Robust Combinations of Statistical Procedures

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xiaoqiao Wei

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advisor: Dr. Yuhong Yang

November, 2010

Acknowledgements

I am deeply grateful to my adviser, Professor Yuhong Yang, for helping find the research topics, for guiding me through the conduction of the dissertation research, for reading and commenting on numerous drafts, for patiently revising my writing, and for his help and advise in many aspects of my life. Thank you, Professor Yang, for helping me arrive at this moment.

I am thankful to Professor Dennis Cook for his serving as my defense committee chair. My thanks also go to Professor Hui Zou and Professor Baolin Wu for having given their time and expertise for reviewing this thesis, and for their valuable suggestion on my research.

I am grateful to the School of Statistics, which has provided me the support and great environment for my professional development. I thank every member of the faculty, staff and students in the School of Statistics for their teaching, service, friendship and help, and for one of the most meaningful time periods in my life.

Finally, I thank my parents, who are the source of my strength and coverage. I thank my wife Minfen and son Alan, for their love and support, and for having always been there with me.

Dedication

Dedicated to my parents: Jichuan Wei and Lianmei Tao.

Abstract

Forecast outliers commonly occur in economic, financial, and other areas of application. In the literature of forecast combinations, there have been only a few studies exploring how to deal with outliers. In the first part of the dissertation, we propose two robust combining methods, which build on the AFTER algorithm by using robust loss functions for reducing influence of outliers. Oracle inequalities for certain versions of these methods are obtained, which show that the combined forecasts automatically perform as well as the best individual among the pool of forecast candidates. Systematic simulations and data examples show that the robust methods outperform AFTER when outliers are likely to occur and perform similarly to AFTER when there are no outliers. Comparison of the robust AFTERs with some commonly used combining methods also shows their potential advantages.

Model selection is a fundamental problem in Statistics. Many model selection methods have been proposed. However, it is often very difficult to judge which one is the best to use for the given data. In the second part of the dissertation, for the purpose of estimating the regression function or prediction, we propose a method, l_1-ARM, to robustly combine model selection methods, which performs well adaptively. In numerical work, we consider the LASSO, SCAD, and adaptive LASSO in representative scenarios as well as cases of randomly generated models. The l_1-ARM automatically performs like the best among them and consequently provides a better estimation/prediction in an overall sense, especially when outliers are likely to occur.

Least squares combinations are an important development in the forecast combination literature. However, ordinary least squares methods often perform poorly in real application due to the variability of coefficient/weight estimations. In the third part of the dissertation, we propose a novel method to simultaneously stabilize and shrink the coefficient/weights estimates. The proposed methods can be applied to various combination methods to improve prediction as long as their weights are determined based on ordinary least squares.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Model selection with a number of predictors has been one of the major research areas of Statistics for decades. Many selection methods have been proposed with their own advantages, but it is true that none of the selection methods can uniformly outperform the others. Thus this brings a challenge to a statistics user: How should he/she select a model selection method with his/her statistical applications?

When the goal is to estimate the regression function, alternative to the use of a single model based on a model selection method, model combination considers a group of candidate models, and produces a combined result of these individual models to share their strengths. There are two directions of model combinations in the literature: combining for improvement and combining for adaptation (see, Yang, 2004). The first direction aims to outperform all of the individual candidates, while the second one is to behave like the best candidate. Juditsky and Nemirovski (2000), Yang (2004) and Tsybakov (2003) have showed that the aggressive first direction comes with a high cost, that is, its convergence rate is substantially slower than in the second direction. Thus unless there is a substantial gain in bias reduction by combining the candidate models, it is more applicable to combine for adaptation in real applications to share the strengths of the individual candidates with a low cost.

In the direction of combining for adaptation, Yang (2001, 2004) proposed an ARM algorithm for regression analysis and an AFTER algorithm for forecast combinations. In

these algorithms, the combining weights are apportioned according to the relative prediction/forecast errors of the candidate models in an exponential form, and the quadratic loss there makes it sensitive to prediction/forecast outliers. In this dissertation, we propose robust combination of statistical procedures. The theoretical properties of the robust combination methods are obtained, which show that the combined result automatically performs as well as the best one among the candidate models. Systematic simulations and data examples show that the robust methods outperform square-loss-based ARM/AFTER when outliers are likely to occur and perform similarly to them when there are no outliers. Comparison of the robust methods with some commonly used combining and selection methods also shows their advantages.

In Chapter 2, we address robust forecast combinations. The last three decades have seen an exciting amount of research and application of forecast combinations, producing thousands of research articles. A number of methods have been proposed, including those based on optimization under variance-covariance estimation of the forecast candidates (e.g., Bates and Granger, 1969), Bayesian methods (e.g., Min and Zellner, 1993), regression on the forecast candidates (e.g., Granger and Ramanathan, 1984), and methods that take into account structural breaks (e.g., Timmermann, 2006). Most combining methods passively take outliers into combinations, which may have detrimental effects. A few trimming or screening and robust regression methods have been proposed to remove or reduce the effects of the worst models (see, e.g., Timmermann, 2006). Since in many practical situations, forecast outliers commonly occur, it is important to have theoretically proven forecast combining methods that are robust to outliers.

To effectively deal with possible outliers, we propose two new weighting forms, L_1-AFTER and h-AFTER. In the extensive simulations for fair and informative comparison results, we see that the new methods perform very close to AFTER under normal error assumptions, but better or much better than AFTER when outliers are likely to occur. In addition, the new methods provide superior forecasts not only in terms of the mean and median, but also of some other quantiles. Two real data examples are provided to demonstrate the advantage of the new methods. A nice feature of the robust approach is that the gain in accuracy for situations with outliers is accompanied by only a slight

damage of performance when no outliers occur. This suggests that L_1-AFTER and h-AFTER can be generally applied. On the theoretical side, we show that the combined forecasts have an attractive oracle property that guarantees their performances to be close to the best candidate under some conditions that do not require model-specific natures of the candidates. It is worth pointing out that the non-quadratic nature of the robust losses for L_1- and h-AFTER makes the derivation of the oracle inequalities significantly different from that for the original AFTER.

In Chapter 3, we address robust combination of model selection methods for prediction. Many successes of combining different predictions in real applications have prompted interests on combining statistical procedures from a practical perspective. For instance, in the well-known Netflix competition, ensemble of different methods is a key idea employed by top teams (see, e.g., http://www.netflixprize.com/leaderboard). The previous theoretically proven combining methods in e.g., Yang (2001) and Catoni (2004), use quadratic-type loss in determining weights for the candidates and show that the combined regression estimator achieves the best performance offered by the candidates in an accumulated risk. The quadratic-type of loss is also used in combining methods by e.g., Juditsky and Nemirovski (2000), and Tsybakov (2003) for larger target classes of combinations.

The mathematically convenient quadratic loss for weighting regression estimators works very well under Gaussian noise. However, when the noise has a heavier tail, as commonly occurs in reality, a few outliers often destabilize the weights. A robust combination of estimates or predictions is thus sought. We propose a robust method, called l_1-ARM, to combine regression estimates/predictions obtained by a list of model selection methods. More specifically, the quadratic loss in the ARM is replaced by absolute loss, and an oracle risk bound is presented that also allows a screening step to be incorporated to remove poor model selection methods, which can be very helpful when a large number of methods are considered. In our numerical work, we focus on combining the LASSO, SCAD, and adaptive LASSO as applied in the linear regression setting.

In Chapter 4, we address regression based forecast combination methods. Granger and Ramanathan (1984) expanded the early developed variance-covariance forecast combination methods into a regression framework. Since then, many regression based forecast combination methods have been proposed in the literature. For example, Diebold (1988)

considered serial correlation in the least squares framework. Coulson and Robins (1993) included a lagged dependent variable besides the forecast candidates. Deutsch et al. (1994) addressed regime switches when estimating coefficients/weights. Recently, researchers have worked on forecast combinations of a large number of forecasts in hope to take advantages of many different sources or models (e.g., Stock & Watson, 2004). It has been shown, however, that the ordinary regression combination is not optimal for this kind of scenarios due to high variance. The empirical evidence of poor performances of the large-number regression combinations is provided by Rapach & Strauss (2008), among many others.

We propose a novel regression based combination method, the decreasingly averaging method. The decreasingly averaging method retains all forecast candidates, but simultaneously stabilizes and slowly shrinks their coefficients/weights according to their order of appearance in the process of sequential selections. The less significant the candidate, the more to be shrunk, thus the less effect on the combined forecasts. This method can be easily implemented. Actually, they can be tools to help other combination methods improve prediction accuracy especially in the large-number combination cases as long as their weights are determined based on ordinary least squares.

In Chapter 5, we summarize the main chapters of the dissertation, and briefly discuss some directions of the future research.

Chapter 2

Robust Forecast Combinations

2.1 Introduction

2.1.1 Forecast combinations

The last three decades have seen an exciting amount of research and application of forecast combinations, producing thousands of research articles (see, e.g., Clemen (1989) and Timmermann (2006) for reviews of articles mostly published in forecasting and econometrics journals). A number of methods have been proposed, including those based on optimization under variance-covariance estimation of the forecast candidates (e.g., Bates and Granger, 1969), Bayesian methods (e.g., Min and Zellner, 1993), regression on the forecast candidates (e.g., Granger and Ramanathan, 1984), and methods that take into account structural breaks (e.g., Timmermann, 2006).

Applications have led to interesting but also somewhat perplexing results (e.g., Palm and Zellner, 1992). For instance, it has been observed again and again that combinations of forecasts beat the individual ones, but it has also been reported that simple methods, such as simple averaging of the candidates, worked better than complicated alternatives that were intended to improve them. When attempting to summarize the empirical findings, the fact that the positive stories do not generally hold, together with a lack of theoretical understanding of the issues involved, makes it very difficult for a forecaster to have a

well-guided choice among many different ways to take advantage of multiple forecasts.

Indeed, to the best of our knowledge, the publications on application of forecast combinations so far give little hint when one combination method should be preferred to another (on choosing between model selection and model combination with time series models, Zou and Yang, 2004 and Chen and Yang, 2007 showed that instability measures play an important role). Without a useful statistical characterization of when the simple averaging method works well, its better or even much better performance on some data sets over a linear regression method (say) should not be taken as a strong vote for simple averaging because clearly there are various situations where simple averaging can fail miserably. For instance, if one candidate has a non-vanishing bias while the others perform very well, then simple averaging cannot avoid a non-vanishing bias unless the number of candidate forecasts is large.

2.1.2 Non-asymptotic theoretical developments

In the past decade or so, non-asymptotic theoretical results on combination/selection of forecasts or models have emerged in statistics and machine learning literature. Some of the results yield theoretical insight on performance of estimation/prediction from model selection or model combination based on finite dimensional models (see, Barron, Birgé and Massart, 1999; Yang, 2001 for references), and others emphasize performance bounds (in terms of statistical risk) that do not depend on assumptions on specific nature of the candidate forecasts or models except boundedness of the difference between the conditional mean of the observation and the forecasts. In recent years, the latter has generated a lot of excitement in the learning community (see, section 9 of Birgé, 2006). First, the candidates can be anything that satisfies mild non-model-specific conditions and thus allow almost arbitrary forecasts to be combined. Second, the combined forecast has a certain optimal performance characterized by an oracle inequality or an index of resolvability bound, which shows that the combined forecast achieves the best performance in a rigorous non-asymptotic mathematical framework (see, e.g., Yang, 2004a; Bunea, Tsybakov and Wegkamp, 2007; Sancetta, 2007 and references therein).

Despite an apparent lack of interaction between the applied forecasters and the theoreticians (statisticians/machine learners), some main research findings on application and theoretical sides of forecast combinations have interesting connections. For example, the phenomenon that complicated combining methods often work much worse than the best individual is closely related to the theoretical finding that pursuing the best linear or convex combination of a number of candidates has an unavoidable high price in terms of forecasting accuracy from a minimax perspective (see, Yang, 2004a for results and references). Indeed, there are two directions of combining forecasts: one for the purpose of behaving like the best candidate (which is of course unknown), called combining for adaptation, and the other for beating all of the individuals, called combining for improvement. The first direction includes typical Bayesian model averaging methods, AIC model averaging (Buckland et al, 1997) and AFTER (Yang, 2004a); and the second includes variance-covariance based and regression based methods, and machine learning and related methods (see e.g., Cesa-Bianchi and Lugosi (2006) for results and references and Sancetta (2010)).

It turns out that a procedure that tries to achieve the performance of best convex or linear combination of the candidate forecasts necessarily has a much higher difficulty due to estimating the best combination coefficients than that for achieving the best individual performance. More specifically, Juditsky and Nemirovski (2000), Yang (2004b) and Tsybakov (2003) showed that the convergence rate in the second direction need necessarily be substantially slower than in the first direction in a suitable uniform sense. Even under unbiasedness assumption of the forecast candidates, opposite to its intention, the variability of the combined forecast by the variance-covariance methods can actually be much larger than the best candidate due to difficulty in estimating the optimal coefficients, especially when the forecasts are highly dependent. These results clearly suggest that one must be careful when choosing between the two directions of forecast combinations. When candidate forecasts are based on time series models of similar nature (such as ARIMA) or regression models from different choice of predictors, it is usually the case that combining for adaptation is the proper goal and combining for improvement is unnecessarily aggressive because linear combination of the candidate models does not lead to reduction of modeling bias in such situations. However, when the best time series model or the best expert in an on-line prediction problem changes over time, gradient-based algorithms designed in

the direction of combining for improvement can bring in significant advantages. See, e.g., Haussler et al (1998), Vovk (1998), Cesa-Bianchi and Lugosi (2006), Sancetta (2010) for very interesting results and more references.

The AFTER algorithm, aiming at combining for adaptation, possesses an optimality in risk under some conditions. When applying AFTER, one does not need to know or estimate the covariances between the individual forecasts. Simulations and real data examples have demonstrated the advantages of AFTER when common model selection or hypothesis testing methods are unstable to choose the best model (Zou & Yang, 2004, and Chen & Yang, 2007). Some applications of AFTER in the literature include Altavilla & Grauwe (2006), Rapach & Strauss (2005, 2008), Sánchez (2008), and Fan, Chen and Lee (2008).

2.1.3 Issues that motivate this study

A forecast is called an outlier when the forecast error is much larger than typical ones. There are two sources that cause unusually large forecast errors: spikes in observations and severe errors by forecasters. When the data generating process has structural breaks (e.g., Pesaran and Timmermann, 2007), outliers are more likely to occur. The AFTER algorithm apportions the combining weights according to the relative forecast errors of the candidate models in an exponential form, and the quadratic loss there makes it sensitive to forecast outliers. Indeed, when the AFTER algorithm encounters an outlier from an overall greatly performing forecasting model, the following weight of the model is dramatically shrunk down. Moreover, the influence of the outlier may last through several stages before the weight comes back. Most combining methods passively take outliers into combinations, which may have detrimental effects. One approach to reduce the effects of outliers is to adjust the learning rate in the weighting formulas of the candidates. A few trimming or screening and robust regression methods have been proposed to remove or reduce the effects of the worst models (see, e.g., Hallman and Kamstra, 1989; Timmermann, 2006). Since in many practical situations, forecast outliers commonly occur, it is important to have theoretically proven forecast combining methods that are robust to outliers.

Researchers have studied forecast combinations under asymmetric loss functions, which are known to affect optimality and even admissibility of estimators (e.g., Zellner, 1985).

Elliott and Timmermann (2004) pointed out that the optimal combining weights under asymmetric loss can be very different from those under squared error loss when the underlying forecast error distribution is skewed. Lee and Yang (2006) showed that under asymmetric L_1 loss functions, the combined binary forecasts based on bagging can improve the predictive ability.

In this work, to effectively deal with possible outliers, we propose two new weighting forms, L_1-AFTER (based on absolute error loss) and h-AFTER (based on Huber loss). Through randomly generated models in our systematic simulations for fair and informative comparison results, we see that the new methods perform very close to AFTER under normal error assumptions, but better or much better than AFTER when outliers are likely to occur. In addition, the new methods provide superior forecasts not only in terms of the mean and median, but also of some other quantiles. Two real data examples are provided to demonstrate the advantage of the new methods. A nice feature of the robust approach is that the gain in accuracy for situations with outliers is accompanied by only a slight damage of performance when no outliers occur. This suggests that L_1-AFTER and h-AFTER can be generally applied. On the theoretical side, we show that the combined forecasts have an attractive oracle property that guarantees their performances to be close to the best candidate under some conditions that do not require model-specific natures of the candidates. This also holds when an asymmetric Huber loss is used for weighting the candidate forecasts. It is worth pointing out that the non-quadratic nature of the robust losses for L_1- and h-AFTER makes the derivation of the oracle inequalities significantly different from that for the original AFTER. However, our present theory still does not handle error distributions with tails heavier than exponential decay, although improved performance of our methods hold under heavy-tailed t_4-distribution in our numerical work.

Our approach complements traditional ones: when a simple statistical model based on which a parametric method (e.g., the variance-covariance method) is derived describes the situation very well, it may have an advantage over ours in terms of accuracy as well as interpretation with reliable confidence/prediction interval and other uncertainty measures; in contrast, when the data generating process (DGP) and relationships between the forecast candidates are difficult to describe, our approach offers more robustness and wider applicability.

The plan of this paper is as follows. In section 2, we set up the problem and briefly review the AFTER algorithm. In section 3, we propose L_1-AFTER and h-AFTER and obtain their theoretical properties under mild regularity conditions. In section 4, we systematically compare the three combining methods and also compare them with other ones through simulations under random model settings and real data examples. Concluding remarks are given in section 5. Technical proofs are in the Appendix.

2.2 Problem setup

Suppose that there is a time series which we are interested in for forecasting, $y_1, y_2, \cdots, y_n$. At each time $i \geq 1$, the explanatory variable $\boldsymbol{x}_i$ is observed prior to the occurrence of y_i. The predictor $\boldsymbol{x}_i$ is multidimensional and may include the past realizations of the response variable. Assume that for $i \geq 1$, the conditional distribution of y_i has mean m_i and variance v_i given $\{(y_t, \boldsymbol{x}_t) : t < i\}$ and $\boldsymbol{x}_i$. That is, $y_i = m_i + e_i$, where e_i is the random discrepancy of y_i from the mean m_i. Let E_i denote the conditional expectation given $\{\boldsymbol{x}_t : t \leq i\}$. Let $\hat{y}_i$ be a predicted value of y_i. The conditional mean square prediction error $E_i(y_i - \hat{y}_i)^2$ can be decomposed into the squared bias and conditional variance: $E_i(y_i - \hat{y}_i)^2 = (m_i - \hat{y}_i)^2 + v_i$. The quantity v_i in the decomposition is always present and is the same for all forecasts. Thus for theoretical considerations, it can be ignored in measuring the performance of the forecasting models, and we consider the net loss: $L(m_i, \hat{y}_i) = (m_i - \hat{y}_i)^2$. When evaluating the performance of a forecast $\hat{y}_i$ over a time period, $i = 1, 2, \cdots, l$, we consider the average forecasting risk

$$\text{Ave. Risk} = \frac{1}{l}\sum_{i=1}^{l} E(m_i - \hat{y}_i)^2,$$

which will be used in our simulation work. For measuring performance on real data, since the above risk is unknowable, we consider the mean square prediction error

$$\text{MSE} = \frac{1}{l}\sum_{i=1}^{l}(y_i - \hat{y}_i)^2.$$

Besides the squared error loss, one may consider other loss functions. For robust estimations, the absolute error loss and Huber loss are commonly used (Huber, 1981). To

allow asymmetry, the modified Huber loss $\varphi(x)$ considered in this work is of the form

$$\varphi_s(x) = \begin{cases} x^2 & \text{if } -1 \le x \le s \\ 2sx - s^2 & \text{if } x > s \\ -2x - 1 & \text{o.w.} \end{cases}$$

for some $s > 0$. Clearly, $\varphi_s(x)$ is symmetric when $s = 1$ and asymmetric otherwise. A graph that compares the absolute error loss, Huber loss with $s = 1.5$ and squared error loss is in Figure 1. In this work we shall examine some theoretical properties of the

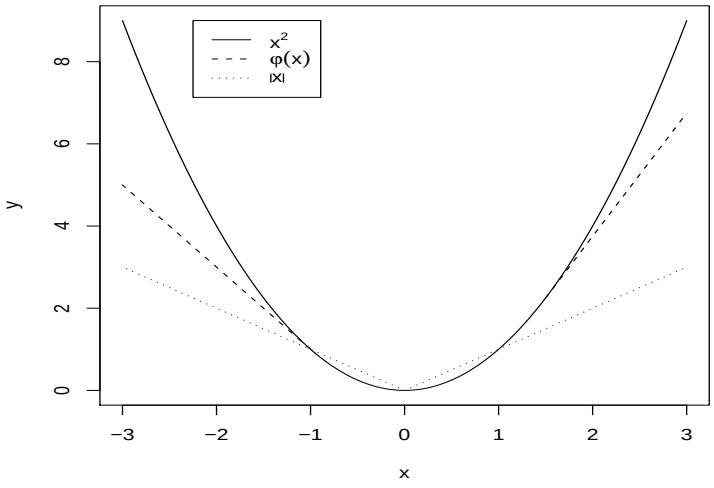

Figure 2.1: Different shapes of x^2, $\varphi(x)$, and $|x|$.

proposed methods under the absolute error and Huber losses.

Assume that there are M forecasting procedures/models. Let $\hat{y}_{j,i}$ denote the forecast of y_i and $\hat{v}_{j,i}$ an estimate of v_i from model j at time i. Let $W_{j,i}$ denote the combining weight for the corresponding forecast candidate. Naturally, we require $W_{j,i}$ be dependent on the information available before y_i is revealed and the nonnegative weights satisfy $\sum_{j=1}^{M} W_{j,i} = 1$ in our proposed methods. The first n_0 observations are treated as initial training data and they are used to obtain the first M forecasts, and we start to combine the forecasts at time $i = n_0 + 1$, and so on. Let the combined forecast be $\hat{y}_{\cdot,i} = \sum_{j=1}^{M} W_{j,i}\hat{y}_{j,i}$.

The core step of the AFTER algorithm (Yang, 2004a) is to assign the weight

$$W_{j,t_1+1} = \frac{\Pi_{i=n_0+1}^{t_1} \hat{v}_{j,i}^{-1/2} \exp(-\sum_{i=n_0+1}^{t_1}(y_i - \hat{y}_{j,i})^2/2\hat{v}_{j,i})}{\sum_{j'=1}^{M} \Pi_{i=n_0+1}^{t_1} \hat{v}_{j',i}^{-1/2} \exp(-\sum_{i=n_0+1}^{t_1}(y_i - \hat{y}_{j',i})^2/2\hat{v}_{j',i})}, \tag{2.1}$$

for $t_1 \geq n_0 + 1$ starting with $W_{j,n_0+1} = \frac{1}{M}$. Numerical results show that when model selection uncertainty is high, AFTER typically improves over model selection methods such as AIC, BIC, HQ and testing (Zou & Yang, 2004, and Chen & Yang, 2007).

A challenge to the AFTER algorithm occurs when forecast outliers appear. Indeed, the quadratic form in (1) can substantially reduce the weight of the best candidate in the presence of an outlier, which makes the weights on the models unstable and hurts the performance of the combined forecast, as will be seen later in section 4.

2.3 Robust AFTER methods

To address the weakness of the original AFTER algorithm in dealing with outliers, we propose two solutions by using the absolute error loss and Huber loss, respectively. In the next section, we will show their superior performance through structured simulations that go beyond a few favorable examples.

2.3.1 Technical conditions

Condition 1: The forecasts satisfy that $\sup_{j\geq 1, i\geq 1} |\hat{y}_{j,i} - m_i| \leq A$ for some positive constant A with probability one.

Condition 2: The conditional variance of e_i is uniformly bounded above by some positive constant B for all $i \geq 1$ with probability one.

Condition 3: There exist a constant $r_0 > 0$ and a monotone function $0 < H(r) < \infty$ on $[0, r_0]$ such that for all $i \geq 1$ and and $0 \leq r \leq r_0$, $E_i \exp(r|e_i|) \leq H(r)$, with probability one.

Condition 3 is satisfied by many error distributions such as normal, gamma, double exponential and certainly bounded distributions, but it does not allow error distributions with tails heavier than exponential decay (e.g., Levy distribution). Condition 1 excludes

some familiar time series models (such as AR(1), see section 3.2 in Sancetta (2010)) and needs more discussion. First, it holds when the observation and forecasts are all bounded, which is true in some applications. For instance, the unemployment rate and the employment growth (e.g., in Rapach and Strauss, 2008), are certainly bounded. In such a case, forecast outliers can occur and our proposed methods are applicable. A demonstration by simulation will be given later in section 4.1 to make this clear. Second, it is important to note that Condition 1 does not require the boundedness of y_i, which thus also allows outliers to occur in the observations. This boundedness requirement holds if the conditional mean of y_i is known to live in a range and the forecasts are accordingly restricted. For instance, for a qualitative threshold ARCH model (Gouriéroux and Monfort, 1992), one may restrict the parameters to satisfy Condition 1. In other applications, e.g., forecasting the amount of seasonal precipitation in a given county for predicting price of crops, based on historical data, suitable bounds on the mean of the response can be obtained, yet extreme response values can occasionally occur. In such a situation, while the forecasts are unlikely to be, say, three standard deviations away from a proper historical mean, it can be better to use a distribution with heavier tail than normal (e.g., double-exponential) to model the innovation error distribution. Third, Condition 1 is technically needed for our proof of the risk bounds in that the tuning parameter λ should be chosen small enough according to the value of A. From a practical perspective, however, when we make use of the estimated $v_{j,i}$ or $d_{j,i}$, we have found that the choice of $\lambda = 1$ works very well as will be seen in section 4. This choice makes the weight of a candidate forecasting procedure interpretable as the Bayesian update of its posterior probability (see, e.g, Haussler et al, 1998). Finally, we point out that the conditions 1-3 do not require any other model-dependent characteristics. Actually our methods work even without knowing the forecasting models as long as the forecasts are available and $v_{j,i}$ or $d_{j,i}$ are estimated based on the observations and forecasts. Some combining methods, such as the approximate Bayesian model averaging method (Garratt et al, 2003), have to know some specific model information like the BIC value to combine the forecasts.

2.3.2 L_1-AFTER

In L_1-AFTER, we replace the quadratic form with absolute value:

$$W_{j,t_1+1} = \frac{\Pi_{i=n_0+1}^{t_1} \hat{d}_{j,i}^{-1} \exp(-\lambda \sum_{i=n_0+1}^{t_1} |y_i - \hat{y}_{j,i}|/\hat{d}_{j,i})}{\sum_{j'=1}^{J} \Pi_{i=n_0+1}^{t_1} \hat{d}_{j',i}^{-1} \exp(-\lambda \sum_{i=n_0+1}^{t_1} |y_i - \hat{y}_{j',i}|/\hat{d}_{j',i})},$$

where $\hat{d}_{j,i}$ is the mean absolute forecast error of model j at time i, or an estimate of $v_{j,i}^{1/2}$. When the variance estimation is difficult, we can set $\hat{d}_{j,i}$ to be 1 in which case the tuning parameter $\lambda > 0$ plays an important role. Our theoretical development focuses on L_1-AFTER without variance estimation (i.e., $\hat{d}_{j,i} = 1$).

Theorem 1. *Under Conditions 1-3, when the tuning parameter $\lambda \le \lambda_0 = \frac{r_0}{2}$, we have*

$$\frac{1}{n-n_0} \sum_{i=n_0+1}^{n} E|y_i - \hat{y}_{\bullet,i}| \le \inf_j \left(\frac{1}{n-n_0} \sum_{i=n_0+1}^{n} E|y_i - \hat{y}_{j,i}| \right) + \frac{\log(M)}{\lambda(n-n_0)} + \frac{a_\lambda(A^2+B)}{2}, \qquad (2)$$

where $a_\lambda = \lambda e^{r_0 A} H(r_0)$. In particular, with the choice of $\lambda = (\frac{2\log(M)}{e^{r_0 A} H(r_0)(A^2+B)(n-n_0)})^{1/2}$, we have

$$\frac{1}{n-n_0} \sum_{i=n_0+1}^{n} E|y_i - \hat{y}_{\bullet,i}| \le \inf_j \left(\frac{1}{n-n_0} \sum_{i=n_0+1}^{n} E|y_i - \hat{y}_{j,i}| \right) + C\sqrt{\frac{\log(M)}{n-n_0}},$$

where C is a constant depending on r_0, A, and B.

Remarks.

1. The theorem shows that the combined forecast automatically behaves like the best one among the candidates plus an additive penalty term of order $\sqrt{\frac{\log(M)}{n-n_0}}$, which cannot be improved for L_1 type of risk. When M is huge, we need to screen out many poor candidates to make the $\log M$ term reasonably small.

2. The derivation of the risk bound does not involve the covariances between the individual forecasts. For combining for improvement, the covariance information is essential and has to be used one way or the other.

2.3.3 H-AFTER

We use the Huber loss in the h-AFTER algorithm and the combining weight becomes

$$W_{j,t_1+1} = \frac{\Pi_{i=n_0+1}^{t_1} \hat{v}_{j,i}^{-1/2} \exp(-\lambda \sum_{i=n_0+1}^{t_1} \varphi_s((y_i - \hat{y}_{j,i})/\sqrt{2\hat{v}_{j,i}}))}{\sum_{j'=1}^{M} \Pi_{i=n_0+1}^{t_1} \hat{v}_{j',i}^{-1/2} \exp(-\lambda \sum_{i=n_0+1}^{t_1} \varphi_s((y_i - \hat{y}_{j',i})/\sqrt{2\hat{v}_{j',i}}))}.$$

In h-AFTER, we keep the quadratic form when the relative forecasting error is not large, and substitute the quadratic term with the tangent lines at the points of (s, s^2) and $(-1, 1)$ otherwise. Clearly, h-AFTER reduces the influence of cases with large forecast errors in the computation of the weights, and by setting $s \neq 1$, we treat substantial under- and over-prediction errors differently. Let $\tau = \max\{s, 1\}$. H-AFTER has the following oracle inequality when $\hat{v}_{j,i}$ are taken to be 1.

Theorem 2. *Under Conditions 1-3, when the tuning parameter $\lambda \leq \lambda_0 = \frac{r_0}{4\tau}$, we have*

$$\frac{1}{n-n_0}\sum_{i=n_0+1}^{n} E\varphi_s(y_i - \hat{y}_{\cdot,i}) \leq \inf_j \left(\frac{1}{n-n_0}\sum_{i=n_0+1}^{n} E\varphi_s(y_i - \hat{y}_{j,i})\right) + \frac{\log(M)}{\lambda(n-n_0)} + \frac{c_\lambda(A^2+B)}{2},$$

where $c_\lambda = 4\lambda\tau^2 e^{r_0 A} H(r_0)$, In particular, with the choice of $\lambda = (\frac{\log(M)}{2\tau^2 e^{r_0 A} H(r_0)(A^2+B)(n-n_0)})^{1/2}$, we have

$$\frac{1}{n-n_0}\sum_{i=n_0+1}^{n} E\varphi_s(y_i - \hat{y}_{\cdot,i}) \leq \inf_j \left(\frac{1}{n-n_0}\sum_{i=n_0+1}^{n} E\varphi_s(y_i - \hat{y}_{j,i})\right) + \bar{C}\sqrt{\frac{\log(M)}{n-n_0}},$$

where $\bar{C}$ is a constant depending on r_0, τ, A, and B.

Remark. The two theorems apply not only for one-step-ahead forecasting, but also for multiple-step-ahead ones.

To guarantee the theoretical properties of L_1-AFTER and h-AFTER, the tuning parameter λ needs to be suitably small. In general, the quantities r_0, A, B are unknown, which makes it impractical to choose λ in accordance with the conditions. In the following simulations and real data examples, we estimate $v_{j,i}$ and $d_{j,i}$ using the observations and forecasts up to current time and simply set $\lambda = 1$. This helps to separate the issue of choosing the best λ from the investigation of the roles of the L_1 and Huber losses in our approach to deal with potential outliers. Similarly we focus on the case with $s = 1$ in the following section (unless stated otherwise) to better illustrate our main points. Data driven selection of λ and s will remain for future work.

2.4 Numerical results

In this section, we extensively compare the three AFTER algorithms with $\lambda = 1$ through simulations and examine their performances in two real data examples. Some other combining methods will be considered as well. The performances are measured under the

squared error: Ave. Risk for simulation and MSE for real data (see section 2 for detail).

We first consider cases that have bounded observations and forecasts. Then we consider two kinds of models, AR and multiple regression models, each under four different error structures: normal, shifted gamma, double exponential, and t, all with mean 0. The (asymmetric) shifted gamma errors are generated from a gamma distribution with shape parameter 3 and scale parameter 1 and then shifted to have zero mean. The t errors are generated from a t distribution with 4 degrees of freedom.

Since we are interested in combining forecasts for adaptation, our focus is the setting where the true model and the true parameters stay the same for each data set (although the true model and the true parameters are randomly generated in some cases). When data are generated from time series models with time varying coefficients, combining for improvement may be the better goal and gradient-based algorithms designed in that direction can perform better than our methods (see e.g., Sancetta (2010)).

2.4.1 Bounded observations and forecasts examples

We consider two simple settings where the observations and forecasts are all bounded, yet forecast outliers can occur. Both symmetric and asymmetric error distributions are considered. We generate a series of y_i, $i = 1, 2, \cdots, 125$. Let m_1 be a random variable uniformly distributed on $[-1, 1]$, and for $i = 1, 2, \cdots, 124$, let $m_{i+1} = 0.2m_i + 0.8z_i$, where z_i is an independent random variable uniformly distributed on $[-1, 1]$ as well. Thus the means of the observations are bounded in $[-1, 1]$. The random noise e_i of y_i follows a bounded symmetric density or a bounded asymmetric density (shifted to have zero mean) as shown in Figure 2. In the left panel of Figure 2, e_i is bounded between $[-10, 10]$ and mainly concentrates on the range $[-0.9, 0.9]$ with probability 90%, but it has some chances to reach much larger values. In the right, e_i is bounded between $[-0.7, 5]$ and concentrates on the range $[-0.7, 0.25]$ with probability 90%.

We consider two forecasting models: AR(1) and random walk. The AR(1) model makes forecasts by R function *ar.yw* after the first 100 observations. By checking the absolute forecast errors of the two candidates, for the realizations tried, we found that the largest values are often away from the means by more than 3 standard deviations. Thus even if

the observations and forecasts are all bounded, forecast outliers still occur. We evaluate the three AFTERs on the last 20 observations and repeat 300 times. The initial estimated $v_{j,i}$ or $d_{j,i}$ of the random walk method is computed on the recent 20 observations, and the following ones are done recursively.

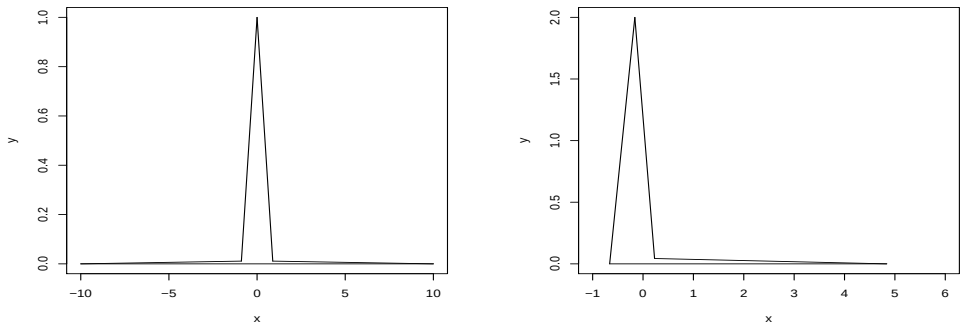

Figure 2.2: Two bounded densities: one symmetric and the other asymmetric.

Table 1 represents the average ratios of the net losses of the robust AFTERs over AFTER. We consider three different values of s in h-AFTER: standard Huber, $s = 0.5$, and $s = 1$. For the symmetric random error case, L_1-AFTER and the symmetric h-AFTER do better than AFTER, while the two asymmetric h-AFTER do slightly worse than AFTER. For the asymmetric random error case, the asymmetric h-AFTER with $s = 0.5$ outperforms the L_1-AFTER and symmetric h-AFTER (only slightly).

Table 2.1: Comparisons when observations and forecasts are bounded.

	L_1-AFTER	h-AFTER$_{s=1}$	h-AFTER$_{s=0.5}$	h-AFTER$_{s=1.5}$
Symmetric case	$0.93_{(0.038)}$	$0.89_{(0.025)}$	$1.06_{(0.085)}$	$1.03_{(0.043)}$
Asymmetric case	$0.97_{(0.013)}$	$0.93_{(0.012)}$	$0.92_{(0.014)}$	$1.04_{(0.016)}$

Note: The numbers in parentheses are the corresponding standard errors.

2.4.2 AR models

To get a general picture of the performances of competing methods, we mainly consider random model settings rather than some specific models. Note that for a reasonable model selection/combination method, it is usually the case that one can find a few cases where the method compares favorably against given alternatives. Hence randomly generated models are important to obtain fair and informative simulation results.

We consider three scenarios. The first is that the true model has a fixed order AR(3) with candidate models ranging from AR(1) to AR(5). The second is that the order of the random model uniformly varies from AR(1) to AR(5) with candidate models ranging from AR(1) to AR(7). The third is that the true model has a structural break, changing from AR(2) to AR(5), with candidate models being AR(1) up to AR(5). In each scenario, we consider each type of error structures with each of eleven different variances $\{0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$. All the coefficients of the true models are randomly generated from the uniform distribution on $[-1, 1]$ (non-stationary coefficients are discarded). The AR processes are generated by R function *arima.sim*.

Given each error distribution, we generate 100 models with randomly generated coefficients as described above, and replicate each one and the candidate forecasts 300 times to simulate the average forecasting risks. In each replication, we generate a sample with size 125. The candidate models start to generate forecasts after 100 observations, then recursively do so once every additional observation is made. The three combining methods start weighting right after the initial forecasts are generated, and the last 20 observations are used for evaluating the competing methods.

Scenario 1: Random models with a fixed order

Figure 3 shows the trends of means of the risk ratios of the proposed methods to AFTER versus the variance under the four types of error structure. From the figure, we have the following observations:

1. When the error is normal, as is expected, the robust methods have no advantage compared to AFTER. When the error is asymmetric (shifted gamma) or when the outliers are likely to happen (double exponential or t), L_1-AFTER and h-AFTER significantly outperform AFTER. In particular, when the error is t, they make improvements by 18-20%.

It seems that, the more frequently outliers occur, the larger the improvements generated by the two proposed methods.

2. When the error is normal or double exponential, L_1-AFTER steadily performs better than h-AFTER. When the error is shifted gamma, h-AFTER is slightly better than L_1-AFTER. When the error is t, L_1-AFTER and h-AFTER have similar performance.

Besides the mean ratios, we are also interested in other summary statistics of the ratios over the 100 random models. Table 2 presents the means plus five-number summaries of the risk ratios (median, minimum, maximum, first quartile, and third quartile) under the four error structures with variance 3. From Table 2, when the error is shifted gamma or double exponential, the maximum ratios of the proposed methods are all ranging from 1.00 to 1.04. That is, in the worst cases, the proposed methods perform almost the same as AFTER, while the minimum ratios indicate that they can earn gains of 30% or more in the most favorable cases. The first quartiles indicate that the alternatives outperform AFTER by more than 10% for at least 25% of the random models, while the third quartiles indicate that the alternatives perform as well as or better than AFTER for at least 75% of the random models.

When the error is t, the maximum ratios of L_1-AFTER and h-AFTER are a little larger: around 1.08. However, the minimum ratios of the two methods indicate possibly more than 60% gains in the most favorable cases. The first quartiles indicate that the two methods outperform AFTER by more than 30% for at least 25% of the random models, while the third quartiles indicate they significantly outperform AFTER for at least 75% of the random models.

Scenario 2: Random models with varying orders

Figure 4 shows the trends of means of the risk ratios of the proposed methods versus the variance when the order of the 100 random models uniformly varies from AR(1) to AR(5) (other aspects are the same as before). From Figure 4, the main message is similar to that of the fixed order setting.

Similarly, Table 3 gives the means plus five-number summaries of the risk ratios over the random models under the four error structures with variance 3. The results show the advantages of the robust AFTER methods across models of different orders under the

		Mean	Median	Min	Max	1st q.	3rd q.
normal	L_1-AFTER	$1.06_{(0.004)}$	1.06	0.93	1.18	1.03	1.09
	h-AFTER	$1.16_{(0.009)}$	1.15	1.00	1.37	1.08	1.22
shifted gamma	L_1-AFTER	$0.93_{(0.009)}$	0.95	0.66	1.04	0.89	1.00
	h-AFTER	$0.92_{(0.008)}$	0.94	0.68	1.04	0.88	1.00
double exponential	L_1-AFTER	$0.86_{(0.010)}$	0.86	0.56	1.01	0.79	0.94
	h-AFTER	$0.91_{(0.009)}$	0.92	0.66	1.04	0.87	0.98
t	L_1-AFTER	$0.78_{(0.018)}$	0.82	0.40	1.06	0.62	0.95
	h-AFTER	$0.78_{(0.019)}$	0.80	0.39	1.08	0.64	0.95

Table 2.2: Comparisons when the true model is a random AR(3).

non-normal error distributions.

		Mean	Median	Min	Max	1st q.	3rd q.
normal	L_1-AFTER	$1.04_{(0.004)}$	1.04	0.96	1.17	1.01	1.08
	h-AFTER	$1.10_{(0.008)}$	1.09	0.99	1.30	1.02	1.17
shifted gamma	L_1-AFTER	$0.95_{(0.007)}$	0.98	0.71	1.05	0.92	1.00
	h-AFTER	$0.93_{(0.007)}$	0.96	0.76	1.02	0.88	0.98
double exponential	L_1-AFTER	$0.90_{(0.008)}$	0.93	0.59	1.03	0.87	0.97
	h-AFTER	$0.94_{(0.007)}$	0.97	0.65	1.02	0.90	0.99
t	L_1-AFTER	$0.84_{(0.015)}$	0.89	0.42	1.02	0.77	0.97
	h-AFTER	$0.85_{(0.016)}$	0.89	0.42	1.08	0.75	0.98

Table 2.3: Comparisons when the true model uniformly varies from AR(1) to AR(5).

Scenario 3: Random models with a structural break

The DGP begins with an AR(2) for the first 115 observations, but changes to an AR(5) afterwards based on $y_{116} = \alpha_1 y_{115} + \alpha_2 y_{114} + \alpha_3 y_{113} + \alpha_4 y_{112} + \alpha_5 y_{111} + e_{116}$, and so on, where α_1, α_2, and the error distribution are the same as in the AR(2), and α_3, α_4, and α_5 are randomly generated from the uniform distribution on $[-1, 1]$. There are 10 observations after the structural break. The other simulation settings are the same as in the preceding scenarios. To save space, in Table 4, we present the mean risk ratios of the proposed methods to AFTER under the normal and t error distributions with variance 3.

The proposed methods on average are almost identical to AFTER when the error is normal and show slight advantages to AFTER when the error follows the t distribution.

	h-AFTER	L_1-AFTER
normal	$0.98_{(0.007)}$	$1.00_{(0.007)}$
t	$0.97_{(0.007)}$	$0.98_{(0.007)}$

Table 2.4: Comparisons when the true model has a structural break.

2.4.3 Regression models

We consider three scenarios similar to those considered in section 4.2. Assume there are eight independent predictors, $x_1, \ldots, x_8$, uniformly distributed on $[0, 1]$. For scenario 1, the random regression model is of the form,

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + e_i.$$

For scenario 2, the random regression model is uniformly drawn from the nested models,

$$\begin{aligned} y_i =& \beta_1 x_{1i} + e_i, \\ y_i =& \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \\ & \ldots \\ y_i =& \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + e_i. \end{aligned}$$

For scenario 3, the underlying model is the second one above for certain periods and then becomes the last above for the remaining observations. In each of the three scenarios, $\beta_1, \beta_2, \ldots, \beta_5$, are independently drawn from the uniform distribution on $[-1, 1]$, and the error term e_i follows each type of the four distributions with eleven different variances $\{0.1,0.5,1,2,3,4,5,6,7,8,9\}$ respectively. The forecast candidates are generated from nested models, where the smallest model only has one predictor x_1, and the largest has all eight predictors.

As before, under each error distribution, we randomly generate 100 models, and replicate each model 300 times to simulate the average forecasting risks. In each replication, we

generate a sample of 65 independent observations. The candidate models start to generate forecasts after 40 observations, then recursively get the following forecasts respectively. The three combining methods start weighting immediately after the initial forecasts are generated, and the last 20 are used for evaluating the competing methods.

Scenario 1: Random models with a fixed number of predictors

From Figure 5, when the error is normal, the two alternatives perform slightly worse than AFTER. When the error is shifted gamma, h-AFTER outperforms AFTER by up to 3% and is better than L_1-AFTER. When the error is double exponential or t, L_1-AFTER outperforms AFTER by up to 6% and is better than h-AFTER.

Scenario 2: Random models with varying number of predictors

Since the trends of means of the risk ratios versus the variance in this case are similar as in Figure 5, they are not presented here. Table 5 shows an example of the means plus five-number summaries of the ratios over 100 random models with variance 5.

It is interesting to compare Table 3 and Table 5. In Table 5, the two alternatives on average perform much closer to AFTER than in Table 3 when the error is normal. However, when the error is of the other three types of distributions, the two alternatives on average make relatively more improvements in Table 3 than in Table 5.

		Mean	Median	Min	Max	1st q.	3rd q.
normal	L_1-AFTER	$1.00_{(0.001)}$	1.00	0.96	1.03	1.00	1.01
	h-AFTER	$1.01_{(0.001)}$	1.01	0.99	1.04	1.00	1.03
shifted gamma	L_1-AFTER	$0.99_{(0.001)}$	1.00	0.96	1.03	0.98	1.00
	h-AFTER	$0.97_{(0.001)}$	0.97	0.94	1.00	0.97	0.98
double exponential	L_1-AFTER	$0.94_{(0.002)}$	0.94	0.89	0.97	0.92	0.95
	h-AFTER	$0.99_{(0.001)}$	0.99	0.95	1.02	0.98	1.00
t	L_1-AFTER	$0.93_{(0.002)}$	0.94	0.86	0.98	0.92	0.95
	h-AFTER	$0.98_{(0.002)}$	0.98	0.91	1.01	0.97	0.99

Table 2.5: Comparisons when the true models have different number of predictors.

Scenario 3: Random models with a structural break

The true model begins with the x_1 and x_2 predictors for the first 55 observations, and

then changes with the predictors, $x_1, \ldots, x_5$, for the last 10 observations. The coefficients of x_1 and x_2 and the error distribution remain the same after the structural break. The coefficients of x_3, x_4, and x_5 are independently drawn from the uniform distribution on $[-1, 1]$. The other settings are the same as in the preceding two scenarios.

In Table 6, we present the mean risk ratios of the proposed methods to AFTER under the normal and t error distributions with variance 5. The results are similar to the earlier comparisons of the AFTERs.

In general, since the Huber and absolute losses differ substantially from squared error loss only when the error is large, we expect that the structural breaks, unless they are really irregular, would not much affect the performances of the robust AFTERs relative to AFTER.

	h-AFTER	L_1-AFTER
normal	$1.01_{(0.001)}$	$1.01_{(0.001)}$
t	$0.98_{(0.001)}$	$0.95_{(0.001)}$

Table 2.6: Comparisons when the true model has a structural break.

2.4.4 Combining when no candidate model fully captures the underlying DGP

Hendry and Clements (2004) theoretically and numerically justified the usefulness of combining forecasts when no candidate model fully captures the underlying data generation process. They mainly considered two combining strategies: the simple averaging method (SA) and the variance-covariance methods (Bates & Granger, 1969). We investigate the performance of the AFTERs in a similar but slightly more complicated context. As in Hendry and Clements (2004), suppose

$$y_i = \beta_1 x_{1,i-1} + \beta_2 x_{2,i-1} + e_i,$$

where x_1 and x_2 are two independent AR(1) processes, each with an autoregressive coefficient of 0.9 and an error term following $N(0, 1)$. The two candidate models are $y_i =$

$a_0 + a_1 x_{1,i-1} + u_i$ and $y_i = b_0 + b_1 x_{2,i-1} + v_i$ respectively. Thus, each of the two models only provides partial descriptions of the underlying DGP. The variance of e_i is set to be 0.16 (as in Hendry and Clements, 2004) in one setting and $var(e_i) = 1$ in a second setting. Differently from Hendry and Clements (2004), the true parameters β_1 and β_2 are randomly drawn from the uniform distribution on $[-1, 1]$. The other settings remain the same as in the previous experiments. We evaluate the performance of the two aforementioned combining strategies, taking the same form of the variance-covariance method (denoted by BG) as in Hansen (2008):

$$W_{j,i+1} = \frac{\hat{v}_{j,i}^{-1}}{\sum_{j'=1}^{J} \hat{v}_{j',i}^{-1}},$$

where $\hat{v}_{j,i}$ is the estimated forecast error variance of model j at time i. To save space, we only present the mean risk ratios of the individual models and the combining methods to AFTER when e_i follows normal and t distributions.

		Model 1	Model 2	SA	h-AFTER	L_1-AFTER	BG
$\sigma_{e_i} = .4$	normal	$15.5_{(4.79)}$	$10.9_{(2.59)}$	$6.63_{(1.30)}$	$1.07_{(0.009)}$	$0.99_{(0.004)}$	$8.82_{(1.57)}$
	t	$4.57_{(0.80)}$	$5.82_{(1.09)}$	$2.61_{(0.30)}$	$0.93_{(0.017)}$	$0.89_{(0.018)}$	$3.86_{(0.35)}$
$\sigma_{e_i} = 1$	normal	$3.75_{(0.67)}$	$3.53_{(0.53)}$	$1.85_{(0.19)}$	$1.08_{(0.007)}$	$1.02_{(0.002)}$	$3.08_{(0.27)}$
	t	$2.91_{(0.43)}$	$3.44_{(0.45)}$	$1.62_{(0.13)}$	$0.94_{(0.012)}$	$0.90_{(0.014)}$	$2.69_{(0.18)}$

Table 2.7: Comparisons when no candidate model fully captures the DGP.

In Table 7, all of the combining methods substantially improve over the individual forecasting models. The AFTERs outperform the simple averaging and variance-covariance methods (note that since the relative performances of L_1-AFTER and h-AFTER to AFTER are quite stable, their standard errors are much smaller than those of the other methods). Again, for the heavy-tailed t distribution, L_1-AFTER and h-AFTER demonstrate a clear predictive gain over AFTER. As for a general comparison between the AFTERs and the other combining methods, the outcome is expected to depend on how much the candidates capture the DGP, the degree of information overlap, the sample size and the error distribution. More work is needed to reach a clear general conclusion.

2.4.5 Comparing weights of the combining methods and some other observations

In the preceding simulations, the degrees of freedom of the t distribution are 4. We observed that if we increase the degrees of freedom of the t distribution, the advantages of the proposed methods decrease which is expected because the t distribution with large degrees of freedom behaves like a normal distribution.

To investigate closely why the alternative methods outperform AFTER, we conduct a single run in the random AR(3) setting under the t distribution error. Table 8 provides partial combining weights of the 20 stages for h-AFTER and AFTER during the single run. Among the five candidates, the AR(3) model is the true model. Although in general for the

	stage	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
AFTER	...	...	...	...	...	...
	6	0.00	0.19	0.26	0.28	0.27
	7	0.12	0.32	0.18	0.19	0.20
	8	0.06	0.28	0.23	0.21	0.22
	9	0.03	0.24	0.25	0.24	0.24
	...	...	...	...	...	...
	20	0.00	0.09	0.35	0.29	0.27
h-AFTER	...	...	...	...	...	...
	6	0.00	0.19	0.26	0.28	0.27
	7	0.05	0.23	0.23	0.24	0.25
	8	0.02	0.19	0.27	0.26	0.26
	9	0.01	0.15	0.29	0.27	0.27
	...	...	...	...	...	...
	20	0.00	0.05	0.36	0.30	0.28

Table 2.8: Weight distributions of AFTER and h-AFTER.

prediction purpose the true model is not necessarily the best (e.g., Zou and Yang, 2004), it is in this case. Since h-AFTER and AFTER share some common components, they assign exactly the same weights to each candidate for the first six stages. However, the 6th observation is spiky, and all the AR models yield poor predictions (AR(3) is still relatively better than the other candidates). Due to the squared error loss, the following weight of

AR(3) by AFTER is significantly reduced; in contrast, the weight of the same candidate by h-AFTER is less affected. In Table 8, we can further observe that the influence of the outlier lasts two more stages in AFTER.

So far we have focused on the forecast outliers due to the spiky observations that occur with very large random noises. To evaluate the performance of the proposed methods when the outliers come from severely poor forecasts, we conduct a very simple simulation. In the random AR(3) model setting, when the error is normal, the alternatives do not have advantages compared to AFTER. Now we randomly add a constant 6 to two out of the twenty forecasts of each candidate model, mimicking the severe errors in forecasts. Table 9 shows the means and five-number summaries of the risk ratios of 100 random models with variance 3.

	Mean	Median	Min	Max	1st q.	3rd q.
L_1-AFTER	$0.93_{(0.004)}$	0.92	0.90	1.26	0.91	0.94
h-AFTER	$1.00_{(0.001)}$	1.00	0.97	1.08	0.99	1.01

Table 2.9: Comparisons when some forecasts are severely poor.

In Table 9, h-AFTER in general performs as well as AFTER, while L_1-AFTER makes significant improvements compared to AFTER, showing the robustness of this method to severely poor forecasts.

2.4.6 The AFTERs vs some other combining methods

We compare the different AFTER algorithms with five other combining methods in random model settings with randomly generated coefficients as before. The first two are SA and BG as in section 4.4. The third is a trimmed mean method (denoted by TM) that removes the largest and smallest forecasts before averaging. The fourth is the median of the candidate forecasts (denoted by MED). The fifth is the ordinary least squares method (Granger & Ramanathan, 1984, denoted by GR).

We generate 100 random AR(3) models with variance 0.5 and sample size 75. The forecast candidates are AR(1) up to AR(5) plus the random walk method. The AR forecasts

are made after 50 observations, and the combining methods are evaluated on the last 20 time points. The estimates $\hat{v}_1$ and $\hat{d}_1$ of the random walk forecasts are obtained on the most recent 15 observations. The GR method uses 20 observations to calculate the first combining weights.

Table 10 shows the mean risk ratios of the combining methods to AFTER under the four error structures. Clearly, the robust AFTERs consistently outperform the other

	normal	shifted gamma	d. exponential	t
L_1-AFTER	$1.05_{(0.004)}$	$0.83_{(0.015)}$	$0.76_{(0.016)}$	$0.67_{(0.023)}$
h-AFTER	$1.13_{(0.006)}$	$0.82_{(0.013)}$	$0.79_{(0.014)}$	$0.66_{(0.023)}$
SA	$3.60_{(0.294)}$	$2.51_{(0.254)}$	$2.65_{(0.272)}$	$2.80_{(0.408)}$
TM	$1.68_{(0.111)}$	$1.39_{(0.107)}$	$1.26_{(0.104)}$	$1.38_{(0.153)}$
MED	$1.18_{(0.048)}$	$0.99_{(0.059)}$	$0.86_{(0.047)}$	$0.87_{(0.080)}$
BG	$1.36_{(0.044)}$	$1.07_{(0.044)}$	$0.99_{(0.043)}$	$0.85_{(0.051)}$
GR	$4.05_{(0.111)}$	$3.75_{(0.216)}$	$3.56_{(0.130)}$	$5.17_{(0.778)}$

Table 2.10: The AFTERs vs some other combining methods.

methods under the four error structures, and the gains are often substantial. The trimmed mean improves the simple averaging, and the median further improves the performance. The BG method ranks between TM and MED. The GR method performs rather poorly under the four error structures. The relative performances of the robust AFTERs to that of AFTER, again, confirm their advantages when dealing with outliers.

2.4.7 Data examples

The U.S. income data

Consider the U.S. aggregated disposable income data with $n = 127$ (Greene, 2000). Chen & Yang (2007) studied this data set to compare the AFTER algorithm with three hypothesis testing procedures, showing an advantage of AFTER. Graphical inspections suggest differencing the observations, and then AR(1) up to AR(5) are considered as the candidate models. When checking the absolute forecast errors of the candidate models, we found evidence of forecast outliers. We also consider the approximate Bayesian model averaging

method (ABMA) (Garratt et al, 2003), where the weight takes the form

$$W_{j,i+1} = \frac{\exp(-\frac{1}{2}\mathrm{BIC}_{j,i})}{\sum_{j'=1}^{J}\exp(-\frac{1}{2}\mathrm{BIC}_{j',i})},$$

where BIC is the BIC value of model j at time i (see also Hansen, 2008). The ABMA method was not investigated in section 4.6 as the random walk forecasts do not have the associated BIC values.

All the different combining methods are evaluated over the last 20 observations. The AFTER algorithms start weighting based on the first 100 observations. The GR method uses the most recent 40 observations to calculate the initial weights. Table 11 shows that the new AFTER methods outperform AFTER by up to 6%. The ABMA and GR methods perform poorly compared to AFTER. The trimmed mean and median methods are comparable to AFTER, but worse than the robust AFTERs. The SA and BG methods are comparable to the robust AFTERs.

Table 2.11: Comparisons for the U.S. income data

AFTER	L_1-AFTER	h-AFTER	ABMA	SA	TM	MED	BG	GR
98.15	92.98	91.97	103.06	92.26	97.89	98.15	93.60	113.81
1.00	0.95	0.94	1.05	0.94	1.00	1.00	0.95	1.16

Note: the second row is the MSEs and the third is the MSE ratios to AFTER.

The U.S. real GDP growth data

Stock and Watson (2003) made forecasts on output growth and inflation of seven developed countries using many likely relevant financial variables. They (and references therein) showed that the forecastibility of many financial variables is much weaker after the mid-1980s due to possible structural breaks (e.g., Rapach and Weber, 2004). They also pointed out an interesting phenomenon: simple combination forecasts reliably improve upon an AR benchmark, while individual forecasting models are unable to do so. Under the same setting, considering 10 of the U.S. financial variables used in Stock and Watson (2003), Rapach and Weber (2004) showed that while the individual forecasting models that included the financial variables provided insignificant forecastibility relative to the AR benchmark

over the 1985:1-1999:4 period, the encompassing test results indicated that many variables still contained useful information.

Under Rapach and Weber's setting (data obtained from http://pages.slu.edu/faculty/rapachde/Research.html), we found that for all the candidate models, the maximums of the absolute forecast errors over the time period are away from the corresponding means by more than two standard deviations, indicating forecast outliers or structural breaks indeed occur. We investigate the performance of the AFTERs and the other competing methods on the U.S. real GDP growth over the 1988:1-1999:4 period, with the 1985:1-1987:4 period being used for the GR method to compute the initial weights. Table 12 presents the MSEs of the AR benchmark and the combining methods when the forecasts are made one quarter ahead. L_1-AFTER improves over AFTER and is among the few that do better than the AR benchmark.

AR	AFTER	L_1-AFTER	h-AFTER	ABMA	SA	TM	MED	BG	GR
4.42	4.48	4.31	4.48	7.90	4.36	4.35	4.28	4.39	5.35

Table 2.12: Comparisons for the U.S. real GDP growth data

2.5 Concluding remarks

There are two directions of forecast combinations. One is combining for adaptation that takes advantage of the candidates so that the combined forecast performs adaptively well in the sense that no matter which candidate happens to be favored by the true DGP, the combined one will perform almost equally well. The other is combining for improvement, that is, the combined forecast is intended to perform better than every candidate. Theoretical work has shown that the cost of combining for improvement due to parameter estimation is substantially higher than that of combining for adaptation. Therefore, combining for improvement can often lead to very poor performance, as is often observed in real application and in our numerical results as well.

In this paper, to achieve the goal of combining for adaptation, we propose two robust AFTER algorithms: L_1-AFTER and h-AFTER. Non-asymptotic risk bounds hold for these

methods. In contrast, asymptotic results (e.g., based on normal approximations) are not necessarily trustworthy. For example, variance-covariance and regression-based methods are supposed to converge to the best convex or linear combination and thus perform better than the best individual forecast. However, as seen in our numerical work, they actually performed rather poorly in some cases.

The original AFTER algorithm is based on the squared error loss. L_1-AFTER uses the absolute forecast error loss and h-AFTER is based on the Huber loss, which combines the behaviors of the squared forecast error loss around zero and the absolute forecast error loss when the error is large. The new methods alleviate the influence of forecast outliers. The simulation results suggest that they significantly outperform AFTER when the error is asymmetric or when outliers commonly occur. Very importantly, when the errors are normally distributed, they usually perform only slightly worse than AFTER. The real data examples also support this finding.

In summary, based on the theoretical and numerical investigations, the robust AFTER methods have very stable and reliable performance when the goal is combining forecasts for adaptation. A future direction is to extend the present theoretical work to deal with error distributions with polynomially decaying (or even heavier) tails.

2.6 Appendix

The proofs of Theorems 1 and 2 follow the same basic ideas, of which, the main challenge is dealing with the non-quadratic nature of the absolute and Huber losses. Since the Huber loss case is more complicated, we provide the details in the proof of Theorem 2 and skip some similar materials in the proof of Theorem 1.

2.6.1 Proof of Theorem 2:

Define $L(x) = x^2 1_{-1\le x\le s} + (2x-1)1_{x>s} + (-2x-1)1_{x<-1}$, and $h(x) = \exp(-\lambda L(x))$. Define $Q_{n-n_0} = \frac{1}{M}\sum_{j=1}^{M}\Pi_{i=n_{0+1}}^{n} h(y_i - \hat{y}_{j,i})$. Then $\forall\ j$,

$$-\log(Q_{n-n_0}) \le \log(M) + \lambda \sum_{i=n_{0+1}}^{n} L(y_i - \hat{y}_{j,i}).$$

On the other hand,

$$Q_{n-n_0} = \sum_{j=1}^{M} \frac{1}{M} h(y_{n_0+1} - \hat{y}_{j,n_0+1}) \times \frac{\sum_{j=1}^{M} h(y_{n_0+1} - \hat{y}_{j,n_0+1}) h(y_{n_0+2} - \hat{y}_{j,n_0+2})}{\sum_{j=1}^{M} h(y_{n_0+1} - \hat{y}_{j,n_0+1})} \times \cdots \times \frac{\sum_{j=1}^{M} \Pi_{i=n_0+1}^{n} h(y_i - \hat{y}_{j,i})}{\sum_{j=1}^{M} \Pi_{i=n_0+1}^{n-1} h(y_i - \hat{y}_{j,i})}.$$

Then $Q_{n-n_0} = \Pi_{i=n_0+1}^{n} \sum_{j=1}^{M} W_{j,i} h(y_i - \hat{y}_{j,i})$. Accordingly, $-\log(Q_{n-n_0}) = -\sum_{i=n_0+1}^{n} \log\left(\sum_{j=1}^{M} W_{j,i} h(y_i - \hat{y}_{j,i})\right) = -\sum_{i=n_0+1}^{n} \log(E^J h(y_i - \hat{y}_{J,i}))$, where E^J denotes the expectation with respect to J under the distribution $P(J = j) = W_{j,i}$ given each i. Then $\hat{y}_{\bullet,i} = E^J \hat{y}_{J,i}$. Define $V = L(y_i - \hat{y}_{J,i}) - E^J L(y_i - \hat{y}_{J,i})$. Therefore, $E^J V^2 = E^J (L(y_i - \hat{y}_{J,i}) - E^J L(y_i - \hat{y}_{J,i}))^2 \leq E^J (L(y_i - \hat{y}_{J,i}) - L(y_i - E^J \hat{y}_{J,i}))^2 \leq 4\tau^2 E^J (\hat{y}_{J,i} - E^J \hat{y}_{J,i})^2 = 4\tau^2 E^J (\hat{y}_{J,i} - \hat{y}_{\bullet,i})^2$, where the second inequality holds because the first derivatives of $L(x)$ are bounded between -2τ and 2τ. By Lemma 3.6.1 of Catoni (2004, p.85), we get $\log(E^J h(y_i - \hat{y}_{J,i})) \leq -\lambda E^J L(y_i - \hat{y}_{J,i}) + I$, where $I \leq \frac{\lambda^2}{2} \exp(\lambda(|L(y_i - \hat{y}_{J,i})| + \sup_{J \geq 1} |L(y_i - \hat{y}_{J,i})|)) E^J V^2$. Then

$$\begin{aligned} I \leq & \frac{\lambda^2}{2} \exp(2\lambda \sup_{J \geq 1} |L(y_i - \hat{y}_{J,i})|) E^J V^2 \leq \frac{\lambda^2}{2} \exp(4\tau\lambda \sup_{J \geq 1} |y_i - \hat{y}_{J,i}|) E^J V^2 \\ \leq & \frac{\lambda^2}{2} \exp(4\tau\lambda(|y_i - m_i| + \sup_{J \geq 1} |m_i - \hat{y}_{J,i}|)) E^J V^2 \leq \frac{\lambda^2}{2} e^{4\tau\lambda A} \exp(4\tau\lambda |e_i|) E^J V^2 \\ \leq & 2\tau^2 \lambda^2 e^{4\tau\lambda A} \exp(4\tau\lambda |e_i|) E^J (\hat{y}_{J,i} - \hat{y}_{\bullet,i})^2, \end{aligned}$$

where the second to last inequality holds because of Condition 1. Under Condition 3, taking conditional expectation with respect to the random noise e_i given $\{(y_t, \boldsymbol{x}_t) : t < i\}$ and $\boldsymbol{x}_i$, we have that when $4\tau\lambda \leq r_0$, $E_i(I) \leq 2\tau^2 \lambda^2 e^{4\tau\lambda A} H(4\tau\lambda) E_i(E^J (\hat{y}_{J,i} - \hat{y}_{\bullet,i})^2)$. Define a surrogate loss function $L_s(x) = L(x) + cx^2$, $c > 0$. Let $b_0 = y_i - \hat{y}_{\bullet,i}$ and $b = y_i - \hat{y}_{J,i}$.

Then,

$$
\begin{aligned}
&L_s(b) - (2cb_0 + 2s1_{b_0>s} - 21_{b_0<-1} + 2b_0 1_{-1\le b_0\le s})(b-b_0) - L_s(b_0)\\
=&c(b-b_0)^2 + b^2 1_{-1\le b\le s} + (2sb-s^2)1_{b>s} + (-2b-1)1_{b<-1} - b_0^2 1_{-1\le b_0\le s}\\
&- (2sb-s^2)1_{b_0>s} + (2b+1)1_{b_0<-1} - 2b_0(b-b_0)1_{-1\le b_0\le s}\\
=&c(b-b_0)^2 + \begin{cases}
0 & \text{if } b>s \text{ and } b_0>s\\
2sb-s^2+2b+1 & \text{if } b>s \text{ and } b_0<-1\\
(s-b_0)(2b-b_0-s) & \text{if } b>s \text{ and } -1\le b_0\le s\\
(s+1)(s-2b-1) & \text{if } b<-1 \text{ and } b_0>s\\
0 & \text{if } b<-1 \text{ and } b_0<-1\\
(b_0+1)(b_0-1-2b) & \text{if } b<-1 \text{ and } -1\le b_0\le s\\
(b-s)^2 & \text{if } -1\le b\le s \text{ and } b_0>s\\
(b+1)^2 & \text{if } -1\le b\le s \text{ and } b_0<-1\\
(b-b_0)^2 & \text{if } -1\le b\le s \text{ and } -1\le b_0\le s
\end{cases}\\
\ge&c(b-b_0)^2.
\end{aligned}
$$

Since $E^J(b-b_0)=0$, we have $E^J L_s(b) - L_s(b_0) \ge cE^J(b-b_0)^2$ and

$$E_i(E^J L_s(y_i - \hat{y}_{J,i}) - L_s(y_i - \hat{y}_{\bullet,i})) \ge cE_i(E^J(\hat{y}_{J,i} - \hat{y}_{\bullet,i})^2).$$

Then

$$E_i(I) \le 2c^{-1}\tau^2\lambda^2 e^{4\tau\lambda A} H(4\tau\lambda) E_i(E^J L_s(y_i - \hat{y}_{J,i}) - L_s(y_i - \hat{y}_{\bullet,i})).$$

Choose λ small enough such that $\frac{\lambda}{2} \ge 2c^{-1}\tau^2\lambda^2 e^{4\tau\lambda A} H(4\tau\lambda)$. Then $E_i(I) \le \frac{\lambda}{2} E_i(E^J L_s(y_i - \hat{y}_{J,i}) - L_s(y_i - \hat{y}_{\bullet,i}))$. So far there have been two constraints on λ: $4\tau\lambda \le r_0$ and $\frac{\lambda}{2} \ge 2c^{-1}\tau^2\lambda^2 e^{4\tau\lambda A} H(4\tau\lambda)$. When we choose c and λ so that $\lambda \le \frac{r_0}{4\tau}$ and $c \ge 4\lambda\tau^2 e^{4\tau\lambda A} H(4\tau\lambda)$,

the constraints are met. Let $c_\lambda = 4\lambda\tau^2 e^{r_0 A} H(r_0)$. With such a choice of λ and c_λ, we have,

$$\begin{aligned}
&E_i(\log E^J \exp(-\lambda L(y_i - \hat{y}_{J,i}))) \\
\leq& -\lambda E_i(L(y_i - \hat{y}_{\bullet,i})) + \lambda E_i(L(y_i - \hat{y}_{\bullet,i}) - E_i(\lambda E^J L(y_i - \hat{y}_{J,i})) \\
&+ \frac{\lambda}{2} E_i(E^J L_s(y_i - \hat{y}_{J,i}) - L_s(y_i - \hat{y}_{\bullet,i})) \\
\leq& -\lambda E_i(L(y_i - \hat{y}_{\bullet,i})) - \frac{\lambda}{2} E_i(E^J L(y_i - \hat{y}_{J,i}) - L(y_i - \hat{y}_{\bullet,i})) \\
&+ \frac{c_\lambda \lambda}{2} E_i(E^J (y_i - \hat{y}_{J,i})^2) - \frac{c_\lambda \lambda}{2} E_i((y_i - \hat{y}_{J,i})^2) \\
\leq& -\lambda E_i(L(y_i - \hat{y}_{\bullet,i})) + \frac{c_\lambda \lambda (A^2 + B)}{2}.
\end{aligned}$$

The last inequality above holds because of the convexity of function $L(x)$ and condition 1 and 2. Since j is arbitrary, we have

$$\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} EL(y_i - \hat{y}_{\bullet,i}) \leq \inf_j \left(\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} EL(y_i - \hat{y}_{j,i}) \right) + \frac{\log(M)}{\lambda(n - n_0)} + \frac{c_\lambda (A^2 + B)}{2}.$$

To determine an optimal λ, let $\frac{\log(M)}{\lambda} = \frac{c_\lambda (A^2+B)(n-n_0)}{2}$. We get $\lambda' = \sqrt{\frac{\log(M)}{2\tau^2 e^{r_0 A} H(r_0)(A^2+B)(n-n_0)}}$. It is clear that the constraint $4\tau\lambda' \leq r_0$ is satisfied when $n - n_0$ is large enough. With the choice of λ', we have

$$\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} EL(y_i - \hat{y}_{\bullet,i}) \leq \inf_j \left(\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} EL(y_i - \hat{y}_{j,i}) \right) + \bar{C} \sqrt{\frac{\log(M)}{n - n_0}},$$

where $\bar{C}$ is a constant depending on τ, r_0, A, and B. Since $\varphi(x) = L(x)$, we have

$$\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} E\varphi(y_i - \hat{y}_{\bullet,i}) \leq \inf_j \left(\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} E\varphi(y_i - \hat{y}_{j,i}) \right) + \bar{C} \sqrt{\frac{\log(M)}{n - n_0}}.$$

This completes the proof.

2.6.2 Proof of Theorem 1:

Unless directly defined, the meanings of the notations applied in the proof are the same as in the proof of Theorem 2. Define $L(x) = x1_{x \geq 0} - x1_{x<0}$, and $h(x) = \exp(-\lambda L(x))$. Define Q_{n-n_0} similarly as before. Then $-\log(Q_{n-n_0}) = -\sum_{i=n_0+1}^{n} \log(E^J h(y_i - \hat{y}_{J,i}))$, and

$$\log(E^J h(y_i - \hat{y}_{J,i})) \leq -\lambda E^J L(y_i - \hat{y}_{J,i}) + I,$$

where $I \leq \frac{\lambda^2}{2} e^{2\lambda A} \exp(2\lambda |e_i|) E^J (\hat{y}_{J,i} - \hat{y}_{\bullet,i})^2$. Define a surrogate loss function $L_s(x) = L(x) + ax^2$, $a > 0$. Then $L_s(b) - (2ab_0 + 1_{b_0 \geq 0} - 1_{b_0 < 0})(b - b_0) - L_s(b_0) \geq a(b - b_0)^2$. This leads to $E^J L_s(y_i - \hat{y}_{J,i}) - L_s(y_i - \hat{y}_{\bullet,i}) \geq a E^J (\hat{y}_{J,i} - \hat{y}_{\bullet,i})^2$. Similarly, with the constraint on λ, $2\lambda \leq r_0$, and $\frac{\lambda}{2} \geq \frac{a^{-1}\lambda^2}{2} e^{2\lambda A} H(2\lambda)$, we have $E_i(I) \leq \frac{\lambda}{2} E_i(E^J L_s(y_i - \hat{y}_{J,i}) - L_s(y_i - \hat{y}_{\bullet,i}))$. Let $a_\lambda = \lambda e^{r_0 A} H(r_0)$. Then $E_i(\log E^J \exp(-\lambda L(y_i - \hat{y}_{J,i}))) \leq -\lambda E_i(L(y_i - \hat{y}_{\bullet,i})) + \frac{a_\lambda \lambda (A^2 + B)}{2}$. Since j is arbitrary, we have

$$\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} EL(y_i - \hat{y}_{\bullet,i}) \leq \inf_j \left(\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} EL(y_i - \hat{y}_{j,i}) \right) + \frac{\log(M)}{\lambda(n - n_0)} + \frac{a_\lambda (A^2 + B)}{2}.$$

With an optimal choice of λ, $\lambda'' = \sqrt{\frac{2\log(M)}{e^{r_0 A} H(r_0)(A^2+B)(n-n_0)}}$, we have

$$\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} EL(y_i - \hat{y}_{\bullet,i}) \leq \inf_j \left(\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} EL(y_i - \hat{y}_{j,i}) \right) + C \sqrt{\frac{\log(M)}{n - n_0}},$$

where C is a constant depending on r_0, A, and B. Since $|x| = L(x)$, we have

$$\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} E|y_i - \hat{y}_{\bullet,i}| \leq \inf_j \left(\frac{1}{n - n_0} \sum_{i=n_0+1}^{n} E|y_i - \hat{y}_{j,i}| \right) + C \sqrt{\frac{\log(M)}{n - n_0}}.$$

This completes the proof.

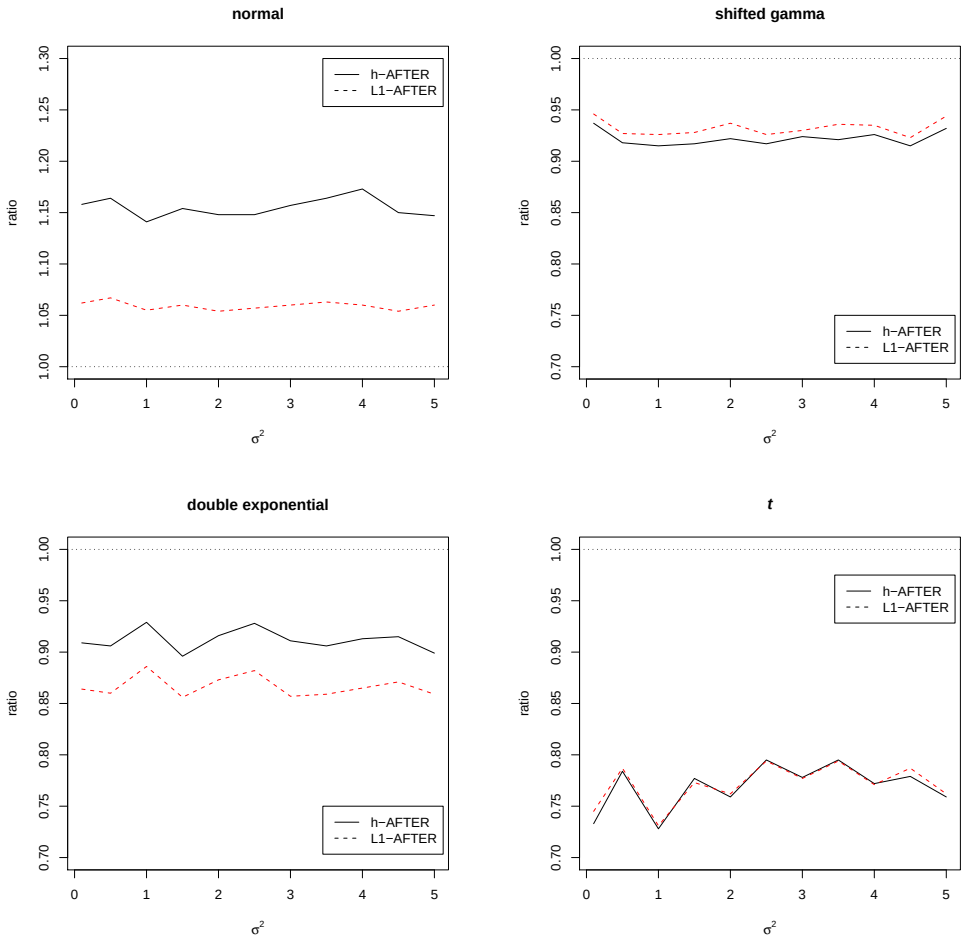

Figure 2.3: Relative risks to that of AFTER under AR(3) models.

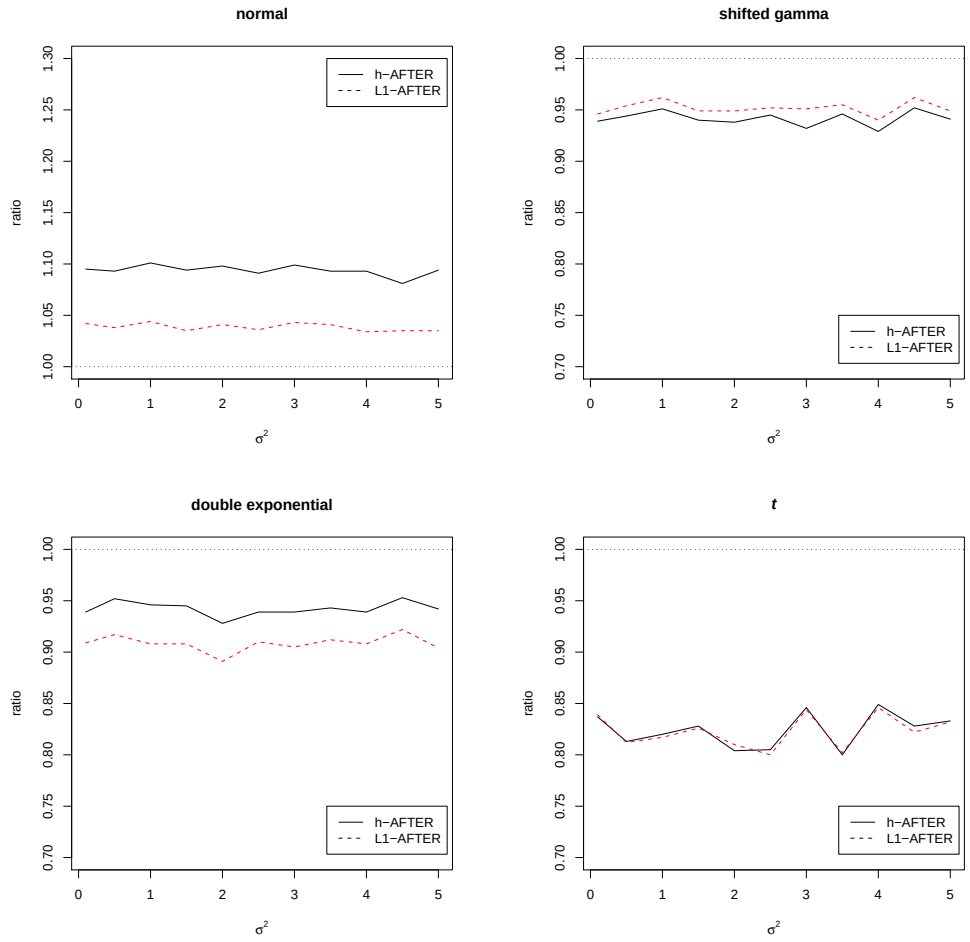

Figure 2.4: Relative risks to that of AFTER under AR(1) to AR(5) models.

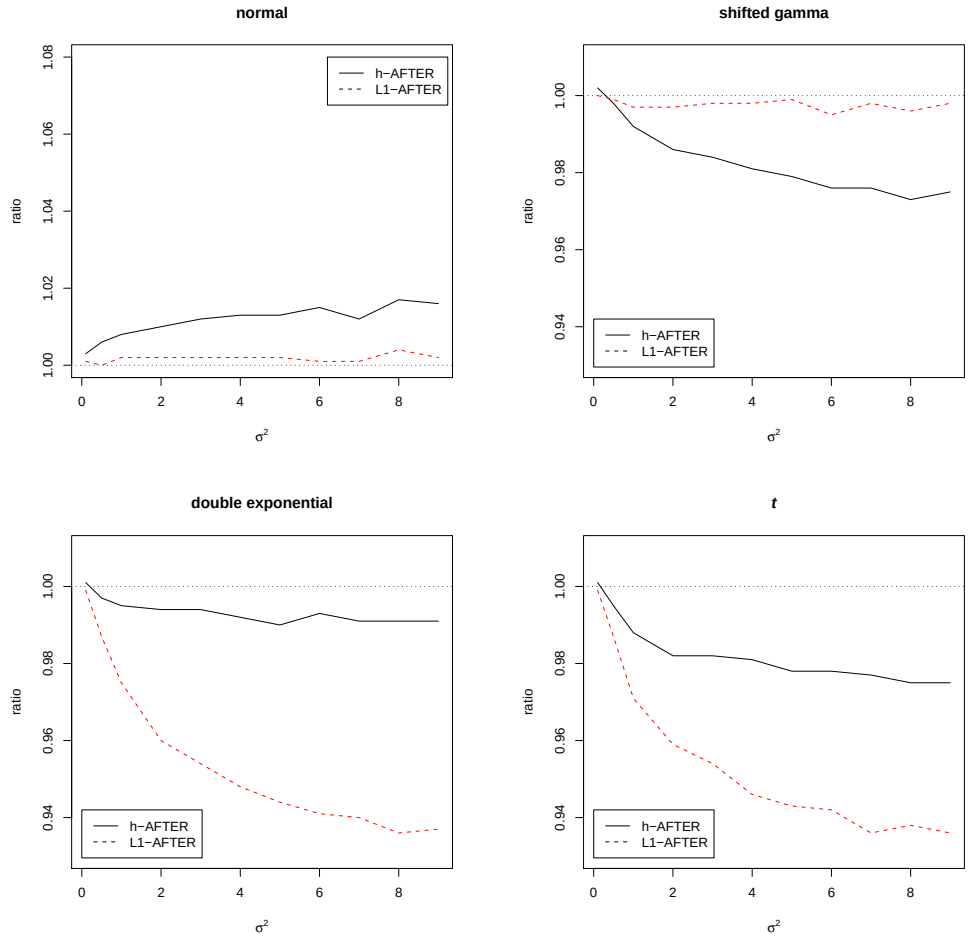

Figure 2.5: Relative risks to that of AFTER when the true regression models have 5 predictors.

Chapter 3

Robust Combination of Model Selection Methods for Prediction

3.1 Introduction

Model selection with a number of predictors has been an exciting research area. Methods have been proposed in recent years to conduct variable selection with computationally feasible algorithms, sometimes maintaining familiar statistical properties of traditional information criteria. These methods have been increasingly used and numerous numerical results demonstrate their advantages in some settings. With multiple such model selection tools available, a question a statistics user faces is: How should one select a model selection method for his/her data?

Obviously, we should not expect a single choice to perform the best in different scenarios. Some insights have been offered in the literature on this issue. For example, sparsity of the underlying regression function in terms of the number of explanatory variables involved is regarded a key feature that makes some methods perform better than others. Fan and Li (2001) pointed out that the SCAD outperforms the LASSO (Tibshirani, 1996) when the model noise level is low while the LASSO does better than the SCAD when the noise level is high. Zou (2006) also observed that the LASSO outperforms the SCAD and adaptive LASSO when the signal-noise-ratio (SNR) is small while the SCAD and adaptive

LASSO methods do better than the LASSO when the SNR is large. However, in real applications, with the model noise level and true regression function unknown, much more needs to be done both theoretically and through systematic numerical investigations before satisfactory conclusions can be reached to provide statistical characterizations of the data that determine the relative performance of the different methods. Intuitively, if one gets to know when to use which model selection method, then there is a great advantage to consider a list of distinct model selection rules so that at least one of them is optimal or well-behaving for the unknown underlying data generating process (DGP).

For moderate or high dimensional regression problems, however, with a small or moderate sample, the task of identifying the best among a group of model selection methods is typically very difficult. Thus there is a serious challenge to realize the potential advantage of sharing strengths of a number of model selection rules in a pool. For the goal of prediction or estimating the regression function (in contrast to identifying the important variables), as is the focus in our paper, one approach to overcome the challenge is to combine the model selection methods by a proper weighting of the predictions or estimates from them. If the combination leads to a performance similar or close to the best method in each scenario of the underlying DGP, then because the methods perform the best in different scenarios, the combined estimator or prediction can outperform all the candidate model selection methods in repeated applications across different scenarios of the DGP. This will be seen in our numerical results later.

The topic of combining regression procedures has been studied, which yields various interesting theoretical properties. Oracle inequalities show that properly combining arbitrary regression procedures leads to a risk close to the best among a target class of combinations of the candidate estimators/predictions plus a minimax-rate optimal "price of combining" that reflects the largeness of the class of allowed combinations. See Chen and Yang (2010) for a literature review on this research area. Successes of combining different predictions in real applications have also prompted more interest on combining statistical procedures from a practical perspective. For instance, in the well-known Netflix competition, ensemble of different methods is a key idea employed by top teams (see, e.g., http://www.netflixprize.com/leaderboard).

The previous theoretically proven combining methods in e.g., Yang (2001) and Catoni

(2004), use quadratic-type loss in determining weights for the candidates and show that the combined regression estimator achieves the best performance offered by the candidates in an accumulated risk. The quadratic-type of loss is also used in combining methods by e.g., Juditsky and Nemirovski (2000), Yang (2004) and Tsybakov (2003) for larger target classes of combinations.

The mathematically convenient quadratic loss for weighting regression estimators works very well under Gaussian noise. However, when the noise has a heavier tail, as commonly occurs in reality, a few outliers often destabilize the weights. A robust combination of estimates or predictions is thus sought.

With the above background, in this paper, for the purpose of estimating the regression function or prediction based on a collection of models, we propose a robust method, called l_1-ARM, to combine regression estimates/predictions obtained by a list of model selection methods. More specifically, the quadratic loss in the ARM (adaptive regression by mixing, see, Yang, 2001) is replaced by absolute loss, and an oracle risk bound is presented that also allows a screening step to be incorporated to remove poor model selection methods, which can be very helpful when a large number of methods are considered. In our numerical work, we focus on combining the LASSO, SCAD (Fan and Li, 2001), and adaptive LASSO (Zou, 2006) as applied in the linear regression setting. The results are highlighted as follows.

1. Several representative linear expressions with different degree of sparsity and multiple noise levels are considered to compare the performance of the model selection methods and l_1-ARM. The results show that the l_1-ARM performs like the best model selection method in the different scenarios and thus indeed share the predictive strengths of the candidate model selection methods in an overall sense.

2. The results show the advantage of the l_1-ARM over the original ARM: when the noise is Gaussian, they perform similarly; but when the noise has a heavy tail, the l_1-ARM performs significantly better. Thus we recommend the l_1-ARM for application.

3. To produce fair and insightful comparisons of the competing methods, we randomly generate models in terms of coefficients and the number of non-zero terms. A simple investigation is done to see how the relative performances of the LASSO, SCAD and adaptive LASSO depend on the sparsity of true regression function and the SNR.

The paper is organized as follows. In Section 2 we propose the l_1-ARM algorithm. In Section 3 we investigate the LASSO, SCAD, adaptive LASSO, and l_1-ARM via various simulation settings and real data examples. In Section 4 we present a theoretical result of the l_1-ARM. We give concluding remarks in Section 5. The technical proof of the theorem in Section 4 is in the Appendix.

3.2 The proposed method

Consider a general regression problem $Y_i = f(\mathbf{x}_i) + \varepsilon_i$, $i = 1, \ldots, n$, where $f(\cdot)$ is the true regression function. For estimating $f(\cdot)$, a number of models are considered. For instance, one may consider the collection of all the subset models with terms chosen from a list of predictors (with possible transformations and/or interaction terms). We will focus on linear models in our numerical work, although our proposed method and the theoretical result are applicable for more general regression models. Being unsure about which model and which model selection rule works the best, we apply K model selection methods on the data. The model selection method j yields an estimator $\widehat{f}_{j,n}(\mathbf{x})$, which takes the form of $f_{\hat{m}}(\mathbf{x}, \hat{\boldsymbol{\beta}}_{\hat{m}})$ with $\hat{m}$ being the model chosen and $\hat{\boldsymbol{\beta}}_{\hat{m}}$ being the parameter estimate by the method. Denote the set of the K candidate methods by Γ.

3.2.1 The l_1-ARM algorithm

Yang (2001) proposed the ARM algorithm for combining a group of regression models. Yuan and Yang (2005) extended the ARM with model screening. In the ARM, the l_2-norm was used in the core step to apportion the weights to each candidate. Under the quadratic loss, however, when the underlying model generates outliers, the weight of the best candidate model can easily be diluted and other models can unexpectedly obtain more weights. To overcome this drawback, we propose the l_1-ARM in hope that it performs similarly as the ARM when the noise is normally distributed and outperforms the ARM when the noise distribution has a heavy tail.

The l_1-ARM algorithm is as follows.

- Step 1. Apply the model selection methods on the whole data to get their recommended models and the regression estimates $\hat{f}_{j,n}(\mathbf{x})$.

- Step 2. Split the data into two parts, $Z^{(1)} = (\mathbf{x}_i, Y_i)$, $1 \le i \le n/2$, and $Z^{(2)} = (\mathbf{x}_i, Y_i)$, $n/2 + 1 \le i \le n$.

- Step 3. Based on $Z^{(1)}$, compute the mean absolute prediction error $\hat{d}_j = \frac{2}{n} \sum_1^{n/2} |Y_i - \hat{f}_{j,n}(\mathbf{x}_i)|$ respectively for each candidate model j.

- Step 4. For each model $j \in \Gamma$, predict Y_i by $\hat{f}_{j,n}(\mathbf{x}_i)$ for $Z^{(2)}$. Compute an overall measure of discrepancy,
$$D_j = \sum_{i=n/2+1}^{n} |Y_i - \hat{f}_{j,n}(\mathbf{x}_i)|.$$

- Step 5. Compute the convex weight for model j,
$$W_j = \frac{\hat{d}_j^{-n/2} \exp(-\eta D_j / \hat{d}_j)}{\sum_{k \in \Gamma} \hat{d}_k^{-n/2} \exp(-\eta D_k / \hat{d}_k)}.$$

- Step 6. Randomly permute the order of the data $N - 1$ times. Repeat Step 2 - Step 5 and let $W_{j,r}$ denote the weight of method j computed at the rth permutation for $0 \le r \le N - 1$. Let $\hat{W}_j = \frac{1}{N} \sum_{r=0}^{N-1} W_{j,r}$.

- Step 7. Let
$$\hat{f}_n(\mathbf{x}) = \sum_{j \in \Gamma} \hat{W}_j \hat{f}_{j,n}(\mathbf{x})$$
be the final l_1-ARM estimate of the true regression function f. At a new $\mathbf{x}'$, the combined prediction is $\hat{Y} = \hat{f}_n(\mathbf{x}')$.

Remarks.

1. When one considers a large number of model selection methods (e.g., one may explore a number of distinct values of a tuning parameter of a model selection rule and treat each of them as a candidate), it may be helpful to combine a reduced candidate set rather than the full set Γ both for better accuracy and for saving computation cost. We will address this issue later in Section 4.

2. One may, of course, consider data splitting ratios other than half/half. For ensuring optimal rate of convergence, a fixed splitting ratio such as half/half works. In our experience, half/half splitting typically works well for our purpose of regression estimation/prediction.

3. In Step 5, there is a tuning parameter η to control the degree of reliance of weighting on the predictive performance (note that when $\eta = 0$, the prediction errors D_j have no effect at all on weighting). In our numerical work, we set $\eta = 1$ and it worked very well.

3.2.2 Combining SCAD, LASSO, and adaptive LASSO

In the numerical work of this paper, we will focus on combining some recently proposed selection methods: the SCAD, LASSO, and adaptive LASSO (a-LASSO). Below we describe some details about applying them.

There are two tuning parameters a and λ in the SCAD. Following Fan and Li (2001), we set $a = 3.7$ in this work. As suggested by Zou (2006), we estimate the weight vector in the a-LASSO by $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}|^{\gamma}$, where $\hat{\boldsymbol{\beta}}$ is the OLS estimator and γ is selected from $\{.5, 1, 2\}$ by fivefold cross validation. The two selection options of the tuning parameter λ of the three methods are described as follows. Denote the full data set by D and the testing set by D_m, $m = 1, \ldots, 5$. For each λ and m, we get the estimate $\hat{\boldsymbol{\beta}}_{\lambda}^{m}$ using the training set $D - D_m$. Then the fivefold cross validation selection is to minimize

$$CV_{\lambda} = \sum_{m=1}^{5} \sum_{(\mathbf{x}_i, Y_i) \in D_m} \left(Y_i - f(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_{\lambda}^{m}) \right)^2 .$$

The BIC selection is to minimize

$$BIC_{\lambda} = \log \hat{\sigma}_{\lambda}^2 + df_{\lambda} \log(n)/n,$$

where df_{λ} is the number of nonzero coefficients of the fitted model (see, Wang, Li and Tsai, 2007; Zou, 2008).

In the following numerical work, the LASSO and a-LASSO are computed by using the R package *lars*, where the optimal λ is chosen from 100 candidates along the entire

solution path by using fivefold cross validation and BIC selections respectively. The SCAD is computed by using the one-step SCAD program provided by Zou and Li (2008), where the optimal λ is chosen from 100 discretized values by using fivefold cross validation and BIC selections respectively.

The comparison of different estimators is done under regression estimation loss $E(f(\mathbf{x})-\hat{f}(\mathbf{x}))^2$ via simulation, where the expectation is taken on the new observation $\mathbf{x}$ that has the same distribution as in the data. For the empirical comparison using real data sets, we consider predictive mean squared error as an objective measure.

3.3 Numerical results

We shall focus on the performance of the selection and combining methods in linear regression: $Y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$. Assume that $\mathbf{x}$ follows a multivariate normal distribution with zero mean and a covariance matrix which will be defined respectively in the following subsections, and that the random noise ε is iid $N(0, \sigma^2)$ or with a contamination.

3.3.1 Some representative examples

Assume that there are 12 predictor variables in $\mathbf{x}$ and that the covariance between x_k and x_l is $\rho^{|k-l|}$ with $\rho = 0.5$, $1 \leq x_k, x_l \leq 12$. The sample size n is set to be 50 and 100, and the performance of the competing methods is evaluated at 1000 independently generated observations from the same distribution. We replicate the estimation process 100 times in each case. In each replication, we set $N = 100$ to calculate the combining weights. We consider the relative loss, which is the ratio of the loss of a competing method over that of the OLS estimate from the full model, and the median of the relative loss over the 100 replicates is reported as in Fan and Li (2001). Another measure, the mean of the relative loss, gives similar results, but the combining methods usually have smaller standard errors than the selection methods (not reported here due to space limitation), indicating that they are more robust. Since the results of n equal to 50 and 100 are similar for the following simulations, we only present the $n = 100$ case below.

High sparsity. In this example, $\boldsymbol{\beta}$ is set to $(3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)^T$, identical to

a model investigated in Zou and Li (2008). Note that this example indicates a case that the underlying model is very sparse (3 nonzero coefficients out of 12 potential predictors). In Table 3.1, the SCAD methods perform the best when the model noise level is low. However, when the model noise level is high, the $\text{LASSO}_{\text{BIC}}$ is more accurate than the SCAD and the a-LASSO methods. Note that BIC consistently outperforms fivefold CV in this case. The two combining methods are automatically close to or outperform the best selection method.

	Scad_{cv}	Scad_{bic}	Lasso_{cv}	$\text{Lasso}_{\text{bic}}$	$\text{Alasso}_{\text{cv}}$	$\text{Alasso}_{\text{bic}}$	Arm	l_1-Arm
$\sigma = 1$	0.25	0.22	0.60	0.53	0.44	0.37	0.34	0.35
$\sigma = 3$	0.43	0.36	0.61	0.53	0.49	0.39	0.39	0.41
$\sigma = 5$	0.58	0.54	0.60	0.51	0.55	0.55	0.49	0.49

Table 3.1: Risk comparison for the high sparsity case

Low sparsity. In this example, $\boldsymbol{\beta}$ is set to $(3, 1.5, 0, 1.1, 2, 0, .9, .8, .6, 0, 0, 0)^T$, expanding the model complexity by adding more nonzero coefficients. There are 7 nonzero coefficients, indicating a low sparsity situation. In Table 3.2, the performance of the a-LASSO is comparable to or better than the SCAD. The LASSO performs relatively better than the others when the model noise level is moderate or high. It can be seen that fivefold CV starts to show its advantage in some settings. The two combining methods tend to be close to the best procedure, and they perform similarly.

	Scad_{cv}	Scad_{bic}	Lasso_{cv}	$\text{Lasso}_{\text{bic}}$	$\text{Alasso}_{\text{cv}}$	$\text{Alasso}_{\text{bic}}$	Arm	l_1-Arm
$\sigma = 1$	0.79	0.76	0.89	0.92	0.83	0.77	0.77	0.77
$\sigma = 3$	1.00	1.08	0.85	0.90	0.98	1.00	0.91	0.92
$\sigma = 5$	1.00	1.09	0.79	0.84	0.97	0.94	0.81	0.82

Table 3.2: Risk comparison for the low sparsity case

Non-sparsity. In this example, $\boldsymbol{\beta}$ is set to $(.5, .5, 0, .5, .5, .5, 0, .5, .5, .5, 0, .5)^T$. There are 9 nonzero coefficients. In Table 3.3, when the model noise level is low, all the three selection methods perform similarly as the OLS estimator. When the model noise level is moderate, the SCAD and a-LASSO perform worse than the LASSO. When the model noise

level is high, all the three selection methods with fivefold CV do better or much better than the OLS estimator, but those with BIC give poor results, much worse than those from fivefold CV. The two combining methods, again, automatically perform like the best procedure, and their performances are very close to each other.

	Scad_{cv}	Scad_{bic}	Lasso_{cv}	$\text{Lasso}_{\text{bic}}$	$\text{Alasso}_{\text{cv}}$	$\text{Alasso}_{\text{bic}}$	Arm	l_1-Arm
$\sigma = 1$	1.00	1.10	1.00	1.00	0.98	0.98	1.00	1.00
$\sigma = 3$	1.00	1.08	0.80	0.86	1.01	1.04	0.86	0.86
$\sigma = 5$	0.90	1.06	0.63	1.18	0.72	0.93	0.69	0.68

Table 3.3: Risk comparison for the non-sparsity case

Heavy-tailed noise case. We investigate the robustness of the selection and combining methods. As in Fan and Li (2001), let ε be mixed with 90% standard normal and 10% standard Cauchy distributions in the above three examples respectively. In Table 3.4, the three selection methods show different ranking in the three examples, i.e., the SCAD does the best with the first example, the a-LASSO the best with the second, and the LASSO the best with the third. We observe that the strengths of the selection methods are getting weaker when the degree of the model sparsity is reducing. BIC favors sparse models, while fivefold CV favors non-sparse models with the existence of outliers. Unlike the previous examples, the two combining methods perform differently, and the l_1-ARM shows a significant improvement over the ARM.

	Scad_{cv}	Scad_{bic}	Lasso_{cv}	$\text{Lasso}_{\text{bic}}$	$\text{Alasso}_{\text{cv}}$	$\text{Alasso}_{\text{bic}}$	Arm	l_1-Arm
Example 1	0.32	0.32	0.54	0.56	0.40	0.42	0.39	0.32
Example 2	0.89	0.86	0.86	0.86	0.83	0.78	0.80	0.73
Example 3	1.00	1.05	0.89	0.96	0.99	1.01	0.99	0.93

Table 3.4: The robustness of the selection and combining methods

Highly correlated predictors. For establishing consistency and efficiency properties of the LASSO, a-LASSO and the one-step SCAD, regularity conditions are used, one of which is that the design matrix behaves nicely (see, e.g., Zou, 2006; Zhao and Yu, 2006; and Meinshausen and Bühlmann, 2006; Zou and Li, 2008; Zhang and Huang, 2008). In

this example, we simulate a model in the opposite direction and compare the performance of the above methods plus forward selections by AIC and BIC. The coefficients of this model are the same as those in the second example. The predictors $\mathbf{x}_i$, $i = 1, \ldots, n$, follow a multivariate normal distribution with mean $\mathbf{0}$ and a covariance matrix which is a random realization of a Wishart distribution with $df = 12$ and scale matrix being the identity matrix. The maximum eigenvalue of the covariance matrix is 46.032, while the minimum eigenvalue is 0.0002. The error ε follows a standard normal distribution. The other simulation settings remain the same as in the previous examples.

In Table 3.5, the SCAD and a-LASSO show no advantage compared to the OLS. For instance, the SCAD with BIC selection on average generates 2 zero coefficients for this model and only one of them is correct. The forward selections by AIC and BIC work more favorably than the other selection methods in this situation. As before, the performance of the combining methods is close to the best candidate.

Aic	Bic	Scad_{cv}	Scad_{bic}	Lasso_{cv}	$\text{Lasso}_{\text{bic}}$	$\text{Alasso}_{\text{cv}}$	$\text{Alasso}_{\text{bic}}$	Arm	l_1-Arm
0.84	0.74	1.00	1.00	0.99	0.93	1.00	0.98	0.79	0.79

Table 3.5: Risk comparison for highly correlated predictors

3.3.2 Randomly generated models

Relative performance. The purpose of random model settings is to try to get an unbiased understanding on the competing methods. We randomly generated 100 models. The number of zero coefficients is uniformly distributed from 2 to 8, and their orders in the 12 potential predictors are also uniformly distributed. The nonzero coefficients are uniformly generated from [0, 3]. Other settings remain the same as in the previous examples. For each model, we calculate the mean of the relative losses. Strictly speaking, for the mixture error case, the risk does not exist since the Cauchy distribution does not even have the first moment. However, our performance measure is the relative losses to the ordinary least square. The median of the mean relative losses of the 100 models is demonstrated in Table 3.6. The a-LASSO with the BIC selection performs the best with $\sigma = 1, 3$, and

the heavy-tail noise case. The LASSO with the fivefold CV selection performs the best with $\sigma = 5$. Interestingly fivefold CV does better than BIC for the SCAD and LASSO while BIC does better than fivefold CV for the a-LASSO in the random model settings. The SCAD is close to the best under low or heavy-tail noise. The two combining methods perform very well consistently. When the underlying model noise is normally distributed, they perform almost identically. However, when outliers are present, the l_1-ARM has the edge.

	Scad_{cv}	Scad_{bic}	Lasso_{cv}	$\text{Lasso}_{\text{bic}}$	$\text{Alasso}_{\text{cv}}$	$\text{Alasso}_{\text{bic}}$	Arm	l_1-Arm
$\sigma = 1$	0.81	0.85	0.92	0.97	0.87	0.77	0.79	0.79
$\sigma = 3$	0.93	0.99	0.89	0.95	0.97	0.85	0.81	0.82
$\sigma = 5$	1.00	1.12	0.86	0.97	0.96	0.92	0.83	0.83
heavy tail	0.78	0.84	0.85	0.93	0.84	0.74	0.77	0.69

Table 3.6: Risk comparison based on randomly generated models

How do the SNR and sparsity affect the relative performances? In the simulation above, the number of zero coefficients can be treated as a measure of the model sparsity. We estimate the variance of the mean function $\mathbf{x}^T\boldsymbol{\beta}$, V_s, with sample size equal to 1100 and then obtain the SNR of the model by taking the ratio of V_s/σ^2. To get some insight on how the SNR and sparsity are associated with the performance of the three selection methods, we regress the mean relative losses of the random models on the SNR and sparsity. To save space, we only consider the case where the tuning parameters are selected by BIC with $\sigma = 3$. The SNR of the 100 random models ranges from 0.94 to 13.2. For ease of notation, the subscript bic is omitted in Table 3.7. We also consider the ratios of the mean relative losses among the three selection methods as the response variables. Table 3.7 gives the coefficients of the SNR and sparsity of each model (the intercept is not presented here). For each model, we check the linear model assumptions by applying the diagnostic means (e.g., residual plots). It shows that all the simple linear models look proper. Note that all of these coefficients are significant at $\alpha = 0.01$ level and that the interaction effects of these two predictors in these models are not significant. Thus in our setting, the SNR and sparsity have "additive" effects on the performance of these selection methods.

	SNR	sparsity
Scad	-0.026	-0.131
Lasso	0.004	-0.035
Alasso	-0.008	-0.092
$\frac{\text{Scad}}{\text{Lasso}}$	-0.033	-0.104
$\frac{\text{Scad}}{\text{Alasso}}$	-0.020	-0.033
$\frac{\text{Alasso}}{\text{Lasso}}$	-0.013	-0.066

Table 3.7: The sparsity vs the signal-to-noise ratio

For the first three rows of Table 3.7, all the coefficients of the sparsity are negative, which indicates the three methods (relative to the full model) perform better with sparse models. The SNR coefficients of the SCAD and a-LASSO are negative, but that of the LASSO is positive, which indicates the LASSO does better when the SNR is small. For the second three rows of Table 3.7, all the coefficients of the SNR and sparsity are negative. It seems that for the large SNR and high sparsity cases the SCAD dominates the LASSO and a-LASSO while the a-LASSO dominates the LASSO. This suggests that the accuracy of the SCAD relative to LASSO and a-LASSO increases with higher SNR and sparsity. The same can be said on a-LASSO relative to LASSO.

3.3.3 High dimensional cases

Forty predictors. Consider high dimensional cases where the number of the predictors is set to be 40. Assume that the covariance between x_k and x_l is $\rho^{|k-l|}$ with $\rho = 0.75$, $1 \leq x_k, x_l \leq 40$. In case 1, there are 5 nonzero coefficients, in case 2, there are 10 nonzero coefficients, and in case 3, there are 20 nonzero coefficients. Note in all the three cases, the nonzero coefficients are uniformly generated from $[0, 3]$ and their orders are uniformly distributed in the model. We consider two scenarios of the model noise for each case: $\sigma = 2$ and having a heavy tail as in the previous examples. We repeat each case 50 times. To ease the computation burden, for each replication, we generate 50 random samples, and for each sample, we set $N = 50$ to get the combining weights. The other simulation settings remain the same as before. The median of the relative losses over the 50 replicates

is presented in Table 3.8, where the two rows in each case correspond to the two model noises respectively. For the different high dimensional cases, we can see that the selection methods have different performances. The two combining methods are close to the best performance among the candidates when the model noise follows a normal distribution. With heavy-tail noise, the l_1-ARM outperforms all the selection methods and the ARM.

	Scad_{cv}	Scad_{bic}	Lasso_{cv}	$\text{Lasso}_{\text{bic}}$	$\text{Alasso}_{\text{cv}}$	$\text{Alasso}_{\text{bic}}$	Arm	l_1-Arm
Case 1	0.22	0.22	0.33	0.30	0.28	0.28	0.22	0.22
	0.24	0.24	0.30	0.28	0.25	0.25	0.25	0.18
Case 2	0.42	0.41	0.52	0.42	0.57	0.42	0.37	0.37
	0.40	0.41	0.46	0.39	0.46	0.36	0.39	0.32
Case 3	0.77	0.75	0.75	0.62	1.35	0.68	0.62	0.62
	0.73	0.73	0.67	0.60	1.11	0.63	0.64	0.55

Table 3.8: Risk comparison for the high dimensional data

Eighty predictors. A useful strategy of dealing with high dimensional cases is to add a screening step, so only very significant variables are included in the later modeling process. In this example, assume that $\mathbf{x}_i$ consists of 80 variables and follow the same distribution as in the above example. The first 8 true coefficients are $(2, 2, 2, 2, 2, 2, 2, 2)^T$, and the remaining true coefficients are all zeros. We apply the Sure Independence Screening (Fan and Lv, 2008) to reduce the dimension from 80 to 40. Table 3.9 shows the median relative losses over 50 random samples of the competing methods. With the screening step, the selection methods have different finite sample performances. The combining methods perform as if they knew which selection method is the best for a specific situation. When there are outliers, the l_1-ARMS shows a clear advantage compared to the selection methods and the ARMS.

p > n. Consider a high dimensional case with $n = 100$ and $p = 300$. Assume that the covariance between x_k and x_l is $\rho^{|k-l|}$ with $\rho = 0.5$, $1 \leq x_k, x_l \leq p$. The true coefficients are all zeros except the first positions being $(3, 1.5, 0, 1.1, 2, 0, .9, .8, .6)^T$. Adaptive LASSO is not applicable for this situation. Instead, we consider SCAD, LASSO, SICA (Lv and Fan, 2009) and MCP (Zhang, 2007), which all can handle p > n cases.

	Scad_{cv}	Scad_{bic}	Lasso_{cv}	$\text{Lasso}_{\text{bic}}$	$\text{Alasso}_{\text{cv}}$	$\text{Alasso}_{\text{bic}}$	Arms	l_1-Arms
$\sigma = 1$	0.14	0.13	0.47	0.29	0.22	0.26	0.16	0.15
$\sigma = 3$	0.43	0.47	0.31	0.29	0.39	0.38	0.33	0.32
heavy tail	0.32	0.33	0.33	0.31	0.29	0.33	0.29	0.22

Table 3.9: Risk comparison for the high dimensional data with screening

We apply the selection methods with a R package 'EZPATH' (Yang and Zou, 2010, http://www.stat.umn.edu/∼yi) and use the default non-convex penalty parameters in the R package for SCAD, MCP and SICA. The tuning parameter of each method is selected by fivefold CV. Table 3.10 shows the performances of the selection and combining methods. For the p > n case, the performance of each selection method varies in different scenarios. On the other hand, the combining methods are always among the best. When there are outliers, they significantly outperform the selection methods.

	Scad	Lasso	Mcp	Sica	Arm	l_1-Arm
$\sigma = 1$	0.16	0.33	0.11	0.31	0.15	0.15
$\sigma = 3$	3.25	2.85	3.41	1.48	1.43	1.43
heavy tail	0.55	0.69	0.50	0.43	0.38	0.37

Table 3.10: Risk comparison for the high dimensional data (p > n)

3.3.4 Data examples

Data example 1. This data set is from the Berkeley Guidance Study (Weisberg 1985, p. 55-57). It contains 10 predictors, and the response variable is measurements of fatness for 32 girls at age 18. The training sample size is set to be 26 and the competing methods are evaluated by the remaining 6 observations. This process is repeated 100 times with random data splittings. The medians of the relative prediction MSE, i.e., the ratio of the MSE of these methods over that of the OLS estimator, is shown in Table 3.11. The a-LASSO with the fivefold CV selection performs the best, and the l_1-ARM performs almost as well as it.

Data example 2. This data set is taken from Johnson (1996). It originally contains

Scad_{cv}	Scad_{bic}	Lasso_{cv}	$\text{Lasso}_{\text{bic}}$	$\text{Alasso}_{\text{cv}}$	$\text{Alasso}_{\text{bic}}$	Arm	l_1-Arm
0.96	0.99	0.85	0.85	0.79	0.82	0.86	0.82

Table 3.11: Results for Data example 1

17 predictors and 252 observations. We take 16 predictors, removing the body density variable since it varies very narrowly, and 251 observations since the 42nd observation is apparently incorrect. The training sample size n is set to be 60 and 120 respectively. The median of the relative prediction MSE is represented in Table 3.12. When $n = 60$, the SCAD with the fivefold CV selection performs the best among the selection methods, while the l_1-ARM performs as well as it. When $n = 120$, the two SCAD selections outperform the other selections with a small margin. The l_1-ARM even improves over the SCAD methods. Note that when the training sample size increases, the OLS method performs better so that the relative advantage of the selection methods decreases. In these two real examples, the l_1-ARM outperforms the ARM.

	Scad_{cv}	Scad_{bic}	Lasso_{cv}	$\text{Lasso}_{\text{bic}}$	$\text{Alasso}_{\text{cv}}$	$\text{Alasso}_{\text{bic}}$	Arms	l_1-Arms
$n = 60$	0.77	0.82	0.91	0.84	0.91	0.93	0.90	0.77
$n = 120$	0.93	0.93	0.96	0.94	0.95	0.95	0.95	0.90

Table 3.12: Results for Data example 2

3.4 Theory

Differently from some previous work on consistency of model selection, our theoretical focus here is risk of prediction. Assume that $(\mathbf{x}, Y)$, $(\mathbf{x}_i, Y_i)$, $1 \leq i \leq n$, are independent and identically distributed. Let $\varepsilon = Y - f(\mathbf{x})$, where f is the regression function (conditional mean of Y given $\mathbf{x}$). To derive the theoretical result, we study a somewhat different version of the algorithm in section 2. The differences are: 1) A screening step is allowed to reduce the candidate model selection methods to be combined for save of computational cost and/or improving prediction accuracy; 2) The model selection methods are applied on part

of the data to come up with the models to be combined, which is mathematically tractable for theoretical investigation in terms of independence; 3) The weights are sequentially averaged. We recommend the earlier algorithm (with screening if so desired) in practice. First, with data partition the selected model by each method depends on the data splitting, which is undesirable. Second, when model selection methods are applied on the whole data (as in the algorithm in section 2) and the weighting is based on the same data, although there is the legitimate concern of data over-use, the step of data splitting and cross evaluation in that algorithm can guard against any serious bias. The previous numerical results clearly show that the practical algorithm works very well compared to the model selection methods, and a limited investigation found that while being much simpler, it performs similarly to or better than the theoretical algorithm given below.

We first split the data into two parts: $Z^{(1)}$ and $Z^{(2)}$. We obtain $\hat{\boldsymbol{\beta}}_j$ and $\hat{d}_j$ on $Z^{(1)}$ and D_j on $Z^{(2)}$ respectively for each candidate model j. Let $\hat{f}_j$ be the estimate of the regression function based on $Z^{(1)}$ from method j, $1 \le j \le K$. Consider a screening procedure that is applied on $Z^{(1)}$ to get a reduced list of candidate model selection methods, denoted by Γ_s with size K_s. Note that K_s is allowed to be random (depending on $Z^{(1)}$). For $i = n/2+1$, let $W_{j,i} = 1/K_s$ for $j \in \Gamma_s$ and for $n/2+1 \le i \le n$, let

$$W_{j,i} = \frac{(\hat{d}_j)^{-(i-n/2-1)} \exp(-\eta \sum_{l=n/2+1}^{i-1} |Y_l - \hat{f}_j(\mathbf{x}_l)|/\hat{d}_j)}{\sum_{k\in\Gamma_s} (\hat{d}_k)^{-(i-n/2-1)} \exp(-\eta \sum_{l=n/2+1}^{i-1} |Y_l - \hat{f}_k(\mathbf{x}_l)|/\hat{d}_k)}.$$

Define

$$\tilde{W}_j = \frac{1}{n/2} \sum_{i=n/2+1}^{n} W_{j,i},$$

and let

$$\tilde{f}(\mathbf{x}) = \sum_{j\in\Gamma_s} \tilde{W}_j \hat{f}_j(\mathbf{x})$$

be the combined estimator.

Condition 1: The true regression function $f(\cdot)$ is bounded by $\frac{A}{2}$ in absolute value for some positive constant A and the estimators $\hat{f}_j$, $j \in \Gamma$, are clipped accordingly.

Condition 2: The conditional variance $E\left(\varepsilon^2|\mathbf{x}\right)$ is uniformly upper bounded by some positive constant B^2 with probability 1.

Condition 3: There exist a constant $t_0 > 0$ and a monotone function $0 < H(t) < \infty$ on $[-t_0, t_0]$ such that for $-t_0 \leq t \leq t_0$, $E\left(\exp(t|\varepsilon|)|\mathbf{x}\right) \leq H(t)$ with probability 1.

Condition 4: Zero is a median of the distribution of the error ε conditional on $\mathbf{x}$.

For simplicity, we take $\hat{d}_{j,i} = 1$ in the following theorem. Let j_* be the minimizer of the risk $E|Y - \hat{f}_j|$ in Γ.

Theorem 3. *Assume Conditions 1-3 hold.*

1. *When the tuning parameter η is chosen small enough, we have*

$$E|Y - \tilde{f}(\mathbf{x})| \leq (A + B)\, P(j_* \notin \Gamma_s) + E|Y - \hat{f}_{j_*}(\mathbf{x})| + CE\left(\sqrt{\frac{\log(K_s)}{n}}\right),$$

   where C is a positive constant that depends on t_0, A, and B.

2. *If in addition, Condition 4 holds, we have*

$$E|Y - \tilde{f}(\mathbf{x})| \leq AP(j_* \notin \Gamma_s) + E|Y - \hat{f}_{j_*}(\mathbf{x})| + CE\left(\sqrt{\frac{\log(K_s)}{n}}\right).$$

Remarks:

1. The same risk bound holds (due to convexity) when data are randomly splitted multiple times and the resulting estimates $\tilde{f}(\mathbf{x})$ are averaged.

2. When no screening is done, the risk bounds becomes $E|Y - \tilde{f}(\mathbf{x})| \leq \inf_j (E|Y - \hat{f}_j(\mathbf{x})|) + C\sqrt{\frac{\log(K)}{n}}$. The theorem shows that the combined estimator behaves like the best $\hat{f}_j$ adaptively up to an additive penalty term of order $\sqrt{\frac{\log(K)}{n}}$, which cannot be generally improved in order.

3. The screening of model selection methods is different from screening of the original predictors. The predictions from the model selection methods can be highly correlated because they are based on the same data and models, which makes screening methods that rely on weak dependence of the predictors not applicable. Cross validation can be used, e.g., keep the top m methods for a pre-determined integer m. With a proper data splitting ratio, as shown in Yang (2007), under some conditions,

for any choice of $m \geq 1$, $P(j_* \notin \Gamma_s) \to 0$. If one method is taken as a reference and the methods that perform much worse are removed, the exclusion probability of j_* can be exponentially small.

4. The improvement of the risk bound under Condition 4 can be important when $E\left(\varepsilon^2|\mathbf{x}\right)$ is large.

3.5 Concluding remarks

Exciting new model selection methods have been derived from various perspectives. One may naturally consider a number of such methods so that there is a better chance that the best one works well for the data at hand.

For the goal of regression function estimation or prediction, through extensive simulations, we observe that three popular model selection methods, LASSO, SCAD, a-LASSO, behave the best in different aarios in terms of the model sparsity and model noise level. For tuning parameter selection, the use of BIC is not always better than fivefold CV, and in fact, it is sometimes much worse. In real applications, when the sample size is not large relative to the number of predictors, it is difficult to determine which selection method should be used and how to choose the tuning parameter in a specific case.

We propose robust adaptive regression by mixing, l_1-ARM, to aggregate the predictive strengths of the different selection methods so that it performs as if one knew which selection method is the best for each scenario in advance.

Both the theory and the numerical work support that for estimating the regression function or prediction, the l_1-ARM performs similarly well as the best candidate method in the individual scenarios. When various scenarios are considered, the l_1-ARM then shows its predictive advantage over the individual model selection methods. A contribution of the l_1-ARM is that it is more robust than the ARM when the underlying model tends to generate outliers. Furthermore, when there is no outlier it does not lose much efficiency compared to the ARM and they perform almost identically.

Finally, it should be pointed out that the l_1-ARM loses the model interpretability and it does not perform variable selection as the recent important model selection methods do.

3.6 Appendix: Proof of Theorem 3

Let $L(u) = |u|$, and $h(u) = \exp(-\eta L(u))$. Let $n_1 = n_2 = n/2$. Define $Q^{n_2} = \sum_{j\in\Gamma_s} \frac{1}{K_s}\Pi_{i=n_1+1}^{n} h(Y_i - \hat{f}_j(\mathbf{x}_i))$, where $\hat{f}_j(\mathbf{x}_i)$ is the prediction from the j^{th} method at time i. Then

$$-\log(Q^{n_2}) \le \log(K_s) + \eta \sum_{i=n_1+1}^{n} L(Y_i - \hat{f}_j(\mathbf{x}_i)).$$

On the other hand,

$$Q^{n_2} = \sum_{j\in\Gamma_s} \frac{1}{K_s} h(Y_{n_1+1} - \hat{f}_j(\mathbf{x}_{n_1+1})) \times \frac{\sum_{j\in\Gamma_s} h(Y_{n_1+1} - \hat{f}_j(\mathbf{x}_{n_1+1}))h(Y_{n_1+2} - \hat{f}_j(\mathbf{x}_{n_1+2}))}{\sum_{j\in\Gamma_s} h(Y_{n_1+1} - \hat{f}_j(\mathbf{x}_{n_1+1}))} \times$$
$$\cdots \times \frac{\sum_{j\in\Gamma_s} \Pi_{i=n_1+1}^{n} h(Y_i - \hat{f}_j(\mathbf{x}_i))}{\sum_{j\in\Gamma_s} \Pi_{i=n_1+1}^{n-1} h(Y_i - \hat{f}_j(\mathbf{x}_i))}.$$

That is, $Q^{n_2} = \Pi_{i=n_1+1}^{n} \sum_{j\in\Gamma_s} W_{j,i} h(Y_i - \hat{f}_j(\mathbf{x}_i))$. Accordingly, $-\log(Q^{n_2})$ $= -\sum_{i=n_1+1}^{n} \log\left(\sum_{j\in\Gamma_s} W_{j,i} h(Y_i - \hat{f}_j(\mathbf{x}_i))\right) = -\sum_{i=n_1+1}^{n} \log(E^J h(Y_i - \hat{f}_J(\mathbf{x}_i)))$, where E^J denotes the expectation with respect to J under the distribution $P(J = j) = W_{j,i}$ given each i for $j \in \Gamma_s$. Define $V = L(Y_i - \hat{f}_J(\mathbf{x}_i)) - E^J L(Y_i - \hat{f}_J(\mathbf{x}_i))$, and $\bar{f}_i(\mathbf{x}) = \sum_{j\in\Gamma_s} W_{j,i}\hat{f}_j(\mathbf{x})$. Then

$$\begin{aligned} E^J V^2 &= E^J(L(Y_i - \hat{f}_J(\mathbf{x}_i)) - E^J L(Y_i - \hat{f}_J(\mathbf{x}_i)))^2 \\ &\le E^J(L(Y_i - \hat{f}_J(\mathbf{x}_i)) - L(Y_i - E^J \hat{f}_J(\mathbf{x}_i)))^2 \\ &\le E^J(\hat{f}_J(\mathbf{x}_i) - E^J \hat{f}_J(\mathbf{x}_i))^2 \\ &= E^J(\hat{f}_J(\mathbf{x}_i) - \bar{f}_i(\mathbf{x}_i))^2. \end{aligned}$$

By Lemma 3.6.1 of Catoni (2004, p. 85), we get

$$\log(E^J h(Y_i - \hat{f}_J(\mathbf{x}_i))) \le -\eta E^J L(Y_i - \hat{f}_J(\mathbf{x}_i)) + I,$$

where $I \leq \frac{\eta^2}{2}\exp(\eta(|L(Y_i - \hat{f}_J(\mathbf{x}_i))| + \sup_{j\geq 1}|L(Y_i - \hat{f}_J(\mathbf{x}_i))|))E^J V^2$. Observe that

$$\begin{aligned}
I \leq & \frac{\eta^2}{2}\exp(2\eta \sup_{j\geq 1}|L(Y_i - \hat{f}_J(\mathbf{x}_i))|)E^J V^2 \\
\leq & \frac{\eta^2}{2}\exp(2\eta \sup_{j\geq 1}|Y_i - \hat{f}_J(\mathbf{x}_i)|)E^J V^2 \\
\leq & \frac{\eta^2}{2}\exp(2\eta(|Y_i - f(\mathbf{x}_i)| + \sup_{j\geq 1}|f(\mathbf{x}_i) - \hat{f}_J(\mathbf{x}_i)|))E^J V^2 \\
\leq & \frac{\eta^2}{2}e^{2\eta A}\exp(2\eta|\varepsilon_i|)E^J V^2 \\
\leq & \frac{\eta^2}{2}e^{2\eta A}\exp(2\eta|\varepsilon_i|)E^J(\hat{f}_J(\mathbf{x}_i) - \bar{f}_i(\mathbf{x}_i))^2,
\end{aligned}$$

where the second last inequality holds because of Condition 1. Under Condition 3, taking expectation with respect to the randomness of the errors ε_i and $\mathbf{x}_i$ for $n_1 + 1 \leq i \leq n$ conditional on the first n_1 observations, we have that when $2\eta \leq t_0$,

$$E_{n_1}(I) \leq \frac{\eta^2}{2}e^{2\eta A}H(2\eta)E_{n_1}(E^J(\hat{f}_J(\mathbf{x}_i) - \bar{f}_i(\mathbf{x}_i))^2).$$

Define a surrogate loss function $L_s(u) = L(u) + au^2$, $a > 0$. Let $b_0 = Y_i - \bar{f}_i(\mathbf{x}_i)$ and $b = Y_i - \hat{f}_J(\mathbf{x}_i)$. Then, as in Shan and Yang (2009),

$$\begin{aligned}
& L_s(b) - (2ab_0 + 1_{b_0\geq 0} - 1_{b_0<0})(b - b_0) - L_s(b_0) \\
& \quad = a(b - b_0)^2 + b(1_{b\geq 0} - 1_{b<0} + 1_{b_0<0} - 1_{b_0\geq 0}) \\
& \quad = a(b - b_0)^2 + \begin{cases} 0 & \text{if } b, b_0 \geq 0 \\ 0 & \text{if } b, b_0 < 0 \\ 2b & \text{if } b \geq 0 \text{ and } b_0 < 0 \\ -2b & \text{if } b < 0 \text{ and } b_0 \geq 0 \end{cases} \\
& \quad \geq a(b - b_0)^2.
\end{aligned}$$

Since $E^J(b - b_0) = 0$, we have $E^J L_s(b) - L_s(b_0) \geq aE^J(b - b_0)^2$, namely,

$$E^J L_s(Y_i - \hat{f}_J(\mathbf{x}_i)) - L_s(Y_i - \bar{f}_i(\mathbf{x}_i)) \geq aE^J(\hat{f}_J(\mathbf{x}_i) - \bar{f}_i(\mathbf{x}_i))^2.$$

Thus,

$$E_{n_1}(E^J L_s(Y_i - \hat{f}_J(\mathbf{x}_i)) - L_s(Y_i - \bar{f}_i(\mathbf{x}_i))) \geq aE_{n_1}(E^J(\hat{f}_J(\mathbf{x}_i) - \bar{f}_i(\mathbf{x}_i))^2).$$

Then we have

$$E_{n_1}(I) \leq \frac{a^{-1}\eta^2}{2} e^{2\eta A} H(2\eta) E_{n_1}(E^J L_s(Y_i - \hat{f}_J(\mathbf{x}_i)) - L_s(Y_i - \bar{f}_i(\mathbf{x}_i))).$$

Let $\frac{\eta}{2} \geq \frac{a^{-1}\eta^2}{2} e^{2\eta A} H(2\eta)$. Then we have

$$E_{n_1}(I) \leq \frac{\eta}{2} E_{n_1}(E^J L_s(Y_i - \hat{f}_J(\mathbf{x}_i)) - L_s(Y_i - \bar{f}_i(\mathbf{x}_i))).$$

Recall there have been two constraints on η:

$$\begin{aligned} 2\eta &\leq t_0, \\ \frac{\eta}{2} &\geq \frac{a^{-1}\eta^2}{2} e^{2\eta A} H(2\eta). \end{aligned}$$

When we choose a and η so that $\eta \leq \frac{t_0}{2}$ and $a \geq \eta e^{2\eta A} H(2\eta)$, the constraints are met. Let $a_\eta = \eta e^{t_0 A} H(t_0)$. With such a choice of η and a_η, we have,

$$\begin{aligned} & E_{n_1}(\log E^J \exp(-\eta L(Y_i - \hat{f}_J(\mathbf{x}_i)))) \\ \leq & -\eta E_{n_1}(L(Y_i - \bar{f}_i(\mathbf{x}_i))) + \eta E_{n_1}(L(Y_i - \bar{f}_i(\mathbf{x}_i)) - E^J L(Y_i - \hat{f}_J(\mathbf{x}_i))) \\ & + \frac{\eta}{2} E_{n_1}(E^J L_s(Y_i - \hat{f}_J(\mathbf{x}_i)) - L_s(Y_i - \bar{f}_i(\mathbf{x}_i))) \\ \leq & -\eta E_{n_1}(L(Y_i - \bar{f}_i(\mathbf{x}_i))) - \frac{\eta}{2} E_{n_1}(E^J L(Y_i - \hat{f}_J(\mathbf{x}_i)) - L(Y_i - \bar{f}_i(\mathbf{x}_i))) \\ & + \frac{a_\eta \eta}{2} E_{n_1}(E^J (Y_i - \hat{f}_J(\mathbf{x}_i))^2) - \frac{a_\eta \eta}{2} E_{n_1}((Y_i - \hat{f}_J(\mathbf{x}_i))^2) \\ \leq & -\eta E_{n_1}(L(Y_i - \bar{f}_i(\mathbf{x}_i))) + \frac{a_\eta \eta (A^2 + B^2)}{2}. \end{aligned}$$

The last inequality above holds because of the convexity of the function L and Conditions 1 and 2. Assume $j \in \Gamma_s$ for the moment. Then we have

$$\sum_{i=n_1+1}^{n} E_{n_1} L(Y_i - \bar{f}_i(\mathbf{x}_i)) \leq \frac{\log(K_s)}{\eta} + \frac{a_\eta (A^2 + B^2) n_2}{2} + \sum_{i=n_1+1}^{n} E_{n_1} L(Y_i - \hat{f}_j(\mathbf{x}_i)).$$

Under the iid assumption on the data and $n_2 = n/2$, we have

$$\frac{1}{n_2} \sum_{i=n_1+1}^{n} E_{n_1} L(Y - \bar{f}_i(\mathbf{x})) \leq E_{n_1} L(Y - \hat{f}_j(\mathbf{x})) + \frac{2\log(K_s)}{\eta n} + \frac{a_\eta (A^2 + B^2)}{2}.$$

With an optimal choice of η, $\eta' = \sqrt{\frac{4\log(K_s)}{ne^{t_0 A}H(t_0)(A^2+B^2)}}$, we have

$$\frac{1}{n_2}\sum_{i=n_1+1}^{n} E_{n_1} L(Y - \bar{f}_i(\mathbf{x})) \leq E_{n_1} L(Y - \hat{f}_j(\mathbf{x})) + C\sqrt{\frac{\log(K_s)}{n}},$$

where C is a positive constant depending on t_0, A, and B. By convexity of the function $L(x)$, together with that $\tilde{f} = \frac{1}{n_2}\sum_{i=n_1+1}^{n} \bar{f}_i$, we have for each $j \in \Gamma_s$, we have

$$E_{n_1}|Y - \tilde{f}(\mathbf{x})| \leq E_{n_1}|Y - \hat{f}_j(\mathbf{x})| + C\sqrt{\frac{\log(K_s)}{n}}.$$

Therefore, if $j^* \in \Gamma_s$, we have

$$E_{n_1}|Y - \tilde{f}(\mathbf{x})| \leq E_{n_1}|Y - \hat{f}_{j^*}(\mathbf{x})| + C\sqrt{\frac{\log(K_s)}{n}}.$$

When $j^* \notin \Gamma_s$, $E_{n_1}|Y - \tilde{f}(\mathbf{x})| \leq E_{n_1}|\varepsilon| + E_{n_1}|f(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq B + A$. Thus

$$E|Y - \tilde{f}(\mathbf{x})| \leq (B+A)\, P\left(j^* \in \Gamma_s^c\right) + E|Y - \hat{f}_{j^*}(\mathbf{x})| + CE\left(\sqrt{\frac{\log(K_s)}{n}}\right).$$

This risk bound can be improved if the error ε has a distribution with median 0 given $\mathbf{x}$, which is shown below. From before, if $j^* \in \Gamma_s$,

$$E_{n_1}|Y - \tilde{f}(\mathbf{x})| - E_{n_1}|\varepsilon| \leq E_{n_1}|Y - \hat{f}_{j^*}(\mathbf{x})| - E_{n_1}|\varepsilon| + C\sqrt{\frac{\log(K_s)}{n}},$$

and if $j^* \notin \Gamma_s$, $E_{n_1}|Y - \tilde{f}(\mathbf{x})| - E_{n_1}|\varepsilon| \leq E_{n_1}|f(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq A$. Then,

$$E_{n_1}|Y - \tilde{f}(\mathbf{x})| - E_{n_1}|\varepsilon| \leq AI_{j^* \in \Gamma_s^c} + \left(E_{n_1}|Y - \hat{f}_j(\mathbf{x})| - E_{n_1}|\varepsilon| + C\sqrt{\frac{\log(K_s)}{n}}\right) I_{j^* \in \Gamma_s},$$

where I denotes the indictor function. If the error distribution (given $\mathbf{x}$) has median zero, we must have $E_{n_1}|Y - \tilde{f}(\mathbf{x})| - E_{n_1}|\varepsilon| \geq 0$ with probability 1 and thus

$$E_{n_1}|Y - \tilde{f}(\mathbf{x})| - E_{n_1}|\varepsilon| \leq AI_{j^* \in \Gamma_s^c} + E_{n_1}|Y - \hat{f}_j(\mathbf{x})| - E_{n_1}|\varepsilon| + C\sqrt{\frac{\log(K_s)}{n}}.$$

That is,

$$E_{n_1}|Y - \tilde{f}(\mathbf{x})| \leq AI_{j^* \in \Gamma_s^c} + E_{n_1}|Y - \hat{f}_{j^*}(\mathbf{x})| + C\sqrt{\frac{\log(K_s)}{n}}.$$

Thus,

$$E|Y - \tilde{f}(\mathbf{x})| \leq AP(j^* \notin \Gamma_s) + E|Y - \hat{f}_{j^*}(\mathbf{x})| + CE\left(\sqrt{\frac{\log(K_s)}{n}}\right).$$

When there is no screening step, $\Gamma_s = \Gamma$ and $K_s = K$. Then we have

$$E|Y - \tilde{f}(\mathbf{x})| \leq E|Y - \hat{f}_{j_*}(\mathbf{x})| + C\sqrt{\frac{\log(K)}{n}}.$$

This completes the proof.

Chapter 4

Regression Based Forecast Combination Methods

4.1 Introduction

In their original papers, Bates and Granger (1969), and Newbold and Granger (1974) showed that combined forecasts may reduce prediction variability under the conditions that the forecasts are unbiased. They proposed several combining rules based on estimated variance-covariances of the forecast candidates. Granger and Ramanathan (1984) later expanded these variance-covariance methods in a regression framework. They argued that the variance-covariance methods can be treated as the least squares solutions under two constraints: one is that there is no constant term in the least squares formulation, the other is that all coefficients/weights are nonnegative and sum up to 1. They advocated use of regression methods that loosen these constraints for smaller mean squared prediction errors.

Since then, many regression based forecast combination methods have been proposed in the literature. For example, Diebold (1988) considered serial correlation in the least squares framework. Coulson and Robins (1993) included a lagged dependent variable besides the forecast candidates. Deutsch et al. (1994) addressed regime switches when estimating coefficients/weights. Interested readers are referred to Timmermann (2006) and references

therein.

Recently, researchers have worked on forecast combinations of a large number of forecasts in hope to take advantages of many different sources or models (e.g., Chan et al., 1999, Stock & Watson, 2003 and 2004, and Rapach & Strauss, 2005 and 2008). It has been shown, however, that the ordinary regression combination is not optimal for this kind of scenarios due to high variance. The empirical evidence of poor performances of the large-number regression combinations is provided by Rapach & Strauss (2005, 2008), among many others.

Chan et al. (1999) proposed the use of James-Stein estimation, ridge regression, and principle components regression as alternatives to ordinary least squares. Swanson and Zeng (2001) proposed regression combinations based on subset selections using AIC (Akaike, 1973) or BIC (Schwarz, 1978) to choose the best subset of all possible forecast candidates. Assume that there are p forecast candidates, by Swanson and Zeng's method, one has to select among in total $2^p - 1$ models, which is not practical when p is relatively large. Alternative to using AIC or BIC, one may apply subset selections by using t-statistic (e.g., Swanson and Zeng, 2001). However, this method often performs poorly in real applications (e.g., Rapach & Strauss, 2005 and 2008).

In this work, we first consider sequential subset selections in contrast to all subset selections, which reduce the number of models fitted to be at most $p(p+1)/2$. Simulations and real data examples show that sequential subset selections substantially improve upon ordinary least squares. Although sequential subset selections are commonly used to choose explanatory variables or orders in ARIMA modeling (see Zou & Yang, 2004, for related issues), to our knowledge, they have not been discussed in the forecast combination framework. From our numerical studies, sequential subset selections are a valuable technique for large-number forecast combinations.

We also propose a novel regression based combination method, the decreasingly averaging method. Sequential subset selections discard insignificant forecast candidates using AIC or BIC. In contrast, the decreasingly averaging method retains all forecast candidates, but simultaneously stabilizes and slowly shrinks their coefficients/weights according to their order of appearance in the process of sequential selections. The less significant the candidate, the more to be shrunk, thus the less effect on the combined forecasts. This

is different from the existing Bayesian shrinkage methods (Stock & Watson, 2004), which shrink towards equal weights. Sequential subset selections and the decreasingly averaging method can be easily implemented. Actually, they can be tools to help other combination methods improve prediction accuracy especially in the large-number combination cases as long as their weights are determined based on ordinary least squares. For instance, it has been pointed out that Bayesian shrinkage methods have variable performance across different occasions or forecast horizons (Stock & Watson, 2004, and Rapach & Strauss, 2005 and 2008). In one real example which follows in section 4, we show that the proposed methods can help Bayesian shrinkage methods perform more stably and competitively compared to other combination methods.

Recently, Hansen (2008) proposed Mallows Model Averaging (MMA) for forecast combinations and found that the MMA method compared favorably with other feasible forecasting methods in terms of the one-step-ahead mean squared forecast error. However, it is not applicable in the present setting of nonnested forecasting models, which are often encountered in real applications.

Yang (2004) pointed out there are two main directions of forecast combinations in the literature: combining for adaptation and combining for improvement. The first one targets the best individual performance among the pool of forecast candidates. The second one aims at significantly outperforming each individual forecast candidate. The early variance-covariance and regression based combination methods (including the proposed methods in this work) fall in the second direction. Interested readers are referred to Wei and Yang (2008) and references therein for some relevant work in the first direction. Different Bayesian combination methods can be categorized into either the first direction (e.g., Wright, 2003) or the second direction (e.g., Palm & Zellner, 1992).

The rest of the paper is organized as follows. In section 2, we propose sequential subset selections and the decreasingly averaging method. In section 3, we present simulation results for the proposed methods. In section 4, the proposed methods are examined through three data examples. Concluding remarks are given in section 5.

4.2 Methodologies

In this section, we first propose sequential subset selections and then the decreasingly averaging method. We shall also discuss the performance measures that will be used in next simulations and real data examples.

4.2.1 Sequential subset selections

Assume that there is a time series which we are interested in for forecasting, $y_t, t = 1, 2, \cdots$, and there are p forecast candidates, $x_{1t}, x_{2t}, \cdots, x_{pt}, t = 1, 2, \cdots$. Granger and Ramanathan (1984) suggested the forecast combination:

$$y_{t+h} = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_{it} + \varepsilon_{t+h},$$

where h is the forecast horizon, and the coefficients/weights can be estimated by least squares, possibly under some constraints. Then the combined forecast is given by

$$x_t^c = \hat{\alpha}_0 + \sum_{i=1}^{p} \hat{\alpha}_i x_{it}.$$

When p is large relative to the forecast sample size, the total variability of the coefficients/weights estimations is substantial and thus hurts the performance of the combined forecast x_t^c. A natural solution to this issue is to select the subset of the most significant forecast candidates, discarding others, in the regression formulation. The most commonly used selection criteria are AIC and BIC. AIC measures the discrepancy between the true model and a fitted model, while BIC approximates the posterior probabilities in a Bayesian framework. Assume that the number of parameters in the fitted model is k, and the sample size is n. AIC and BIC are both of the form

$$\text{-log(maximized likelihood)} + \text{penalty},$$

where the penalty is k in AIC, or $k*\log(n)/2$ in BIC. The model that minimizes the criterion is selected.

When p is large, direct applications of AIC or BIC over all subset models is infeasible. Sequential selections are a practical approach to proceed. To apply sequential subset

selections to the p forecast candidates, we first choose the most significant one which minimizes AIC or BIC among the p candidates. Then we update the previous model by adding the second most significant one among the remaining $p-1$ candidates. If the AIC (or BIC) of the updated model is greater than that of the previous model, we stop and the previous model is our final choice. Otherwise, we continue to sequentially add one candidate a time until all of the p forecast candidates are exhausted.

4.2.2 The decreasingly averaging method

In this method, similarly to above, we first determine the most significant candidate using AIC or BIC, denoting it by $x_{(1)}$. Based on $x_{(1)}$, we determine the second most significant cadidate, denoting it by $x_{(2)}$, and so on, until the least significant candidate, denoting it by $x_{(p)}$. Note that AIC and BIC provide exactly the same order of the p candidates. Then we fit the following p nested models:

$$\begin{aligned}
y_{t+h} =& \alpha_0 + \alpha_1 x_{(1)t} + \varepsilon_{t+h}, \\
y_{t+h} =& \alpha_0 + \alpha_1 x_{(1)t} + \alpha_2 x_{(2)t} + \varepsilon_{t+h}, \\
& \cdots \\
y_{t+h} =& \alpha_0 + \alpha_1 x_{(1)t} + \alpha_2 x_{(2)t} + \alpha_3 x_{(3)t} + \cdots + \alpha_p x_{(p)t} + \varepsilon_{t+h}.
\end{aligned}$$

We obtain a p by $(p+1)$ matrix of the estimated coefficients:

$$\begin{pmatrix}
\hat{\alpha}_0^1 & \hat{\alpha}_1^1 & 0 & \cdots & 0 \\
\hat{\alpha}_0^2 & \hat{\alpha}_1^2 & \hat{\alpha}_2^2 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\hat{\alpha}_0^p & \hat{\alpha}_1^p & \hat{\alpha}_2^p & \cdots & \hat{\alpha}_p^p
\end{pmatrix}$$

where the superscript $i, i = 1, 2, \cdots, p$, denotes the ith nested model, and the ith row has $p-i$ zeros to represent the coefficients of the candidates which are not in the model. We then take arithmetic averages over each of the columns, and denote them by

$$\begin{pmatrix} \tilde{\alpha}_0 & \tilde{\alpha}_1 & \tilde{\alpha}_2 & \cdots & \tilde{\alpha}_p \end{pmatrix}$$

Finally, the combining weights of the p candidates (plus an intercept) are given by the averaged coefficients. The intercept and the coefficient of the most significant candidate

are stabilized through averaging the p models. Other coefficients are not only stabilized but forced to slightly shrink. The shrinkage occurs when the zeros are taken into the average calculations. The degrees of shrinkage of the coefficients of the candidates are in some sense proportional to the degrees of their insignificance. For instance, $\tilde{\alpha}_p = \frac{1}{p}\hat{\alpha}_p^p$. The final weight of the least significant candidate is $\frac{1}{p}$ of its coefficient in the ordinary regression combination. Therefore, the insignificant candidates are of less importance, but still playing a role, in the forecast combinations. Note that the decreasingly averaging method is very different from simply averaging the forecast candidates.

4.2.3 Performance measures

In our simulations, the time series $y_t, t = 1, 2, \cdots$, is generated by a mean function plus a random error. The conditional mean squared h-step-ahead forecasting error $E(y_{t+h} - x_t^c)^2$ on current time t is actually the sum of the squared conditional bias and conditional variance:

$$E(y_{t+h} - x_t^c)^2 = (m_{t+h|t} - x_t^c)^2 + v_{t+h|t},$$

where $m_{t+h|t}$ is the mean, and $v_{t+h|t}$ is the error variance, conditional on current time t. Since the conditional variance $v_{t+h|t}$ is always the same no matter which combination method is used for prediction, when comparing the performance of different combination methods, we may only consider the squared conditional bias as the net loss:

$$L(m_{t+h|t}, x_t^c) = (m_{t+h|t} - x_t^c)^2.$$

Accordingly, the corresponding net risk may be defined by $E(m_{t+h|t} - x_t^c)^2$. When evaluating the performance of a series of combined forecasts x_t^c, $t = 1, 2, \cdots, l$, we consider the average forecasting risk

$$\text{Ave. Risk} = \frac{1}{l}\sum_{t=1}^{l} E(m_{t+h|t} - x_t^c)^2.$$

This risk will be used as the performance measure in the following simulation investigations. In real data applications, since the above risk is unkownable, we instead consider the mean square prediction error

$$\text{MSE} = \frac{1}{l}\sum_{t=1}^{l}(y_{t+h} - x_t^c)^2.$$

4.3 Simulations

In this section, we extensively examine the performances of the proposed methods under random model settings. We do not limit our focus on some individual models. In contrast, we evaluate the behavior of the proposed methods on a number of randomly generated models. The simulation results under random model settings can provide us more fair and informative understanding of the proposed methods than those of some specific models.

We consider two kinds of scenarios. One is that the random model has a fixed order, and the other is that the random model has various orders. Under each scenario, we consider two cases. One is that the true model is in the forecast candidate set, and the other is that the true model is not. Sequential subset selections and the decreasingly averaging method can be easily applied to various combination methods based on ordinary least squares for prediction improvement. For instance, in Coulson and Robin's (1993) method, where the coefficients/weights of the forecast candidates plus a lagged dependent variable are determined by ordinary least squares, one can apply sequential subset selections to leave out insignificant terms to reduce the total estimation variability, or the decreasingly averaging method to simultaneously stabilize and shrink the coefficients/weights to improve upon the original versions. In this and following sections, we focus on applying the proposed methods to the ordinary regression combinations and Bayesian shrinkage methods.

Stock and Watson (2004) proposed Bayesian shrinkage methods as follows

$$w_i = \lambda\hat{\alpha}_i + (1-\lambda)(1/p)$$

where w_i is the weight of the ith candidate, and $\hat{\alpha}_i$ is the estimated coefficient of the ith candidate in an ordinary regression combination which does not include an intercept. λ is a shrinkage tuning parameter, and is equal to $max\{0, 1-\kappa[p/(n-p)]\}$. In the following simulations and data examples, we consider κ equal to 0.5 and 1 respectively. When κ is large (λ is small), the weight shrinks toward equal weights from the least squares estimation. Diebold and Pauly (1990) pointed out that this kind of weight w_i can be interpreted as a Bayesian estimator.

4.3.1 Random models with a fixed order

In this sub-section, we first consider that the true model has a fixed order AR(4) with 9 candidate models which are a white noise, AR(1) to AR(4), and MA(1) to MA(4), respectively. Note that the true model is in the candidate set. For this and the following simulations, the error term of the true model follows a normal distribution with mean zero and variance 4. We generate 100 models with coefficients randomly generated from the uniform distribution on $[-1, 1]$ (non-stationary coefficients are discarded). We replicate each model and the candidate forecasts 100 times to simulate the forecasting risks. In each replication, we generate a sample with size 140. The candidate models start to generate one-step-ahead forecasts after 80 observations, then recursively do so once every additional observation is made. Thus there are in total 60 forecasts for each model. The combination methods use 40 forecasts to start weighting, evaluated on last 20 observations.

We consider two types of ordinary least squares combinations. One is with recursive fitted sizes, which start from 40 and sequentially increase by one until 59. The other is with rolling fitted sizes, which always use the most recent 40 forecasts. We denote the two methods by re-OLS and ro-OLS respectively. Furthermore, we consider two Bayesian shrinkage methods as described above (denoted by Shk-1 and Shk-2 respectively). We apply the four methods to determine the combination weights of the 9 forecast candidates. We also apply the proposed methods to these four methods. When applying AIC or BIC selection, some insignificant forecast candidates are left out, and the four existing methods determine the coefficients/weights based on the remaining forecast candidates. When applying the decreasingly averaging method (denoted by DA), all of the 9 forecast candidates remain in the four existing methods, but their (OLS) coefficients are adjusted by DA. Table 1 shows the results of the comparisons. The second column gives the means of the 100 forecasting risks of the four existing combination methods. The rest of the columns give respectively the means of the 100 forecasting risks of the proposed methods applied to the existing methods. The numbers in parentheses are the improvement pencentages of the proposed methods over the existing methods.

From Table 1, the recursive OLS outperforms the rolling OLS, while Bayesian shrinkage methods outperform the recursive OLS. All of the proposed methods can significantly

		DA	AIC	BIC
re-OLS	1.53	0.82 (46%)	0.84 (45%)	0.69 (55%)
ro-OLS	2.12	1.01 (52%)	0.98 (54%)	0.79 (63%)
Shk-1	1.08	0.60 (45%)	0.61 (44%)	0.51 (53%)
Shk-2	1.02	0.66 (35%)	0.58 (43%)	0.50 (51%)

Table 4.1: Comparisons when the true model AR(4) is in the candidate set.

improve upon the existing combination methods. When the true model is in the candidate set, BIC selection performs the best among the proposed methods, while the decreasingly averaging method performs similarly as AIC selection.

We then consider that the random true model has a fixed order AR(6) with the same 9 candidate models as in the previous experiment. Note that in this experiment the true model is not in the candidate set. Other settings remains the same as before. Table 2 shows the results of the comparisons.

		DA	AIC	BIC
re-OLS	3.18	2.28 (28%)	2.68 (16%)	2.63 (17%)
ro-OLS	3.80	2.44 (36%)	2.93 (23%)	2.77 (27%)
Shk-1	2.62	2.07 (21%)	2.40 (8%)	2.46 (6%)
Shk-2	2.53	2.19 (14%)	2.35 (7%)	2.43 (4%)

Table 4.2: Comparisons when the true model AR(6) is not in the candidate set.

The numbers in Table 2 are substantially bigger than those in Table 1 because in this experiment the true model is not in the candidate set. Every combination method produces bigger forecasting risks in this situation. The proposed methods, however, still significantly improve upon the existing methods. Furthermore, there are a couple of interesting phenomena worth mentioning. When the true model is not in the candidate set, the decreasingly averaging method performs the best among the proposed methods, while AIC and BIC selections perform similarly. In reality, it is usually the case where no model in the candidate set truly describes the underlying data generating process. Thus one may usually have many candidate models to entertain but without the true model. From this experiment,

the decreasingly averaging method shows potential advantages in real applications.

4.3.2 Random models with various orders

In previous sub-section, the random true model has a fixed order (either AR(4) or AR(6)). In this sub-section, we consider the true model has different orders since in reality the underlying data generating process may have a structural change. We first consider that the true model uniformly varies from AR(1) to AR(4), while the candidate models are the same as in previous experiments. Note that even though the true model varies, it is still in the candidate set. Other settings remains the same as in the previous experiments. Table 3 shows the results of the comparisons.

		DA	AIC	BIC
re-OLS	1.40	0.77 (45%)	0.76 (46%)	0.62 (56%)
ro-OLS	1.91	0.95 (50%)	0.90 (53%)	0.72 (62%)
Shk-1	0.95	0.54 (44%)	0.54 (44%)	0.44 (54%)
Shk-2	0.86	0.55 (36%)	0.52 (39%)	0.43 (50%)

Table 4.3: Comparisons when the true model varies and is in the candidate set.

It is of interest to compare Table 3 with Table 1. Even though the four existing methods yield different means of forecasting risks in the two tables, the proposed methods make similar improvements over the existing methods. Again in Table 3, when the true model is in the candidate set, BIC selection performs the best among the proposed methods, while the other two perform similarly.

We then consider that the random true model uniformly varies from AR(5) to AR(7) with the same 9 candidate models. Note that the true model is not in the candidate set. Other settings remains the same as before. Table 4 shows the results of the comparisons.

From Table 4, the proposed methods significantly improve upon the four existing methods. As in Table 2, when the true model is not in the candidate set, AIC and BIC selections perform similarly, while the decreasingly averaging method significantly outperforms both of them, which again shows its potential advantages in real applications.

		DA	AIC	BIC
re-OLS	2.64	1.87 (29%)	2.17 (18%)	2.10 (20%)
ro-OLS	3.14	2.01 (36%)	2.34 (26%)	2.22 (29%)
Shk-1	2.15	1.68 (22%)	1.94 (10%)	1.93 (10%)
Shk-2	2.13	1.83 (14%)	1.90 (11%)	1.90 (11%)

Table 4.4: Comparisons when the true model varies and is not in the candidate set.

So far we have dealt with the cases where the random error of the true model follows a normal distribution with mean zero and variance 4. Alternatively, we also consider the random error taking different variances from 0.5 to 7 and different distributions such as shifted gamma (with mean zero), double exponential, and t. We obtain similar results as presented in this paper, which are available upon request.

4.4 Data examples

In this section, we apply the proposed methods to three real data sets with a focus on the third one where we compare the proposed methods with other existing combination methods across different forecast horizons.

4.4.1 Data set 1

The data with $n = 98$ are levels of Lake Huron measured in each July from 1875 through 1972 (Brockwell & Davis, 1991). Graphical inspectations suggest differencing the data. The candidates are ARMA(p,q) models with $p, q = 0, 1, 2$. The training sample size for the candidate models is 57. Then we obtain 40 one-step-ahead forecasts for each model. The combination methods use the beginning 20 forecasts to calculate the initial coefficients/weights. We compare the performance of the combination methods over the last 20 observations. Table 5 gives the comparison results. The second column gives the MSEs of the existing combination methods. The rest of the columns give respectively the MSEs of the proposed methods applied to the existing methods.

In Table 5, the three proposed methods dramatically improve upon the recursive OLS

		DA	AIC	BIC
re-OLS	1.16	0.78 (33%)	0.76 (35%)	0.82 (30%)
ro-OLS	1.74	1.07 (38%)	1.04 (40%)	0.73 (58%)
Shk-1	0.85	0.73 (15%)	0.77 (9.8%)	0.85 (0.4%)
Shk-2	0.72	0.68 (4.9%)	0.76 (-5.9%)	0.84 (-17%)

Table 4.5: Comparison results of data set 1.

method, and they perform comparably. The proposed methods also dramatically improve upon the rolling OLS method with BIC selection standing out. The decreasingly averaging method incorporates Bayesian shrinkage methods favorably compared to sequential subset selections.

4.4.2 Data set 2

The data are aggregated Australian clay brick quarter productions (in million units) from March 1956 through September 1994 (Makridakis et al., 1998). The data set consists of 155 observations. After taking a log transformation, we difference the data to improve the stationarity. The candidates are ARMA(p,q) models with $p, q = 0, 1, 2, 3, 4, 5$ (discard the case if the AR parts are not stationary). We obtain 20 candidate models. The training sample size for the candidates models is 100. There are 54 one-step-ahead forecasts for each model. The combination methods use the beginning 34 forecasts to calculate the initial coefficients/weights. We compare the performance of the combination methods over the last 20 observations. Table 6 gives the comparison results (MSE$\times 10^3$).

		DA	AIC	BIC
re-OLS	6.5	4.9 (25%)	5.3 (18%)	6.7 (-3.1%)
ro-OLS	37.2	7.4 (80%)	5.4 (85%)	5.3 (86%)
Shk-1	5.4	5.1 (5.6%)	5.7 (-5.6%)	4.4 (19%)
Shk-2	6.0	6.0 (0.0%)	5.6 (6.7%)	4.4 (27%)

Table 4.6: Comparison results of data set 2.

In Table 6, most of the proposed methods significantly outperform the two regression combination methods. For this data set, BIC selection incorporates Bayesian shrinkage methods favorably, while the decreasingly averaging method make an improvement by 5.6% or remains the same performance as the original Bayesian shrinkage method.

4.4.3 Data set 3

Rapach and Strauss (2005) studied the large-number combinations for forecasting employment growth in Missouri using 22 candidate models. They examined in total 20 different combination methods, where the recursive and rolling OLS and Bayesian shrinkage methods are all included (for more details about the data set, forecast candidates, and combination methods, please check their article). The Missouri employment growth data set spans from January 1976 to January 2005, and the combination methods are evaluated over the last 10 years. There are four forecast horizons considered, $3, 6, 12$, and 24 months. Before we discuss the proposed methods, we present the MSEs of the best candidate, best combination, and simple average across the four horizons in Table 7. [1] To be conformable to Rapach and Strauss (2005), the entries in Table 7 and the following Table 8 are ratios of the MSEs of the methods to that of an AR benchmark model.

	Best ind.	Best com.	Simple average
h=3	0.90	0.94	0.96
h=6	0.83	0.91	0.92
h=12	0.70	0.71	0.84
h=24	0.76	0.51	0.83

Table 4.7: The MSEs of some methods of data set 3.

From Table 7, we can find that the simple average method performed very well at short horizons (h = 3 or 6), and the best combined forecast (which is Shk-2) significantly outperformed the best individual forecast candidate at h = 24. Table 8 shows the results when we apply the proposed methods to the recursive and rolling OLS and Bayesian

[1] We rewrote the whole program in R statistical software, and found that there are very minor differences between our numerical results and those of Rapach and Strauss (2005).

shrinkage methods.

			DA	AIC	BIC
h=3	re-OLS	1.44	1.15 (20%)	1.24 (14%)	0.93 (36%)
	ro-OLS	2.28	1.75 (23%)	2.05 (10%)	1.51 (34%)
	Shk-1	1.18	1.04 (12%)	1.18 (0.0%)	0.89 (24%)
	Shk-2	1.08	0.99 (9.1%)	1.13 (-4.6%)	0.89 (18%)
h=6	re-OLS	1.56	1.24 (21%)	1.27 (19%)	1.10 (29%)
	ro-OLS	2.45	1.89 (23%)	2.15 (12%)	1.68 (31%)
	Shk-1	1.21	1.01 (17%)	1.05 (13%)	0.97 (20%)
	Shk-2	1.02	0.91 (11%)	1.00 (2.0%)	0.96 (5.9%)
h=12	re-OLS	1.27	1.11 (13%)	1.24 (2.4%)	1.13 (11%)
	ro-OLS	2.81	2.13 (24%)	2.49 (11%)	2.30 (18%)
	Shk-1	0.85	0.79 (7.1%)	0.95 (-12%)	0.92 (-8.2%)
	Shk-2	0.71	0.71 (0.0%)	0.91 (-28%)	0.91 (-28%)
h=24	re-OLS	0.95	0.75 (21%)	0.83 (13%)	0.75 (21%)
	ro-OLS	1.62	1.28 (21%)	1.54 (4.9%)	1.25 (23%)
	Shk-1	0.51	0.55 (-7.8%)	0.54 (-5.9%)	0.52 (-2.0%)
	Shk-2	0.53	0.57 (-7.5%)	0.53 (0.0%)	0.52 (1.9%)

Table 4.8: Comparison results of data set 3 across different forecast horizons.

From Table 8, we can find that the proposed method can significantly improve upon the recursive and rolling OLS methods, and at $h = 24$, they perform very well, reaching 0.75. More interesting things happen to Bayesian shrinkage methods. We can find that Bayesian shrinkage methods have variable performance across different horizons. They performed very well at long horizons ($h = 12$ or 24), but poorly at short horizons ($h = 3$ or 6). However, BIC selection plus Bayesian shrinkage can reach 0.89 at $h = 3$, and the decreasingly averaging method plus Bayesian shrinkage can reach 0.91 at $h = 6$. The decreasingly averaging method plus Bayesian shrinkage makes improvement by 6.7% or remains the same as the original Bayesian shrinkage method at $h = 12$. The proposed methods plus Bayesian shrinkage methods have slightly worse performance than the original Bayesian shrinkage methods at $h = 24$, but still significantly outperform other combination methods. If we simply incorporate the decreasingly averaging method with the second Bayesian shrinkage method, we will reach $0.99, 0.91, 0.71$, and 0.57, at the four horizons

respectively, which makes the Bayesian shrinkage method the most attractive method out of the 20 combination methods across different forecast horizons.

4.5 Concluding remarks

The Least squares combinations (Granger & Ramanathan, 1984) are an important development in the forecast combination literature. The methods include the early variance-covariance methods as their special cases in some sense. Recently, researchers have worked on large-number forecast combinations. It has been shown that ordinary least squares combinations of all forecast candidates may have very poor performance in such situations. Due to computational difficulty, all subset selections are unattractive. As a solution, we propose two approaches, sequential subset selections and the decreasingly averaging method. The proposed methods are easily implemented, and can be tools to help various combination methods to improve prediction accuracy as long as their coefficients/weights are determined based on ordinary least squares. In this work, we focus on applying the proposed methods on the ordinary regression combinations and Bayesian shrinkage methods.

Sequential subset selections discard insignificant forecast candidates to reduce the variability of coefficient/weight estimations, leading to possibly improved predictions. The decreasingly averaging method retains all the candidates, but simultaneously stabilizes and slowly shrinks the coefficients/weights according to their significance, which is different from Bayesian shrinkage methods, which shrink towards equal weights.

We conduct structured simulations to examine the performance of sequential subset selections and the decreasingly averaging method. The numerical results show the proposed methods can significantly improve upon the recursive and rolling OLS and Bayesian shrinkage methods. When the true model is in the candidate set, BIC performs the best among the proposed methods, while the other two perform similarly. When the true model is not in the candidate set, the decreasingly averaging method significantly outperforms AIC and BIC selections, while AIC and BIC selections perform similarly. Three real data examples also confirm the potential advantages of the proposed methods. Especially, in data set 3, we examine their performance in a comprehensive setting, comparing them with 20 different combination methods. We find that the proposed methods can help Bayesian

shrinkage methods improve prediction accuracy at short horizons. In particular, the decreasingly averaging method can help the second Bayesian shrinkage method be the most attractive combination method out of the 20 combination methods across different forecast horizons.

The theoretical understanding of the decreasingly averaging method remains future investigations. Another direction of future work is to examine the proposed methods on other combination methods based on ordinary least squares.

Chapter 5

Conclusion and Discussion

There are two directions of model combinations. One is combining for adaptation and the other is combining for improvement, The second direction comes with a much slower convergence rate compared with the first direction. In this work, to achieve the goal of combining forecasts for adaptation, we propose two robust AFTER algorithms: L_1-AFTER and h-AFTER. Non-asymptotic risk bounds hold for these methods.

The original AFTER algorithm is based on the squared error loss. L_1-AFTER uses the absolute forecast error loss and h-AFTER is based on the Huber loss, which combines the behaviors of the squared forecast error loss and the absolute forecast error loss. The new methods alleviate the influence of forecast outliers. The simulation results and real data examples suggest that they significantly outperform AFTER when the error is asymmetric or when outliers commonly occur. Very importantly, when the errors are normally distributed, they usually perform only slightly worse than AFTER.

Based on the theoretical and numerical investigations, the robust AFTER methods have very stable and reliable performance when the goal is combining forecasts for adaptation. A future direction is to extend the present theoretical work to deal with error distributions with polynomially decaying (or even heavier) tails.

Exciting new model selection methods have been derived from various perspectives, among which, LASSO, SCAD, and adaptive LASSO are three popular ones. For the goal of regression function estimation or prediction, through extensive simulations, we observe

that the three popular model selection methods behave the best in different scenarios in terms of the model sparsity and model noise level. For choosing the tuning parameter, BIC selection is not always better than fivefold CV. In real applications, when the sample size is not large relative to the number of predictors, it is difficult to determine which selection method should be used and how to choose the tuning parameter in a specific case.

We propose robust adaptive regression by mixing, l_1-ARM, to aggregate the predictive strengths of the different selection methods so that it performs as if one knew which selection method is the best for each scenario in advance. The numerical results confirm that the l_1-ARM performs similarly well as the best candidate method in the individual scenarios. When various scenarios are considered, the l_1-ARM then shows its predictive advantage over the individual model selection methods.

A contribution of the l_1-ARM is that it is more robust than the ARM when the underlying model tends to generate outliers. Furthermore, when there is no outlier it does not lose much efficiency compared to the ARM and they perform almost identically. Finally, it should be pointed out that the l_1-ARM loses the model interpretability and it does not perform variable selection as the recent important model selection methods do.

Granger and Ramanathan (1984) proposed least squares forecast combinations, which were an important development in the forecast combination research areas. The methods included the early developed forecast combination methods as their special cases and built up connections with rich regression analysis theories and applications. However, it has been shown that ordinary least squares combinations may have very poor performance in some situations. In the third part of the dissertation, we propose a novel regression based combination method, the decreasingly averaging method. The proposed method is easily implemented, and can be tools to help various combination methods to improve prediction accuracy as long as their coefficients/weights are determined based on ordinary least squares.

The decreasingly averaging method retains all the candidates, but simultaneously stabilizes and slowly shrinks the coefficients/weights according to their significance, which is different from Bayesian shrinkage methods, which shrink towards equal weights. We conduct structured simulations to examine the performance of the decreasingly averaging method. The numerical results show the proposed method can significantly improve

upon the recursive and rolling OLS and Bayesian shrinkage methods. The theoretical understanding of the decreasingly averaging method remains future investigations. Another direction of future work of the decreasingly averaging method is to investigate its performance on other combination methods based on ordinary least squares.

Chapter 6

Reference

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle". In: B. N. Petrov, & F. Csaki (Eds.), Proceedings of the 2nd International Symposium on Information Theory, Budapest: Akademia Kiado.

Altavilla, C. and Grauwe, P. D. (2006). Forecasting and Combining Competing Models of Exchange Rate Determination. CESifo Working Paper No. 1747.

Barron, A. R., Birgé, L., and Massart, P. (1999). Risk Bounds for Model Selection via Penalization. Probability Theory and Related Fields, V.113, 301-413.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. Operations Research Quarterly, V.20, 451-468.

Birgé, L. (2006). Model Selection via Testing: An Alternative to (Penalized) Maximum Likelihood Estimators. Ann. I. H. Poincaré, PR 42, 273-325.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model Selection: An Integral Part of Inference. Biometrics, V.53, 603-618.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). Aggregation for Gaussian regression. The Annals of Statistics, V.35, 1674-1697.

Catoni, O. (2004). Statistical Learning Theory and Stochastic Optimization. New York: Springer.

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

Chan, Y. Z., Stock, J. H., and Watson, M. W. (1999). "A dynamic factor model framework for forecast combination". Spanish Economic Review, V.1, 91-121.

Chen, L. and Yang, Y. (2010). Combining Statistical Procedures. In *Frontiers of Statistics* book series.

Chen, Z. and Yang, Y. (2007). Time Series Models for Forecasting: Testing or Combining?. Studies in Nonlinear Dynamics & Econometrics, Vol.11, No.1, Article 3.

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

Coulson, N. E. and Robins, R. P. (1993). "Forecasting combination in a dynamic setting". Journal of Forecasting, V.12, 63-67.

Diebold, F. X. (1988). "Serial Correlation and the Combination of Forecasts". Journal of Business & Economic Statistics, V.6, 105-111.

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

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

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

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

Fan, J. and Li, J. (2001). Variable Selection via Nonconave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space (with discussion). *JRSSB* **70**, 849-911.

Fan, S., Chen, L., and Lee, W. J. (2008). Short-term Load Forecasting Using Comprehensive Combination based on Multi- Meteorological Information. IAS Annual Conference.

Garratt, A., Lee, K., Pesaran, H. M., and Shin, Y. (2003). Forecast Uncertainties in Macroeconomic Modeling: An Application to the U.K. Economy. JASA, V.98, 829-838.

Gouriéroux, C. and Monfort, A. (1992). Qualitative Threshold ARCH Models. Journal of Econometrics, V.52, 159-199.

Granger, C. W. J. and Ramanathan, R. (1984). Improved methods of Forecasting. Journal of Forecasting, V.3, 197-204.

Greene, W. H. (2000). Econometric Analysis (4th edition). New York: Prentice Hall.

Hallman, J. and Kamstra, M. (1989). Combining Algorithms Based on Robust Estimation Techniques and Co-Integrating Restrictions. Journal of Forecasting, V.8, 189-198.

Hansen, B. E. (2008). Least Squares Forecast Averaging. Journal of Econometrics, V.146, 342-350.

Haussler, D., Kivinen, J. and Warmuth, M. (1998) Sequential Prediction of Individual Sequences under General Loss Functions. IEEE Transactions on Information Theory, V.44, 1906-1925.

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

Huber, P. J. (1981). Robust Statistics. New York: Wiley.

Johnson, R. W. (1996). Fitting Percentage of Body Fat to Simple Body Measurements. *Journal of Statistics Education* **4**, n.1.

Lee, A.-H. and Yang, Y. (2006). Bagging Binary and Quantile Predictors for Time Series. Journal of Econometrics, V.135, 465-497.

Juditsky, A. and Nemirovski, A. (2000). Functional Aggregation for Nonparametric Estimation. The Annals of Statistics, V.28, 681-712.

Lv, J. and Fan, Y. (2009). A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares. *Ann. Statist.* **37**, 3498-3528.

Makridakis, S., Wheelwright, S., and Hyndman, R. J. (1998). "Forecasting: Methods and Applications" (3rd edition). New York: Wiley.

Min, C.-K. and Zellner, A. (1993). Bayesian and non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates. Journal of Econometrics, V.56, 89-118.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436-1462.

Newbold, P. and Granger, C. W. J. (1974). "Experience with Forecasting Univariate Time Series and the Combination of Forecasts". JRSSA, V.137, 131-165.

Palm, F. C. and Zellner, A. (1992) To Combine or Not to Combine? Issues of Combining Forecasts. Journal of Forecasting, V.11, 687-701.

Pesaran, M. H. and Timmermann, A. (2007) Selection of Estimation Window in the Presense of Breaks. Journal of Econometrics, V.137, 134-161.

Rapach, D. E. and Weber, C. E. (2004). Financial Variables and the Simulated Out-of-Sample Forecastability of U.S. Output Growth Since 1985: An Encompassing Approach. Economic Inquiry, V.42, 717-738.

Rapach, D. E. and Strauss, J. K. (2005). Forecasting Employment Growth in Missouri with Many Potentially Relevant Predictors: An Analysis of Forecast Combining Methods. Federal Reserve Bank of St. Louis Regional Economic Development, V.1, 97-112.

Rapach, D. E. and Strauss, J. K. (2008). Forecasting US Employment Growth Using Forecasting Combining Methods. Journal of Forecasting, V.27, 75-93.

Sancetta, A. (2007). Online Forecast Combinations of Distributions: Worst Case Bounds. Journal of Econometrics, V.141, 621-651.

Johnson, R. W. (1996). Fitting Percentage of Body Fat to Simple Body Measurements. *Journal of Statistics Education* **4**, n.1.

Sancetta, A. (2010). Recursive Forecast Combination for Dependent Heterogeneous Data. Econometric Theory, V.26, 598-631.

Sánchez, I. (2008). Adaptive Combination of Forecasts with Application to Wind Energy. International Journal of Forecasting, V.24, 679-693.

Schwarz, G. (1978). "Estimating the Dimension of a Model". The Annals of Statistics, V.6, 461-464.

Stock, J. H. and Watson, M. W. (2003). Forecasting Output and Inflation: The Role of Asset Prices. Journal of Economic Literature, V.41, 788-829.

Shan, K. and Yang, Y. (2009). Combining Regression Quantile Estimators. *Statistica Sinica* **19**, 1171-1191.

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

Swanson, N. R. and Zeng, T. (2001). "Choosing among Competing Econometric Forecasts: Regression-based Forecast Combination Using Model Selection". Journal of Forecasting, V.20, 425-440.

Tibshirani, R. J. (1996). Regression Shrinkage and Selection via the Lasso. *JRSSB* **58**, 267-288.

Timmermann, A. (2006). Forecast Combinations. Handbook of Economic Forecasting (Edited by G. Elliott, C. W. J. Granger and A. Timmermann). Amsterdam: Elsevier.

Tsybakov, A. B. (2003). Optimal rates of aggregation. Learning Theory and Kernel Machines, Lecture Notes in Artificial Intelligence 2777 303–313. Heidelberg: Springer.

Vovk, V. (1998) A Game of Prediction with Expert Advice. Journal of Computer and System Sciences, V.56, 153-173.

Wang, H., Li, R., and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

Wei, X. and Yang, Y. (2008). "Robust Forecast Combinations". Submitted.

Weisberg, S. (1985). Applied Linear Regression. New York: Wiley.

Wright, J. H. (2003). "Forecasting U.S. Inflation by Bayesian Model Averaging". International Finance Discussion Papers, No.780, Board of Governors of the Federal Reserve System.

Yang, Y. (2001). Adaptive Regression by Mixing. JASA, V.96, 574-588.

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

Yang, Y. (2004b). Aggregating Regression Procedures to Improve Performance. Bernoulli, V.10, 25-47.

Yang, Y. (2007). Consistency of Cross Validation for Comparing Regression Procedures. *Ann. Statist.* **35**, 2450-2473.

Yuan, Z. and Yang, Y. (2005). Combining Linear Regression Models: When and How? *Journal of the American Statistical Association* **100**, 1202-1214.

Zellner, A. (1986). Bayesian Estimation and Prediction Using Asymmetric Loss Functions. JASA, V.81, 446-451.

Zhang, C. (2007). Penalized Linear Unbiased Selection. Technical Report, Dept. Statistics, Rutgers Univ.

Zhang, C. and Huang, J. (2008). The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression. *Ann. Statist.* **36**, 1567-1594.

Zhao, P. and Yu, B. (2006) On Model Selection Consistency of Lasso. *Journal of Machine Learning Research* **7**, 2541-2563.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* **101**, 1418-1429.

Zou, H. (2008). On the "Degrees of Freedom" of the LASSO. *The Annals of Statistics* **35**, 2173-2192.

Zou, H. and Li, R. (2008). One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models. *Ann. Statist.* **36**, 1509-1533.

Zou, H. and Yang, Y. (2004). Combining Time Series Models for Forecasting. International Journal of Forecasting, V.20, 69-84.