# Small Sample Improvement Over Bayes Prediction Under Model Uncertainty

by

Hubert Wong

B.A.Sc, UBC 1992

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

**Doctor of Philosophy**

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

we accept this thesis as conforming
to the required standard

**The University of British Columbia**

August 2000

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of STATISTICS

The University of British Columbia
Vancouver, Canada

Date 08/08/00

# Abstract

Existing criteria for evaluating the adequacy of a predictive model are model-based (e.g. AIC, BIC, MSPE) or empirical (e.g. PRESS and other cross-validation type criteria). We introduce a new class of "mongrel" criteria for on-line prediction that evaluates candidate predictors based on both model information and past empirical performance. Simulation results showed that the mongrel procedure produced more accurate predictions than the standard Bayes procedure for small sample sizes. This improvement was observed over a wide range of data-generators for the problem of variable selection in normal linear models.

# Contents

# List of Tables

# List of Figures

# Acknowledgements

I would like to thank Professors Jim Zidek, John Petkau, and Paul Gustafson for their invaluable advice and support all of these past years, even when it was not clear what this thesis was about. Special thanks go to my supervisor, Professor Bertrand Clarke, for being even crazier than me.

Partial funding for this project was provided by NSERC Operating Grant OGP-0138122.

HUBERT WONG

*The University of British Columbia*
*August 2000*

# Chapter 1

# Introduction

This dissertation describes a new approach to on-line prediction in the presence of model uncertainty. Here, a prediction, or forecast – we use the two terms interchangeably – is a statement about the outcome of an as yet unobserved random variable. The term "on-line" indicates that we are predicting a sequence $\mathbf{Y} = (Y_1, Y_2, \ldots)$ wherein the prediction for the outcome $Y_{n+1}$, is made at time-point $n$ using only the information that is available at that time. The term "model uncertainty" means that we have a collection of models (i.e. parametric families) $\mathcal{M} = \{M_{\theta_k} : \theta_k \in \Theta_k, k \in \mathcal{K}\}$ as candidates for the distribution of $\mathbf{Y}$. Each $M_{\theta_k}$ generates it's own candidate predictor $\hat{Y}_{k,n+1}$ for $Y_{n+1}$. However, because of model uncertainty, we require a criterion for evaluating the worth of each model in order to obtain the prediction that will be used, either by choosing one model from or averaging over the models in $\mathcal{M}$. In existing literature, the criteria for evaluating the candidate models can be classified into one of two main types. We label them "model-based" and "empirical".

The value of a model-based criterion depends the structure of an as-

sumed probability model. This dependence typically manifests as a measure of fit of the data to this model. For instance, a likelihood or an expected risk of the predictor computed conditionally on this model and the data would be model-based. In general, two different models will generate a different values for the criterion even if both models had generated the same sequence of predictions in the past. Well-known examples of model-based criteria used for model choice include the Akaike Information Criterion (AIC) and variants on it such as the Bayes Information Criterion (BIC). These criteria are computed as minus two times the maximized log-likelihood plus a penalty term that depends on the number of fitted parameters in the model and sample size. For model averaging, the most active research area currently is in Bayesian model averaging wherein the adequacy of a model is judged on it's posterior probability. Recent references include Raftery et al (1997), Clyde (1999), Hoeting et al (1999). One criticism of these criteria is that they essentially measure the fit of the data to the model rather than evaluate the expected accuracy of the current prediction. To overcome this limitation, we can use a (decision-theoretic) risk criterion instead. Here, the risk $\rho_i(\hat{Y}_{k,n+1})$ of candidate predictor $\hat{Y}_{k,n+1}$, assuming model $i$ is true and *conditional on the data*, is evaluated for each $i$ in turn. The value of the criterion is taken to be the average risk $\sum_i \alpha_i \rho_i(\hat{Y}_{k,n+1})$ where $\alpha_i$ is the weight assigned to model $i$. However, whichever of these model-based criteria is used, all of them can be criticized on the basis that they do not consider the past predictive performance of the candidate models.

In contrast, an empirical criterion assesses the worth of a candidate model strictly on its *observed* predictive performance. That is, the worth of the predictor $\hat{Y}_{k,j}$ from model $k$ is given by a loss function $L(\hat{Y}_{k,j}, Y_j)$ that depends only on the observed values of $\hat{Y}_{k,j}$ and $Y_j$. All other information is ignored. Thus if two models had generated the same set of predictions,

the value of the empirical criterion would be the same for both models regarless of their underlying structure. The paradigmatic empirical approach in non-sequential settings is "leave-one-out" cross-validation. Here, individual observations $Y_j$ are omitted one at a time and $\hat{Y}_{k,j}$ is obtained from fitting the model using the remaining data. The criterion for assessing model $k$ is the total $T_k = \sum_j L(\hat{Y}_{k,j}, Y_j)$. A smaller value of $T_k$ indicates a better model. For squared error loss this criterion is the well-known predicted residual sum of squares, or PRESS, statistic (Allen, 1974). In a sequential setting, the prediction at a given time-point must be issued without knowledge of later data and hence the PRESS criterion is artificial. Dawid (1984) suggested that an appropriate modification is to base the criterion on the sequence of one-step ahead prediction losses $L(\hat{Y}_{k,n+1}, Y_{n+1})$ that already have been incurred where the forecast $\hat{Y}_{k,n+1}$ is based on only data known at time-point $n$. Subsequent work by Dawid and others (e.g., Dawid 1992, Sellier-Moiseiwitsch and Dawid 1993, Skouras 1996) developed the asymptotic theory underlying this "prequential approach" to forecasting.

The empirical approach has several attractive features. Consider, for example, two sequences of on-line forecasts for the next-day maximum temperature where one of the sequences was generated by a meteorologist's statistical model and the other sequence was generated by an old man based on "the feel in his bones". The performance of these two sequences would be non-comparable using a model-based approach since no model is available for the old man's sequence. But this comparison is easy to make using an empirical approach; simply define the criterion as $\sum_j L(\hat{Y}_{k,j}, Y_j)$, say. While the model-free nature of an empirical criterion allows for comparison of arbitrary forecasting procedures, we do not view this property as an advantage necessarily. Indeed, we will indicate shortly why it is a disadvantage in small samples.

Instead, we feel that the most attractive feature of an empirical criterion is that it provides an evaluation based directly on past predictive performance and we seek to preserve this feature in our new approach.

Initially, our work was motivated by the possibility that judicious use of the observed one-step ahead losses might allow for better predictions in small-samples while retaining the desirable asymptotic properties. The intuition is that losses incurred early in the sequence are likely to have less bearing on the quality of a candidate predictor than losses later in the sequence. Early on, the models are poorly estimated and therefore yield less precise information about performance. Dawid (1992) recognized this possibility and omitted the losses from early time-points in his simulations when computing the total loss. We, however, were interested specifically in such early losses since our focus was on small-sample performance.

Through simulations, we investigated the influence of early losses in *ad hoc* fashion by downweighting the earlier losses when computing the total loss, that is, rather than using $\sum_{i=1}^{n} L(\hat{Y}_{k,i}, Y_i)$ as the criterion when predicting for time-point $n+1$, we used $\sum_{i=1}^{n} w_i L(\hat{Y}_{k,i}, Y_i)$ where the weights satisfied $w_i \leq w_{i+1}$. Some weighting choices we considered were: (1) $w_i = i$ and (2) $w_i = 0$ if $i < n/2$, $w_i = 1$ if $n/2 \leq i \leq n$. Many of our choices yielded statistically significant improvements over using the simple total loss. However the magnitude of the improvement seemed small. Typically the reduction in the squared error prediction loss was around 0.5%.

We felt that there were two major limitations to this initial approach. First, by taking a purely empirical view, we were left without a probabilistic framework. This meant we could not quantify the relative importance of residuals at different time-points and thus could not determine optimal weighting strategies. Second, the initial specification of candidate models and prior be-

liefs about their plausibility represented information that may not have been used fully by looking only at the incurred losses. This information would be particularly valuable early in the sequence when little data has accumulated. Hence, we concluded that a purely empirical approach was not suitable for evaluating candidate models in small samples.

In Chapter 2, we describe a novel approach that combines aspects of both the model-based and the empirical approaches: we assume a probability framework for computing an expected risk but the expectation is computed conditional on a statistic $\mathbf{S}_n$ that reflects *the observed predictive performance* of the candidate models rather than conditional on the data values. To emphasize the dual aspects of the approach, we give the label "mongrel risk" to the resulting criteria. Examples of $\mathbf{S}_n$ that reflect empirical performance include past losses $L(\hat{Y}_{k,i}, Y_i)$, or past "predictuals" $Y_i - \hat{Y}_{k,i}$ (the residuals that would arise from using the predictions from model $k$), from the candidate models. Different choices for $\mathbf{S}_n$ generate different members in the class of mongrel risk criteria. The task is to determine good strategies for selecting $\mathbf{S}_n$ at any given time-point. A simple strategy is to set $\mathbf{S}_n$ equal to the last, say, $n/2$ losses or predictuals always. But such a rule may be too naive as it ignores both information intrinsic to the model structure and supplied by the data. More sophisticated strategies, which we label as "global", would incorporate model and covariate information but does not use the outcomes of the response $\mathbf{Y}_{(n)}$ in selecting $\mathbf{S}_n$. If a strategy also uses the outcomes of $\mathbf{Y}_{(n)}$ to select $\mathbf{S}_n$, we call it "adaptive". In adaptive selection, we define a "meta-risk" for assessing the adequacy of each candidate $\mathbf{S}_n$. We describe two examples of such meta-risks: a minimax meta-risk and a weighted average meta-risk.

In Chapter 3, we apply the mongrel risk criterion to prediction based on normal linear models. Explicit formulae for computing the mongrel risk

are given for the class $\mathbf{S}_n = \mathbf{U}^T\mathbf{Y}_{(n)} + \mathbf{c}$ where $\mathbf{U}$ and $\mathbf{c}$ do not depend on $\mathbf{Y}_{(n)}$; that is, $\mathbf{S}_n$ is affine in the response vector. This class of $\mathbf{S}_n$ includes past predictuals. We illustrate the implementation of the approach using a simulation study which uses naive choices of $\mathbf{S}_n$. We observed that setting $\mathbf{S}_n$ equal to a few recent predictuals resulted in more accurate predictions than the Bayes procedure (obtained by setting $\mathbf{S}_n = \mathbf{Y}_{(n)}$) in small samples for many but not all data-generating models.

We investigate the adaptive selection of $\mathbf{S}_n$ for small-samples in Chapter 4. Specifically, we seek the optimal number of predictuals to include in $\mathbf{S}_n$. We argue that adaptive selection maximizes the improvement in predictive accuracy. We implemented both the minimax and the weighted average versions of the meta-risk based on the simulated data used in Chapter 3. Our simulation results showed that in a model averaging context, the minimax meta-risk criterion produced more accurate predictions than the Bayes procedure uniformly over all of the scenarios tested. Moreover, the magnitude of the improvement was substantially greater than that seen in Chapter 3 where global choices of $\mathbf{S}_n$ were used.

In Chapter 5, we use robustness considerations to suggest global strategies. Unfortunately, these strategies are difficult to implement since the computations require integration over the distribution of $\mathbf{Y}_{(n+1)}$ that are not tractable typically. We implemented one strategy for selecting $\mathbf{S}_n$ in our simulation study by approximating the needed quantities. Our results indicated that the performance of this strategy beat out the Bayes strategy in many scenarios but also lost badly in a few cases.

The asymptotic theory for $\mathbf{S}_n$ is developed in Chapter 6. We characterize the sub-class of affine $\mathbf{S}_n$ that will yield asymptotic consistency of model weights. This consistency condition ensures that if one of the candidate mod-

els is true then as $n \to \infty$, the predictor from the true model will always be chosen (in a model choice approach) or that the weight assigned to the true model tends to 1 (in a model averaging approach).

In Chapter 7, we discuss additional simulation results that support the use of the mongrel procedure. In addition, we indicate areas of current and future work.

# Chapter 2

# Mongrel Risk

Let $\mathbf{Y} = (Y_1, Y_2, \ldots)$ be the sequence of random variables that is to be predicted. At each time point $n$, we must issue a prediction concerning the value of $Y_{n+1}$.

To aid us in constructing the prediction for time point $n+1$, typically the following information is available:

1. a $p$-vector of covariates $\mathbf{X}_{n+1}$ whose elements may be related to $Y_{n+1}$,
2. the outcomes and covariates already observed up to time point $n$, i.e., $\mathbf{Y}_{(n)} = (Y_1, \ldots, Y_n)$ and $\mathbf{X}_{(n)}$, the $n \times p$ matrix with row $i$ equal to $\mathbf{X}_i$,
3. prior information about the structure, which we call the model, that describes the probabilistic dependence of the outcomes on the covariates and the set of unknown parameters $\theta$ that indexes the model, and
4. prior distributions on the values of the unknown $\theta$.

When the model posited in 3. is uncertain, we often entertain a collection of candidate models $\mathcal{M} = \{M_{\theta_k} : \theta_k \in \Theta_k, k \in \mathcal{K}\}$. Each candidate model

$k$ is assigned a prior probability $\alpha_{k,o}$ that reflects the plausibility that it is the true model. We say that model $k$ is true if it contains the true distribution of $\mathbf{Y}$ and no sub-model (still in $\mathcal{M}$) of model $k$ contains this distribution. For each model $k$, we can proceed in a variety of ways to obtain a function of the observed data which we will use as the point prediction of $Y_{n+1}$. If the loss function $L(a, y_{n+1})$ describes the loss incurred by predicting using the value $a = a(\mathbf{Y}_{(n)})$ when $Y_{n+1} = y_{n+1}$ obtains, the Bayes predictor $\hat{Y}_{k,n+1}$, conditional on model $k$ being true, is the value of $a$ minimizing the posterior risk, i.e.,

$$\hat{Y}_{k,n+1} = \arg\min_a \mathbf{E}_{k|\mathbf{Y}_{(n)}} L(a, Y_{n+1}) \tag{2.1}$$

Here, $\mathbf{E}_{k|\mathbf{S}}$ indicates the conditional expectation given a statistic $\mathbf{S}$ assuming model $k$, marginalized with respect to the unknown parameters, is true. That is, if we let $\mathbf{T}$ be a minimal extension of $\mathbf{S}$ to $\mathbf{Y}_{(n+1)}$, then for any function $g(\mathbf{Y}_{(n+1)})$,

$$\mathbf{E}_{k|\mathbf{S}} g(\mathbf{Y}_{(n+1)}) \equiv \int g(\mathbf{y}_{(n+1)}) p(\mathbf{y}_{(n+1)}|\theta_k, \mathbf{S}) p(\theta_k|\mathbf{S}) d\theta_k d\mathbf{t} \tag{2.2}$$

where $p(\cdot)$ denotes the appropriate conditional density and $d\mathbf{t}$ represents integration over the sigma-field generated by $\mathbf{T}$. Analogously, the notations $\mathbf{C}_{k|\mathbf{S}}$ and $\mathbf{V}_{k|\mathbf{S}}$ will indicate the covariance and variance operators respectively.

Each candidate model generates a forecast but we must give a single forecast that will be used. This problem is usually solved either by choosing one of the candidate forecasts or by using the forecast generated from a mixture over the candidate models. We use the term "model choice" to describe the first approach and the term "model averaging" for the second.

To implement a model choice strategy, we require a criterion to assess the risk of the forecast derived from each candidate model. If model $k$ were

true, the posterior risk of using $\hat{Y}_{k,n+1}$ is

$$\rho_k(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) = \mathbf{E}_{k|\mathbf{Y}_{(n)}} L(\hat{Y}_{k,n+1}, Y_{n+1}). \tag{2.3}$$

But because the true model is uncertain, we should also consider the posterior risk of using $\hat{Y}_{k,n+1}$ when a different candidate model $i \neq k$ is true. That is, we also consider

$$\rho_i(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) = \mathbf{E}_{i|\mathbf{Y}_{(n)}} L(\hat{Y}_{k,n+1}, Y_{n+1}). \tag{2.4}$$

In the Bayes decision approach, the overall assessment of risk for each candidate forecast is given by the weighted average over the collection of posterior risks that would be incurred by this forecast under different true models. That is, the adequacy of the predictor $\hat{Y}_{k,n+1}$ would be the average posterior risk

$$\bar{\rho}(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) = \sum_{i \in \mathcal{K}} \alpha_i(\mathbf{Y}_{(n)}) \rho_i(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) \tag{2.5}$$

where the model weights are

$$\alpha_i(\mathbf{Y}_{(n)}) \equiv P(M_i|\mathbf{Y}_{(n)}), \tag{2.6}$$

i.e., $\alpha_i(\mathbf{Y}_{(n)})$ is the posterior probability that model $i$ is true.

Based on the criterion $\bar{\rho}(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)})$, the best choice for the predictor of $Y_{n+1}$ is $\hat{Y}_{k^*,n+1}$ where $k^*$ satisfies

$$k^* = \arg\min_{k \in \mathcal{K}} \bar{\rho}(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}). \tag{2.7}$$

For the special case of squared error loss, $k^*$ is the index of the model with the highest posterior probability, i.e., (2.7) reduces to

$$k^* = \arg\max_{k \in \mathcal{K}} \alpha_k(\mathbf{Y}_{(n)}). \tag{2.8}$$

In the case of a model averaging strategy, the Bayes decision approach is to construct the predictive mixture distribution for $Y_{n+1}$ given by

$$F_{m|\mathbf{Y}_{(n)}} = \sum_{i \in \mathcal{K}} \alpha_i(\mathbf{Y}_{(n)}) F_{i|\mathbf{Y}_{(n)}} \tag{2.9}$$

where $F_{i|\mathbf{Y}_{(n)}}$ is the posterior distribution of $Y_{n+1}$ given $\mathbf{Y}_{(n)}$ assuming model $i$ is true and the weight $\alpha_i(\mathbf{Y}_{(n)})$ is determined as in (2.6), the same as for choice strategies. (Here, the subscript $m$ refers to the mixture rather than a candidate model.) The risk of using $a = a(\mathbf{Y}_{(n)})$ to forecast $Y_{n+1}$ under the mixture is

$$\begin{aligned} \bar{\rho}(a; \mathbf{Y}_{(n)}) &= \mathbf{E}_{m|\mathbf{Y}_{(n)}} L(a, Y_{n+1}) && (2.10) \\ &= \sum_{i \in \mathcal{K}} \alpha_i(\mathbf{Y}_{(n)}) \rho_i(a, \mathbf{Y}_{(n)}) && (2.11) \end{aligned}$$

and the optimal forecast is now

$$\hat{Y}_{m,n+1} = \arg\min_a \bar{\rho}(a; \mathbf{Y}_{(n)}). \tag{2.12}$$

For the special case of squared error loss, (2.12) reduces to

$$\hat{Y}_{m,n+1} = \sum_{k \in \mathcal{K}} \alpha_k(\mathbf{Y}_{(n)}) \hat{Y}_{k,n+1}. \tag{2.13}$$

We will refer to the solution defined by (2.1) through (2.13) as the Bayes procedure (for model choice or model averaging, as appropriate).

The problem with the Bayes procedure is that none of the assessments of risk it uses reflect the true risk. First, the evaluation of $\rho_i(\cdot)$ assumes a distribution for $\mathbf{Y}_{(n+1)}$ obtained by averaging over the distribution of the parameters whereas the true parameter values are fixed numbers which almost certainly do not equal those implied in the averaging. The Bayesian's position is that this averaging represents the best that one can do according to the rules

of probability. However this does not make the assessed risk true. Moreover, since only one of the models can in fact be true, at least some of the risks $\rho_i(\cdot)$ are based on an incorrect model. These observations raise two issues: (i) Are these risk assessments valid? and (ii) Can the criteria for assessing risk be modified to produce better predictors? As to (i), it is clear that we cannot avoid making some "incorrect" assessments of risk since the true distribution is unknown. So pending a better resolution we must be satisfied with the process of weighting the risks or the models according to beliefs about the merit of each model as a reasonable means to handle the model uncertainty. This procedure is no different than what we use to handle parameter uncertainty in the strictly parametric case. However, we answer (ii) by showing that better predictors can be obtained by changing the way we calculate $\alpha_i(\cdot)$ and $\rho_i(\cdot)$.

To put our proposed approach in context, let $\mathbf{S}_n = \mathbf{S}_n(\mathbf{Y}_{(n)})$ be any statistic and consider generalizing the Bayes decision procedure by replacing occurences of $\mathbf{Y}_{(n)}$ by $\mathbf{S}_n$ in (2.4) through (2.13). The choice for $\mathbf{S}_n$ need not be the same for each instance. We let $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$ denote the choices used when evaluating $\alpha_i(\cdot)$ and $\rho_i(\cdot)$, respectively. (More generally, we are free to choose a different $\mathbf{S}_n$ for every instance of $\mathbf{Y}_{(n)}$. For example, the choice of $\mathbf{S}_n^\alpha$ in (2.6) could be different for $\alpha_i(\mathbf{S}_n^\alpha)$ than for $\alpha_{i'}(\mathbf{S}_n^\alpha)$.) Our thesis is that better predictors can be obtained by choosing $\mathbf{S}_n^\alpha$ and/or $\mathbf{S}_n^\rho$ to be different from $\mathbf{Y}_{(n)}$.

In this more general framework, the formula for the weights becomes

$$\alpha_i(\mathbf{S}_n^\alpha) \equiv P(M_i|\mathbf{S}_n^\alpha). \tag{2.14}$$

For a choice strategy, equations (2.4), (2.5), and (2.7) become

$$\rho_i(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho) = \mathbf{E}_{i|\mathbf{S}_n^\rho} L(\hat{Y}_{k,n+1}, Y_{n+1}), \tag{2.15}$$

$$\bar{\rho}(\hat{Y}_{k,n+1}; \mathbf{S}_n^\alpha, \mathbf{S}_n^\rho) = \sum_{i \in \mathcal{K}} \alpha_i(\mathbf{S}_n^\alpha) \rho_i(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho), \tag{2.16}$$

$$k^* = \arg\min_{k\in\mathcal{K}} \bar{\rho}(\hat{Y}_{k,n+1}; \mathbf{S}_n^\alpha, \mathbf{S}_n^\rho). \tag{2.17}$$

If the loss is squared error and $\mathbf{S}_n^\rho = \mathbf{Y}_{(n)}$, then it can easily be shown that (2.17) reduces to

$$k^* = \arg\max_{i\in\mathcal{K}} \alpha_i(\mathbf{S}_n^\alpha). \tag{2.18}$$

For averaging strategies, (2.9) through (2.12) become

$$F_{m|\mathbf{S}_n^\alpha,\mathbf{S}_n^\rho} = \sum_{i\in\mathcal{K}} \alpha_i(\mathbf{S}_n^\alpha) F_{i|\mathbf{S}_n^\rho}, \tag{2.19}$$

$$\bar{\rho}(a; \mathbf{S}_n^\alpha, \mathbf{S}_n^\rho) = \mathbf{E}_{m|\mathbf{S}_n^\alpha,\mathbf{S}_n^\rho} L(a, Y_{n+1}), \tag{2.20}$$

$$\hat{Y}_{m,n+1} = \arg\min_a \bar{\rho}(a; \mathbf{S}_n^\alpha, \mathbf{S}_n^\rho). \tag{2.21}$$

For squared error loss (2.13) becomes

$$\hat{Y}_{m,n+1} = \sum_{k\in\mathcal{K}} \alpha_k(\mathbf{S}_n^\alpha) \mathbf{E}_{k|\mathbf{S}_n^\rho} Y_{n+1}. \tag{2.22}$$

When $\mathbf{S}_n^\rho = \mathbf{Y}_{(n)}$, the expectations in (2.22) are posterior means and so (2.22) further reduces to

$$\hat{Y}_{m,n+1} = \sum_{i\in\mathcal{K}} \alpha_k(\mathbf{S}_n^\alpha) \hat{Y}_{k,n+1}. \tag{2.23}$$

The collection (2.14) through (2.23) defines a new class of risk criteria, indexed by the choice of a pair of $\sigma$-fields $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$, for obtaining a predictor $\hat{Y}_{k^*,n+1}$ (model choice) or $\hat{Y}_{m,n+1}$ (model averaging). We will use $\hat{Y}_{n+1} \equiv \hat{Y}_{n+1}(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ to refer to a predictor of either form ($\hat{Y}_{k^*,n+1}$ or $\hat{Y}_{m,n+1}$). We conjecture that suitable choices for $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$ are vectors of statistics that reflect how well each of the candidate models have performed in predicting earlier data points. Such choices of $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$ motivate the following definitions.

**Definition 2.1** *The* mongrel risk *of the predictor $\hat{Y}_{n+1}$ when model $i$ is true is*

$$\rho_i(\hat{Y}_{n+1}; \mathbf{S}_n^\rho) = \mathbf{E}_{i|\mathbf{S}_n^\rho} L(\hat{Y}_{n+1}, Y_{n+1}). \tag{2.24}$$

**Definition 2.2** *The* average mongrel risk *of the predictor* $\hat{Y}_{n+1}$ *is*

$$\bar{\rho}(\hat{Y}_{n+1}; \mathbf{S}_n^\alpha, \mathbf{S}_n^\rho) = \sum_{i \in \mathcal{K}} \alpha_i(\mathbf{S}_n^\alpha) \rho_i(\hat{Y}_{n+1}; \mathbf{S}_n^\rho). \tag{2.25}$$

The label "mongrel" is intended to reflect the hybridization of a model-based framework with empirical performance. The Bayes criterion is a special case in which we set $\mathbf{S}_n^\alpha = \mathbf{S}_n^\rho = \mathbf{Y}_{(n)}$.

The losses $L(\hat{Y}_{k,j}; Y_j)$ from previous time-points $j$ that would have been incurred had the predictor $\hat{Y}_{k,j}$ been used are natural candidates for statistics to be included in $\mathbf{S}_n^\alpha$ or $\mathbf{S}_n^\rho$. The following statistics are of particular interest and so we name them specifically.

**Definition 2.3** *The* predictual *resulting from using the predictor* $\hat{Y}_{k,j}$ *to predict* $Y_j$ *is*

$$R_{k,j} = Y_j - \hat{Y}_{k,j}. \tag{2.26}$$

(The index $k$ in the predictuals or the losses need not be the same as the index for the model that is under evaluation). By choosing $\mathbf{S}_n^\alpha$ and/or $\mathbf{S}_n^\rho$ of this form and conditioning on them, we obtain predictors that are functions of the actual performance of the candidate models rather than simply on data values.

It is not obvious which predictuals or losses should be included and one of our main goals is to find good choices for $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. We will focus on using predictuals rather than losses for two reasons. The pragmatic reason is that in the normal linear models with which we will be working, predictuals (more generally, any affine functions of $\mathbf{Y}_{(n)}$) allow us to evaluate the risks in (2.14) through (2.23) analytically. The conceptual reason is that losses, typically, do not distinguish between the bias and variance aspects in the error and this information may be relevant to assessing the quality of the candidates.

Since it is intuitively reasonable to expect that more recent predictuals contain more information (if only because recently fitted models are more stable), we prefer the inclusion of more recent predictuals in $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. Therefore, we will consider only collections of predictuals wherein the inclusion of a predictual from time-point $j$ implies that the predictuals from all time-points greater than $j$ are also included. Note that the special case of using the predictuals from all past time-points is equivalent to using the Bayes procedure since the $\sigma$-field generated by all predictuals and the $\sigma$-field generated by the data are equivalent. (The value of $Y_1$ can be recovered from the predictual at the first time-point. At any time-point $n > 1$, the value of $Y_n$ can be recovered given the $n$-th predictual and $\mathbf{Y}_{(n-1)}$. By induction, $\mathbf{Y}_{(n)}$ can be recovered from the set of all past predictuals.)

The formulae (2.14) through (2.23) serve as a means for obtaining the predictor for a given choice of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ but say nothing about what $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ to use. (This step is unnecessary in the Bayes approach since in that case $\mathbf{S}_n^\alpha = \mathbf{S}_n^\rho = \mathbf{Y}_{(n)}$ always.) One simple specification for $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ is to use always, say, only the $n/2$ (or some other pre-specified function of $n$) most recent predictuals when predicting for time-point $n$. This "naive" specification does not consider the structure of the underlying models or the observed data. In Chapter 3, we will see that this choice often yields better predictions than those obtained using the Bayes procedure but does not do so consistently. A more sophisticated selection procedure could take into account the model structure and the covariate values $\mathbf{X}_{(n)}$ but not the outcomes of $\mathbf{Y}_{(n)}$. This "global" approach, discussed in Chapter 5, is difficult to implement because the computations require an integration over the distribution of $\mathbf{Y}_{(n+1)}$ that is not tractable typically. However, we place less importance on the global approach because we believe that to obtain the greatest improvement, $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ must be

chosen *adaptively*, that is, rather than using the same $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ regardless of the outcomes of $\mathbf{Y}_{(n)}$, a different $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ is chosen for each sequence. The reasons for this belief will be developed more fully in Chapter 4. For now, we describe only the basic mathematical framework.

In order to compare different $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$'s, we need to assess the adequacy of each $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$.

**Definition 2.4** *The* meta-risk *of* $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ *is a number that assesses the adequacy of* $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$.

The optimal $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ minimizes this meta-risk.

There are two ways of interpreting meta-risk. The first way involves noting that any given $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ determines completely $\hat{Y}_{n+1}$. Thus, it seems reasonable that the meta-risk should be computed conditional on that choice of $\hat{Y}_{n+1}$. In this case, potential definitions for the meta-risk include the maximum risk of $\hat{Y}_{n+1}$ over different true models

$$\rho_\vee(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho; \mathbf{T}_n^\rho) = \max_i \rho_i(\hat{Y}_{n+1}; \mathbf{T}_n^\rho), \tag{2.27}$$

or the weighted average risk of $\hat{Y}_{n+1}$ over different true models

$$\rho_+(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho; \mathbf{T}_n^\alpha, \mathbf{T}_n^\rho) = \sum_i \alpha_i(\mathbf{T}_n^\alpha)\rho_i(\hat{Y}_{n+1}; \mathbf{T}_n^\rho) \tag{2.28}$$

where we have introduced the statistics $\mathbf{T}_n^\alpha$ and $\mathbf{T}_n^\rho$ to emphasize that these statistics, which are used to evaluate $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$, are distinct from the statistics $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$, which are used to evaluate the candidate predictors. Taking $\mathbf{T}_n^\alpha = \mathbf{S}_n^\alpha$ and $\mathbf{T}_n^\rho = \mathbf{S}_n^\rho$ would be natural because it would mean that we are using the same model weights and the same mongrel risks in both deriving and assessing $\hat{Y}_{n+1}$. However, we are not constrained to do so. Note that the dependence on $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ is implicit to the construction of $\hat{Y}_{n+1}$.

The second way of interpreting the meta-risk takes the view that the choice of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ precedes consideration of how the predictor is subsequently obtained for the selected $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. Thus, the meta-risk should reflect the risk of different candidate predictors that might arise (as opposed to does arise) from using $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. This type of assessment is straightforward in the model choice context. For example, we could define the meta-risk as, say, the maximum average mongrel risk over candidate predictors

$$\rho_\vee(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho) = \max_k \bar{\rho}(\hat{Y}_{k,n+1}; \mathbf{S}_n^\alpha, \mathbf{S}_n^\rho), \tag{2.29}$$

or the weighted average of the average mongrel risk over candidate predictors

$$\rho_\vee(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho; \mathbf{T}_n^\alpha) = \sum_k \alpha_i(\mathbf{T}_n^\alpha)\bar{\rho}(\hat{Y}_{k,n+1}; \mathbf{S}_n^\alpha, \mathbf{S}_n^\rho). \tag{2.30}$$

Again, it would be natural, but not necessary, to take $\mathbf{T}_n^\alpha = \mathbf{S}_n^\alpha$. Unfortunately, the analog to (2.29) or (2.30) in model averaging is not obvious. The class of "candidate predictors" in model averaging is the space of all functions of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. This class is problematic because of its large size and there does not appear to be a natural restriction.

We will use the first interpretation of meta-risk, in part because it is simpler to compute, but also because we find the second one odd in giving weight to the risk of a candidate predictor when it is known that that predictor will not be used.

# Chapter 3

# Application to Normal Linear Models

In this chapter, we derive the computational formulae needed to implement the criteria defined in (2.14) through (2.20) in the context of normal linear models. The candidate models are subset regression models. For expository simplicity, we will use the symbol $\mathbf{S}_n$ in statements that are applicable to both $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$. Except where noted, in this chapter and for the remainder of thesis, we assume that $\mathbf{S}_n$ is a $J = J(n)$ vector obtained as an affine transformation of $\mathbf{Y}_{(n)}$, i.e.,

$$\mathbf{S}_n = \mathbf{U}^T \left( \mathbf{Y}_{(n)} + \mathbf{c} \right), \tag{3.1}$$

where the $n \times J$ matrix $\mathbf{U}$ and the $n$-vector $\mathbf{c}$ do not depend on $\mathbf{Y}_{(n)}$. Without loss of generality, we can assume that $\mathbf{U}$ is of full rank; if it is not, we simply remove linearly dependent rows until it is full rank. Ultimately, we are interested in $\mathbf{S}_n$ only for the $\sigma$-field it generates. Choices for $\mathbf{S}_n$ that satisfy (3.1) include the special cases where $\mathbf{S}_n$ is a constant, $\mathbf{S}_n = \mathbf{Y}_{(n)}$, and $\mathbf{S}_n$ is a vector of past predictuals. The inclusion of past predictuals in this group

follows from the fact that Bayes predictors in linear models are affine in $\mathbf{Y}_{(n)}$.

The manner in which predictuals impact mongrel risk assessments can be rather subtle. So we will provide some basic results in order to develop an intuitive understanding. We conclude the chapter by presenting simulation results for some simple forms of $\mathbf{S}_n$. Our primary intent here is to illustrate the procedure; the identification of "good" choices for $\mathbf{S}_n$ is taken up in the subsequent chapters.

## 3.1 Notation

We consider a collection of subset regression models of the form

$$\mathbf{Y}_{(n)}|\mathbf{X}_{(n)}, \beta_k \sim \mathcal{N}\left(\mathbf{X}_{(n)}\mathbf{D}_k\beta_k, \sigma^2\mathbf{I}\right) \tag{3.2}$$

as candidate models where $\mathbf{D}_k$ is a $p \times p_k$ matrix that "picks out" the $p_k$ covariates associated with model $k$. For simplicity, assume that $\sigma^2$ is known. The parameter vector, $\beta_k$, is assumed to have a prior distribution $\pi_k$ given by

$$\beta_k \sim \mathcal{N}\left(\mathbf{b}_k, \Gamma_k\right). \tag{3.3}$$

For notational convenience, let

$$\mathbf{Z}_{k,(n)} = \mathbf{X}_{(n)}\mathbf{D}_k. \tag{3.4}$$

Then the marginal density for $\mathbf{Y}_{(n)}$ after mixing over the prior is $\mathcal{N}\left(\nu_{k,n}, \Psi_{k,n}\right)$ where

$$\nu_{k,n} = \mathbf{Z}_{k,(n)}\mathbf{b}_k \tag{3.5}$$

$$\Psi_{k,n} = \sigma^2\mathbf{I} + \mathbf{Z}_{k,(n)}\Gamma_k\mathbf{Z}_{k,(n)}^T. \tag{3.6}$$

For each model $k$ and given data $\mathbf{Y}_{(n)}$ at time $n$, the Bayes rule with respect to $\pi_k$ and under squared error loss is

- for estimating $\beta_k$:

$$\begin{aligned}\hat{\beta}_k &= \arg\min_{\mathbf{a}} \mathbf{E}_{k|\mathbf{Y}_{(n)}} (\beta_k - \mathbf{a})^2 \\ &= \mathbf{E}_{k|\mathbf{Y}_{(n)}} \beta_k \\ &= C_{k,n}\Psi_{k,n}^{-1}\mathbf{Y}_{(n)} + (\mathbf{b}_k - C_{k,n}\Psi_{k,n}^{-1}\mathbf{Z}_{k,(n)}\mathbf{b}_k) \qquad (3.7)\end{aligned}$$

where $C_{k,n} = \Gamma_k \mathbf{Z}_{k,(n)}^T$ is the covariance between $\beta_k$ and $\mathbf{Y}_{(n)}$.

- for predicting $Y_{n+1}$:

$$\begin{aligned}\hat{Y}_{k,n+1} &= \mathbf{Z}_{k,n+1}^T \hat{\beta}_k \\ &= \mathbf{Z}_{k,n+1}^T C_{k,n}\Psi_{k,n}^{-1}\mathbf{Y}_{(n)} + \mathbf{Z}_{k,n+1}^T \left(\mathbf{b}_k - C_{k,n}\Psi_{k,n}^{-1}\mathbf{Z}_{k,(n)}\mathbf{b}_k\right) \\ &= \mathbf{u}_{k,n+1}^T \mathbf{Y}_{(n)} + \left(\mathbf{Z}_{k,n+1} - \mathbf{u}_{k,n+1}^T \mathbf{Z}_{k,(n)}\right)\mathbf{b}_k \qquad (3.8)\end{aligned}$$

where

$$\mathbf{u}_{k,n+1}^T = \mathbf{Z}_{k,n+1}^T C_{k,n}\Psi_{k,n}^{-1} \qquad (3.9)$$

can be recognized as (marginalized with respect to $\beta_k$) the covariance of $Y_{n+1}$ and $\mathbf{Y}_{(n)}$ multiplied by the inverse of the variance of $\mathbf{Y}_{(n)}$.

The Bayes predictors $\hat{Y}_{k,n+1}$ arising from considering different models will constitute the collection of candidate forecasts.

The predictual arising from using $\hat{Y}_{k,n+1}$, the predictor from model $k$, to predict $Y_{n+1}$ is

$$\begin{aligned}R_{k,n+1} &= Y_{n+1} - \hat{Y}_{k,n+1} \\ &= Y_{n+1} - \mathbf{u}_{k,n+1}^T \mathbf{Y}_{(n)} - \left(\mathbf{Z}_{k,n+1} - \mathbf{u}_{k,n+1}^T \mathbf{Z}_{k,(n)}\right)\mathbf{b}_k \\ &= \mathbf{u}_{k,n+1}^{*T} \left(\mathbf{Y}_{(n+1)} - \mathbf{Z}_{k,(n+1)}\mathbf{b}_k\right) \qquad (3.10)\end{aligned}$$

where $\mathbf{u}_{k,n+1}^{*T} = (-\mathbf{u}_{k,n+1}^T, 1)$.

Given data up to time-point $n$, it can be useful to express the predictuals for all time points less than $n$ in terms of the full data set $(\mathbf{Y}_{(n)}, \mathbf{Z}_{k,(n)})$, i.e., we write the predictual from time-point $j$ as

$$R_{k,j} = \mathbf{u}_{k,j}^{oT}\left(\mathbf{Y}_{(n)} - \mathbf{Z}_{k,(n)}\mathbf{b}_k\right) \tag{3.11}$$

where $\mathbf{u}_{k,j}^{oT} = \left(\mathbf{u}_{k,j}^{*T}, 0, \ldots, 0\right)$. The difference between (3.10) and (3.11) is purely notational. However, when $\mathbf{S}_n$ is composed of predictuals only, the matrix $\mathbf{U}$ in (3.1) is easy to contruct if the predictuals are in the form (3.11); simply set each row in $\mathbf{U}$ to be $\mathbf{u}_{k,j}^{oT}$ with the desired choice of $k$ and $j$.

When model $i$ is true, $R_{k,n+1}$ is normally distributed with mean and variance given by, respectively,

$$\mathbf{E}_i R_{k,j} = \mathbf{u}_{k,j}^{oT}\left(\mathbf{Z}_{i,(n)}\mathbf{b}_i - \mathbf{Z}_{k,(n)}\mathbf{b}_k\right) \tag{3.12}$$

$$\mathbf{V}_i R_{k,j} = \mathbf{u}_{k,j}^{oT}\Psi_{i,n}\mathbf{u}_{k,j}^{o}. \tag{3.13}$$

## 3.2 Formulae for Choice and Weighting Strategies

To implement the criteria (2.14) through (2.20), we require explicit formulae for $\alpha_i(\mathbf{S}_n^\alpha)$, $\rho_i(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho)$ and $F_{i|\mathbf{S}_n^\rho}$. Since $\mathbf{S}_n$ is affine in $\mathbf{Y}_{(n)}$, these formulae can be obtained in a straightforward manner using the properties of the multivariate normal distribution. If model $i$ is assumed to be true, then $\mathbf{S}_n$ is distributed as a $\mathcal{N}(\mu_i, \Sigma_i)$ with mean and variance

$$\mu_i = \mathbf{U}^T\mathbf{Z}_{i,(n)}\mathbf{b}_i + \mathbf{U}^T\mathbf{c}, \tag{3.14}$$

$$\Sigma_i = \sigma^2\mathbf{U}^T\mathbf{U} + \mathbf{U}^T\mathbf{Z}_{i,(n)}\Gamma_i\mathbf{Z}_{i,(n)}^T\mathbf{U}. \tag{3.15}$$

Then the mongrel risk of the predictor $\hat{Y}_{k,n+1}$ is

$$\begin{aligned}\rho_i(\hat{Y}_{k,n+1};\mathbf{S}_n^\rho) &= \mathbf{E}_{i|\mathbf{S}_n^\rho} R_{k,n+1}^2 \\ &= \mathbf{V}_{i|\mathbf{S}_n^\rho} R_{k,n+1} + \mathbf{E}_{i|\mathbf{S}_n^\rho}^2 R_{k,n+1} \\ &= \mathbf{V}_i R_{k,n+1} - \Xi_{k,i}\Sigma_i^{-1}\Xi_{k,i}^T \\ &\quad + \left(\mathbf{E}_i R_{k,n+1} + \Xi_{k,i}\Sigma_i^{-1}\left(\mathbf{S}_n^\rho - \mu_i\right)\right)^2 \end{aligned} \tag{3.16}$$

where the covariance between the predictual $R_{k,n+1}$ and $\mathbf{S}_n^\rho$ is

$$\Xi_{k,i} \equiv \mathbf{C}_i(R_{k,n+1}, \mathbf{S}_n^\rho) = D_{k,i}^T \Psi_{i,n}\mathbf{U} \tag{3.17}$$

and

$$D_{k,i} \equiv \Psi_{i,n}^{-1}\mathbf{Z}_{i,(n)}\Gamma_i\mathbf{Z}_{i,n+1} - \Psi_{k,n}^{-1}\mathbf{Z}_{k,(n)}\Gamma_k\mathbf{Z}_{k,n+1}. \tag{3.18}$$

The predictive distribution under model $i$ conditional on $\mathbf{S}_n^\rho$ is given by,

$$F_{i|\mathbf{S}_n^\rho} \sim \mathcal{N}(\psi_i, \Upsilon_i) \tag{3.19}$$

where

$$\begin{aligned}\psi_i &= \mathbf{E}_{i|\mathbf{S}_n^\rho} Y_{n+1} \\ &= \mathbf{Z}_{i,n+1}^T\mathbf{b}_i + \mathbf{Z}_{i,n+1}^T\Gamma_i\mathbf{Z}_{i,(n)}^T\mathbf{U}\Sigma_i^{-1}\left(\mathbf{S}_n^\rho - \mu_i\right), \end{aligned} \tag{3.20}$$

$$\begin{aligned}\Upsilon_i &= \mathbf{V}_{i|\mathbf{S}_n^\rho} Y_{n+1} \\ &= \sigma^2 + \mathbf{Z}_{i,n+1}^T\Gamma_i\mathbf{Z}_{i,n+1} - \mathbf{Z}_{i,n+1}^T\Gamma_i\mathbf{Z}_{i,(n)}^T\mathbf{U}\Sigma_i^{-1}\mathbf{U}^T\mathbf{Z}_{i,(n)}\Gamma_i\mathbf{Z}_{i,n+1} \\ &= \sigma^2 + \mathbf{Z}_{i,n+1}^T\left(\Gamma_i^{-1} + \sigma^{-2}\mathbf{Z}_{i,(n)}^T\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Z}_{i,(n)}\right)^{-1}\mathbf{Z}_{i,n+1}. \end{aligned} \tag{3.21}$$

The model weights $\alpha_i(\mathbf{S}_n^\alpha)$ are obtained using Bayes theorem which yields

$$\alpha_i(\mathbf{S}_n^\alpha) = \frac{\alpha_{i,o}m_k(\mathbf{S}_n^\alpha)}{\sum_k \alpha_{k,o}m_k(\mathbf{S}_n^\alpha)}, \tag{3.22}$$

where $\alpha_{i,o}$ denotes the prior weight given to model $i$ and

$$m_i(\mathbf{S}_n^\alpha) = (2\pi)^{J/2}|\Sigma_i|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{S}_n^\alpha - \mu_i)^T\Sigma_i^{-1}(\mathbf{S}_n^\alpha - \mu_i)\right\} \tag{3.23}$$

is the marginal density of $\mathbf{S}_n^\alpha$ when model $i$ is true.

## 3.3 Some Intuition

It is worthwhile at this point to develop some intuition as to how past predictuals represent information. The key point is that predictuals are relevant only when model uncertainty is present. Indeed, in Corollary 3.1 below, we see that when a single model is taken to be true *a priori*, the assessed quality of the predictor from this model does not depend on any affine function of $\mathbf{Y}_{(n)}$. One important notion relevant to choosing $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$ is risk-sufficiency.

**Definition 3.1** *(a) A $\sigma$-field $\mathbf{S}_n^\rho$ is* risk-sufficient for $\alpha$ *if $\alpha_i(\mathbf{S}_n^\rho) = \alpha_i(\mathbf{Y}_{(n)})$ for every $i$. (b) A $\sigma$-field $\mathbf{S}_n^\rho$ is* risk-sufficient for $\rho$ *if $\rho_i(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho) = \rho_i(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)})$ for every pair $(k, i)$.*

In taking a mongrel risk approach, we need to *avoid* choosing a pair $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ that is risk-sufficient since then our procedure devolves to the Bayes procedure that we are trying to beat. The following Lemma can be useful for verifying risk-sufficiency of a given $\mathbf{S}_n^\rho$.

**Lemma 3.1** *$\rho_i(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho) = \rho_i(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) \iff$ at least one of the following holds:*

*(i)* $\Xi_{k,i} = \mathbf{0}$,

*(ii)* $\text{rank}(\mathbf{U}) = n$, *or*

*(iii)* $D_{k,i}$ *lies in the column space of* $\mathbf{U}$, *i.e., there exists a vector* $\mathbf{d}$ *such* $D_{k,i} = \mathbf{U}\mathbf{d}$.

**Proof** The sufficiency of condition (i) is obvious from inspection of (3.16). If condition (i) does not hold, we need $\Delta_\rho \equiv \rho_i(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho) - \rho_i(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) = 0$.

Let $\chi_{k,i} = \mathbf{C}_i(R_{k,n+1}, \mathbf{Y}_{(n)})$. Applying (3.16) and simplifying, we obtain

$$\begin{aligned}\Delta_\rho &= -\Xi_{k,i}\Sigma_i^{-1}\Xi_{k,i}^T + \left(\mathbf{E}_i R_{k,n+1} + \Xi_{k,i}\Sigma_i^{-1}(\mathbf{S}_n^\rho - \mu_i)\right)^2 \\ &\quad +\chi_{k,i}\Psi_{i,n}^{-1}\chi_{k,i}^T - \left(\mathbf{E}_i R_{k,n+1} + \chi_{k,i}\Psi_{i,n}^{-1}(\mathbf{Y}_{(n)} - \nu_i)\right)^2 \\ &= D_{k,i}^T\Psi_i(\mathbf{I}-\mathbf{P})D_{k,i} + c\,(\mathbf{Y}_{(n)} - \nu_i)^T(\mathbf{I}-\mathbf{P})D_{k,i} \qquad (3.24)\end{aligned}$$

where

$$\mathbf{P} = \mathbf{U}(\mathbf{U}^T\Psi_i\mathbf{U})^{-1}\mathbf{U}^T\Psi_i \qquad (3.25)$$

is a projection matrix onto the column space of $\mathbf{U}$ and $c = \mathbf{E}_i R_{k,n+1} + (\mathbf{Y}_{(n)} - \nu_i)^T(\mathbf{I}+\mathbf{P})D_{k,i}$. Clearly $\Delta_\rho = 0$ iff

$$(\mathbf{I}-\mathbf{P})D_{k,i} = \mathbf{0}. \qquad (3.26)$$

Since $\text{rank}(\mathbf{U}) = n$ implies $\mathbf{I}-\mathbf{P} = \mathbf{0}$, condition (ii) is sufficient. Otherwise, we need the null space of $\mathbf{I}-\mathbf{P}$ to contain $D_{k,i}$. But the null space of $\mathbf{I}-\mathbf{P}$ is equal to the column space of $\mathbf{U}$ and hence we need $D_{k,i}$ to lie in the column space of $\mathbf{U}$. This is condition (iii). ∎

As an example consider that, conditional on a given model $i$, the minimal sufficient statistic for the parameter $\beta_i$ is $\mathbf{Z}_{i,(n)}^T\Psi_{i,n}^{-1}\mathbf{Y}_{(n)}$. This suggests that a natural candidate for risk-sufficiency (for $\rho$) in a collection of $K$ models is $\mathbf{S}_n^\rho = \mathbf{U}^T\mathbf{Y}_{(n)}$ where $\mathbf{U} = \left(\Psi_{1,n}^{-1}\mathbf{Z}_{1,(n)} \mid \cdots \mid \Psi_{K,n}^{-1}\mathbf{Z}_{K,(n)}\right)$. Indeed, by taking components of $\mathbf{d}$ to be 0 or of the form $\Gamma_i\mathbf{Z}_{i,n+1}$, it is clear from (3.18) that this $\mathbf{U}$ can generate any $D_{k,k'}$ for any choice of $k$ and $k'$. This shows that $\mathbf{S}_n^\rho$ is risk-sufficient. Note that Lemma 3.1 also suggests that risk-sufficiency is a weaker notion of sufficiency than parametric sufficiency. If we have, say, only two candidate models, then we can obtain one-dimensional risk-sufficient statistic simply by setting $\mathbf{U} = D_{k,i}$. Moreover, this choice can be made irrespective of the number of parameters in the models. This result contrasts

with parametric sufficiency where the minimal sufficient statistic usually has dimension equal to the number of parameters. (This phenomena may also explain what's happening in Lemma 3.2 below where only a few predictuals is needed to get risk-sufficiency.) This observation leads us to conjecture that given $K$ candidate models, it should be possible to construct a $K-1$ dimensional risk-sufficient statistic.

**Corollary 3.1** *The risk* $\rho_k(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho)$ *is constant in* $\mathbf{S}_n^\rho$.

**Proof** Setting $i = k$ in (3.17), (3.18) gives $\Xi_{k,k} = \mathbf{0}$ and the result follows from condition (i) in Lemma 3.1. ∎

This result is not surprising really. Conditional on a fixed model, the Bayes predictor is the optimal predictor so the corresponding predictual must be uncorrelated with any affine function of $\mathbf{Y}_{(n)}$. (Otherwise one can construct a better predictor and thereby contradict the optimality of the Bayes predictor.)

The characterization of risk-sufficiency for $\alpha$ is more complicated and we are unable to provide a simple test for general choices of $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$. However, for our purposes, we are more concerned with the behaviour of $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$ that are comprised of predictuals. In our computational work, we have established that when model $i$ is true, the mongrel risk of the predictor from a different model, $k$ say, can depend nontrivially on the past predictuals generated by model $i$ or model $k$, that is,

**Fact 3.1** *Suppose model* $i$ *is true and let* $\mathbf{S}_n^\alpha$ *and* $\mathbf{S}_n^\rho$ *each consist of a set of past predictuals generated by model* $i$. *Then in general,*

$$\rho_i(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho) \neq \rho_i(\hat{Y}_{k,n+1}; 0). \tag{3.27}$$

Thus, our mongrel risk procedure indeed generates predictions that are different from the Bayes procedure. But our computational results also indicate that care must be exercised to ensure that we do not choose a set of risk-sufficient predictuals.

**Fact 3.2** *Suppose $k$ and $k'$ are a pair of nested models where model $k$ contains $\tilde{p}$ additional predictors. Let $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$ each consist of the $\tilde{p}$ most recent predictuals from each model. Then*

$$\rho_k(\hat{Y}_{k',n+1}; \mathbf{S}_n^\rho) = \rho_k(\hat{Y}_{k',n+1}; \mathbf{Y}_{(n)}), \tag{3.28}$$

$$\rho_{k'}(\hat{Y}_{k',n+1}; \mathbf{S}_n^\rho) = \rho_{k'}(\hat{Y}_{k',n+1}; \mathbf{Y}_{(n)}), \tag{3.29}$$

*and*

$$\alpha_i(\mathbf{S}_n^\alpha) = \alpha_i(\mathbf{Y}_{(n)}) \text{ for both } i = k, k' \tag{3.30}$$

*That is, this choice of $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$ is risk-sufficient.*

These results also suggest that predictive risk assessments using predictuals have a Markov property; which of two models is better is determined by only the most recent predictual(s) from each model. It can also be shown that including predictuals from only the larger model has no impact on any of the risk assessments. The combination of these two facts suggests that we include only the predictuals from the smaller model. This is what we do in all of our simulations.

## 3.4 A Simulation Example

Through simulation we assessed the forecasting performance for several choices of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. Data sequences of length forty were randomly generated according

to the model

$$Y_n = \gamma_0 + \gamma_1 X_{1,n} + \gamma_2 X_{2,n} + \epsilon_n \tag{3.31}$$

where $X_{1,n}, X_{2,n}, \epsilon_n$ were all independent standard normal variables. We fixed $\gamma_0 = 1$ and $\gamma_1 = 0.8$, but varied the value of $\gamma_2$ to be 0, 0.2, or 0.4 in different simulations.

We feel that these three choices for $\gamma_2$ cover a large enough range for testing our methodology. With unit variances for $X_{2,n}$ and $\epsilon_n$, the correlation between $Y$ and $X_2$, conditional on $X_1$, is $\gamma_2/\sqrt{1+\gamma_2^2}$. When $\gamma_2 = 0.4$, this correlation is $\approx 0.37$, a reasonably strong association. We assume that any stronger association typically would have been evident, or at least suspected, prior to the analysis. Consequently, there would have been no doubt that such a covariate should be included in all of the candidate models, i.e., such a covariate would not be subject to our predictor selection procedures.

The collection of candidate models contained two models:

- Model 1 (the "reduced model"): containing the intercept and $X_1$ only, i.e.,

$$Y_n = \beta_0^* + \beta_1^* X_{1,n} + \epsilon_n \tag{3.32}$$

- Model 2 (the "full" model"): containing the intercept and both $X_1$ and $X_2$, i.e.,

$$Y_n = \beta_0 + \beta_1 X_{1,n} + \beta_2 X_{2,n} + \epsilon_n. \tag{3.33}$$

For prior distributions on the parameters, we assumed that $(\beta_0^*, \beta_1^*) \sim \mathcal{N}((1, 0.8), \mathbf{I})$ and $(\beta_0, \beta_1, \beta_2) \sim \mathcal{N}((1, 0.8, 0.2), \mathbf{I})$. We considered three choices for the $\alpha_{2,o}$, the prior probability of Model 2: 0.2, 0.5, and 0.8. Hence a total of 9 scenarios (3 choices for $\gamma_2$ $\times$ 3 choices for $\alpha_{2,o}$) were considered. The number of sequences used in each scenario was $m = 5000$.

In all of the simulations presented here, we set $\mathbf{S}_n^\rho = \mathbf{Y}_{(n)}$. The choices for $\mathbf{S}_n^\alpha$ were

- $\mathbf{R}_1$ (the most recent predictual from Model 1)
- $\mathbf{R}_5$ (the most recent 5 predictuals from Model 1)
- $\mathbf{R}_h$ (the most recent half of (i.e., $n/2$, rounded down) predictuals from Model 1)
- $\mathbf{R}_n \equiv \mathbf{Y}_{(n)}$ (all past predictuals $\equiv$ full data).

### 3.4.1 Model Averaging

Figures 3.1 to 3.9 plot the results from taking a model averaging approach. The top panel in each page plots of the mean squared prediction error

$$\text{MSPE} = \frac{1}{m}\sum_{i=1}^{m}(Y_{n+1} - \hat{Y}_{n+1})^2 \tag{3.34}$$

incurred by using each specified choice of $\mathbf{S}_n^\alpha$. The second panel from the top shows the average weight that was assigned to the full model. The standard which we are trying to beat is the Bayes procedure in which $\mathbf{S}_n^\alpha = \mathbf{Y}_{(n)}$ (labeled 'a2ff' in the plots). The bottom two panels compare the difference in MSPE between the Bayes procedure and the choice $\mathbf{S}_n^\alpha = \mathbf{R}_h$ ('a2hf') or the choice $\mathbf{S}_n^\alpha = \mathbf{R}_5$ ('a25f'). On these two plots, a curve lying above zero indicates that our mongrel approach, 'a2hf' or 'a25f', is beating out the Bayes procedure.

To facilitate comparison across different plots, we classified the performance of 'a2hf' relative to 'a2ff' into the groups shown in Table 3.1. Table 3.2 summarizes the results of this classification in 3 time intervals: 10 to 20, 20 to 30, and 30 to 40.

For small $n$, the mongrel strategy $\mathbf{S}_n^{\alpha} = \mathbf{R}_h$ performed no worse and in most scenarios better than the Bayes strategy. The magnitude of the improvement decreased as the length of the sequence increased. When $n$ is large and $\gamma_2 = 0.4$, the mongrel strategy performed worse than the Bayes strategy.

At this point, one could question whether the better performance by the mongrel strategies is coincidental. That is, could a mongrel strategy be doing better because it gives higher weight to the "right" model than does the Bayes strategy just by chance? Figure 3.4 suggests that this is not the case. Here, the true model has $\gamma_2 = 0$. So a strategy that gives greater weight to the reduced model on average ought to perform better than one that gives less weight to the reduced model. We see in the second panel that the mongrel strategies ('a2hf' and 'a25f') both give less weight to the reduced model than does the Bayes strategy so, on average, the mongrel strategies are at a disadvantage with respect to giving high weight to the right model. Yet, from the third and fourth panels, we see that the mongrel strategies are beating the Bayes strategy. This suggests that the mongrel strategies are more intelligent.

The results for model averaging are qualitatively very similar to the results for model choice. Once again the mongrel strategy is no worse and sometimes better than the Bayes strategy for small $n$. Also, the mongrel strategy is worse when $n$ is large and $\gamma_2 = 0.4$. In general, the mongrel strategy beats the Bayes strategy in more scenarios and by a greater degree than was seen in the model choice approach.

The same counter-intuitive phenomenom seen in the choice approach occurs here. When $\gamma_2 = 0$, the mongrel procedures on average give less weight to the reduced model (the true model) than the Bayes strategy and yet tend to perform better early in the sequence.

### 3.4.2 Model Choice

Figures 3.10 to 3.18 plot the results from taking a model choice approach. The panels contain the same information as in the model choice results except that the second panel now shows the proportion of times that the full model was selected. Table 3.3 summarizes these plots.

The results for model choice are qualitatively very similar to the results for model averaging. Once again the mongrel strategy is never worse and sometimes better than the Bayes strategy for small $n$. Also, the mongrel strategy is worse when $n$ is large and $\gamma_2 = 0.4$. In general, the difference in performance between the mongrel strategies beats the Bayes strategy in the model choice approach was smaller than what was observed in model averaging.

The intelligent behaviour of mongrel strategies that was seen in the model averaging approach manifests here as well. When $\gamma_2 = 0$, the mongrel procedures on average choose the reduced model (the true model) less often than the Bayes strategy and yet tend to perform better (early in the sequence).

### 3.4.3 Summary

The simulation results presented here provide some evidence that taking a mongrel risk approach often beats out the Bayes procedure, particularly at early time points. The advantange gained here, using simple choices for the mongrel risk criteria, decreased as time progressed and, when $\gamma_2 = 0.4$ became a disadvantage. Moreover, the magnitude of the gains was relatively small. However, these choices were only intended to illustrate of the technique. In the next chapter, we show that larger and more lasting gains can be obtained by optimizing the choice of $(\mathbf{S}_n^{\alpha}, \mathbf{S}_n^{\rho})$.

Table 3.1: Key for comparing the mongrel procedures to the Bayes procedure.

| | MSPE(Bayes) - MSPE(mongrel) |
|---|---|
| $+++$ | $> 0.02$ |
| $++$ | between 0.01 and 0.02 |
| $+$ | between 0 and 0.01 |
| 0 | no clear difference |
| $-$ | between 0 and $-0.01$ |
| $--$ | between $-0.01$ and $-0.02$ |
| $---$ | $< -0.02$ |

Table 3.2: Summary comparison of the naive mongrel averaging strategy with $\mathbf{S}_n^\alpha = \mathbf{R}_h$ to the Bayes strategy ($\mathbf{S}_n^\alpha = \mathbf{Y}_{(n)}$).

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| **I** | 0.2 | 0 | 0 | 0 | 0 |
| | | 0.2 | $++$ | $+$ | 0 |
| | | 0.4 | 0 | $-$ | $--$ |
| | 0.5 | 0 | $++$ | $+$ | $+$ |
| | | 0.2 | $++$ | $+$ | $+$ |
| | | 0.4 | $++$ | 0 | $--$ |
| | 0.8 | 0 | $+$ | $+$ | $+$ |
| | | 0.2 | $++$ | $++$ | $++$ |
| | | 0.4 | $++$ | $+$ | 0 |

Table 3.3: Summary comparison of the naive mongrel choice strategy with $\mathbf{S}_n^\alpha = \mathbf{R}_h$ to the Bayes strategy ($\mathbf{S}_n^\alpha = \mathbf{Y}_{(n)}$).

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| **I** | 0.2 | 0 | 0 | 0 | 0 |
| | | 0.2 | ++ | + | + |
| | | 0.4 | 0 | − | − − |
| | 0.5 | 0 | ++ | + | + |
| | | 0.2 | ++ | 0 | 0 |
| | | 0.4 | 0 | − − | − − − |
| | 0.8 | 0 | 0 | 0 | 0 |
| | | 0.2 | 0 | 0 | + |
| | | 0.4 | 0 | 0 | − |

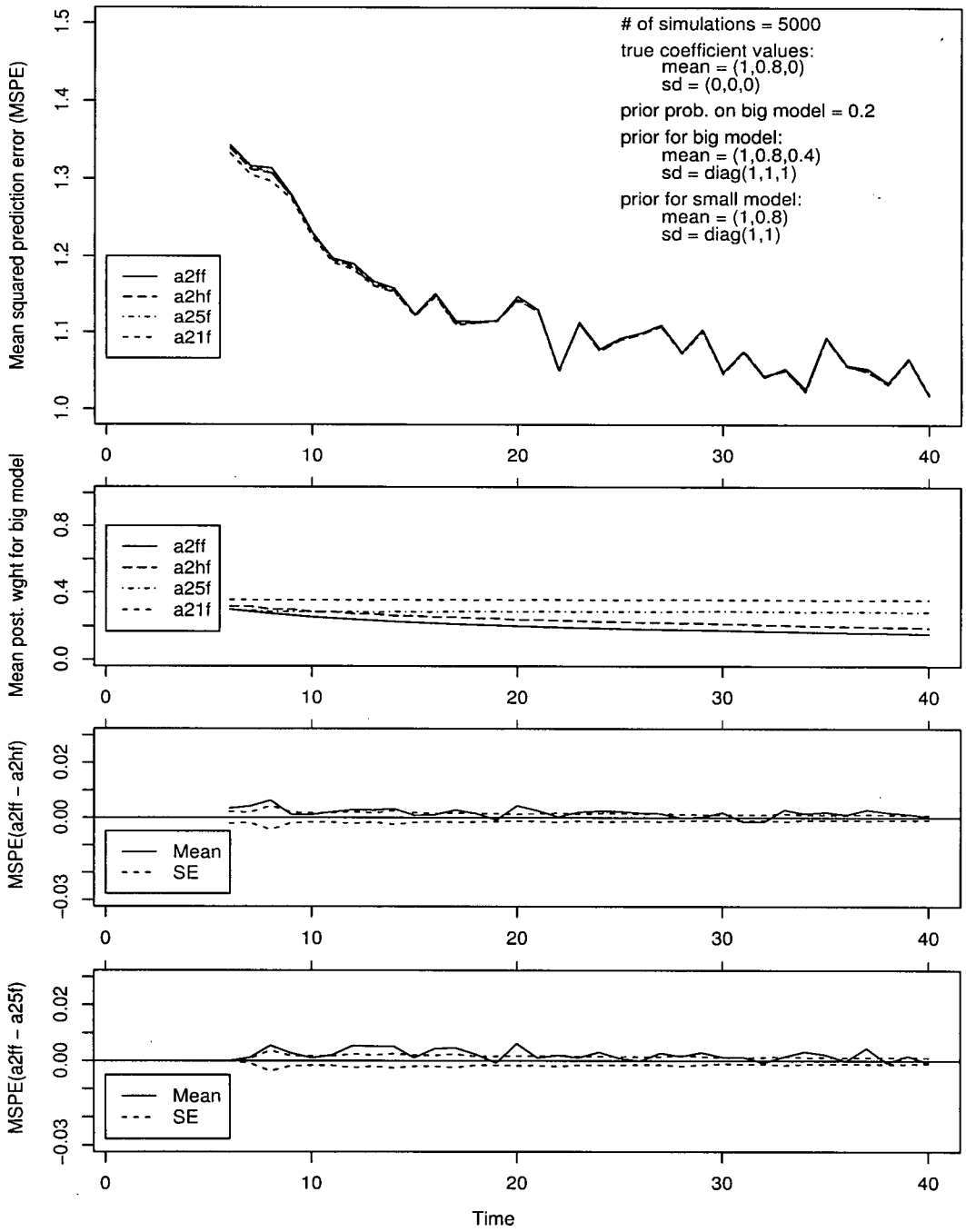


Figure 3.1: Performance of naive averaging strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0$.

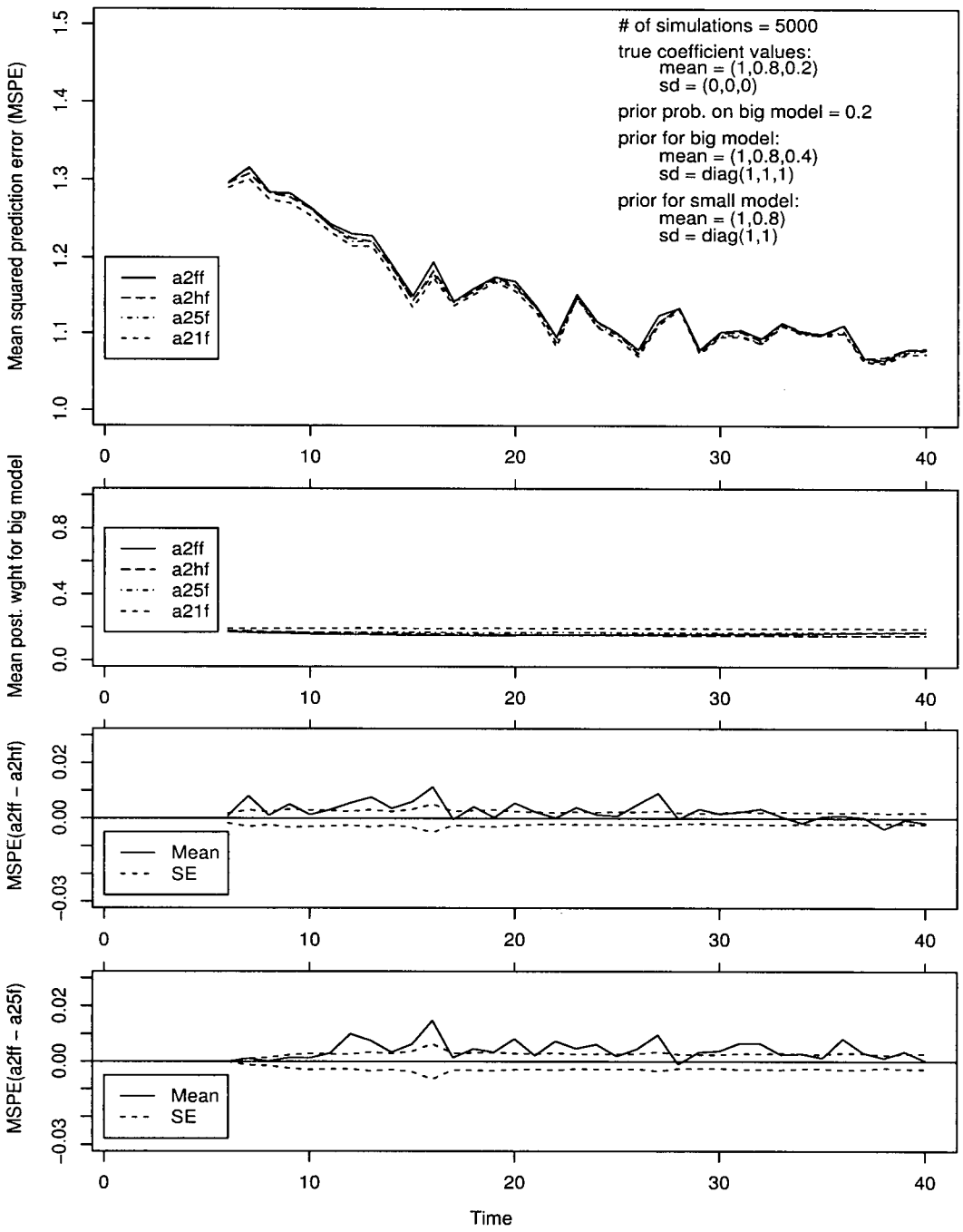


Figure 3.2: Performance of naive averaging strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0.2$.

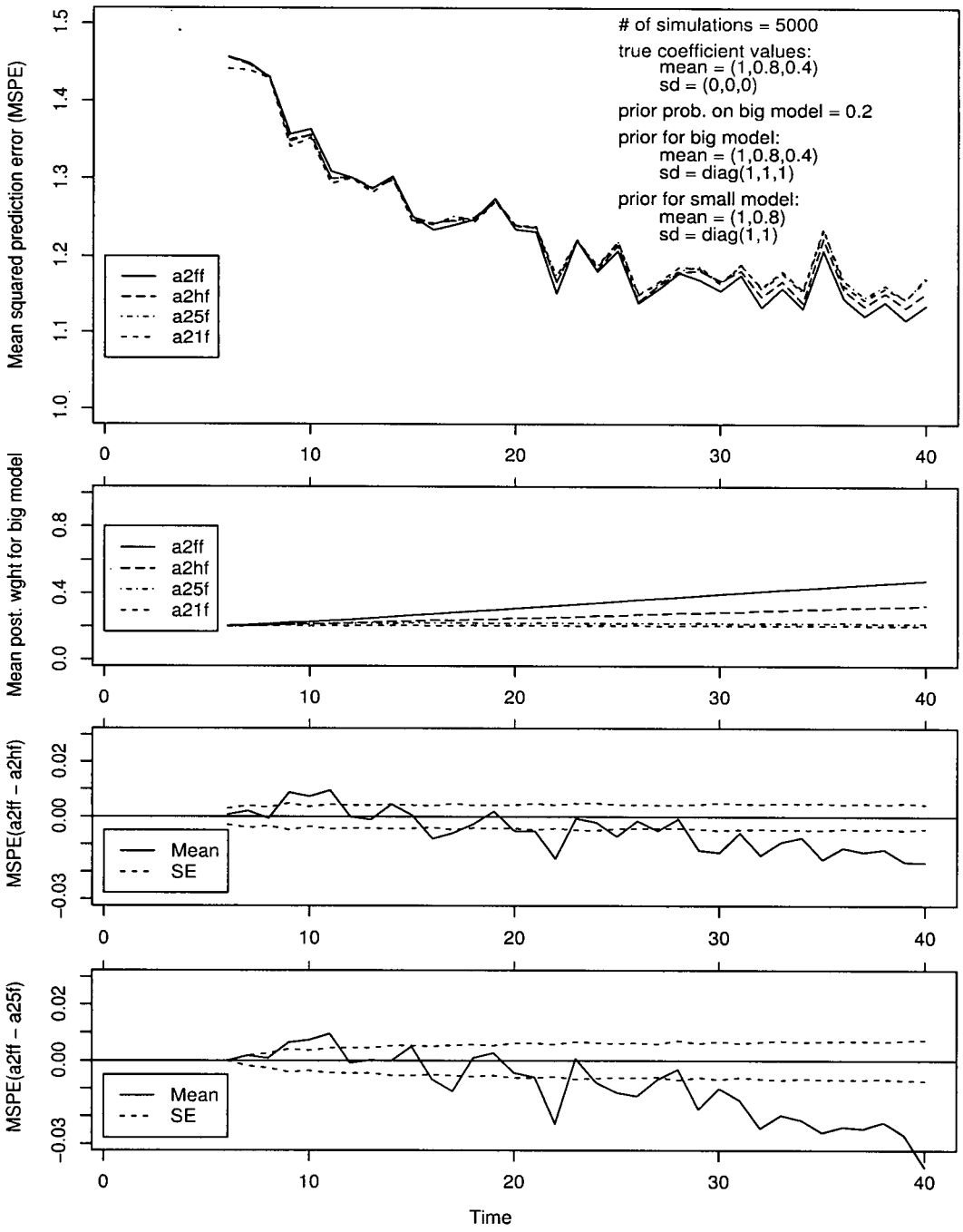


Figure 3.3: Performance of naive averaging strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0.4$.

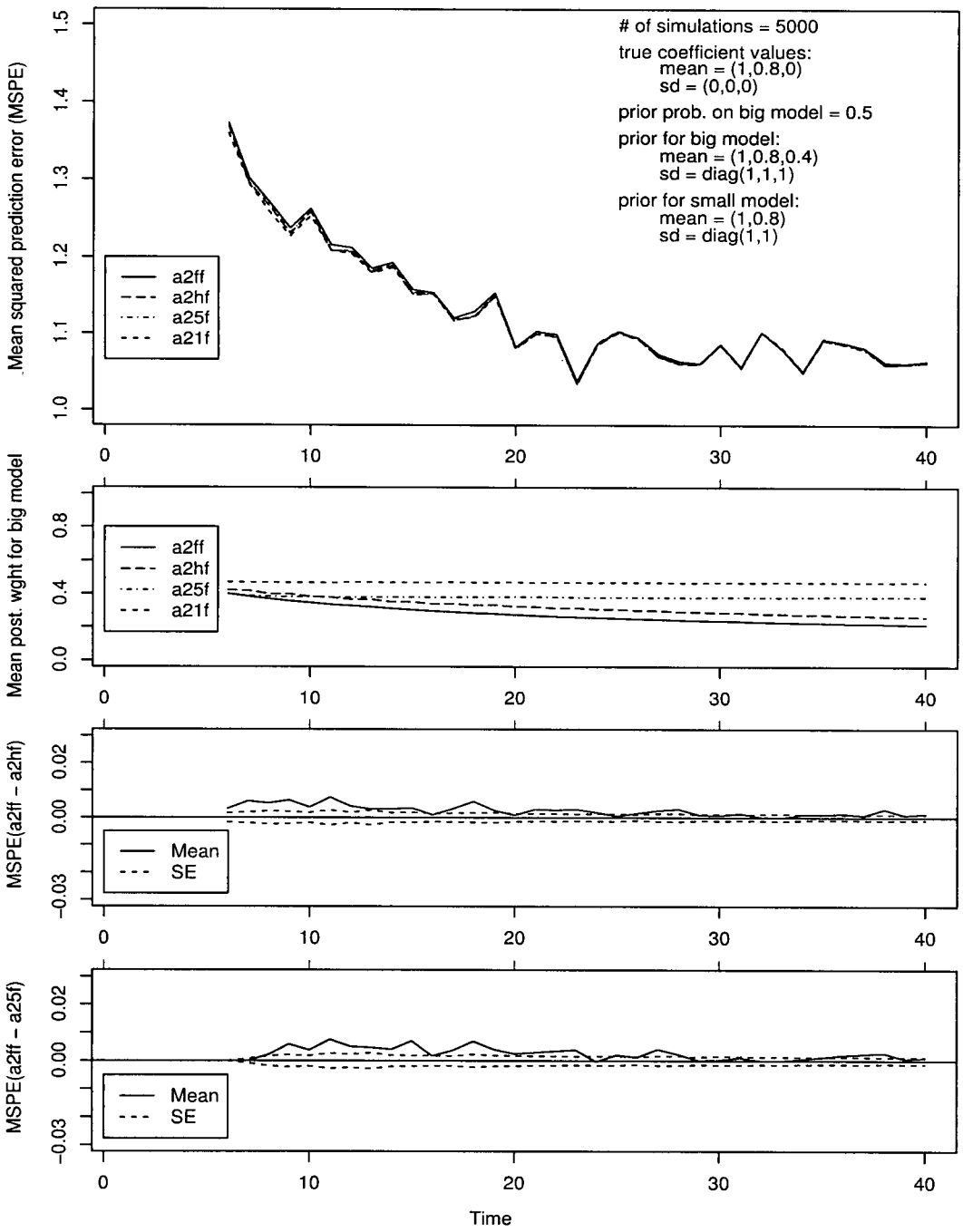


Figure 3.4: Performance of naive averaging strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0$.

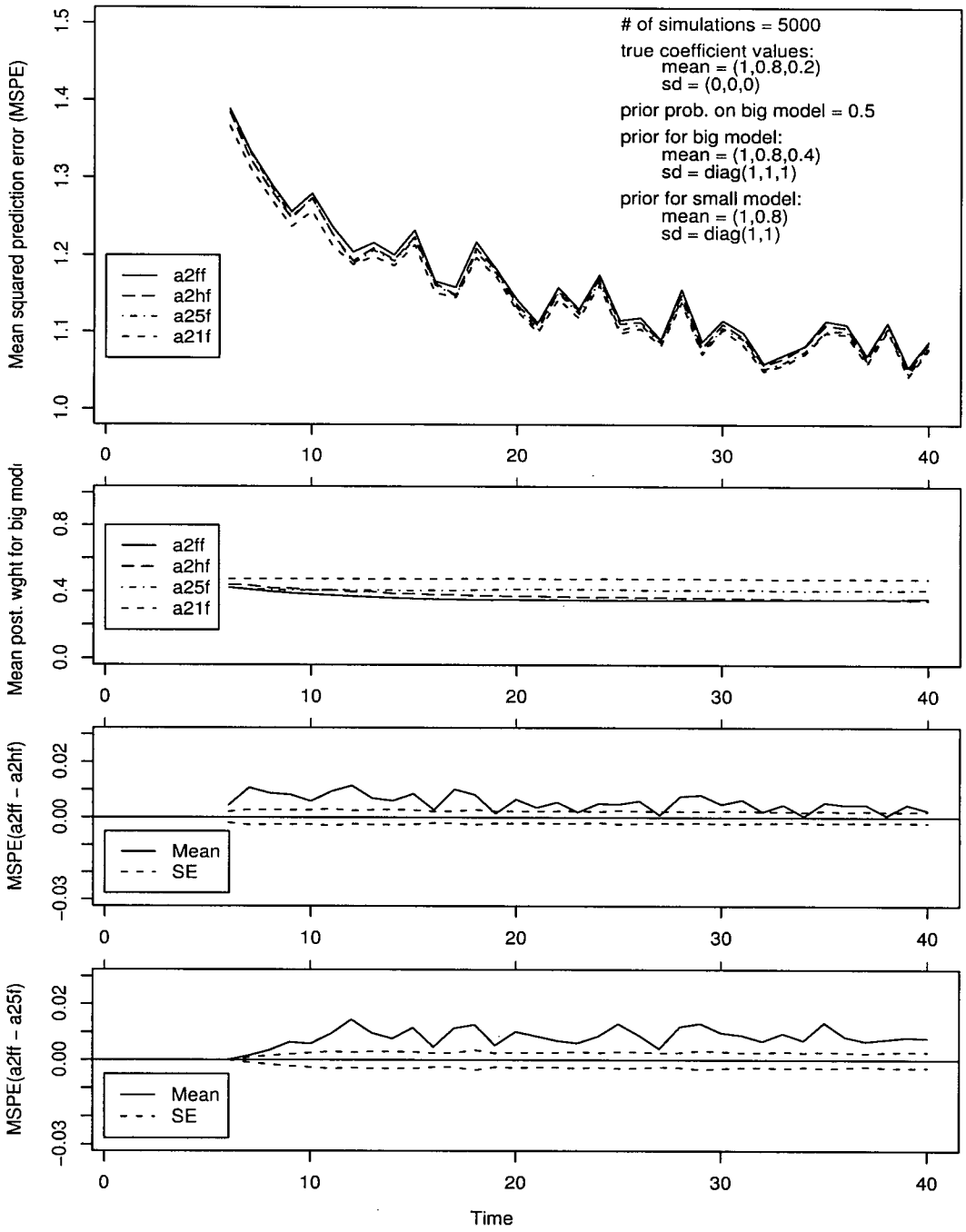


Figure 3.5: Performance of naive averaging strategies: $a_{2,o} = 0.5$, $\gamma_2 = 0.2$.

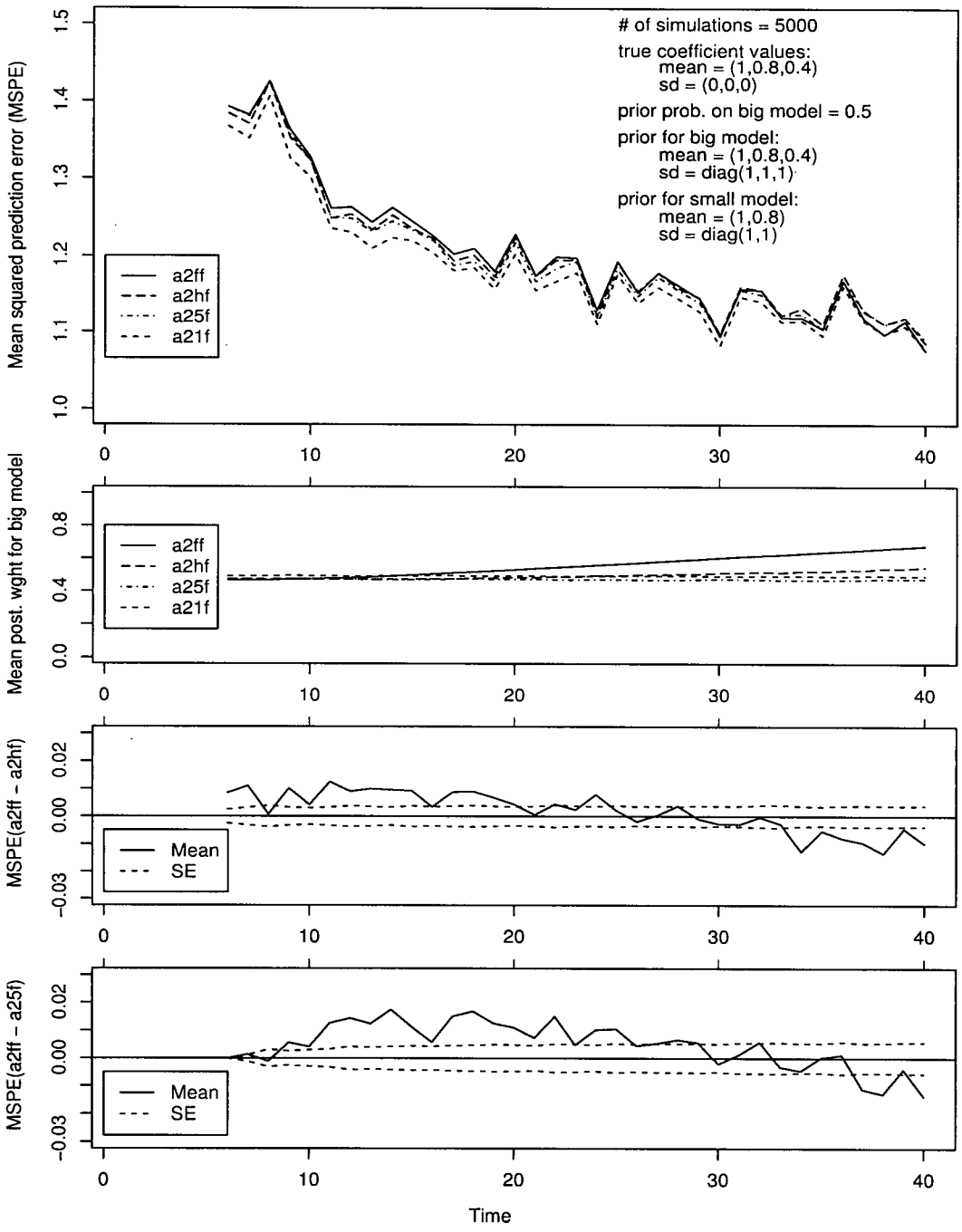


Figure 3.6: Performance of naive averaging strategies: $a_{2,o} = 0.5$, $\gamma_2 = 0.4$.

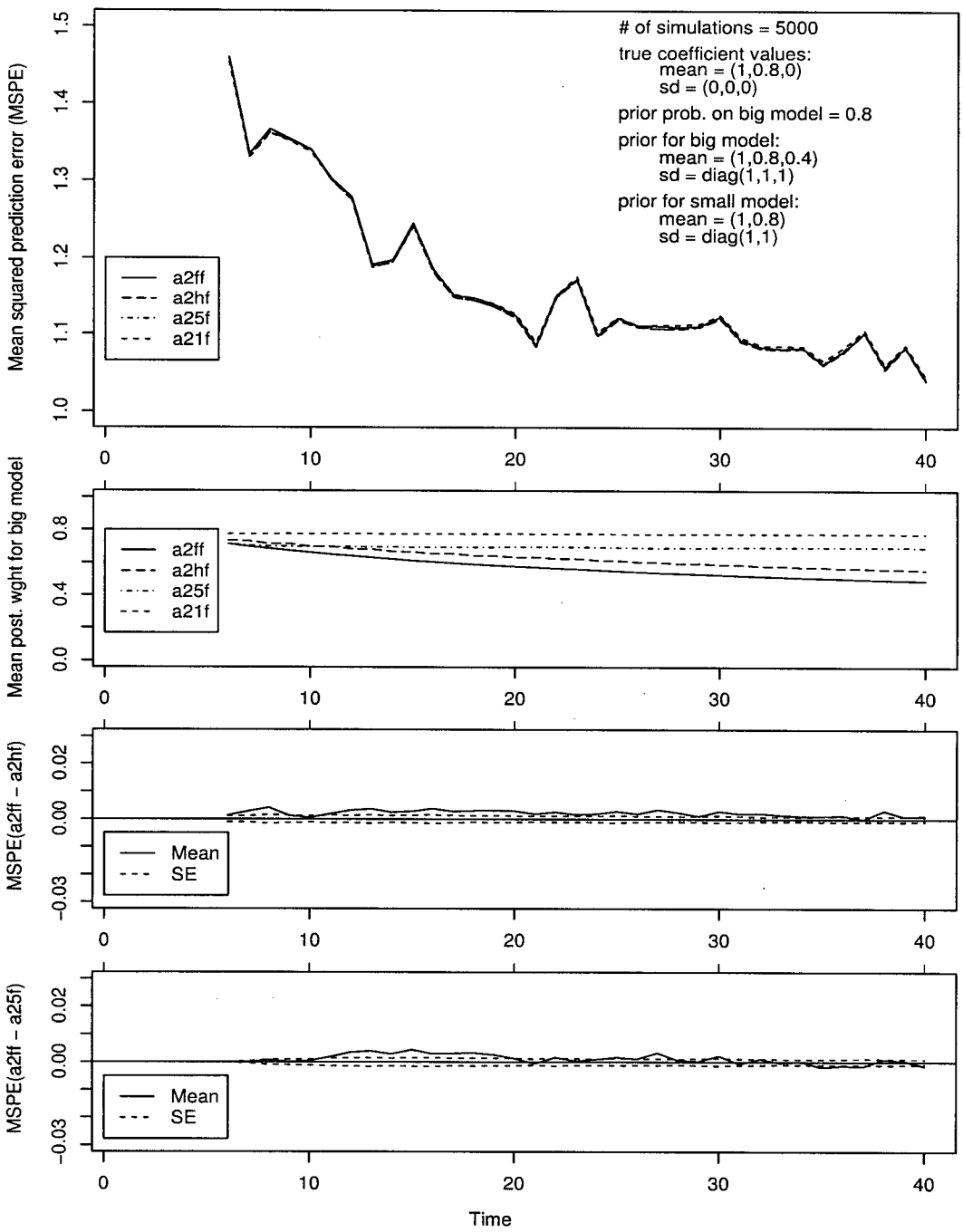


Figure 3.7: Performance of naive averaging strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0$.

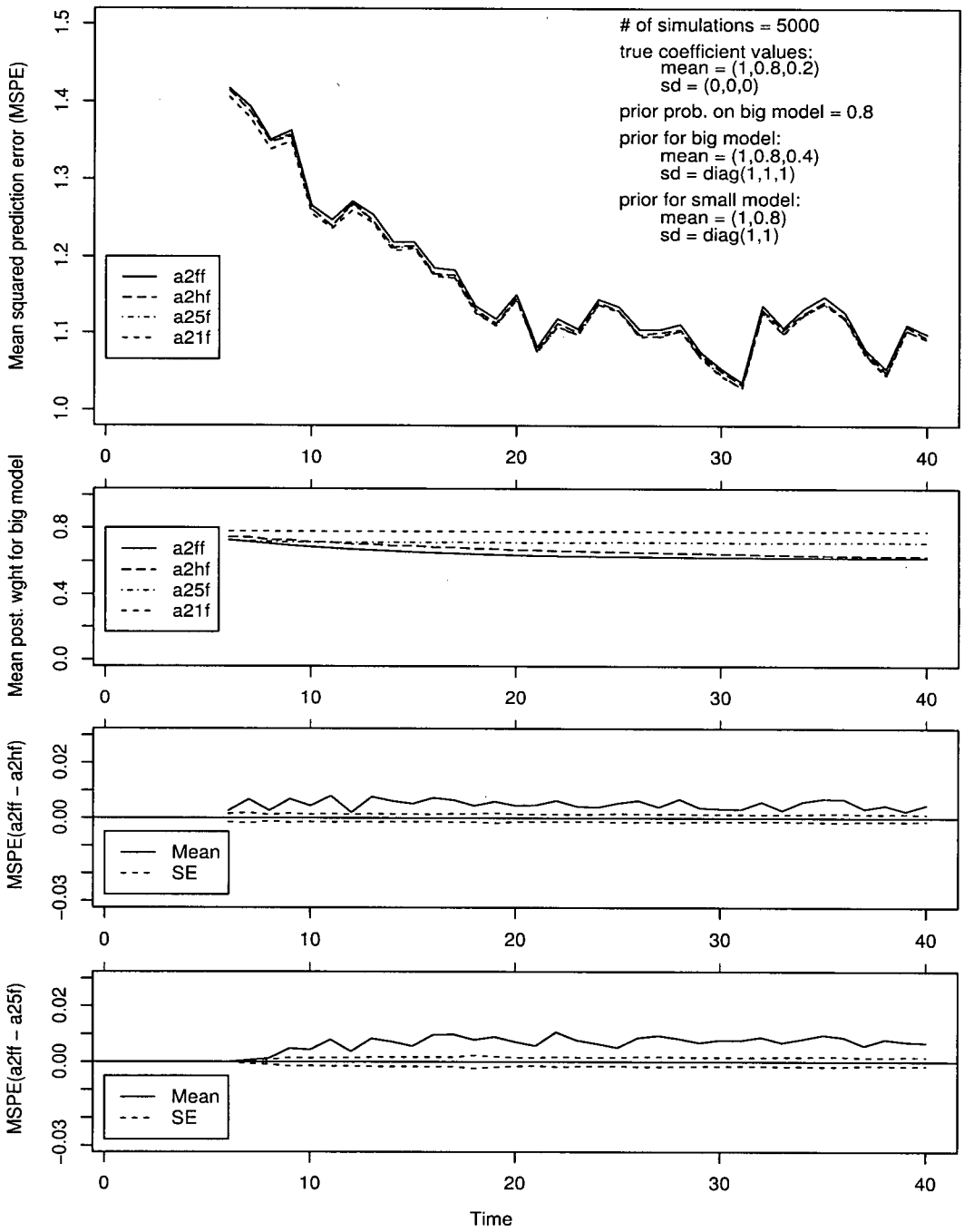


Figure 3.8: Performance of naive averaging strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0.2$.

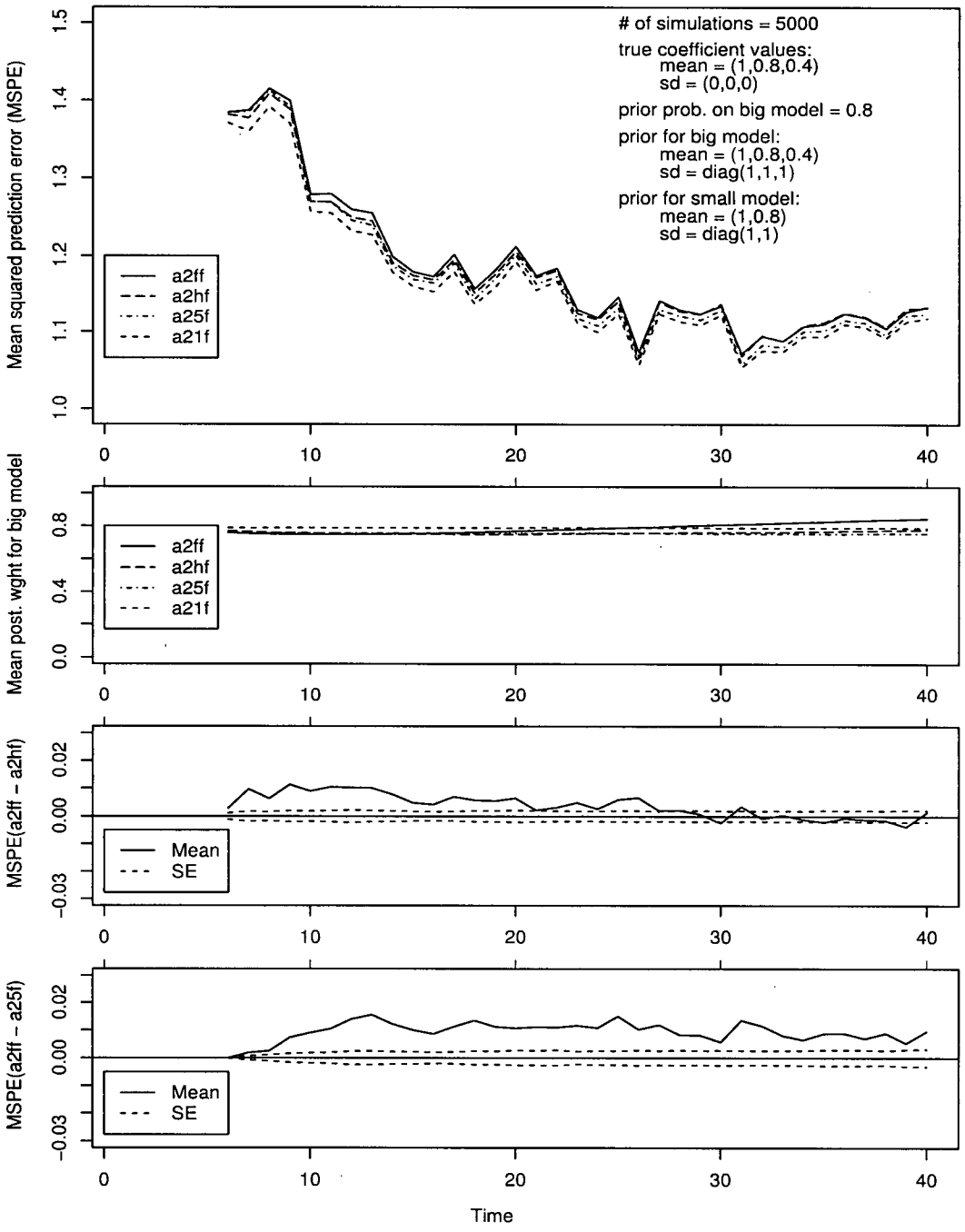


Figure 3.9: Performance of naive averaging strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0.4$.

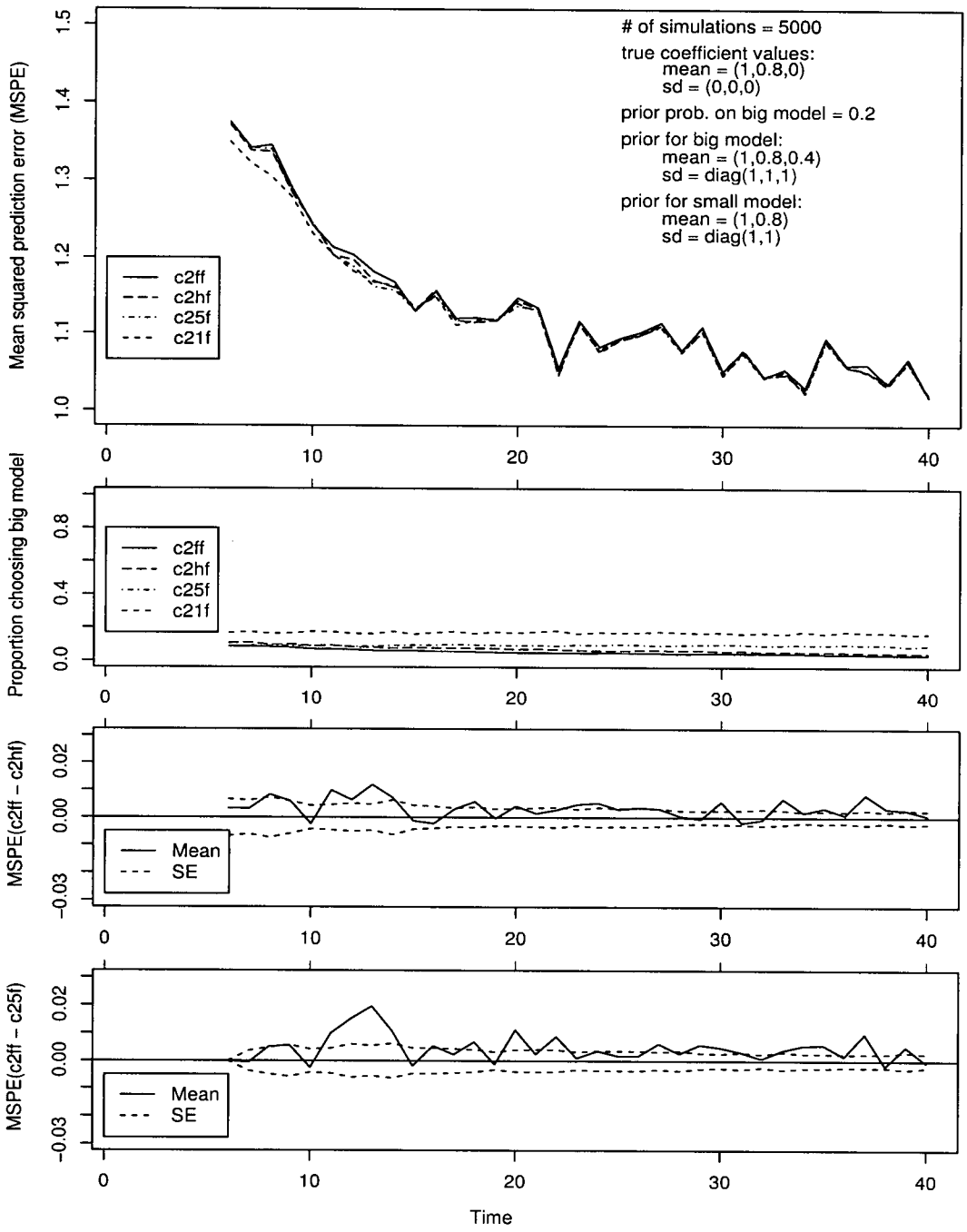


Figure 3.10: Performance of naive choice strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0$.

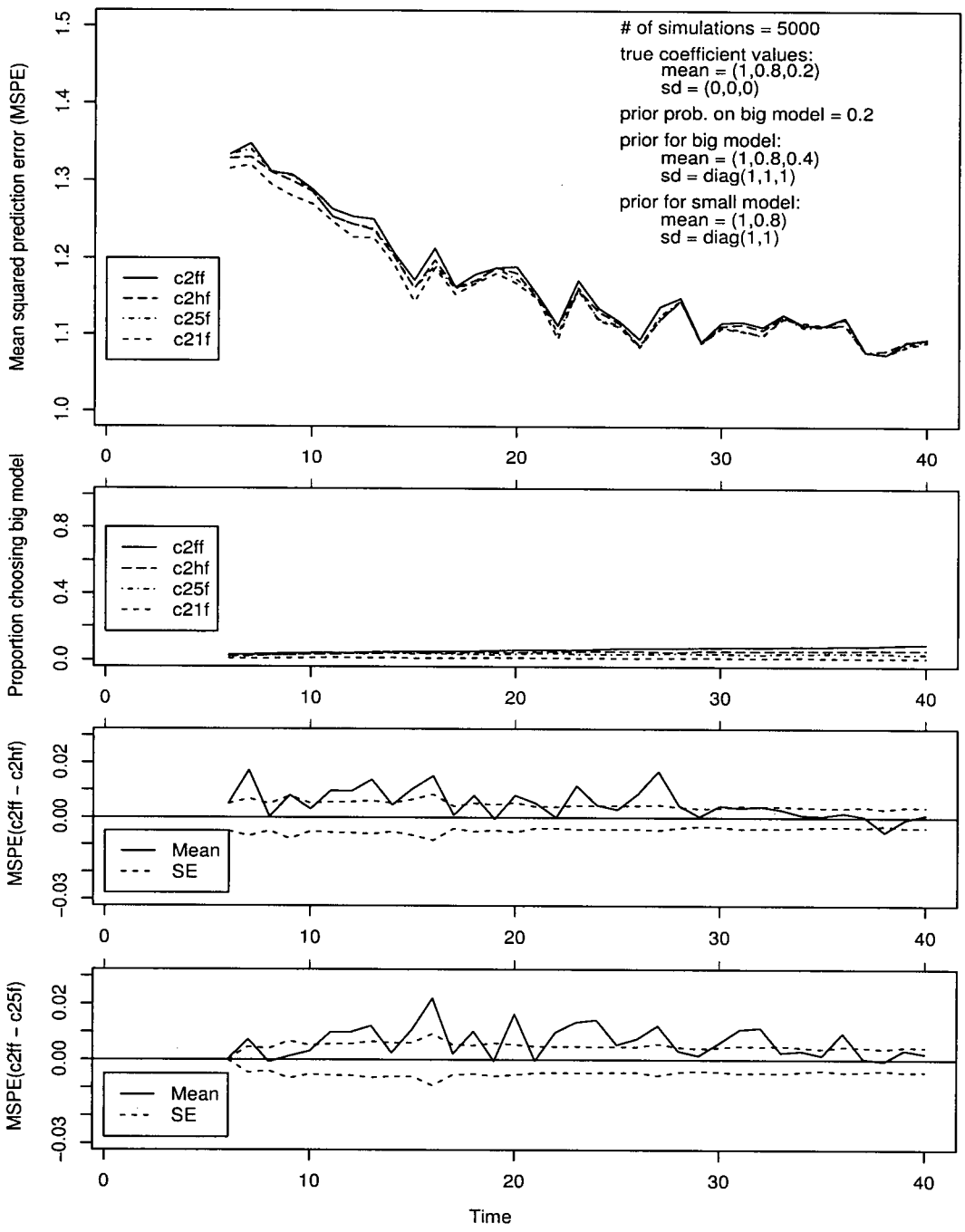


Figure 3.11: Performance of naive choice strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0.2$.

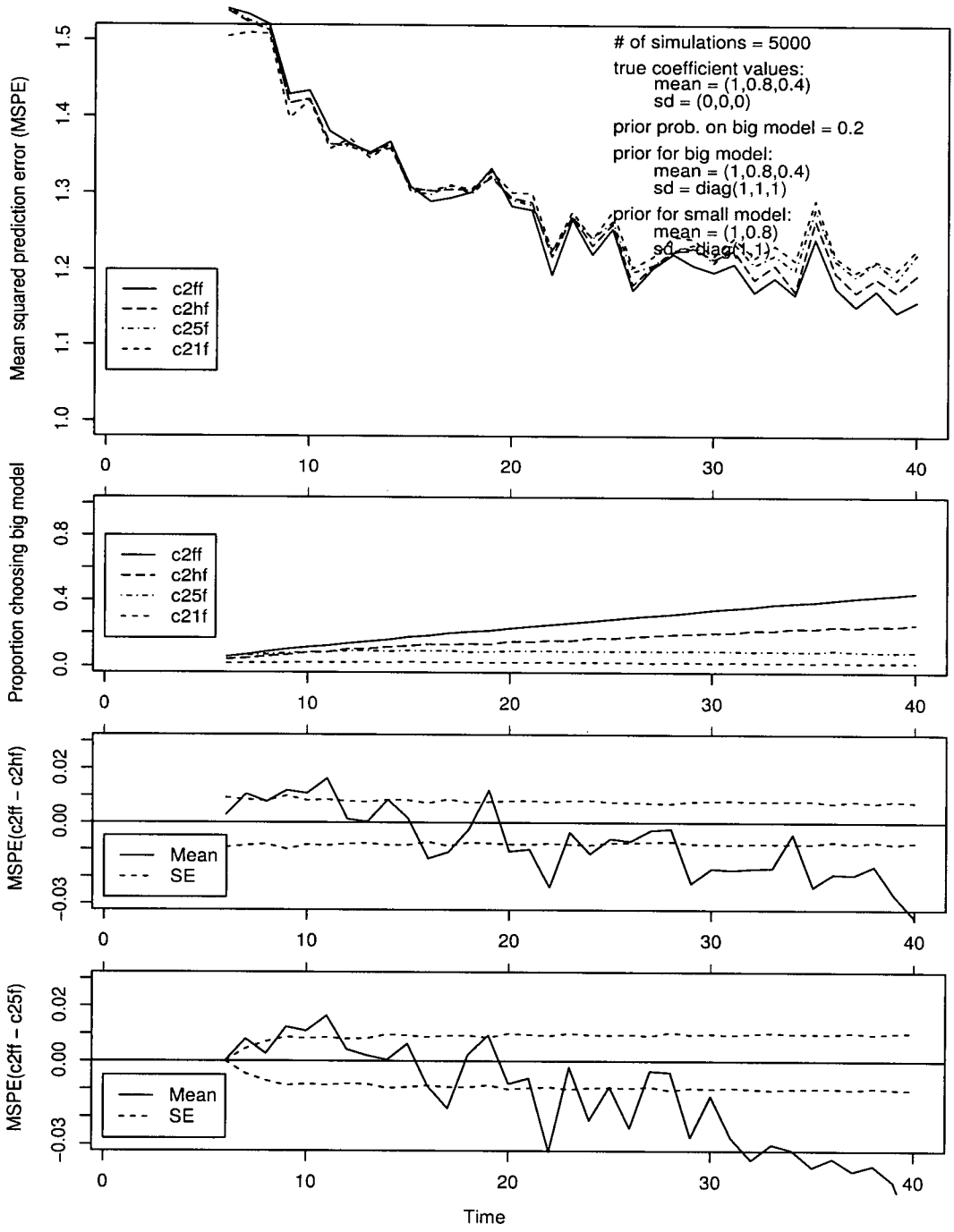


Figure 3.12: Performance of naive choice strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0.4$.

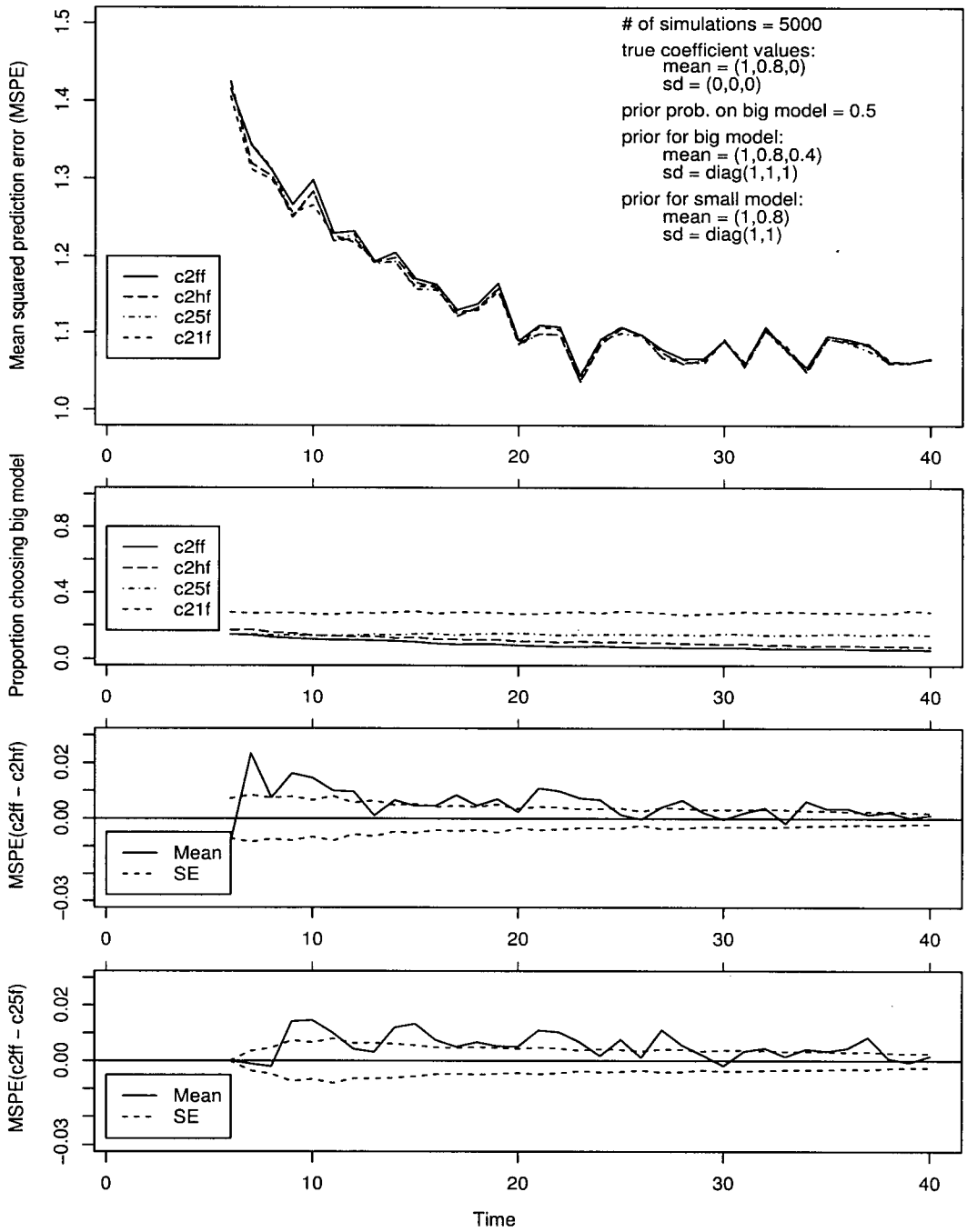


Figure 3.13: Performance of naive choice strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0$.

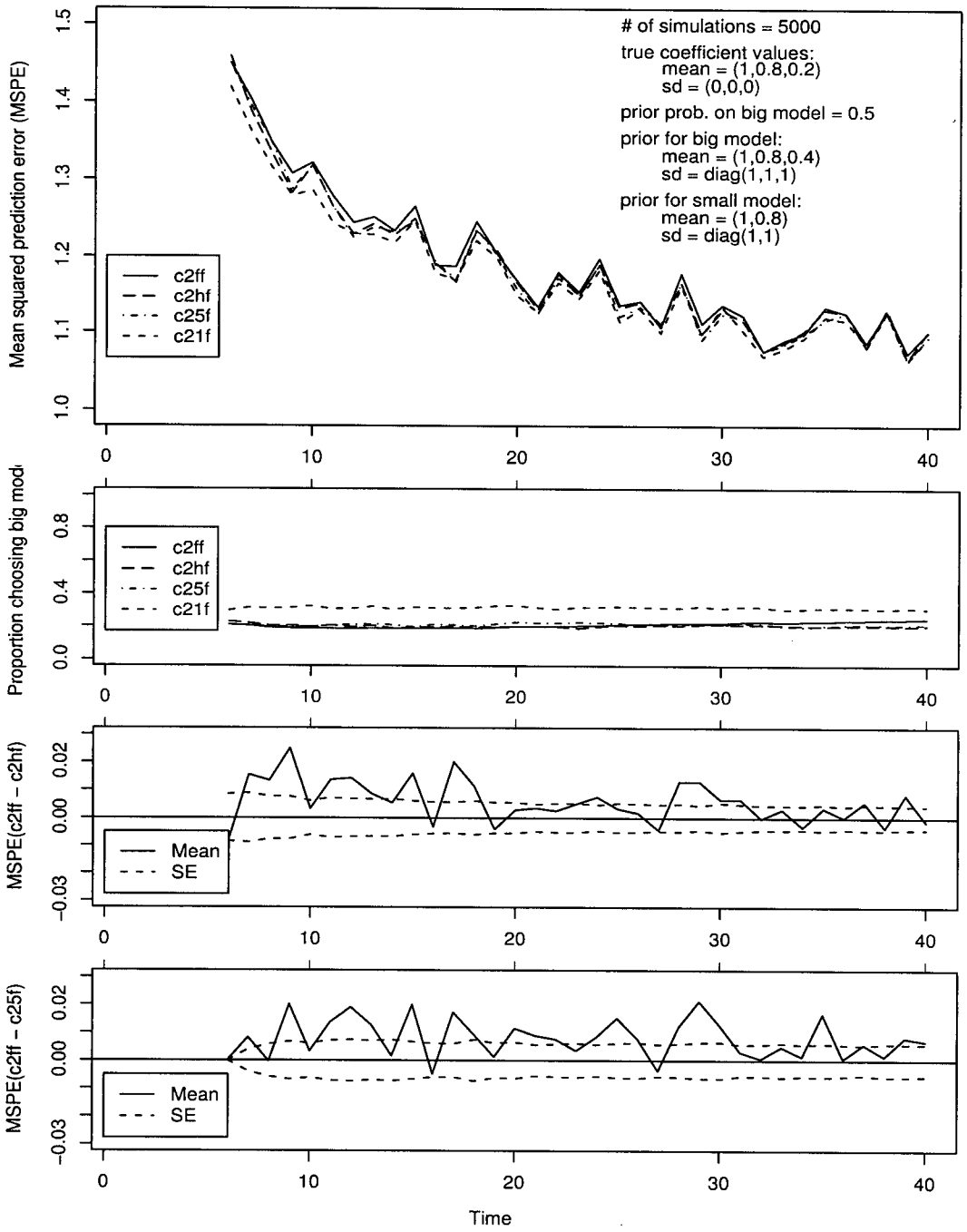


Figure 3.14: Performance of naive choice strategies: $a_{2,o} = 0.5$, $\gamma_2 = 0.2$.

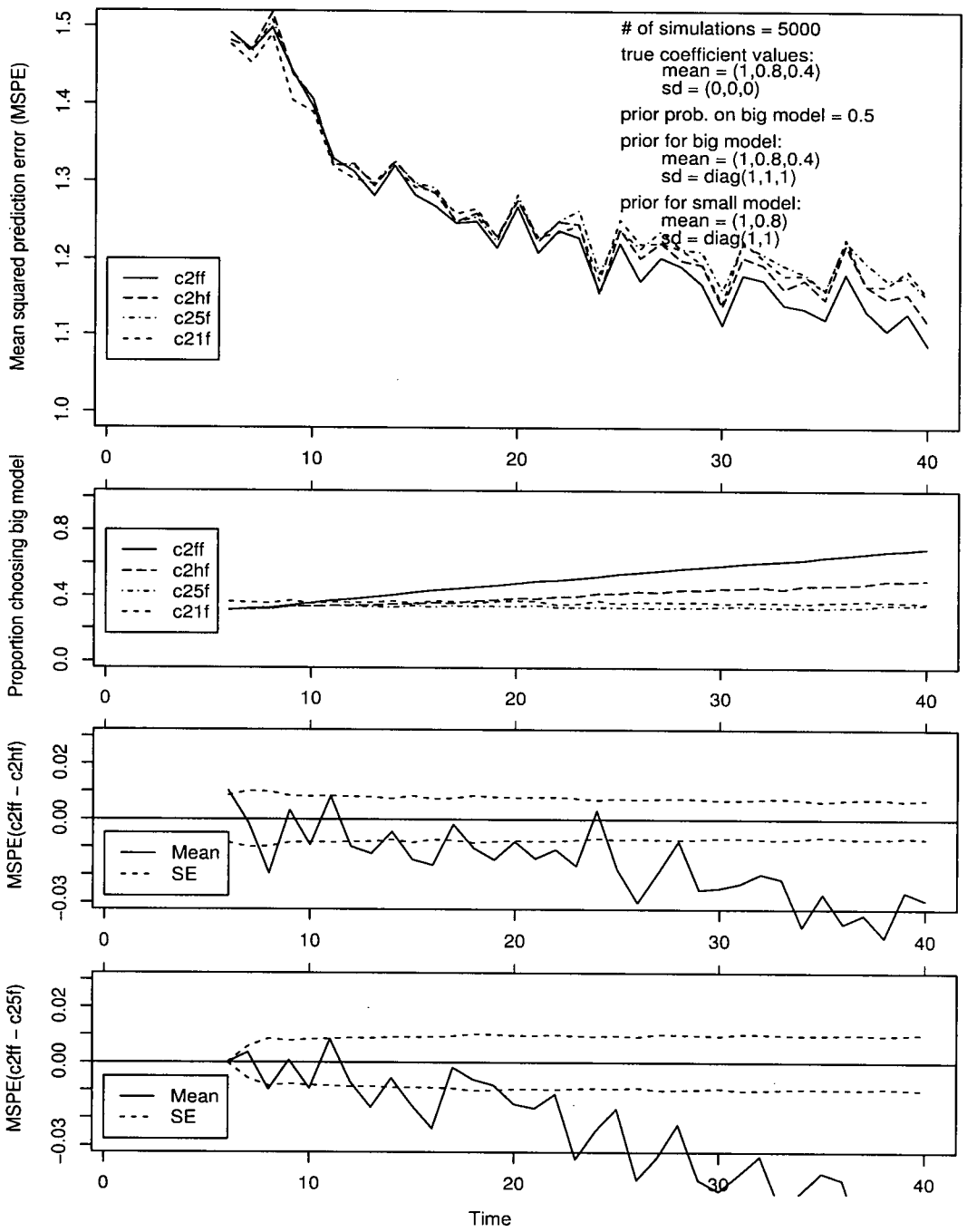


Figure 3.15: Performance of naive choice strategies: $a_{2,o} = 0.5$, $\gamma_2 = 0.4$.

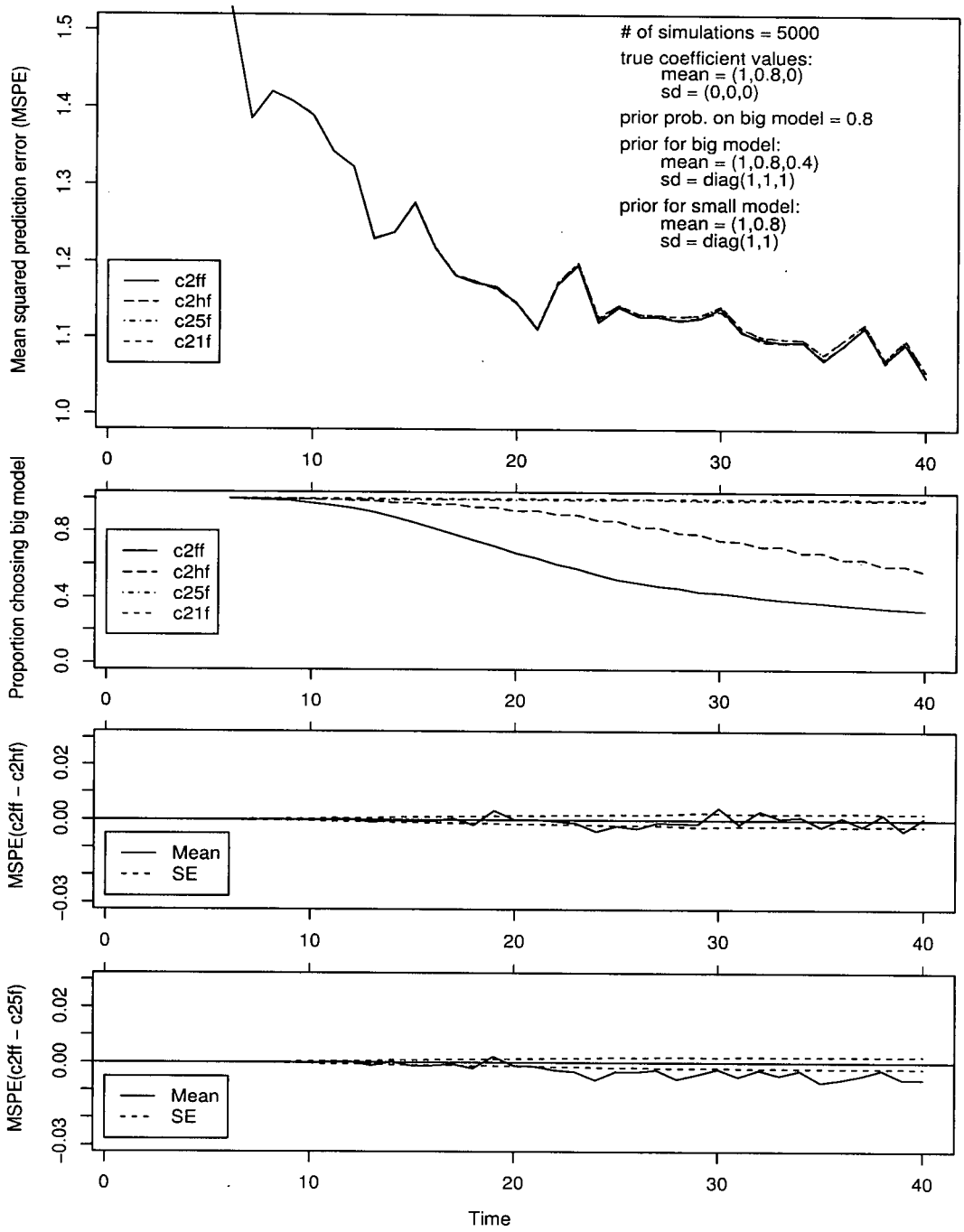


Figure 3.16: Performance of naive choice strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0$.

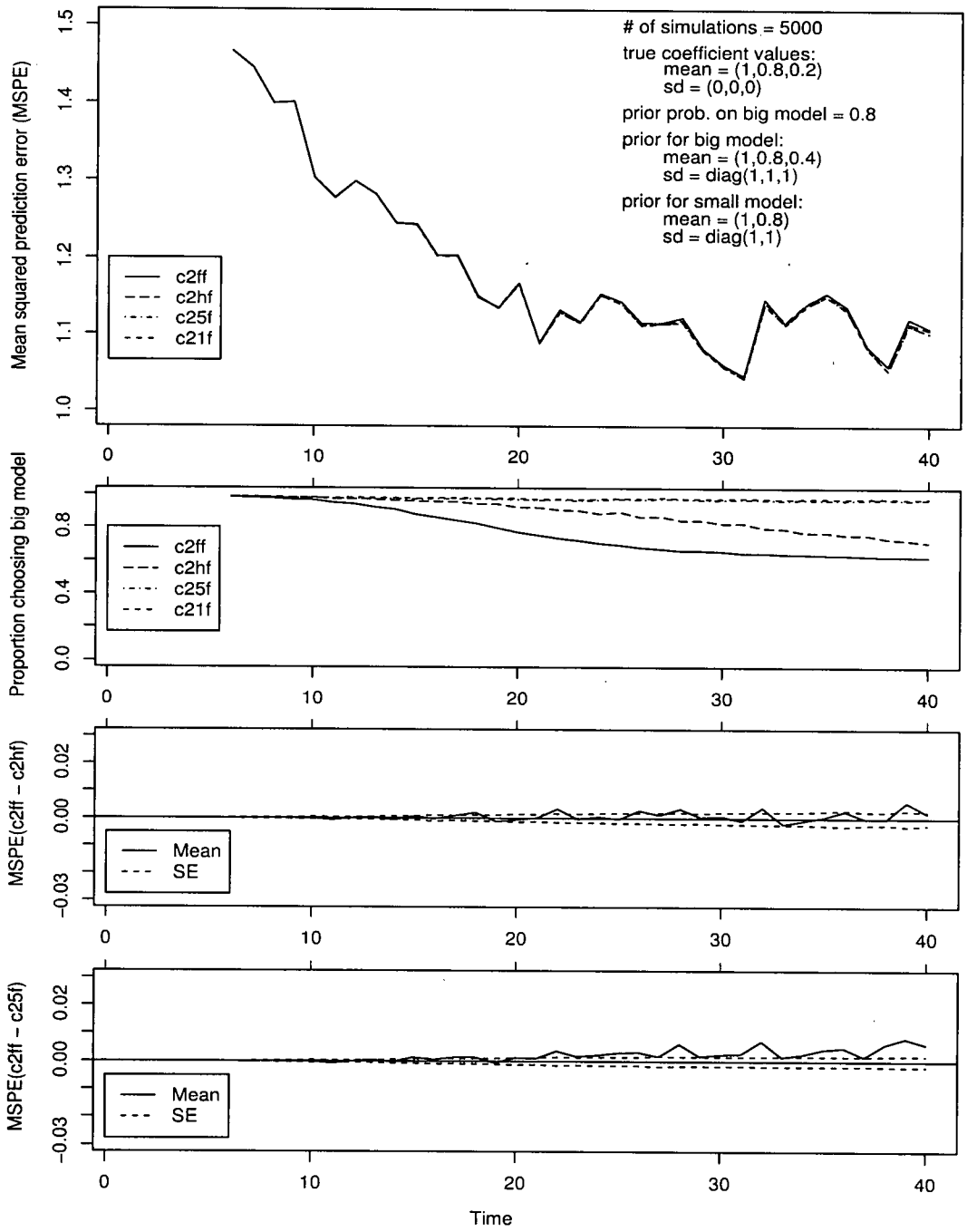


Figure 3.17: Performance of naive choice strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0.2$.

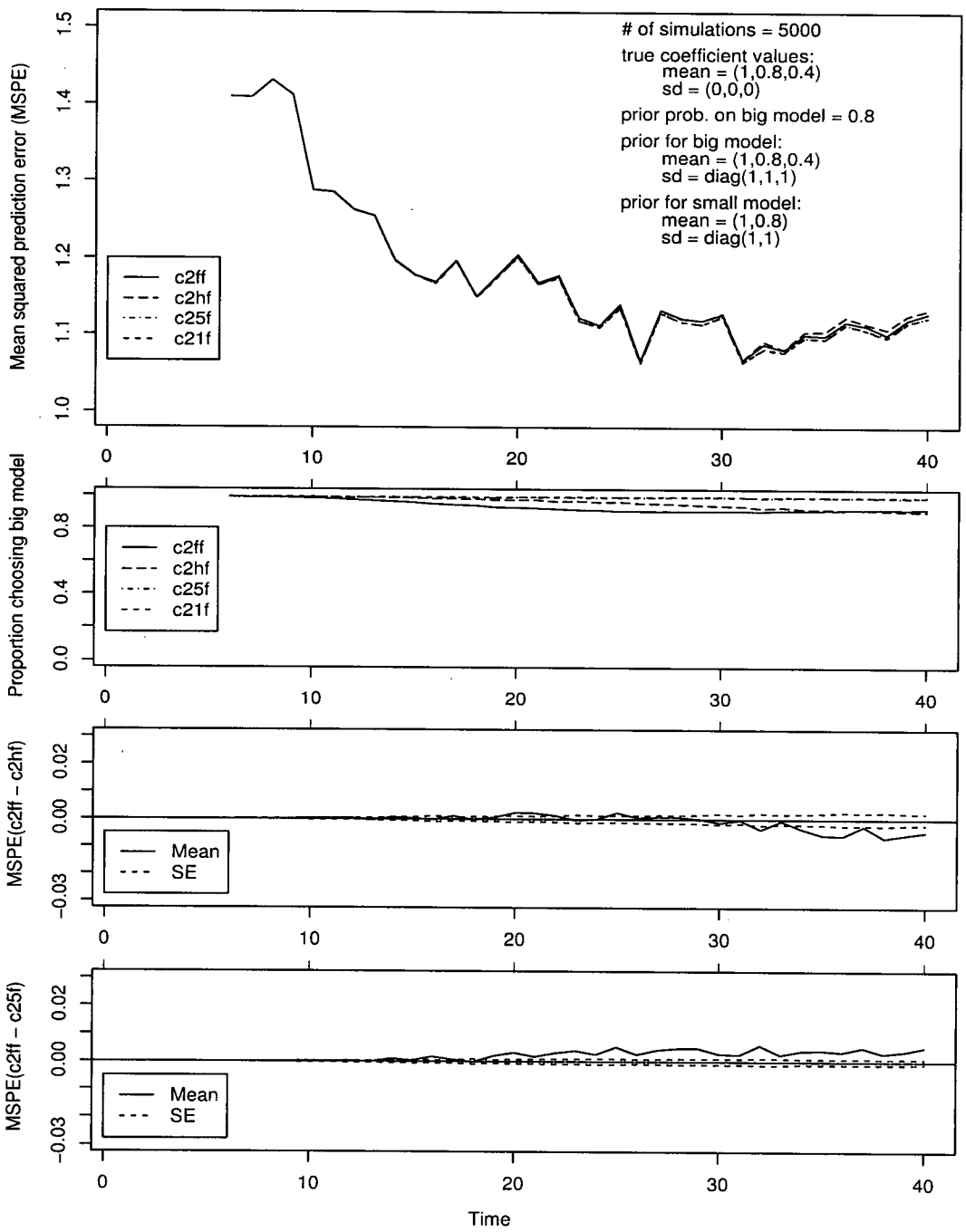


Figure 3.18: Performance of naive choice strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0.4$.

# Chapter 4

# Finite Samples: Adaptive Selection

We classify the methods for choosing $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ as either global or adaptive. In a global method, the choice is made without regard to the past outcomes $\mathbf{Y}_{(n)}$. In contrast, an adaptive method selects a different $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ for each sequence based on $\mathbf{Y}_{(n)}$. The reason for classifying in this way is to differentiate between methods according to whether the method uses past predictive performance. Note that a choice of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ that depends on $\mathbf{X}_{(n)}$ is considered to be global in this classification scheme.

We conjecture that adaptive methods yield better results than global methods. The intuition is that using all of the data, i.e., setting $\mathbf{S}_n^\alpha = \mathbf{S}_n^\rho = \mathbf{Y}_{(n)}$, provides the most accurate risk assessments for the 'typical' sequence but generates misleading assessments in other sequences. An adaptive method identifies and compensates for the misleading results by conditioning on past predictive performance.

To implement an adaptive selection method, we must specify a meta-

risk criterion, such as given by (2.27) or (2.28), that is to be used to evaluate each choice of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. We will use the simulation results from the framework described in Chapter 3 to help illustrate the method. As in Chapter 3, we set $\mathbf{S}_n^\rho = \mathbf{Y}_{(n)}$ and vary the choice for $\mathbf{S}_n^\alpha$. Here, we optimize the choice over the class $\mathbf{S}_n^\alpha \in \{\mathbf{R}_J : J = 0, 1, \ldots, n\}$. The index $J$ represents the number of most recent predictuals from the reduced model that are included in $\mathbf{S}_n^\alpha$. Recall that if we take $J = n$, then the optimal choice corresponds to that from the Bayes procedure.

The intuition underlying the meta-risks $\rho_\vee$ and $\rho_+$ is seen most easily by considering *meta-risk profiles.*

**Definition 4.1** *The* meta-risk profile *assuming model $i$ is true, is the collection of meta-risks $\rho_i(\hat{Y}_{n+1}; \mathbf{T}_n^\rho)$ generated by varying $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$.*

Thus, in our simulation framework, the meta-risk profile under a given model for predicting $Y_{n+1}$ contains the $n+1$ points corresponding to $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho) = (\mathbf{R}_J, \mathbf{Y}_{(n)})$, $J = 0, 1, \ldots, n$. Figure 4.1 displays graphically the meta-risk profiles as a function of $J$ for each of the first 12 simulated sequences of data. The value being predicted here is $Y_{10}$. The prediction $\hat{Y}_{n+1}$ was generated using an averaging strategy and we have set $\mathbf{T}_n^\rho = \mathbf{S}_n^\alpha$. In this particular scenario, $\gamma_2 = 0.4$ in the data-generator. The solid curve connects points in the meta-risk profile assuming the big model true the dashed curve connects points in the meta-risk profile assuming the small model is true.

In specifying a meta-risk criterion, we must consider the values on both these curves since the true model is unknown. On average, we expect that meta-risks computed conditional on more predictuals (larger $J$) tend to be more accurate. But, for the reasons discussed in Chapter 2, none of these meta-risks is the true risk so it cannot be argued that conditioning on full

data ($J = n$) is necessarily optimal. Instead, we should feel free to use fewer predictuals in $\mathbf{S}_n^\alpha$ if the resulting predictor is perceived to have more desirable characteristics. For example, if we find that the meta-risk of the predictor based on full data, say $J = 10$, is large under one of the models, call it $k$, but the meta-risk of the predictor based on $J = 9$ is (relatively) small under all of the models, then we might prefer to select $J = 9$ to avoid the large meta-risk that would be incurred if in fact model $k$ was true.

There are a variety of reasonable ways of reducing the meta-risk profiles to a choice of $J$. If for each $J$, we take the maximum of the meta-risks over all of the models, we get the maximum meta-risk criterion $\rho_\vee$ described by (2.28). Then, we obtain a "minimax" strategy by choosing the number of predictuals in $\mathbf{S}_n^\alpha$ to be

$$J_{\mathrm{mM}} = \arg\min_J \rho_\vee(\mathbf{R}_J, \mathbf{Y}_{(n)}; \mathbf{T}_n^\rho) \tag{4.1}$$

$$= \arg\min_J \max_k \rho_k(\hat{Y}_{n+1}(\mathbf{R}_J, \mathbf{Y}_{(n)}), \mathbf{T}_n^\rho). \tag{4.2}$$

(Note that in contrast to a standard risk function plot in which the parameter appears along the horizontal axis, the meta-risk profile plot places the elements in the decision space, i.e., the choices for $\mathbf{S}_n^\alpha$, on this axis.) Alternatively, if for each $J$, we average the values over all of the models using weights $\alpha_i(\mathbf{T}_n^\alpha)$, we get the weighted average meta-risk criterion $\rho_+$ described by (2.28). Then we obtain a "minimum-weighted-average" strategy by choosing the number of predictuals in $\mathbf{S}_n^\alpha$ to be

$$J_{\mathrm{mWA}} = \arg\min_J \rho_+(\mathbf{R}_J, \mathbf{Y}_{(n)}; \mathbf{T}_n^\alpha, \mathbf{T}_n^\rho) \tag{4.3}$$

$$= \arg\min_J \sum_k \alpha_k(\mathbf{T}_n^\alpha)\rho_k(\hat{Y}_{n+1}(\mathbf{R}_J, \mathbf{Y}_{(n)}), \mathbf{T}_n^\rho). \tag{4.4}$$

Additional strategies are discussed later in this chapter.

In Figure 4.1, the points in the weighted average meta-risk profile, with $\mathbf{T}_n^\rho = \mathbf{S}_n^\alpha$, are connected by the dotted curve. Let $[i, j]$ to denote the panel in the $i$-th row and $j$-th column. We observe that for the five sequences in [1,2], [2,1], [2,2], [4,3], and [5,3] the maximum meta-risk at each $J$ corresponds to the big model and that the minimum over these maxima appears to occur when $J = 9$, that is, when $\mathbf{S}_n^\alpha$ is equivalent to the full data. In contrast, for the sequence in, say, [1,1] the meta-risk profiles cross and the minimum of the maxima appears to occur when $J = 6$. The weighted average meta-risk profiles exhibit similar patterns; the minimum is achieved with $J = 9$ for the sequences in [2,1], [2,3], and [3,3] but with $J < 9$ in all the remaining sequences. We conjecture that when the minimum meta-risk is achieved with $J < 9$, the predictor based on the minimizing $\mathbf{S}_n^\alpha$ is less sensitive to model mis-specification than the Bayes predictor (for which $J = 9$). The general pattern seems to be that the meta-risk profile under the big model tends to decrease as $J$ increases, whereas the profile under the small model remains fairly flat for all $J$ or increases as $J$ increases.

The meta-risk profiles for the choice $\mathbf{T}_n^\rho = \mathbf{Y}_{(n)}$ (see Figure 4.2 are somewhat different in that the profile under the big model does not tend to decrease but rather remains relatively flat or increases slightly with $J$. The profile under the small model behaves in much the same way as when $\mathbf{T}_n^\rho = \mathbf{S}_n^\alpha$. The difference in behaviour for the profile under the big model results in much different values of $J_{\mathrm{mM}}$.

To assess how often the minimum was achieved for each value of $J$, we constructed histograms at time-points $n = 10$, 25, and 40. These histograms are shown in the first column of Figure 4.3 for a minimax strategy with $\mathbf{T}_n^\rho = \mathbf{S}_n^\alpha$. The last bin in each histogram represents the proportion of times that full data was selected for $\mathbf{S}_n^\alpha$. For this particular scenario, we see that the

modal choice was full data, but that the choices are spread over a very wide range with a tendency to picking a large number of predictuals. The second column in Figure 4.3 shows the histograms for when $\mathbf{T}_n^\rho = \mathbf{Y}_{(n)}$. In this case, the modal choice is to use no predictuals and the tendency is to select a smaller number of predictuals. These histograms will be seen to be useful for explaining the characteristics of different strategies; we defer further discussion to the following sections. We will discuss the results from taking a model averaging approach in detail and only comment briefly on the model choice approach. The reason is that we are able to implement the needed computation for model averaging but not for model choice.

## 4.1 Model Averaging

Through simulations, we first examined the performance of the "minimax" and "minimum-weighted-average" strategies from a model averaging approach. Because the results suggested deficiencies in the mWA strategies, we subsequently also considered two modified strategies which we label "Bayes-near-minimum" and "Bayes-factor-decisive". We look at the performance of each of the four strategies in turn.

### 4.1.1 Minimax meta-risk

Figures 4.4 through 4.12 show the performance obtained by using a minimax (mM) meta-risk strategy (see (4.2)). The MSPE obtained using the two strategies $(\mathbf{T}_n^\alpha, \mathbf{T}_n^\rho) = (\mathbf{S}_n^\alpha, \mathbf{S}_n^\alpha)$ (labeled 'a2xxmM') and $(\mathbf{T}_n^\alpha, \mathbf{T}_n^\rho) = (\mathbf{S}_n^\alpha, \mathbf{Y}_{(n)})$ ('a2xfmM') are shown in the top panel. For comparison, we have also plotted the MSPE obtained by the Bayes strategy ('a2ff') and, for reference, the

MSPE that would have been obtained had the full model ('big') or the reduced model ('small') been used at all times. The second panel from the top shows the average weight given to the big model. The third and fourth panels show the difference in MSPE between the Bayes strategy and each of the two mWA strategies. Positive values in these two plots indicate that the mWA strategy is better than the Bayes strategy.

The most important feature of 'a2xxmM' is that it never performs worse and often performs substantially better than Bayes. The performance of 'a2xfmM' is generally similar to 'a2xxmM'. However, 'a2xfmM' performs slightly worse than Bayes when $\gamma_2 = 0$ and $\alpha_{2,o}$ is 0.2 or 0.5. This loss is mitigated somewhat in that the size of the gain, when present, tends to be substantially greater than the gain seen for 'a2xxmM'. Both mM strategies exhibit greatest improvement early in the sequence. Tables 4.1 and 4.2 summarize the comparison of the Bayes strategy to the 'a2xxmM' and the 'a2xfmM' strategies, respectively, over all of the scenarios.

The similarity in performance is rather surprising since the histograms of $J_{\mathrm{mM}}$ (e.g. Figure 4.3) are radically different. Whereas 'a2xxmM' tends to pick all of the data or nearly all of the data, 'a2xfmM' tends to use none or nearly none of the data. At present, we do not have a good explanation for this difference.

### 4.1.2 Minimum-weighted-average meta-risk

Figures 4.13 through 4.21 show the performance obtained by using a minimum-weighted-average strategy (see (4.4). The MSPE obtained using the two strategies $(\mathbf{T}_n^\alpha, \mathbf{T}_n^\rho) = (\mathbf{S}_n^\alpha, \mathbf{S}_n^\alpha)$ (labeled 'a2xxwa') and $(\mathbf{T}_n^\alpha, \mathbf{T}_n^\rho) = (\mathbf{S}_n^\alpha, \mathbf{Y}_{(n)})$ ('a2xfwa') are shown in the top panel. For comparison, we have also plot-

ted the MSPE obtained by the Bayes strategy ('a2ff') and, for reference, the MSPE that would have been obtained had the full model ('big') or the reduced model ('small') been used at all times.

The second panel from the top shows the average weight given to the big model. The third and fourth panels show the difference in MSPE between the Bayes strategy and each of the two mWA strategies. Positive values in these two plots indicate that the mWA strategy is better than the Bayes strategy.

Early in the sequence, the 'a2xxwa' strategy beat the Bayes strategy substantially in 6 of the 9 scenarios, was slightly better in 1 scenario, was slightly worse in the scenario with $\gamma_2 = 0.4$ and $\alpha_{2,o} = 0.5$, and substantially worse in the scenario with $\gamma_2 = 0.4$ and $\alpha_{2,o} = 0.8$ As time increased, the performance of 'a2xxwa' relative to Bayes decreased in every scenario. The trend was consistent in every scenario. By time-point 40, 'a2xxwa' beat Bayes substantially in only 1 scenario. In all 3 scenarios where $\gamma_2 = 0.4$, 'a2xxwa' was substantially worse. In the remaining 5 scenarios, 'a2xxwa' was slightly better in three of them, about the same in one, and slightly worse in one.

The 'a2xfwa' strategy performed similarly to the 'a2xxwa' strategy when $\gamma_2 = 0$. But for most of the other cases, 'a2xfwa' performed worse than 'a2xxwa'. Generally, 'a2xfwa' behaved more like Bayes than 'a2xxwa' did, that is, while the differences in performance between 'a2xxwa' and Bayes were often large the differences between 'a2xfwa' and Bayes were relatively smaller. Table 4.3 (4.4) summarizes the comparisons of the Bayes strategy with the 'a2xxwa' ('a2xfwa') strategy for all of the figures.

In general, the performance of the mWA strategies appears to deteriorate as $\gamma_2$ increases. Additionally, the problem seems to be enhanced when $\alpha_{2,o}$ is small. We conjecture that using only some of the predictuals to update the model weights inhibits the identification of when the big model is true and

this identification is needed for avoiding bias. This conjecture is supported by the plots of the mean posterior weight assigned to the big model (second panel). When $\gamma_2 = 0.4$, $\alpha_{2,o}$ increases relatively rapidly for the Bayes strategy as data accumulates but not for the mWA strategies.

The mWA strategies behave oppositely to the mM strategies in that while mWA strategies work better when $\gamma_2$ is small the mM strategies work better when $\gamma_2$ is large.

### 4.1.3 Bayes-near-minimum meta-risk

Because averaging generally is considered a good way of handling model uncertainty, we thought that it would be nice if the minimum-weighted-average meta-risk strategy could be "patched up" to perform well even when $\gamma_2$ is large. In a large proportion of the sequences, the average meta-risk profile often dropped to its minimum $\rho_{\min} \equiv \rho_+(\hat{Y}_{n+1}(\mathbf{R}_{J_{\mathrm{mWA}}}, \mathbf{Y}_{(n)}); \mathbf{T}_n^\alpha, \mathbf{T}_n^\rho)$ once a small number of predictuals was included in $\mathbf{S}_n^\alpha$ and then remained at relatively flat for all larger $J$. If we take the view that a small difference between $\rho_{\min}$ and the meta-risk $\rho_{\mathrm{B}} \equiv \rho_+(\hat{Y}_{n+1}(\mathbf{R}_n, \mathbf{Y}_{(n)}); \mathbf{T}_n^\alpha, \mathbf{T}_n^\rho)$ for the full data (Bayes) procedure is not important, then we might want to default to using the full data rather than the number of predictuals corresponding to the minimum risk. That is, the number of predictuals to include in $\mathbf{S}_n^\alpha$ is

$$J_{\mathrm{BNM}} = \begin{cases} n & \rho_{\mathrm{B}}/\rho_{\min} < c \\ J_{\mathrm{mWA}} & \text{O.W.} \end{cases} \tag{4.5}$$

for a suitable cut-off value $c$. We call this the "Bayes-near-minimum" (BNM) strategy.

Tables 4.5 and 4.6 summarize the results for BNM strategies with $\mathbf{T}_n^\rho = \mathbf{S}_n^\alpha$ and $\mathbf{T}_n^\rho = \mathbf{Y}_{(n)}$, respectively. (We have not included the Figures that

would be analogous to Figures 4.13 through 4.21 due to length considerations.) The cut-off value was $c = 1.05$. While this approach managed to reduce the magnitude of the poor performance when $\gamma_2 = 0.4$, it did not eliminate it completely. Moreover, it also tended to reduce the size of the gains. Hence, we did not pursue this modification believing it to be ineffective.

This approach yielded another interesting fact. The given cut-off resulted in over 95% of the predictions, typical across all scenarios, being based on using all of the data. Yet, the differences in performance between the Bayes strategy and the NM strategies were sometimes nearly as large as those between the Bayes strategy and the mWA strategies. This result provides further evidence that a small proportion of the sequences generates much of the differences in performance (for better or for worse).

### 4.1.4 Bayes-factor-decisive meta-risk

Another way of assessing whether to default to the full data is to use the Bayes factor (BF) of the small model with respect to the big model, say. The idea here is similar to "Occam's window" – a model that has little support from the data (indicated by a small Bayes factor) ought to be viewed as discredited and discarded from consideration. Since we think that, on average, mongrel criteria give a less precise measure than the Bayes criterion for the true risk, we ought to avoid use of the mongrel criteria when we have confidence in the Bayes criterion (which we assume is reflected in a sufficiently large or small BF). Hence, the number of predictuals to include in $\mathbf{S}_n^\alpha$ is

$$J_{\text{BFD}} = \begin{cases} n & \text{if BF} > c \text{ or BF} < 1/c \\ J_{\text{mWA}} & \text{o.w.} \end{cases} \tag{4.6}$$

for some cut-off value $c$. We call this the "Bayes-factor-decisive" (BFD) strategy.

Tables 4.7 and 4.8 summarize the results for BFD strategies with $\mathbf{T}_n^\rho = \mathbf{S}_n^\alpha$ and $\mathbf{T}_n^\rho = \mathbf{Y}_{(n)}$, respectively. (Again, we have not included the Figures that would be analogous to Figures 4.13 through 4.21 due to length considerations.) The cut-off Bayes factor was $c = 4$. Overall, the BFD strategies were not successful in improving on the mWA strategies. The modification more often reduced the gains than reduced the losses seen originally in the mWA strategies.

## 4.2 Model Choice

Unfortunately, we are unable to optimize meta-risk when using model choice approach because the meta-risks cannot be evaluated in closed form. Specifically, $\hat{Y}_{n+1}$ takes the value $\hat{Y}_{1,n+1}$ if $\mathbf{Y}_{(n)}$ is in the set $\mathcal{S}_1 = \{\bar{\rho}(\hat{Y}_{1,n+1}; \mathbf{S}_n^\alpha, \mathbf{S}_n^\rho) < \bar{\rho}(\hat{Y}_{2,n+1}; \mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)\}$ or the value $\hat{Y}_{2,n+1}$ if $\mathbf{Y}_{(n)}$ is in the complement, $\mathcal{S}_2$, of $\mathcal{S}_1$. Hence the meta-risk is

$$\rho_i(\hat{Y}_{n+1}; \mathbf{T}_n^\rho) = \mathbf{E}_{i|\mathbf{T}_n^\rho} \sum_k (Y_{n+1} - \hat{Y}_{k,n+1})^2 \chi_{\mathcal{S}_k} \tag{4.7}$$

where $\chi_A$ is the indicator of the set $A$. The integral in (4.7) is not tractable analytically.

## 4.3 Relationship between mWA and mM meta-risk

A standard result in estimation theory is that, typically, the minimax decision is equivalent to the Bayes decision under a least favourable prior on the

parameter. A natural question to ask is whether the mM meta-risk choice of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ is equivalent to the mWA meta-risk choice. The following result shows the mM and mWA procedures are equivalent only in a trivial case.

**Lemma 4.1** *Let*

$$\pi^* = \arg\max_{\pi} \min_{(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)} \sum_k \pi_k \rho_k(\hat{Y}_{k,n+1}; \mathbf{T}_n^\rho) \tag{4.8}$$

*where* $\pi = (\pi_1, \ldots, \pi_K)$ *is a probability on the model space. That is,* $\pi^*$ *is a least favourable prior. Then the mM and mWA meta-risks yield the same choice for* $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ *(almost surely in* $\mathbf{Y}_{(n)}$*) iff* $\forall k$*,*

$$\alpha_k(\mathbf{T}_n^\alpha) = \pi_k^*. \tag{4.9}$$

**Proof** The problem can be viewed as a finite zero-sum game in which the choice of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ represents the decision to be made and the model space is the parameter space. Then the Minimax Theorem (Berger, p. 345; 1985) applies so that the game has value

$$V = \inf_\delta \sup_\pi \sum_k \pi_k \rho_k(\hat{Y}_{n+1}; \mathbf{T}_n^\rho) = \sup_\pi \inf_\delta \sum_k \pi_k \rho_k(\hat{Y}_{n+1}; \mathbf{T}_n^\rho) \tag{4.10}$$

where the infimum is over all randomized choices of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ and the supremum is over all probabilities on the model space. By Lemma 1 (Berger, p.318; 1985), the value of the game can also be expressed as

$$V = \inf_{(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)} \sup_k \rho_k(\hat{Y}_{n+1}; \mathbf{T}_n^\rho) = \sup_\pi \inf_{(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)} \sum_k \pi_k \rho_k(\hat{Y}_{n+1}; \mathbf{T}_n^\rho). \tag{4.11}$$

Finally, the infimums and supremums can be replaced by minima and maxima since both the model and decision spaces are finite and therefore closed, so we have

$$\min_{(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)} \max_k \rho_k(\hat{Y}_{n+1}; \mathbf{T}_n^\rho) = \max_\pi \min_{(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)} \sum_k \pi_k \rho_k(\hat{Y}_{n+1}; \mathbf{T}_n^\rho). \tag{4.12}$$

The choice of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ on the L.H.S. of (4.12) corresponds to the choice obtained using the mM strategy. The choice of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ on the R.H.S. of (4.12) will correspond to the choice obtained using the mWA strategy iff $\forall k, \alpha_k(\mathbf{T}_n^\alpha) \equiv \pi_k^*$. ∎

However, since $\pi_k^*$ is constant in $\mathbf{T}_n^\alpha$, it is clear that (4.9) is possible only if $\mathbf{T}_n^\alpha$ is held constant for all $\nu$. We are primarily interested in the case where $\mathbf{T}_n^\alpha$ varies with $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ and so in general the results we obtain using the minimax meta-risk are not equivalent to the results obtained using the minimum weighted average meta-risk for any such choice of $\mathbf{T}_n^\alpha$.

## 4.4 Discussion

At first glance, some of the differences in behaviour of these strategies seem natural while others seem counterintuitive. The results suggest that the mM strategies are more robust than the WA strategies. Whereas 'a2xxmM' never performs worse than Bayes and 'a2xfmM' at most performs slightly worse than Bayes, both 'a2xxwa' and 'a2xfwa' perform much worse than Bayes when $\gamma_2 = 0.4$. This result agrees with the usual sense in which minimax strategies are robust. On the other hand, weighted averaging strategies are expected to perform better overall. This does not seem to occur here – the gains from the mM strategies appear to exceed those seen for the WA strategies in general.

A partial explanation may be that the optimization over the meta-risk is not intended to generate an optimal predictor directly but merely to decide on a good $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ on which the optimal predictor will then be computed. In this case, the most desirable property of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ is that it minimizes the chance that our subsequent choice of a predictor will be poor because it derives

from a bad model. That is, minimizing meta-risk is inherently a robustness issue and that may be why the mM meta-risks generate better results.

A comparison of strategies using $\mathbf{T}_n^\rho = \mathbf{S}_n^\alpha$ to $\mathbf{T}_n^\rho = \mathbf{Y}_{(n)}$ shows a clear pattern. For all four methods of selecting the optimal value of $J$, the choice $\mathbf{T}_n^\rho = \mathbf{S}_n^\alpha$ generated more accurate predictions in nearly all of the scenarios. This pattern would be expected if our reasoning for the advantages of using mongrel risk is correct, i.e., we get better assessments of risk by conditioning on empirical performance rather than the raw data. However, this is only a conjecture at this time.

Note that if the choices for $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$ are varied independently of each other, then we would have $n^2$ potential choices for $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ since each of the two statistics can range from 0 up to $n-1$ predictuals. We have not optimized over all $n^2$ choices for two reasons. First, the expectations in the meta-risks (2.27) or (2.28) can be evaluated analytically only when $\sigma(\mathbf{S}_n^\alpha) \subseteq \sigma(\mathbf{T}_n^\rho)$. (so that the weights $\alpha_k(\mathbf{S}_n^\alpha)$ can be taken outside the expectation $\mathbf{E}_{i|\mathbf{T}_n^\rho}$). In general, $\mathbf{T}_n^\rho$ is not set to be equivalent to the full data so some choices of $\mathbf{S}_n^\alpha$ cannot be handled. The second reason is that we wish to dissociate the effects due to varying $\mathbf{S}_n^\alpha$ from the effects due to varying $\mathbf{S}_n^\rho$. We have examined only the impact of varying $\mathbf{S}_n^\alpha$ because this situation was easiest to implement computationally. The case with $\mathbf{S}_n^\rho$ varied requires additional computational effort and has been left for future investigation.

Table 4.1: Summary comparison of the mM mongrel averaging strategy with $\mathbf{T}_n^{\rho} = \mathbf{S}_n^{\alpha}$ to the Bayes strategy.

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| I | 0.2 | 0 | + | 0 | 0 |
| | | 0.2 | + | 0 | 0 |
| | | 0.4 | + | + | + |
| | 0.5 | 0 | ++ | + | 0 |
| | | 0.2 | ++ | + | + |
| | | 0.4 | +++ | ++ | + |
| | 0.8 | 0 | ++ | ++ | + |
| | | 0.2 | +++ | ++ | ++ |
| | | 0.4 | +++ | ++ | + |

Table 4.2: Summary comparison of the mM mongrel averaging strategy with $\mathbf{T}_n^{\rho} = \mathbf{Y}_{(n)}$ to the Bayes strategy.

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| I | 0.2 | 0 | – | – | – |
| | | 0.2 | ++ | ++ | ++ |
| | | 0.4 | +++ | +++ | +++ |
| | 0.5 | 0 | 0 | – | – |
| | | 0.2 | ++ | ++ | ++ |
| | | 0.4 | +++ | +++ | +++ |
| | 0.8 | 0 | ++ | ++ | + |
| | | 0.2 | +++ | ++ | ++ |
| | | 0.4 | +++ | ++ | + |

Table 4.3: Summary comparison of the mWA mongrel averaging strategy with $(\mathbf{T}_n^{\alpha}, \mathbf{T}_n^{\rho}) = (\mathbf{S}_n^{\alpha}, \mathbf{S}_n^{\alpha})$ to the Bayes strategy.

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| **I** | 0.2 | 0 | ++ | + | + |
| | | 0.2 | + | 0 | − |
| | | 0.4 | − − | − − − | − − − |
| | 0.5 | 0 | +++ | ++ | + |
| | | 0.2 | ++ | + | 0 |
| | | 0.4 | - | - - | - - - |
| | 0.8 | 0 | +++ | ++ | ++ |
| | | 0.2 | +++ | ++ | + |
| | | 0.4 | ++ | - - | - - |

Table 4.4: Summary comparison of the mWA mongrel averaging strategy with $(\mathbf{T}_n^{\alpha}, \mathbf{T}_n^{\rho}) = (\mathbf{S}_n^{\alpha}, \mathbf{Y}_{(n)})$ to the Bayes strategy.

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| **I** | 0.2 | 0 | ++ | + | + |
| | | 0.2 | + | 0 | − |
| | | 0.4 | − − | − − − | − − − |
| | 0.5 | 0 | ++ | + | + |
| | | 0.2 | 0 | − | − − |
| | | 0.4 | − − | − − − | − − − |
| | 0.8 | 0 | − | − | 0 |
| | | 0.2 | − | − | − |
| | | 0.4 | − | − | − |

Table 4.5: Summary comparison of the BNM mongrel averaging strategy with $(\mathbf{T}_n^{\alpha}, \mathbf{T}_n^{\rho}) = (\mathbf{S}_n^{\alpha}, \mathbf{S}_n^{\alpha})$ to the Bayes strategy.

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| I | 0.2 | 0 | ++ | + | + |
| | | 0.2 | ++ | + | 0 |
| | | 0.4 | − − | − − − | − − − |
| | 0.5 | 0 | ++ | ++ | + |
| | | 0.2 | ++ | 0 | 0 |
| | | 0.4 | 0 | − − | − − − |
| | 0.8 | 0 | ++ | + | + |
| | | 0.2 | ++ | + | 0 |
| | | 0.4 | ++ | − | − |

Table 4.6: Summary comparison of the BNM mongrel averaging strategy with $(\mathbf{T}_n^{\alpha}, \mathbf{T}_n^{\rho}) = (\mathbf{S}_n^{\alpha}, \mathbf{Y}_{(n)})$ to the Bayes strategy.

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| I | 0.2 | 0 | ++ | + | + |
| | | 0.2 | ++ | + | 0 |
| | | 0.4 | − − | − − − | − − − |
| | 0.5 | 0 | + | + | + |
| | | 0.2 | 0 | − | 0 |
| | | 0.4 | − | − − | − − |
| | 0.8 | 0 | 0 | 0 | 0 |
| | | 0.2 | 0 | 0 | 0 |
| | | 0.4 | 0 | 0 | 0 |

Table 4.7: Summary comparison of the BFD mongrel averaging strategy with $(\mathbf{T}_n^{\alpha}, \mathbf{T}_n^{\rho}) = (\mathbf{S}_n^{\alpha}, \mathbf{S}_n^{\alpha})$ to the Bayes strategy.

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| **I** | 0.2 | 0 | + | + | + |
| | | 0.2 | 0 | − | − |
| | | 0.4 | − − − | − − − | − − − |
| | 0.5 | 0 | ++ | ++ | + |
| | | 0.2 | 0 | 0 | − |
| | | 0.4 | − − | − − − | − − − |
| | 0.8 | 0 | ++ | ++ | + |
| | | 0.2 | ++ | + | + |
| | | 0.4 | 0 | − − | − − |

Table 4.8: Summary comparison of the BFD mongrel averaging strategy with $(\mathbf{T}_n^{\alpha}, \mathbf{T}_n^{\rho}) = (\mathbf{S}_n^{\alpha}, \mathbf{Y}_{(n)})$ to the Bayes strategy.

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| **I** | 0.2 | 0 | + | + | + |
| | | 0.2 | 0 | − | − |
| | | 0.4 | − − − | − − − | − − − |
| | 0.5 | 0 | ++ | ++ | + |
| | | 0.2 | 0 | 0 | − |
| | | 0.4 | − − | − − − | − − − |
| | 0.8 | 0 | − | − | + |
| | | 0.2 | − − | − | − |
| | | 0.4 | − | − | 0 |

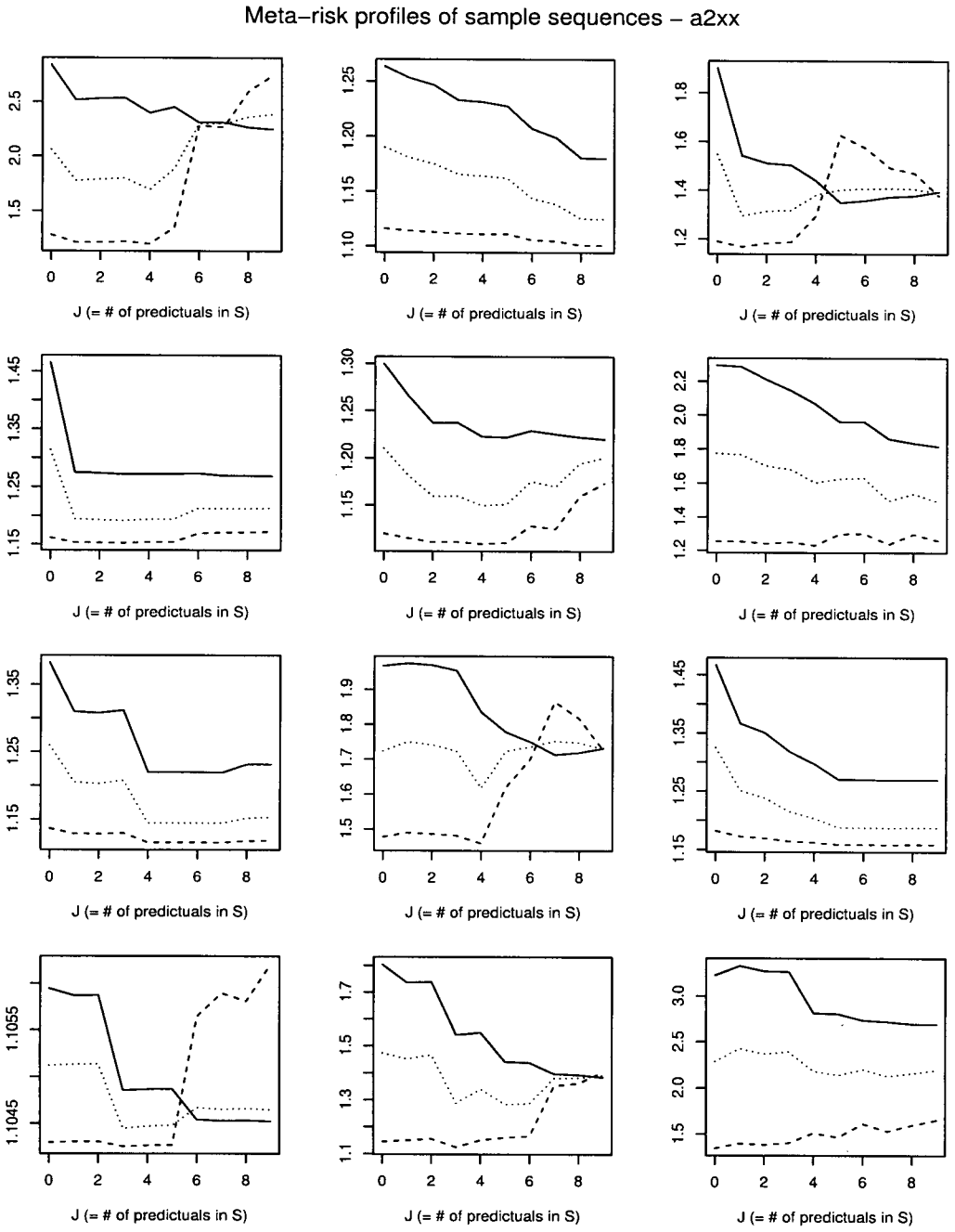


Figure 4.1: Meta-risk profiles for the first 12 sequences from an averaging strategy with $\mathbf{T}_n^{\alpha} = \mathbf{T}_n^{\rho} = \mathbf{S}_n^{\alpha}$: $a_{2,o} = 0.2$, $\gamma_2 = 0$. The solid (dashed) curve assumes the full (reduced) model is true. The dotted curve is a weighted average.

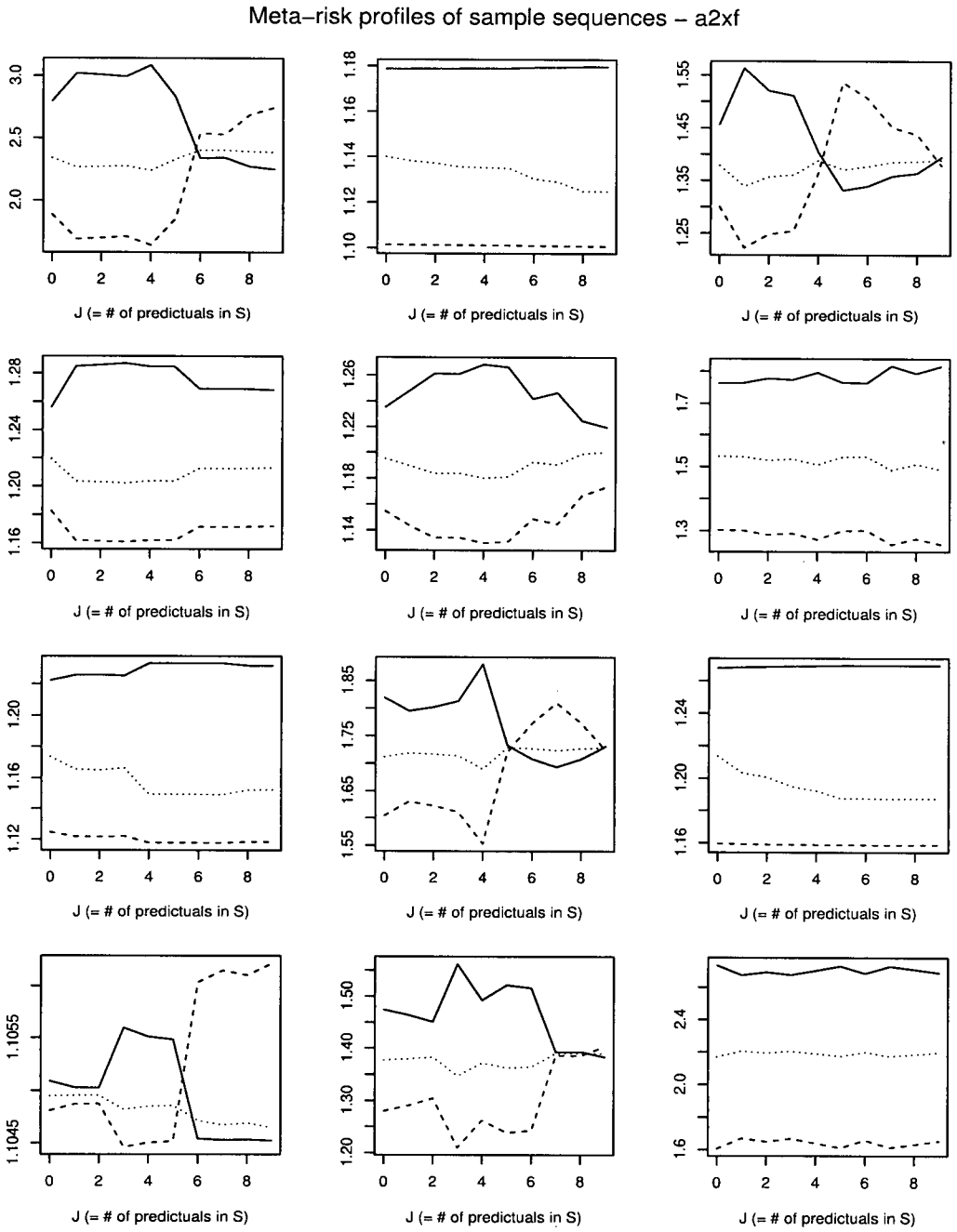


Figure 4.2: Meta-risk profiles for the first 12 sequences from an averaging strategy with $\mathbf{T}_n^{\alpha} = \mathbf{T}_n^{\rho} = \mathbf{S}_n^{\alpha}$: $a_{2,o} = 0.2$, $\gamma_2 = 0$. The solid (dashed) curve assumes the full (reduced) model is true. The dotted curve is a weighted average.

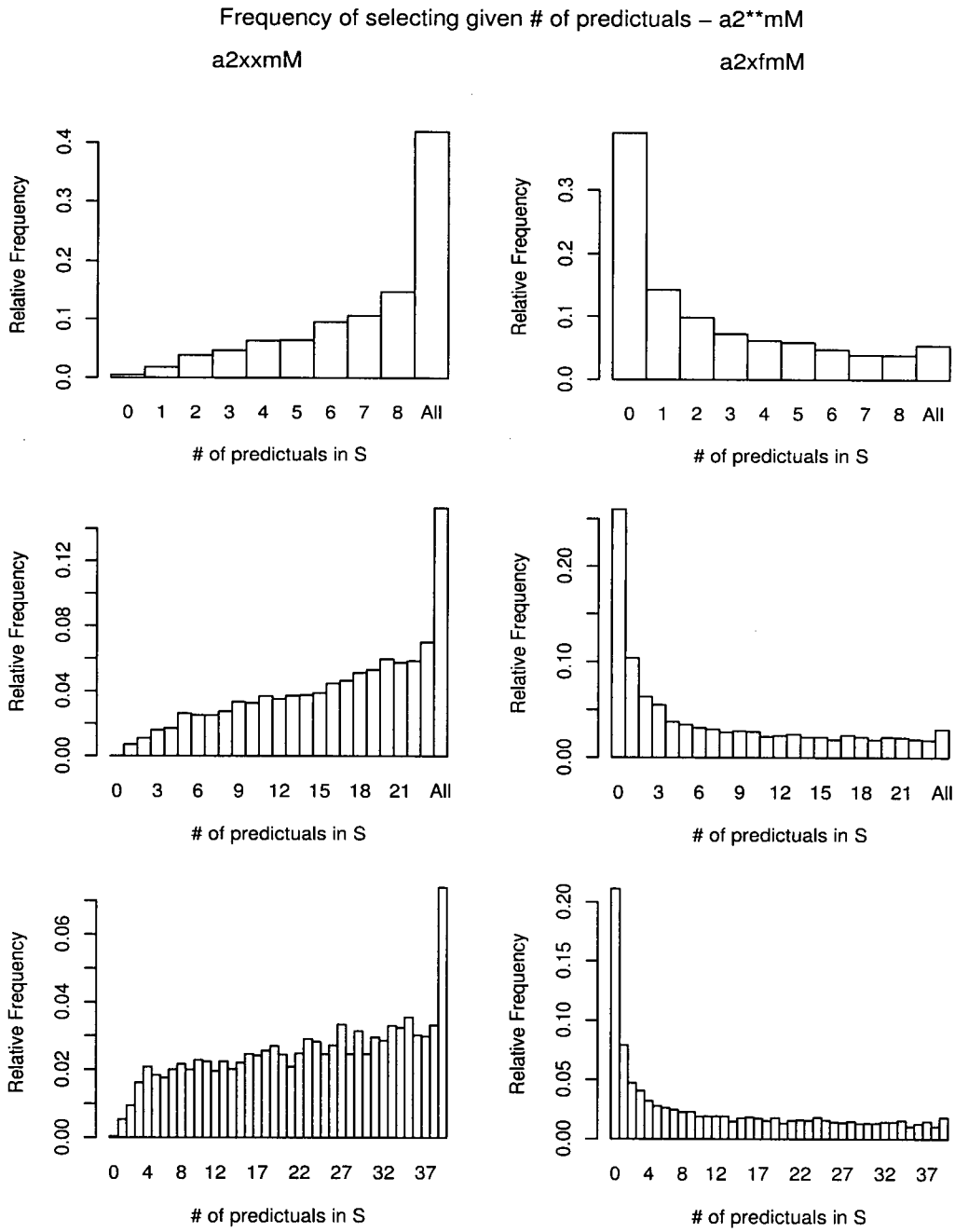


Figure 4.3: Histograms of the number of predictuals to include in $\mathbf{S}_n^\alpha$ that was selected by the minimax meta-risk procedure with $\mathbf{T}_n^\rho = \mathbf{S}_n^\alpha$: $a_{2,o} = 0.2$, $\gamma_2 = 0$.

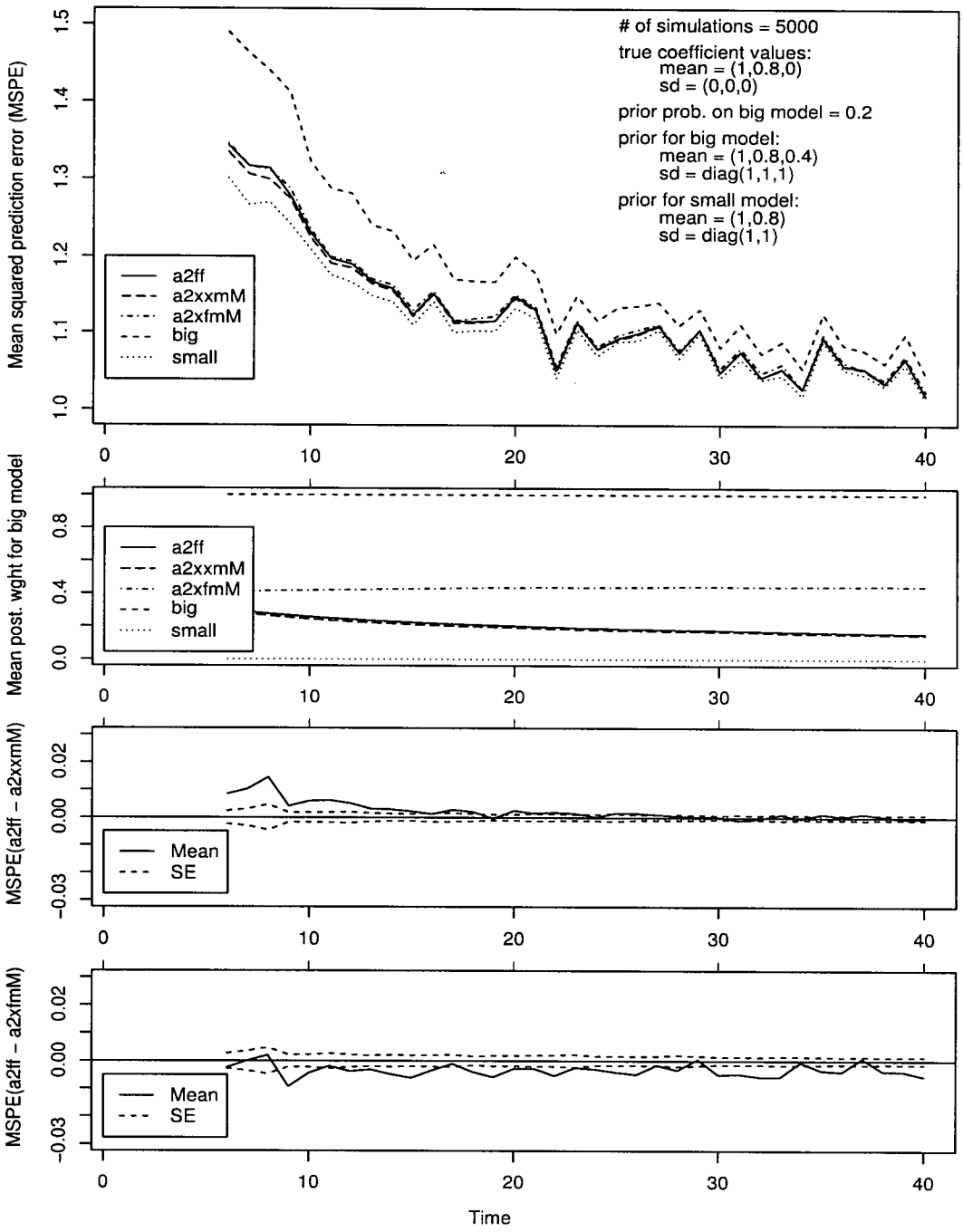


Figure 4.4: Performance of mM averaging strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0$.

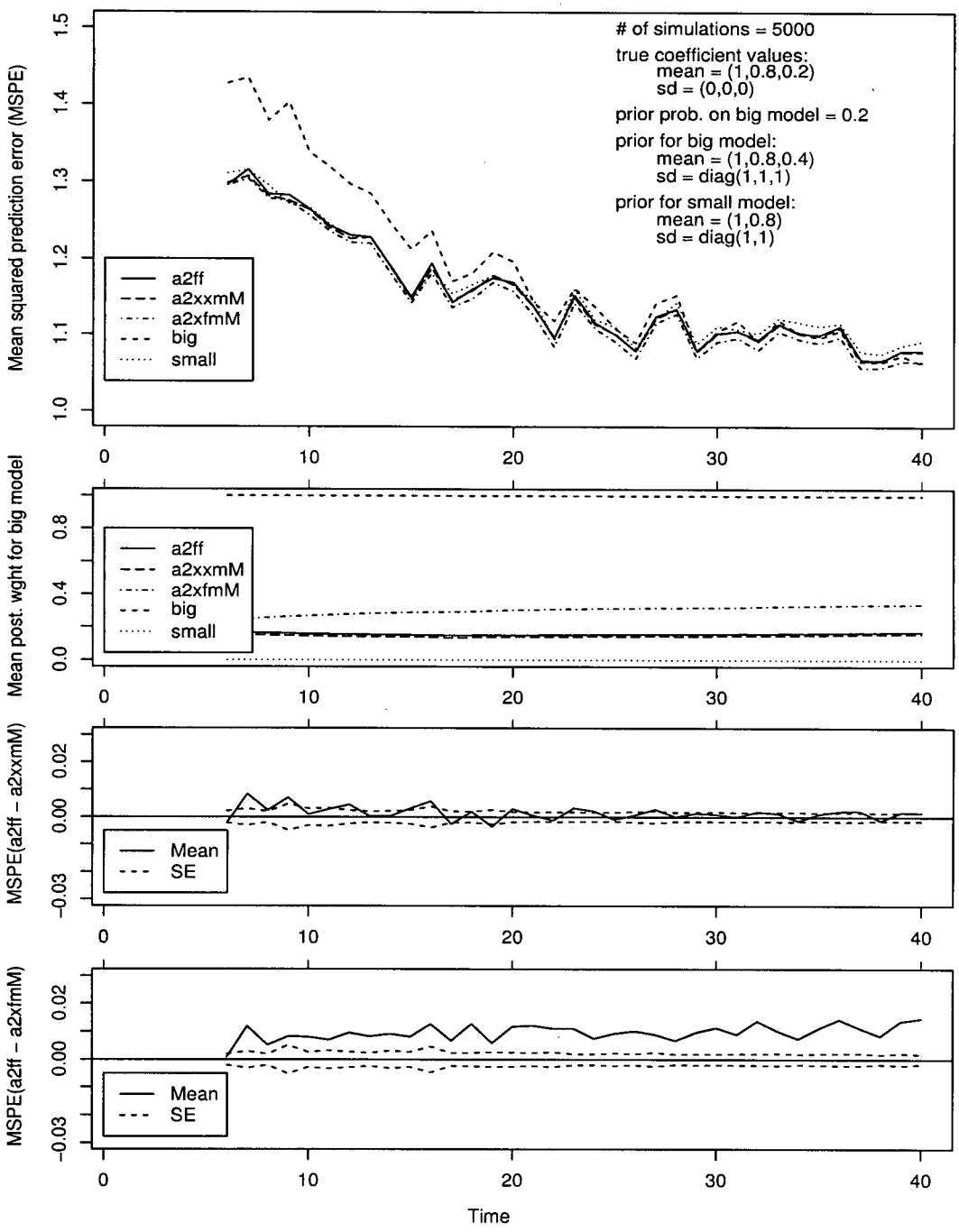


Figure 4.5: Performance of mM averaging strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0.2$.

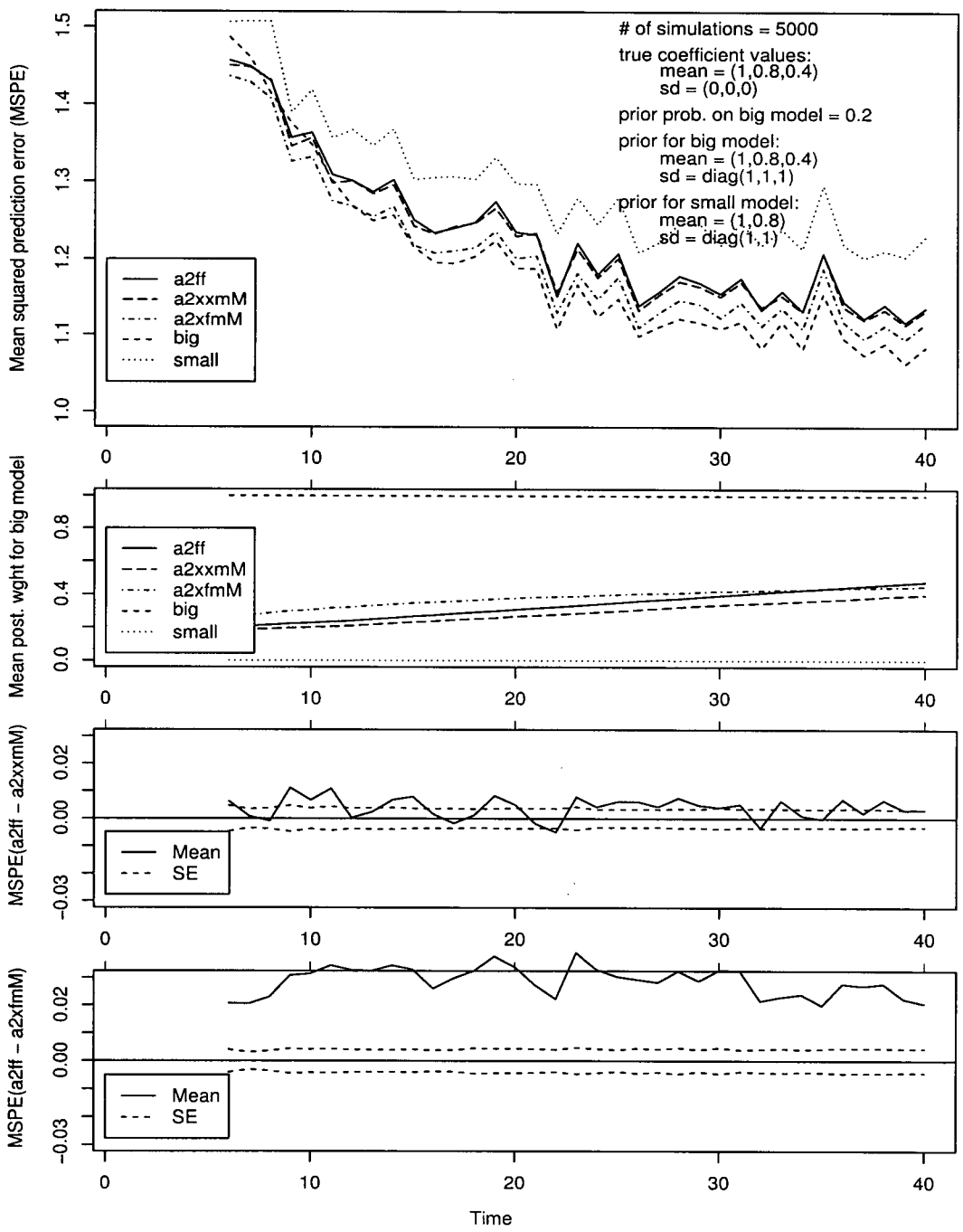


Figure 4.6: Performance of mM averaging strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0.4$.

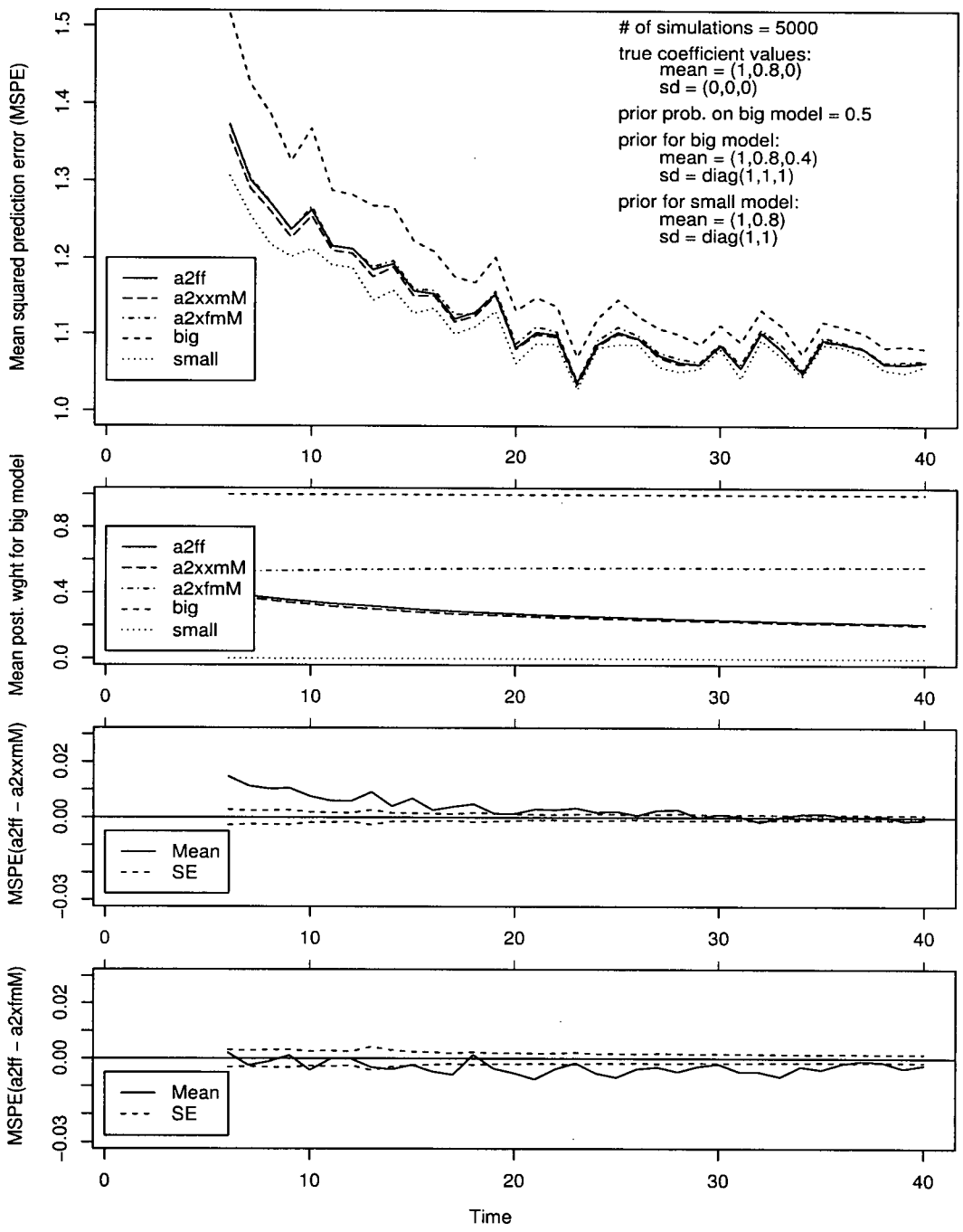


Figure 4.7: Performance of mM averaging strategies: $a_{2,o} = 0.5$, $\gamma_2 = 0$.

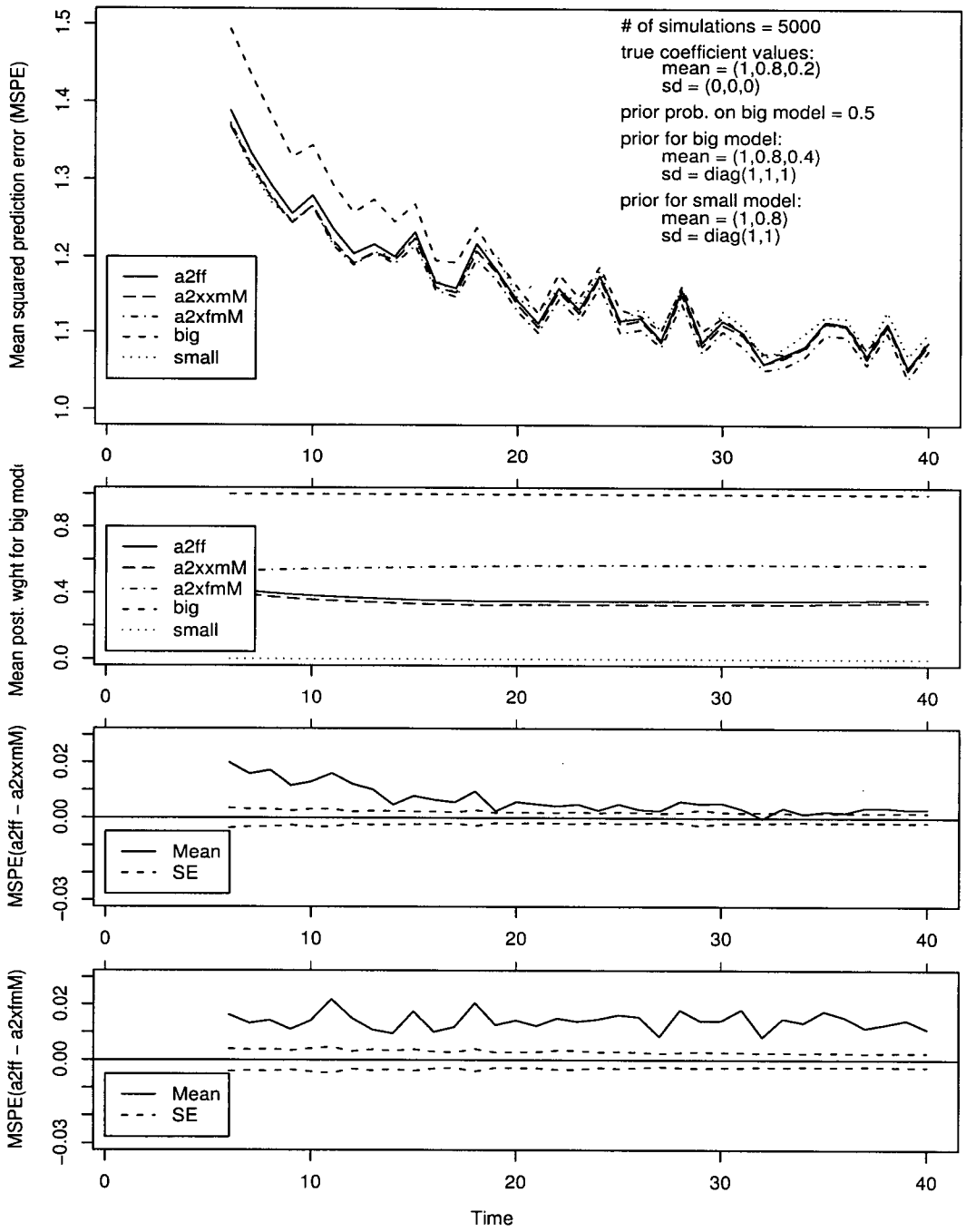


Figure 4.8: Performance of mM averaging strategies: $a_{2,o} = 0.5$, $\gamma_2 = 0.2$.

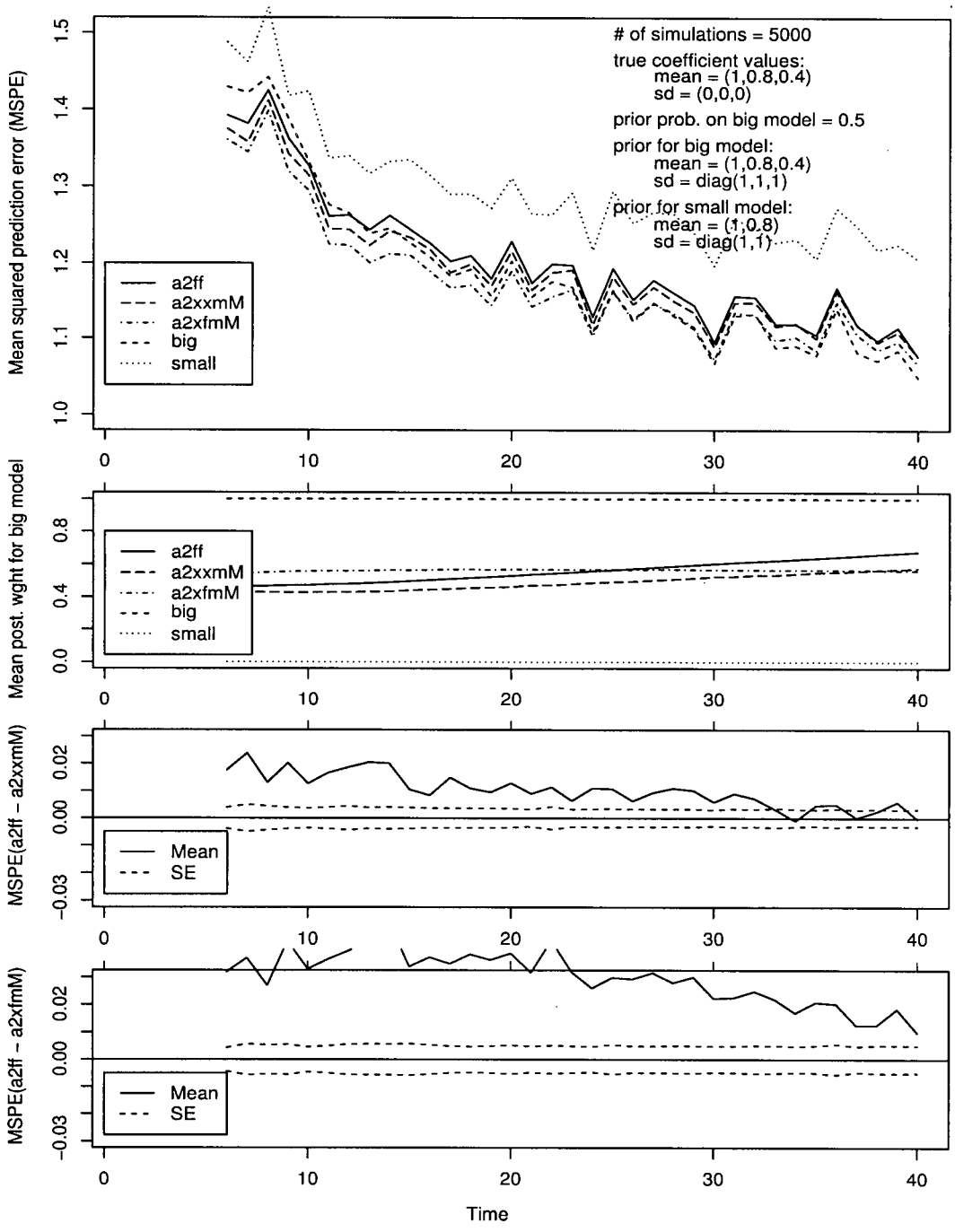


Figure 4.9: Performance of mM averaging strategies: $a_{2,o} = 0.5$, $\gamma_2 = 0.4$.

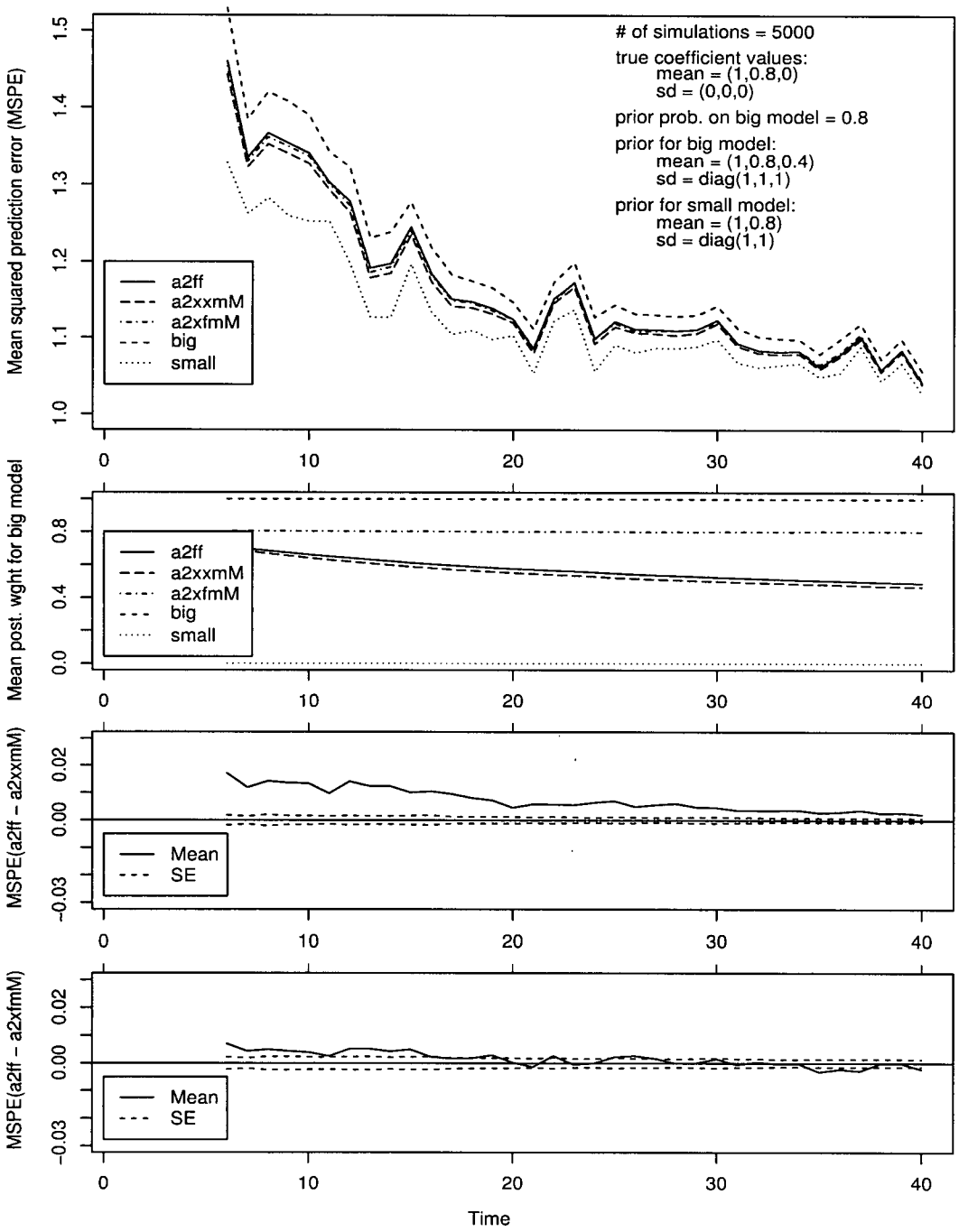


Figure 4.10: Performance of mM averaging strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0$.

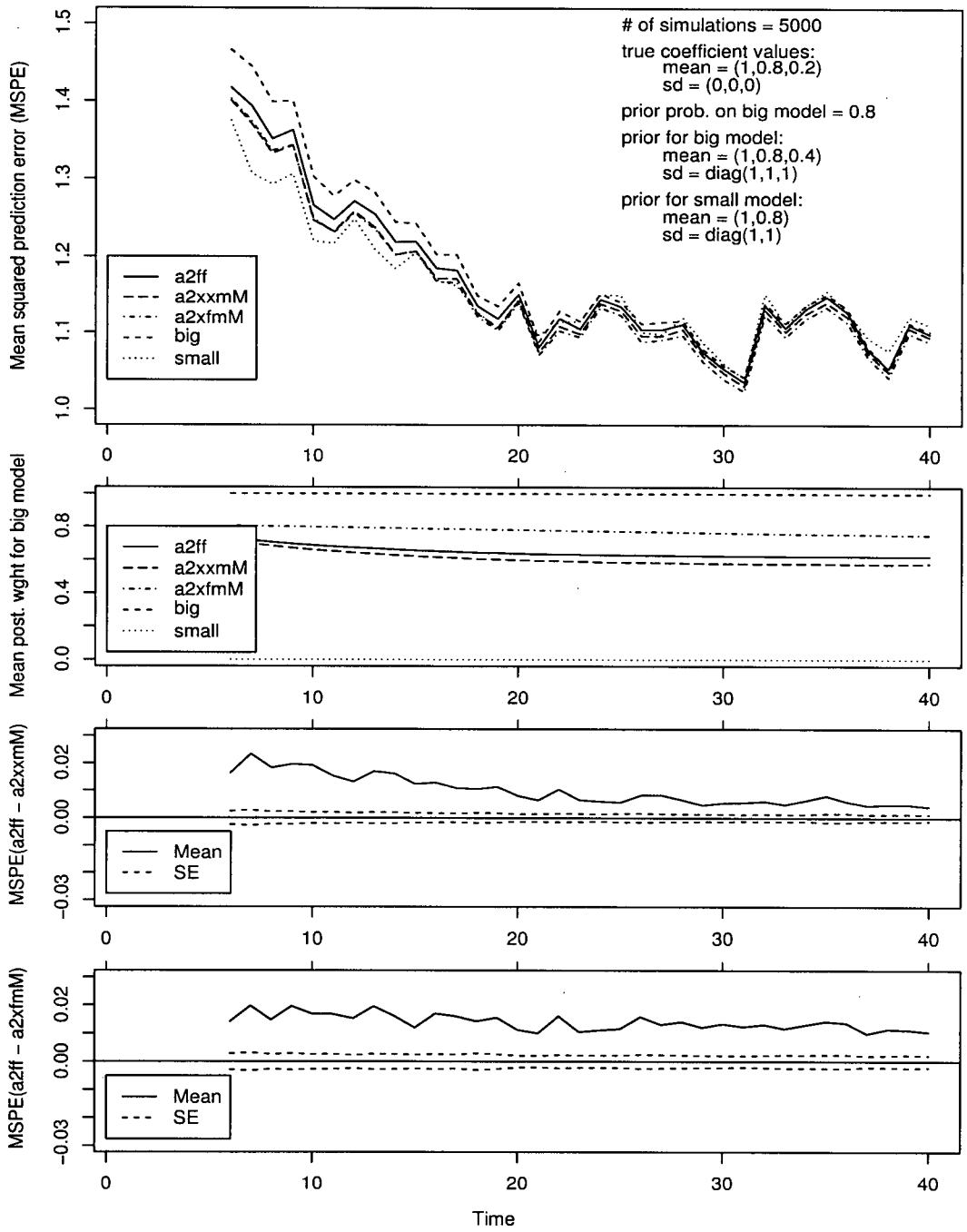


Figure 4.11: Performance of mM averaging strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0.2$.

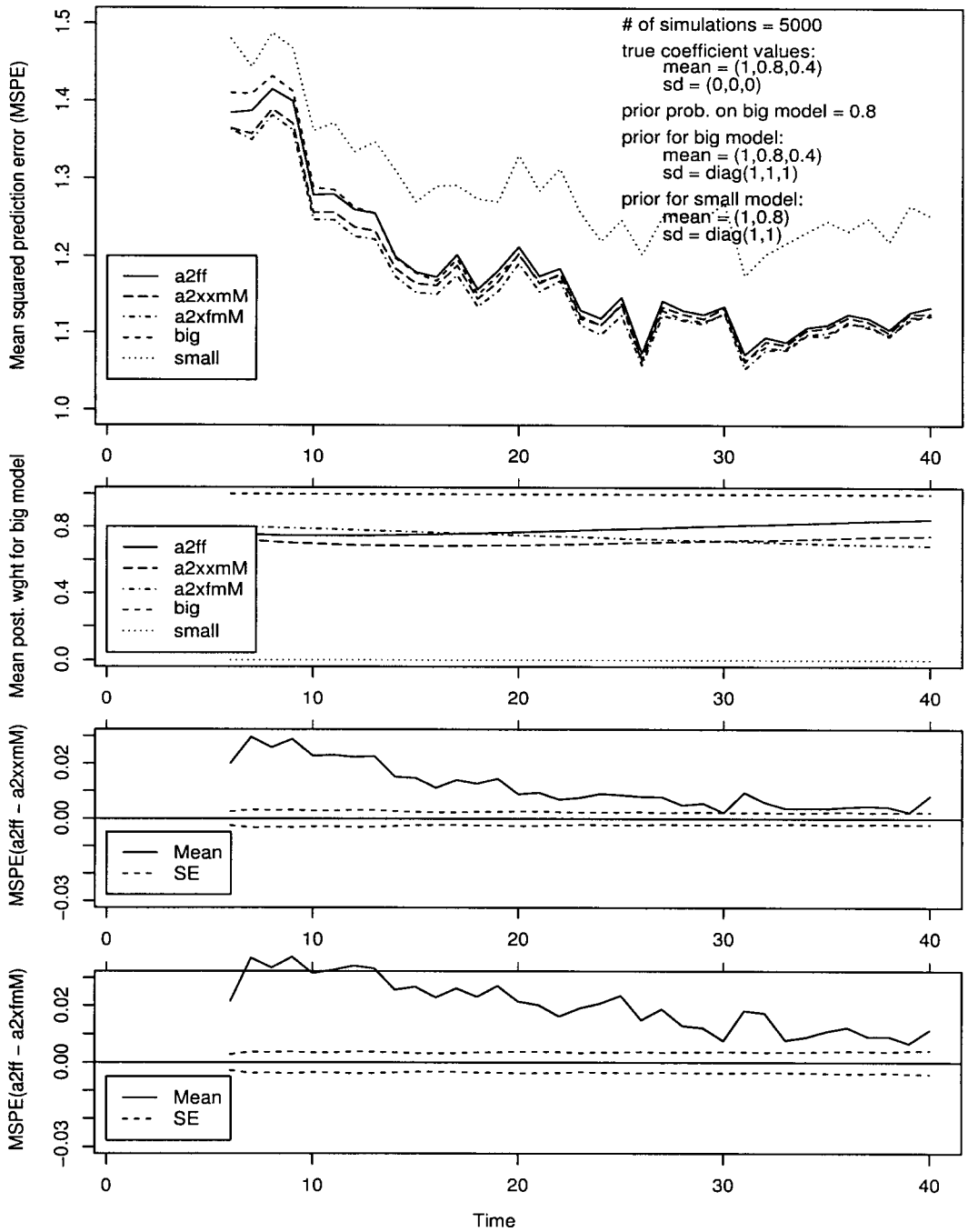


Figure 4.12: Performance of mM averaging strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0.4$.

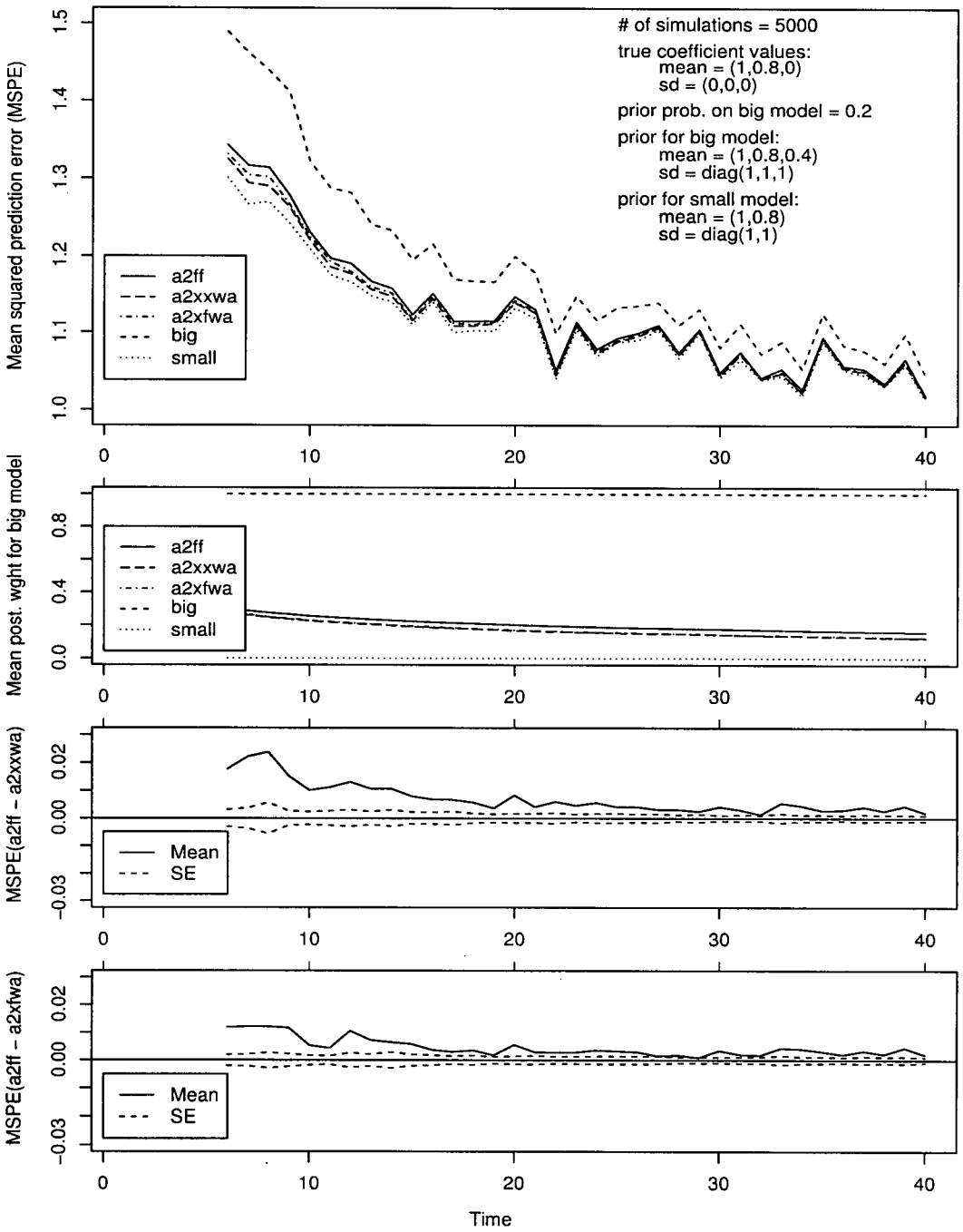


Figure 4.13: Performance of mWA averaging strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0$.

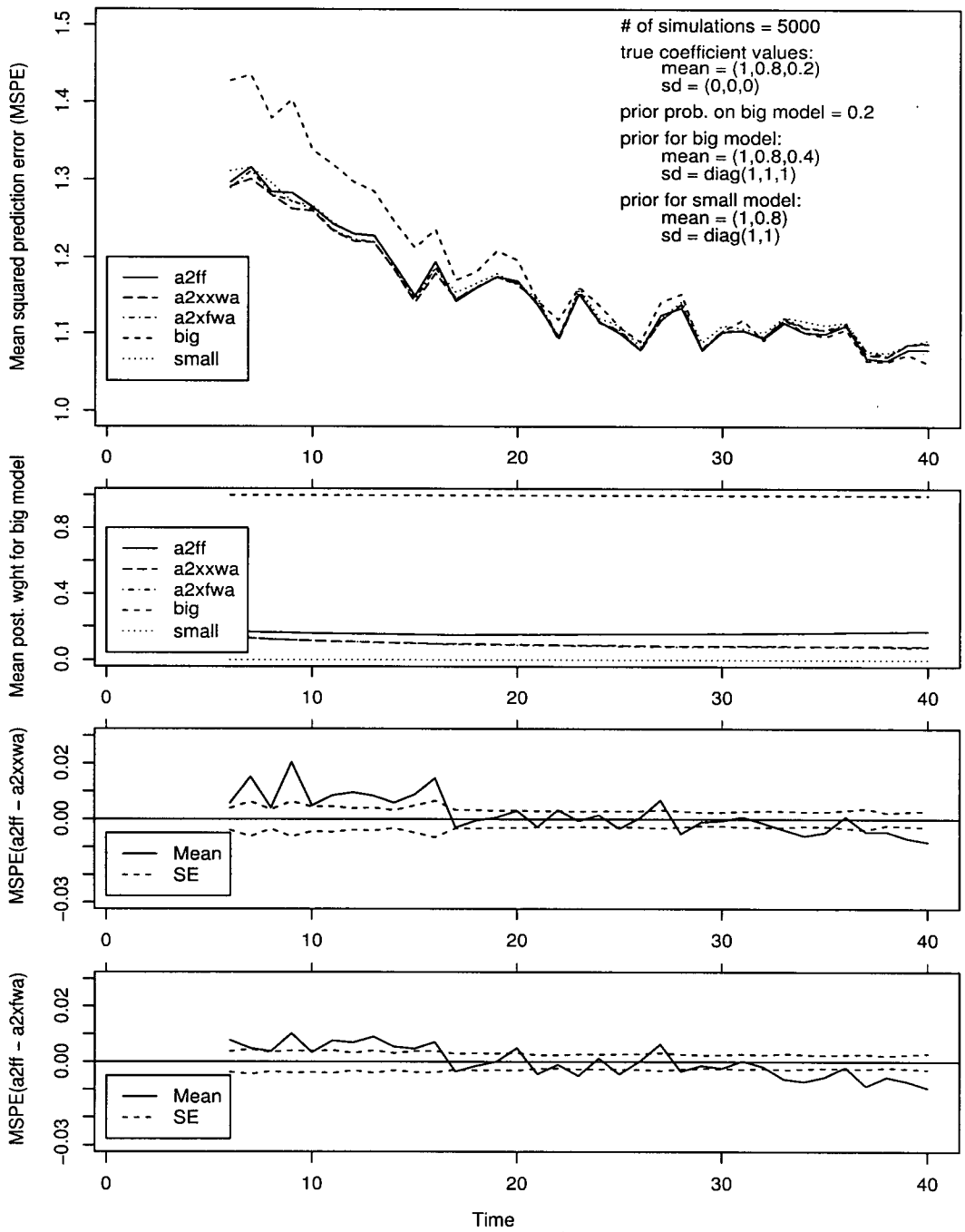


Figure 4.14: Performance of mWA averaging strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0.2$.

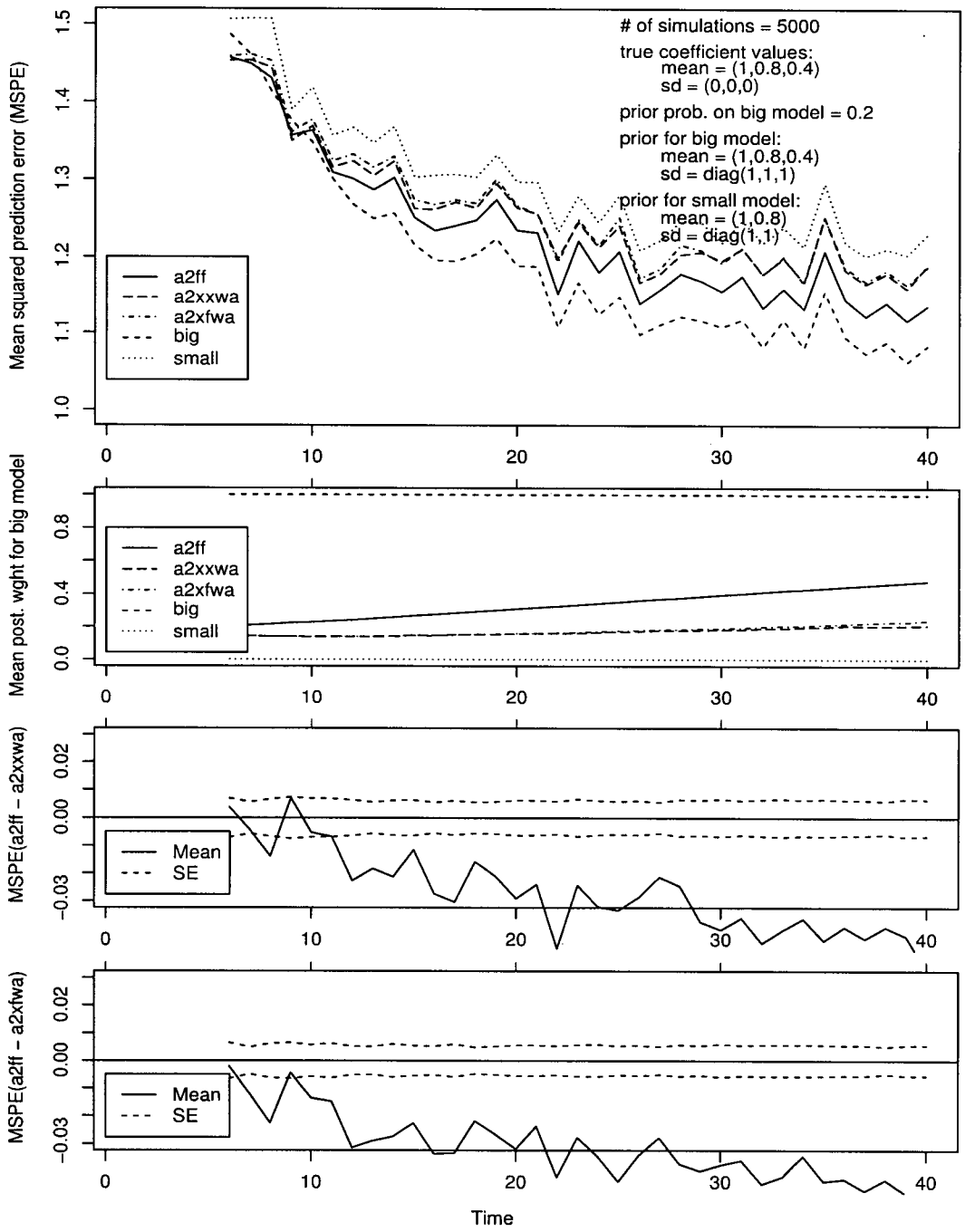


Figure 4.15: Performance of mWA averaging strategies: $a_{2,o} = 0.2$, $\gamma_2 = 0.4$.

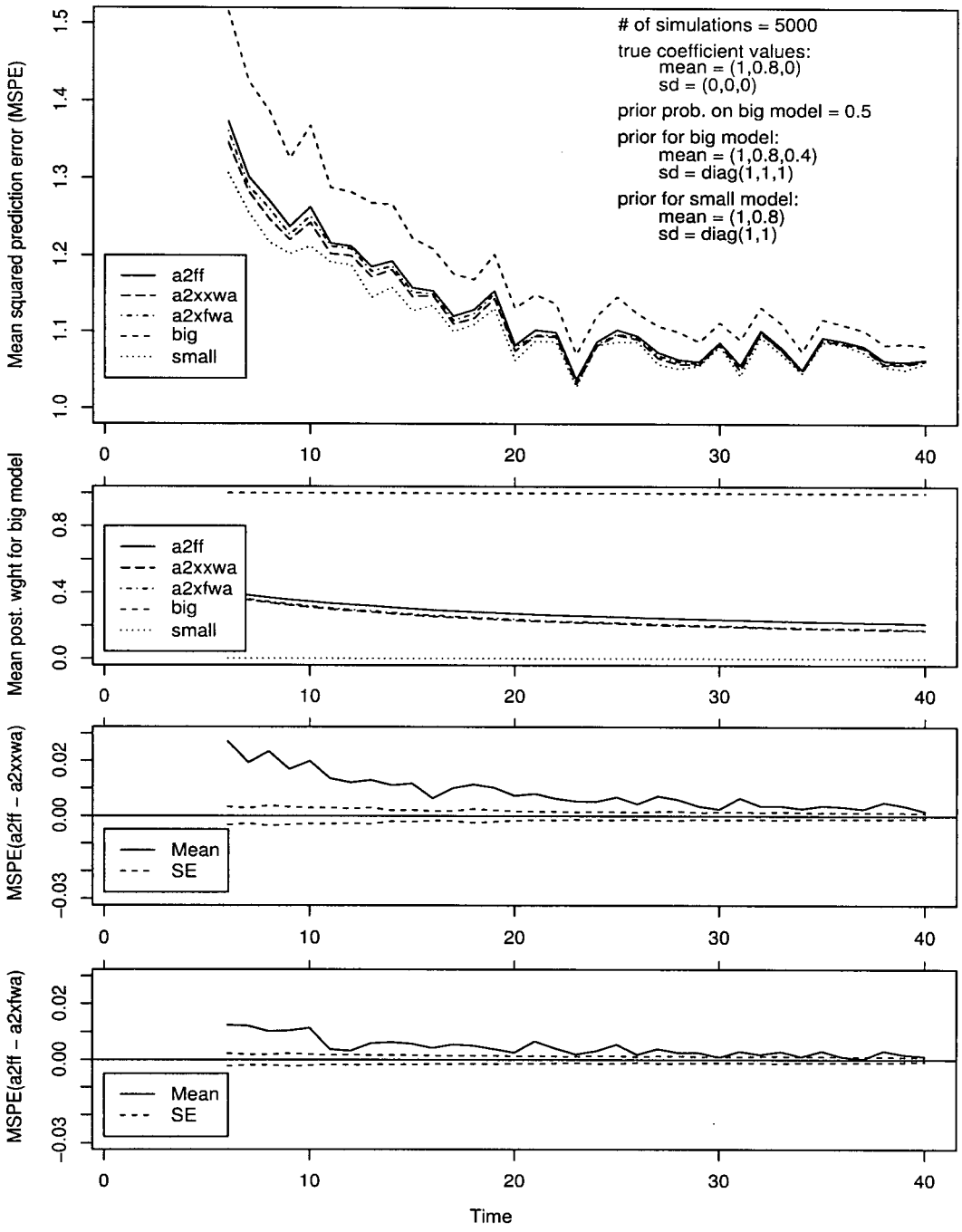


Figure 4.16: Performance of mWA averaging strategies: $a_{2,o} = 0.5$, $\gamma_2 = 0$.

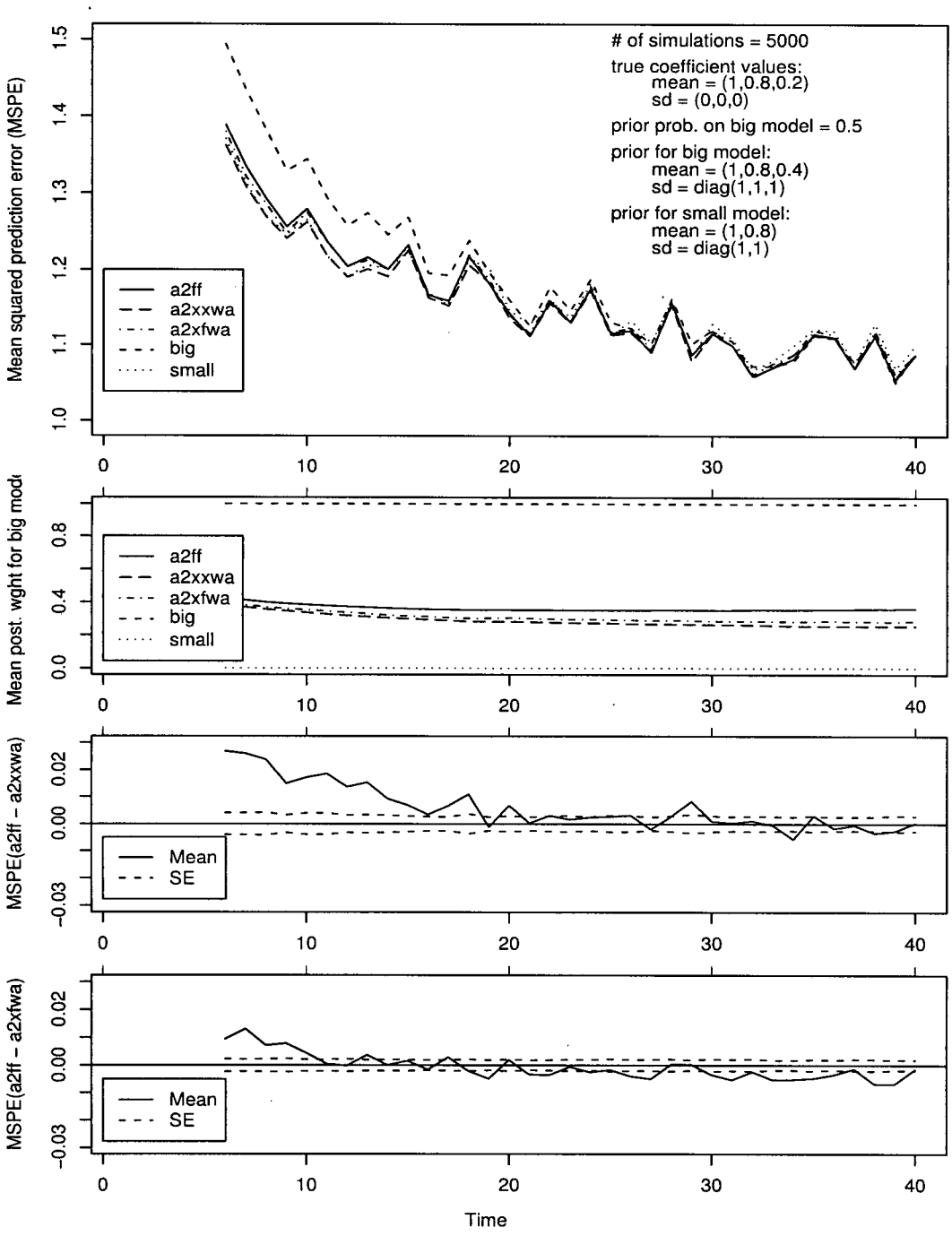


Figure 4.17: Performance of mWA averaging strategies: $a_{2,o} = 0.5$, $\gamma_2 = 0.2$.

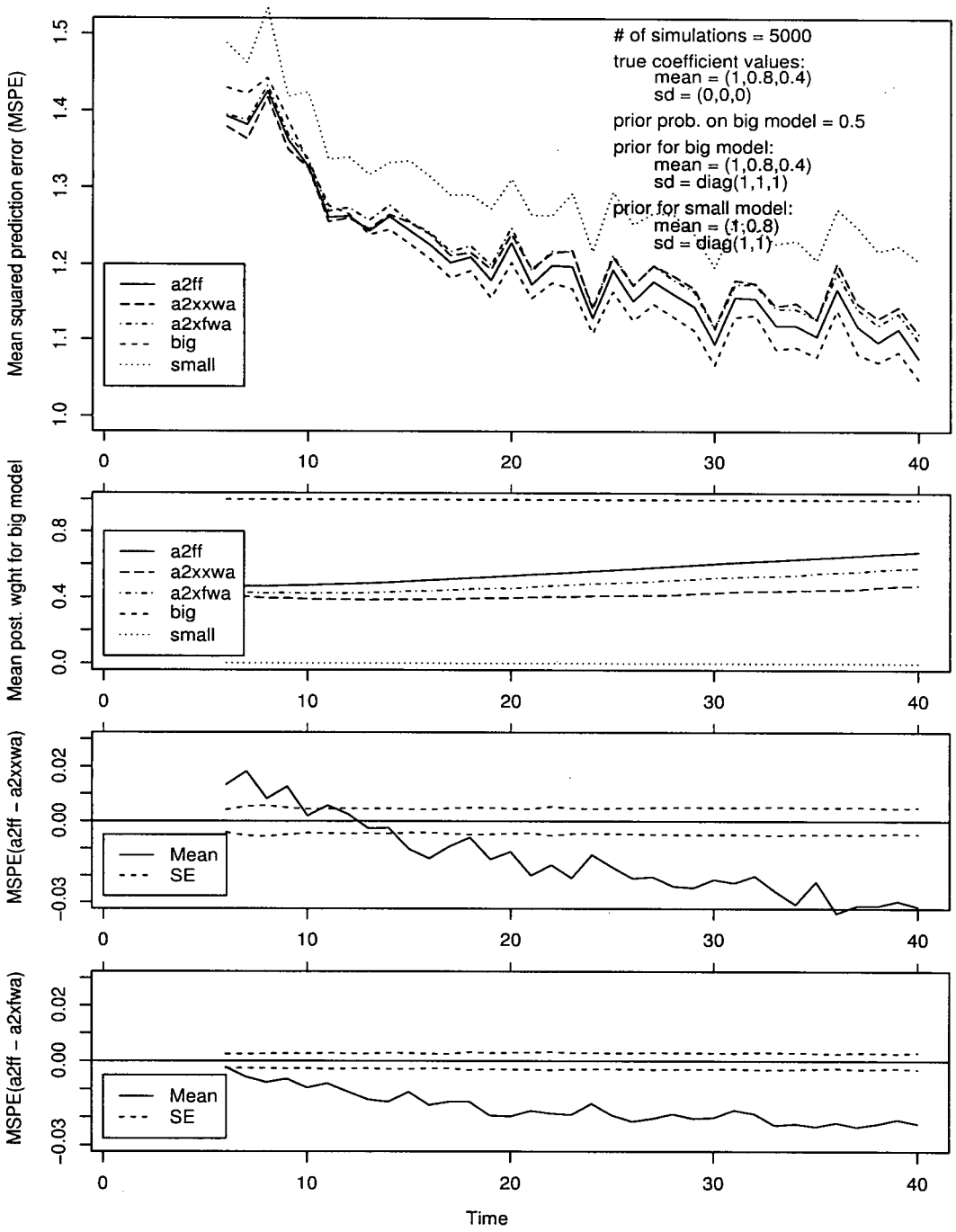


Figure 4.18: Performance of mWA averaging strategies: $a_{2,o} = 0.5$, $\gamma_2 = 0.4$.

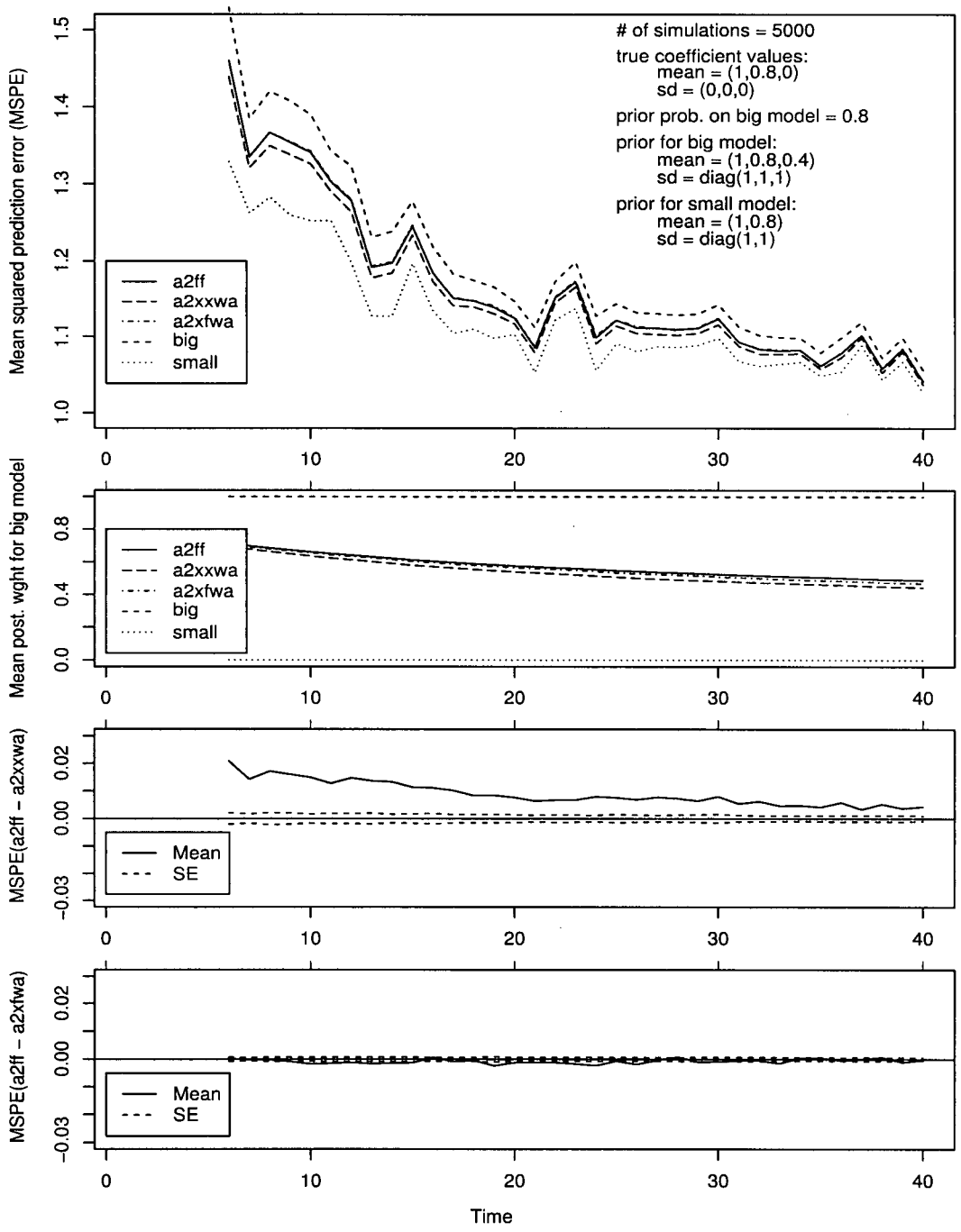


Figure 4.19: Performance of mWA averaging strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0$.

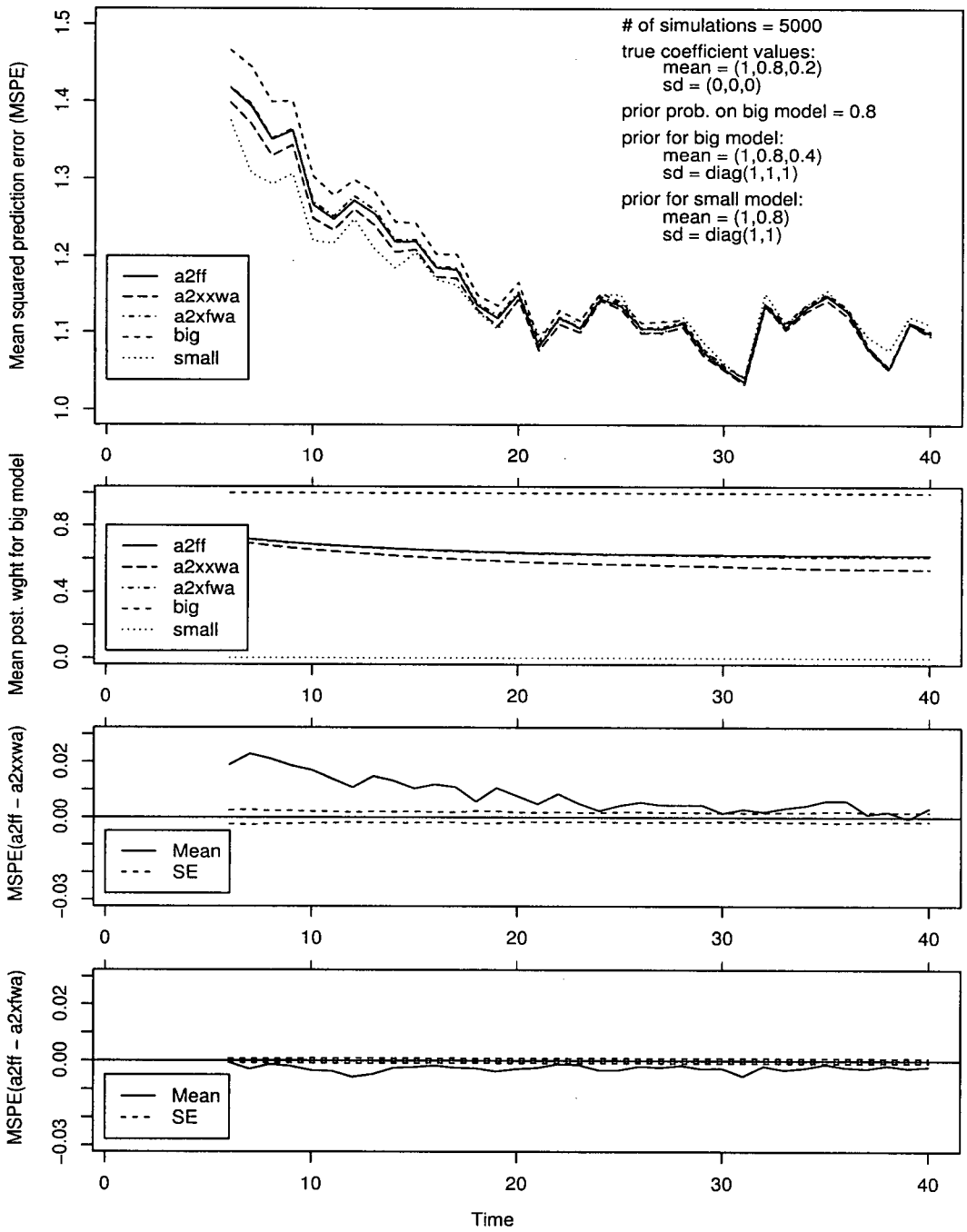


Figure 4.20: Performance of mWA averaging strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0.2$.

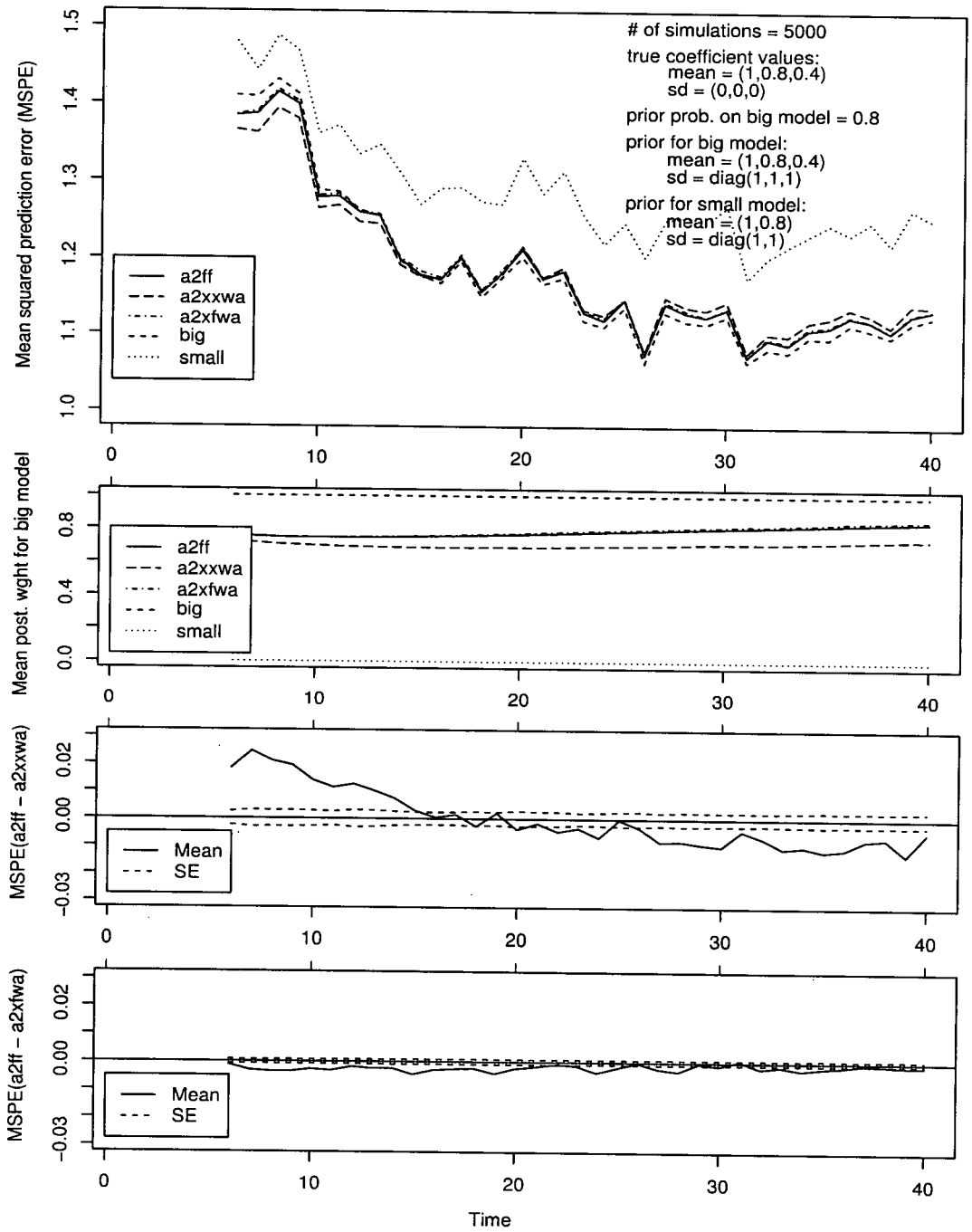


Figure 4.21: Performance of mWA averaging strategies: $a_{2,o} = 0.8$, $\gamma_2 = 0.4$.

# Chapter 5

# Finite Samples: Global Selection

The specification of a global selection of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ can be approached in a variety of ways. The basic principle that we assume here is that we hope to capture sufficient information to obtain an accurate assessment of risk but simultaneously we do not want the the assessment to be responsive to "bad data".

For example, consider the selection of $\mathbf{S}_n^\rho$. Intuitively, as the size of the sigma field $\sigma(\mathbf{S}_n^\rho)$ increases, the risk assessment becomes evermore sensitive to the data. Treating $\rho_i(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho)$ as a random variable and using $\mathbf{V}_i\rho_i(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho)$ as a measure of the sensitivity, the following result formalizes this intuition.

**Lemma 5.1** *Let $\mathbf{S}_n^\rho$, $\mathbf{T}_n^\rho$ be two statistics such that $\mathbf{T}_n^\rho = g(\mathbf{S}_n^\rho)$, or equivalently, $\sigma(\mathbf{T}_n^\rho) \subseteq \sigma(\mathbf{S}_n^\rho)$. Then $\rho_i(\hat{Y}_{k,n+1}; \mathbf{T}_n^\rho)$ is not more sensitive to the data than $\rho_i(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho)$, that is,*

$$\mathbf{V}_i\rho_i(\hat{Y}_{k,n+1}; \mathbf{T}_n^\rho) \leq \mathbf{V}_i\rho_i(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho). \tag{5.1}$$

**Proof** Iterating in the definition of $\rho_i$ we see

$$\begin{aligned}
\mathbf{V}_i\rho_i(\hat{Y}_{k,n+1};\mathbf{T}_n^\rho) &= \mathbf{V}_i\mathbf{E}_{i|\mathbf{T}_n^\rho}L(Y_{n+1},\hat{Y}_{k,n+1}) \\
&= \mathbf{V}_i\mathbf{E}_{i|\mathbf{T}_n^\rho}\mathbf{E}_{i|\mathbf{S}_n^\rho}L(Y_{n+1},\hat{Y}_{k,n+1}) \\
&= \mathbf{V}_i\mathbf{E}_{i|\mathbf{T}_n^\rho}\rho_i(\hat{Y}_{k,n+1};\mathbf{S}_n^\rho) \\
&\leq \mathbf{V}_i\rho_i(\hat{Y}_{k,n+1};\mathbf{S}_n^\rho)
\end{aligned}$$

since $\mathbf{V}_i\rho_i(\hat{Y}_{k,n+1};\mathbf{S}_n^\rho) = \mathbf{V}_i\mathbf{E}_{i|\mathbf{T}_n^\rho}\rho_i(\hat{Y}_{k,n+1};\mathbf{S}_n^\rho) + \mathbf{E}_i\mathbf{V}_{i|\mathbf{T}_n^\rho}\rho_i(\hat{Y}_{k,n+1};\mathbf{S}_n^\rho)$. ∎

This result suggests that the choice $\mathbf{S}_n^\rho = \mathbf{Y}_{(n)}$ yields a mongrel risk that is maximally sensitive to the data, while the choice $\mathbf{S}_n^\rho = 0$ yields a minimally sensitive mongrel risk. In the case where $\mathbf{S}_n^\rho$ is a collection of past predicted predictuals, the inclusion of more predictuals increases sensitivity.

## 5.1 Closeness to the Bayes solution

One type of robustness is that the risk criterion should be insensitive to incorrectly specified models. We illustrate for the case of model choice with two candidate models. In choosing between the two models, we base our decision on the difference in our assessments of the risk of the predictors from the two models. The essential idea is that we would like this difference to be "close" to the true difference in risk irrespective of the data-generator. But because the data-generator is unknown, we use instead the posterior distributions from the Bayes procedure as surrogates.

For two models indexed by $k$ and $k'$, we set the target for the difference in risks between the two models to be

$$\delta_k \equiv \rho_k(\hat{Y}_{k,n+1};\mathbf{Y}_{(n)}) - \rho_k(\hat{Y}_{k',n+1};\mathbf{Y}_{(n)}) \tag{5.2}$$

when model $k$ is true, or

$$\delta_{k'} \equiv \rho_{k'}(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) - \rho_{k'}(\hat{Y}_{k',n+1}; \mathbf{Y}_{(n)}) \tag{5.3}$$

when model $k'$ is true. Meanwhile, we select $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ based on our assessment of the difference in average mongrel risk

$$\bar{\delta} \equiv \bar{\rho}(\hat{Y}_{k,n+1}; \mathbf{S}_n^\alpha, \mathbf{S}_n^\rho) - \bar{\rho}(\hat{Y}_{k',n+1}; \mathbf{S}_n^\alpha, \mathbf{S}_n^\rho). \tag{5.4}$$

Ideally, we would like to select $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ such that $\bar{\delta}$ to be close to both $\delta_k$ and $\delta_{k'}$.

If we are willing to weight the models according to numbers $w_k$ and $w_{k'}$, then we might try selecting $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ to minimize, say,

$$w_k(\bar{\delta} - \delta_k)^2 + w_{k'}(\bar{\delta} - \delta_{k'})^2 \tag{5.5}$$

(pointwise in $\mathbf{Y}_{(n)}$), or,

$$\tilde{\rho} \equiv w_k \mathbf{E}_k(\bar{\delta} - \delta_k)^2 + w_{k'} \mathbf{E}_{k'}(\bar{\delta} - \delta_{k'})^2. \tag{5.6}$$

(The obvious choice for the weights would be $w_i = \alpha_i(\mathbf{y}_{(n)})$.) It can be shown that the minimum for (5.5) can be obtained by setting $\mathbf{S}_n^\alpha = \mathbf{S}_n^\rho = \mathbf{Y}_{(n)}$ so this criterion is not useful for selecting $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. The criterion (5.6) can be thought of as a robustness criterion. We are trying to select the $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ that keeps $\bar{\delta}$ close to both $\delta_k$ and $\delta_{k'}$ (in a weighted average sense) regardless of the value of $\mathbf{Y}_{(n)}$ that obtains. Unfortunately, the evaluation of the expectations in $\tilde{\rho}$ are not tractable analytically and so the criterion is not easily implemented. We attempt to circumvent this problem by finding an approximation $\hat{\rho}$ to $\tilde{\rho}$ on which we base our selection of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$.

### 5.1.1 Approximating (5.6)

Note that both $\bar{\delta} - \delta_k$ and $\bar{\delta} - \delta_{k'}$ is a weighted sum of six mongrel risk terms which we can write as, for $\bar{\delta} - \delta_k$ say,

$$\bar{\delta} - \delta_k = \sum_{j=1}^{6} \kappa_j r_j \tag{5.7}$$

where $\kappa_j \in \{\pm 1, \pm\alpha_k(\mathbf{S}_n^\alpha), \pm\alpha_{k'}(\mathbf{S}_n^\alpha)\}$ is a weight and $r_j$ is a mongrel risk $\rho_i(\cdot;\cdot)$. In the normal case, $r_j$ can be evaluated using (3.12) through (3.16) and these yield expressions that are quadratic in $\mathbf{S}_n^\rho$. By substituting for $\mathbf{S}_n^\rho$ by its function of $\mathbf{Y}_{(n)}$ these expressions can be written in the form

$$r_j = \zeta_j + (\gamma_j + \delta_j^T \mathbf{Y}_{(n)})^2 \tag{5.8}$$

where $\zeta_j$ and $\gamma_j$ are numbers and $\delta_j$ is a vector.

**Lemma 5.2** *Suppose the weights $\alpha_i(\mathbf{S}_n^\alpha)$ in (5.7) are replaced by their observed values $\alpha_i(\mathbf{s}_n^\alpha)$. Then an approximation to $\mathbf{E}_k(\bar{\delta} - \delta_k)^2$ is given by*

$$\begin{aligned} \hat{E}_k \equiv\ & 2tr(\mathbf{A}\Psi_k\mathbf{A}\Psi_k) + 4\|\nu_k\|_{\mathbf{A}}^2 + 8\nu_k^T\mathbf{A}\Psi_k\xi + 4\|\xi\|_{\Psi_k}^2 \\ & + \left(tr(\mathbf{A}\Psi_k) + \|\nu_k\|_{\mathbf{A}}^2 + 2\nu_k^T\xi + \sum_j \kappa_j(\zeta_j + \gamma_j^2)\right)^2 \end{aligned} \tag{5.9}$$

*where $\xi = \sum_j^6 \gamma_j\kappa_j\delta_j$, $\mathbf{A} = \Delta\Delta^T$, and $\Delta$ is the matrix with columns $\sqrt{\kappa_j}\delta_j$.*

**Proof** Treating the $\kappa_j$ as numbers instead of random variables and applying (A7), we have

$$\sum_{j=1}^{r} \kappa_j r_j = (\mathbf{Y} + \mathbf{b})^T \mathbf{A} (\mathbf{Y} + \mathbf{b}) + c \tag{5.10}$$

where $\mathbf{b}$ and $c$ are defined in (A9) and (A10). Applying (A4) and (A5), we obtain

$$\mathbf{E}_k\left(\sum_{j=1}^{r} \kappa_j r_j\right)^2 = \mathbf{V}_k\left(\sum_{j=1}^{r} \kappa_j r_j\right) + \left(\mathbf{E}_k\left(\sum_{j=1}^{r} \kappa_j r_j\right)\right)^2$$

$$= 2tr(\mathbf{A}\Psi_k\mathbf{A}\Psi_k) + 4\|\nu_k + \mathbf{b}\|^2_{\mathbf{A}\Psi_k\mathbf{A}} + \Big(tr(\mathbf{A}\Psi_k) + \|\nu_k + \mathbf{b}\|^2_{\mathbf{A}} + c\Big)^2.$$

The lemma follows from expanding the terms and applying the relation $\mathbf{Ab} = \xi$. ∎

Thus, we can approximate $\tilde{\rho}$ by

$$\hat{\rho} \equiv w_k\hat{E}_k + w_{k'}\hat{E}_{k'}, \tag{5.11}$$

and then the optimal choice for $(\mathbf{S}^\alpha_n, \mathbf{S}^\rho_n)$ is

$$(\mathbf{S}^\alpha_n, \mathbf{S}^\rho_n)^* = \arg\min_{(\mathbf{S}^\alpha_n, \mathbf{S}^\rho_n)} \hat{\rho}. \tag{5.12}$$

The computations in Lemma 5.2 can be can be avoided in special cases such as the following:

**Lemma 5.3** *Suppose* $\mathbf{S}^\rho_n = \mathbf{Y}_{(n)}$ *and that the weights* $\alpha_i(\mathbf{S}^\alpha_n)$ *in (5.7) are replaced by their observed values* $\alpha_i(\mathbf{s}^\alpha_n)$. *Then the optimal choice for* $\mathbf{S}^\alpha_n$ *minimizes*

$$\hat{\rho} = w_k\alpha^2_{k'}(\mathbf{s}^\alpha_n)\mathbf{E}_kB^4 + w_{k'}\alpha^2_k(\mathbf{s}^\alpha_n)\mathbf{E}_{k'}B^4 \tag{5.13}$$

*where* $B^2 = (\hat{Y}_{k,n+1} - \hat{Y}_{k',n+1})^2$.

**Proof** Expanding $\bar{\delta} - \delta_k$ using the definition of $\bar{\rho}$ with $\mathbf{S}^\rho_n = \mathbf{Y}_{(n)}$, we get

$$\begin{aligned}
\bar{\delta} - \delta_k &= \alpha_k\rho_k(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) + \alpha_{k'}\rho_{k'}(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) \\
&\quad -(\alpha_k\rho_k(\hat{Y}_{k',n+1}; \mathbf{Y}_{(n)}) + \alpha_{k'}\rho_{k'}(\hat{Y}_{k',n+1}; \mathbf{Y}_{(n)})) \\
&\quad +\rho_k(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) - \rho_k(\hat{Y}_{k',n+1}; \mathbf{Y}_{(n)}) \\
&= \alpha_k\rho_k(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) + \alpha_{k'}(\rho_{k'}(\hat{Y}_{k',n+1}; \mathbf{Y}_{(n)}) + B^2) \\
&\quad -(\alpha_k\rho_k(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) + B^2) - \alpha_{k'}\rho_{k'}(\hat{Y}_{k',n+1}; \mathbf{Y}_{(n)}) \\
&\quad +\rho_k(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) - (\rho_k(\hat{Y}_{k,n+1}; \mathbf{Y}_{(n)}) + B^2) \\
&= 2\alpha_{k'}B^2
\end{aligned} \tag{5.14}$$

where the second equality follows from the property $\rho_k(\hat{Y}_{k',n+1}) = \rho_k(\hat{Y}_{k,n+1}) + (\hat{Y}_{k,n+1} - \hat{Y}_{k',n+1})^2$. Similarly, $\bar{\delta} - \delta_{k'} = -2\alpha_k B^2$. The Lemma then follows by treating the weights $\alpha_k$, $\alpha_{k'}$ as numbers so that they can be taken outside the expectations in (5.6). ∎

### 5.1.2 Simulation Results

We applied the approximation provided by (5.9) to optimize the choice for $J_g$, the number of predictuals to be included in $\mathbf{S}_n^\rho$. (The choice $J_g = 0$ was omitted in order to avoid certain additional computations not readily available.) We set $\mathbf{S}_n^\alpha = \mathbf{Y}_{(n)}$. We call this method of choosing $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ the "ROB" strategy. (Note that the optimization is over choices of $\mathbf{S}_n^\rho$ rather than $\mathbf{S}_n^\alpha$ as was the case in earlier chapters; the reason is that the needed computations for optimizing over $\mathbf{S}_n^\alpha$ were not readily available.)

Figure 5.1 plots each of the two expectations in (5.11) and their weighted sum $\hat{\rho}$ as a function of $J_g$ for each of the first 12 sequences at time-point 10 from the scenario $\gamma_2 = 0.2$ and $\alpha_{2,o} = 0.5$. The expectation with respect to the full model is much larger by than the one with respect to small model always. One possible explanation for these differences is that, intuitively, large models are more sensitive to data than small models. But we are uncertain whether this explanation can justify such dramatic differences or if there is an error in our computations. At this time, we have not been able to find any coding errors so we shall proceed on the assumption that the computations are correct but are subject to verification.

The choice histogram for the scenario with $\gamma_2 = 0.2$ and $\alpha_{2,o} = 0.5$ (Figure 5.2) is bimodal and concentrates at the extremes, i.e., there was a tendency to use either very few or almost all of the predictuals. This shape

was observed all of the scenarios, albeit with different splits over the two extremes; typically, a larger value of $\gamma_2$ was associated with a greater chance that few predictuals would be selected.

Figures 5.3 through 5.11 compare the performance of the ROB strategy based on model averaging (labeled 'a2robwa') relative to the Bayes strategy. Table 5.1 summarizes these comparisons for all of the scenarios. Figures 5.12 through 5.20 compare the performance the the ROB strategy based on model choice (labeled 'c2robwa') relative to the Bayes strategy. Table 5.2 summarizes the comparisons for all of the scenarios.

The performance of both the ROB averaging and choice strategies are similar in that the ROB strategy tends to do better than the Bayes strategy when $\gamma_2$ is small (0 or 0.2) but substantially worse when $\gamma_2 = 0.4$ and $\alpha_{2,o} = 0.2$ or 0.5. Also, the performance of the ROB strategies relative to Bayes is greatest when $n$ is small.

## 5.2 Equalizing meta-risk

Rather than focusing on robustness to model misspecification, we might focus instead on the robustness of the predictors $\hat{Y}_{n+1}(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. One measure of the robustness of $\hat{Y}_{n+1}(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ is that its meta-risk should be relatively constant across candidate predictors in an averaged (over $\mathbf{Y}_{(n)}$) sense. This constancy property should hold regardless of which model is true. For instance, when there are only two candidate predictors, we might select the $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ that minimizes

$$\max_k \mathbf{V}_k \left(\rho_1(\hat{Y}_{n+1}; \mathbf{T}_n^\rho) - \rho_2(\hat{Y}_{n+1}; \mathbf{T}_n^\rho).\right) \tag{5.15}$$

This criterion leads to a solution that conceptually is similar to an averaged (over $\mathbf{Y}_{(n)}$) version of the adaptive minimax criterion. The plots of the meta-

risk profiles from Chapter 4 suggest that the minimax solution typically is found where the profile under the big model crosses the profile under the small model (if such a crossing occurs) or at the point where the two profiles are nearest to each other (if no crossing occurs). If these profiles reflected average (over $\mathbf{Y}_{(n)}$) rather than per-sequence properties, then these solutions would correspond typically to the locations where the minimum of (5.15) is achieved.

Once again, the difficulty with implementing a criterion based on (5.15) is that the needed expectations cannot be evaluated simply.

## 5.3 Discussion

In both the previous and the current chapters, we have assumed that the same $\mathbf{S}_n^\rho$ is used in evaluating $\rho_k(\hat{Y}_{k,n+1}; \mathbf{S}_n^\rho)$ for all $i$. However, it may be desirable to use a different $\mathbf{S}_n^\rho$ for each model depending on, say, its posterior weight. For example, suppose we are taking a model averaging approach and consider the possible influence of a model that is larger than the true model. The chance of obtaining a poor predictive distribution is relatively high since the predictive distribution would be quite sensitive to the data for this model. The sensitivity would be transferred to the mixture distribution and might result in a poor predictor. Hence, as the posterior weight of a model decreases, it may be beneficial to reduce the number of predictuals in $\mathbf{S}_n^\rho$ in order to limit the impact of data. Conversely, if the posterior weight increases we feel evermore certain that we have the correct model and so it may be beneficial to increase the number of predictuals in $\mathbf{S}_n$ to take more advantage of the data.

Table 5.1: Summary comparison of the ROB mongrel averaging strategy to the Bayes strategy.

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| **I** | 0.2 | 0 | + | + | + |
| | | 0.2 | ++ | + | + |
| | | 0.4 | + | − | − − − |
| | 0.5 | 0 | ++ | + | + |
| | | 0.2 | ++ | ++ | + |
| | | 0.4 | +++ | ++ | 0 |
| | 0.8 | 0 | + | 0 | − |
| | | 0.2 | ++ | ++ | ++ |
| | | 0.4 | +++ | ++ | ++ |

Table 5.2: Summary comparison of the ROB mongrel choice strategy to the Bayes strategy.

| | | | n | | |
|---|---|---|---|---|---|
| $\Gamma_2$ | $\alpha_{2,o}$ | $\gamma_2$ | 10 to 20 | 20 to 30 | 30 to 40 |
| **I** | 0.2 | 0 | +++ | ++ | + |
| | | 0.2 | ++ | ++ | + |
| | | 0.4 | 0 | − − − | − − − |
| | 0.5 | 0 | +++ | ++ | + |
| | | 0.2 | +++ | ++ | + |
| | | 0.4 | − − | − − − | − − − |
| | 0.8 | 0 | 0 | − | − |
| | | 0.2 | 0 | + | + |
| | | 0.4 | 0 | + | + |

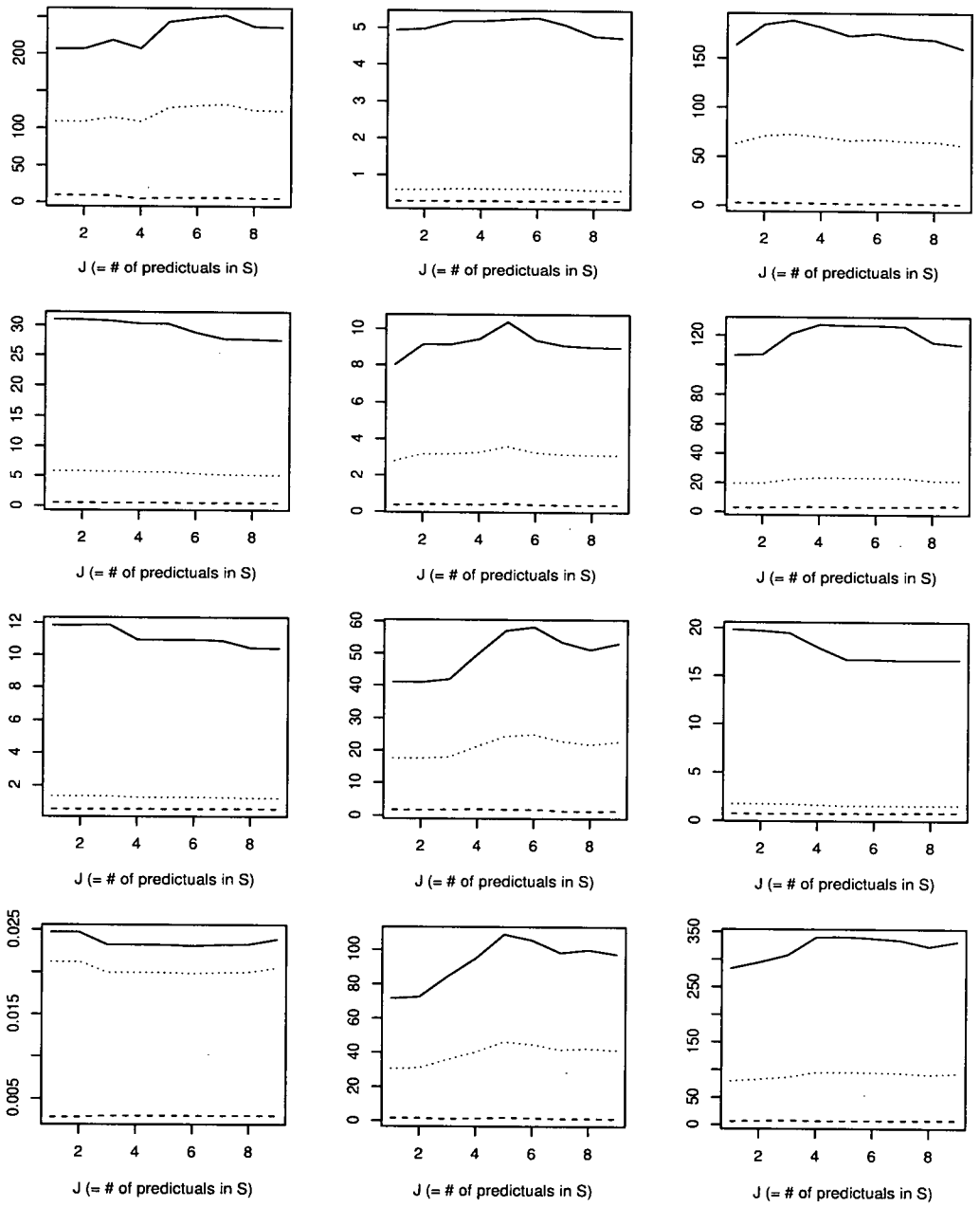


Figure 5.1: Robustness profiles as a function of the number of predictuals included in $\mathbf{S}_n^\rho$ for ROB strategy: $a_{2,o} = 0.5$, $\gamma_2 = 0.2$.

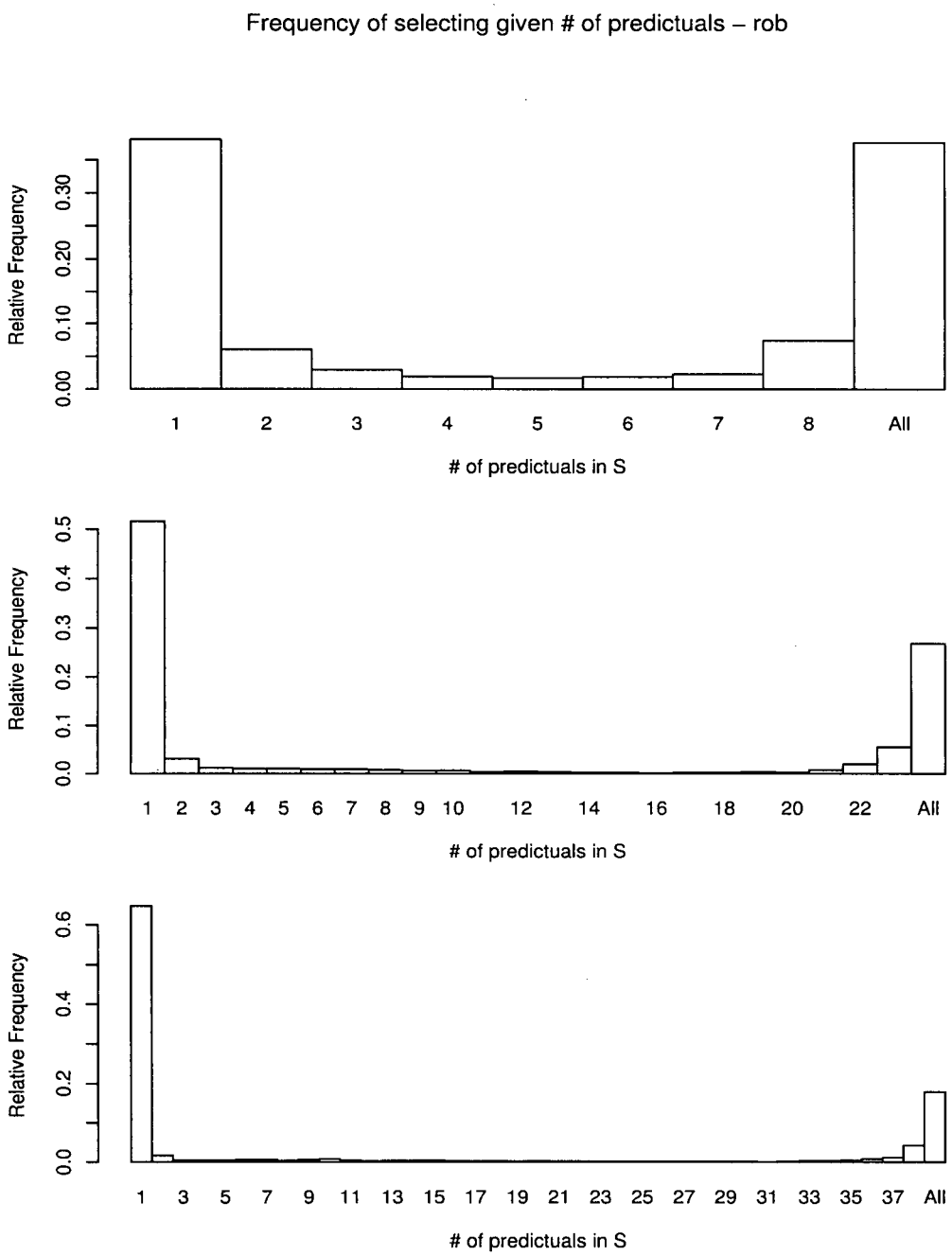


Figure 5.2: Histograms of the number of predictuals to include in $\mathbf{S}_n^{\rho}$ as selected by ROB strategy: $a_{2,o} = 0.5$, $\gamma_2 = 0.2$.

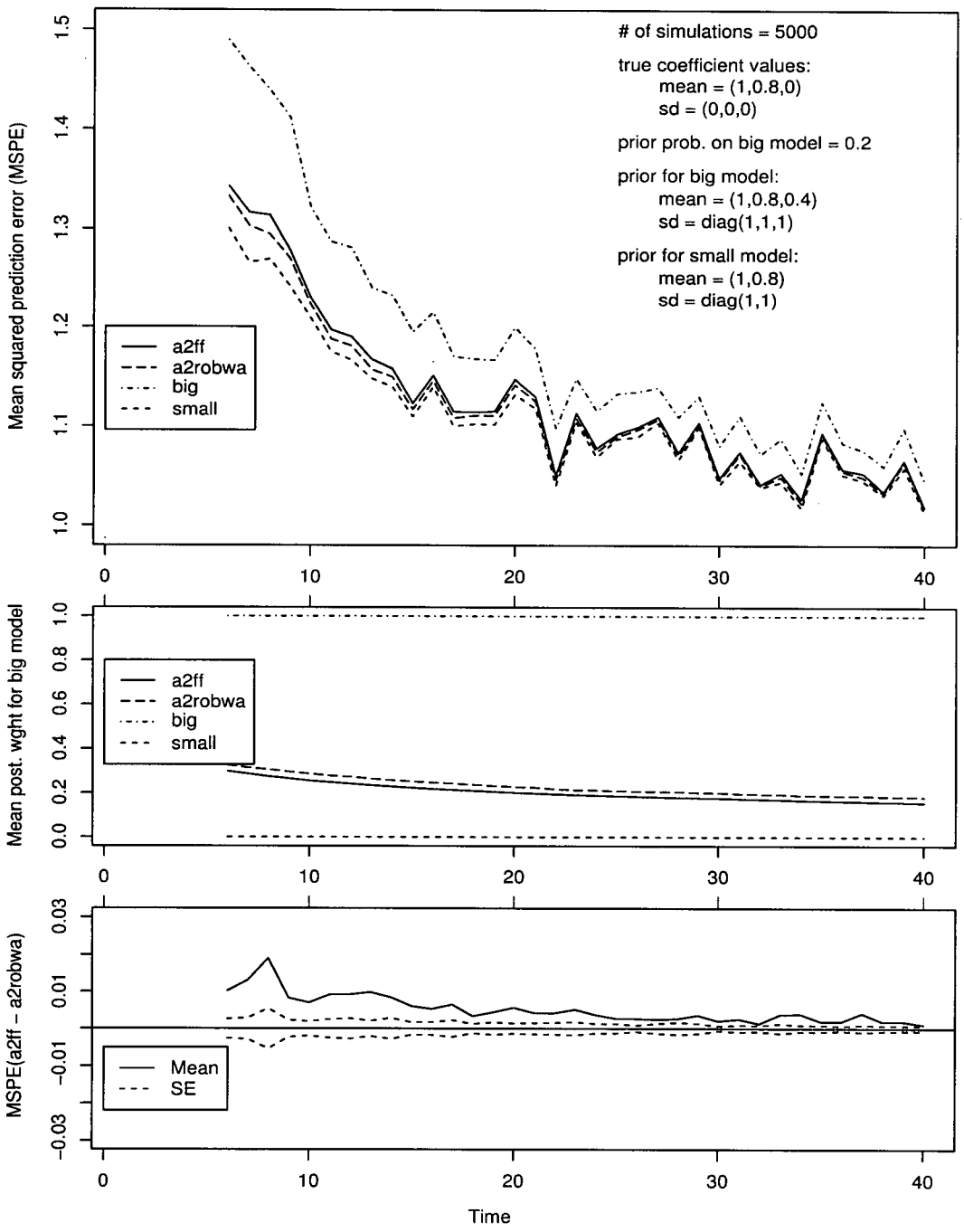


Figure 5.3: Performance of ROB averaging strategy: $a_{2,o} = 0.2$, $\gamma_2 = 0$.

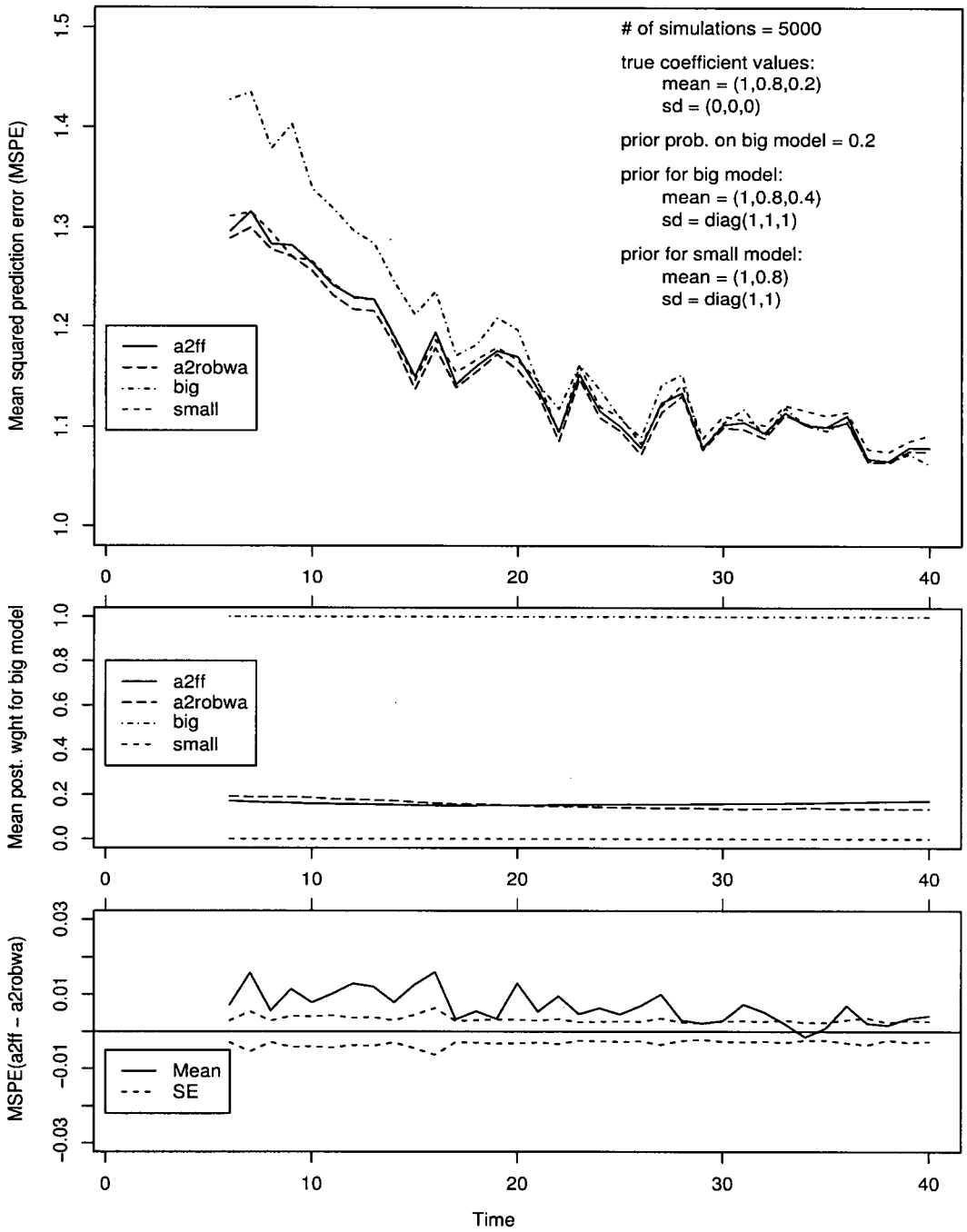


Figure 5.4: Performance of ROB averaging strategy: $a_{2,o} = 0.2$, $\gamma_2 = 0.2$.

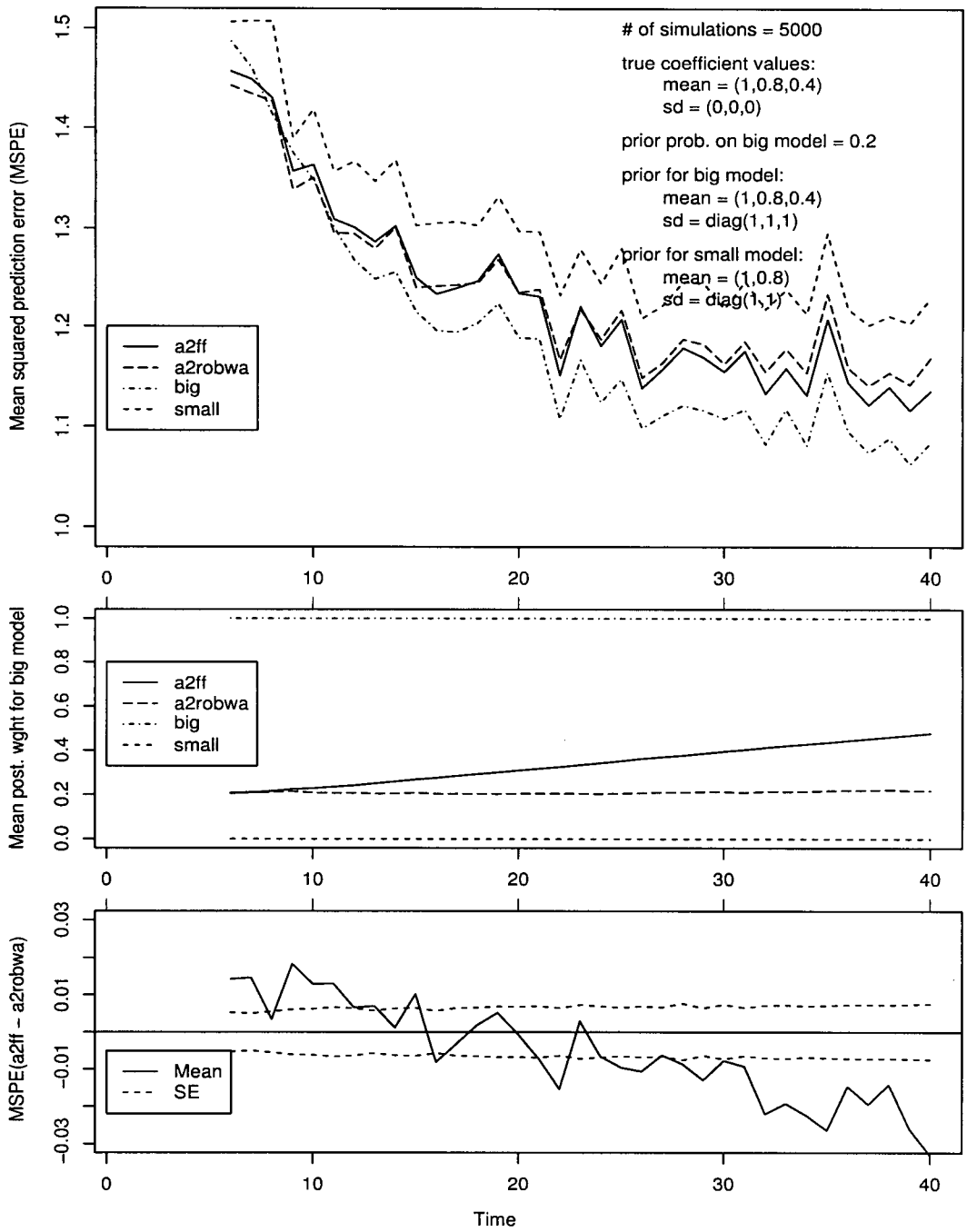


Figure 5.5: Performance of ROB averaging strategy: $a_{2,o} = 0.2$, $\gamma_2 = 0.4$.

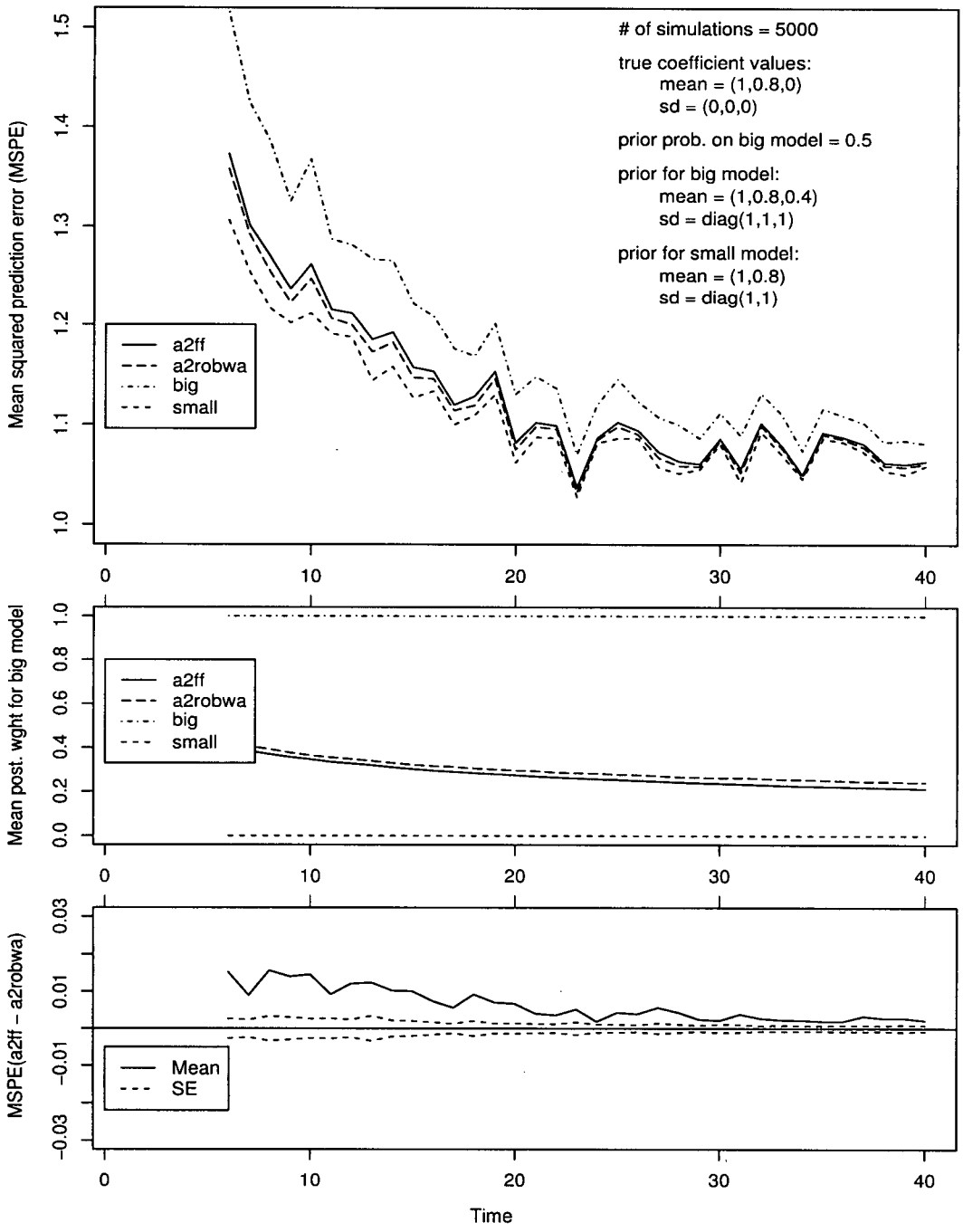


Figure 5.6: Performance of ROB averaging strategy: $a_{2,o} = 0.5$, $\gamma_2 = 0$.

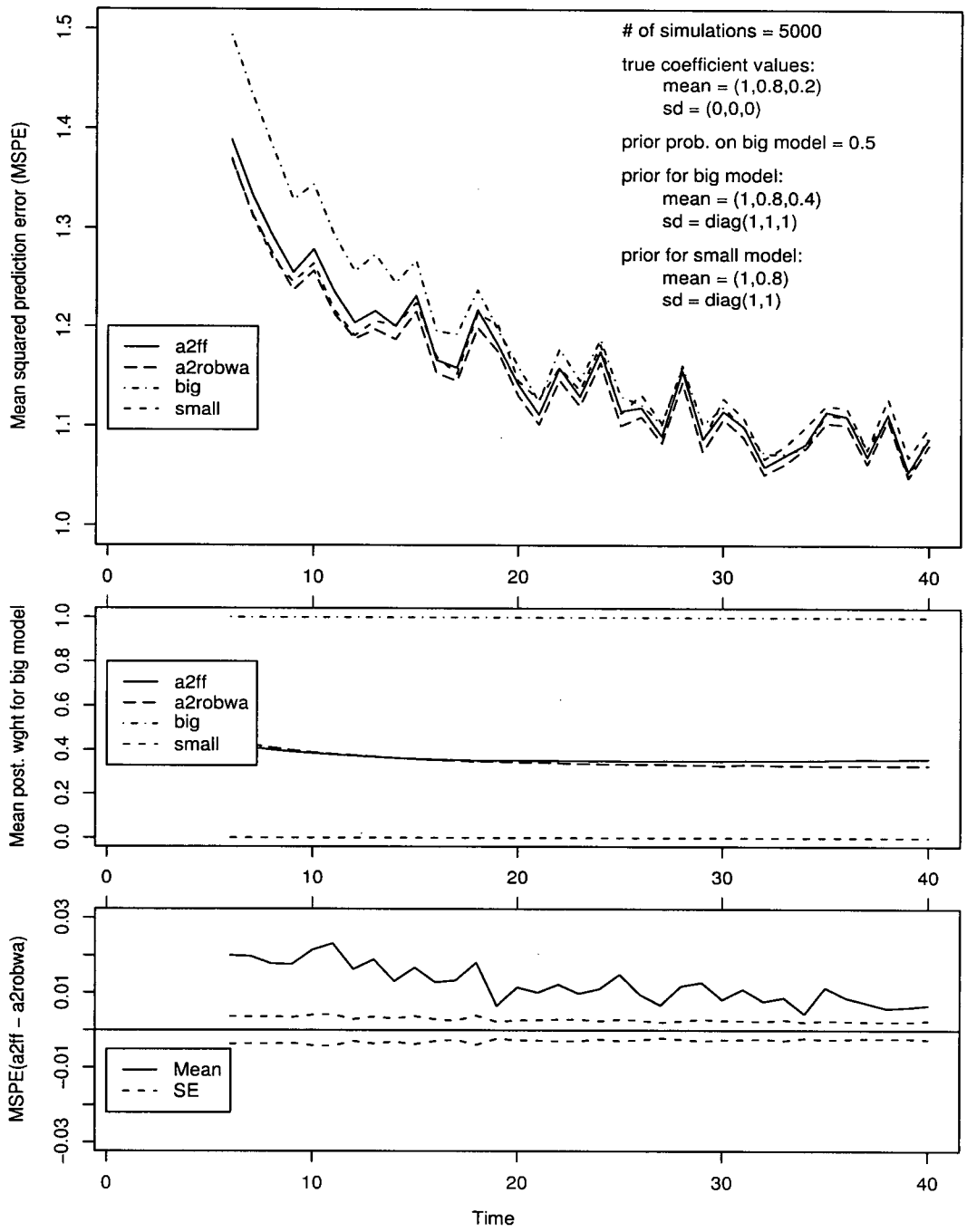


Figure 5.7: Performance of ROB averaging strategy: $a_{2,o} = 0.5$, $\gamma_2 = 0.2$.

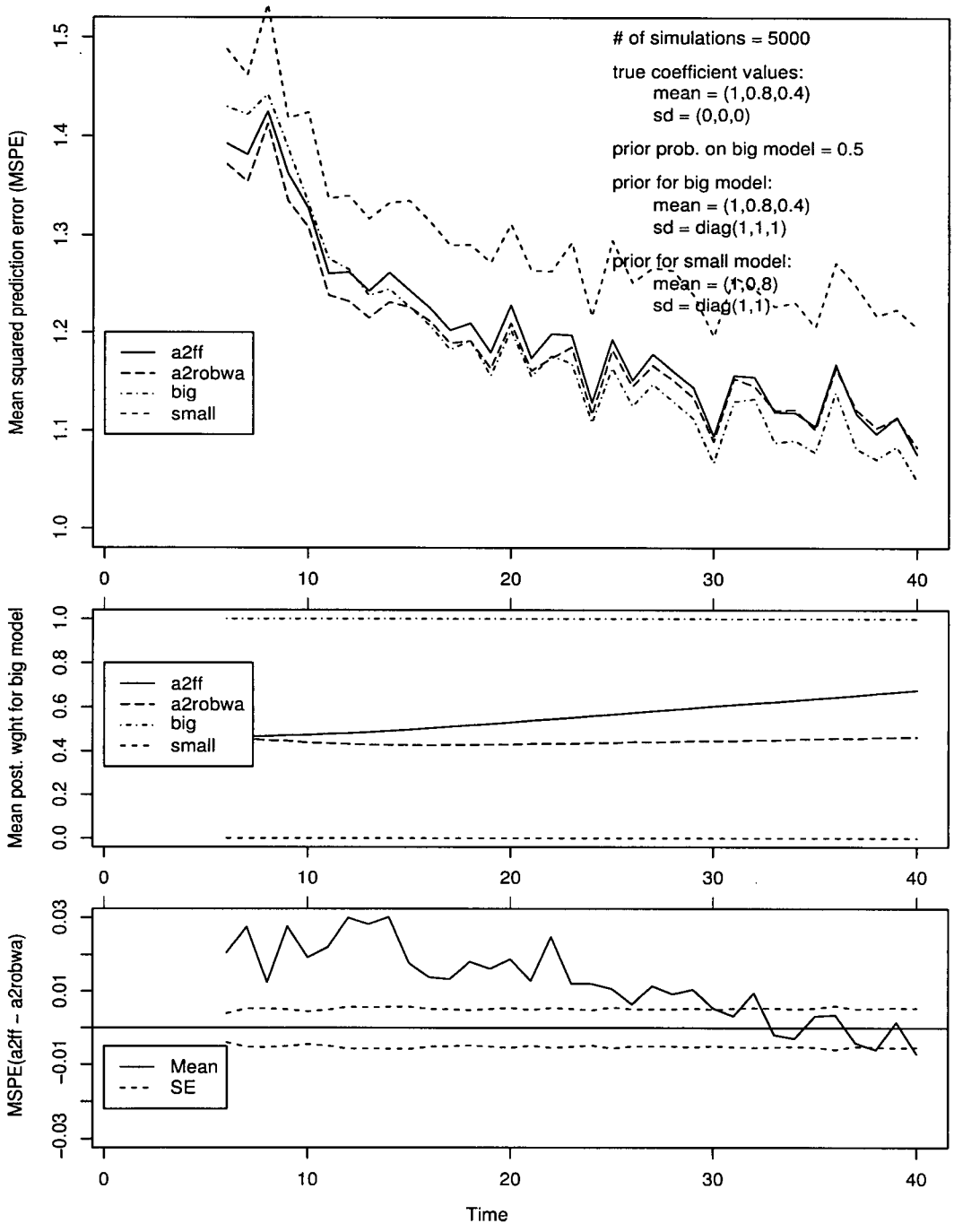


Figure 5.8: Performance of ROB averaging strategy: $a_{2,o} = 0.5$, $\gamma_2 = 0.4$.

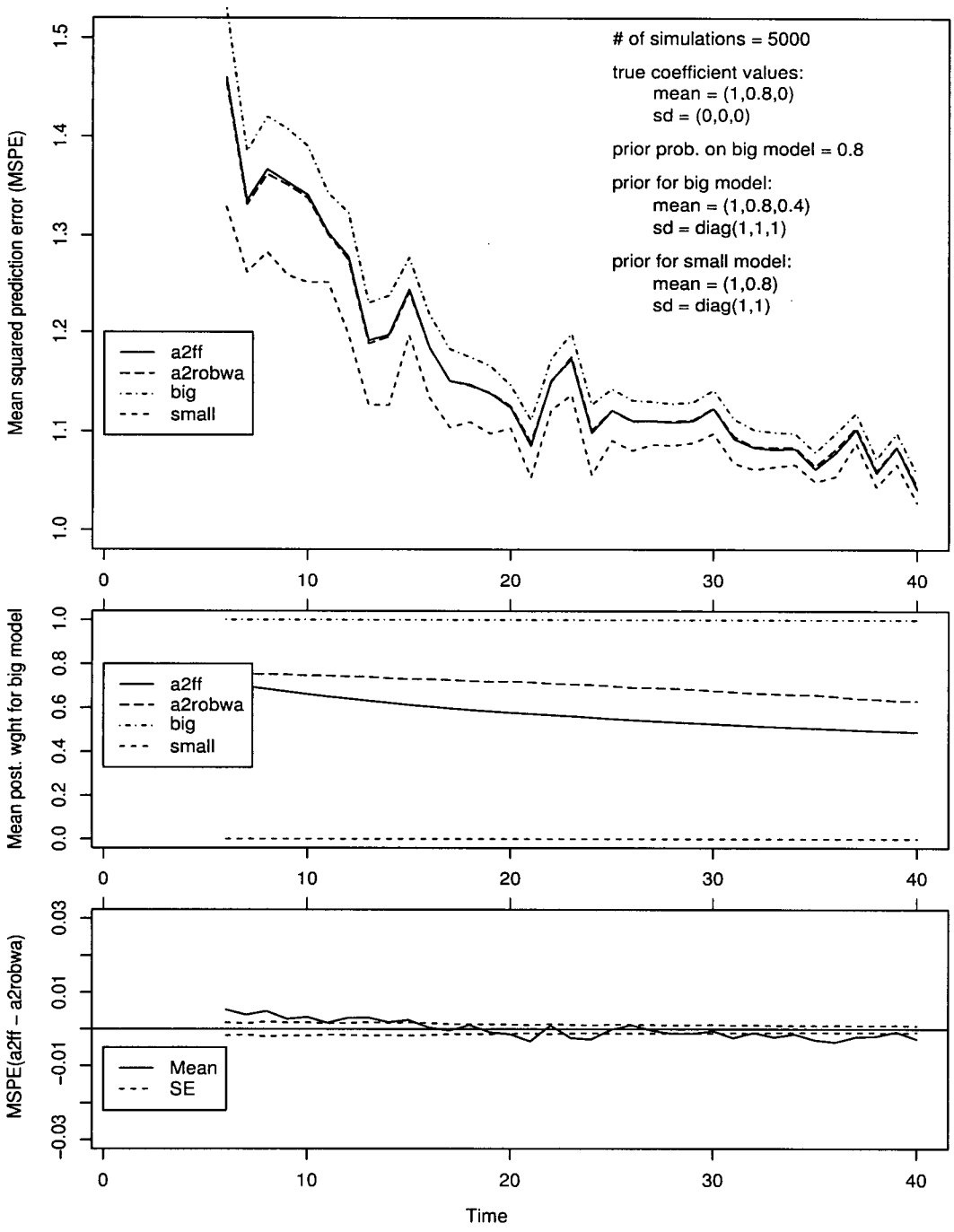


Figure 5.9: Performance of ROB averaging strategy: $a_{2,o} = 0.8$, $\gamma_2 = 0$.

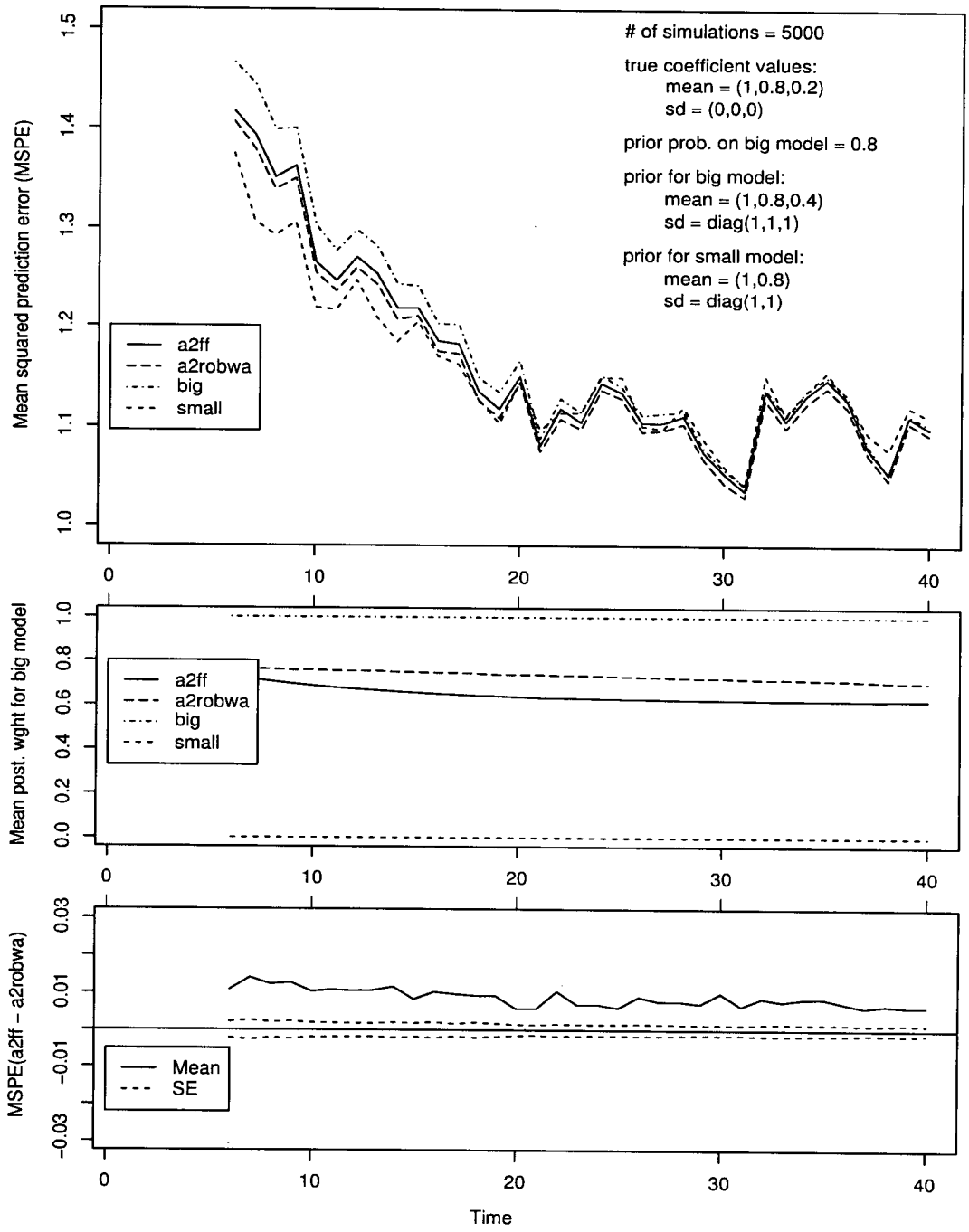


Figure 5.10: Performance of ROB averaging strategy: $a_{2,o} = 0.8$, $\gamma_2 = 0.2$.

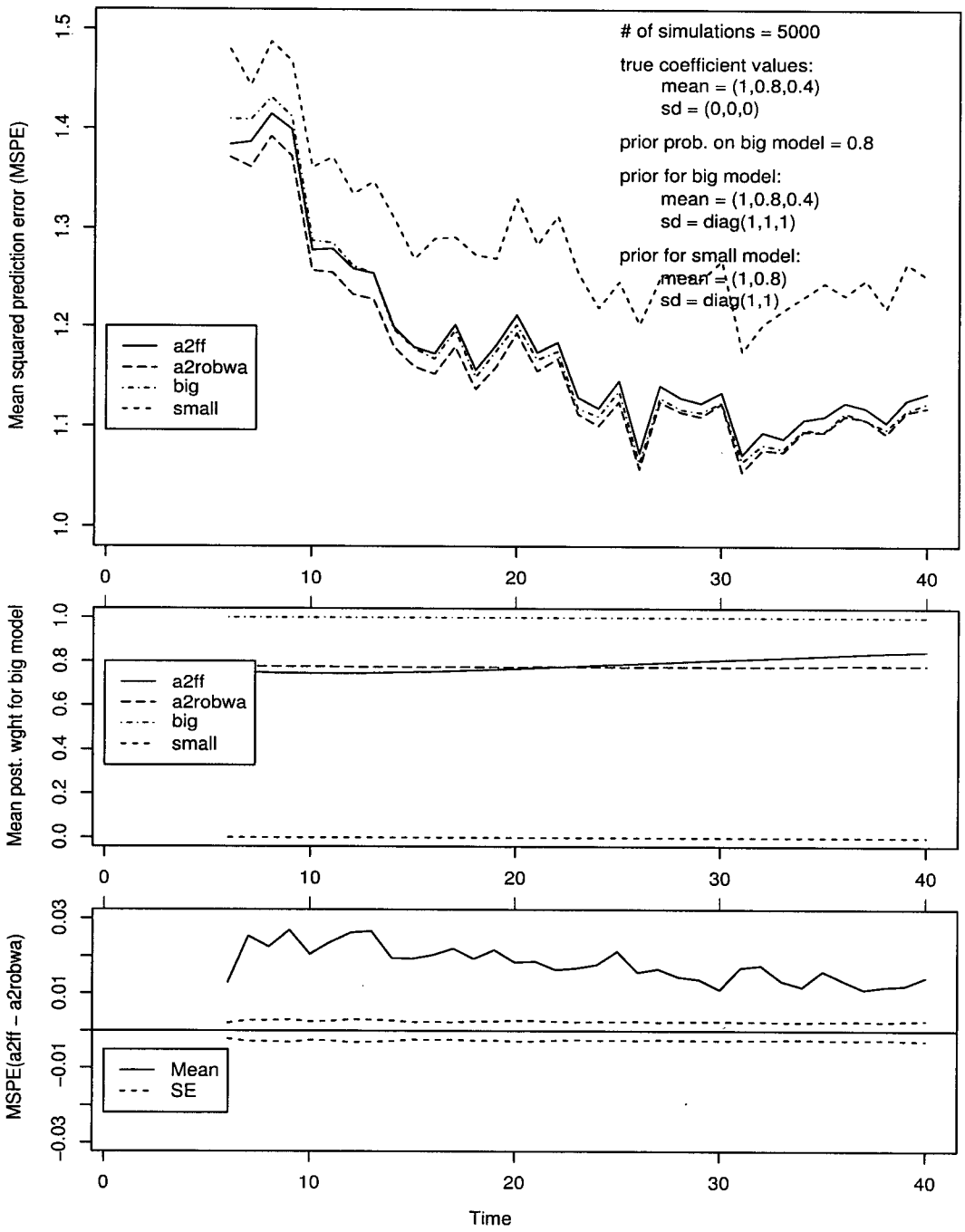


Figure 5.11: Performance of ROB averaging strategy: $a_{2,o} = 0.8$, $\gamma_2 = 0.4$.

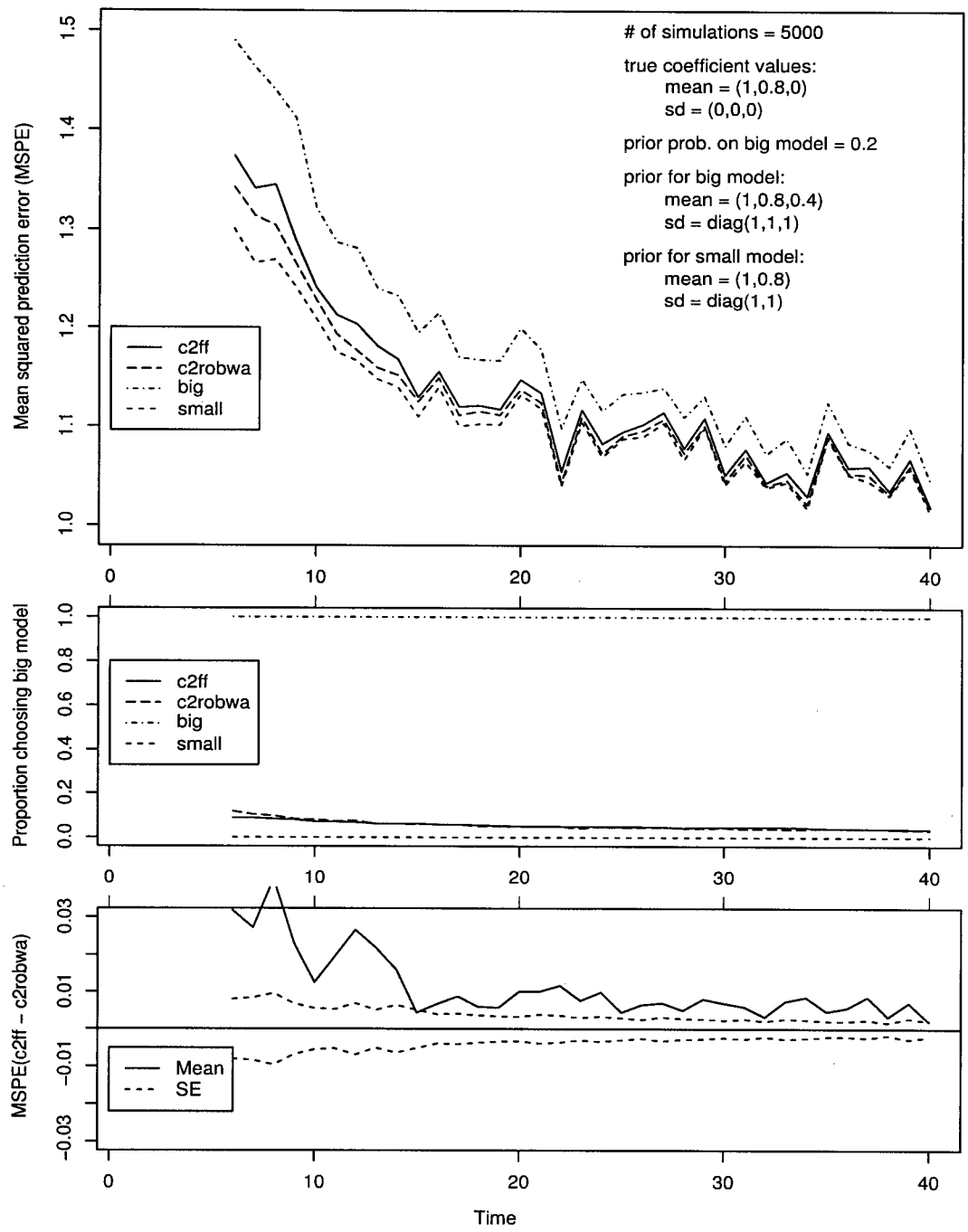


Figure 5.12: Performance of ROB choice strategy: $a_{2,o} = 0.2$, $\gamma_2 = 0$.

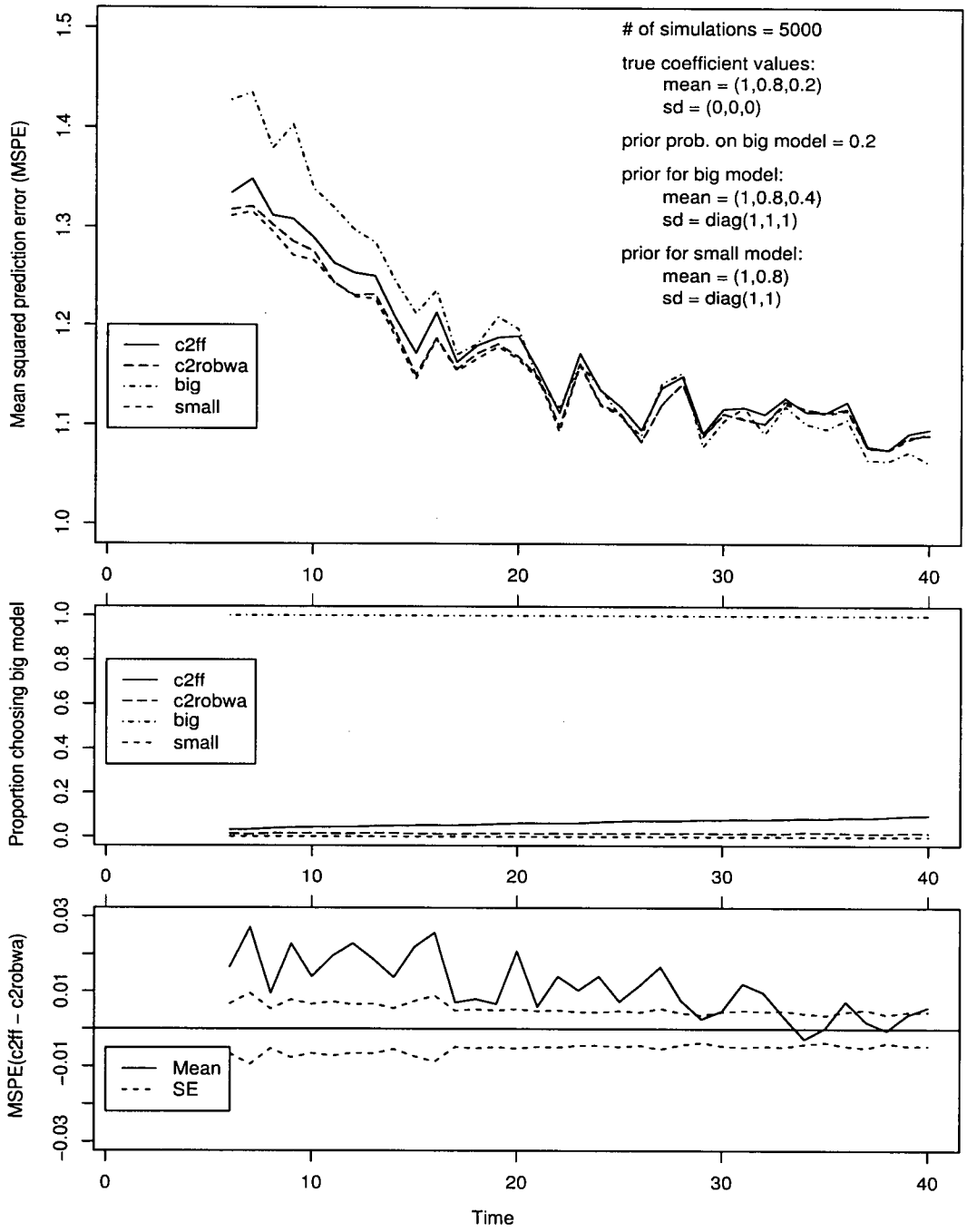


Figure 5.13: Performance of ROB choice strategy: $a_{2,o} = 0.2$, $\gamma_2 = 0.2$.

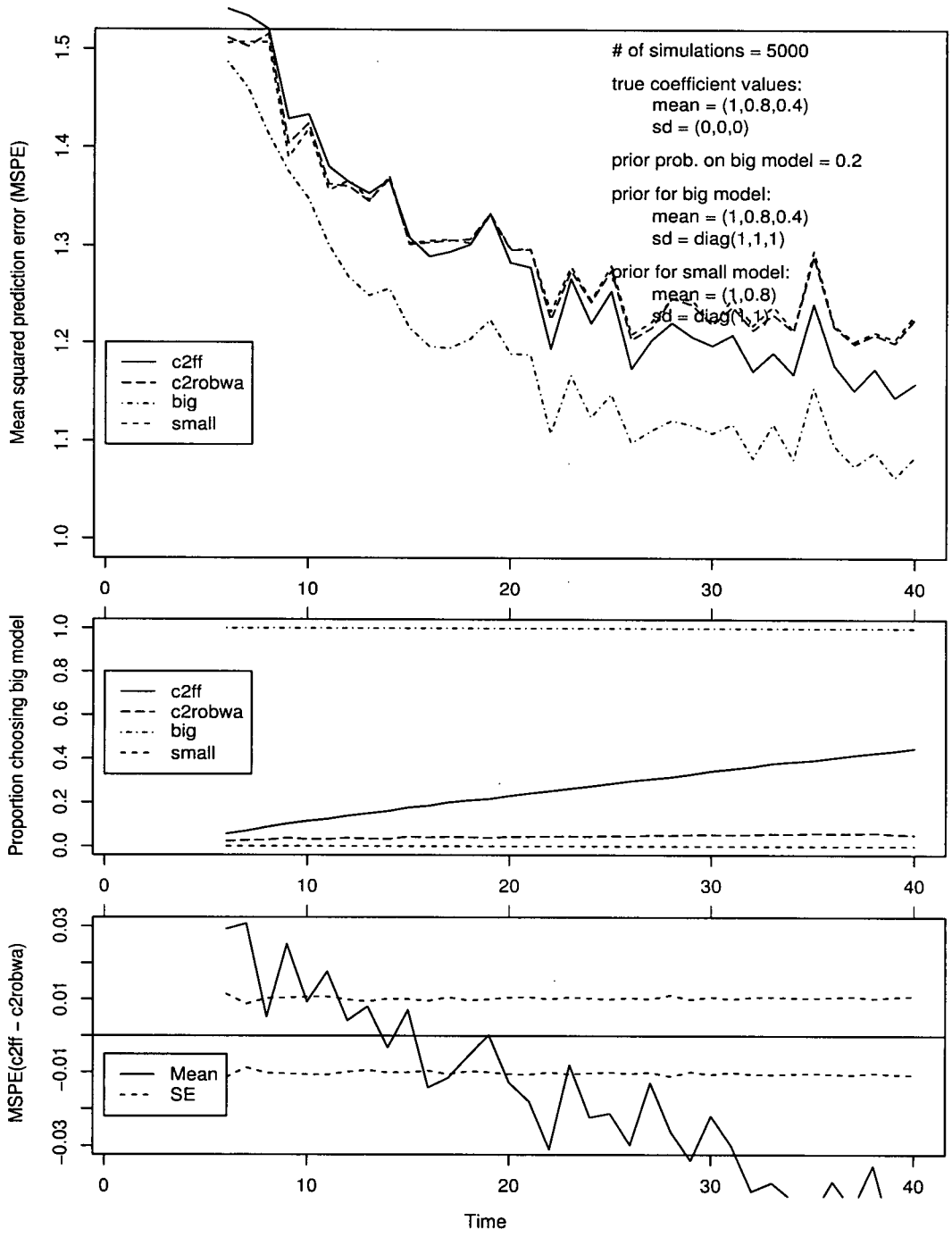


Figure 5.14: Performance of ROB choice strategy: $a_{2,o} = 0.2$, $\gamma_2 = 0.4$.

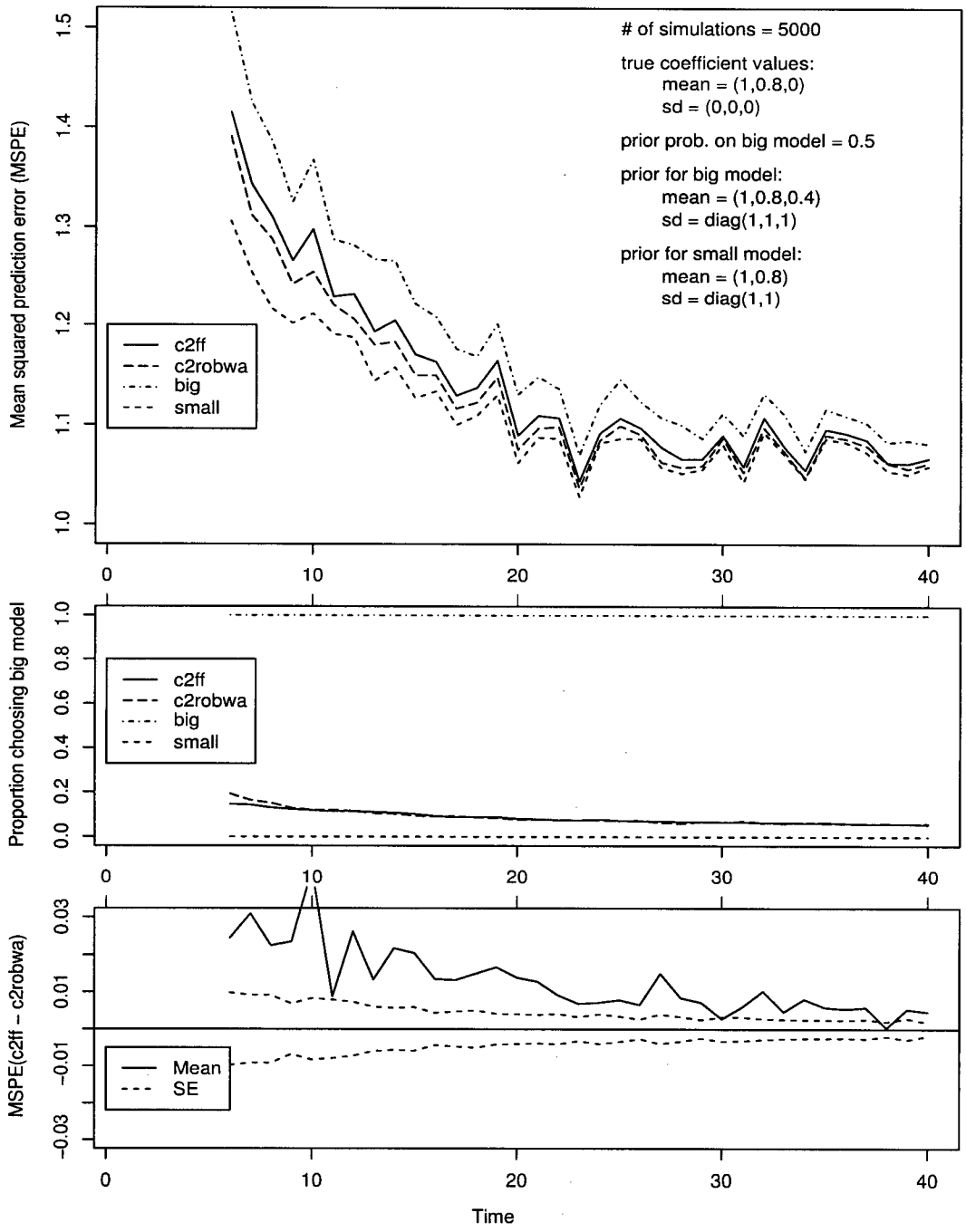


Figure 5.15: Performance of ROB choice strategy: $a_{2,o} = 0.5$, $\gamma_2 = 0$.

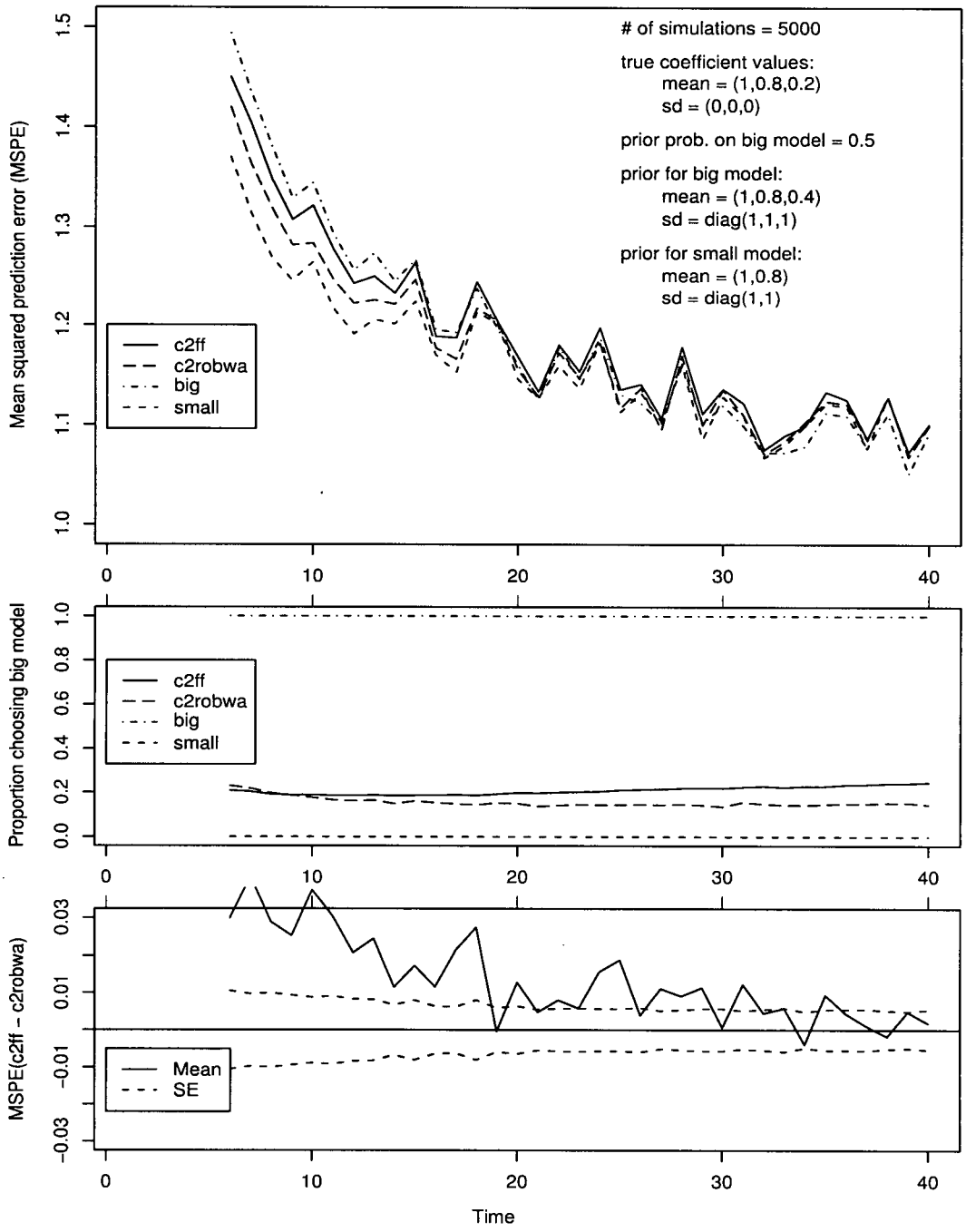


Figure 5.16: Performance of ROB choice strategy: $a_{2,o} = 0.5$, $\gamma_2 = 0.2$.

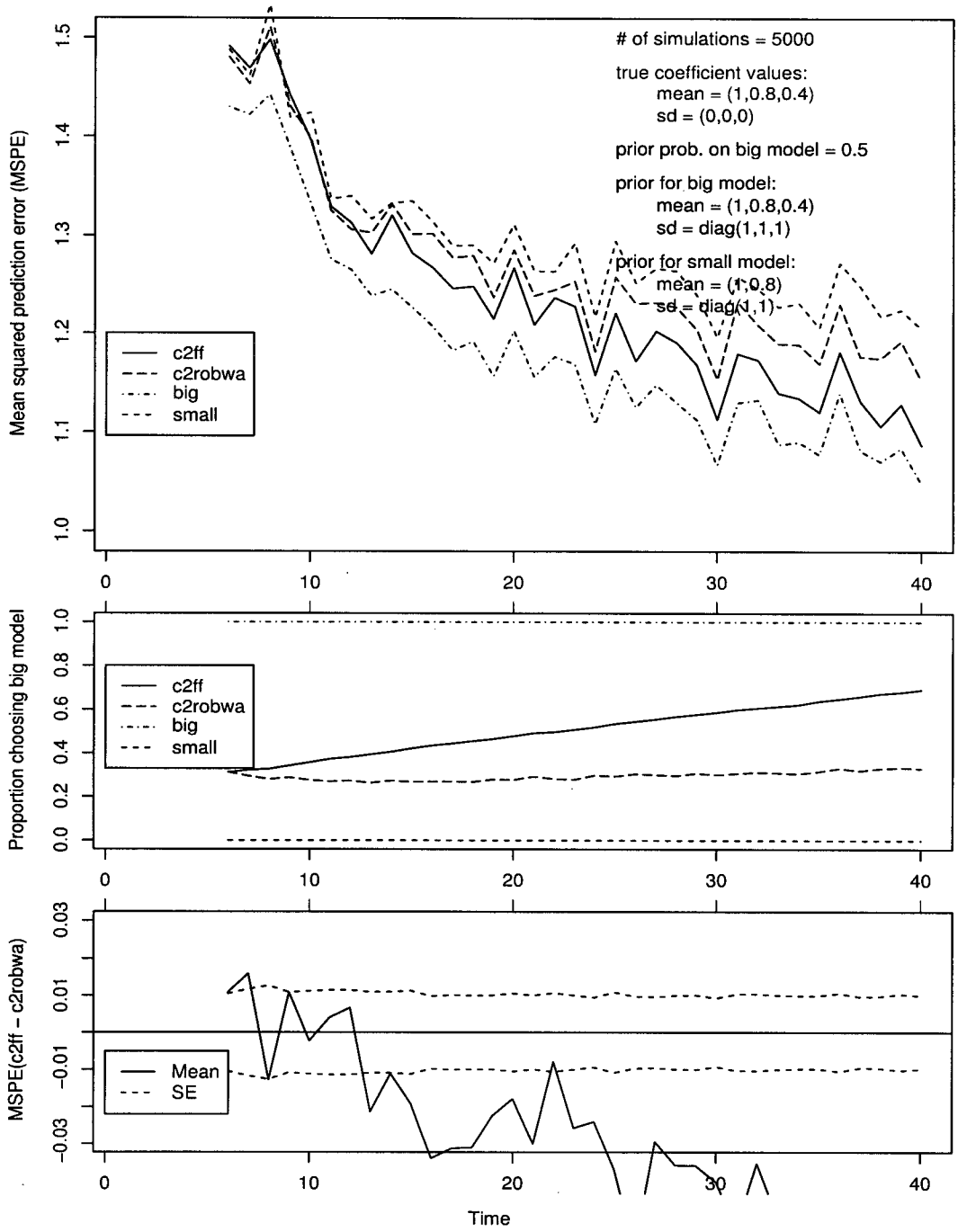


Figure 5.17: Performance of ROB choice strategy: $a_{2,o} = 0.5$, $\gamma_2 = 0.4$.

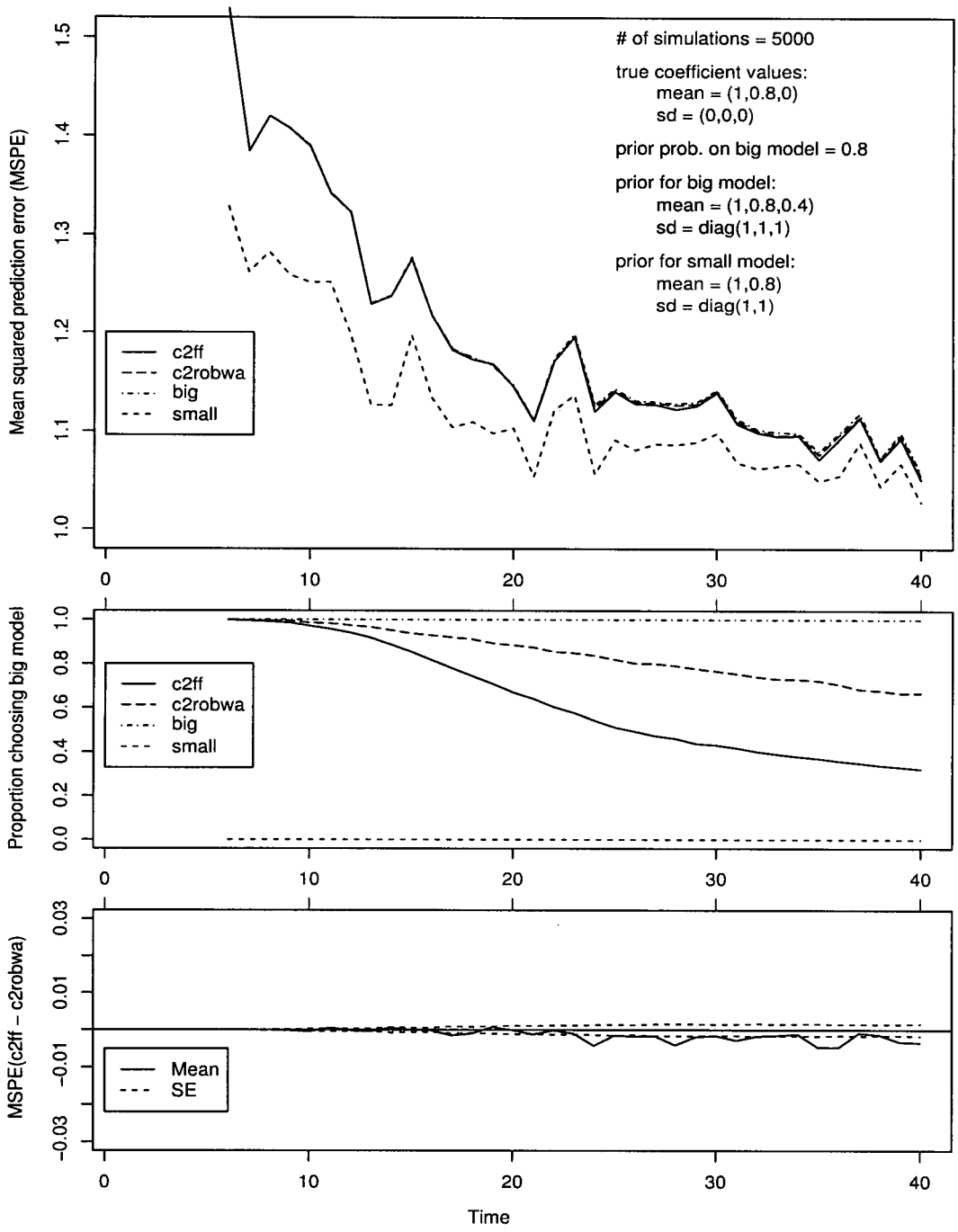


Figure 5.18: Performance of ROB choice strategy: $a_{2,o} = 0.8$, $\gamma_2 = 0$.

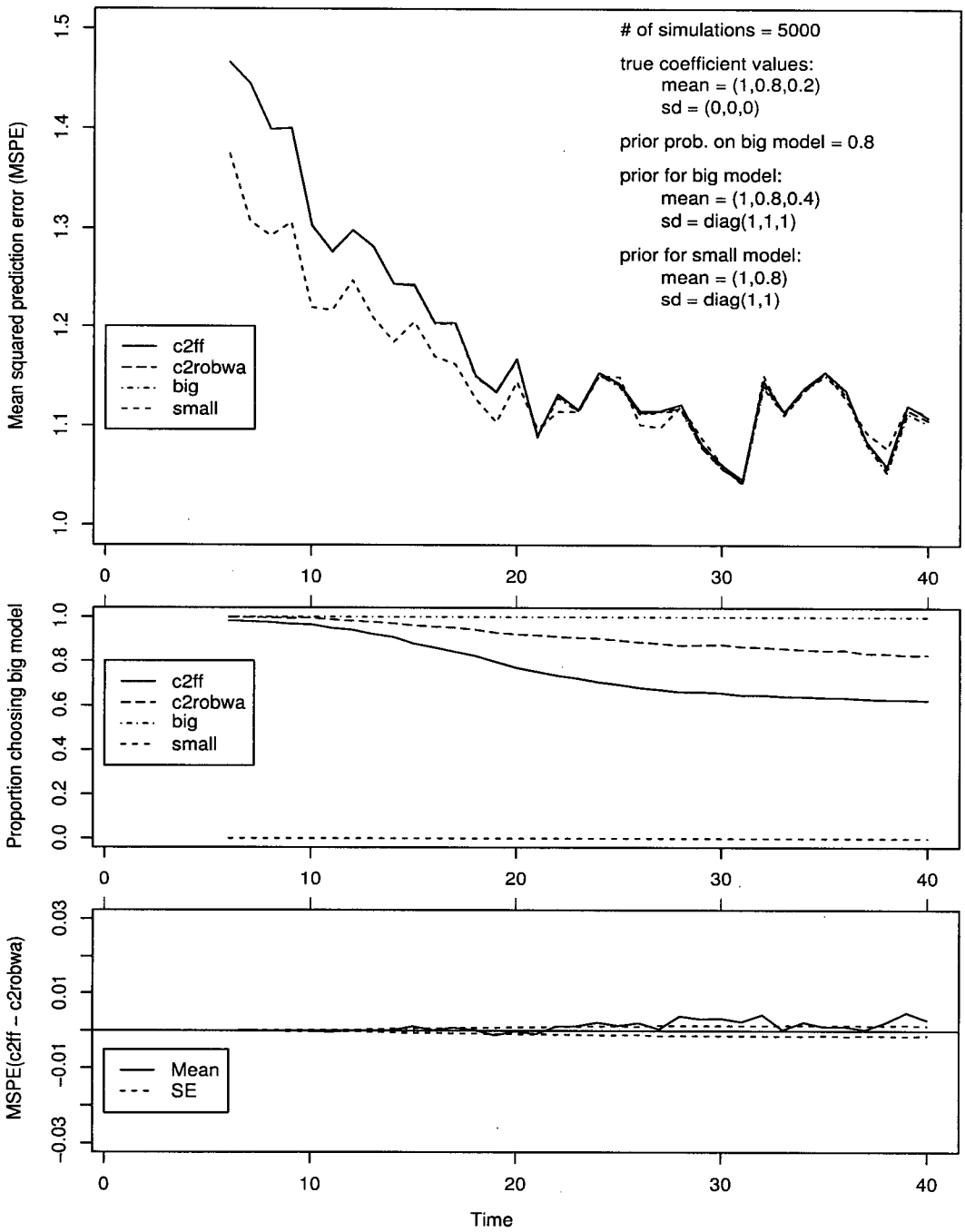


Figure 5.19: Performance of ROB choice strategy: $a_{2,o} = 0.8$, $\gamma_2 = 0.2$.

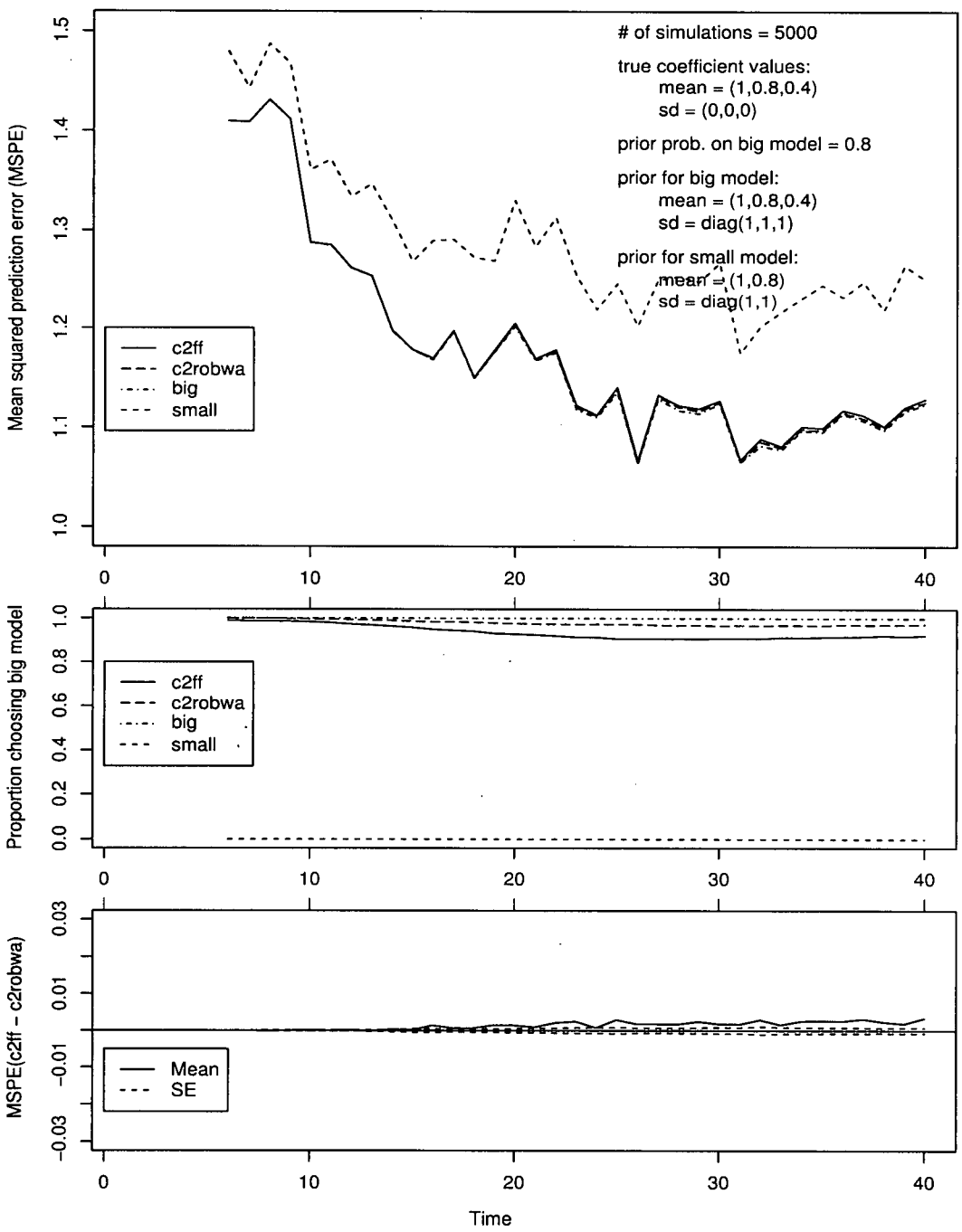


Figure 5.20: Performance of ROB choice strategy: $a_{2,o} = 0.8$, $\gamma_2 = 0.4$.

# Chapter 6

# Asymptotics

In this chapter, we establish guidelines for ensuring that the conditioning statistics $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$ have good asymptotic properties. Once again the context will be a pair of normal linear models with $\mathbf{S}_n^\alpha$ and $\mathbf{S}_n^\rho$ restricted to affine functions of the response. We characterize the class of sequences that guarantee that the prediction generated by a model choice or averaging strategy will be asymptotically equivalent to the Bayes predictor from the true model.

## 6.1 Consistency of Model Weights

Regardless of whether a model choice or a model averaging approach is taken, we may wish to require that the model weights, $\alpha_k$, be consistently estimated, that is, if model $k$, say, were true, then we would like $\alpha_k \to 1$ (weakly or strongly) as more data become available. Otherwise, in the model choice approach, the components of the risk that are computed under the wrong model (which, as such, are suspect) continue to influence the overall assessment even asymptotically. In the model averaging context, consistent estimation of

$\alpha_k$ ensures that the predictor is derived ultimately from only the correct model. We give a complete proof for weak convergence using Chebyshev's inequality and then outline a Wald-type proof for strong convergence.

Let $k$, $k'$ index normal linear models with model $k'$ nested within model $k$. Partition the design matrix as $\mathbf{Z}_{k,(n)} = (\mathbf{Z}_{k',(n)} \mid \tilde{\mathbf{Z}})$ where $\tilde{\mathbf{Z}}$ consists of the $\tilde{p} = p_k - p_{k'}$ covariates that are present in model $k$ but not in model $k'$. Suppose that $\Gamma_k$ is block diagonal with respect to the partitioned parameter vector, i.e.,

$$\Gamma_k = \begin{pmatrix} \Gamma_{k'} & \mathbf{0} \\ \mathbf{0} & \tilde{\Gamma} \end{pmatrix}. \tag{6.1}$$

We will establish conditions on $\mathbf{U}$ that yield consistent estimation of $\alpha_i$ for $i = k$ and $i = k'$.

Let $\lambda_i(\mathbf{A})$, $\lambda_{\min}(\mathbf{A})$, $\lambda_{\max}(\mathbf{A})$ denote respectively, the $i$-th, the minimum, and the maximum eigenvalues of the matrix $\mathbf{A}$. Let $\|\mathbf{x}\|^2_{\mathbf{A}} \equiv \mathbf{x}^T\mathbf{A}\mathbf{x}$. For any non-negative definite matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimension, we write $\mathbf{A} \leq \mathbf{B}$ iff $\|\mathbf{x}\|^2_{\mathbf{A}} \leq \|\mathbf{x}\|^2_{\mathbf{B}}$ for all vectors $\mathbf{x}$.

It will be useful to re-express (3.22) as

$$\alpha_k(\mathbf{S}^\alpha_n) = \frac{\alpha_{k,o}}{\alpha_{k,o} + \alpha_{k',o}M} \tag{6.2}$$

where

$$M = \frac{m_{k'}(\mathbf{S}^\alpha_n)}{m_k(\mathbf{S}^\alpha_n)} \tag{6.3}$$

is the ratio of the marginal densities for $\mathbf{S}^\alpha_n$ from the two models and to define

$$-2\log M = \Delta - \log D \tag{6.4}$$

where

$$D = \frac{|\Sigma_k|}{|\Sigma_{k'}|} = |\Sigma^{-1/2}_{k'}\Sigma_k\Sigma^{-1/2}_{k'}| \tag{6.5}$$

$$\Delta = \|\mathbf{S}_n^\alpha - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}} - \|\mathbf{S}_n^\alpha - \mu_k\|^2_{\Sigma_k^{-1}} \quad (6.6)$$

$$= \|\mathbf{S}_n^\alpha - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}-\Sigma_k^{-1}} - 2(\mathbf{S}_n - \mu_{k'})^T\Sigma_k^{-1}(\mu_{k'} - \mu_k) - \|\mu_{k'} - \mu_k\|^2_{\Sigma_k^{-1}} \quad (6.7)$$

Denote the true parameter value by $\beta_o = (\beta_o^*, \tilde{\beta})$ where $\tilde{\beta} \equiv \mathbf{0}$ when model $k'$ is true. Let

$$\mu_o = \mathbf{U}^T\mathbf{Z}_{k,(n)}\beta_o$$

$$= \mathbf{U}^T\left(\mathbf{Z}_{k',(n)}\beta_o^* + \tilde{\mathbf{Z}}\tilde{\beta}\right) \quad (6.8)$$

denote the expected value of $\mathbf{S}_n^\alpha$.

Applying (A4) to (6.6),

$$\mathbf{E}_i\Delta = \left[\mathrm{tr}\,(\Sigma_{k'}^{-1}\Sigma_i) + \|\mu_o - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}\Sigma_i\Sigma_{k'}^{-1}}\right] - \left[\mathrm{tr}\,(\Sigma_k^{-1}\Sigma_i) + \|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}\Sigma_i\Sigma_k^{-1}}\right]$$

$$= \mathrm{tr}\,(\Sigma_{k'}^{-1}\Sigma_i - \Sigma_k^{-1}\Sigma_i) + \|\mu_o - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}\Sigma_i\Sigma_{k'}^{-1}} - \|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}\Sigma_i\Sigma_k^{-1}}. \quad (6.9)$$

To obtain a bound on the variance of $-2\log M$, we apply the relation $\mathbf{V}(A+B) \leq 2(\mathbf{V}(A)+\mathbf{V}(B))$ to (6.7) and then use (A5) to obtain

$$\mathbf{V}_i\Delta \leq 2\left[\mathbf{V}_i(\|\mathbf{S}_n - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}-\Sigma_k^{-1}} + \mathbf{V}_i(2(\mathbf{S}_n - \mu_k)^T\Sigma_k^{-1}(\mu_{k'} - \mu_k))\right]$$

$$= 4\left[\mathrm{tr}\,((\Sigma_{k'}^{-1} - \Sigma_k^{-1})\Sigma_i)^2 + 2\|\mu_o - \mu_{k'}\|^2_{(\Sigma_{k'}^{-1}-\Sigma_k^{-1})\Sigma_i(\Sigma_{k'}^{-1}-\Sigma_k^{-1})}\right.$$

$$\left. + \|\mu_{k'} - \mu_k\|^2_{\Sigma_k^{-1}\Sigma_i\Sigma_k^{-1}}\right]. \quad (6.10)$$

**Lemma 6.1** *The quantities* $\|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}}$, $\|\mu_o - \mu_{k'}\|^2_{\Sigma_k^{-1}}$, *and* $\|\mu_{k'} - \mu_k\|^2_{\Sigma_k^{-1}}$ *are bounded irrespective of whether model* $k$ *or model* $k'$ *is true.*

**Proof** Substituting for $\Sigma_k^{-1}$ using (3.15) and applying (A3),

$$\|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}} = \|\beta_o - \mathbf{b}_k\|_{\mathbf{Z}_{k,(n)}^T\mathbf{U}\Sigma_k^{-1}\mathbf{U}^T\mathbf{Z}_{k,(n)}}$$

$$\leq \|\beta_o - \mathbf{b}_k\|_{\Gamma_k^{-1}}. \quad (6.11)$$

The expression on the final line involves only constants and hence is bounded. The proofs for the other two quantities follow analogously. ∎

**Lemma 6.2** *When model $k'$ is true, $\|\mu_o - \mu_{k'}\|^2_{\Sigma^{-1}_{k'}}$ is bounded.*

**Proof** Making use of the fact that $\tilde{\beta} = \mathbf{0}$, substituting for $\Sigma^{-1}_{k'}$ using (3.6) and applying (A3),

$$\begin{aligned} \|\mu_o - \mu_{k'}\|^2_{\Sigma^{-1}_{k'}} &= \|\mathbf{U}^T\mathbf{Z}_{k,(n)}(\beta^*_o - \mathbf{b}_{k'})\|_{\Sigma^{-1}_{k'}} \\ &= \|\beta^*_o - \mathbf{b}_{k'}\|_{\mathbf{Z}^T_{k,(n)}\mathbf{U}\Sigma^{-1}_{k'}\mathbf{U}^T\mathbf{Z}_{k,(n)}} \\ &\leq \|\beta^*_o - \mathbf{b}_k\|_{\Gamma^{-1}_{k'}}. \end{aligned} \tag{6.12}$$

The expression on the last line involves only constants and hence is bounded. ∎

**Lemma 6.3** *Let* $\mathbf{G} = \Sigma^{-1/2}_{k'}\mathbf{U}^T\tilde{\mathbf{Z}} = \left(\mathbf{U}^T\Psi_{k'}\mathbf{U}\right)^{-1/2}\mathbf{U}^T\tilde{\mathbf{Z}}$ *and let* $\mathbf{H} = \mathbf{G}\tilde{\Gamma}\mathbf{G}^T$. *The following are equivalent:*

$$(i) \qquad \lambda_{\max}(\mathbf{G}^T\mathbf{G}) \to \infty \tag{6.13}$$

$$(ii) \qquad \lambda_{\max}(\mathbf{H}) \to \infty \tag{6.14}$$

$$(iii) \qquad \mathrm{tr}\,(\mathbf{H}) \to \infty. \tag{6.15}$$

**Proof** (i) $\Longleftrightarrow$ (ii): Observe that $\lambda_{\min}(\tilde{\Gamma})\mathbf{G}\mathbf{G}^T \leq \mathbf{H} \leq \lambda_{\max}(\tilde{\Gamma})\mathbf{G}\mathbf{G}^T$ implies $\lambda_{\min}(\tilde{\Gamma})\lambda_{\max}(\mathbf{G}\mathbf{G}^T) \leq \lambda_{\max}(\mathbf{H}) \leq \lambda_{\max}(\tilde{\Gamma})\lambda_{\max}(\mathbf{G}\mathbf{G}^T)$. The result then follows from the fact that the non-zero eigenvalues of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ are equal for any matrix $\mathbf{A}$. (ii) $\Longleftrightarrow$ (iii): Obvious since $\mathbf{H}$ has at most $\tilde{p}$ non-zero eigenvalues and $\mathrm{tr}\,(\mathbf{H})$ equals the sum of its eigenvalues. ∎

**Theorem 6.1** *If model $k'$ is true, then a necessary and sufficient condition for $\alpha_{k'} \to 1$ is that* $\mathbf{U}$ *satisfies (6.13).*

**Proof** Clearly, $\alpha_{k'} \to 1 \iff \alpha_k \to 0 \iff (-2\log M) \to -\infty$. It is sufficient to show that $D \to \infty \iff$ (6.13) holds while both $\mathbf{E}_{k'}\Delta$ and $\mathbf{V}_{k'}\Delta$ are bounded irrespective of whether (6.13) holds.

Observe that

$$\Sigma_k = \Sigma_{k'} + \mathbf{U}^T\tilde{\mathbf{Z}}\tilde{\Gamma}\tilde{\mathbf{Z}}^T\mathbf{U} \tag{6.16}$$

implies $D = |\mathbf{I} + \mathbf{H}|$. Clearly, $\mathbf{I} + \mathbf{H}$ has at most $\tilde{p}$ eigenvalues greater than one and the remaining eigenvalues are equal to one. Since the determinant of a matrix equals the product of its eigenvalues, it follows that (1) $D \geq 1$, and (2) by Lemma 6.3 $D \to \infty \iff$ (6.13) holds.

Setting $i = k'$ in (6.9) and simplifying,

$$\mathbf{E}_{k'}\Delta = \mathrm{tr}\,(\mathbf{I} - \Sigma_{k'}^{-1}\Sigma_k) + \|\mu_o - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}} - \|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}\Sigma_{k'}\Sigma_k^{-1}}. \tag{6.17}$$

The first term in (6.17) is bounded since $\mathrm{tr}\,(\mathbf{I}-\Sigma_k^{-1}\Sigma_{k'}) = \mathrm{tr}\,(\Sigma_k^{-1}\mathbf{U}^T\tilde{\mathbf{Z}}\tilde{\Gamma}\tilde{\mathbf{Z}}^T\mathbf{U}) = \mathrm{tr}\,(\tilde{\Gamma}^{1/2}\tilde{\mathbf{Z}}^T\mathbf{U}\Sigma_k^{-1}\mathbf{U}^T\tilde{\mathbf{Z}}\tilde{\Gamma}^{1/2}) \leq \mathrm{tr}\,(\tilde{\Gamma}^{1/2}\tilde{\Gamma}\tilde{\Gamma}^{1/2}) = \tilde{p}$ where the inequality follows from applying (A3) after substituting for $\Sigma_k$ using (6.16). The second term is bounded by Lemma 6.2 and the third term is bounded by Lemma 6.1 since $\Sigma_{k'} \leq \Sigma_k$ implies $\|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}\Sigma_{k'}^{-1}\Sigma_k^{-1}} \leq \|\mu_o - \mu_k\|_{\Sigma_k^{-1}}$.

Setting $i = k'$ in (6.10) and simplifying,

$$\begin{aligned}\mathbf{V}_{k'}\Delta \;\leq\; 4\Big[&\mathrm{tr}\,[(\mathbf{I} - \Sigma_{k'}^{-1}\Sigma_k)^2] + 2\|\mu_o - \mu_{k'}\|^2_{(\Sigma_{k'}^{-1}-\Sigma_k^{-1})\Sigma_{k'}^{-1}(\Sigma_{k'}-\Sigma_k^{-1})} \\ &+\|\mu_{k'} - \mu_k\|^2_{\Sigma_k^{-1}\Sigma_{k'}\Sigma_k^{-1}}\Big].\end{aligned} \tag{6.18}$$

The first term is bounded since $\mathrm{tr}\,(\mathbf{A}^2) \leq (\mathrm{tr}\,(\mathbf{A}))^2$. Since $(\Sigma_{k'}^{-1}-\Sigma_k^{-1})\Sigma_{k'}^{-1}(\Sigma_{k'}-\Sigma_k^{-1}) = \Sigma_{k'}^{-1/2}(\mathbf{I} - (\mathbf{I} + \mathbf{H})^{-1})\Sigma_{k'}^{-1/2} \leq \Sigma_{k'}^{-1}$, applying Lemma 6.2 shows that the second term is bounded. Finally, the third term is bounded by Lemma 6.1 and the fact $\Sigma_{k'} \leq \Sigma_k$. ∎

**Theorem 6.2** *If model $k$ is true, then (6.13) is a necessary condition for $\alpha_k \to 1$.*

**Proof** Since $\alpha_k \to 1 \iff (-2\log M) \to \infty$, it is sufficient to show that both $\mathbf{E}_k(-2\log M)$ and $\mathbf{V}_k(-2\log M)$ are bounded when (6.13) fails to hold. This

task reduces to showing that both $\mathbf{E}_k\Delta$ and $\mathbf{V}_k\Delta$ are bounded since it has already been seen that $\log(D)$ is bounded if (6.13) fails to hold. So suppose (6.13) does not hold.

Setting $i = k$ in (6.9) and simplifying, we have

$$\mathbf{E}_k\Delta = \mathrm{tr}\,(\mathbf{H}) + \|\mu_o - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}\Sigma_k\Sigma_{k'}^{-1}} - \|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}}. \tag{6.19}$$

The first term is bounded by supposition and the third term is bounded according to Lemma 6.1. To show that the second term is bounded, let $c = 1 + \lambda_{\max}(\mathbf{H})$ and observe that $\Sigma_{k'}^{-1}\Sigma_k\Sigma_{k'}^{-1} = \Sigma_{k'}^{-1/2}(\mathbf{I} + \mathbf{H})\Sigma_{k'}^{-1/2} \leq c\Sigma_{k'}^{-1}$. Then

$$\begin{aligned}
\|\mu_o - \mu_{k'}\|_{\Sigma_{k'}^{-1}\Sigma_k\Sigma_{k'}^{-1}} &\leq c\|\mu_o - \mu_{k'}\|_{\Sigma_{k'}^{-1}} \\
&= c\|\mathbf{Z}_{k',(n)}(\beta_o^* - \mathbf{b}_{k'}) + \tilde{\mathbf{Z}}\tilde{\beta}\|_{\mathbf{U}\Sigma_{k'}^{-1}\mathbf{U}^T} \\
&\leq c\|\mathbf{Z}_{k',(n)}(\beta_o^* - \mathbf{b}_{k'})\|_{\mathbf{U}\Sigma_{k'}^{-1}\mathbf{U}^T} + c\|\tilde{\mathbf{Z}}\tilde{\beta}\|_{\mathbf{U}\Sigma_{k'}^{-1}\mathbf{U}^T} \\
&= c\|\beta_o^* - \mathbf{b}_{k'}\|_{\mathbf{Z}_{k',(n)}^T\mathbf{U}\Sigma_{k'}^{-1}\mathbf{U}^T\mathbf{Z}_{k',(n)}} + c\|\tilde{\beta}\|_{\mathbf{G}^T\mathbf{G}} \\
&\leq c\|\beta_o^* - \mathbf{b}_{k'}\|_{\Gamma_{k'}^{-1}} + c\|\tilde{\beta}\|_{\mathbf{G}^T\mathbf{G}}.
\end{aligned} \tag{6.20}$$

The second inequality follows from Cauchy-Schwarz and the third one from an application of (A3). Both of the terms in (6.21) are bounded by supposition.

Setting $i = k$ in (6.10) and simplifying, we have

$$\begin{aligned}
\mathbf{V}_k\Delta &\leq 4\left[\mathrm{tr}\,(\mathbf{H}^2) + 2\|\mu_o - \mu_{k'}\|^2_{(\Sigma_{k'}^{-1}-\Sigma_k^{-1})\Sigma_k(\Sigma_{k'}^{-1}-\Sigma_k^{-1})} + \|\mu_{k'} - \mu_k\|^2_{\Sigma_k^{-1}}\right] \\
&\leq 4\left[\mathrm{tr}\,(\mathbf{H}^2) + 2\|\mu_o - \mu_{k'}\|_{\Sigma_{k'}^{-1}\Sigma_k\Sigma_{k'}^{-1}} + \|\mu_{k'} - \mu_k\|^2_{\Sigma_k^{-1}}\right]
\end{aligned} \tag{6.21}$$

where the last inequality follows from the fact $(\Sigma_{k'}^{-1} - \Sigma_k^{-1})\Sigma_k(\Sigma_{k'}^{-1} - \Sigma_k^{-1}) \leq \Sigma_{k'}^{-1}\Sigma_k\Sigma_{k'}^{-1}$. All of the terms in in (6.21) are bounded by supposition or have already have been shown to be bounded. ∎

**Theorem 6.3** *If model k is true, then (6.13) together with the condition that*

$$\lambda_{\max}(\mathbf{G}^T\mathbf{G}) \leq K\lambda_{\min}(\mathbf{G}^T\mathbf{G}) \tag{6.22}$$

*for all n and some constant K is sufficient for* $\alpha_k \to 1$.

**Proof** It is sufficient to show that (6.13) and (6.22) together imply that $\mathbf{E}_k(-2\log M) \to \infty$ and $\mathbf{V}_k(-2\log M)/\mathbf{E}_k^2(-2\log M) \to 0$ since Chebyshev's inequality then implies $(-2\log M) \to \infty$.

Clearly, from (6.19),

$$\begin{aligned}\mathbf{E}_k(-2\log M) &= \operatorname{tr}(\mathbf{H}) + \|\mu_o - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}\Sigma_k\Sigma_{k'}^{-1}} - \|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}} - \log|\mathbf{I}+\mathbf{H}| \\ &\geq \|\mu_o - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}\Sigma_k\Sigma_{k'}^{-1}} \end{aligned} \tag{6.23}$$

since $\operatorname{tr}(\mathbf{H}) - \log|\mathbf{I}+\mathbf{H}| = \sum(\lambda_i(\mathbf{H}) - \log(1+\lambda_i(\mathbf{H})) \geq 0$. We now show that $\|\mu_o - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}\Sigma_k\Sigma_{k'}^{-1}}$ increases at a rate of at least $O\left(\lambda_{\min}((\mathbf{G}^T\mathbf{G})^2)\right)$. Since $\Sigma_{k'}^{-1}\Sigma_k\Sigma_{k'}^{-1} = \Sigma_{k'}^{-1} + \Sigma_{k'}^{-1}\mathbf{U}^T\tilde{\mathbf{Z}}\tilde{\Gamma}\tilde{\mathbf{Z}}^T\mathbf{U}\Sigma_{k'}^{-1}$, we have

$$\begin{aligned}\|\mu_o - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}\Sigma_k\Sigma_{k'}^{-1}} &\geq \|\mu_o - \mu_{k'}\|^2_{\Sigma_{k'}^{-1}\mathbf{U}^T\tilde{\mathbf{Z}}\tilde{\Gamma}\tilde{\mathbf{Z}}^T\mathbf{U}\Sigma_{k'}^{-1}} \\ &= \|\beta_o^*\|^2_{\mathbf{Z}^T_{k',(n)}\mathbf{U}\Sigma_{k'}^{-1/2}\mathbf{H}\Sigma_{k'}^{-1/2}\mathbf{U}^T\mathbf{Z}_{k',(n)}} + \|\tilde{\beta}\|^2_{\mathbf{G}^T\mathbf{G}\tilde{\Gamma}\mathbf{G}^T\mathbf{G}} \\ &\quad -2\beta_o^{*T}\mathbf{Z}_{k',(n)}\Sigma_{k'}^{-1}\mathbf{U}^T\tilde{\mathbf{Z}}\tilde{\Gamma}\mathbf{G}^T\mathbf{G}\tilde{\beta}. \end{aligned} \tag{6.24}$$

The second term in (6.24) is at least $O\left(\lambda_{\min}((\mathbf{G}^T\mathbf{G})^2)\right)$ since

$$\begin{aligned}\|\tilde{\beta}\|^2_{\mathbf{G}^T\mathbf{G}\tilde{\Gamma}\mathbf{G}^T\mathbf{G}} &\geq \lambda_{\min}(\tilde{\Gamma})\|\tilde{\beta}\|^2_{(\mathbf{G}^T\mathbf{G})^2} \\ &\geq \lambda_{\min}(\tilde{\Gamma})\lambda_{\min}((\mathbf{G}^T\mathbf{G})^2)\|\tilde{\beta}\|^2. \end{aligned} \tag{6.25}$$

In contrast, the first term in (6.24) increases at a rate of at most $O\left(\lambda_{\min}(\mathbf{G}^T\mathbf{G})\right)$ since

$$\begin{aligned}\|\beta_o^*\|^2_{\mathbf{Z}^T_{k,(n)}\mathbf{U}\Sigma_{k'}^{-1/2}\mathbf{H}\Sigma_{k'}^{-1/2}\mathbf{U}^T\mathbf{Z}_{k,(n)}} &\leq \lambda_{\max}(\mathbf{G}\tilde{\Gamma}\mathbf{G}^T)\|\beta_o^*\|^2_{\mathbf{Z}^T_{k',(n)}\mathbf{U}\Sigma_{k'}^{-1}\mathbf{U}^T\mathbf{Z}_{k',(n)}} \\ &\leq \lambda_{\max}(\tilde{\Gamma})\lambda_{\max}(\mathbf{G}^T\mathbf{G})K' \\ &\leq \lambda_{\max}(\tilde{\Gamma})K\lambda_{\min}(\mathbf{G}^T\mathbf{G})K' \end{aligned} \tag{6.26}$$

where condition (6.22) has been used in the last inequality. Hence the second term in (6.24) dominates the first term. By the Cauchy-Schwarz inequality, the the first term also dominates the third term. Therefore $\mathbf{E}_k(-2\log M) \to \infty$ at a rate of at least $O\left(\lambda_{\min}((\mathbf{G}^T\mathbf{G})^2)\right)$.

To show that $\mathbf{V}_k(-2\log M)$ increases at a rate of at most $O(\lambda_{\min}((\mathbf{G}^T\mathbf{G})^2)$, it is sufficient to show that the first term in (6.21) is at most $O(\lambda_{\min}((\mathbf{G}^T\mathbf{G})^2)$ since it has been shown already that the third term in (6.21) is bounded and that the second term is at most $O\left(\lambda_{\min}((\mathbf{G}^T\mathbf{G})^2)\right)$. But $\mathrm{tr}\,(\mathbf{H}^2) \leq (\mathrm{tr}\,(\mathbf{H}))^2 \leq \left(\lambda_{\max}(\tilde{\Gamma})\mathrm{tr}\,(\mathbf{G}^T\mathbf{G})\right)^2 \leq \left(\lambda_{\max}(\tilde{\Gamma})\tilde{p}\lambda_{\max}(\mathbf{G}^T\mathbf{G})\right)^2 \leq \left(\lambda_{\max}(\tilde{\Gamma})\tilde{p}K\lambda_{\min}(\mathbf{G}^T\mathbf{G})\right)^2$ is clearly $O\left(\lambda_{\min}((\mathbf{G}^T\mathbf{G})^2)\right)$. ∎

We conjecture that an alternative simpler condition to (6.22) is that $J \geq \tilde{p}$ since we believe this condition implies (6.22). (Clearly $J \geq \tilde{p}$ is a necessary condition for (6.22) since otherwise $\mathbf{G}^T\mathbf{G}$ has a zero eigenvalue.) Another perhaps even more attractive alternative would be to remove (6.22) and instead impose a condition on $(\mathbf{Z}_{k',(n)}, \tilde{\mathbf{Z}})$ directly. So long as $\mathbf{G}^T\mathbf{G}$ is full rank, the lower bound obtained in (6.25) obtains only for exceptional sequences of $(\mathbf{Z}_{k',(n)}, \tilde{\mathbf{Z}})$. For typical sequences, the lower bound is $O(\lambda_{\max}((\mathbf{G}^T\mathbf{G})^2)$ so if we exclude the exceptional sequences from consideration, no additional conditions need be added (beyond requiring $\mathbf{G}^T\mathbf{G}$ to be full rank). Condition (6.22) aside, Theorems 6.1 through 6.3 essentially state that weak consistency is characterized by whether or not $\mathbf{U}$ satisfies condition (6.13).

The remaining material in this subsection gives a Wald-type argument that gives sufficient conditions for strong consistency.

**Conjecture 6.1** *A sufficient condition for $\alpha_i \to 1$ a.s. when model $i$ is true for both $i = k$ and $i = k'$ is that*

$$\lim_{n\to\infty} \mathrm{rank}(\mathbf{U}) \to \infty. \tag{6.27}$$

**Sketch of proof when model $k$ true:** Let $\Omega_{k'}$ and $\Omega_k$ denote the parameter spaces under models $k'$ and $k$ respectively. We need to show that (6.27) implies that for any $\epsilon > 0$, $L \equiv P_{\beta_o}(M > \epsilon) \to 0$ when $\beta_o \in \Omega_k$ but $\beta_o \notin \Omega_{k'}$. The density of $\mathbf{S}_n^\alpha$ given the parameter value $\beta_i$ under model $i$ is

$$p_i = (2\pi)^{(-J/2)}|\sigma^2\mathbf{U}^T\mathbf{U}|^{-1/2}\exp\left\{-\frac{1}{2}\|\mathbf{s}_{(n)} - \mathbf{U}^T\mathbf{Z}_{i,(n)}\beta_i\|^2_{(\sigma^2\mathbf{U}^T\mathbf{U})^{-1}}\right\}. \quad (6.28)$$

Let the $p_o$ denote the density $p_i$ evaluated at the true parameter value $\beta_o$. Let $B(\beta_o, \delta)$ denote the open ball centred at $\beta_o$ with radius $\delta$ such that $B(\beta_o, \delta) \cap \Omega_{k'} = \emptyset$ and let $B^c(\beta_o, \delta)$ denote its complement.

Let $\nu_i(\beta_i)$ denote the prior density on $\beta_i$. Then

$$M = \frac{\int_{\Omega_{k'}} \nu_{k'}(\beta_{k'})p_{k'}d\beta_{k'}}{\int_{\Omega_k} \nu_k(\beta_k)p_k d\beta_k} \quad (6.29)$$

and

$$\begin{aligned} L &= P_{\beta_o}(M > \epsilon) \\ &\leq P_{\beta_o}\left(\frac{\sup_{B^c(\beta_o,\delta)} p_k}{p_o}\frac{p_o}{\int_{B(\beta_o,\delta)} \nu_k(\beta_k)p_k d\beta_k} > \epsilon\right) \\ &\leq P_{\beta_o}\left(\frac{\sup_{B^c(\beta_o,\delta)} p_k}{p_o} > e^{-n\alpha}\right) + P_{\beta_o}\left(\int_{B(\beta_o,\delta)} \nu_k(\beta_k)p_k d\beta_k < \frac{e^{-n\alpha}}{\epsilon}p_o\right). \end{aligned} \quad (6.30)$$

The first inequality follows from recognizing that $\sup_{B^c(\beta_o,\delta)} p_k \geq \sup_{\Omega_{k'}} p_k = \sup_{\Omega_{k'}} p_{k'} \geq \int_{\Omega_{k'}} \nu_{k'}(\beta_{k'})p_{k'}d\beta_{k'}$ and $\int_{B(\beta_o,\delta)} \nu_k(\beta_k)p_k d\beta_k \leq \int_{\Omega_k} \nu_k(\beta_k)p_k d\beta_k$. The second inequality follows from first using the relation $E = (E \cap A) \cup (E \cap A^c) \subset A \cup (E \cap A^c)$ where $E = \{M > \epsilon\}$ and

$$A = \left\{\frac{\sup_{B^c(\beta_o,\delta)} p_k}{p_o} > e^{-n\alpha}\right\} \quad (6.31)$$

and then applying the relation $\{xy > p\} \cap \{x \leq q\} \subset \{qy > p\}$ to $E \cap A^c$ and re-arranging terms. Now let $\mathbf{W}_n = \mathbf{A}^{-1/2}\mathbf{S}_n^\alpha$ where $\mathbf{A} = \sigma^2\mathbf{U}^T\mathbf{U}$. Then the

elements of $\mathbf{W}_n$ are independent normals with unit variance and hence $p_k$ as a function of $\mathbf{W}_n$ is given by

$$\begin{aligned} p_k &= (2\pi)^{-J/2}|\sigma^2\mathbf{U}^T\mathbf{U}|^{-1/2}\exp\left\{-\frac{1}{2}\|\mathbf{w}_n - \mathbf{A}^{-1/2}\mathbf{U}^T\mathbf{Z}_{k,(n)}\beta_k\|^2\right\} \\ &= (2\pi)^{-J/2}|\sigma^2\mathbf{U}^T\mathbf{U}|^{-1/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{J}(w_i - g_i(\beta_k))^2\right\} \end{aligned} \tag{6.32}$$

where $w_i$ and $g_i(\beta_k)$ are the $i$-th elements of $\mathbf{w}_n$ and $\mathbf{A}^{-1/2}\mathbf{U}^T\mathbf{Z}_{k,(n)}\beta_k$ respectively. As a function of $\mathbf{w}_n$, the expression for $p_k$ is not a density due to the additional factor $|\sigma^2\mathbf{U}^T\mathbf{U}|^{-1/2}$. However this factor appears for each instance of $p_k$ and cancel in each term in (6.30). Hence therein $p_k$ can be treated as a true density. If $g_i(\beta_k)$ were independent and identically distributed (i.i.d.) for all $i$, then as $J \to \infty$, the first term of (6.30) converges to 0 by Wolfowitz's (1949) result and the second term is $O(1/J)$ by a result of Clarke and Barron (1990). Hence if these two results were extended to cover the case of independent and not identically distributed (i.n.i.d.) data as arising in this problem, then the theorem is proved. The extension from i.i.d. to i.n.i.d. is usually routine although often tedious and involving many conditions. For example, Hoadley (1971) extends Wolfowitz's result by generalizing standard regularity conditions. The extension of the Clarke and Barron work is expected to be similar. ∎

**Sketch of proof when model $k'$ true:** Let $B(\beta_o, \delta_1, \delta_2)$ denote an open rectangular neighborhood of $\beta_o$ formed as the Cartesian product of $B(\beta_o, \delta_1)$ and $B(\beta_o, \delta_2)$ where $B(\beta_o, \delta_1)$ and $B(\beta_o, \delta_2)$ are open rectangular neighborhoods of $\beta_o$ in the subspaces of $\beta_{k'}$ and $\tilde{\beta}$ respectively. Let $\delta_1$ ($\delta_2$) denote the length of each side of of the rectangle. Let $v(B)$ denote the volume of the neighborhood $B$. Then

$$v(B(\beta_o, \delta_1, \delta_2)) = v(B(\beta_o, \delta_1))v(B(\beta_o, \delta_2)). \tag{6.33}$$

Suppose for the moment that $\forall \delta_1, \delta_2$ there exists a sufficiently large $N$ such that

$$\int_{B(\beta_o,\delta_1,\delta_2)} \nu_k p_k d\beta_k \geq \int_{B^c(\beta_o,\delta_1,\delta_2)} \nu_k p_k d\beta_k \tag{6.34}$$

for all $n > N$. Then for sufficiently large $n$,

$$\begin{aligned} M &= \frac{\int_{\Omega_{k'}} \nu_{k'}(\beta_{k'}) p_{k'} d\beta_{k'}}{\int_{\Omega_k} \nu_k(\beta_k) p_k d\beta_k} \\ &\geq \frac{v(B(\beta_o,\delta_1)) \int_{B(\beta_o,\delta_1,\delta_2)} \frac{\nu_{k'} p_{k'}}{v(B(\beta_o,\delta_1))} d\beta_{k'}}{2v(B(\beta_o,\delta_1,\delta_2)) \int_{B(\beta_o,\delta_1,\delta_2)} \frac{\nu_k p_k}{v(B(\beta_o,\delta_1,\delta_2))} d\beta_k} \\ &= \frac{1}{v(B(\beta_o,\delta_2))} \frac{\int_{B(\beta_o,\delta_1,\delta_2)} \frac{\nu_{k'} p_{k'}}{v(B(\beta_o,\delta_1))} d\beta_{k'}}{\int_{B(\beta_o,\delta_1,\delta_2)} \frac{\nu_k p_k}{v(B(\beta_o,\delta_1,\delta_2))} d\beta_k}. \end{aligned} \tag{6.35}$$

As $\delta_1$ and $\delta_2$ go to 0, the integrals in the numerator and denominator of (6.35) both converge to $p_k(\beta_o)$ and hence $M \to \infty$. So the theorem is proved if it is shown that (6.27) implies (6.34). Roughly speaking, 6.34 is requiring that the density concentrates around $\beta_o$ as $n$ gets large. If we were dealing with i.i.d. observations then (6.27) would be sufficient so again the key is an extension to i.n.i.d. data. ∎

For a model choice strategy, consistency of the model weights guarantees that the final prediction matches the Bayes predictor from the true model asymptotically. Roughly speaking, if model $i$ is correct (and so $\alpha_i \to 1$), then asymptotically $\rho^*(k; \mathbf{S}_n^\rho) \approx \rho(k; i, \mathbf{S}_n^\rho)$ and obviously $\rho(k; i, \mathbf{S}_n)$ is minimized by taking $k = i$.

## 6.2 Conditional Predictive Distributions

Consistency of model weights in a model averaging procedure does not guarantee that the final prediction matches the Bayes predictor from the correct model in general. In model averaging the final prediction is derived from $F_{i|\mathbf{S}_n^\rho}$

(when model $i$ is true) whereas the Bayes predictive distribution is $F_{i|\mathbf{Y}_{(n)}}$. For point prediction, these two distributions will generate the same predictor asymptotically if and only if the means $\mu(F_{i|\mathbf{S}_n^\rho})$ and $\mu(F_{i|\mathbf{Y}_{(n)}})$ for the two distributions are equal asymptotically, i.e.,

$$\Delta_\mu \equiv \mu(F_{i|\mathbf{Y}_{(n)}}) - \mu(F_{i|\mathbf{S}_n^\rho}) \to 0. \tag{6.36}$$

A generalization of Lemma 3.1 provides one characterization of the solution to 6.36.

**Theorem 6.4** *Asymptotically, the distributions $F_{i|\mathbf{S}_n^\rho}$ and $F_{i|\mathbf{Y}_{(n)}}$ have the same mean iff the vector $\Psi_i^{-1}\mathbf{Z}_{i,(n)}\Gamma_i\mathbf{Z}_{i,n+1}$ lies in the null space of $\mathbf{U}$.*

**Proof** Substituting in (6.36) using (3.8) and (3.20), and simplifying, we get

$$\begin{aligned}\Delta_\mu &= (\mathbf{Y}_{(n)} - \mathbf{Z}_{i,(n)}\mathbf{b}_i)^T[\Psi_i^{-1} - \mathbf{U}(\mathbf{U}^T\Psi_i\mathbf{U})^{-1}\mathbf{U}^T]\mathbf{Z}_{i,(n)}\Gamma_i\mathbf{Z}_{i,n+1} \\ &= (\mathbf{Y}_{(n)} - \mathbf{Z}_{i,(n)}\mathbf{b}_i)^T(\mathbf{I} - \mathbf{P})\Psi_i^{-1}\mathbf{Z}_{i,(n)}\Gamma_i\mathbf{Z}_{i,n+1} \qquad (6.37)\end{aligned}$$

where $\mathbf{P}$ is given by (3.25). By the argument used in the proof of Lemma 3.1, $\Delta_\mu = 0$ iff the condition in the statement of the theorem holds. ∎

# Chapter 7

# Discussion

We have proposed a new class of criteria, the mongrel risk, for selecting online predictors. The mongrel risk combines both model information and past empirical performance to evaluate the candidate predictors. The application of the mongrel risk requires a rule for selecting the conditioning statistic $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. The selection can be made in an adaptive or global approach.

Our simulations show that an adaptive mongrel approach beats out the Bayes procedure uniformly over a wide range of data-generators in small samples. An analytic proof of this result would, of course, be desirable but it is unlikely to be obtainable given that the expressions needed to assess the performance of a mongrel procedure cannot be evaluated analytically. Moreover, because we are dealing strictly with small sample performance, we cannot appeal to the approximation techniques that are often employed in proving asymptotic results.

We believe that we have investigated a sufficient variety of simulation parameters to conclude that our results hold in enough generality to be compelling. The choices for the coefficient, $\gamma_2$, of $X_2$ cover a practically meaningful

range of values and the choices for the prior probability, $\alpha_{2,o}$ on the full model span a range that seems reasonable for practical applications.

We also investigated the impact of the choice of prior distributions on the parameters. For our main simulation work, we set the prior variances, $\Gamma_1$ & $\Gamma_2$, on the regression parameters to be identity matrices because we felt that such values reflected the typical (mild to moderate) amount of prior information available in practice. In simulation results not presented here, the use of very weak priors ($\Gamma_i = 25\mathbf{I}$) gave qualitatively the same results. Our choice of prior means on the parameters may also seem fortuitous in that when $\gamma_2$ is 0 or 0.4, one of the models will have prior means that completely match the data-generator. In practice, we would expect that neither of the prior means would match the means in the data-generator. Hence, we also conducted simulations in which a different set of coefficients for the data-generator was generated for each sequence randomly. That is, we used a random effects model for generating the data. The distribution of the random coefficients was normal with means equal to the values from the fixed coefficients model and $0.2\mathbf{I}$ as the variance. Once again, we found that the results (not presented here) were qualitatively similar to what we found using the fixed coefficients model.

As an alternative to using the MSPE (3.34), we evaluated the performance of the naive mongrel strategies 'a2hf' and 'c2hf' using the difference in relative entropies

$$\Delta D \equiv D(p_T \| p_B) - D(p_T \| p_M), \tag{7.1}$$

where $p_T$ is the true predictive density (from the data-generator), $p_B$ is the density from taking a Bayes strategy, and $p_M$ is the density from the mongrel strategy. We limited the investigation to scenarios with $\alpha_{2,o} = 0.5$ and $\gamma_2 = 0$,

0.2, or 0.4. The plots (not shown) of the $\Delta D$ in all three cases exhibited patterns very similar to their analogs under MSPE loss (the third panel in Figures 3.4, 3.5, 3.6 (model averaging) or 3.13, 3.14, 3.15 (model choice)). The mongrel strategies beat out or lost out to the Bayes strategy in the same scenarios and across roughly the same time intervals regardless of whether relative entropy or MSPE loss was used. These results suggest that the gains seen from taking a mongrel approach generalize to loss functions other than MSPE (though this claim is perhaps mitigated by the use of normal models for which relative entropy is closely related to squared error loss).

Relative entropy can also be used to select $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. Consider the case of model averaging. The candidate predictive distributions are of the form of the mixture distribution (2.19) for different choices of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$. Let $p_{m|(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)}$ denote the density of the mixture distribution. The adequacy of $p(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ as a predictive distribution in relative entropy distance is $D(p_T \| p_{m|(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)})$. We cannot use this measure to compare different choices of $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ because the true density $p_T$ is unknown. Instead, we can compare them based on the relative entropy distance with respect to the posterior density $p_{i|\mathbf{Y}_{(n)}}$ of each candidate model, i.e.,

$$D_i \equiv D(p_{i|\mathbf{Y}_{(n)}} \| p_{m|(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)}) \tag{7.2}$$

for each candidate model $i$. As in Chapter 4, we might summarize over all models by taking the maximum or taking a weighted average over the $D_i$'s. This leads to select $(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)$ as

$$\mathbf{S}_{\text{mM}} \equiv \arg \min_{(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)} \max_i D_i \tag{7.3}$$

or

$$\mathbf{S}_{\text{mWA}} \equiv \arg \min_{(\mathbf{S}_n^\alpha, \mathbf{S}_n^\rho)} \sum_i \omega_i D_i \tag{7.4}$$

for a minimax or weighted average strategy (with weights $\omega_i$) respectively. The choice of predictive distribution is then given by $p_{m|\mathbf{S}_{\mathrm{mM}}}$ or $p_{m|\mathbf{S}_{\mathrm{mWA}}}$. In general, $D_i$ cannot be evaluated analytically. However, it's evaluation requires only a one-dimensional integral so numerical integration is feasible. Relative entropy is generally viewed as a "natural" measure of loss so it may be preferred over squared error loss outside of normal models.

Both practical and conceptual difficulties have limited our work to only two candidate models. On a practical level, the computational time required to complete a simulation increases roughly as the square of the number of candidate models so that going from two to three models would more than double the time needed. More importantly, we have yet to develop fully the idea of risk sufficiency so that it is unclear which predictuals should be considered for inclusion in $(\mathbf{S}_n^{\alpha}, \mathbf{S}_n^{\rho})$. For instance, should we still use only the predictuals from the smallest model? a combination of predictuals from all but the largest model? or some entirely different set of predictuals? So long as $(\mathbf{S}_n^{\alpha}, \mathbf{S}_n^{\rho})$ is not risk-sufficient, the generalization of most of our techniques to three or more models is straightforward. But based on our experience with two models, it seems not too difficult to unintentionally achieve risk-sufficiency. We view the development of a greater understanding of risk-sufficiency as important future work.

Our arguments for why appropriate use of the mongrel risk criterion generates more accurate predictions than the Bayes procedure have been, for the most part, heuristic. Much additional work is needed to explain fully the properties of the mongrel procedure that give it the advantage. We are currently examining the results for individual sequences to identify the circumstances under which the mongrel approach is better. For example, Figure 7.1 displays a histogram of the difference MSPE(a2ff) − MSPE(a2xfmM) in

performance between the Bayes averaging strategy (a2ff) and the minimax mongrel averaging strategy for the 5000 individual sequences. (The predictions are for $Y_{10}$ with simulation parameters $\gamma_2 = 0.2$ and $\alpha_{2,o} = 0.5$.) The distribution is tightly concentrated about zero but nearly all of the large deviations are positive. Thus it appears that the mongrel and Bayes strategies typically perform about equally well, but in a small fraction of sequences the mongrel procedure performs much better. This result agrees well with our intuition that the mongrel approach is good at identifying the exceptional sequences where using all of the data produces misleading risk assessments. Our efforts in this area are ongoing.

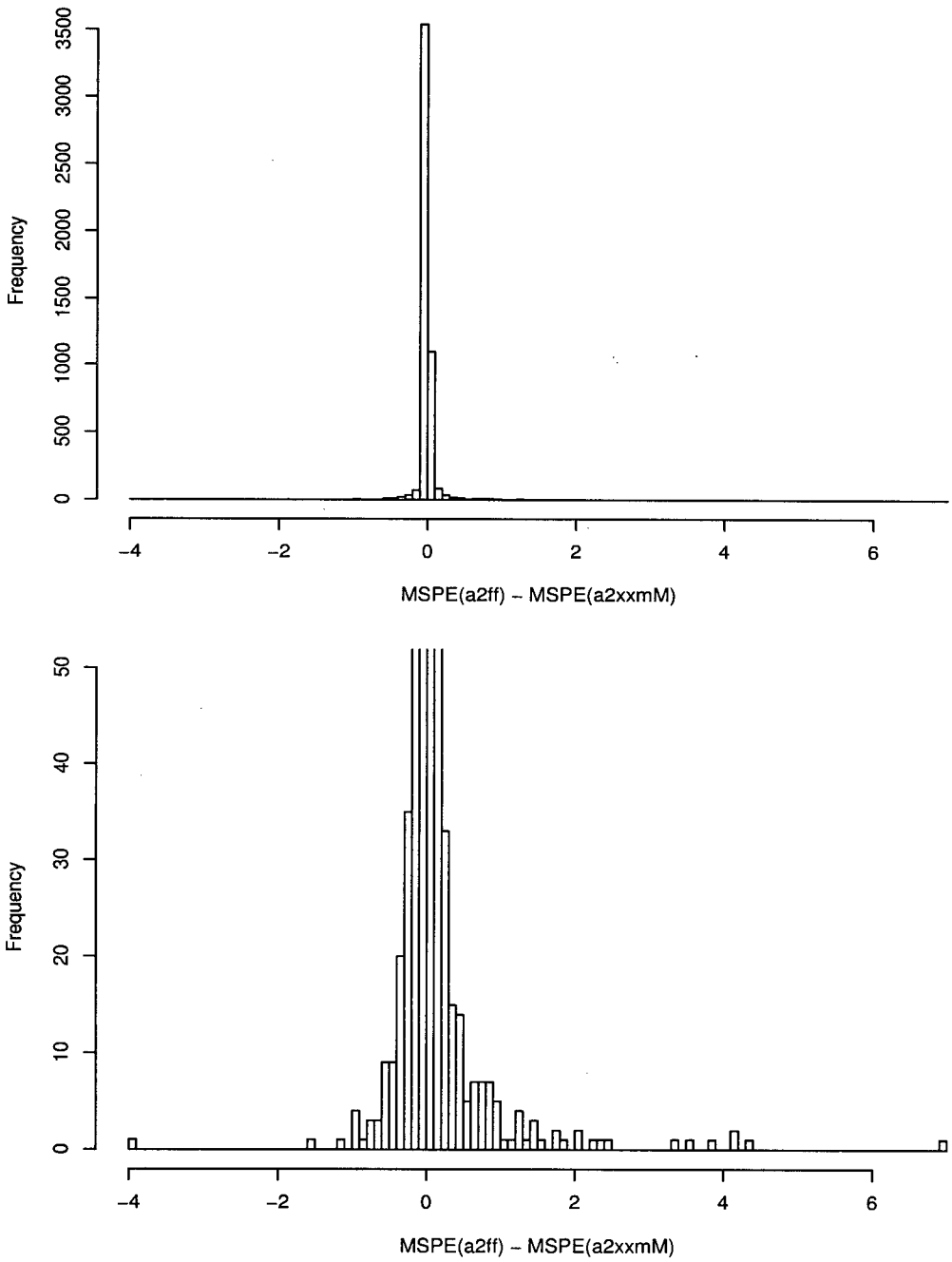


Figure 7.1: Distribution of MSPE(aff) − MSPE(a2xfmM) for predicting $Y_{10}$ with simulation parameters $a_{2,o} = 0.5$, $\gamma_2 = 0.2$. (Bottom panel displays the smaller frequency range on an expanded scale.)

# Bibliography

Allen, D.M. (1974). The relationship between variable selection and prediction. *Technometrics*, **16**, 125-127.

Berger, J.O. (1985), Statistical Decision Theory and Bayesian Analysis (2nd ed.), New York, Springer-Verlag.

Clarke, B.S. and Barron, A.R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory*, **36**, 453-471.

Clyde, M.A. (1999). Bayesian model averaging and model search strategies. In J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds), Bayesian Statistics 6, 157-185. Oxford University Press.

Dawid, A.P. (1984). Statistical theory: the prequential approach (with discussion). *J. Roy. Statist. Soc. A*, **147**, 278-292.

Dawid, A.P. (1992). Prequential Data Analysis. In M. Ghosh and P.K. Pathak (eds), *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, 113-126. IMS Lecture Notes — Monograph Ser. 17. Hayward, CA, Institute of Mathematical Statistics.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Ann. Math. Statist.*, **42**, 1977-1991.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statist. Sci.*, **14**, 382-417.

Raftery, A.E., Madigan, D. and Hoeting, J.A. (1997). Bayesian model averaging for linear regression models. *J. Am. Statist. Assoc.*, **92**, 179-191.

Seillier-Moiseiwitsch, F. and Dawid, A.P. (1993). On testing the validity of sequential probability forecasts. *J. Am. Statist. Assoc.*, **88**, 355-359.

Skouras, K. (1998). On the Optimal Performace of Forecasting Systems: The Prequential Approach. Ph.D. Thesis, Department of statistical Science, University College London.

Wolfowitz, J. (1949). On Wald's proof of the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, **20**, 601-602

# Appendix

**A.1** Let $\mathbf{A}$ and $\mathbf{D}$ be nonsingular matrices. Then

$$\left(\mathbf{A}+\mathbf{BDB}^T\right)^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D}^{-1}+\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{A}^{-1}. \tag{A1}$$

Rearranging this equality and left and right multiplying by $\mathbf{A}$, we obtain

$$\mathbf{B}\left(\mathbf{D}^{-1}+\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T+\mathbf{A}\left(\mathbf{A}^{-1}+\mathbf{BDB}^T\right)^{-1}\mathbf{A}=\mathbf{A} \tag{A2}$$

so that, if $\mathbf{A}$ and $\mathbf{D}$ are p.d., then each term in (A2) is n.n.d. and hence

$$\mathbf{B}\left(\mathbf{D}^{-1}+\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T\leq\mathbf{A}. \tag{A3}$$

**A.2** Let $\mathbf{Y}\sim\mathcal{N}(\mathbf{m},\mathbf{S})$. Then for any n.n.d. matrix $\mathbf{A}$

$$\mathbf{E}(\mathbf{Y}^T\mathbf{AY}) = \mathrm{tr}\,(\mathbf{AS})+\mathbf{m}^T\mathbf{Am}, \tag{A4}$$

$$\mathbf{V}(\mathbf{Y}^T\mathbf{AY}) = 2\mathrm{tr}\,(\mathbf{ASAS})+4\mathbf{m}^T\mathbf{ASAm}. \tag{A5}$$

Note that for any n.n.d. matrix $\mathbf{B}$, $\mathrm{tr}\,(\mathbf{B}^2)=\sum_i\sum_j b_{ij}^2\leq\sum_i\sum_j b_{ii}b_{jj}=(\mathrm{tr}\,\mathbf{B})^2$ where $b_{ij}$ is the $i,j$-th element of $\mathbf{B}$. Setting $\mathbf{B}=\mathbf{A}^{1/2}\mathbf{SA}^{1/2}$, we have $\mathrm{tr}\,(\mathbf{ASAS})=\mathrm{tr}\,(\mathbf{B}^2)\leq(\mathrm{tr}\,\mathbf{B})^2=(\mathrm{tr}\,\mathbf{AS})^2$ and hence

$$\mathbf{V}(\mathbf{Y}^T\mathbf{AY})\leq 2(\mathrm{tr}\,\mathbf{AS})^2+4\mathbf{m}^T\mathbf{ASAm}. \tag{A6}$$

**A.3** Let $\rho_i = \zeta_i + \left(\gamma_i + \delta_i^T \mathbf{Y}\right)^2$, where $\zeta_i$, $\gamma_i$ are scalars and $\delta_i$, $\mathbf{Y}$ are $n$-vectors. If $r \leq n$, then for any scalars $\nu_i$,

$$\sum_{i=1}^{r} \nu_i \rho_i = (\mathbf{Y} + \mathbf{b})^T \mathbf{A} (\mathbf{Y} + \mathbf{b}) + c \tag{A7}$$

where

$$\mathbf{A} = \sum_{i=1}^{r} \nu_i \delta_i \delta_i^T, \tag{A8}$$

$$\mathbf{b} = \mathbf{A}^- \sum_{i=1}^{r} \gamma_i \nu_i \delta_i \tag{A9}$$

$$c = \sum_{i=1}^{r} \nu_i \left(\zeta_i + \gamma_i^2\right) - \mathbf{b}^T \mathbf{A} \mathbf{b} \tag{A10}$$

and $\mathbf{A}^-$ is a generalized inverse of $\mathbf{A}$. Note also that $\mathbf{Ab} = \sum_{i=1}^{r} \gamma_i \nu_i \delta_i$.