Online Forecast Combination for Dependent Heterogeneous Data

Alessio Sancetta*
Faculty of Economics, University of Cambridge

April 18, 2007

Abstract

This paper studies a procedure to combine individual forecasts that achieve theoretical optimal performance. The results apply to a wide variety of loss functions and no conditions are imposed on the data sequences and the individual forecasts apart from a tail condition. The theoretical results show that the bounds are also valid in the case of time varying combination weights, under specific conditions on the nature of time variation. Some experimental evidence to confirm the results is provided.

Keywords: Forecast Combination, Model Selection, Multiplicative Update, Non-asymptotic Bound, On-line Learning.

JEL: C53, C14.

1 Introduction

This paper considers the problem of online combination of individual forecasts to improve the prediction error in terms of a variety of loss functions. The forecast combination problem has attracted much attention in the econometrics literature and the review article of Timmermann (2004) discusses the most recent advances. The present study is directly related to Yang (2004). The combination weights are derived using one of the recursive algorithms introduced in the machine learning literature (Kivinen and Warmuth, 1997, Cesa-Bianchi, 1999, and Herbster and

*Address for correspondence: Alessio Sancetta, Faculty of Economics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DE, UK. E-mail: alessio.sancetta@econ.cam.ac.uk.

Warmuth, 2001, and Chapter 11 in the recent monograph Cesa-Bianchi and Lugosi, 2006) and differs slightly from Yang (2004). In Yang (2004) the prediction error is compared to the prediction error of the best individual forecast. The prediction error of the present algorithm is compared to the best achievable combination weights and this is an improvement. Yang (2004) derives bounds that hold in expectation and do not require bounded random variables, but the conditions used are somehow restrictive. In particular, his results require the existence of the moment generating function for the forecast errors and are restricted to loss functions that are basically quadratic (or powers of a quadratic, but imposing additional restrictive tail conditions). The present results hold for more general data series (e.g. the moment generating function does not need to exist) and a wide variety of loss functions are allowed. Moreover, the optimal combination weights are allowed to change overtime. Clearly, all these extra flexibility comes at a price: the bounds derived are much weaker than the ones in Yang (2004, Theorem 4). Given that also the algorithms are different, the results given here are complementary and particularly useful when the more restrictive conditions in Yang (2004) do not apply. This paper also complements Sancetta (2006) in three ways: (1.) interest lies on forecast combination of individual sequences (2.) the procedure is compared to the best forecast combination with insight and not just to the best individual forecast, (3.) the theoretical bounds are expressed in terms of probabilities and expectations to avoid the assumption of bounded sequences.

The best combination of individual forecasts may lead to improvement over all the single individual forecasts. Several studies have shown that combining forecasts can be a useful hedge against structural breaks, and forecast combinations are often more stable than single forecasts (e.g. Hendry and Clements, 2004, Stock and Watson, 2004). Hence, it makes sense to try to approximate the best forecast combination. Moreover, while optimal forecast combinations are often derived by minimization of the user's expected loss over all possible decisions (e.g. Elliott and Timmermann, 2004), the presence of structural breaks might invalidate empirical estimation when we replace unknown expectations with sample ones. A similar remark applies when the noise level or persistence is quite high relatively to the sample size. In this case, it is often suggested to shrink the weights towards the equally weighted combination weights (e.g. Diebold and Pauly, 1990, Aiolfi and Timmermann, 2004). The procedure discussed here does not require any stability of the system besides a tail condition and it works in the presence of dependent observation. Moreover, in order to account for possible breaks or instability, a non-zero weight is retained for all individual forecasts, effectively performing some form of shrinkage as discussed in the above references. We show that this allow us to track the best time varying combination weights. Hence, the combination weights will be time varying, but there is no need to estimate changes in regime (e.g. Deutsch et al., 1994).

In particular, Section 2 states the algorithm used and carries out a theoretical analysis of its properties. Section 3 provides some experimental evidence for the validity of the theoretical results. Section 4 contains some further remarks and discussion. Proofs can be found in Section 5.

2 Online Forecast Combination

Suppose $(Y_t, X_t)_{t\in\mathbb{N}}$ are a sequence of random variables with values in $\mathcal{Y}\times\mathcal{X}$, where $\mathcal{Y}\subseteq\mathbb{R}$ and $\mathcal{X}\subseteq\mathbb{R}^K$ $(K>1)$. We interpret Y_t to be a quantity to be forecasted and X_t to be individual forecasts of it. We do not discuss the nature of these individual forecasts: they could be exogenous to the econometrician's decision rule. We use the forecasts X_t to construct the combined forecast $\hat{P}_t := \langle w_t, X_t\rangle \in \mathcal{P}$ to predict Y_t where $w_t \in \mathcal{S}^K$, and $\mathcal{S}^K$ is the K dimensional unit simplex ($\langle ..., ...\rangle$ is the inner product). For $\hat{P}_t$ to be a valid predictor, the weights w_t should only depend on the past and be independent of the present and future, i.e. we may only use $(Y_s, X_s)_{s<t}$ to construct w_t. The quality of the prediction $\hat{P}_t$ is evaluated by a loss function $l:\mathcal{Y}\times\mathcal{P}\to\mathbb{R}$, which is convex in the second argument. The cumulative loss based on using $\left(\hat{P}_t\right)_{t\leq T}$ is defined by $\hat{L}_T := \sum_{t=1}^T l\left(Y_t, \hat{P}_t\right)$. We shall write $l_t(w) = l(Y_t, \langle w, X_t\rangle)$ to stress dependence on $w\in\mathcal{S}^K$. We may use whatever information up to time $t-1$ to construct w_t and we will compare the loss from this approach to the loss incurred by $\inf_{u\in\mathcal{S}^K}\sum_{t=1}^T l_t(u)$. More generally, we will compare our loss with the loss incurred by using arbitrary weights $u_1, ..., u_T \in \mathcal{S}^K$, i.e. $L_T(u_1, ..., u_T) := \sum_{t=1}^T l_t(u_t)$. Since $u_1, ..., u_T$ are arbitrary, we may choose them such that $u_t := \arg\inf_{v\in\mathcal{S}^K} l_t(v)$. Clearly, in an arbitrary framework, we cannot expect to find any algorithm that produces results nearly as good as this choice of combination weights, but we still can do fairly well if the $u_1, ..., u_T$ are not allowed to change too often. Throughout, $u_1, ..., u_T$ will denote the optimal, but unfeasible combination weights from time 1 to T. Note that the time frame, i.e. T, does not need to be known in advance and the analysis is carried out for any arbitrary possibly unknown T. After all, when we carry out online estimation, we may not know in advance for how long we will use the same individual forecasts.

2.1 Recursive Choice of Regression Coefficients

The choice of combination weights is given by the algorithm in Exhibit 1. There, the gradient of the loss function with respect to w is denoted by $\nabla l_t(w)$ and $\nabla_k l_t(w)$ is its k^{th} element. The algorithm depends on the so called learning rate $\eta_t := \eta t^{-\alpha}$ that is a function of a coefficient $\eta > 0$ and time with exponent $\alpha \in (0, 1/2]$. The exponent α is related to the frequency of change in $u_1, ..., u_T$. For any K dimensional vector v, define $|v|_p = \left(\sum_{k=1}^{K} |v_k|^p\right)^{1/p}$, (with obvious modification when $p = \infty$). Under the square loss, if $\sum_{t=1}^{T-1} |u_t - u_{t+1}|_1 = O\left(T^{1-\epsilon}\right)$ $\epsilon \in [0, 1]$, α should satisfy $\alpha = \epsilon/2$ (see Corollary 3).[1] In both cases, the algorithm uses a second update that shrinks ("projects") the parameters onto a constrained set where the unfeasible parameters $u_1, ..., u_T$ are suppose to lie. This "projection" can be interpreted as shrinkage and allows to control the loss in case of changes in the underlying parameters. As discussed in the Introduction, shrinkage is a commonly used technique in econometric forecasting.

```
Exhibit 1.
 Set
 $w_{11} = ... = w_{1K} = 1/K$;
 $\eta > 0$;
 $\alpha \in (0, 1/2]$;
 $\gamma \geq 0$;
 $\hat{L}_0 := 0$;
 For $t = 1, ..., T$
 $\hat{L}_t = \hat{L}_{t-1} + l_t(w_t)$
 $\eta_t = \eta t^{-\alpha}$;
 $w'_{t,k} := w_{tk} \exp\{-\eta_t \nabla_k l_t(w_t)\} / \sum_{k=1}^{K} w_{tk} \exp\{-\eta_t \nabla_k l_t(w_t)\}$;
 $\psi := \sum_{k=1}^{K} I_{\{w'_{t,k} < \gamma/K\}}$
 $\omega := \sum_{k=1}^{K} I_{\{w'_{t,k} < \gamma/K\}} w'_{tk}$;
 $w_{t+1,k} = w'_{tk} [1 - (\psi\gamma)/K] / [1 - \omega]$ if $w'_{tk} \geq \gamma/K$, $w_{t+1,k} = \gamma/K$ otherwise.
```

The algorithm uses two updates which will be motivated in the next subsection. The first update produces the ex post combination weight $w'_{t,k}$. This is called ex post because the update depends on the gradient of the observed loss at time t. This ex post weight is then projected on a subset of $\mathcal{S}^K$ in order to make sure

[1] For simplicity we only consider polynomial growth.

that no weight is less than the user's prespecified level γ/K. In particular, all weights less than γ/K are set equal to the lower bound level γ/K. To impose the constraint that the weights are in $\mathcal{S}^K$ the weights that are greater than γ/K are then shrunk towards γ/K by an amount $[1-(\psi\gamma)/K]/[1-\omega]$ to make sure that the constraint is satisfied. Note that setting $\gamma=1$ produces the equally weighted forecast combination. Below we provide motivation for these update.

2.2 Motivation

The algorithm can be motivated using the approach first suggested in Kivinen and Warmuth (1997). Define $D(u,w) := \sum_{k=1}^K u_k \ln(u_k/w_k)$, which is the relative entropy. The ex post weight w_t' is such that, under regularity conditions for w close to w_t,

$$\begin{aligned} & \arg\inf_{w\in\mathcal{S}^K} \left[D(w,w_t) + \eta_t l(Y_t, \langle w, X_t\rangle)\right] \\ \simeq\ & \arg\inf_{w\in\mathcal{S}^K} \left\{D(w,w_t) + \eta_t \left[l(Y_t,\langle w_t,X_t\rangle) + l'(Y_t,\langle w_t,X_t\rangle)(w-w_t)\right]\right\} \quad (1) \\ =\ & : w_t', \end{aligned}$$

where $l'(y,p) := (\partial/\partial p)\, l(y,p)$ throughout. We could use some "distance" function other than the relative entropy, and this would produce different updates. In the computation of w_t', the loss function l is replaced by its first order Taylor series approximation in order to allow for simple closed form solution. The optimization problem has two terms that act in opposite directions: the first ones requires the ex post weight w_t' to be close to the ex ante one, the second term, requires the ex post weight to be an update of the ex ante weight such that the loss at time t would have been minimized. The coefficient $\eta_t > 0$ controls these two opposing aspects. When η_t is large, the update changes the weight towards the would have been optimal weight. A small η_t implies smoother ex post updates (i.e. more dependence on past observations). The coefficient η_t will be called learning rate. The combination weights at time t are obtained by a further update. We use a function $Q_{(\mathcal{G},D)} : int\left(\mathcal{S}^K\right) \rightarrow \mathcal{S}^K$ that depends on some closed convex subset $\mathcal{G}\subset\mathcal{S}^K$ and on the "distance" D ($int\left(\mathcal{S}^K\right)$ denotes the interior of $\mathcal{S}^K$). In particular, $Q_{(\mathcal{G},D)}(w_t') = \arg\min_{u\in\mathcal{G}} D(u,w_t')$ and $w_t = Q_{(\mathcal{G},D)}(w_t')$ is the projection of w_t' onto

$$\mathcal{G} := \left\{u \in \mathcal{S}^K \cap [\gamma/K, 1]^K\right\} \quad (2)$$

in terms of D. To constrain the parameters in $\mathcal{S}^K$ we add the constraint $\sum_{k=1}^K w_k = 1$ and the update to w_t' is obtained by simple optimization of the Lagrangian obtained from (1). The second update to w_t is obtained from $Q_{(\mathcal{G},D)}$. It is easy to see that the solution lies in $\mathcal{G}$. A proof that $Q_{(\mathcal{G},D)}(w_t') = w_{t+1}$ as in Exhibit 1 can be found in Herbster and Warmuth (2001).

2.3 Theoretical Performance of the Algorithms

A theoretical justification for the algorithm in Exhibit 1 is given. In particular, it is shown that reasonable bounds hold uniformly in the time horizon. If we have a rough idea of the frequency of breaks, we can use this extra knowledge to improve the bound. Moreover, the bounds avoid the usual condition of bounded random variables in favour of tails conditions. Because of this, the bound is weaker than the ones usually derived in the machine learning literature. We shall use the following conditions.

Condition 1 *$l(y,p)$ is convex in the second argument and has a derivative with respect to the second argument and there exists a continuous function f such that for any constant $m > 0$*

$$\max_{|y|\leq m, |p|\leq m} |(\partial/\partial p)\, l(y,p)| \leq f(m).$$

Remark 1 *Condition 1 is satisfied by any loss function $l(y,p) = l(y-p)$ which is convex, with continuous derivatives. Common examples are the square loss and LinEx. However, this condition is also satisfied by discontinuous functions as long as they are dominated by some function with continuous first derivative. A common example are ϵ insensitive loss functions, i.e. equal to $l(y,p) = l(y-p)$ if $|y-p| > \epsilon$, zero otherwise.*

Condition 2 *There are positive constants c_1, c_2 and c_3 such that*

$$\begin{aligned} \Pr(|Y_t| > x) &\leq c_1 \exp\{-c_2 x^{c_3}\} \\ \Pr(|X_{tk}| > x) &\leq c_1 \exp\{-c_2 x^{c_3}\}, \ k = 1, ..., K. \end{aligned}$$

Remark 2 *In Condition 2 the exponent c_3 can be arbitrarily small so that the moment generating function does not need to exist. Moreover, the condition is in terms of the target sequence $(Y_t)_{t\in\mathbb{N}}$ and the individual forecasts $(X_t)_{t\in\mathbb{N}}$ and not in terms of the forecast error, whose distribution might be more difficult to derive. For loss functions with polynomial growth, the results are also valid for polynomial tails, but with bounds worse than the ones to be shown. For any random variable Z define $\|Z\|_r := (\mathbb{E}|Z|^r)^{1/r}$ (the L_r norm, $r > 0$). Then, using the square loss,* *it can be shown that error/T in Corollary 1 is* $o\left(T^{-\frac{8-\gamma}{2\gamma}}\right)$ *as soon as* $\|X_{tk}\|_{\gamma+\epsilon}$ *and* $\|Y_t\|_{\gamma+\epsilon}$ *are finite for* $\epsilon > 0$*, requiring the existence of an* $8+\epsilon$ *moment. For the sake of simplicity the polynomial case is not considered. These details are left to the interested reader who will be able to discover that the dependence on* K*, in Theorem 1, also deteriorates from logarithmic to polynomial growth (with exponent less than one).*

Remark 3 *Condition 8 in Yang (2004) requires* $\mathbb{E}|R|^{2\beta}\exp\left\{t|R|^{\beta}\right\} < \infty$*, where* R *is the forecast error and* $\beta > 0$ *is as determined in his Condition 7. His Condition 7 requires* $\beta \geq 1$*, hence the forecast errors need to have a distribution with tails that decay at least as fast as an exponential.*

The bounds to be shown are of the kind

$$\sum_{t=1}^{T}\left[l_t(w_t) - l_t(u_t)\right] \leq error$$

with some high probability, where *error* refers to a bound for the algorithm in Exhibit 1 for any arbitrary $u_1, ..., u_T \in \mathcal{G}$. Hence, note that the unfeasible weights we compare to are also constrained to lie in (2). Besides this, they are arbitrary and we could choose them at time T, with hindsight. This will not be mentioned again. We will also show that these bounds hold in expectation.

Theorem 1 *Define* $m_T(\tau) := \left[\frac{1}{c_2}\tau + \frac{1}{c_2}\ln\left(c_1(K+1)T\right)\right]^{1/c_3}$*. Under Condition 1 and 2, with probability at least* $1 - e^{-\tau}$

$$\sum_{t=1}^{T}\left[l_t(w_t) - l_t(u_t)\right] \leq error$$

where

$$error := \frac{T^{\alpha}}{\eta}\left(6\ln(K/\gamma) + \left[\ln(K/\gamma) + K\right]\sum_{t=1}^{T-1}|u_t - u_{t+1}|_1\right) + \frac{\eta T^{1-\alpha}}{8}\left[f\left(m_T(\tau)\right)m_T(\tau)\right]^2.$$

The bound shows how the choice of $u_1, ..., u_T$ affects the relative performance. Note that we did not impose any dependence conditions. For loss functions that grow polynomially (e.g. the square loss), the above bounds only grow polynomially in τ being violated with exponentially small probability $e^{-\tau}$. The constant 6 in the first term is much larger than necessary, but leads to a tidier bound. Note that the bound depends on γ, i.e. the set $\mathcal{G}$. Setting γ small would increase the set of allowed weights, but also increase the error in the bound.

The bounds of Theorem 1 require some further comment in order to be fully appreciated. To this end, we use a series of corollaries to show its applications. The first important case is when we compare to the best time invariant combination weights with hindsight.

Corollary 1 *Suppose* $u_t = u_{t+1}$ $\forall t$*. Using the conditions and notation of Theorem 1, with probability at least* $1 - e^{-\tau}$,

$$\sum_{t=1}^{T} l_t \left(w_t\right) - \inf_{u \in \mathcal{G}} \sum_{t=1}^{T} l_t \left(u\right) \leq error$$

where

$$error \leq T^{1/2} \left[6 \ln \left(K/\gamma\right) + \frac{1}{8} \left[f \left(m_T \left(\tau\right)\right) m_T \left(\tau\right)\right]^2\right],$$

when we use the learning rate $\eta_t = t^{-1/2}$.

For time invariant combination weights, the error only grows logarithmically in the number of combination weights K. Moreover, $error/T \to 0$ if $\left[f \left(m_T \left(\tau\right)\right) m_T \left(\tau\right)\right]^2 = o\left(T^{1/2}\right)$. This depends on the loss function and on the tails of the random variables. To avoid convoluted technical conditions, the following only gives two special simple examples. To ease notation, dependence on K (finite and fixed) will be suppressed.

Corollary 2 *Under Condition 2, with probability at least* $1 - e^{-\tau}$,
(i.) if $l\left(y, p\right) = \left|y - p\right|^2$,

$$\frac{error}{T} = O\left(T^{-1/2} \left[\tau + \ln T\right]^{4/c_3}\right);$$

(ii.) if $l\left(y, p\right) = \exp\left\{\theta\left(y - p\right)\right\} - \theta\left(y - p\right) - 1$*, and* $c_3 > 1$,

$$\frac{error}{T} = o\left(e^{b\tau^{1/c_3}}\right),$$

for some finite constant b *depending on* θ, c_2 *and* c_3.

As a last application, we show that the algorithms discussed here allow to hedge against changes in the regression coefficients. For simplicity, we only consider the square loss case.

Corollary 3 *Suppose* $\sum_{t=1}^{T-1} \left|u_t - u_{t+1}\right|_1 = O\left(T^{1-\epsilon}\right)$ $(\epsilon \in [0,1])$ *and* $l\left(y, p\right) = \left|y - p\right|^2$*. Then, under Condition 2, with probability at least* $1 - e^{-\tau}$,

$$error = O\left(T^{1-(\epsilon/2)} \left[\tau + \ln T\right]^{4/c_3}\right),$$

when we choose $\eta_t = O\left(t^{-\frac{\epsilon}{2}}\right)$.

The result shows that we can allow for changes in parameters as frequent as $O\left(T^{1-\epsilon}\right)$ with $\epsilon > 0$ (note that $(\ln T)^{4/c_3} = o\left(T^{\epsilon/2}\right)$) and still obtain a reasonable performance, though only asymptotically. In practice, if we believe that changes might be somehow frequent (e.g. $\epsilon \to 0$), then we should choose α in Exhibit 1 to be smaller than $1/2$.

While it is quite common to state probability bounds, in econometrics, bounds are usually stated in terms of expectations. Note that a probability bound plus uniform integrability implies a bound in expectations. Hence, we have the following.

Corollary 4 *Suppose* $\max_{t\leq T} \left\| l\left(Y_t, \langle w_t, X_t\rangle\right)\right\|_r \leq A < \infty$ *with* $r = 2$*. Then, all the previous bounds hold taking expectation on the left hand side, adding* $2A\sqrt{T}$ *and setting* $\tau = 0$ *in* $m_T(\tau)$*, i.e.*

$$\sum_{t=1}^{T} \mathbb{E}\left[l_t(w_t) - l_t(u_t)\right] \leq error + 2A\sqrt{T}.$$

Clearly, square integrability of the loss function is not required. If we can only assume a finite bound with $r \in (1,2)$, we need to modify the definition of m_T in order to balance all the terms. This can be done by slight modification of the proof of Corollary 4 and details are left to the interested reader.

The results in this section can be summarized as follows: (1.) recursive forecast combination may allow us to perform almost as well as if we knew the combination weights for the whole sample; (2.) *error* only grows logarithmically in the number K of individual forecasts under suitable tail conditions when we compare to the best time invariant combination weights.

3 Experimental Results

We consider the problem of dynamic model selection. Suppose we can use different models to generate a conditional forecast of some series of random variables. We suggest to use a combined forecast instead of identifying the best model using some cross-validation procedure that estimates the prediction error. Prediction error estimation might be unstable in some circumstances, hence, combination of forecasts from different models might lead to more stable results. This is conceptually similar to the procedures studied in Breiman (1996) though he focuses on the linear model. As shown in the theoretical results, combining forecasts may lead to performance superior to the choice of the best individual forecast when the loss function is convex. Our goal is to show the use of forecast combination in the context of model selection (the two are intimately related) using both simulated data-sets and empirical data.

3.1 Combining Nonparametric Estimators to Approximate Continuous Functions

The goal of this application is to approximate an unknown univariate continuous function subject to additive noise by nonparametric estimators with different degrees of smoothing (estimated using past observations only). In particular we will use nearest neighbor (NN) estimators (Cover, 1968, and Kohler et al., 2006, for recent results and further references). These are very simple estimators and each estimator will be identified by a different proportion of data used to construct it. Since the performance of NN regression strongly depends on the proportion of data used (or number of neighbors), different neighbors' sizes lead to quite different forecasts. These forecasts are then combined using the algorithm in Exhibit 1. This kind of experiment is also interesting because, as the number of forecasts increases, we have more past observations at our disposal so that the optimal degree of smoothness in the NN regression decreases at each trial. These implies that the optimal combination weights need also to be time varying, which is exactly what we can allow for by choosing $\alpha < 1/2$ in Exhibit 1.

The results of the simulation study would strongly depend on the Monte Carlo design chosen, e.g. on the function we choose to approximate. To mitigate this problem, it is a good idea to perform forecast prediction for many continuous functions. Hence we will simulate a relatively large number of functions (as in Friedman, 2001). Consider a function $F : \mathbb{R} \rightarrow \mathbb{R}$ that admits the following representation

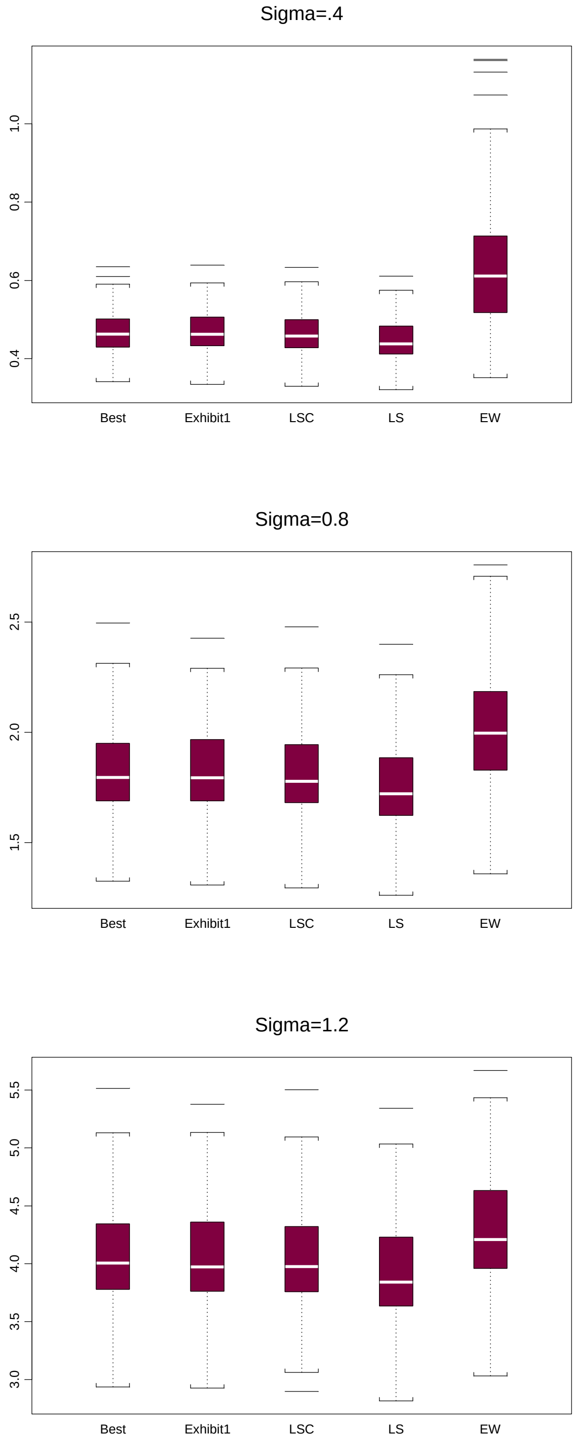
$$\begin{aligned} F(z) &= \sum_{i=1}^{I} a_i g_i(z) \\ g_i(z) &= \exp\left\{ -\frac{(z-b_i)^2}{2c_i^2} \right\}, \end{aligned} \tag{3}$$

where $|a_i|$ is a bounded real and $b_i, c_i \in \mathbb{R}$, $i = 1, ..., I$. For $I \rightarrow \infty$ the class of functions parametrized by a_i, b_i, c_i, $(i = 1, ..., I)$ is dense in the class of continuous bounded functions on compact subsets of $\mathbb{R}$ (e.g. Ripley, 1996). For the simulation study, we shall consider $I = 20$, $(a_i)_{i\in\{1,...,I\}}$ iid uniform in $[-1, 1]$, and $(b_i)_{i\in\{1,...,I\}}$ iid normal with mean zero and variance one $(N(0,1))$. For simplicity, $c_i = c$ $(\forall i)$ is also $N(0,1)$. The scaling parameters c_i are set all equal in order to avoid particularly irregular functions that might be very uncommon in any practical application. One hundred functions are obtained by simulation of the coefficients in (3) and 100 samples $\left(Y_t^{(r)}, Z_t^{(r)}\right)_{t\in\mathbb{N}}$ of size n $(r = 1, ..., 100)$ are simulated as follows:

$$\begin{aligned} Y_t^{(r)} &= F^{(r)}\left(Z_t^{(r)}\right) + U_t^{(r)} \\ Z_t^{(r)} &= .9Z_{t-1}^{(r)} + \xi_t^{(r)} \\ U_t^{(r)} &= .8U_{t-1}^{(r)} + \varepsilon_t^{(r)}, \end{aligned}$$

where, for each r, $\left(\xi_t^{(r)}\right)_{t\in\mathbb{N}}$ and $\left(\varepsilon_t^{(r)}\right)_{t\in\mathbb{N}}$ are, respectively, sequences of iid $N(0,1)$ and $N(0,\sigma^2)$. Hence, each r corresponds to a function $F^{(r)}$ which is identified by the simulated parameters a_s, b_s, $c_s = c$ in (3). Predictions are made constructing neighbors based on a fixed proportion of past observations. At each new data point the following proportions of sample data are used: $h = 0.025, .05, .1, .2, ..., .6$. Hence, each h identifies a different nonparametric estimator, with h large implying higher degree of smoothness so that $X_t^{(r)} := \left(X_{t1}^{(r)}, ..., X_{tK}^{(r)}\right)$ is the estimator for different h's constructed using $\left(Y_s^{(r)}, Z_s^{(r)}\right)_{s<t}$ and evaluated at $Z_t^{(r)}$. For each of these functions, results are tested on $\sigma = .4, .8, 1.2$ and sample size $n = 700$. We start forecast combination from the 101 observation so that 100 observations are available for estimation of the predictors on the first round. Clearly, as the number of observations increases, we should decrease h in order to achieve consistency in the limit. This should be done slowly enough with respect to the sample size. In this example the sample size increases from 101 to 700 hence, the combination weights should be time varying: at the beginning more weight should be given to estimators with h fairly large. We construct the forecast combination $\hat{P}_t := \langle w_t, X_t\rangle$, and assess its performance using the square loss $l(y,p) = |y-p|^2$. This exercise is of interest because the complex nonlinear structure of the data requires a small h to deliver a good approximation, but the relatively small sample size requires a large h to minimize the estimation error. This is the typical trade off between bias (approximation) and variance (estimation) error. Hence, as the sample size increases, we expect the combination weights to change, gradually preferring forecasts based on smaller h. In order to allow for meaningful time variation, we assume that $1-\epsilon = 1/3$ in Corollary 3 is a reasonable speed of change. This implies that $\alpha = 1/3$ in Exhibit 1. We also set $\gamma = .05$, and $\eta = 1$ for lack of better choice. (Had we chosen to work with data with ticker tails, the theoretical results and simulations carried out by the author suggest that a small $\eta\,(< 1)$ would be a better choice.) We compare to the ex post best individual forecast (Best), the ex post least square combination constrained in $\mathcal{G}$ (LSC), the ex post unconstrained least square combination (LS) and the equally weighted (EW) forecast combination. The results are summarized in the boxplots in Figure 1. These boxplots are based on 100 average losses from samples of size n: each average loss corresponds to a simulated target function. Overall the results are quite promising. As the noise level σ increases, the optimal h's tend to be in the middle of our chosen range $[.025, .6]$. This improves overall performance of forecast combination and particularly for EW. In fact, if the optimal h's are very small (e.g. close to or less than .025), then we have less good individual forecasts that can be used (as we only allow the range $[.025, .6]$). As shown by the boxplots, this is a big problem for EW forecasts combination. One consequence of this observation is that choosing individual forecasts carefully is very important especially when we use shrinkage or like to use EW forecast combinations. Similar and important comments in relation to the use of EW forecasts can be found in the simulation study of Timmermann (2004).

Figure 1. Boxplots of Prediction Errors



3.2 Which GARCH Model to Choose under the Absolute Loss?

This example focuses on volatility estimation using different GARCH models. We use 200 initial observations to estimate the models and a prediction is made for observation 201, then the models are re-estimated using 201 observations and a prediction is made for observation 202 and so on.

The experiments are carried out using the log returns on the front month of the following financial futures: FTSE (11/10/94-22/09/06), DAX (04/01/99-22/09/06), S&P (14/03/89-19/06/01) and Dow Jones (06/10/97-22/09/06). We consider GARCH, PGARCH (power GARCH with power 1), TGARCH (threshold GARCH), EGARCH (exponential GARCH), Two-Component GARCH and FIGARCH (fractionally integrated GARCH). In all cases, the order is (1,1) and the conditional distributions considered are Gaussian, double exponential and Student t. We also consider the case where the conditional mean is zero (standard case) and when it follows an ARMA(1,1) with an intercept. For the FIGARCH the conditional distribution is Gaussian. Estimation was carried out in S+FinMetrics$^{\text{TM}}$ (Zivot and Wang, 2005, for details on these GARCH specifications and the implementation in S+FinMetrics$^{\text{TM}}$). Considering different conditional distributions for the errors and conditional mean functions, we end up with 34 volatility forecasts. For sufficiently small degrees of freedom in the t-distribution, the 4th moment of the log returns might be undefined. Since we put no restriction on the degrees of freedom, we shall not use the square loss. The absolute loss for the difference between squared log returns and the volatility forecasts is employed. Moreover, the absolute loss is more robust to outliers, which might be a problem given that we are already using squared returns. Note that Condition 2 does not allow for power tails (e.g. a t-distribution), though Remark 2 suggests that moment conditions are sufficient (a $4+\epsilon$ would be sufficient when using the absolute loss). We do not worry about these issues and report Results in Table 2 where we used $\alpha = 1/2$ and $\gamma = .05$. Note that the in sample least square estimator of the combination weights does not need to be optimal under the absolute loss, but it is computed anyway due to its simplicity. This is confirmed by the results. The performance of the algorithm in Exhibit 1 is comparable to the best ex post forecast and slightly superior to the EW combined forecast. However, the large standard errors show that no definite claim should be made about this last remark. In particular, given the small variability in performance among the GARCH specification, the EW forecast combination appears to be a simple and robust alternative. For the sake of curiosity, we mention that the choice of double exponential density for the conditional distribution leads to mediocre sample performance and that TGARCH models with conditional Gaussian and/or t-distribution received on average the largest combination weights, i.e. produce good forecasts.

Table 2. Loss of Predictions for GARCH Predictions.

		Worst	Best	Exhibit1	LS	EW
FTSE	Mean	1.41	1.26	1.27	1.23	1.30
	S.E.	0.04	0.04	0.04	0.04	0.04
DAX	Mean	2.70	2.41	2.42	2.42	2.50
	S.E.	0.10	0.10	0.10	0.09	0.10
SP	Mean	1.15	1.08	1.09	1.11	1.10
	S.E.	0.04	0.04	0.04	0.04	0.04
DJ	Mean	1.48	1.31	1.33	1.34	1.38
	S.E.	0.06	0.05	0.06	0.05	0.05

3.3 Choosing the Best Exponentially Weighted Average Forecast: Combination of Many Individual Forecasts

The goal of this experiments is to use a large number (160) of individual forecasts generated by exponential moving average. Due to the large number of forecasts, the performance should deteriorate, but according to the theoretical results not too much. To investigate this issue, we consider absolute values of log returns on different financial futures. Predictability of absolute returns of stock indexes appears to be strong (e.g. Ding et al. 1993, Ding and Granger, 1996). Note that prediction of powers of absolute returns is a way to predict volatility, and the use of absolute returns appears good not only because of the stronger linear time dependence, but also because of more regularity as opposed to square returns (Mercurio and Spokoiny, 2004).

To generate a large number of different individual forecasts, predictions based on exponentially weighted moving averages are used. This is a simple estimator that depends on one parameter only. Suppose $(Y_t)_{t\in\mathbb{N}}$ are absolute values of log returns on some financial asset, then forecasts are constructed as

$$X_{tk} = (1-\lambda_k)\sum_{s=0}^{\infty}\lambda_k^s Y_{t-1-s},$$

where $\lambda_k := (m_k-1)\,/\,(m_k+1)$, with $m_k = 5 : 800\,(5)$, (i.e. $5, 10, ..., 800$), so that we have a total of 160 individual forecasts. Taking it to the extreme (because of our goal in this subsection), this could be the problem faced by someone trying to estimate volatility on financial assets using exponentially weighted moving averages (as in RiskMetrics), but being agnostic about the smoothing parameter. The square loss between the absolute returns and the forecasts is used to assess the performance (e.g. Fan and Gu, 2003, p.270).

In particular, we consider log returns on the securities of the previous example. In the case of the S&P futures we decided to use the extended period 25/03/88-22/09/06 to investigate the behavior of the algorithm over longer time spans. (In the previous example, a shorter period was chosen because of computational reasons.) Table 2 reports the mean and standard errors after running the algorithm with the same parameters as in the previous experiment. We also consider the ex post worst prediction just to give an idea of the variability of the individual forecasts performances. There is not excessive variability in performance. Hence, most of the individual forecasts are almost redundant being strongly correlated. The large number of forecasts and the strong collinearity require the use of generalized inverses to compute the in sample least square estimator. Results show that the algorithm achieves a performance almost similar to the best individual forecast and the ex post least square estimator, as predicted by the theory despite the large number of forecasts. The EW forecast also produces forecasts that are reasonably good. This is somehow to be expected because of the strong correlation among individual forecasts.

Table 1. Loss of Predictions for Absolute Values

		Worst	Best	Exhibit1	LS	EW
FTSE	Mean	0.57	0.48	0.49	0.48	0.53
	S.E.	0.03	0.02	0.02	0.02	0.03
DAX	Mean	1.09	0.91	0.92	0.90	1.00
	S.E.	0.07	0.05	0.05	0.05	0.06
SP	Mean	0.57	0.45	0.45	0.44	0.49
	S.E.	0.03	0.02	0.03	0.03	0.03
DJ	Mean	0.59	0.52	0.53	0.51	0.55
	S.E.	0.03	0.03	0.04	0.03	0.04

4 Further Remarks

4.1 Related Application: Universal Portfolios

A nice application of the forecast combination problem is the case of unsupervised learning, i.e. when the target is unobservable. The most notable example is the portfolio choice problem, where there is no target sequence and the goal is to maximize wealth. Algorithms that allow us to construct portfolios that have performance comparable to the best constantly rebalanced portfolio with hindsight have attracted the attention of the mathematical finance literature. The results of this paper allows to derive theoretical bounds for this problem. For the sake of concreteness, let $X_t := (X_1, ..., X_K)$ be a vector of relative returns (e.g. closing price divided opening price at time t) and let $-\ln(p) : \mathcal{P} \to \mathbb{R}$ be a loss function. Minimization of the cumulated negative log loss is equivalent to maximization of relative wealth (Breiman, 1961, Cover, 1991, and Samuelson, 1979, for a critique of this criterion). Clearly, our framework works just as well with other loss functions (i.e. any change sign of a utility function), but to better relate to the existing literature in this area, $\ln(p)$ will be used, and the problem is turned into a maximization one. To this end, we state the following.

Theorem 2 *For any sequence of relative returns $(X_t)_{t\in\mathbb{N}}$, and $u_1, ..., u_T \in \mathcal{G}$, using the algorithm in Exhibit 1,*

$$\sum_{t=1}^{T} \ln(\langle w_t, X_t\rangle) \geq \sum_{t=1}^{T} \ln(\langle u_t, X_t\rangle) - error,$$

where

$$error < \frac{T^{\alpha}}{\eta}\left(6\ln(K/\gamma) + [\ln(K/\gamma) + K]\sum_{t=1}^{T-1}|u_t - u_{t+1}|_1\right) + \frac{\eta T^{1-\alpha}}{8}\left(\frac{K}{\gamma}\right)^2$$

and for the constantly rebalanced portfolio, i.e. $u_t = u_{t+1}$, $\forall t$,

$$error < error < 2\frac{K}{\gamma}(T\ln(K/\gamma))^{1/2},$$

choosing $\eta = \frac{\gamma}{K}(48\ln(K/\gamma))^{1/2}$, $\alpha = 1/2$. Hence, the portfolio constructed from Exhibit 1 is universal for portfolio weights in $\mathcal{G}$.

Note that this result is quite strong: no probabilistic assumption is necessary. This result also allows for nonstatioanry portfolios (i.e. time varying rebalanced portfolios) and extends results in the literature on universal portfolios improving the bounds for the best constantly rebalanced case (e.g. Helmbold et al., 1998, gives a bound $O\left(T^{3/4}\right)$ as compared to the $O\left(T^{1/2}\right)$ derived here, though, Cover, 1991, gives a bound $O(\ln T)$, using a different algorithm, which is difficult to implement in practice). However, we require the portfolio weights to be in $\mathcal{G}$ and this is restrictive. The bound in Theorem 2, is somehow loose, especially for the second term outside the parenthesis, hence the derived η might not be the best choice. For the sake of presentation, a tidy bound is preferred. Because of the very low signal to noise ratio in relative returns, the choice of learning rate seems to be crucial. Further study seems to be required to make these algorithms usable in real trading. Empirical results reported in the literature are very encouraging (e.g. Cover, 1991, Helmbold et al., 1998, Györfi et al., 2006), but also very dependent on the dataset used (e.g. Nikandrova, 2005). Empirical results carried out by the author, but not reported here, also suggest that more work is required in the direction of practical implementation, casting doubts on the finite sample performance. Nevertheless, this remains an interesting application that might deserve further study.

4.2 Final Comments

The algorithm in Exhibit 1 allows us to carry out online combination of individual forecasts. The analysis of the algorithm shows that producing forecasts by this method leads, with high probability, to asymptotic optimal performance. However, the experimental results show that one should be somehow cautious about the theoretical results. The experimental performance appeared to be good, but we were unable to perform better than the best time invariant forecast combination. The theoretical results suggested that a bit more could have been achieved.

One short coming of the present procedure is that we do not allow for the number K of forecasts to change over time. This would be particularly useful in the case of survey data (e.g. the Philadelphia' Fed Survey of Professional Forecasters). Nevertheless, time varying combination weights allows us to drop a forecaster and replace it by another. While the total number of forecasts is fixed, the kind of individual forecasts may change over time (at a rate slower than T). Overcoming the problem of a fixed number of forecasts should be paramount for many economic applications. Nevertheless, the fact that the bounds are uniform in T suggests that we could always reset the algorithm every time there is a change in K and still be able to somehow control the cumulative error.

The current procedure uses weights in the unit simplex, hence, forecast combination can be interpreted as model averaging: instead of selecting the best model, we average across them and this seems to lead to good empirical performance (e.g. non-negative garotte in the case of linear regression, Breiman, 1996). However, if forecasts are biased, it is well known that a way to improve is to allow weights to be negative or to use an intercept (e.g. Timmermann, 2004). This means that we would perform some kind of optimization over a set larger than the constrained unit simplex $\mathcal{G}$ in (2). In theory this improves the performance. However, in practice, optimization over a larger set leads to a larger estimation error and possibly poor performance. Optimizing (1) using the square loss and the Euclidean distance for the distance function gives

$$\begin{aligned} w_t' &= w_t - 2\eta_t \left(\langle w_t, X_t\rangle - Y_t\right) X_t \\ w_{t+1} &= w_t' \text{ if } w_t' \in \mathcal{G}_2,\ sw_t'/\left|w_t'\right|_2 \text{ otherwise,} \end{aligned} \tag{4}$$

where $\mathcal{G}_2 := \left\{w \in \mathbb{R}^K, |w|_2 \leq s\right\}$, and s is a prespecified shrinkage parameter. Therefore, the weights are possibly negative but constrained in their sum as in ridge regression. Experimental work carried out by the author, but not reported here, suggested that there could be considerable loss in performance using (4) relatively to Exhibit 1. In particular the performance appeared to be more dependent on the learning rate when (4) was used. Hence, it would be important to investigate extensions that allow combination weights to lie in e.g. $\mathcal{G}_2$ and still lead to good empirical performance. Interestingly, theoretical results in the statistical literature (e.g. Duflo, 1997) suggest the learning rate to be of smaller order than the results discussed here and in the machine learning literature. Hence, an indepth study is required to shed light on the difference both from a theoretical and empirical point of view.

A problem related to iterative procedures is that they do not allow for multiple steps ahead predictions. Often interest lies in h-step ahead predictions and not one step ahead. This does not seem to be contemplated in Exhibit 1. Clearly, we could just change frequency so that one step ahead corresponds to h-step ahead in the original frequency. However, this appears quite wasteful. For this reason, overlapping predictions might be the best alternative, e.g. daily data to make a 5-day ahead prediction on volatility. In this case, we run into problems if we want to use a supervised learning algorithm as in Exhibit 1. We can clearly run 5 separate weakly predictions, one for each working day of the week. This would allow us to apply the algorithm as if we were doing it to five different data series. However, this seems to be intuitively inefficient: why waiting a week to make a weight update when we have data available for losses on other days of the week? This issue needs to be carefully addressed in future research.

5 Proofs

The proof of Theorem 1 will use a few lemmas stated and proved in the next subsection.

Proof of Theorem 1. Note that by Condition 1,

$$g_t := \max_{|y|\leq m_t, |p|\leq m_t} l'(y,p) \leq f(m_t) \tag{5}$$

With probability at least $1-\epsilon$, where $\epsilon := e^{-\tau}$,

$$\begin{aligned}
\sum_{t=1}^{T}[l_t(w_t) - l_t(u_t)] &= \sum_{t=1}^{T}[l_t(w_t) I_{M_t} - l_t(u_t)] \\
&\quad \text{[by Lemma 1]} \\
&\leq \sum_{t=1}^{T}\frac{D(u_t,w_t) - D(u_t,w_t')}{\eta_t} I_{M_t} + \frac{\eta}{8}\sum_{t=1}^{T}\frac{\left|g_t X_t I_{\{|X_t|\leq m_t\}}\right|_\infty^2}{t^\alpha} \\
&\quad \text{[by Lemma 2]}
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{t=1}^{T}\frac{D(u_t,w_t) - D(u_t,w_t')}{\eta_t} I_{M_t} + \frac{\eta}{8}\sum_{t=1}^{T}\frac{[f(m_t(\tau))m_t(\tau)]^2}{t^\alpha} \\
&\quad \text{[by (5) and using the constraint]} \\
&\leq \sum_{t=1}^{T}\frac{D(u_t,w_t) - D(u_t,w_{t+1})}{\eta_t} I_{M_t} + \frac{\eta}{8}\sum_{t=1}^{T}\frac{[f(m_t(\tau))m_t(\tau)]^2}{t^\alpha} \\
&\quad \text{[by Lemma 3]}
\end{aligned}$$

$$\begin{aligned}
&\leq 6\frac{T^\alpha}{\eta}(1+\ln(K/\gamma)) + [\ln(K/\gamma)+K]\frac{T^\alpha}{\eta}\sum_{t=1}^{T}|u_t - u_{t+1}|_1 + \frac{\eta}{8}\sum_{t=1}^{T}\frac{[f(m_t)m_t]^2}{t^\alpha} \\
&\quad \text{[by Lemma 4]} \\
&\leq \frac{T^\alpha}{\eta}\left(6(1+\ln(K/\gamma)) + \sum_{t=1}^{T}[\ln(\gamma/K)+K]|u_t - u_{t+1}|_1\right) + \frac{\eta}{8}[f(m_T(\tau))m_T(\tau)]^2\sum_{t=1}^{T}t^{-\alpha} \\
&\leq \frac{T^\alpha}{\eta}\left(6(1+\ln(K/\gamma)) + \sum_{t=1}^{T-1}[\ln(\gamma/K)+K]|u_t - u_{t+1}|_1\right) + \frac{\eta T^{1-\alpha}}{8}[f(m_t(\tau))m_t(\tau)]^2
\end{aligned}$$

summing over t and noting that there is no prediction at time $T+1$, hence we can choose $u_{T+1} = u_T$ and the last term in the summation (inside the parenthesis) drops. ∎

The proof of the Corollaries is next.

Proof of Corollary 2. By direct computation,

$$\left|\frac{\partial l(y,p)}{\partial p}\right| = 2|y-p| \leq 2(|y|\vee|p|),$$

and Condition 1 applies with $f(m) = 2m$, so that

$$[f(m_t(\tau))m_t(\tau)]^2 = 4m_T(\tau)^4 = O\left([\tau + \ln T]^{4/c_3}\right)$$

and we obtain the first result by application of Corollary 1. For LinEx loss,

$$\left|\frac{\partial l\left(y,p\right)}{\partial p}\right| \leq 2\left|\theta\right| \exp\left\{\left|\theta\right|\left|y-p\right|\right\} \leq 2\left|\theta\right| \exp\left\{2\left|\theta\right|\left(\left|y\right| \vee \left|p\right|\right)\right\}$$

so that, by Condition 1,

$$\begin{aligned}\left[f\left(m_t\left(\tau\right)\right)m_t\left(\tau\right)\right]^2 &= 4\left|\theta\right|^2 \exp\left\{2\left|\theta\right| m_T\left(\tau\right)\right\} m_T\left(\tau\right)^2 \leq 4\left|\theta\right|^2 \exp\left\{4\left|\theta\right| m_T\left(\tau\right)\right\} \\ &= O\left(\exp\left\{4\left|\theta\right|\left[\left(\frac{\tau}{c_2}\right)^{1/c_3} + \left(\frac{\ln T}{c_2}\right)^{1/c_3}\right]\right\}\right) \\ &= o\left(T\exp\left\{4\left|\theta\right|\left(\frac{\tau}{c_2}\right)^{1/c_3}\right\}\right),\end{aligned}$$

using the fact that $4\left|\theta\right|\left(\ln T/c_2\right)^{1/c_3} = o\left(\ln T\right)$ for $c_3 > 1$. The result follows by application of Corollary 1. ■

Proof of Corollary 3. From the proof of Corollary 2,

$$\left[f\left(m_t\left(\tau\right)\right)m_t\left(\tau\right)\right]^2 = O\left(\left[\tau + \ln T\right]^{4/c_3}\right).$$

Hence, solving $T^{\alpha+(1-\epsilon)} = T^{1-\alpha}$ w.r.t. α and substituting back in Theorem 1 gives the result. ■

Proof of Corollary 4. Take expectation on both sides of (6) in the proof of Lemma 1 in the next subsection. The first term on the right hand side is bounded by Theorem 1 and we shall only bound the second term using Holder inequality

$$\sum_{t=0}^{T} \mathbb{E}l\left(Y_t, w_t X_t\right) I_{M_t^c} \leq A \sum_{t=0}^{T} \Pr\left(M_t^c\right)^{1/2}.$$

To bound $\Pr\left(M_t^c\right)$, note that the event $\left\{\max_{1\leq k\leq K}\left|X_{tk}\right| \leq m_t\right\}$ implies the event $\left\{\left|\hat{P}_t\right| \leq m_t\right\}$ (see the proof of Lemma 1). Hence, by the arguments in the proof of Lemma 1,

$$\begin{aligned}\Pr\left(M_t^c\right) &\leq \Pr\left(\left|Y_t\right| > m_t\right) + \Pr\left(\max_{1\leq k\leq K}\left|X_{tk}\right| > m_t\right) \\ &\leq \left(K+1\right)c_1 \exp\left\{-c_2 m_t^{c_3}\right\} \\ &= t^{-1}\end{aligned}$$

choosing

$$m_t := \left[\frac{1}{c_2}\ln\left(c_1\left(K+1\right)t\right)\right]^{1/c_3}.$$

Hence,

$$\mathbb{E}l\left(Y_t, w_t X_t\right) I_{M_t^c} \leq A \Pr\left(M_t^c\right)^{1/2} \leq A\sqrt{t^{-1}}.$$

Summing over t gives the result. ■

Proof of Theorem 2. By a sign change we can consider $-\ln(p)$, as for the previous results. Hence, follow the proof of Theorem 1 in the following order: apply Lemma 2 with no need to apply Lemma 1 first. Then, Lemma 2 gives a bound in terms of $|l'(y,\langle w,x\rangle)x|_\infty^2$, which, for the specific choice of loss function corresponds to

$$\max_k \left|\frac{x_{tk}}{\langle w_t, x_t\rangle}\right|^2 \leq \max_k \left|\frac{x_{tk}}{(\gamma/K)\, x_{tk}}\right|^2 \leq \left|\frac{K}{\gamma}\right|^2$$

using the constraint. Then, the proof proceeds as for Theorem 1 and the Corollaries. ■

5.1 Technical Lemmas

These lemmas are used in the proof of Theorem 1.

Lemma 1 *Define* $m_t = \left[\frac{1}{c_2}\ln\left(\frac{(K+1)c_1 t}{\epsilon}\right)\right]^{1/c_3}$ *and the set*

$$M_t := \left\{|Y_t| \leq m_t, \left|\hat{P}_t\right| \leq m_t\right\}.$$

Under Condition 1 and 2, with probability at least $1-\epsilon$*,*

$$\sum_{t=1}^T l\left(Y_t, \hat{P}_t\right) = \sum_{t=1}^T l\left(Y_t, \hat{P}_t\right) I_{M_t}.$$

Proof. Let M_t^c be the complement of M_t. Consider the following identity

$$\begin{aligned} \sum_{t=1}^T l\left(Y_t, \hat{P}_t\right) &= \sum_{t=1}^T l\left(Y_t, \hat{P}_t\right) I_{M_t} + \sum_{t=1}^T l\left(Y_t, \hat{P}_t\right) I_{M_t^c} \qquad (6)\\ &= \mathrm{I} + \mathrm{II}. \end{aligned}$$

Note that

$$\begin{aligned} \Pr\left(\left|\hat{P}_t\right| > m_t\right) &\leq \Pr\left(|w_t|_1 \max_{1\leq k\leq K} |X_{tk}| > m_t\right) \\ &= \Pr\left(\max_{1\leq k\leq K} |X_{tk}| > m_t\right) \qquad (7)\\ \text{[because } w_t &\in \mathcal{S}^K] \\ &\leq \sum_{k=1}^K \Pr\left(|X_{tk}| > m_t\right), \end{aligned}$$

by the union bound, so that the event $\{\max_{1\leq k\leq K}|X_{tk}|\leq m_t\}$ implies the event $\left\{\left|\hat{P}_t\right|\leq m_t\right\}$. Hence,

$$\begin{aligned}
\Pr\left(\sum_{t=1}^{T} l_t\left(Y_t,\hat{P}_t\right)I_{M_t^c}>0\right) &\leq \Pr\left(\max_{1\leq t\leq T}|Y_t|>m_t\right)+\Pr\left(\max_{1\leq t\leq T}\left|\hat{P}_t\right|>m_t\right)\\
&\leq \sum_{t=1}^{T}\left[\Pr\left(|Y_t|>m_t\right)+\sum_{k=1}^{K}\Pr\left(|X_{tk}|>m_t\right)\right]\\
&\quad\text{[by (7) and the union bound]}\\
&\leq (K+1)\,c_1\sum_{t=1}^{T}\exp\left\{-c_2m_t^{c_3}\right\}\\
&\quad\text{[by Condition 2]}\\
&= \epsilon \text{ choosing } m_t=\left[\frac{1}{c_2}\ln\left(\frac{(K+1)\,c_1 t}{\epsilon}\right)\right]^{1/c_3}.
\end{aligned}$$

Hence with this choice of m_t, the result holds with probability at least $1-\epsilon$. ■

The next follows from Lemma 2 in Cesa-Bianchi (1999). A sketch of proof is provided for the sake of completeness and also because it is a key ingredient.

Lemma 2 *For any $y\in\mathbb{R}$, $x\in\mathbb{R}^K$, and $\eta>0$, $u\in\mathcal{S}^K$, $w\in int\left(\mathcal{S}^K\right)$ and*

$$w_k':=\frac{w_k\exp\left\{-\eta l'\left(y,\langle w,x\rangle\right)x_k\right\}}{\sum_{k=1}^{K}w_k\exp\left\{-\eta l'\left(y,\langle w,x\rangle\right)x_k\right\}},$$

the following holds

$$l\left(y,\langle w,x\rangle\right)-l\left(y,\langle u,x\rangle\right)\leq\frac{D\left(u,w\right)-D\left(u,w'\right)}{\eta}+\eta\frac{\left|l'\left(y,\langle w,x\rangle\right)x\right|_{\infty}^{2}}{8}$$

Proof. Define $l'(y,p):=(\partial/\partial p)\,l(y,p)$ and $l''(y,p)$ for the second order derivative with respect to p. Hence,

$$\begin{aligned}
l\left(y,\langle w,x\rangle\right)-l\left(y,\langle u,x\rangle\right) &= \langle w-u,x\rangle\,l'\left(y,\langle w,x\rangle\right)-\langle w-u,x\rangle^{2}\,l''\left(y,\langle w,x\rangle\right)\\
&\leq \langle w-u,x\rangle\,l'\left(y,\langle w,x\rangle\right) \qquad (8)
\end{aligned}$$

by convexity of $l\left(y,p\right)$ with respect to p. Recall that $D\left(w,u\right):=\sum_{k}w_{k}\ln\left(w_{k}/u_{k}\right)$, and define $z_{k}:=-\eta l'\left(y,\langle w,x\rangle\right)x_{k}$. Then,

$$\begin{aligned}
D\left(u,w\right)-D\left(u,w'\right) &= \sum_{k=1}^{K}u_{k}\ln\left(w_{k}'/w_{k}\right)\\
&= \sum_{k=1}^{K}u_{k}\left[\ln\frac{w_{k}\exp\left\{z_{k}\right\}}{\sum_{l=1}^{K}w_{l}\exp\left\{z_{l}\right\}}-\ln w_{k}\right]\\
&\geq \sum_{k=1}^{K}u_{k}z_{k}-\ln\left(\sum_{l=1}^{K}w_{l}\exp\left\{z_{l}\right\}\right)\\
&\quad \text{[because }-\ln w_{k}\text{ is non-negative]}\\
&= \sum_{k=1}^{K}\left(u_{k}-w_{k}\right)z_{k}-\ln\left(\sum_{l=1}^{K}w_{l}\exp\left\{z_{l}-\sum_{k=1}^{K}w_{k}z_{k}\right\}\right)\\
&\geq \sum_{k=1}^{K}\left(u_{k}-w_{k}\right)z_{k}-\frac{\left|z\right|_{\infty}^{2}}{8},
\end{aligned}$$

using Hoeffding's inequality (e.g. Devroye et al., 1996) for the moment generating function of bounded random variables in the last step. Substituting in (8), we obtain

$$\langle w-u,x\rangle l'\left(y,\langle w,x\rangle\right)\leq\frac{D\left(u,w\right)-D\left(u,w'\right)}{\eta}+\eta\frac{\left|l'\left(y,\langle w,x\rangle\right)x\right|_{\infty}^{2}}{8}$$

implying the last result. ■

The next is Corollary 3 in Herbster and Warmuth (2001). A concise proof for general "distance" functions can be found in Lemma 11.3 of Cesa-Bianchi and Lugosi (2006).

Lemma 3 *For any vector $u_{t}\in\mathcal{G}$ ($\mathcal{G}$ as in (2)), $w_{t+1}=\arg\min_{v\in\mathcal{G}}D\left(v,w_{t}'\right)$ (i.e. w_{t+1} as in Exhibit 1) and $w_{t}'\in int\left(\mathcal{S}^{K}\right)$,*

$$D\left(u_{t},w_{t}'\right)\geq D\left(u_{t},w_{t+1}\right).$$

For $u\in\mathcal{S}^{K}$ and $w\in$ *interior* $\left(\mathcal{S}^{K}\right)$, define

$$D_{F}\left(u,w\right)=F\left(u\right)-F\left(w\right)-\langle\left(u-w\right),\nabla F\left(w\right)\rangle.$$

We note that when $F\left(u\right):=\sum_{k=1}^{K}u_{k}\left(\ln u_{k}-1\right)$, $D_{F}\left(u,w\right)=D\left(u,w\right)$. In order to remind us of the above decomposition, we will use $D_{F}\left(u,w\right)$ instead of $D\left(u,w\right)$ in the last two lemmas. Different choices of F, would lead to different "distance" functions and different updates via (1) (e.g. Cesa-Bianchi and Lugosi, 2006, Ch.11, for details).

Lemma 4 *For* $u_t \in \mathcal{S}^K$, η_t *and* w_t *as in Exhibit 1* $(t = 1, ..., T)$,

$$\begin{aligned} & \sum_{t=1}^{T} \frac{D_F(u_t, w_t) - D_F(u_t, w_{t+1})}{\eta_t} \\ \leq \quad & 6\frac{T^\alpha}{\eta}[1 + \ln(K/\gamma)] + [\ln(K/\gamma) + K]\frac{T^\alpha}{\eta}\sum_{t=1}^{T}|u_t - u_{t+1}|_1 . \end{aligned}$$

Proof. To prove the Lemma we need to find a telescoping sum. To this end

$$\begin{aligned} \frac{D_F(u_t, w_t) - D_F(u_t, w_{t+1})}{\eta_t} &= \frac{D_F(u_t, w_t) - D_F(u_{t+1}, w_{t+1})}{\eta_t} - \frac{D_F(u_t, w_{t+1}) - D_F(u_{t+1}, w_{t+1})}{\eta_t} \\ & \quad [\text{adding and subtracting } D_F(u_{t+1}, w_{t+1})] \\ &= \mathrm{I} + \mathrm{II}. \end{aligned}$$

We shall deal with the sum over t of each term separately.

Sum over I.

$$\begin{aligned} \sum_{t=1}^{T}\mathrm{I} &= \frac{1}{\eta_1}D_F(u_1, w_1) - \frac{1}{\eta_T}D_F(u_{T+1}, w_{T+1}) - \sum_{t=1}^{T-1}D_F(u_{t+1}, w_{t+1})\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t+1}}\right) \\ &= \frac{1}{\eta}D_F(u_1, w_1) - \frac{T^\alpha}{\eta}D_F(u_{T+1}, w_{T+1}) - \sum_{t=1}^{T-1}D_F(u_{t+1}, w_{t+1})\frac{t^\alpha - (t+1)^\alpha}{\eta} \\ & \quad [\text{by definition of } \eta_t] \\ &\leq \frac{1}{\eta}D_F(u_1, w_1) + \max_{1\leq t\leq T-1}D_F(u_{t+1}, w_{t+1})\sum_{t=1}^{T-1}\frac{(t+1)^\alpha - t^\alpha}{\eta} \\ & \quad [\text{using the fact that } D_F \text{ is non-negative}] \\ &= \frac{1}{\eta}D_F(u_1, w_1) + \max_{1\leq t\leq T-1}D_F(u_{t+1}, w_{t+1})\frac{T^\alpha}{\eta} \\ & \quad [\text{because the sum telescopes}] \\ &\leq 2\max_{1\leq t\leq T}D_F(u_t, w_t)\frac{T^\alpha}{\eta}. \end{aligned}$$

Sum over II.

By definition of D_F,

$$-D_F(u_t, w_{t+1}) + D_F(u_{t+1}, w_{t+1}) = F(u_{t+1}) - F(u_t) + \langle (u_t - u_{t+1}), \nabla F(w_{t+1}) \rangle .$$

Hence,

$$\begin{aligned} \mathrm{II} &= \frac{F(u_{t+1}) - F(u_t)}{\eta_t} + \frac{\langle (u_t - u_{t+1}), \nabla F(w_{t+1}) \rangle}{\eta_t} \\ &= \mathrm{IV} + \mathrm{V}. \end{aligned}$$

Note that $F(u) \in [0,-1]$ for $u \in [0,1]$. Hence, summing over IV,

$$\begin{aligned}
\sum_{t=1}^{T} \mathrm{IV} &= -\sum_{t=1}^{T} \frac{F(u_t) - F(u_{t+1})}{\eta_t} \\
&= -\frac{1}{\eta} F(u_1) + \frac{1}{\eta_T} F(u_{T+1}) + \sum_{t=1}^{T-1} F(u_{t+1}) \frac{t^\alpha - (t+1)^\alpha}{\eta} \\
&\le -\frac{1}{\eta} F(u_1) + \max_{1 \le t \le T-1} [-F(u_{t+1})] \sum_{t=1}^{T-1} \frac{(t+1)^\alpha - t^\alpha}{\eta} \\
&\quad \text{[because } F(u) \text{ is non-positive]} \\
&\le -\frac{1}{\eta} F(u_1) + \max_{1 \le t \le T-1} [-F(u_{t+1})] \frac{T^\alpha}{\eta} \\
&\quad \text{[because the sum telescopes]} \\
&\le 2 \max_{1 \le t \le T} [-F(u_t)] \frac{T^\alpha}{\eta}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\sum_{t=1}^{T} \mathrm{V} &\le \frac{T^\alpha}{\eta} \sum_{t=1}^{T} \left| \langle (u_t - u_{t+1}), \nabla F(w_{t+1}) \rangle \right| \\
&\le \frac{T^\alpha}{\eta} \sum_{t=1}^{T} |u_t - u_{t+1}|_1 \, |\nabla F(w_{t+1})|_\infty .
\end{aligned}$$

Hence

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{D_F(u_t, w_t) - D_F(u_t, w_{t+1})}{\eta_t} \\
\le\ & 2\frac{T^\alpha}{\eta} \max_{1 \le t \le T} D_F(u_t, w_t) + 2\frac{T^\alpha}{\eta} \max_{1 \le t \le T} [-F(u_t)] + \frac{T^\alpha}{\eta} \sum_{t=1}^{T} \left| \langle (u_t - u_{t+1}), \nabla F(w_{t+1}) \rangle \right| \\
=\ & 2\frac{T^\alpha}{\eta} \max_{1 \le t \le T} \left(F(u_t) - F(w_t) - \langle (u_t - w_t), \nabla F(w_t) \rangle \right) \\
& +2\frac{T^\alpha}{\eta} \max_{1 \le t \le T} [-F(u_t)] + \frac{T^\alpha}{\eta} \sum_{t=1}^{T} |u_t - u_{t+1}|_1 \, |\nabla F(w_t)|_\infty \\
& \text{[by definition of } D_F(u_t, w_t) \text{]} \\
\le\ & 2\frac{T^\alpha}{\eta} \max_{1 \le t \le T} \langle (w_t - u_t), \nabla F(w_t) \rangle + 2\frac{T^\alpha}{\eta} \max_{1 \le t \le T} [-F(u_t)] \\
& +2\frac{T^\alpha}{\eta} \max_{1 \le t \le T} [-F(w_t)] + \frac{T^\alpha}{\eta} \sum_{t=1}^{T} |u_t - u_{t+1}|_1 \, |\nabla F(w_t)|_\infty \\
=\ & : \mathrm{VI}.
\end{aligned}$$

Using Lemma 5 (stated next) the above display is bounded by

$$\mathrm{VI} \leq 2\frac{T^{\alpha}}{\eta}\ln\left(K/\gamma\right) + 4\frac{T^{\alpha}}{\eta}\left[\ln\left(K/\gamma\right) + 1\right] + \left[\ln\left(K/\gamma\right) + K\right]\frac{T^{\alpha}}{\eta}\sum_{t=1}^{T}\left|u_t - u_{t+1}\right|_1$$

hence the result. ■

Using the constraint $\mathcal{G}$, we can derive the following bounds used above.

Lemma 5 *For* $F\left(u\right) := \sum_{k=1}^{K} u_k\left(\ln u_k - 1\right)$ *and* $u,w \in \mathcal{G}$

$$\left\langle\left(w_t - u_t\right), \nabla F\left(w_t\right)\right\rangle \leq \ln\left(K/\gamma\right),$$

$$\left|u_t - u_{t+1}\right|_1\left|\nabla F\left(w_{t+1}\right)\right|_{\infty} \leq \left[\ln\left(K/\gamma\right) + K\right]\left|u_t - u_{t+1}\right|_1,$$

and

$$F\left(u_t\right) \leq \ln\left(\gamma/K\right) - 1$$

Proof. By direct calculation,

$$\begin{aligned}\left\langle\left(w_t - u_t\right), \nabla F\left(w_t\right)\right\rangle &= \sum_{k=1}^{K}\left(w_{tk} - u_{tk}\right)\left[\ln w_{tk} + \left(K - 1\right)\right] \\ &= \sum_{k=1}^{K}\left(w_{tk} - u_{tk}\right)\ln w_{tk} \\ &\quad \text{[because the coefficients add up to one]} \\ &\leq -\sum_{k=1}^{K} u_{tk}\ln w_{tk} \\ &\quad \text{[because the first term is negative]} \\ &\leq -\ln\left(\gamma/K\right),\end{aligned}$$

using the constraint. Finally,

$$\left|u_t - u_{t+1}\right|_1\left|\nabla F\left(w_{t+1}\right)\right|_{\infty} \leq \left[-\ln\left(\gamma/K\right) + K - 1\right]\left|u_t - u_{t+1}\right|_1,$$

using the constraint. ■

References

Aiolfi, M. and A. Timmermann (2004) Persistence in Forecasting Performance and Conditional Combination Strategies. Forthcoming in Journal of Econometrics.

Breiman, L. (1961) Optimal Gambling Systems for Favorable Games. Proceeding 4th Berkeley Symposium on Mathematical Statistics and Probability 1, 65-78.

Breiman, L. (1996) Heuristics of Instability and Stabilization in Model Selection. Annals of Statistics, 24, 2350-2383.

Cesa-Bianchi, N. (1999) Analysis of Two Gradient-Based Algorithms for Online Regression. Journal of Computer and System Sciences 59, 392-411.

Cesa-Bianchi, N. and G. Lugosi (2006) Prediction, Learning , and Games. Cambridge: Cambridge University Press.

Cheng, M.-Y., J. Fan and V. Spokoiny (2003) Dynamic Nonparametric Filtering with Application to Volatility Estimation. In M.G. Akritas and D.N. Politis (eds.), Recent Advances and Trends in Nonparametric Statistics, Elsevier (North Holland), 315-333.

Cover, T.M. (1968) Estimation by the Nearest Neighbor Rule. IEEE Transactions on Information Theory 14, 21-27.

Cover, T. (1991) Universal Portfolios. Mathematical Finance 1, 1-29.

Deutsch, M., C.W.J. Granger and T. Teräsvirta (1994) The Combination of Forecasts Using Changing Weights. International Journal of Forecasting 10, 47-57.

Devroye, L., L. Györfi and G. Lugosi (1996) A Probabilistic Theory of Pattern Recognition. New York: Springer.

Diebold, F.X. and P. Pauly (1990). The Use of Prior Information in Forecast Combination. International Journal of Forecasting 6, 503-508.

Ding, Z., C.W.J. Granger (1996) Modeling Volatility Persistence of Speculative Returns: A New Approach. Journal of Econometrics 73, 185-215.

Ding, Z., C.W.J. Granger and R.F. Engle (1993) A Long Memory Property of Stock Market Returns and a New Model. Journal of Empirical Finance 1, 83-106.

Duflo, M. (1997) Random Iterative Models. Berlin: Springer.

Elliott, G. and A. Timmermann (2004) Optimal Forecast Combinations under General Loss Functions and Forecast Error Distributions. Journal of Econometrics 122, 47-79.

Fan, J. and Gu, J. (2003). Semiparametric Estimation of Value-at-Risk. Econometrics Journal 6, 261-290.

Friedman, J.H. (2001)Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics 29, 1189-1232.

Györfi, L., G. Lugosi and F. Udina (2006) Nonparametric Kernel-Based Sequential Investment Strategies. Mathematical Finance 16, 337-358.

Kivinen, J. and M.K. Warmuth (1997) Exponentiated Gradient versus Gradient Descent for Linear Predictors. Information and Computation 132, 1-63.

Kohler, M., Krzyżak, A. and H. Walk (2006) Rates of Convergence for Partitioning and Nearest Neighbor Regression Estimates with Unbounded Data. Journal of Multivariate Analisys 97, 311-323.

Helmbold, D.P., R.E. Schapire, Y. Singer and M.K. Warmuth (1998) Online Portfolio Selection using Multiplicative Updates. Mathematical Finance 8, 325-347.

Hendry, D.F. and M.P. Clements (2004) Pooling of Forecasts. Econometrics Journal 7, 1-31.

Herbster, M. and M.K. Warmuth (2001) Tracking the Best Linear Predictor. Journal of Machine Learning Research 1, 281-309.

Mercurio, D. and V. Spokoiny (2004) Statistical Inference for Time-Inhomogeneous Volatility Models. Annals of Statistics 32, 577-602.

Nikandrova, A. (2005) Universal Portfolios Selection from a Practical Perspective. Master Dissertation, Faculty of Economics, University of Cambridge.

Ripley, B.D. (1996) Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press.

Samuelson, P.A. (1979) Why We Should Not Make Mean Log of Wealth Big Though Years to Act Are Long. Journal of Banking and Finance 3, 305-307.

Sancetta, A. (2006) Online Forecast Combinations of Distributions: Worst Case Bounds. Forthcoming in Journal of Econometrics.

Stock, J.H. and M.W. Watson (2004) Combination Forecasts of Output growth in a Seven-Country Data Set. Journal of Forecasting 23, 405-430.

Timmermann, A. (2004) Forecast Combinations. Forthcoming in G. Elliott, C.W.J Granger and A. Timmermann (eds.) Handbook of Economic Forecasting. North Holland.

Yang, Y. (2004) Combining Forecasting Procedures: Some Theoretical Results. Econometric Theory 20, 176-222.

Zivot, E. and J. Wang (2005) Modeling Financial Time Series with S-PLUS®. New York: Springer.