}\right) \tag{7}$$

Equation (7) is differentiated using equation (9).

$$EDF = Trace\left(X\left(\left(\hat{\beta}^T \otimes I_p\right)\frac{\partial q}{\partial Y} + \left(I_1 \otimes q\right)\frac{\partial \hat{\beta}}{\partial Y}\right)\right). \tag{10}$$

$$EDF = Trace\left(Xq\frac{\partial \hat{\beta}}{\partial Y}\right) + Trace\left(X\left(\hat{\beta}^T \otimes I_p\right)\frac{\partial q}{\partial Y}\right). \tag{11}$$

$$EDF = Trace\left(A\frac{\partial \hat{\beta}}{\partial Y}\right) + Trace\left(X\left(\hat{\beta}^T \otimes I_p\right)\frac{\partial q}{\partial Y}\right), \tag{11.1}$$

where A is an n by p block matrix $[X_{m_1}q_{m_1}, \quad \cdots \quad X_{m_M}q_{m_M}]$ of M blocks $X_{m_i}q_{m_i}$ of dimension n by p_{m_i} and where $q_{m_1}, \ldots q_{m_M}$ are scalars.

The matrix $\frac{\partial \hat{\beta}}{\partial Y}$ is a p by n block matrix $\begin{bmatrix} \frac{\partial \hat{\beta}_{m_1}}{\partial Y} \\ \vdots \\ \frac{\partial \hat{\beta}_{m_M}}{\partial Y} \end{bmatrix}$ of M blocks $\frac{\partial \hat{\beta}_{m_i}}{\partial Y}$ of dimension p_{m_i} by n.

By the properties of block multiplication, $A\frac{\partial \hat{\beta}}{\partial Y} = \sum_{i=1}^{M} q_{m_i}X_{m_i}\frac{\partial \hat{\beta}_{m_i}}{\partial Y}$,

where the scalar q_i commutes to the left side of X_{m_i}.

$$EDF = Trace\left(\sum_{i=1}^{M} q_{m_i}X_{m_i}\frac{\partial \hat{\beta}_{m_i}}{\partial Y}\right) + Trace\left(X\left(\hat{\beta}^T \otimes I_p\right)\frac{\partial q}{\partial Y}\right). \tag{11.2}$$

$$EDF = Trace\left(\sum_{i=1}^{M} q_{m_i}\frac{\partial X_{m_i}\hat{\beta}_{m_i}}{\partial Y}\right) + Trace\left(X\left(\hat{\beta}^T \otimes I_p\right)\frac{\partial q}{\partial Y}\right). \tag{11.3}$$

Using the properties of the Trace function and substituting $X_{m_i}\hat{\beta}_{m_i}$ $with$ $\hat{Y}_{m_i}$,

$$EDF = \sum_{i=1}^{M} q_{m_i} Trace\left(\frac{\partial \hat{Y}_{m_i}}{\partial Y}\right) + Trace\left(X(\hat{\beta}^T \otimes I_p)\frac{\partial q}{\partial Y}\right). \quad (12)$$

Substituting $Trace\left(\frac{\partial \hat{Y}_{m_i}}{\partial Y}\right)$ with EDF_{m_i} by using equation (6) results in Proposition I.

Proposition I equation

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace\left(X(\hat{\beta}^T \otimes I_p)\frac{\partial q}{\partial Y}\right). \quad (13)$$

Proposition II is derived from Proposition I.

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace\left(D\frac{\partial q}{\partial Y}\right), \quad (13.1)$$

where $D = X(\hat{\beta}^T \otimes I_p)$ is an n by p^2 block matrix $[X\hat{\beta}_1 \quad \cdots \quad X\hat{\beta}_P]$, of P blocks $X\hat{\beta}_i$ of dimensions n by p with scalars $\hat{\beta}_1, \ldots, \hat{\beta}_p$.

The matrix $\frac{\partial q}{\partial Y}$ is a p^2 by n matrix as a result of the diagonal p by p matrix q being vectorized and differentiated using the convention articulated in equation (8).

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace\,(E), \quad (13.2)$$

where $E = D\frac{\partial q}{\partial Y}$ is an n by n matrix $\begin{bmatrix} \sum_{j=1}^{P} X_{1j}\hat{\beta}_j \frac{\partial q_j}{\partial Y_1} & \cdots & \sum_{i=1}^{P} X_{1j}\hat{\beta}_j \frac{\partial q_j}{\partial Y_n} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{P} X_{nj}\hat{\beta}_j \frac{\partial q_j}{\partial Y_1} & \cdots & \sum_{i=1}^{P} X_{nj}\hat{\beta}_j \frac{\partial q_j}{\partial Y_n} \end{bmatrix}$ (13.3)

and X_{1j} is the 1 j entry of matrix X.

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace\left(\begin{bmatrix} \sum_{j=1}^{P} X_{1j}\hat{\beta}_j \frac{\partial q_j}{\partial Y_1} & \cdots & \sum_{i=1}^{P} X_{1j}\hat{\beta}_j \frac{\partial q_j}{\partial Y_n} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{P} X_{nj}\hat{\beta}_j \frac{\partial q_j}{\partial Y_1} & \cdots & \sum_{i=1}^{P} X_{nj}\hat{\beta}_j \frac{\partial q_j}{\partial Y_n} \end{bmatrix}\right). \quad (14)$$

By grouping terms by underlying auxiliary models and using the fact that $\frac{\partial q_j}{\partial Y_1}$ is constant within the sum for each auxiliary model,

$$\sum_{j=1}^{P} X_{1j}\hat{\beta}_j \frac{\partial q_j}{\partial Y_1} = \sum_{i=1}^{M}\sum_{k=1}^{p_{m_i}} \left(X_{1k}\hat{\beta}_k \frac{\partial q_k}{\partial Y_1}\right) = \sum_{i=1}^{M}\sum_{k=1}^{p_{m_i}} \left(X_{1k}\hat{\beta}_k\right) \frac{\partial q_{m_i}}{\partial Y_1} = \sum_{i=1}^{M} \hat{Y}_{1i} \frac{\partial q_{m_i}}{\partial Y_1}, \tag{14.1}$$

where $\hat{Y}_{1i}$ *is* a fitted value of the ith auxiliary model.

E can then be restated as $\begin{bmatrix} \sum_{i=1}^{M} \hat{Y}_{1i} \frac{\partial q_{m_i}}{\partial Y_1} & \cdots & \sum_{i=1}^{M} \hat{Y}_{1i} \frac{\partial q_{m_i}}{\partial Y_n} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{M} \hat{Y}_{ni} \frac{\partial q_{m_i}}{\partial Y_1} & \cdots & \sum_{i=1}^{M} \hat{Y}_{ni} \frac{\partial q_{m_i}}{\partial Y_n} \end{bmatrix}$. (14.2)

E can then be restated as $E = X_A \frac{\partial \hat{\beta}_q}{\partial Y}$, (14.3)

where X_A is an n by m matrix of auxiliary model fitted values $\begin{bmatrix} \hat{Y}_{11} & \cdots & \hat{Y}_{1m} \\ \vdots & \ddots & \vdots \\ \hat{Y}_{n1} & \cdots & \hat{Y}_{nm} \end{bmatrix}$, (14.4)

$\hat{\beta}_q$is an m by 1 vector of $q'_m s$ $\begin{bmatrix} q_{m_1} \\ \vdots \\ q_{m_M} \end{bmatrix}$, (14.5)

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace \left(\begin{bmatrix} \sum_{i=1}^{M} \hat{Y}_{1i} \frac{\partial q_{m_i}}{\partial Y_1} & \cdots & \sum_{i=1}^{M} \hat{Y}_{1i} \frac{\partial q_{m_i}}{\partial Y_n} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{M} \hat{Y}_{ni} \frac{\partial q_{m_i}}{\partial Y_1} & \cdots & \sum_{i=1}^{M} \hat{Y}_{ni} \frac{\partial q_{m_i}}{\partial Y_n} \end{bmatrix} \right), \text{and} \tag{15}$$

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace \left(X_A \frac{\partial \hat{\beta}_q}{\partial Y} \right). \tag{16}$$

Utilizing the condition that $\hat{\beta}_q$ is estimated by a single equation linear in specification as a function of X_A results in the form $\hat{\beta}_q = F(X_A)Y$, which allows equation (16) to be rewritten as (16.1):

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace \left(X_A \frac{\partial F(X_A)Y}{\partial Y} \right). \tag{16.1}$$

Equation (16.1) is differentiated using equation (9):

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace\left(X_A\left((Y^T \otimes I_m)\frac{\partial F(X_A)}{\partial Y} + (I_1 \otimes F(X_A))I_n\right)\right). \tag{16.2}$$

This simplifies to equation (17):

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + Trace(X_A F(X_A)) + Trace\left(X_A\left((Y^T \otimes I_m)\frac{\partial F(X_A)}{\partial Y}\right)\right). \tag{17}$$

The notation for $(Y^T \otimes I_m)\frac{\partial F(X_A)}{\partial Y}$ is rewritten as $\hat{\beta}_{\Delta q}$, which is an m by n matrix.

$$\hat{\beta}_{\Delta q} = \left(Y^T \otimes I_m\right)\frac{\partial F(X_A)}{\partial Y} = \begin{bmatrix} Y_1 & 0 & 0 & \cdots & Y_n & 0 & 0 \\ 0 & Y_1 & 0 & \cdots & 0 & Y_n & 0 \\ 0 & 0 & Y_1 & \cdots & 0 & 0 & Y_n \end{bmatrix} \begin{bmatrix} \frac{\partial F(X_A)_{11}}{\partial Y_1} & \cdots & \frac{\partial F(X_A)_{11}}{\partial Y_n} \\ \frac{\partial F(X_A)_{21}}{\partial Y_1} & & \frac{\partial F(X_A)_{21}}{\partial Y_n} \\ \vdots & & \vdots \\ \frac{\partial F(X_A)_{m1}}{\partial Y_1} & \cdots & \frac{\partial F(X_A)_{m1}}{\partial Y_n} \\ \frac{\partial F(X_A)_{12}}{\partial Y_1} & & \frac{\partial F(X_A)_{12}}{\partial Y_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial F(X_A)_{mn}}{\partial Y_1} & & \frac{\partial F(X_A)_{mn}}{\partial Y_n} \end{bmatrix}. \tag{17.1}$$

$$\hat{\beta}_{\Delta q} = \begin{bmatrix} \sum_{l=1}^{n} \frac{Y_l \partial F(X_A)_{1l}}{\partial Y_1} & \cdots & \sum_{l=1}^{n} \frac{Y_l \partial F(X_A)_{1l}}{\partial Y_n} \\ \vdots & \ddots & \vdots \\ \sum_{l=1}^{n} \frac{Y_l \partial F(X_A)_{ml}}{\partial Y_1} & \cdots & \sum_{l=1}^{n} \frac{Y_l \partial F(X_A)_{ml}}{\partial Y_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial F(X_A)_{11}}{\partial Y_1} & \cdots & \frac{\partial F(X_A)_{1n}}{\partial Y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial F(X_A)_{m1}}{\partial Y_1} & \cdots & \frac{\partial F(X_A)_{mn}}{\partial Y_n} \end{bmatrix} Y. \tag{17.2}$$

$$\hat{\beta}_{\Delta q} = \begin{bmatrix} \frac{\partial F(X_A)_{11}}{\partial Y_1} & \cdots & \frac{\partial F(X_A)_{1n}}{\partial Y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial F(X_A)_{m1}}{\partial Y_1} & \cdots & \frac{\partial F(X_A)_{mn}}{\partial Y_n} \end{bmatrix} Y. \tag{17.3}$$

As shown in (17.3), $\hat{\beta}_{\Delta q}$ is the change in $F(X_A)Y$ resulting from X_A being a function of Y.

*A*s a result, the final term in equation (17) can be written as $Trace\left(X_A \hat{\beta}_{\Delta q}\right)$, resulting in Proposition II.

Proposition II equation

$$EDF = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + EDF_w + Trace\left(X_A \hat{\beta}_{\Delta q}\right). \tag{18}$$

Computing the limit of $Trace\left(X_A \hat{\beta}_{\Delta q}\right)\ as\ \hat{Y}'s \to Y$:

$$\text{As } \hat{Y}'s \to Y, \text{the matrix } X_A = \begin{bmatrix} \hat{Y}_{11} & \cdots & \hat{Y}_{1m} \\ \vdots & \ddots & \vdots \\ \hat{Y}_{n1} & \cdots & \hat{Y}_{nm} \end{bmatrix} \to \begin{bmatrix} Y_1 & \cdots & Y_1 \\ \vdots & \ddots & \vdots \\ Y_n & \cdots & Y_n \end{bmatrix}. \tag{18.1}$$

$$\text{As } X_A \to \begin{bmatrix} Y_1 & \cdots & Y_1 \\ \vdots & \ddots & \vdots \\ Y_n & \cdots & Y_n \end{bmatrix}, \text{the vector } \hat{\beta}_q = F(X_A)Y \ \to \phi, \text{where } \phi \text{ is a vector of length } m. \tag{18.2}$$

$$\text{As } \hat{\beta}_q = F(X_A)Y \ \to \phi, \text{the sum } \sum_{i=1}^{m} \hat{\beta}_{q_i} \to 1. \tag{18.3}$$

$$\text{As } \sum_{i=1}^{m} \hat{\beta}_{q_i} \to 1, \text{any changes in the elements of } \hat{\beta}_q \text{ are offsetting; therefore,}$$

$$\sum_{i=1}^{m} \frac{\partial \hat{\beta}_{q_i}}{\partial \tau} \to 0 \text{ for an arbitrary } \tau. \tag{18.4}$$

Using the substitution $\hat{\beta}_q = F(X_A)Y$, equation (18.4) can be restated as (18.5):

$$\sum_{i=1}^{m} \frac{\partial \hat{\beta}_{q_i}}{\partial \tau} = \sum_{i=1}^{m} \frac{\partial (F(X_A)Y)_i}{\partial \tau} = \sum_{i=1}^{m} \frac{\partial (\sum_{l=1}^{n} Y_l F(X_A)_{il})_i}{\partial \tau} \to 0 \text{ for an arbitrary } \tau. \tag{18.5}$$

Taking the derivative in equation (18.5) results in (18.6):

$$\sum_{i=1}^{m} \frac{\partial (\sum_{l=1}^{n} Y_l F(X_A)_{il})_i}{\partial \tau} = \sum_{i=1}^{m} \sum_{l=1}^{n} \left(\frac{F(X_A)_{il} \partial Y_l}{\partial \tau} + \frac{Y_l \partial F(X_A)_{il}}{\partial \tau} \right) \to 0 \text{ for an arbitrary } \tau. \tag{18.6}$$

$$\text{For well behaved cases, } \sum_{i=1}^{m} \sum_{l=1}^{n} \left(\frac{F(X_A)_{il} \partial Y_l}{\partial \tau} \right) \neq -1 \sum_{i=1}^{m} \sum_{l=1}^{n} \left(\frac{Y_l \partial F(X_A)_{il}}{\partial \tau} \right), \text{which results}$$

in equations (18.7) and (18.8):

$$\sum_{i=1}^{m} \sum_{l=1}^{n} \left(\frac{F(X_A)_{il} \partial Y_l}{\partial \tau} \right) \to 0 \text{ for an arbitrary } \tau. \tag{18.7}$$

$$\sum_{i=1}^{m}\sum_{l=1}^{n}\left(\frac{Y_l \partial F(X_A)_{il}}{\partial \tau}\right) \to 0 \text{ for an arbitrary } \tau. \tag{18.8}$$

Utilizing the (17.2) representation of $\hat{\beta}_{\Delta q} = \begin{bmatrix} \sum_{l=1}^{n} \frac{Y_l \partial F(X_A)_{1l}}{\partial Y_1} & \cdots & \sum_{l=1}^{n} \frac{Y_l \partial F(X_A)_{1l}}{\partial Y_n} \\ \vdots & \ddots & \vdots \\ \sum_{l=1}^{n} \frac{Y_l \partial F(X_A)_{ml}}{\partial Y_1} & \cdots & \sum_{l=1}^{n} \frac{Y_l \partial F(X_A)_{ml}}{\partial Y_n} \end{bmatrix}$

together with equation (18.8) implies that in the matrix $\hat{\beta}_{\Delta q}$, the sum of each column $\to 0$.

Therefore, as $\sum_{i=1}^{m}\sum_{l=1}^{n}\left(\frac{Y_l \partial F(X_A)_{il}}{\partial \tau}\right) \to 0$ for an arbitrary τ, $\hat{\beta}_{\Delta q} \to \lambda$, where λ is an m by n

matrix with columns that sum to zero. (18.9)

Substituting X_A and $\hat{\beta}_{\Delta q}$ with their respective limits from equations (18.1) and (18.9) results in equation (18.10):

$$\lim_{\hat{Y}'s \to Y} Trace\left(X_A \hat{\beta}_{\Delta q}\right) = Trace\left(\begin{bmatrix} Y_1 & \cdots & Y_1 \\ \vdots & \ddots & \vdots \\ Y_n & \cdots & Y_n \end{bmatrix} \lambda\right). \tag{18.10}$$

Given that the rows of $\begin{bmatrix} Y_1 & \cdots & Y_1 \\ \vdots & \ddots & \vdots \\ Y_n & \cdots & Y_n \end{bmatrix}$ are constant and that the columns of λ sum

to zero, their product is an n by n zero matrix.

$$\lim_{\hat{Y}'s \to Y} Trace\left(X_A \hat{\beta}_{\Delta q}\right) = Trace\left(\begin{bmatrix} Y_1 & \cdots & Y_1 \\ \vdots & \ddots & \vdots \\ Y_n & \cdots & Y_n \end{bmatrix} \lambda\right) = Trace \begin{bmatrix} 0_{11} & \cdots & 0_{n1} \\ \vdots & \ddots & \vdots \\ 0_{n1} & \cdots & 0_{nn} \end{bmatrix} = 0. \tag{18.11}$$

$$\lim_{\hat{Y}'s \to Y} Trace\left(X_A \hat{\beta}_{\Delta q}\right) = 0. \tag{18.12}$$

Therefore, the limit of $Trace\left(X_A \hat{\beta}_{\Delta q}\right)$ as $\hat{Y}'s \to Y$ is zero for well behaved cases.

Approximating $Trace\left(X_A \hat{\beta}_{\Delta q}\right)$ with its limit of zero reduces equation (18) to equation (19):

$$EDF \approx \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + EDF_w + \lim_{\hat{Y}'s \to Y} Trace\left(X_A \hat{\beta}_{\Delta q}\right) = \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + EDF_w + 0. \tag{18.13}$$

$$EDF \approx \sum_{i=1}^{M} q_{m_i} EDF_{m_i} + EDF_w . \tag{19}$$

Appendix 2: Number of ways to group variables into forecast combinations

To compute the number of ways v variables can be grouped into forecast combinations, $\binom{v}{i}$ provides the number of possible groupings of v variables into auxiliary models of i variables, where $\binom{v}{i}$ is the combination operator. Equation (20) computes the total number of possible auxiliary model variable groupings g, where auxiliary models range in size from 1 to v variables.

$$g = \sum_{i=1}^{v} \binom{v}{i}. \tag{20}$$

Then the number of ways the auxiliary model groupings can be arranged into forecast combinations c_0 is arrived at by computing all possible subsets of g by putting this number to a base of 2, provided that forecast combinations of forecast combinations are not included.

$$c_0 = 2^g. \tag{21}$$

To allow for a single generation of forecast combinations of forecast combinations, the initial set is increased from g elements to $g + c_0$ elements, and the total of all possible subsets is c_1.

$$c_1 = 2^{g+c_0}. \tag{22}$$

Appendix 3: Simulation procedure for section 4

The simulation procedure for section 4 is itemized below.

Step 1: Randomly generate p correlated explanatory variables x_i, where $i = 1, \dots p$, such that each x_i has 100 observations drawn from a mean 0 variance 1 normal distribution, with a correlation of 0.8 across $x_i's$.

Step 2: Randomly generate the dependent variable as $y = \frac{1}{p}x_1 + \cdots + \frac{1}{p}x_p + \varepsilon$, where ε is randomly generated noise from a mean 0 variance 1 normal distribution.

Step 3: Repeat steps 1 and 2 for the four cases $p = 4, 9, 25, 100$.

Step 4: Using observations 1 to 99, estimate a forecast combination model of y for each of the four cases of p, where the auxiliary models are estimated with ordinary least squares of the forms provided below and the weighting scheme is estimated with ordinary least squares under the constraint $\sum_{i=1}^{M} q_{m_i} = 1$:

- for $p = 4$, two auxiliary models of two variables each $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
- for $p = 9$, three auxiliary models of three variables each $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_3 x_3$
- for $p = 25$, five auxiliary models of five variables each $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_5 x_5$
- for $p = 100$, ten auxiliary models of ten variables each $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{10} x_{10}$

Step 5: Using Mallows' C_p formula,[10] $C_p = \overline{err} + 2\frac{DF}{n}\hat{\sigma}_\varepsilon^2$, where C_p is the C_p statistic, $\overline{err}$ is the training error, $\hat{\sigma}_\varepsilon^2$ is the noise variance. For each of the four cases, compute two estimates of the out-of-sample mean squared forecast error using *DF = equation 19* and $DF = count\ of\ estimated\ parameters$. For both estimates, set $\hat{\sigma}_\varepsilon^2$ equal to its true value of 1.

Step 6: For each of the four cases, use the models from step 4 estimated on observations 1 to 99 and the 100th observation of the $x_i's$, generate out of sample forecasts for the 100th observation of y, and calculate the out-of-sample squared forecast error.

Step 7: Repeat steps 1 to 6 100,000 times and compare the average performance of the two estimates from step 5 with true out-of-sample squared forecast errors resulting from step 6.

[10] Hastie, Tibshirani and Friedman (2001, chapter 7) discuss using this formula to estimate out-of-sample mean squared forecast errors.

References

Bates, J., and C. Granger. 1969. "The Combination of Forecasts." *Operations Research Quarterly* 20 (4): 451–468.

Breiman, L. 1996. "Stacked Regressions," *Machine Learning* 24: 49–64.

Claeskens, G., J. Magnus, A. Vasnev and W. Wang. 2016. "The Forecast Combination Puzzle: A Simple Theoretical Explanation." *International Journal of Forecasting* 32 (3): 754–762.

Diebold, F. and M. Shin. 2019. "Machine Learning for Regularized Survey Forecast Combination: Partially-Egalitarian LASSO and its Derivatives." *International Journal of Forecasting* 35 (4): 1679–1691.

Efron, B. 2004. "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation." *Journal of the American Statistical Association* 99 (467): 619–642.

Efron, B., T. Hastie, I. Johnstone and R. Tibshirani. 2004. "Least Angle Regression." *Annals of Statistics* 32 (2): 407–499.

Elliott, G. and Timmermann. 2005. "Optimal Forecast Combination Under Regime Switching." *International Economic Review* 46 (4): 1081–1102.

Hansen, B. 2007. "Least Squares Model Averaging." *Econometrica* 75 (4): 1175–1189.

Harville, D. 2008. *Matrix Algebra from A Statistician's Perspective*. New York: Springer Science + Business Media.

Hastie, T., R. Tibshirani and J. Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York: Springer Science + Business Media.

Hsiao, C. and S. Wan. 2014. "Is There an Optimal Forecast Combination?" *Journal of Econometrics* 178 (P2): 294–309.

Kato, K. 2009. "On the Degrees of Freedom in Shrinkage Estimation." *Journal of Multivariate Analysis* 100 (7): 1338–1352.

LeBlanc, M. and R. Tibshirani. 1996. "Combining Estimates in Regression and Classification." *Journal of the American Statistical Association* 91 (436): 1641–1650.

Magnus, J. (2010). "On the Concept of Matrix Derivative." *Journal of Multivariate Analysis* 101 (9): 2200–2206.

Mukherjee, A., K. Chen, N. Wang and J. Zhu. 2015. "On the Degrees of Freedom of Reduced-Rank Estimators in Multivariate Regression." *Biometrika* 102 (2): 457–477.

Samuels, J. and R. Sekkel. 2017. "Model Confidence Sets and Forecast Combination." *International Journal of Forecasting* 33 (1): 48–60.

Stein, C. 1981. "Estimation of the Mean of a Multivariate Normal Distribution." *Annals of Statistics* 9 (6): 1135–1151.

Stock, J. and M. Watson. 2004. “Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting* 23 (6): 405–430.

Tibshirani, R. and J. Taylor. 2012. “Degrees of Freedom in Lasso Problems.” *Annals of Statistics* 40 (2): 1198–1232.

Zou, H., T. Hastie and R. Tibshirani. 2007. “On the ‘Degrees of Freedom’ of the Lasso.” *Annals of Statistics* 35 (5): 2173–2192.