Linear Plus Quadratic Approach to the Mean Square Error Optimal Combination of Forecasts

Sven-Oliver Troschke and Götz Trenkler
Department of Statistics, University of Dortmund,
44221 Dortmund, Germany
troschke@statistik.uni-dortmund.de

Abstract: This paper deals with linear plus quadratic approaches aiming to find a combined forecast for a scalar random variable from several individual forecasts for that variable. When combining forecasts linear approaches have been used predominantly. One reason may be the well-known fact that the linear approach with constant term is optimal with respect to the mean square prediction error loss, if the single forecasts and the target variable follow a joint normal distribution. In this paper no assumption is made on the type of the joint distribution. Its moments up to order four, however, are assumed to be given for the derivation of the optimal combination parameters. Three versions for the quadratic part of the combined forecast are discussed. As a by-product a linear plus quadratic adjustment of a single forecast is obtained. In order to apply these methods to empirical data the moments of the joint distribution have to be estimated.

Keywords: Combination of forecasts, Linear plus quadratic combination.

AMS 2000 Subject Classification: 62M20

1 Introduction

Suppose that we are given k forecasts $f_1, \dots, f_k$ for a scalar random variable y. The forecasts are gathered in a random vector $\mathbf{f}$, i.e. $\mathbf{f} = (f_1, \dots, f_k)^{\mathrm{T}}$. Our aim is to obtain combined forecasts f_{comb} from the single forecasts f_i which are optimal within certain given classes of combinations.

Optimality in this paper is always understood as optimality with respect to the mean square prediction error (MSPE). Given a forecast f for a random variable y

the MSPE is given by

$$\mathrm{MSPE}(f,y) = \mathrm{E}[(y-f)^2] = \mathrm{Var}(y-f)+[\mathrm{E}(y-f)]^2 \ . \tag{1.1}$$

It is a well-known fact (see e.g. THIELE, 1993) that a linear combination $f_{\mathbf{b},c} = \mathbf{b}^{\mathrm{T}}\mathbf{f} + c$ with suitably chosen $\mathbf{b} = (b_1, \ldots, b_k)^{\mathrm{T}} \in \mathbb{R}^k$ and $c \in \mathbb{R}$ is optimal among all combinations if y and $\mathbf{f}$ follow a joint normal distribution.
In the absence of joint normality, however, it is worthwhile to consider nonlinear forecast combinations. Stimulated by TAYLOR's series expansion formula we may try to 'approximate' the target variable y by a *linear plus quadratic* function in $\mathbf{f}$

$$f_{\mathbf{A},\mathbf{b},c} = \mathbf{f}^{\mathrm{T}}\mathbf{A}\mathbf{f} + \mathbf{b}^{\mathrm{T}}\mathbf{f} + c \ , \tag{1.2}$$

rather than by a linear function. Here $c \in \mathbb{R}$, $\mathbf{b} = (b_1, \ldots, b_k)^{\mathrm{T}} \in \mathbb{R}^k$ and

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ a_{12} & a_{22} & \ldots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1k} & a_{2k} & \ldots & a_{kk} \end{pmatrix} \in \mathbb{R}^{k \times k} \tag{1.3}$$

may be assumed to be symmetric without loss of generality, because it only appears within the quadratic form $\mathbf{f}^{\mathrm{T}}\mathbf{A}\mathbf{f}$.
In order to apply such a linear plus quadratic combination of forecasts *two steps* have to be taken:
In the *first step* we derive the theoretically optimal combination parameters $\mathbf{A}_{\mathrm{opt}}$, $\mathbf{b}_{\mathrm{opt}}$ and c_{opt} such that

$$\mathrm{MSPE}(f_{\mathbf{A}_{\mathrm{opt}},\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}, y) \leq \mathrm{MSPE}(f_{\mathbf{A},\mathbf{b},c}, y) \tag{1.4}$$

for all symmetric matrices $\mathbf{A}$, vectors $\mathbf{b}$ and scalars c. Clearly, the optimal linear plus quadratic combination also outperforms the optimal linear combination $\mathbf{f}_{\mathbf{b}^*_{\mathrm{opt}},c^*_{\mathrm{opt}}}$ since the latter may be regarded as a linear plus quadratic combination with $\mathbf{A} = \mathbf{0}$, $\mathbf{b} = \mathbf{b}^*_{\mathrm{opt}}$ and $c = c^*_{\mathrm{opt}}$.
For the determination of the optimal combination parameters we will assume that the first to fourth order moments of the joint distribution of y and $\mathbf{f}$ exist. If this is not the case, e.g. if the target variable y and the single forecasts f_i are trended, then appropriate transformations of y and $\mathbf{f}$ should be undertaken, e.g. differencing of the time series of observations or consideration of relative changes. Since $f_1, \ldots, f_k$ are forecasts of y the same transformation should be appropriate for both, target variable and forecasts.

Furthermore, we have to assume that we *know* the first to fourth order moments of the joint distribution of y and $\mathbf{f}$. (This describes a state of knowledge between states 1 and 2 in the classification scheme by HARVILLE (1985). Here state 1 means complete knowledge about the distribution whereas state 2 is described by knowledge of the first and second order moments.) We will see that the optimal linear plus quadratic combination parameters depend on these first to fourth order moments.
In practical applications, however, such moments will hardly ever be known. (Thus our knowledge falls even behind state 4 of knowledge in HARVILLE's scheme, where some assumptions on the first order moments are made.) Consequently, in the *second step* we have to estimate the necessary moments from a sample of observations on the variables of interest. Then we plug these estimators into the formulae for the optimal combinations.
The focus of this paper will be the *first step*, i.e. the derivation of the optimal combination parameters from known first to fourth order moments of the joint distribution of y and $\mathbf{f}$. We will investigate three versions of the linear plus quadratic approach: The combined forecast in Equation (1.2) with no additional restriction (besides symmetry) imposed on the matrix $\mathbf{A}$ is referred to as the *strong version*. Consequently, the strong linear plus quadratic approach involves $k(k+1)/2$ parameters for the quadratic part and $(k+1)(k+2)/2$ parameters in total.
Since the number of observations from which the unknown parameters are to be estimated is not so large in general, it is reasonable to consider reduced linear plus quadratic approaches as well, which involve less parameters. In order to achieve this goal we may restrict $\mathbf{A}$ to be a diagonal matrix or even to be a multiple of the $k \times k$ identity matrix.
Restricting $\mathbf{A}$ to be diagonal leads to the *medium version* of the linear plus quadratic approach

$$f_{\mathbf{a},\mathbf{b},c} = \mathbf{f}^{\mathrm{T}} \operatorname{dg}(\mathbf{a})\mathbf{f} + \mathbf{b}^{\mathrm{T}}\mathbf{f} + c = \sum_{i=1}^{k} a_i f_i^2 + \mathbf{b}^{\mathrm{T}}\mathbf{f} + c \,, \tag{1.5}$$

where $\mathbf{a} = (a_1, \ldots, a_k)^{\mathrm{T}} \in \mathbb{R}^k$,

$$\operatorname{dg}(\mathbf{a}) = \begin{pmatrix} a_1 & 0 & \ldots & 0 \\ 0 & a_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_k \end{pmatrix} \in \mathbb{R}^{k \times k} \,, \tag{1.6}$$

is a diagonal matrix, $\mathbf{b} = (b_1, \ldots, b_k)^{\mathrm{T}} \in \mathbb{R}^k$ and $c \in \mathbb{R}$. Thereby the number of the elements in $\mathbf{A}$ is reduced to k and the total of unknown parameters is reduced to $2k+1$.

Restricting $\mathbf{A}$ to be a multiple of the identity matrix, i.e. $\mathbf{A} = \alpha \mathbf{I}_k$, leads to the *weak version* of the linear plus quadratic approach which is

$$f_{\alpha,\mathbf{b},c} = \alpha \mathbf{f}^{\mathrm{T}} \mathbf{f} + \mathbf{b}^{\mathrm{T}} \mathbf{f} + c \ , \tag{1.7}$$

where $\alpha \in \mathbb{R}$, $\mathbf{b} = (b_1, \ldots, b_k)^{\mathrm{T}} \in \mathbb{R}^k$ and $c \in \mathbb{R}$. Thus there only remains one single parameter for the quadratic part and $k+2$ unknown parameters in total.
As we will see later on the optimal choice of the combination parameters within the three approaches requires different levels of knowledge about the moments of the joint distribution of y and $\mathbf{f}$. In each case, however, moments up to order four are involved. We will now introduce our notations:
Extending the approach from Harville (1985) and utilizing the notations from Rao and Kleffe (1988) we will assume the following setting: The expectations of y and $\mathbf{f}$ are given by $\mathrm{E}(y) = \mu_0$ and $\mathrm{E}(\mathbf{f}) = \boldsymbol{\mu}_{\mathbf{f}} := (\mu_1, \ldots, \mu_k)^{\mathrm{T}}$, respectively, which gives rise to the model:

$$\begin{pmatrix} y \\ \mathbf{f} \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \boldsymbol{\mu}_{\mathbf{f}} \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \boldsymbol{\varepsilon}_{\mathbf{f}} \end{pmatrix} =: \boldsymbol{\mu} + \boldsymbol{\varepsilon} \ , \tag{1.8}$$

where $\boldsymbol{\varepsilon}_{\mathbf{f}} := (\varepsilon_1, \ldots, \varepsilon_k)^{\mathrm{T}}$. Consequently, $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and the higher order moments of $\boldsymbol{\varepsilon}$ are the centered moments of $(y, \mathbf{f}^{\mathrm{T}})^{\mathrm{T}}$.
First, let us turn to the second order moments:

$$\boldsymbol{\Sigma} := \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \mathrm{E}\left[\begin{pmatrix} \varepsilon_0 \\ \boldsymbol{\varepsilon}_{\mathbf{f}} \end{pmatrix} \begin{pmatrix} \varepsilon_0 \\ \boldsymbol{\varepsilon}_{\mathbf{f}} \end{pmatrix}^{\mathrm{T}} \right] =: \begin{pmatrix} \Sigma_{00} & \boldsymbol{\Sigma}_{0\mathbf{f}} \\ \boldsymbol{\Sigma}_{\mathbf{f}0} & \boldsymbol{\Sigma}_{\mathbf{ff}} \end{pmatrix} \tag{1.9}$$

and

$$\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \mathrm{E}\left[\left(\begin{pmatrix} y \\ \mathbf{f} \end{pmatrix} - \begin{pmatrix} \mu_0 \\ \boldsymbol{\mu}_{\mathbf{f}} \end{pmatrix} \right) \left(\begin{pmatrix} y \\ \mathbf{f} \end{pmatrix} - \begin{pmatrix} \mu_0 \\ \boldsymbol{\mu}_{\mathbf{f}} \end{pmatrix} \right)^{\mathrm{T}} \right] = \mathrm{Cov} \begin{pmatrix} y \\ \mathbf{f} \end{pmatrix} . \tag{1.10}$$

The lower left $(k \times 1)$–submatrix $\boldsymbol{\Sigma}_{\mathbf{f}0}$ and the lower right $(k \times k)$–submatrix $\boldsymbol{\Sigma}_{\mathbf{ff}}$ of $\boldsymbol{\Sigma}$ read explicitly

$$\boldsymbol{\Sigma}_{\mathbf{f}0} = \begin{pmatrix} \Sigma_{10} \\ \Sigma_{20} \\ \vdots \\ \Sigma_{k0} \end{pmatrix} \quad \text{and } \boldsymbol{\Sigma}_{\mathbf{ff}} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1k} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{k1} & \Sigma_{k2} & \cdots & \Sigma_{kk} \end{pmatrix} . \tag{1.11}$$

We will assume invertibility of the centered second order moment matrix of $\mathbf{f}$ throughout, i.e. we assume invertibility of $\boldsymbol{\Sigma}_{\mathbf{ff}} = \mathrm{Cov}(\mathbf{f})$ and hence also invertibility

of the non-centered second order moment matrix $\mathbf{\Sigma}_{\mathbf{ff}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}} = \mathrm{E}(\mathbf{f}\mathbf{f}^{\mathrm{T}})$ is granted. Note that vectors and matrices are represented by bold face letters.
Analogously, the third order moments of $\boldsymbol{\varepsilon}$ are given by

$$\mathbf{\Phi} := \mathrm{E}(\boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \begin{pmatrix} \mathbf{\Phi}_0 \\ \mathbf{\Phi}_1 \\ \vdots \\ \mathbf{\Phi}_k \end{pmatrix}, \tag{1.12}$$

where

$$\mathbf{\Phi}_i = \mathrm{E}(\varepsilon_i \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \begin{pmatrix} \Phi_{i00} & \mathbf{\Phi}_{i0\mathbf{f}} \\ \mathbf{\Phi}_{i\mathbf{f}0} & \mathbf{\Phi}_{i\mathbf{ff}} \end{pmatrix}, \; i = 0, \dots, k \tag{1.13}$$

and the fourth order moments are given by

$$\mathbf{\Psi} = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \otimes \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \begin{pmatrix} \mathbf{\Psi}_{00} & \mathbf{\Psi}_{01} & \dots & \mathbf{\Psi}_{0k} \\ \mathbf{\Psi}_{10} & \mathbf{\Psi}_{11} & \dots & \mathbf{\Psi}_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Psi}_{k0} & \mathbf{\Psi}_{k1} & \dots & \mathbf{\Psi}_{kk} \end{pmatrix}, \tag{1.14}$$

where

$$\mathbf{\Psi}_{ij} = \mathrm{E}(\varepsilon_i \varepsilon_j \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \begin{pmatrix} \Psi_{ij00} & \mathbf{\Psi}_{ij0\mathbf{f}} \\ \mathbf{\Psi}_{ij\mathbf{f}0} & \mathbf{\Psi}_{ij\mathbf{ff}} \end{pmatrix}, \; i, j = 0, \dots, k. \tag{1.15}$$

Note that $\mathbf{\Sigma}$, $\mathbf{\Phi}_i$ and $\mathbf{\Psi}_{ij}$ are symmetric matrices of order $(k+1)\times(k+1)$. Furthermore $\mathbf{\Psi}_{ij} = \mathbf{\Psi}_{ji}$ such that the matrix $\mathbf{\Psi}$ is symmetric as well. The elements of $\mathbf{\Phi}$ are $\Phi_{ijl} = \mathrm{E}(\varepsilon_i\varepsilon_j\varepsilon_l)$ and the elements of $\mathbf{\Psi}$ are $\Psi_{ijlm} = \mathrm{E}(\varepsilon_i\varepsilon_j\varepsilon_l\varepsilon_m)$.
Section 2 deals with the classical linear approaches within the framework of this paper whereas Sections 3, 4 and 5 investigate the respective linear plus quadratic approaches. Section 6 considers the special case of combining $k = 2$ forecasts. Setting $k = 1$ we obtain and investigate adjustments of individual forecasts in Section 7. The question in how far the various methods are sensitive to the chosen coordinate system is discussed in Section 8. Section 9 concludes the paper.
Section A in the appendix lists some results mostly from the theory of matrix differential calculus which will be useful in the subsequent sections.

2 The linear approach

Linearly combined forecasts are of the form $\mathbf{b}^{\mathrm{T}}\mathbf{f} + c$, where it may be appropriate to impose certain restrictions on the combination parameters $\mathbf{b}$ and c. Linear ap-

proaches have been widely discussed in the literature, compare e.g. CLEMEN (1989) or THIELE (1993) for good overviews on the topic.
To derive the theoretically optimal combination parameters within the linear approaches we only need the first and second order moments of the joint distribution of y and $\mathbf{f}$ to exist and to be known.
We will consider four versions of the linear approach: The first is

$$f_{\mathbf{b},c} = \mathbf{b}^{\mathrm{T}}\mathbf{f} + c \ . \tag{2.1}$$

As stated in Section 1, with suitably chosen parameters, this version leads to the MSPE-optimal combined forecast under joint normality of y and $\mathbf{f}$.
A simpler approach is to define the combined forecast to be a weighted average of the single forecasts

$$f_{\mathbf{b}} = \mathbf{b}^{\mathrm{T}}\mathbf{f} \ . \tag{2.2}$$

If each of the single forecasts is unbiased it is a well-known fact that the combined forecast is unbiased as well if, in the second approach, the parameters are chosen such that they sum up to unity, i.e. $\mathbf{b}^{\mathrm{T}}\mathbf{1} = 1$. This leads to the third version of the linear approach which utilizes this restriction:

$$f_{\mathbf{b},\mathrm{rest}} = \mathbf{b}^{\mathrm{T}}\mathbf{f} \ , \ \text{where } \mathbf{b}^{\mathrm{T}}\mathbf{1} = 1 \ . \tag{2.3}$$

If the individual forecasts f_i are biased it is reasonable to perform a bias correction $f_i - \mu_i + \mu_0$ before combining them. After the correction the individual forecasts are unbiased and, hence, they should be combined with weights summing up to unity. This leads to the restricted linear combination with absolute term:

$$f_{\mathbf{b},c,\mathrm{rest}} = \mathbf{b}^{\mathrm{T}}\mathbf{f} + c \ , \ \text{where } \mathbf{b}^{\mathrm{T}}\mathbf{1} = 1 \ . \tag{2.4}$$

For each of the four versions we now want to state how the combination parameters should be chosen in order to minimize the mean square prediction error of such a combined forecast and we will provide the respective minimal values.
When considering the last two approaches which utilize the restriction on the vector $\mathbf{b}$ it is common practice in the literature to do this by using the moments of the distribution of errors $\mathbf{e} = \mathbf{f} - y\mathbf{1}_k$, i.e. $\mathbf{e} = (e_1, \ldots, e_k)$ is the vector containing the single forecast errors.
In linear forecast combination under the restriction $\mathbf{b}^{\mathrm{T}}\mathbf{1} = 1$ consideration of the errors is appealing. In this case (and only in this case) the same weights b_i are

assigned to the single forecasts f_i to yield the combined forecast *as well as* to the single errors e_i to yield the error of the combined forecast:

$$e_{\mathbf{b},\mathrm{rest}} = f_{\mathbf{b},\mathrm{rest}} - y = \mathbf{b}^{\mathrm{T}}\mathbf{f} - \mathbf{b}^{\mathrm{T}}\mathbf{1}y = \mathbf{b}^{\mathrm{T}}\mathbf{e} \tag{2.5}$$

and

$$e_{\mathbf{b},c,\mathrm{rest}} = f_{\mathbf{b},c,\mathrm{rest}} - y = \mathbf{b}^{\mathrm{T}}\mathbf{f} + c - \mathbf{b}^{\mathrm{T}}\mathbf{1}y = \mathbf{b}^{\mathrm{T}}\mathbf{e} + c\ . \tag{2.6}$$

Consequently, we may consider the forecast errors instead of the forecasts in order to obtain the optimal combination weights which are to be assigned to the single forecasts. No similar result holds for non-linear combinations (like the linear plus quadratic combinations considered here) or for linear combinations without the restriction $\mathbf{b}^{\mathrm{T}}\mathbf{1} = 1$.
In linear forecast combination under the restriction $\mathbf{b}^{\mathrm{T}}\mathbf{1} = 1$ it is, however, equivalent to base our derivations on the moments of $(y, \mathbf{f}^{\mathrm{T}})^{\mathrm{T}}$ or on the moments of $\mathbf{e}$, if we assume that the first and second order moments of $(y, \mathbf{f}^{\mathrm{T}})^{\mathrm{T}}$ exist. Note that $\mathrm{E}[(\mathbf{b}^{\mathrm{T}}\mathbf{e})^2] = \mathrm{E}(e^2_{\mathbf{b},\mathrm{rest}}) = \mathrm{MSPE}(f_{\mathbf{b},\mathrm{rest}}, y) = \mathrm{Var}(y - \mathbf{b}^{\mathrm{T}}\mathbf{f}) + [\mathrm{E}(y - \mathbf{b}^{\mathrm{T}}\mathbf{f})]^2$.
For these reasons we use the moments of the joint distribution of y and $\mathbf{f}$ throughout (like e.g. in HARVILLE, 1985). By doing so we ensure comparability of the results from the various approaches.

First we will consider the linear approach with constant term c and without restrictions on the vector $\mathbf{b}$, i.e. we consider $f_{\mathbf{b},c} = \mathbf{b}^{\mathrm{T}}\mathbf{f} + c$ with expectation

$$\mathrm{E}(f_{\mathbf{b},c}) = \mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + c\ . \tag{2.7}$$

From HARVILLE (1985), Equation (2.1) we know that the optimal choices for $\mathbf{b}$ and c are given by

$$\mathbf{b}_{\mathrm{opt}} = \boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}0} \quad \text{and} \quad c_{\mathrm{opt}} = \mu_0 - \boldsymbol{\Sigma}_{\mathbf{f}0}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\mu}_{\mathbf{f}} \tag{2.8}$$

leading to the optimal value of the MSPE-function

$$\mathrm{MSPE}(f_{\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}, y) = \Sigma_{00} - \boldsymbol{\Sigma}_{\mathbf{f}0}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}0}\ . \tag{2.9}$$

Obviously, the combined forecast $f_{\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}$ is unbiased even if the single forecasts are biased.

Now we turn to the linear approach without constant term and without restrictions on the vector $\mathbf{b}$, i.e. we consider $f_{\mathbf{b}} = \mathbf{b}^{\mathrm{T}}\mathbf{f}$ with expectation

$$\mathrm{E}(f_{\mathbf{b}}) = \mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}}\ . \tag{2.10}$$

The mean square prediction error of a forecast combination $f_{\mathbf{b}}$ is given by

$$\mathrm{MSPE}(f_{\mathbf{b}}, y) = \mathbf{b}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathbf{ff}}+\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}})\mathbf{b}-2\mathbf{b}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathbf{f}0}+\mu_0\boldsymbol{\mu}_{\mathbf{f}})+\Sigma_{00}+\mu_0^2 \ . \tag{2.11}$$

Differentiating this function with respect to $\mathbf{b}$ and setting the derivative equal to zero we arrive at the optimal choice for $\mathbf{b}$ within this approach, namely

$$\mathbf{b}_{\mathrm{opt}} = \left(\boldsymbol{\Sigma}_{\mathbf{ff}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{\Sigma}_{\mathbf{f}0} + \mu_0\boldsymbol{\mu}_{\mathbf{f}}\right) \ . \tag{2.12}$$

Inserting this optimal weight vector into Equation (2.10) it can be seen that the corresponding linear combination is not necessarily unbiased even if the individual forecasts are unbiased. The optimal MSPE-value is given by

$$\mathrm{MSPE}(f_{\mathbf{b}_{\mathrm{opt}}}, y) = \Sigma_{00}+\mu_0^2-(\boldsymbol{\Sigma}_{\mathbf{f}0}+\mu_0\boldsymbol{\mu}_{\mathbf{f}})^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathbf{ff}}+\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}})^{-1}(\boldsymbol{\Sigma}_{\mathbf{f}0}+\mu_0\boldsymbol{\mu}_{\mathbf{f}}) \ . \tag{2.13}$$

Using Lemma A.1 this may be rewritten as

$$\mathrm{MSPE}(f_{\mathbf{b}_{\mathrm{opt}}}, y) = \Sigma_{00}-\boldsymbol{\Sigma}_{\mathbf{f}0}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}0}+\frac{\left(\mu_0 - \boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}0}\right)^2}{1+\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\mu}_{\mathbf{f}}} \tag{2.14}$$

such that from comparing this formula to Equation (2.9) the loss caused by dropping the constant term becomes evident.

Next, we investigate the linear approach without constant term and with restriction on the vector $\mathbf{b}$, i.e. we consider $f_{\mathbf{b},\mathrm{rest}} = \mathbf{b}^{\mathrm{T}}\mathbf{f}$ with $\mathbf{b}^{\mathrm{T}}\mathbf{1} = 1$. This combination approach is designed for the situation where each single forecast is unbiased, i.e. $\boldsymbol{\mu}_{\mathbf{f}} = \mathrm{E}(\mathbf{f}) = \mathrm{E}(y)\mathbf{1} = \mu_0\mathbf{1}$. Namely, under the unbiasedness assumption $f_{\mathbf{b},\mathrm{rest}}$ is unbiased as well:

$$\mathrm{E}(f_{\mathbf{b},\mathrm{rest}}) = \mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} = \mu_0\mathbf{b}^{\mathrm{T}}\mathbf{1} = \mathrm{E}(y) \ . \tag{2.15}$$

Consequently, also the calculation of the optimal combination weights and the corresponding optimal MSPE-value for $f_{\mathbf{b},\mathrm{rest}}$ are performed under the unbiasedness assumption:
Evidently, the MSPE-function is the same as that given in Equation (2.11), but using $\boldsymbol{\mu}_{\mathbf{f}} = \mu_0\mathbf{1}$ this may be rewritten as

$$\mathrm{MSPE}(f_{\mathbf{b},\mathrm{rest}}, y) = \mathbf{b}^{\mathrm{T}}(\breve{\boldsymbol{\Sigma}}_{\mathbf{ff}}+\mu_0^2\mathbf{1}\mathbf{1}^{\mathrm{T}})\mathbf{b}-2\mathbf{b}^{\mathrm{T}}(\breve{\boldsymbol{\Sigma}}_{\mathbf{f}0}+\mu_0^2\mathbf{1})+\breve{\Sigma}_{00}+\mu_0^2 \ , \tag{2.16}$$

where

$$\breve{\boldsymbol{\Sigma}} = \mathrm{E}\left[\left(\begin{pmatrix} y \\ \mathbf{f} \end{pmatrix} - \mu_0\mathbf{1}\right)\left(\begin{pmatrix} y \\ \mathbf{f} \end{pmatrix} - \mu_0\mathbf{1}\right)^{\mathrm{T}}\right] \tag{2.17}$$

is the covariance matrix of $(y, \mathbf{f}^{\mathrm{T}})^{\mathrm{T}}$ under the unbiasedness assumption.
In order to minimize this function with respect to $\mathbf{b}$ under the restriction $\mathbf{b}^{\mathrm{T}}\mathbf{1} = 1$ we follow a LAGRANGE multiplier approach to obtain

$$\mathbf{b}_{\mathrm{opt}} = \breve{\boldsymbol{\Sigma}}_{\mathbf{ff}}^{-1}\breve{\boldsymbol{\Sigma}}_{\mathbf{f}0} + \frac{1 - \mathbf{1}^{\mathrm{T}}\breve{\boldsymbol{\Sigma}}_{\mathbf{ff}}^{-1}\breve{\boldsymbol{\Sigma}}_{\mathbf{f}0}}{\mathbf{1}^{\mathrm{T}}\breve{\boldsymbol{\Sigma}}_{\mathbf{ff}}^{-1}\mathbf{1}}\breve{\boldsymbol{\Sigma}}_{\mathbf{ff}}^{-1}\mathbf{1}\ , \tag{2.18}$$

which leads to the optimal MSPE-value

$$\mathrm{MSPE}(f_{\mathbf{b}_{\mathrm{opt}},\mathrm{rest}}, y) = \breve{\Sigma}_{00} - \breve{\boldsymbol{\Sigma}}_{\mathbf{f}0}^{\mathrm{T}}\breve{\boldsymbol{\Sigma}}_{\mathbf{ff}}^{-1}\breve{\boldsymbol{\Sigma}}_{\mathbf{f}0} + \frac{(1 - \mathbf{1}^{\mathrm{T}}\breve{\boldsymbol{\Sigma}}_{\mathbf{ff}}^{-1}\breve{\boldsymbol{\Sigma}}_{\mathbf{f}0})^2}{\mathbf{1}^{\mathrm{T}}\breve{\boldsymbol{\Sigma}}_{\mathbf{ff}}^{-1}\mathbf{1}}\ . \tag{2.19}$$

If the assumption of unbiasedness is satisfied, then, of course, the matrix $\breve{\boldsymbol{\Sigma}}$ coincides with the true covariance matrix $\boldsymbol{\Sigma}$. If, however, the unbiasedness assumption is violated and $f_{\mathbf{b}_{\mathrm{opt}},\mathrm{rest}}$ is applied nevertheless, the combined forecast is based on an incorrect covariance matrix.
In the context of considering $f_{\mathbf{b}_{\mathrm{opt}},\mathrm{rest}}$ the following two observations are interesting. They are proven in Appendix B.

Assertion 2.1 *If the unbiasedness assumption is incorrect it is obvious that the true optimal* MSPE*-value* $\mathrm{MSPE}(f_{\mathbf{b}_{opt},rest}, y)$ *should be calculated by inserting* $\mathbf{b}_{opt}$ *from Equation (2.18) into the general Equation (2.11), which is valid for any linear combination of the type* $\mathbf{b}^{\mathrm{T}}\mathbf{f}$*. We obtain, however, the same Result (2.19) from inserting* $\mathbf{b}_{opt}$ *into the (now invalid) Equation (2.16).*

Assertion 2.2 *The optimal parameter vector* $\mathbf{b}_{opt}$ *is not changed if we use any other constant than* μ_0 *in the calculation of the covariance matrix* $\breve{\boldsymbol{\Sigma}}$ *in Formula (2.17). The covariance matrix* $\breve{\boldsymbol{\Sigma}}$ *itself is changed, though.*
An important consequence for practical applications is that we need not worry about which estimate of μ_0 *should be used when estimating* $\breve{\boldsymbol{\Sigma}}$*: We may use the arithmetic mean of the observations on the target variable* y*, the arithmetic mean of all observations on the target variable* y and *the single forecasts* f_i*, both of which are reasonable estimates, or we may even use 0.*

Equation (2.15) confirms that the combined forecast $f_{\mathbf{b},\mathrm{rest}}$ is unbiased if all single forecasts are unbiased as well. The assumption of unbiasedness for each single forecast seems at least doubtful. GRANGER and RAMANATHAN (1984, p. 200) point out:

> There is nothing sacred about the weights adding up to unity, although that seems to be the common practice. Furthermore, there is no reason to believe that every alternative forecast will be unbiased.

The linear combination $f_{\mathbf{b},c}$, however, is always unbiased whenever the combination parameters $\mathbf{b}$ and c are determined according to the above optimal choice. This is also true for any of the considered combination methods involving a constant term, including the three linear plus quadratic approaches, as we will see later on.

We now turn to the linear approach with constant term and with restriction on the vector $\mathbf{b}$, i.e. we consider $f_{\mathbf{b},c,\mathrm{rest}} = \mathbf{b}^{\mathrm{T}}\mathbf{f} + c$ with $\mathbf{b}^{\mathrm{T}}\mathbf{1} = 1$. Its expectation is given by

$$\mathrm{E}(f_{\mathbf{b},c,\mathrm{rest}}) = \mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + c\,. \tag{2.20}$$

The optimal choices for $\mathbf{b}$ and c can be calculated to be

$$\mathbf{b}_{\mathrm{opt}} = \boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}0} + \frac{1 - \mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}0}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\mathbf{1}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\mathbf{1} \quad \text{and} \quad c_{\mathrm{opt}} = \mu_0 - \mathbf{b}_{\mathrm{opt}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}}\,. \tag{2.21}$$

The corresponding optimal value of the MSPE-function is

$$\mathrm{MSPE}(f_{\mathbf{b}_{\mathrm{opt}},\mathbf{c}_{\mathrm{opt}},\mathrm{rest}}, y) = \Sigma_{00} - \boldsymbol{\Sigma}_{\mathbf{f}0}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}0} + \frac{(1 - \mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}0})^2}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\mathbf{1}}\,. \tag{2.22}$$

Comparing this formula to Equation (2.9) we see which loss is caused by placing the restriction on $\mathbf{b}$.
The optimal weight vector $\mathbf{b}_{\mathrm{opt}}$ and the optimal MSPE-value are in the same form as in the previous approach, but they are calculated from the covariance matrix $\boldsymbol{\Sigma}$ instead of $\breve{\boldsymbol{\Sigma}}$. Regarding Equations (2.20) and (2.21) it becomes evident that the combined forecast $f_{\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}},\mathrm{rest}}$ is unbiased even if the single forecasts are biased.
Finally, we will also include the arithmetic mean of the individual forecasts in our considerations, since it is a very simple and empirically very powerful statistic:

$$f_{\mathrm{am}} = \frac{1}{k}\sum_{i=1}^{k} f_i = \frac{1}{k}\mathbf{1}^{\mathrm{T}}\mathbf{f}\,. \tag{2.23}$$

Its expectation is

$$\mathrm{E}(f_{\mathrm{am}}) = \frac{1}{k}\mathbf{1}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} \tag{2.24}$$

and thus the unweighted average is not unbiased in general. If, however, each individual forecast is unbiased, then also f_{am} is. The corresponding MSPE-value is given by

$$\mathrm{MSPE}(f_{\mathrm{am}}, y) = \Sigma_{00} - \frac{2}{k}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{f}0} + \frac{1}{k^2}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}\mathbf{1} + \left(\frac{1}{k}\mathbf{1}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} - \mu_0\right)^2 . \tag{2.25}$$

We now turn to the linear plus quadratic approaches to the combination of forecasts. They are of the general form $\mathbf{f}^{\mathrm{T}}\mathbf{A}\mathbf{f} + \mathbf{b}^{\mathrm{T}}\mathbf{f} + c$, and the versions analyzed here differ with respect to the choice of the matrix $\mathbf{A}$ in the quadratic part of this expression. They will be dealt with in Sections 3, 4 and 5 respectively. Since the linear combination $f_{\mathbf{b},c} = \mathbf{b}^{\mathrm{T}}\mathbf{f} + c$ with weights chosen according to Equations (2.8) is MSPE-optimal among all combined forecasts under joint normality of y and $\mathbf{f}$, employment of linear plus quadratic approaches only deserves attention under non-normality. Hence we will assume non-normality in the following.

3 The linear plus quadratic approach with full matrix A

The strong linear plus quadratic approach $f_{\mathbf{A},\mathbf{b},c} = \mathbf{f}^{\mathrm{T}}\mathbf{A}\mathbf{f} + \mathbf{b}^{\mathrm{T}}\mathbf{f} + c$ is based on a full $k \times k$ real symmetric matrix $\mathbf{A}$ to build the quadratic part, a k–dimensional real vector $\mathbf{b}$ as well as a real constant term c.
The expectation of $f_{\mathbf{A},\mathbf{b},c}$ is immediately derived from Lemma A.4 (a). Setting $\tilde{\mathbf{Y}} = \mathbf{f}$, $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\mathbf{f}}$ and $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}_{\mathbf{f}}$ we obtain $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{\mathbf{ff}}$. Setting further $\tilde{\mathbf{A}} = \mathbf{A}$ and $\tilde{\mathbf{a}} = \mathbf{b}$ we arrive at

$$\mathrm{E}(f_{\mathbf{A},\mathbf{b},c}) = \boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}} + \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{ff}}) + \mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + c\ . \tag{3.1}$$

We now want to determine how the combination parameters $\mathbf{A}$, $\mathbf{b}$ and c should be chosen in order to minimize the mean square prediction error of such a combined forecast. To achieve this goal we will perform the following three steps: In the first step we will explicitly calculate the general MSPE-function of a combined forecast $f_{\mathbf{A},\mathbf{b},c}$. In the second step we will differentiate this function with respect to $\mathbf{A}$, $\mathbf{b}$ and c. In the final step we will simultaneously equate these derivatives to zero which results in a linear equation system. The unique solution $(\mathbf{A}_{\mathrm{opt}}, \mathbf{b}_{\mathrm{opt}}, c_{\mathrm{opt}})$ of this equation system yields the desired minimum of the MSPE-function.

Step 1: Explicit calculation of the MSPE**-function.** Since $\mathrm{MSPE}(f_{\mathbf{A},\mathbf{b},c}, y) = \mathrm{E}[(y - f_{\mathbf{A},\mathbf{b},c})^2] = \mathrm{Var}(y - f_{\mathbf{A},\mathbf{b},c}) + [\mathrm{E}(y - f_{\mathbf{A},\mathbf{b},c})]^2$ we may split the necessary calculations in two parts.

While the calculation of $[\mathrm{E}(y - f_{\mathbf{A},\mathbf{b},c})]^2$ is quite easily done with the help of (3.1) and $\mathrm{E}(y) = \mu_0$, the calculation of $\mathrm{Var}(y - f_{\mathbf{A},\mathbf{b},c})$ requires much more effort.
Setting

$$\tilde{\mathbf{Y}} = \begin{pmatrix} y \\ \mathbf{f} \end{pmatrix} , \; \tilde{\boldsymbol{\mu}} = \begin{pmatrix} \mu_0 \\ \boldsymbol{\mu}_{\mathbf{f}} \end{pmatrix} = \boldsymbol{\mu} \text{ and } \tilde{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_0 \\ \boldsymbol{\varepsilon}_{\mathbf{f}} \end{pmatrix} = \boldsymbol{\varepsilon} \tag{3.2}$$

we obtain

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} \, , \; \tilde{\boldsymbol{\Phi}} = \boldsymbol{\Phi} \text{ and } \tilde{\boldsymbol{\Psi}} = \boldsymbol{\Psi} \tag{3.3}$$

as defined in (1.9) and (1.12) – (1.15). Setting further

$$\tilde{\mathbf{A}} = \tilde{\mathbf{B}} = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{0} & -\mathbf{A} \end{pmatrix} \text{ and } \tilde{\mathbf{a}} = \tilde{\mathbf{b}} = \begin{pmatrix} 1 \\ -\mathbf{b} \end{pmatrix} \tag{3.4}$$

we may then apply Lemma A.4 (b).
Joining the two parts of the calculation and performing some simplifications we finally arrive at the following expression for the mean square prediction error of $f_{\mathbf{A},\mathbf{b},c}$, where the terms have been ordered with respect to the occurring unknowns:

$$\begin{aligned}
&\mathrm{MSPE}(f_{\mathbf{A},\mathbf{b},c}, y) = \\
&= \; 4\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{ff}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}} + 4\boldsymbol{\varphi}_{\mathbf{A}}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}} + \mathrm{tr}(\mathbf{A}\boldsymbol{\psi}_{\mathbf{A}}) + (\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}})^2 + 2\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}}\,\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{ff}}) \\
&\quad - 4\boldsymbol{\Sigma}_{\mathbf{f}0}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}} - 2\,\mathrm{tr}(\mathbf{A}\boldsymbol{\Phi}_{0\mathbf{ff}}) - 2\mu_0\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}} - 2\mu_0\,\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{ff}}) \\
&\quad + 4\mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}} + 2\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_{\mathbf{A}} + 2\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}}\mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + 2\,\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{ff}})\mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} \\
&\quad + \mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}\mathbf{b} + \mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{b} \\
&\quad - 2\mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{f}0} - 2\mu_0\mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} \\
&\quad + 2\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}}c + 2\,\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{ff}})c \\
&\quad + 2\mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}}c \\
&\quad + c^2 \\
&\quad - 2\mu_0 c \\
&\quad + \Sigma_{00} + \mu_0^2 \, ,
\end{aligned} \tag{3.5}$$

where

$$\boldsymbol{\varphi}_{\mathbf{A}} = \begin{pmatrix} \mathrm{tr}(\mathbf{A}\boldsymbol{\Phi}_{1\mathbf{ff}}) \\ \vdots \\ \mathrm{tr}(\mathbf{A}\boldsymbol{\Phi}_{k\mathbf{ff}}) \end{pmatrix} \tag{3.6}$$

is a k–dimensional vector and

$$\boldsymbol{\psi}_{\mathbf{A}} = \begin{pmatrix} \mathrm{tr}(\mathbf{A}\boldsymbol{\Psi}_{11\mathbf{ff}}) & \dots & \mathrm{tr}(\mathbf{A}\boldsymbol{\Psi}_{1k\mathbf{ff}}) \\ \vdots & \ddots & \vdots \\ \mathrm{tr}(\mathbf{A}\boldsymbol{\Psi}_{k1\mathbf{ff}}) & \dots & \mathrm{tr}(\mathbf{A}\boldsymbol{\Psi}_{kk\mathbf{ff}}) \end{pmatrix} \tag{3.7}$$

is a symmetric $k \times k$ matrix.

Step 2: Differentiation. Applying common differential calculus we immediately get

$$\frac{\partial \mathrm{MSPE}(f_{\mathbf{A},\mathbf{b},c}, y)}{\partial c} = 2\left[c - \mu_0 + \mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + \boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}} + \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{ff}})\right] \ . \tag{3.8}$$

With the help of Lemma A.7 it is not difficult to show

$$\begin{aligned} \frac{\partial \mathrm{MSPE}(f_{\mathbf{A},\mathbf{b},c}, y)}{\partial \mathbf{b}} &= 2\left[\boldsymbol{\Sigma}_{\mathbf{ff}}\mathbf{b} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{b} - \boldsymbol{\Sigma}_{\mathbf{f}0} - \mu_0\boldsymbol{\mu}_{\mathbf{f}} + 2\boldsymbol{\Sigma}_{\mathbf{ff}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}} \right. \\ &\quad \left. + \boldsymbol{\varphi}_{\mathbf{A}} + \boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}} + \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{ff}})\boldsymbol{\mu}_{\mathbf{f}} + c\boldsymbol{\mu}_{\mathbf{f}}\right] \ . \end{aligned} \tag{3.9}$$

Differentiation with respect to $\mathbf{A}$ is the hard part of this second step. Since $\mathbf{A}$ is symmetric we have to apply Lemma A.9. Furthermore, Lemma A.8 has to be applied several times and also Lemmas A.2 and A.3 are of value at some stages. We finally arrive at

$$\begin{aligned} &\frac{\partial \mathrm{MSPE}(f_{\mathbf{A},\mathbf{b},c}, y)}{\partial \mathbf{A}} = \\ &= \ 2[\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}(4\boldsymbol{\Sigma}_{\mathbf{ff}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{1}}^{\mathrm{T}}) + (4\boldsymbol{\Sigma}_{\mathbf{ff}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{1}}^{\mathrm{T}})\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}} \\ &\qquad - \mathrm{diag}(\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}(4\boldsymbol{\Sigma}_{\mathbf{ff}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{1}}^{\mathrm{T}}))] \\ &\quad + \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{ff}})[4\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}} - 2\,\mathrm{diag}(\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{1}}^{\mathrm{T}})] \\ &\quad + \mathrm{tr}(\mathbf{A}\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{1}}^{\mathrm{T}})[4\boldsymbol{\Sigma}_{\mathbf{ff}} - 2\,\mathrm{diag}(\boldsymbol{\Sigma}_{\mathbf{ff}})] \\ &\quad + 4[\boldsymbol{\varphi}_{\mathbf{A}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\varphi}_{\mathbf{A}}^{\mathrm{T}} - \mathrm{diag}(\boldsymbol{\varphi}_{\mathbf{A}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}})] \\ &\quad + 4\boldsymbol{\psi}_{\mathbf{A}} - 2\,\mathrm{diag}(\boldsymbol{\psi}_{\mathbf{A}}) \\ &\quad + \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}\mu_j[8\boldsymbol{\Phi}_{i\mathbf{ff}} - 4\,\mathrm{diag}(\boldsymbol{\Phi}_{i\mathbf{ff}})] \\ &\quad + 4[\boldsymbol{\Sigma}_{\mathbf{ff}}\mathbf{b}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}} + \boldsymbol{\mu}_{\mathbf{f}}\mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}} - \mathrm{diag}(\boldsymbol{\Sigma}_{\mathbf{ff}}\mathbf{b}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}})] \\ &\quad + (\mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + c - \mu_0)[4(\boldsymbol{\Sigma}_{\mathbf{ff}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{1}}^{\mathrm{T}}) - 2\,\mathrm{diag}(\boldsymbol{\Sigma}_{\mathbf{ff}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}})] \\ &\quad + \sum_{i=1}^{k} b_i[4\boldsymbol{\Phi}_{i\mathbf{ff}} - 2\,\mathrm{diag}(\boldsymbol{\Phi}_{i\mathbf{ff}})] \\ &\quad - 4[\boldsymbol{\Sigma}_{\mathbf{f}0}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\Sigma}_{\mathbf{f}0}^{\mathrm{T}} - \mathrm{diag}(\boldsymbol{\Sigma}_{\mathbf{f}0}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}})] \\ &\quad - 4\boldsymbol{\Phi}_{0\mathbf{ff}} + 2\,\mathrm{diag}(\boldsymbol{\Phi}_{0\mathbf{ff}}) \ , \end{aligned} \tag{3.10}$$

where for a $k \times k$–matrix $\mathbf{M} = (m_{ij})$ we define

$$\operatorname{diag}(\mathbf{M}) = \begin{pmatrix} m_{11} & 0 & \dots & 0 \\ 0 & m_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & m_{kk} \end{pmatrix} \in \mathbb{R}^{k \times k} . \tag{3.11}$$

Step 3: Equating to zero. Setting Equations (3.8), (3.9) and (3.10) simultaneously to zero and solving the resulting linear equation system for the unknown parameters we obtain the optimal choices for $\mathbf{A}$, $\mathbf{b}$ and c.
From Equation (3.8) we get

$$c_{\text{opt}} = \mu_0 - \mathbf{b}_{\text{opt}}^{\text{T}} \boldsymbol{\mu}_{\mathbf{f}} - \boldsymbol{\mu}_{\mathbf{f}}^{\text{T}} \mathbf{A}_{\text{opt}} \boldsymbol{\mu}_{\mathbf{f}} - \operatorname{tr}(\mathbf{A}_{\text{opt}} \boldsymbol{\Sigma}_{\mathbf{ff}}) . \tag{3.12}$$

Using (3.12) we obtain from (3.9)

$$\mathbf{b}_{\text{opt}} = \boldsymbol{\Sigma}_{\mathbf{ff}}^{-1} (\boldsymbol{\Sigma}_{\mathbf{f}0} - \boldsymbol{\varphi}_{\mathbf{A}_{\text{opt}}}) - 2\mathbf{A}_{\text{opt}} \boldsymbol{\mu}_{\mathbf{f}} . \tag{3.13}$$

Using (3.12) and (3.13) Equation (3.10) is equivalent to

$$\begin{aligned} & 4\boldsymbol{\psi}_{\mathbf{A}_{\text{opt}}} - 2\operatorname{diag}(\boldsymbol{\psi}_{\mathbf{A}_{\text{opt}}}) - 4\boldsymbol{\Phi}_{0\mathbf{ff}} + 2\operatorname{diag}(\boldsymbol{\Phi}_{0\mathbf{ff}}) \\ + & \sum_{i=1}^{k} \boldsymbol{\xi}_i^{(k)\text{T}} \boldsymbol{\Sigma}_{\mathbf{ff}}^{-1} (\boldsymbol{\Sigma}_{\mathbf{f}0} - \boldsymbol{\varphi}_{\mathbf{A}_{\text{opt}}}) [4\boldsymbol{\Phi}_{i\mathbf{ff}} - 2\operatorname{diag}(\boldsymbol{\Phi}_{i\mathbf{ff}})] \\ & - 4\operatorname{tr}(\mathbf{A}_{\text{opt}} \boldsymbol{\Sigma}_{\mathbf{ff}}) \boldsymbol{\Sigma}_{\mathbf{ff}} + 2\operatorname{tr}(\mathbf{A}_{\text{opt}} \boldsymbol{\Sigma}_{\mathbf{ff}}) \operatorname{diag}(\boldsymbol{\Sigma}_{\mathbf{ff}}) \quad = \quad \mathbf{0} . \end{aligned} \tag{3.14}$$

Here $\boldsymbol{\xi}_i^{(k)}$ denote the k-dimensional unit vectors, i.e. the i-th component of $\boldsymbol{\xi}_i^{(k)}$ is equal to 1 whereas the other components are equal to 0.
Equation (3.14) represents a linear equation system with the unknowns being the $k(k+1)/2$ different elements of the symmetric matrix $\mathbf{A}_{\text{opt}}$. Unfortunately, we cannot write down its solution explicitly, and hence we cannot give the optimal combination parameters $(\mathbf{A}_{\text{opt}}, \mathbf{b}_{\text{opt}}, c_{\text{opt}})$ in an explicit form. In practical applications we have to solve Equation (3.14) in order to obtain $\mathbf{A}_{\text{opt}}$, then insert this result into Equation (3.13) and thus get $\mathbf{b}_{\text{opt}}$ and finally insert these two results into Equation (3.12) to obtain c_{opt}.
Provided that Equations (3.14), (3.13) and (3.12) have a common unique solution $(\mathbf{A}_{\text{opt}}, \mathbf{b}_{\text{opt}}, c_{\text{opt}})$, it can be seen that this solution describes a minimum of the MSPE-

function within the considered class of combined forecasts:

$$
\begin{aligned}
\mathrm{MSPE}(f_{\mathbf{A},\mathbf{b},c}, y) &= \mathrm{E}[(y - f_{\mathbf{A},\mathbf{b},c})^2] \\
&= \mathrm{E}[(y - \mathbf{f}^{\mathrm{T}}\mathbf{A}\mathbf{f} - \mathbf{b}^{\mathrm{T}}\mathbf{f} - c)^2] \\
&= \mathrm{E}\left[\left(y - \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} f_i f_j - \sum_{i=1}^{k} b_i f_i - c\right)^2\right] \\
&= \mathrm{E}\left[\left(y - \sum_{i=1}^{k} a_{ii} f_i^2 - 2\sum_{i}\sum_{<\,j} a_{ij} f_i f_j - \sum_{i=1}^{k} b_i f_i - c\right)^2\right]
\end{aligned} \tag{3.15}
$$

is a quadratic function in the unknown parameters bounded below by the value 0 (compare Lemma A.2).

Since we cannot express the optimal combination parameters $\mathbf{A}_{\mathrm{opt}}$, $\mathbf{b}_{\mathrm{opt}}$ and c_{opt} with the help of explicit formulae, we cannot give an explicit expression for the optimal value $\mathrm{MSPE}(f_{\mathbf{A}_{\mathrm{opt}},\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}, y)$ of the MSPE-function either.

We can conclude, however, that $f_{\mathbf{A}_{\mathrm{opt}},\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}$ is an unbiased forecast: Following Equation (3.1) the expectation of $f_{\mathbf{A}_{\mathrm{opt}},\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}$ is given by

$$
\mathrm{E}(f_{\mathbf{A}_{\mathrm{opt}},\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}) = \boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}_{\mathrm{opt}}\boldsymbol{\mu}_{\mathbf{f}} + \mathrm{tr}(\mathbf{A}_{\mathrm{opt}}\boldsymbol{\Sigma}_{\mathbf{ff}}) + \mathbf{b}_{\mathrm{opt}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + c_{\mathrm{opt}}\ . \tag{3.16}
$$

Then unbiasedness is guaranteed by the optimal choice of the constant term as can be seen by inserting

$$
c_{\mathrm{opt}} = \mu_0 - \mathbf{b}_{\mathrm{opt}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} - \boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{A}_{\mathrm{opt}}\boldsymbol{\mu}_{\mathbf{f}} - \mathrm{tr}(\mathbf{A}_{\mathrm{opt}}\boldsymbol{\Sigma}_{\mathbf{ff}}) \tag{3.17}
$$

into Equation (3.16).

For a simple example see Section 6 where the combination of $k = 2$ forecasts is performed.

The fact that the optimal combination parameters $\mathbf{A}_{\mathrm{opt}}$, $\mathbf{b}_{\mathrm{opt}}$ and c_{opt} are not given by explicit formulae, but have to be calculated from Equations (3.12), (3.13) and (3.14) not only hinders further theoretical considerations but also impedes the application of the strong linear plus quadratic combination technique: We can only deal with these numbers of individual forecasts k for which we have made the linear equation system (3.14) explicit. From Section 6 dealing with $k = 2$ it becomes clear that this may be a cumbersome task.

4 The linear plus quadratic approach with diagonal matrix A

The medium linear plus quadratic approach $f_{\mathbf{a},\mathbf{b},c} = \sum_{i=1}^{k} a_i f_i^2 + \mathbf{b}^{\mathrm{T}}\mathbf{f} + c$ emerges from restricting the full matrix $\mathbf{A}$ in the strong approach to a diagonal matrix $\mathrm{dg}(\mathbf{a})$, $\mathbf{a} = (a_1, \ldots, a_k)^{\mathrm{T}} \in \mathbb{R}^k$.
Inserting $\mathbf{A} = \mathrm{dg}(\mathbf{a})$ in Equation (3.1) we obtain the expectation of $f_{\mathbf{a},\mathbf{b},c}$

$$\mathrm{E}(f_{\mathbf{a},\mathbf{b},c}) = \sum_{i=1}^{k} a_i \mu_i^2 + \sum_{i=1}^{k} a_i \Sigma_{ii} + \mathbf{b}^{\mathrm{T}} \boldsymbol{\mu}_{\mathbf{f}} + c \ . \tag{4.1}$$

Unfortunately, the MSPE-optimal choices for the combination parameters $\mathbf{a}$, $\mathbf{b}$ and c cannot be derived directly from the results of the preceding section. Instead, we have to perform the same three steps as before heeding the additional restrictions imposed on the matrix in the quadratic part.
Along the same lines as in Section 3 we obtain the following equations determining the optimal choices for $\mathbf{a}$, $\mathbf{b}$ and c (compare Section C in the appendix):

$$c_{\mathrm{opt}} = \mu_0 - \mathbf{b}_{\mathrm{opt}}^{\mathrm{T}} \boldsymbol{\mu}_{\mathbf{f}} - \boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}} \,\mathrm{dg}(\mathbf{a}_{\mathrm{opt}}) \boldsymbol{\mu}_{\mathbf{f}} - \mathrm{tr}(\mathrm{dg}(\mathbf{a}_{\mathrm{opt}}) \boldsymbol{\Sigma}_{\mathbf{ff}}) \ , \tag{4.2}$$

$$\mathbf{b}_{\mathrm{opt}} = \boldsymbol{\Sigma}_{\mathbf{ff}}^{-1} (\boldsymbol{\Sigma}_{\mathbf{f}0} - \boldsymbol{\varphi}_{\mathbf{a}_{\mathrm{opt}}}) - 2\,\mathrm{dg}(\mathbf{a}_{\mathrm{opt}}) \boldsymbol{\mu}_{\mathbf{f}} \tag{4.3}$$

and

$$\begin{aligned} \sum_{i=1}^{k} \sum_{l=1}^{k} a_{\mathrm{opt},l} \Psi_{llii} \boldsymbol{\xi}_i^{(k)} - \sum_{i=1}^{k} \Phi_{0ii} \boldsymbol{\xi}_i^{(k)} - \mathrm{tr}(\mathrm{dg}(\mathbf{a}_{\mathrm{opt}}) \boldsymbol{\Sigma}_{\mathbf{ff}}) \sum_{i=1}^{k} \Sigma_{ii} \boldsymbol{\xi}_i^{(k)} & \\ + \sum_{i=1}^{k} \sum_{l=1}^{k} \left(\boldsymbol{\xi}_l^{(k)\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{ff}}^{-1} \left(\boldsymbol{\Sigma}_{\mathbf{f}0} - \boldsymbol{\varphi}_{\mathbf{a}_{\mathrm{opt}}} \right) \right) \Phi_{lii} \boldsymbol{\xi}_i^{(k)} & = \mathbf{0} \ , \end{aligned} \tag{4.4}$$

where $\boldsymbol{\xi}_i^{(k)}$ denotes the i-th k-dimensional unit vector.
Equation (4.4) is a linear equation system with the unknowns being the k components of the vector $\mathbf{a}_{\mathrm{opt}}$. Again, we cannot write down its solution explicitly, and hence we cannot give the optimal combination parameters $(\mathbf{a}_{\mathrm{opt}}, \mathbf{b}_{\mathrm{opt}}, c_{\mathrm{opt}})$ in an explicit form. In practice we have to proceed by solving Equation (4.4) in order to obtain $\mathbf{a}_{\mathrm{opt}}$. Then this result is inserted into Equation (4.3) and thus $\mathbf{b}_{\mathrm{opt}}$ is obtained. Finally these two results are inserted into Equation (4.2) and we get c_{opt}.
By the same reasoning as at the end of the previous section we may conclude that the unique solution $(\mathbf{a}_{\mathrm{opt}}, \mathbf{b}_{\mathrm{opt}}, c_{\mathrm{opt}})$ of Equations (4.4), (4.3) and (4.2) leads

to the minimum value of the MSPE-function within the considered class of combined forecasts. Due to the lack of an explicit expression for the optimal combination parameters, again we cannot give an explicit expression for the optimal value $\mathrm{MSPE}(f_{\mathbf{a}_{\mathrm{opt}},\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}, y)$. Just like above we may, however, conclude that $f_{\mathbf{a}_{\mathrm{opt}},\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}$ is an unbiased forecast.

In Section 6 the combination of $k = 2$ forecasts using the medium linear plus quadratic approach is considered as well.

5 The linear plus quadratic approach with A chosen as a scalar multiple of the identity matrix

In the weak linear plus quadratic approach $f_{\alpha,\mathbf{b},c} = \alpha\mathbf{f}^{\mathrm{T}}\mathbf{f} + \mathbf{b}^{\mathrm{T}}\mathbf{f} + c$ the full matrix $\mathbf{A}$ from the strong approach is restricted to $\alpha\mathbf{I}$, a real scalar multiple of the $k \times k$ identity matrix.

It should be pointed out again, that the weak linear plus quadratic combination increases the number of combination parameters by only one with respect to the best linear combination, but it involves $k-1$ parameters less than the medium and even $k(k+1)/2-1$ parameters less than the strong linear plus quadratic combination. Consequently, it may be practical in empirical applications where the number of data available for parameter estimation is not large.

Inserting $\mathbf{A} = \alpha\mathbf{I}$ in Equation (3.1) we obtain the expectation of $f_{\alpha,\mathbf{b},c}$

$$\mathrm{E}(f_{\alpha,\mathbf{b},c}) = \alpha(\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + \mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{ff}})) + \mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + c\ . \tag{5.1}$$

Like in the two sections before the MSPE-optimal choices for the combination parameters α, $\mathbf{b}$ and c have to be determined in three steps (compare Appendix C). Unlike the two sections before we are now able to express these optimal parameters explicitly:

$$\alpha_{\mathrm{opt}} = \frac{\mathrm{tr}(\boldsymbol{\Phi}_{0\mathbf{ff}}) - \boldsymbol{\Sigma}_{\mathbf{f}0}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\varphi}}{\mathrm{tr}(\boldsymbol{\psi}) - \boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\varphi} - [\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{ff}})]^2}\ , \tag{5.2}$$

$$\mathbf{b}_{\mathrm{opt}} = \boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}0} - \alpha_{\mathrm{opt}}(\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\varphi} + 2\boldsymbol{\mu}_{\mathbf{f}}) \tag{5.3}$$

and

$$c_{\mathrm{opt}} = \mu_0 - \mathbf{b}_{\mathrm{opt}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} - \alpha_{\mathrm{opt}}(\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + \mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{ff}}))\ . \tag{5.4}$$

By the same reasoning as in Section 3 we may conclude that the unique solution $(\alpha_{\mathrm{opt}}, \mathbf{b}_{\mathrm{opt}}, c_{\mathrm{opt}})$ given above leads to the minimum value of the MSPE-function within the considered class of combined forecasts. Inserting $(\alpha_{\text{opt}}, \mathbf{b}_{\text{opt}}, c_{\text{opt}})$ into the general function $\text{MSPE}(f_{\alpha,\mathbf{b},c}, y)$ (Equation (C.7) from Appendix C) we may derive that this optimal value is given by

$$\text{MSPE}(f_{\alpha_{\text{opt}},\mathbf{b}_{\text{opt}},c_{\text{opt}}}, y) = \Sigma_{00} - \mathbf{\Sigma}_{\mathbf{f}0}^{\text{T}}\mathbf{\Sigma}_{\mathbf{ff}}^{-1}\mathbf{\Sigma}_{\mathbf{f}0} - \frac{\left(\text{tr}(\mathbf{\Phi}_{0\mathbf{ff}}) - \mathbf{\Sigma}_{\mathbf{f}0}^{\text{T}}\mathbf{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\varphi}\right)^2}{\text{tr}(\boldsymbol{\psi}) - \boldsymbol{\varphi}^{\text{T}}\mathbf{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\varphi} - [\text{tr}(\mathbf{\Sigma}_{\mathbf{ff}})]^2}\,. \tag{5.5}$$

From comparing this equation to Equation (2.9) we may conclude that employing the optimal weak linear plus quadratic combined forecast instead of the optimal linear combined forecast leads to a gain of

$$\frac{\left(\text{tr}(\mathbf{\Phi}_{0\mathbf{ff}}) - \mathbf{\Sigma}_{\mathbf{f}0}^{\text{T}}\mathbf{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\varphi}\right)^2}{\text{tr}(\boldsymbol{\psi}) - \boldsymbol{\varphi}^{\text{T}}\mathbf{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\varphi} - [\text{tr}(\mathbf{\Sigma}_{\mathbf{ff}})]^2} \tag{5.6}$$

with respect to the MSPE-criterion.
Again the optimal choice c_{opt} for the constant term guarantees unbiasedness of the combined forecast $f_{\alpha_{\text{opt}},\mathbf{b}_{\text{opt}},c_{\text{opt}}}$.
The following section deals with the combination of $k = 2$ forecasts via this and the other linear plus quadratic approaches.

6 Combination of $k = 2$ forecasts

In order to see explicitly how the single forecasts are combined using the linear plus quadratic approaches and in order to give a clearer impression of the nature of the equation systems arising in the three previous chapters, we will now consider the simple case of combining $k = 2$ forecasts f_1 and f_2 for the target variable y. Consequently, the strong, medium and weak versions depend on 6, 5 and 4 unknown parameters, respectively.
In any of the linear plus quadratic approaches the difficult equation is the one determining the parameters of the quadratic part. Having solved this equation it is an easy task to derive the parameters $\mathbf{b}$ and c of the respective linear parts. Hence, in this section we will concentrate on making the equations for the respective quadratic parts explicit.
In the situation of $k = 2$ forecasts Equation (3.14) for the determination of the optimal full parameter matrix $\mathbf{A}_{\text{opt}}$ from the strong linear plus quadratic approach is equivalent to the linear equation system $\mathbf{Tx} + \mathbf{s} = \mathbf{0}$. Here the vector $\mathbf{x} = (a_{\text{opt},11}, a_{\text{opt},12}, a_{\text{opt},22})^{\text{T}}$ consists of the unknown different entries in the symmet-

ric matrix $\mathbf{A}_{\rm opt} \in \mathbb{R}^{2\times 2}$, $\mathbf{T}$ is a 3×3 matrix with elements

$$\begin{aligned}
t_{11} &= \Psi_{1111} + \frac{1}{d}[\Phi_{111}(-\Sigma_{22}\Phi_{111} + \Sigma_{12}\Phi_{211}) + \Phi_{211}(-\Sigma_{11}\Phi_{211} + \Sigma_{12}\Phi_{111})] \\
&\quad - \Sigma_{11}^2 \\
t_{12} &= 2(\Psi_{1112} + \frac{1}{d}[\Phi_{111}(-\Sigma_{22}\Phi_{112} + \Sigma_{12}\Phi_{212}) + \Phi_{211}(-\Sigma_{11}\Phi_{212} + \Sigma_{12}\Phi_{112})] \\
&\quad - \Sigma_{11}\Sigma_{12}) \\
t_{13} &= \Psi_{1122} + \frac{1}{d}[\Phi_{111}(-\Sigma_{22}\Phi_{122} + \Sigma_{12}\Phi_{222}) + \Phi_{211}(-\Sigma_{11}\Phi_{222} + \Sigma_{12}\Phi_{122})] \\
&\quad - \Sigma_{11}\Sigma_{22} \\
t_{21} &= \Psi_{1211} + \frac{1}{d}[\Phi_{112}(-\Sigma_{22}\Phi_{111} + \Sigma_{12}\Phi_{211}) + \Phi_{212}(-\Sigma_{11}\Phi_{211} + \Sigma_{12}\Phi_{111})] \\
&\quad - \Sigma_{11}\Sigma_{12} \\
t_{22} &= 2(\Psi_{1212} + \frac{1}{d}[\Phi_{112}(-\Sigma_{22}\Phi_{112} + \Sigma_{12}\Phi_{212}) + \Phi_{212}(-\Sigma_{11}\Phi_{212} + \Sigma_{12}\Phi_{112})] \\
&\quad - \Sigma_{12}^2) \\
t_{23} &= \Psi_{1222} + \frac{1}{d}[\Phi_{112}(-\Sigma_{22}\Phi_{122} + \Sigma_{12}\Phi_{222}) + \Phi_{212}(-\Sigma_{11}\Phi_{222} + \Sigma_{12}\Phi_{122})] \\
&\quad - \Sigma_{12}\Sigma_{22} \\
t_{31} &= \Psi_{2211} + \frac{1}{d}[\Phi_{122}(-\Sigma_{22}\Phi_{111} + \Sigma_{12}\Phi_{211}) + \Phi_{222}(-\Sigma_{11}\Phi_{211} + \Sigma_{12}\Phi_{111})] \\
&\quad - \Sigma_{11}\Sigma_{22} \\
t_{32} &= 2(\Psi_{2212} + \frac{1}{d}[\Phi_{122}(-\Sigma_{22}\Phi_{112} + \Sigma_{12}\Phi_{212}) + \Phi_{222}(-\Sigma_{11}\Phi_{212} + \Sigma_{12}\Phi_{112})] \\
&\quad - \Sigma_{12}\Sigma_{22}) \\
t_{33} &= \Psi_{2222} + \frac{1}{d}[\Phi_{122}(-\Sigma_{22}\Phi_{122} + \Sigma_{12}\Phi_{222}) + \Phi_{222}(-\Sigma_{11}\Phi_{222} + \Sigma_{12}\Phi_{122})] \\
&\quad - \Sigma_{22}^2
\end{aligned} \tag{6.1}$$

and $\mathbf{s}$ is a 3–dimensional vector with components

$$\begin{aligned}
s_1 &= -\Phi_{011} + \frac{1}{d}[\Phi_{111}(\Sigma_{22}\Sigma_{10} - \Sigma_{12}\Sigma_{20}) + \Phi_{211}(\Sigma_{11}\Sigma_{20} - \Sigma_{12}\Sigma_{10})] \\
s_2 &= -\Phi_{012} + \frac{1}{d}[\Phi_{112}(\Sigma_{22}\Sigma_{10} - \Sigma_{12}\Sigma_{20}) + \Phi_{212}(\Sigma_{11}\Sigma_{20} - \Sigma_{12}\Sigma_{10})] \\
s_3 &= -\Phi_{022} + \frac{1}{d}[\Phi_{122}(\Sigma_{22}\Sigma_{10} - \Sigma_{12}\Sigma_{20}) + \Phi_{222}(\Sigma_{11}\Sigma_{20} - \Sigma_{12}\Sigma_{10})] \,.
\end{aligned} \tag{6.2}$$

The scalar d stands for

$$d = \det(\mathbf{\Sigma_{ff}}) = \Sigma_{11}\Sigma_{22} - \Sigma_{12}^2 \,. \tag{6.3}$$

Equation (4.4) for the determination of the optimal parameter vector $\mathbf{a}_{\rm opt} = (a_{{\rm opt},1}, a_{{\rm opt},2})^{\rm T}$ from the medium linear plus quadratic approach is equivalent to the linear equation system

$$\begin{pmatrix} t_{11} & t_{13} \\ t_{31} & t_{33} \end{pmatrix} \begin{pmatrix} a_{\text{opt},1} \\ a_{\text{opt},2} \end{pmatrix} + \begin{pmatrix} s_1 \\ s_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} , \tag{6.4}$$

i.e. the elements of the system matrix of the linear equation system in this approach are identical with the four corner elements of the matrix $\mathbf{T}$ in the approach with full parameter matrix $\mathbf{A}$. Likewise the elements of the vector in this approach are equal to the top and bottom elements of the vector $\mathbf{s}$ in the strong approach.
Equation (C.15) for the determination of the optimal parameter α_{opt} from the weak linear plus quadratic approach is equivalent to the linear equation

$$(t_{11}+t_{13}+t_{31}+t_{33})\alpha_{\text{opt}} + (s_1+s_3) = 0 , \tag{6.5}$$

i.e. the ingredients of this equation are the same as in the medium approach.
If we take a closer look at the elements t_{ij} and s_i involved in each of the linear plus quadratic approaches, it is evident that all approaches depend on moments up to order 4. The strong approach, however, needs fourth order moments which are not used in the medium and weak approaches. The difference between the medium and the weak approach in this respect is that the medium approach utilizes three different fourth order moments individually, while the weak approach only utilizes the weighted sum of the same three quantities. Thus we can say that each version needs a different level of knowledge about the moments of the joint distribution of y and $\mathbf{f}$.
After considering the special case $k = 2$ we will now turn to the special case of $k = 1$ forecast.

7 The special case $k = 1$: Adjustment of forecasts

There is no reason why the special case $k = 1$ should be ruled out in the above considerations. Of course, this "combination of one forecast" should rather be addressed as *adjustment of single forecasts*. Exploiting the moment structure of the joint distribution of the target variable y and a single forecast f_i the performance of f_i can be improved with respect to the mean square prediction error by this kind of adjustment.
The MSPE of the forecast f_i is given by

$$\begin{aligned} \text{MSPE}(f_i, y) &= \text{E}[(y - f_i)^2] \\ &= \text{Var}(y - f_i) + [\text{E}(y - f_i)]^2 \\ &= \Sigma_{00} + \Sigma_{ii} - 2\Sigma_{i0} + \mu_0^2 + \mu_i^2 - 2\mu_0\mu_i . \end{aligned} \tag{7.1}$$

All of the linear and linear plus quadratic combination approaches described above may be employed in this case. Some of them, however, are identical to others, as we will see in the following.
For instance all three linear plus quadratic combined forecasts coincide in the current situation, i.e. we only need to consider one *linear plus quadratic adjustment*

$$(f_i)_{\alpha,b,c} = \alpha f_i^2 + b f_i + c \tag{7.2}$$

with $\alpha, b, c \in \mathbb{R}$. As a special case of Equations (5.2), (5.3) and (5.4) the optimal choices for the unknown parameters may be derived as

$$\alpha_{\text{opt}} = \frac{\Phi_{0ii} - \Sigma_{i0}\Sigma_{ii}^{-1}\Phi_{iii}}{\Psi_{iiii} - \Phi_{iii}^2\Sigma_{ii}^{-1} - \Sigma_{ii}^2}\ , \tag{7.3}$$

$$b_{\text{opt}} = \frac{\Sigma_{i0}}{\Sigma_{ii}} - \alpha_{\text{opt}}\left(\frac{\Phi_{iii}}{\Sigma_{ii}} + 2\mu_i\right) \quad \text{and} \tag{7.4}$$

$$c_{\text{opt}} = \mu_0 - b_{\text{opt}}\mu_i - \alpha_{\text{opt}}\left(\mu_i^2 + \Sigma_{ii}\right) \tag{7.5}$$

leading to the MSPE-value of the optimal linear plus quadratic adjusted forecast

$$\text{MSPE}((f_i)_{\alpha_{\text{opt}},b_{\text{opt}},c_{\text{opt}}}, y) = \Sigma_{00} - \frac{\Sigma_{i0}^2}{\Sigma_{ii}} - \frac{\left(\Phi_{0ii} - \Sigma_{i0}\Sigma_{ii}^{-1}\Phi_{iii}\right)^2}{\Psi_{iiii} - \Phi_{iii}^2\Sigma_{ii}^{-1} - \Sigma_{ii}^2}\ . \tag{7.6}$$

The *unrestricted linear adjustment with constant term* is

$$(f_i)_{b,c} = b f_i + c \tag{7.7}$$

with $b, c \in \mathbb{R}$. Granger (1989, p. 169) points out the usefulness of such an adjustment. The optimal choices for the parameters are obtained as special cases of Equations (2.8), namely

$$b_{\text{opt}} = \frac{\Sigma_{i0}}{\Sigma_{ii}} \quad \text{and} \quad c_{\text{opt}} = \mu_0 - \frac{\Sigma_{i0}}{\Sigma_{ii}}\mu_i \tag{7.8}$$

with corresponding optimal MSPE-value

$$\text{MSPE}((f_i)_{b_{\text{opt}},c_{\text{opt}}}, y) = \Sigma_{00} - \frac{\Sigma_{i0}^2}{\Sigma_{ii}}\ . \tag{7.9}$$

The *unrestricted linear adjustment without constant term* reads

$$(f_i)_b = b f_i \tag{7.10}$$

with $b \in \mathbb{R}$. According to Equation (2.12) the optimal choice for b is given by

$$b_{\text{opt}} = \frac{\Sigma_{i0} + \mu_0\mu_i}{\Sigma_{ii} + \mu_i^2} \tag{7.11}$$

which gives the optimal MSPE-value

$$\begin{aligned} \text{MSPE}((f_i)_{b_{\text{opt}}}, y) &= \Sigma_{00} + \mu_0^2 - \frac{(\Sigma_{i0} + \mu_0\mu_i)^2}{\Sigma_{ii} + \mu_i^2} \\ &= \Sigma_{00} - \frac{\Sigma_{i0}^2}{\Sigma_{ii}} + \frac{\left(\mu_0 - \mu_i\Sigma_{i0}\Sigma_{ii}^{-1}\right)^2}{1 + \mu_i^2\Sigma_{ii}^{-1}} \, . \end{aligned} \tag{7.12}$$

The *linear adjustment with constant term and with the restriction of the weights summing up to unity* is

$$(f_i)_{1,c} = f_i + c \, . \tag{7.13}$$

According to Equation (2.21) the optimal choice for $c \in \mathbb{R}$ is given by

$$c_{\text{opt}} = \mu_0 - \mu_i \tag{7.14}$$

thus resulting in the well known bias corrected forecast. The corresponding optimal MSPE-value is

$$\text{MSPE}((f_i)_{1,c_{\text{opt}}}, y) = \Sigma_{00} - \frac{\Sigma_{i0}^2}{\Sigma_{ii}} + \Sigma_{ii}(1 - \Sigma_{i0}\Sigma_{ii}^{-1})^2 \, . \tag{7.15}$$

Finally, the *linear adjustment without constant term and with the restriction of the weights summing up to unity* as well as the adjustment counterpart of the *arithmetic mean* equal the original single forecast f_i and need no special consideration.
Following the results in Section 2 each of the adjusted forecasts with a constant term c is unbiased.

8 Translations and scale transformations

It is an important question in which way the linear plus quadratic combinations of forecasts are affected by transformations of origin and scale, i.e. in how far the results depend on the chosen coordinate system. We can ask which of the optimal weights or MSPE-values change under translations or scale transformations and, if so, how they do change.
Due to the lack of explicit formulae for the combination parameters within the strong and medium linear plus quadratic approaches we cannot prove all of the

facts stated below for these approaches as we can do for all the other forecasts. Regarding the similar nature of the weak linear plus quadratic approach, however, it may be supposed that the facts are valid for the medium and strong versions as well. This has also been confirmed by all numerical investigations so far.
Let us first consider translations of the data. By this we mean that we add a constant τ to the target variable y as well as to each single forecast f_i, i.e. after the translation we obtain the new variables

$$\begin{pmatrix} \tilde{y} \\ \tilde{\mathbf{f}} \end{pmatrix} = \begin{pmatrix} y \\ \mathbf{f} \end{pmatrix} + \tau \mathbf{1}_{k+1} \ . \tag{8.1}$$

The expectation vector $\tilde{\boldsymbol{\mu}}$ and the centered moment matrices $\tilde{\boldsymbol{\Sigma}}$, $\tilde{\boldsymbol{\Phi}}$ and $\tilde{\boldsymbol{\Psi}}$ of the transformed variables $(\tilde{y}, \tilde{\mathbf{f}}^{\mathrm{T}})^{\mathrm{T}}$ relate to the corresponding quantities $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ of the original variables $(y, \mathbf{f}^{\mathrm{T}})^{\mathrm{T}}$ as follows:

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} + \tau \mathbf{1}_{k+1} \ , \tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} \ , \tilde{\boldsymbol{\Phi}} = \boldsymbol{\Phi} \quad \text{and} \quad \tilde{\boldsymbol{\Psi}} = \boldsymbol{\Psi} \ . \tag{8.2}$$

Consequently, also the quantities $\boldsymbol{\varphi}_{\mathbf{A}}$, $\boldsymbol{\varphi}_{\mathbf{a}}$, $\boldsymbol{\varphi}$, $\boldsymbol{\psi}_{\mathbf{A}}$, $\boldsymbol{\psi}_{\mathbf{a}}$ and $\boldsymbol{\psi}$ are not affected by such a translation. The same is true for the second order moment matrix $\breve{\boldsymbol{\Sigma}}$ which is calculated differently because of the assumption of unbiasedness of each single forecast.
Consulting the equations determining the optimal parameters and the corresponding MSPE-values from the respective sections above we can derive the following facts:
For the linear plus quadratic combinations or adjustments $f_{\mathbf{A}_{\mathrm{opt}},\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}$, $f_{\mathbf{a}_{\mathrm{opt}},\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}$, $f_{\alpha_{\mathrm{opt}},\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}$ and $(f_i)_{\alpha_{\mathrm{opt}},b_{\mathrm{opt}},c_{\mathrm{opt}}}$ the optimal parameter matrix, vector or scalar ($\mathbf{A}_{\mathrm{opt}}$, $\mathbf{a}_{\mathrm{opt}}$ or α_{opt}) corresponding to the quadratic part remains unchanged by the translation of the data, whereas the parameter vector or scalar ($\mathbf{b}_{\mathrm{opt}}$ or b_{opt}) corresponding to the linear part as well as the constant term (c_{opt}) are affected by that translation. This amounts to the effect that the combined or adjusted forecast is translated by the quantity τ as well and, consequently, the MSPE-value is invariant with respect to the translation of the data.
For the linear unrestricted combined or adjusted forecast with constant term $f_{\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}}}$ or $(f_i)_{b_{\mathrm{opt}},c_{\mathrm{opt}}}$ only the constant term c_{opt} is affected by the translation, while for the linear restricted combination with or without constant term $f_{\mathbf{b}_{\mathrm{opt}},c_{\mathrm{opt}},\mathrm{rest}}$ or $f_{\mathbf{b}_{\mathrm{opt}},\mathrm{rest}}$ as well as for the bias corrected forecast $(f_i)_{1,c_{\mathrm{opt}}}$ no combination parameter is changed. In any case the adjusted or combined forecast is also translated by τ such that the MSPE-value is not changed by a translation of the data. The latter is also true for the single forecasts f_i and their arithmetic mean f_{am}.
Only the linear unrestricted combined or adjusted forecast without constant term $f_{\mathbf{b}_{\mathrm{opt}}}$ or $(f_i)_{b_{\mathrm{opt}}}$ exhibits an undesired behaviour under a translation of the data.

The parameter vector or scalar ($\mathbf{b}_{\mathrm{opt}}$ or b_{opt}) is changed in such a way that the combined or adjusted forecast is *not* translated by τ with the consequence that the MSPE-value is changed as well.
The MSPE-values of all the adjustments and combinations involving a constant term c_{opt} are not only invariant with respect to a translation of the data by a constant τ, but they are not even affected by *any* change of the vector $\boldsymbol{\mu} = (\mu_0, \boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}})^{\mathrm{T}}$.

Let us now turn to scale transformations of the data. By this we mean that target variable y as well as each single forecast f_i are multiplied by the same constant λ, i.e. after the translation we obtain the new variables

$$\begin{pmatrix} \tilde{y} \\ \tilde{\mathbf{f}} \end{pmatrix} = \lambda \begin{pmatrix} y \\ \mathbf{f} \end{pmatrix} . \tag{8.3}$$

The moments of the transformed variables $(\tilde{y}, \tilde{\mathbf{f}}^{\mathrm{T}})^{\mathrm{T}}$ relate to the corresponding quantities of the original variables $(y, \mathbf{f}^{\mathrm{T}})^{\mathrm{T}}$ as follows:

$$\tilde{\boldsymbol{\mu}} = \lambda \boldsymbol{\mu} \, , \tilde{\boldsymbol{\Sigma}} = \lambda^2 \boldsymbol{\Sigma} \, , \tilde{\boldsymbol{\Phi}} = \lambda^3 \boldsymbol{\Phi} \quad \text{and} \quad \tilde{\boldsymbol{\Psi}} = \lambda^4 \boldsymbol{\Psi} \, . \tag{8.4}$$

The quantities derived from these moments are affected in the same way, i.e. $\boldsymbol{\varphi}_{\mathbf{A}}$, $\boldsymbol{\varphi}_{\mathbf{a}}$ and $\boldsymbol{\varphi}$ are multiplied by λ^3 and $\boldsymbol{\psi}_{\mathbf{A}}$, $\boldsymbol{\psi}_{\mathbf{a}}$ and $\boldsymbol{\psi}$ are multiplied by λ^4. The special second order moment matrix $\breve{\boldsymbol{\Sigma}}$ is multiplied by λ^2.
Proceeding like above we can derive the following general facts: Whenever a quadratic part is involved in the combination or adjustment, the corresponding optimal parameter matrix, vector or scalar after the scale transformation is $1/\lambda$ times the respective quantity ($\mathbf{A}_{\mathrm{opt}}$, $\mathbf{a}_{\mathrm{opt}}$ or α_{opt}) before the transformation. The optimal parameter vector or scalar for the linear part ($\mathbf{b}_{\mathrm{opt}}$ or b_{opt}) remains unchanged by the transformation. Finally, whenever a constant part (c_{opt}) is involved, the optimal choice after the transformation is λ times the optimal choice before the transformation. Accordingly, after the transformation each (single, adjusted or combined) forecast has been multiplied by the same scale factor λ by which the data have been multiplied. This is a reasonable behaviour. As a consequence after the transformation each MSPE-value is the λ^2–fold of the value before the scale transformation.

Combining the results on translations and scale transformations above we may conclude that only the linear unrestricted combination or adjustment without constant term $f_{\mathbf{b}_{\mathrm{opt}}}$ or $(f_i)_{b_{\mathrm{opt}}}$ are unreasonably sensitive to linear transformations of the data. Consequently, we cannot recommend the use of these techniques since results will depend on the chosen coordinate system.

The linear plus quadratic combined forecasts on the other hand show a reasonable behaviour. Since they include a constant term their MSPE-values are even insensitive with respect to the expectation vector $\boldsymbol{\mu} = (\mu_0, \boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}})^{\mathrm{T}}$ of the joint distribution of $(y, \mathbf{f}^{\mathrm{T}})^{\mathrm{T}}$.

9 Conclusions

In this paper we have introduced the linear plus quadratic approach for the combination of forecasts. Three versions of this approach have been considered. The strong version depends on the largest number of unknown combination parameters followed by the medium and then the weak version. We have derived equation systems from which the respective optimal combination parameters can be calculated. Each of the linear plus quadratic approaches requires knowledge about the moments up to order four of the joint distribution of $\mathbf{y}$ and $\mathbf{f}$. Again, the strong version requires more knowledge than the medium version, and the medium version requires more detailed knowledge than the weak version. We have also considered the classical linear approaches as competitors to the new approaches.

For the special case of $k = 2$ forecasts we have shown how the combination via the linear plus quadratic approaches works in detail. We have also considered the special case $k = 1$ which means adjustment of an individual forecast. Due to the smaller number of parameters involved the weak linear plus quadratic combination seems to be suitable if only a small amount of data is available for combination parameter estimation.

We have seen that the linear plus quadratic approaches show a reasonable behaviour when the coordinate system is changed in which the target variable and the forecasts are measured. From this point of view use of the linear unrestricted combination of forecasts without constant term $f_{\mathbf{b}_{\mathrm{opt}}}$ is not advisable. Thus it does not seem useful to investigate linear plus quadratic combinations $f_{\mathbf{A},\mathbf{b}}$, $f_{\mathbf{a},\mathbf{b}}$ or $f_{\alpha,\mathbf{b}}$ not involving a constant term c.

A detailed analysis of the possible benefits of the linear plus quadratic approaches has to follow. A point of special interest would be to find a guideline for potential users identifying situations beforehand in which linear plus quadratic combination of forecasts is promising. Especially the question of how much data should be available is interesting. Another point is to find out whether it is worthwhile to consider the combination of more than $k = 2$ forecasts via the linear plus quadratic approaches. As stated at the end of Section 3 derivation of the optimal combination parameters for the linear plus quadratic approaches may become quite cumbersome for

$k > 2$ forecasts. Consequently, it is desirable to find an easier way to apply linear plus quadratic combination. This is indeed possible: GRANGER and RAMANATHAN (1984) observe that the linear combination problems from Section 2 may be regarded as regression problems. Analogously, a regression approach may be followed for linear plus quadratic combination, thus allowing for easier implementation for any number k of forecasts and making standard computer software applicable. This regression approach will be dealt with in a follow-up paper by the same authors (TROSCHKE and TRENKLER (2000)).

Appendix

A A collection of useful results

This section lists some basic results which are needed for our considerations. Most of them are well-known from the literature. The others are quite immediate.
The first lemma provides the inverse of a regular matrix modified by a matrix of rank one:

Lemma A.1 (Rao and Bhimasankaram, 1992, p. 145) *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be non-singular and let* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. *Then*

$$(\mathbf{A}+\mathbf{u}\mathbf{v}^{\mathrm{T}})^{-1} = \mathbf{A}^{-1} - \frac{1}{1+\mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{u}}\,\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}\ .$$

The following two lemmas give explicit representations of some matrix or vector expressions in terms of the elements involved.

Lemma A.2 *Let* $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{m\times n}$, $\mathbf{x} = (x_i) \in \mathbb{R}^m$ *and* $\mathbf{y} = (y_j) \in \mathbb{R}^n$. *Then*

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}x_i y_j\ .$$

In the special case where $m = n$ *and* $\mathbf{A} = \mathbf{I}_n$ *we obtain*

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{n} x_i y_i\ .$$

Lemma A.3 *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{X} \in \mathbb{R}^{n\times m}$. *Then*

$$\mathrm{tr}(\mathbf{A}\mathbf{X}) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}x_{ji}\ .$$

In the special case where $m = n$ *and* $\mathbf{A}$ *is symmetric we obtain*

$$\mathrm{tr}(\mathbf{A}\mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_{ij}\ .$$

The next result is concerned with the first and second order moments of quadratic forms. Clearly, it is most important for our derivations. It should be pointed out that no distributional assumption is made. Assuming (multivariate) normality would lead to much simpler formulae on the one hand. But on the other hand the normality assumption would render the whole linear plus quadratic approach to the combination of forecasts unnecessary, as has been made clear in the introduction.

Lemma A.4 (RAO **and** KLEFFE, **1988, p. 32, (iv))** *Let* $\tilde{\mathbf{Y}} = \tilde{\boldsymbol{\mu}} + \tilde{\varepsilon}$ *where* $\tilde{\boldsymbol{\mu}}$ *is a constant vector and* $\tilde{\varepsilon}$ *is a vector random variable with moments* $\mathrm{E}(\tilde{\varepsilon}) = \mathbf{0}$, $\mathrm{E}(\tilde{\varepsilon}\tilde{\varepsilon}^{\mathrm{T}}) = \tilde{\boldsymbol{\Sigma}}$, $\mathrm{E}(\tilde{\varepsilon} \otimes \tilde{\varepsilon}\tilde{\varepsilon}^{\mathrm{T}}) = \tilde{\boldsymbol{\Phi}}$ *and* $\mathrm{E}(\tilde{\varepsilon}\tilde{\varepsilon}^{\mathrm{T}} \otimes \tilde{\varepsilon}\tilde{\varepsilon}^{\mathrm{T}}) = \tilde{\boldsymbol{\Psi}}$. *Further let* $\tilde{\mathbf{a}}$ *and* $\tilde{\mathbf{b}}$ *be vectors and let* $\tilde{\mathbf{A}}$ *and* $\tilde{\mathbf{B}}$ *be symmetric matrices of appropriate dimensions. Then*

(a) $\mathrm{E}(\tilde{\mathbf{a}}^{\mathrm{T}}\tilde{\mathbf{Y}} + \tilde{\mathbf{Y}}^{\mathrm{T}}\tilde{\mathbf{A}}\tilde{\mathbf{Y}}) = \tilde{\mathbf{a}}^{\mathrm{T}}\tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\mu}}^{\mathrm{T}}\tilde{\mathbf{A}}\tilde{\boldsymbol{\mu}} + \mathrm{tr}(\tilde{\mathbf{A}}\tilde{\boldsymbol{\Sigma}})$,

(b) $\mathrm{Cov}(\tilde{\mathbf{a}}^{\mathrm{T}}\tilde{\mathbf{Y}} + \tilde{\mathbf{Y}}^{\mathrm{T}}\tilde{\mathbf{A}}\tilde{\mathbf{Y}}, \tilde{\mathbf{b}}^{\mathrm{T}}\tilde{\mathbf{Y}} + \tilde{\mathbf{Y}}^{\mathrm{T}}\tilde{\mathbf{B}}\tilde{\mathbf{Y}})$

$$\begin{aligned} &= \tilde{\mathbf{b}}^{\mathrm{T}}\left[2\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{A}}\tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{a}} + \tilde{\boldsymbol{\Phi}}^{*}(\tilde{\mathbf{A}})\right] \\ &+ \mathrm{tr}\Big(\tilde{\mathbf{B}}\Big[4\tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^{\mathrm{T}}\tilde{\mathbf{A}}\tilde{\boldsymbol{\Sigma}} + 2\tilde{\boldsymbol{\Phi}}(\tilde{\mathbf{A}}\tilde{\boldsymbol{\mu}}) + 2\tilde{\boldsymbol{\Phi}}^{*}(\tilde{\mathbf{A}})\tilde{\boldsymbol{\mu}}^{\mathrm{T}} \\ &\qquad + \tilde{\boldsymbol{\Psi}}(\tilde{\mathbf{A}}) + 2\tilde{\boldsymbol{\mu}}\tilde{\mathbf{a}}^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}} + \tilde{\boldsymbol{\Phi}}(\tilde{\mathbf{a}}) - \mathrm{tr}(\tilde{\mathbf{A}}\tilde{\boldsymbol{\Sigma}})\tilde{\boldsymbol{\Sigma}}\Big]\Big) \ . \end{aligned}$$

Here the following abbreviations have been used: For a vector $\tilde{\mathbf{c}} = (\tilde{c}_i)$ *and a matrix* $\tilde{\mathbf{C}} = (\tilde{c}_{ij})$ *we define*

$$\begin{aligned} \tilde{\boldsymbol{\Psi}}(\tilde{\mathbf{C}}) &= \sum_i \sum_j \tilde{c}_{ij}\tilde{\boldsymbol{\Psi}}_{ij} \ , \\ \tilde{\boldsymbol{\Phi}}(\tilde{\mathbf{c}}) &= \sum_i \tilde{c}_i\tilde{\boldsymbol{\Phi}}_i \ , \\ \tilde{\boldsymbol{\Phi}}^{*}(\tilde{\mathbf{C}}) &= (\mathrm{tr}(\tilde{\mathbf{C}}\tilde{\boldsymbol{\Phi}}_i))_i \ , \end{aligned}$$

i.e. the first two quantities are matrices, whereas the last one is a vector.

In order to determine the optimal combination parameters within our various approaches differential calculus has to be applied. Since some of the parameters are vectors or even matrices the concept of matrix differential calculus (MAGNUS and NEUDECKER, 1999) proves most helpful.

Definition A.5 *Let* $f(\mathbf{X})$ *be a scalar valued function of a matrix* $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n\times q}$. *Then* f *is called differentiable with respect to* $\mathbf{X}$ *if and only if it is differentiable with respect to each of the elements* x_{ij}. *The derivative of* f *with respect to* $\mathbf{X}$

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} := \begin{pmatrix} \partial f(\mathbf{X})/\partial x_{11} & \dots & \partial f(\mathbf{X})/\partial x_{1q} \\ \vdots & & \vdots \\ \partial f(\mathbf{X})/\partial x_{n1} & \dots & \partial f(\mathbf{X})/\partial x_{nq} \end{pmatrix}$$

is a matrix with the same dimensions as $\mathbf{X}$.

Lemma A.6 *Let $f(\mathbf{X})$ be a differentiable scalar valued function of a matrix $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n\times q}$. Then a necessary condition for f to have a local minimum or a local maximum is*

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{0} \, ,$$

where the derivative of f with respect to $\mathbf{X}$ is given in Definition A.5 above.

The next two lemmas give the derivatives for special scalar valued vector and matrix functions.

Lemma A.7 (Magnus **and** Neudecker**, 1999, p. 177)** *Let $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$. Then*

$$\frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \, ,$$

$$\frac{\partial \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^{\mathrm{T}})\mathbf{x} \, .$$

Lemma A.8 (Magnus **and** Neudecker**, 1999, p. 178)** *Let $\mathbf{A}, \mathbf{B}, \mathbf{X}$ be real matrices of appropriate dimensions. Then*

$$\frac{\partial\, \mathrm{tr}(\mathbf{A}\mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}^{\mathrm{T}} \, ,$$

$$\frac{\partial\, \mathrm{tr}(\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{B})}{\partial \mathbf{X}} = \mathbf{B}^{\mathrm{T}}\mathbf{X}\mathbf{A}^{\mathrm{T}} + \mathbf{B}\mathbf{X}\mathbf{A} \, ,$$

$$\frac{\partial\, \mathrm{tr}(\mathbf{X}\mathbf{A}\mathbf{X}\mathbf{B})}{\partial \mathbf{X}} = \mathbf{B}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} + \mathbf{A}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} \, .$$

It is a special and difficult situation when the derivative is to be taken with respect to a symmetric matrix. The following lemma shows how to proceed correctly in this case.

Lemma A.9 (Rao **and** Rao**, 1998, p. 230)** *Let f be a scalar valued function of a matrix variable $\mathbf{A}$, where $\mathbf{A}$ is symmetric. Then*

$$\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} = \left\{ \frac{\partial f(\mathbf{B})}{\partial \mathbf{B}} + \left(\frac{\partial f(\mathbf{B})}{\partial \mathbf{B}}\right)^{\mathrm{T}} - \mathrm{diag}\left(\frac{\partial f(\mathbf{B})}{\partial \mathbf{B}}\right) \right\}\Bigg|_{\mathbf{B}=\mathbf{A}} \, .$$

This is meant to indicate that f is regarded as a function of an arbitrary matrix $\mathbf{B}$ which has the same size as $\mathbf{A}$, but all the components of $\mathbf{B}$ are regarded as independent variables. Then the derivative of f is formed with respect to $\mathbf{B}$, the above expression is calculated and in this expression $\mathbf{B}$ is replaced by the symmetric matrix $\mathbf{A}$ again.

Here for a square matrix $\mathbf{M} = (m_{ij})$ we define $\mathrm{diag}(\mathbf{M})$ as the diagonal matrix of the same dimension with the elements m_{ii} on its diagonal (compare Equation (3.11)).

The final lemma in this section is concerned with the derivatives of a special kind of function which is of major importance in Section 4 dealing with the medium linear plus quadratic approach.

Lemma A.10 *Let $a_1, \ldots, a_k$ be scalar variables. Further let f be a scalar valued function of two index variables l and m and let $f(l,m)$ be independent of the $a_1, \ldots, a_k$. Finally, let $s \in \{1, \ldots, k\}$ be fixed. Then*

$$\frac{\partial \sum_{l=1}^{k} \sum_{m=1}^{k} a_l a_m f(l,m)}{\partial a_s} = \sum_{l=1}^{k} a_l (f(s,l) + f(l,s)) \; .$$

Proof:

$$\begin{aligned}
&\frac{\partial \sum_{l=1}^{k} \sum_{m=1}^{k} a_l a_m f(l,m)}{\partial a_s} = \\
&= \frac{\partial \sum_{l=1}^{k} a_l^2 f(l,l)}{\partial a_s} + \frac{\partial \sum \sum_{l \neq m} a_l a_m f(l,m)}{\partial a_s} \\
&= \frac{\partial a_s^2 f(s,s)}{\partial a_s} + \frac{\partial \sum_{m \neq s} a_s a_m f(s,m)}{\partial a_s} + \frac{\partial \sum_{l \neq s} a_l a_s f(l,s)}{\partial a_s} \\
&= 2 a_s f(s,s) + \sum_{m \neq s} a_m f(s,m) + \sum_{l \neq s} a_l f(l,s) \\
&= \sum_{m=1}^{k} a_m f(s,m) + \sum_{l=1}^{k} a_l f(l,s) \\
&= \sum_{l=1}^{k} a_l f(s,l) + \sum_{l=1}^{k} a_l f(l,s) \\
&= \sum_{l=1}^{k} a_l (f(s,l) + f(l,s)) \; . \qquad \square
\end{aligned}$$

B Proof of assertions in Section 2

The observations deal with the optimal linear combination $f_{\mathbf{b}_{\text{opt}},\text{rest}}$ without constant term and with the restriction of the combination weights summing up to unity, i.e. $\mathbf{b}_{\text{opt}}^{\mathrm{T}}\mathbf{1} = 1$. This combination has been designed for the case where each individual forecast is unbiased. Consequently, the optimal weight vector $\mathbf{b}_{\text{opt}}$ is calculated on the basis of the covariance matrix $\breve{\boldsymbol{\Sigma}} = \mathrm{E}((\mathbf{Y} - \mu_0\mathbf{1})(\mathbf{Y} - \mu_0\mathbf{1})^{\mathrm{T}})$ making use of the unbiasedness assumption. Here $\mathbf{Y} = (y, \mathbf{f}^{\mathrm{T}})^{\mathrm{T}}$.

Assertion 2.1 *If the unbiasedness assumption is incorrect it is obvious that the true optimal* MSPE*-value* $\mathrm{MSPE}(f_{\mathbf{b}_{opt},rest}, y)$ *should be calculated by inserting* $\mathbf{b}_{opt}$ *from Equation (2.18) into the general Equation (2.11), which is valid for any linear combination of the type* $\mathbf{b}^{\mathrm{T}}\mathbf{f}$*. We obtain, however, the same Result (2.19) from inserting* $\mathbf{b}_{opt}$ *into the (now invalid) Equation (2.16).*

Proof: If the unbiasedness assumption is incorrect the following relation can be established between the true covariance matrix $\boldsymbol{\Sigma} = \mathrm{E}((\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^{\mathrm{T}})$ and the matrix $\breve{\boldsymbol{\Sigma}} = \mathrm{E}((\mathbf{Y} - \mu_0\mathbf{1})(\mathbf{Y} - \mu_0\mathbf{1})^{\mathrm{T}})$ which has been calculated following the incorrect assumption:

$$\boldsymbol{\Sigma} = \breve{\boldsymbol{\Sigma}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \mu_0\boldsymbol{\mu}\mathbf{1}^{\mathrm{T}} + \mu_0\mathbf{1}\boldsymbol{\mu}^{\mathrm{T}} - \mu_0^2\mathbf{1}\mathbf{1}^{\mathrm{T}} \ . \tag{B.1}$$

From this identity we may conclude that

$$\begin{aligned} \Sigma_{00} &= \breve{\Sigma}_{00} \\ \boldsymbol{\Sigma}_{\mathbf{f}0} &= \breve{\boldsymbol{\Sigma}}_{\mathbf{f}0} \\ \boldsymbol{\Sigma}_{\mathbf{ff}} &= \breve{\boldsymbol{\Sigma}}_{\mathbf{ff}} - \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}} + \mu_0\boldsymbol{\mu}_{\mathbf{f}}\mathbf{1}^{\mathrm{T}} + \mu_0\mathbf{1}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}} - \mu_0^2\mathbf{1}\mathbf{1}^{\mathrm{T}} \ . \end{aligned} \tag{B.2}$$

Inserting $\mathbf{b}_{\text{opt}}$ into the valid Equation (2.11), applying the above identities and exploiting the restriction $\mathbf{b}_{\text{opt}}^{\mathrm{T}}\mathbf{1} = 1$ we obtain

$$\begin{aligned} &\mathrm{MSPE}(f_{\mathbf{b}_{\text{opt}},\text{rest}}, y) = \\ &= \mathbf{b}_{\text{opt}}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathbf{ff}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}})\mathbf{b}_{\text{opt}} - 2\mathbf{b}_{\text{opt}}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathbf{f}0} + \mu_0\boldsymbol{\mu}_{\mathbf{f}}) + \Sigma_{00} + \mu_0^2 \\ &= \mathbf{b}_{\text{opt}}^{\mathrm{T}}(\breve{\boldsymbol{\Sigma}}_{\mathbf{ff}} + \mu_0\boldsymbol{\mu}_{\mathbf{f}}\mathbf{1}^{\mathrm{T}} + \mu_0\mathbf{1}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}} - \mu_0^2\mathbf{1}\mathbf{1}^{\mathrm{T}})\mathbf{b}_{\text{opt}} - 2\mathbf{b}_{\text{opt}}^{\mathrm{T}}(\breve{\boldsymbol{\Sigma}}_{\mathbf{f}0} + \mu_0\boldsymbol{\mu}_{\mathbf{f}}) + \breve{\Sigma}_{00} + \mu_0^2 \\ &= \mathbf{b}_{\text{opt}}^{\mathrm{T}}\breve{\boldsymbol{\Sigma}}_{\mathbf{ff}}\mathbf{b}_{\text{opt}} - 2\mathbf{b}_{\text{opt}}^{\mathrm{T}}\breve{\boldsymbol{\Sigma}}_{\mathbf{f}0} + \breve{\Sigma}_{00} \ . \end{aligned} \tag{B.3}$$

On the other hand, inserting $\mathbf{b}_{\text{opt}}$ into the presumably invalid Equation (2.16) for $\mathrm{MSPE}(f_{\mathbf{b}_{\text{opt}},\text{rest}}, y)$ gives

$$\begin{aligned} &\mathbf{b}_{\text{opt}}^{\mathrm{T}}(\breve{\boldsymbol{\Sigma}}_{\mathbf{ff}} + \mu_0^2\mathbf{1}\mathbf{1}^{\mathrm{T}})\mathbf{b}_{\text{opt}} - 2\mathbf{b}_{\text{opt}}^{\mathrm{T}}(\breve{\boldsymbol{\Sigma}}_{\mathbf{f}0} + \mu_0^2\mathbf{1}) + \breve{\Sigma}_{00} + \mu_0^2 \\ &= \mathbf{b}_{\text{opt}}^{\mathrm{T}}\breve{\boldsymbol{\Sigma}}_{\mathbf{ff}}\mathbf{b}_{\text{opt}} - 2\mathbf{b}_{\text{opt}}^{\mathrm{T}}\breve{\boldsymbol{\Sigma}}_{\mathbf{f}0} + \breve{\Sigma}_{00} \ , \end{aligned} \tag{B.4}$$

as well because of the restriction $\mathbf{b}_{\rm opt}^{\rm T}\mathbf{1} = 1$. This completes the proof. ∎

Assertion 2.2 *The optimal parameter vector $\mathbf{b}_{opt}$ is not changed if we use any other constant than μ_0 in the calculation of the covariance matrix $\breve{\mathbf{\Sigma}} = \mathrm{E}((\mathbf{Y} - \mu_0\mathbf{1})(\mathbf{Y} - \mu_0\mathbf{1})^{\rm T})$.*
An important consequence for practical applications is that we need not worry about which estimate of μ_0 should be used when estimating $\breve{\mathbf{\Sigma}}$: We may use the arithmetic mean of the observations on the target variable y, the arithmetic mean of all observations on the target variable y and *the single forecasts f_i, both of which are reasonable estimates, or we may even use 0.*
Proof: We show that the optimal parameter vector $\mathbf{b}_{\rm opt}$ is the same regardless whether we use $\breve{\mathbf{\Sigma}} = \mathrm{E}((\mathbf{Y} - \mu_0\mathbf{1})(\mathbf{Y} - \mu_0\mathbf{1})^{\rm T})$ or $\tilde{\mathbf{\Sigma}} = \mathrm{E}((\mathbf{Y} - \nu_0\mathbf{1})(\mathbf{Y} - \nu_0\mathbf{1})^{\rm T})$ for calculation, where $\nu_0 \in \mathbb{R}$ is arbitrary.
It is convenient to switch to an alternative representation of the optimal weight vector $\mathbf{b}_{\rm opt}$: Since $\mathbf{b}_{\rm opt}^{\rm T}\mathbf{1} = 1$ we may as well use the covariance matrix of the errors $\mathbf{e} = \mathbf{f} - y\mathbf{1}$ instead of the covariance matrix of $\mathbf{Y}$. Under the assumption of unbiasedness of the individual forecasts we have $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ and the covariance matrix of the errors is given by

$$\breve{\mathbf{V}} = \mathrm{E}(\mathbf{e}\mathbf{e}^{\rm T}) \ . \tag{B.5}$$

It is well-known that the optimal parameter vector $\mathbf{b}_{\rm opt}$ is then given by

$$\mathbf{b}_{\rm opt} = \frac{\breve{\mathbf{V}}^{-1}\mathbf{1}}{\mathbf{1}^{\rm T}\breve{\mathbf{V}}^{-1}\mathbf{1}} \tag{B.6}$$

and the corresponding optimal MSPE-value by

$$\mathrm{MSPE}(f_{\mathbf{b}_{\rm opt}}, y) = (\mathbf{1}^{\rm T}\breve{\mathbf{V}}^{-1}\mathbf{1})^{-1} \ , \tag{B.7}$$

which obviously depend on $\breve{\mathbf{V}}$ alone.
Now

$$\begin{aligned} \breve{\mathbf{V}} &= \mathrm{E}[\mathbf{e}\mathbf{e}^{\rm T}] = \mathrm{E}[(\mathbf{f} - y\mathbf{1})(\mathbf{f} - y\mathbf{1})^{\rm T}] \\ &= \mathrm{E}[((\mathbf{f} - \nu\mathbf{1}) + (\nu\mathbf{1} - y\mathbf{1}))((\mathbf{f} - \nu\mathbf{1}) + (\nu\mathbf{1} - y\mathbf{1}))^{\rm T}] \\ &= \mathrm{E}[(\mathbf{f} - \nu\mathbf{1})(\mathbf{f} - \nu\mathbf{1})^{\rm T}] - \mathrm{E}[(y-\nu)(\mathbf{f} - \nu\mathbf{1})]\mathbf{1}^{\rm T} \\ &\quad - \mathbf{1}\,\mathrm{E}[(y-\nu)(\mathbf{f} - \nu\mathbf{1})^{\rm T}] + \mathrm{E}[(y-\nu)^2]\mathbf{1}\mathbf{1}^{\rm T} \ , \end{aligned} \tag{B.8}$$

where $\nu \in \mathbb{R}$ is arbitrary. If we set $\nu = \mu_0$ in the final expression we obtain

$$\breve{\mathbf{V}} = \breve{\mathbf{\Sigma}}_{\mathbf{ff}} - \breve{\mathbf{\Sigma}}_{\mathbf{f}0}\mathbf{1}^{\rm T} - \mathbf{1}\breve{\mathbf{\Sigma}}_{\mathbf{f}0}^{\rm T} + \breve{\Sigma}_{00} \ , \tag{B.9}$$

whereas for $\nu = \nu_0$ we obtain

$$\check{\mathbf{V}} = \tilde{\mathbf{\Sigma}}_{\mathbf{ff}} - \tilde{\mathbf{\Sigma}}_{\mathbf{f}0}\mathbf{1}^{\mathrm{T}} - \mathbf{1}\tilde{\mathbf{\Sigma}}_{\mathbf{f}0}^{\mathrm{T}} + \tilde{\Sigma}_{00} \ . \tag{B.10}$$

Since $\mathbf{b}_{\mathrm{opt}}$ depends solely on $\check{\mathbf{V}}$ this completes the proof. ∎

C Derivation of optimal parameters for medium and weak linear plus quadratic combination

In Sections 4 and 5 the equations determining the optimal combination parameters for the medium and weak linear plus quadratic approaches are given. The purpose of this appendix is to provide some intermediate results from the omitted proofs. Both proofs are carried out along the same three steps which also occurred in the derivations connected with the strong approach.

First we will deal with the medium linear plus quadratic combination:

Step 1: Explicit calculation of the MSPE**-function.** This first step is accomplished by inserting $\mathbf{A} = \mathrm{dg}(\mathbf{a})$ in Equation (3.5):

$$\begin{aligned}
&\mathrm{MSPE}(f_{\mathbf{a},\mathbf{b},c}, y) = \\
&= \sum_{l=1}^{k}\sum_{m=1}^{k} a_l a_m \left(4\Sigma_{lm}\mu_l\mu_m + 4\mu_m\Phi_{mll} + \Psi_{mmll} + \mu_l^2\mu_m^2 + 2\mu_l^2\Sigma_{mm}\right) \\
&\quad - 2\sum_{l=1}^{k} a_l \left(2\Sigma_{l0}\mu_l + \Phi_{0ll} + \mu_0\mu_l^2 + \mu_0\Sigma_{ll}\right) \\
&\quad + 2\sum_{l=1}^{k}\sum_{m=1}^{k} a_l b_m \left(2\mu_l\Sigma_{lm} + \Phi_{mll}\right) + 2\mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}}\sum_{l=1}^{k} a_l(\mu_l^2 + \Sigma_{ll}) \\
&\quad + \mathbf{b}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{ff}}\mathbf{b} + \mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\mathbf{b} \\
&\quad - 2\mathbf{b}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{f}0} - 2\mu_0\mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} \\
&\quad + 2c\sum_{l=1}^{k} a_l(\mu_l^2 + \Sigma_{ll}) \\
&\quad + 2\mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}}c \\
&\quad + c^2 \\
&\quad - 2\mu_0 c \\
&\quad + \Sigma_{00} + \mu_0^2 \ ,
\end{aligned} \tag{C.1}$$

where

$$\boldsymbol{\varphi}_{\mathbf{a}} := \boldsymbol{\varphi}_{\mathrm{dg}(\mathbf{a})} = \left(\sum_{l=1}^{k} a_l \Phi_{1ll}, \ldots, \sum_{l=1}^{k} a_l \Phi_{kll} \right)^{\mathrm{T}} = \sum_{i=1}^{k} \sum_{l=1}^{k} a_l \Phi_{ill} \boldsymbol{\xi}_i^{(k)} \tag{C.2}$$

and

$$\boldsymbol{\psi}_{\mathbf{a}} := \boldsymbol{\psi}_{\mathrm{dg}(\mathbf{a})} = \begin{pmatrix} \sum_{l=1}^{k} a_l \Psi_{11ll} & \ldots & \sum_{l=1}^{k} a_l \Psi_{1kll} \\ \vdots & \ddots & \vdots \\ \sum_{l=1}^{k} a_l \Psi_{k1ll} & \ldots & \sum_{l=1}^{k} a_l \Psi_{kkll} \end{pmatrix} . \tag{C.3}$$

By $\boldsymbol{\xi}_i^{(k)}$ we denote the k–dimensional unit vectors, i.e. the i–th component of $\boldsymbol{\xi}_i^{(k)}$ is equal to 1 whereas the other components are equal to 0.

Step 2: Differentiation. Using common differential calculus and applying Lemma A.7 we get

$$\frac{\partial \mathrm{MSPE}(f_{\mathbf{a},\mathbf{b},c}, y)}{\partial c} = 2 \left[c - \mu_0 + \mathbf{b}^{\mathrm{T}} \boldsymbol{\mu}_{\mathbf{f}} + \sum_{l=1}^{k} a_l (\mu_l^2 + \Sigma_{ll}) \right] \tag{C.4}$$

and

$$\begin{aligned} \frac{\partial \mathrm{MSPE}(f_{\mathbf{a},\mathbf{b},c}, y)}{\partial \mathbf{b}} &= \\ = \ & 2[\boldsymbol{\Sigma}_{\mathbf{ff}} \mathbf{b} + \boldsymbol{\mu}_{\mathbf{f}} \boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}} \mathbf{b} - \boldsymbol{\Sigma}_{\mathbf{f}0} - \mu_0 \boldsymbol{\mu}_{\mathbf{f}} + c \boldsymbol{\mu}_{\mathbf{f}} \\ & + \left(\sum_{l=1}^{k} a_l (\mu_l^2 + \Sigma_{ll}) \right) \boldsymbol{\mu}_{\mathbf{f}} + \boldsymbol{\varphi}_{\mathbf{a}} + 2 \boldsymbol{\Sigma}_{\mathbf{ff}} \, \mathrm{dg}(\mathbf{a}) \boldsymbol{\mu}_{\mathbf{f}}] \, . \end{aligned} \tag{C.5}$$

In order to differentiate $\mathrm{MSPE}(f_{\mathbf{a},\mathbf{b},c}, y)$ with respect to $\mathbf{a}$, we differentiate with respect to a_s, $s = 1, \ldots, k$, and arrange the result in vector form. Lemma A.10 is applied several times. We finally arrive at

$$
\begin{aligned}
&\frac{\partial \mathrm{MSPE}(f_{\mathbf{a},\mathbf{b},c}, y)}{\partial \mathbf{a}} = \\
&= \; 8\,\mathrm{dg}(\boldsymbol{\mu}_{\mathbf{f}})\boldsymbol{\Sigma}_{\mathbf{ff}}\,\mathrm{dg}(\mathbf{a})\boldsymbol{\mu}_{\mathbf{f}} + 4\sum_{i=1}^{k}\sum_{l=1}^{k} a_l \mu_l \Phi_{lii}\boldsymbol{\xi}_i^{(k)} + 4\,\mathrm{dg}(\boldsymbol{\mu}_{\mathbf{f}})\boldsymbol{\varphi}_{\mathbf{a}} \\
&\quad + 2\sum_{i=1}^{k}\sum_{l=1}^{k} a_l \Psi_{llii}\boldsymbol{\xi}_i^{(k)} + 2\,\mathrm{tr}(\mathrm{dg}(\mathbf{a})\boldsymbol{\Sigma}_{\mathbf{ff}})\sum_{i=1}^{k}\mu_i^2\boldsymbol{\xi}_i^{(k)} \\
&\quad - 4\,\mathrm{dg}(\boldsymbol{\mu}_{\mathbf{f}})\boldsymbol{\Sigma}_{\mathbf{f}0} - 2\sum_{i=1}^{k}\Phi_{0ii}\boldsymbol{\xi}_i^{(k)} \\
&\quad + 4\,\mathrm{dg}(\boldsymbol{\mu}_{\mathbf{f}})\boldsymbol{\Sigma}_{\mathbf{ff}}\mathbf{b} + 2\sum_{i=1}^{k}\sum_{l=1}^{k} b_l \Phi_{lii}\boldsymbol{\xi}_i^{(k)} \\
&\quad + 2(c + \mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} - \mu_0 + \boldsymbol{\mu}_{\mathbf{1}}^{\mathrm{T}}\,\mathrm{dg}(\mathbf{a})\boldsymbol{\mu}_{\mathbf{f}})\sum_{i=1}^{k}(\mu_i^2 + \Sigma_{ii})\boldsymbol{\xi}_i^{(k)} \,.
\end{aligned}
\tag{C.6}
$$

Step 3: Equating to zero. Setting Equations (C.4), (C.5) and (C.6) simultaneously to zero we arrive at Equations (4.2), (4.3) and (4.4) from Section 4. They determine the optimal choices for $\mathbf{a}$, $\mathbf{b}$ and c.

Now we turn to the derivation of the optimal parameters for the weak linear plus quadratic combination:
Step 1: Explicit calculation of the MSPE**-function.** By inserting $\mathbf{A} = \alpha\mathbf{I}$ in Equation (3.5) we immediately arrive at

$$
\begin{aligned}
&\mathrm{MSPE}(f_{\alpha,\mathbf{b},c}, y) = \\
&= \; \alpha^2\left(4\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}\boldsymbol{\mu}_{\mathbf{f}} + 4\boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + \mathrm{tr}(\boldsymbol{\psi}) + (\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}})^2 + 2\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}}\,\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{ff}})\right) \\
&\quad + \alpha\left(-4\boldsymbol{\Sigma}_{\mathbf{f}0}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} - 2\,\mathrm{tr}(\boldsymbol{\Phi}_{0\mathbf{ff}}) - 2\mu_0(\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + \mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{ff}}))\right) \\
&\quad + \alpha\mathbf{b}^{\mathrm{T}}\left(4\boldsymbol{\Sigma}_{\mathbf{ff}}\boldsymbol{\mu}_{\mathbf{f}} + 2\boldsymbol{\varphi} + 2(\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + \mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{ff}}))\boldsymbol{\mu}_{\mathbf{f}}\right) \\
&\quad + \mathbf{b}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathbf{ff}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}})\mathbf{b} \\
&\quad + \mathbf{b}^{\mathrm{T}}\left(-2\boldsymbol{\Sigma}_{\mathbf{f}0} - 2\mu_0\boldsymbol{\mu}_{\mathbf{f}}\right) \\
&\quad + \alpha c\left(2(\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + \mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{ff}}))\right) \\
&\quad + c\mathbf{b}^{\mathrm{T}}(2\boldsymbol{\mu}_{\mathbf{f}}) \\
&\quad + c^2 \\
&\quad + c(-2\mu_0) \\
&\quad + \Sigma_{00} + \mu_0^2 \,,
\end{aligned}
\tag{C.7}
$$

since we have

$$\boldsymbol{\varphi}_{\alpha\mathbf{I}} = \alpha(\mathrm{tr}(\boldsymbol{\Phi}_{1\mathbf{ff}}),\ldots,\mathrm{tr}(\boldsymbol{\Phi}_{k\mathbf{ff}}))^{\mathrm{T}} =: \alpha\boldsymbol{\varphi} \tag{C.8}$$

and

$$\boldsymbol{\psi}_{\alpha\mathbf{I}} = \alpha \begin{pmatrix} \mathrm{tr}(\boldsymbol{\Psi}_{11\mathbf{ff}}) & \ldots & \mathrm{tr}(\boldsymbol{\Psi}_{1k\mathbf{ff}}) \\ \vdots & \ddots & \vdots \\ \mathrm{tr}(\boldsymbol{\Psi}_{k1\mathbf{ff}}) & \ldots & \mathrm{tr}(\boldsymbol{\Psi}_{kk\mathbf{ff}}) \end{pmatrix} =: \alpha\boldsymbol{\psi}\ . \tag{C.9}$$

For notational convenience we will abbreviate the coefficient of α^2 in the first line of Equation (C.7) by $d_{\alpha\alpha}$ and the coefficient of α in the second line by d_α. The vector by which $\alpha\mathbf{b}^{\mathrm{T}}$ is multiplied in the third line will subsequently be abbreviated by $\mathbf{d}_{\alpha\mathbf{b}}$. Note that neither of $d_{\alpha\alpha}$, d_α and $\mathbf{d}_{\alpha\mathbf{b}}$ depends on any of the unknown combination parameters.

Step 2: Differentiation. With the help of some differential calculus we derive

$$\frac{\partial \mathrm{MSPE}(f_{\alpha,\mathbf{b},c},y)}{\partial c} = \alpha\,[2\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} + 2\,\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{ff}})] + 2c - 2\mu_0 + 2\mathbf{b}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}}\ , \tag{C.10}$$

$$\frac{\partial \mathrm{MSPE}(f_{\alpha,\mathbf{b},c},y)}{\partial \mathbf{b}} = \alpha\mathbf{d}_{\alpha\mathbf{b}} + 2[\boldsymbol{\Sigma}_{\mathbf{ff}} + \boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}]\mathbf{b} - 2\boldsymbol{\Sigma}_{\mathbf{f}0} - 2\mu_0\boldsymbol{\mu}_{\mathbf{f}} + 2c\boldsymbol{\mu}_{\mathbf{f}} \tag{C.11}$$

and

$$\frac{\partial \mathrm{MSPE}(f_{\alpha,\mathbf{b},c},y)}{\partial \alpha} = 2d_{\alpha\alpha}\alpha + d_\alpha + \mathbf{b}^{\mathrm{T}}\mathbf{d}_{\alpha\mathbf{b}} + c[2\,\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{ff}}) + 2\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}}]\ . \tag{C.12}$$

Step 3: Equating to zero. Finally we set Equations (C.10), (C.11) and (C.12) simultaneously to zero and solve the resulting linear equation system for the unknown parameters. Thus we obtain the optimal choices for α, $\mathbf{b}$ and c.
From Equation (C.10) we obtain

$$c_{\mathrm{opt}} = \mu_0 - \mathbf{b}_{\mathrm{opt}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} - \alpha_{\mathrm{opt}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{f}} - \alpha_{\mathrm{opt}}\,\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{ff}})\ . \tag{C.13}$$

Using (C.13), from (C.11) we derive

$$\mathbf{b}_{\mathrm{opt}} = \boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}0} - \alpha_{\mathrm{opt}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\varphi} - 2\alpha_{\mathrm{opt}}\boldsymbol{\mu}_{\mathbf{f}}\ . \tag{C.14}$$

With the help of (C.13) and (C.14) Equation (C.12) can be equivalently expressed as

$$\alpha_{\mathrm{opt}}\left(\mathrm{tr}(\boldsymbol{\psi}) - \boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\varphi} - [\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{ff}})]^2\right) + \left(-\,\mathrm{tr}(\boldsymbol{\Phi}_{0\mathbf{ff}}) + \boldsymbol{\Sigma}_{\mathbf{f}0}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{ff}}^{-1}\boldsymbol{\varphi}\right) = 0\ . \tag{C.15}$$

Solving Equation (C.15) for α and inserting backwards into Equations (C.14) and (C.13) we arrive at Equations (5.2), (5.3) and (5.4) from Section 5. They give the optimal combination parameters for the current approach explicitly.

Acknowledgements: The authors wish to thank Jürgen Groß for his helpful comments and suggestions. The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

References

Bates, J.M. and Granger, C.W.J. (1969): 'The combination of forecasts'. *Operational Research Quarterly* **20**, 451-468.

Clemen, R.T. (1989): 'Combining forecasts: A review and annotated bibliography'. *International Journal of Forecasting* **5**, 559-583.

Granger, C.W.J. (1989): 'Combining forecasts – Twenty years later'. *Journal of Forecasting* **8**, 167-173.

Granger, C.W.J. and Ramanathan, R. (1984): 'Improved methods of combining forecasts'. *Journal of Forecasting* **3**, 197-204.

Harville, D.A. (1985): 'Decomposition of prediction error'. *Journal of the American Statistical Association* **80**, 132-138.

Magnus, J.R. and Neudecker, H. (1999): Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition. *Wiley, Chichester.*

Rao, A.R. and Bhimasankaram, P. (1992): Linear Algebra. *Tata McGraw-Hill, New Delhi.*

Rao, C.R. and Kleffe, J. (1988): Estimation of Variance Components and Applications. *North-Holland, Amsterdam.*

Rao, C.R. and Rao, M.B. (1998): Matrix Algebra and Its Applications to Statistics and Econometrics. *World Scientific, Singapore.*

Thiele, J. (1993): Kombination von Prognosen. (Wirtschaftswissenschaftliche Beiträge: Band 74). *Physica, Heidelberg.*

Troschke, S.O. and Trenkler, G. (2000): Regression Approach to the Linear Plus Quadratic Combination of Forecasts. *Technical Report 55/2000, Sonderforschungsbereich 475, University of Dortmund.*