TOPICS IN MODEL AVERAGING & TOXICITY MODELS IN COMBINATION THERAPY

BY

PRIYAM MITRA

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Statistics & Biostatistics

Written under the direction of

Dr. Min-ge Xie

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2015

ABSTRACT OF THE DISSERTATION

Topics In Model Averaging & Toxicity Models In Combination Therapy

By PRIYAM MITRA

Dissertation Director:

Dr. Min-ge Xie

In this dissertation we work on two problems. In the first problem we propose a general framework for frequentist model averaging and explore its applications. In the second problem, we propose an adaptive design using a copula model that helps us analyze data from drug combination therapy. It is shown later that these new methods are more efficient than the existing methods.

Model selection methods often ignore the uncertainty introduced in the selection process and there always remains the possibility that the selected model can possibly be a wrong one. A model averaging approach addresses this issue by combining estimators for a set of candidate models so that it incorporates the underlying model uncertainty. In Chapter 2 we establish a general frequentist model averaging framework that greatly broadens the scope of the existing methodologies under the frequentist model averaging development. We propose a set of weights to combine

the individual estimators so that the asymptotic mean squared error of the model average estimator is minimized. Results from simulations and real data analysis show the benefits of the proposed approach over traditional model selection approaches as well as existing model averaging methods.

The early phase clinical studies in drug development are focused on the toxicity and sometimes efficacy of a new treatment or a new combination of treatments. Often the aim is to identify a maximum tolerated dose (MTD), which is the maximum dose combination level that does not cause an unacceptable toxicity. In Chapter 3, we explore the combination of two treatments using a copula model. We combine the individual toxicity profiles of the treatments to develop the combination model framework. The theoretic framework is further extended to a combination of more than two treatments and combination of ordinal toxicity measures. A case study based on a combination oncology trial is presented to demonstrate the proposed dose finding strategy for combination therapy.

Acknowledgements

This thesis wouldn't see the light of day were it not for my advisor Dr. Min-ge Xie. I continue to be amazed by his intellectual acuity and dedication to research. Everything I have learnt and accomplished in these last few years have been through his guidance. His continual support and patience has been instrumental in seeing me through. My heartfelt gratitude to him.

Dr. Tianhui (Helen) Zhou and Dr. Yun Shen have been amazing and helpful mentors throughout my internship at Bristol-Myers Squibb. I thank them for their advice and help. I would also like to thank Dr. Katy Simonsen and Dr. Venkat Sethuraman for their guidance and support.

I would like to thank my dissertation committee members Dr. Dan Yang, Dr. Helen Zhou and Dr. Han Xiao for agreeing to be a part of the committee, their busy schedules notwithstanding. My thanks also to Dr. David Tyler, Dr Harold Sackrowitz, Dr Han Xiao and my Advisor Dr. Xie for appearing in my Ph.D. qualifying oral exam committee. I would like to thank our entire department of statistics and biostatistics and also all the graduate students in our department. They have all been a big part of my graduate life. Our department chair Dr. Regina Liu and our graduate director Dr. John Kolassa have always had an open door and a listening ear for everything I wanted talk about. I thank them for their help and advice.

I would also take a moment to remember Dr. Kesar Singh. When I first came to Rutgers, Dr. Singh would always inquire about my studies and general well being. He also provided a lot of encouragement about my research. His untimely passing was and still is a terrible shock. My heartfelt respect to him.

None of it would be possible without the love and sacrifice of my family. No measure of gratitude can capture the debt I owe to my mother Mrs. Sumitra Mitra for her love and support. I also thank my dad and the rest of my family in India. I would also like to thank my husband Ritwik for his love and support and for always being there for me.

Dedication

To Ma & Baba

and

to my dearest Husband Ritwik

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

With the advancement of modern science and technology, scientists are faced with the proliferation of large amounts of data on almost every facets of human life; from trivial to essential. The race to mine insights from this treasure trove of data is also under way. Statistical science plays a crucial role in this great venture by providing inferential tools that help make the procured knowledge more definitive. More specifically, the goal of statistics is, given a problem, to propose effective and interpretable models and provide statistical guarantees on the veracity of the same. In this dissertation we have tackled two problems and proposed methodologies to address those problems. The usefulness of our proposed methodology is justified through theoretical results and applications to real life datasets.

- The first problem relates to the idea of model selection. When there are several plausible models to choose from but no definite scientific rationale to dictate which one should be used, model selection methods have been used traditionally to determine a 'correct' model for data analysis. Once a model is chosen, further analysis proceeds as if the model selected is the true one. This practice ignores the uncertainty introduced in the process due to model selection, and can often lead to faulty inference. The key idea behind model averaging is that we do not fully accept a single model, then reject all other models, as is usually done with model selection. Instead, we will acknowledge our uncertainty regarding which model is truth, and combine all candidate models to some degree. We quantify our degree of belief, or the relative strength of the evidence in support

of each model, through use of numerical model weights. Thus model averaging incorporates model uncertainty during analysis and provides a solution to the problem faced in model selection. Our research on model averaging is motivated in part by a real life example on a prostate cancer study where the relationship between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy was investigated. We focus on combining different candidate models that uses a different set of clinical measures to predict the level of antigen. Our purpose is to improve upon existing methods and get more efficient results.

- The second problem relates to combination therapy in clinical trials. With the development of new drugs in pharmaceutical industry, clinical trials involving treatments that combine multiple drugs are being introduced. Multiple drugs used together can interact and may enhance the effectiveness of the treatment. The goal of combination therapy is to achieve better patient response, particularly for cancer patients who are non respondent to conventional single agent therapies. However a very important question remains regarding how to do the analysis so that we can extract the benefits from using individual drug information. Our main purpose is to develop an efficient way of combining the effects of multiple drugs.

To solve the first problem we propose a general framework that subsumes all existing frameworks and study asymptotic properties of model average estimators in that framework. In Hjort and Claeskens [2003] frequentist model averaging was developed under the assumption that all candidate models had to be within $O(1/\sqrt{n})$ distance of the true model. We remove this restrictive assumption and develop frequentist model averaging approaches under a much more general framework. Our model averaging scheme allows us to use all the potential candidate models available for analysis. We also discuss developing a set of weights that will help us to build

a combined model average estimator. The weights are based on mean square error of the model average estimator, which takes into account both bias and variance of the estimator. Specifically, a consistent estimate of the mean square error of the model average estimator is proposed, and the weights are chosen such that the MSE estimate is minimized. It is later shown that in most of the cases, weights that are chosen to combine the candidate models highlight the contribution of the true model. To examine the performance of the proposed estimator we compare it with existing model selection and model averaging methos, and show that the proposed method is most effective.

To solve the second problem we model the rates of toxicity of the combined drugs via a copula-type regression. The main advantage of using a copula model for analyzing combination data is the way such a model incorporates individual drug toxicity information. Often, before certain drugs are combined, the toxicity profile of each individual drug is investigated in detail. Hence, one usually has rich prior information about the individual drugs from some early clinical or pre clinical data. This data could help with determining the marginal toxicity profile of the drugs, which would, in turn make the combined model more efficient. In this paper, we explore the possibilities of utilizing the prior information to obtain the individual toxicity probabilities using an adaptive design with a hierarchical Bayesian model. We show that this approach can achieve fast and accurate estimation of the drug- drug interaction pattern. We also develop a strategy for dose escalation and also explore the dose-toxicity space for proposing future dose levels. The proposed method is examined in a simulation study as well as in a case study that is developed by using the data from oncology clinical trials.

Chapter 2
Frequentist Model Averaging

2.1 Introduction

When there are several plausible models to choose from but no definite scientific rationale to dictate which one should be used, a model selection method has been used traditionally to determine a 'correct' model for data analysis. Commonly used model selection methods (such as step wise regression, AIC, BIC, etc.; c.f., Hastie and Tibshirani [2005]) are data driven and different methods may use different criteria. Once a model is chosen, further analysis proceeds as if the model selected is the true one. This practice ignores the uncertainty introduced in the process due to model selection, and can often lead to faulty inference as discussed in Madigan et al. [1994], Draper [1995] and Buckland et al. [1997]. Model averaging methods have been introduced to incorporate model uncertainty during analysis and to provide a solution to the problem [cf., Claeskens and Hjort, 2008]. Instead of deciding which one model is the 'correct' one, a model averaging method uses a set of plausible candidate models and final measures of inference are derived from a combination of all the models. The candidate models are combined using some data-dependent weights to reflect the degree to which each candidate model is trusted.

Our research on model averaging is motivated in part by a real life example on a prostate cancer study where the relationship between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy was investigated. The variables included in the study are log cancer

volume, log prostate weight, age , log of the amount of benign prostatic hyperplasia, seminal vesicle invasion, log of capsular penetration, Gleason score, and percent of Gleason scores 4 or 5. When analyzing some dataset, different model selection methods may choose different models as the 'true' one. For example, AIC and BIC, two commonly used model selection criteria, may pick two different models, as the criteria for selection is different. Such situation would certainly lead many questions in practice. For instance, if the estimator is selected by using a model selection criteria, how would we address the possibility that the selection is a wrong model? Also, if different model selection methods give us different results, one might wonder how trustworthy the model selection procedures are. Instead of choosing one model using a model selection scheme, one can use an average of estimators from different models. The model average estimator then can provide us an estimate of any parameter involved in the study and can be used for providing confidence bounds. The model average estimator can also be used for prediction purposes as well.

Hjort and Claeskens [2003] provided the first formal theoretical treatment of frequentist model averaging approaches, and it was well cited. However, the assumption that any extra parameters not included in the narrowest model will shrink to zero at a $\mathcal{O}(1/\sqrt{n})$ rate is too constraining in practice. It essentially requires that the all candidate models are within a $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model. Although this assumption avoids a technical difficulty of handling biased estimators, in reality we do not know the true model and excluding from consideration those models that are beyond $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model appears to be very restrictive.

In this paper, we remove this restrictive assumption in Hjort and Claeskens [2003] and develop frequentist model averaging approaches under a much more general framework. Our model averaging scheme allows us to use all the potential candidate models available, even the ones with large biases.

The idea of including all models, even that are biased, is motivated by the idea

of bias-variance trade-off. If we are using an overly simple model, the parameter estimates will probably be biased, but it will also have less variance, because there are fewer parameters to estimate. Similarly a bigger model is used, the parameter estimates will have low or no bias but increased variance. In our analysis we do not assume any particular structure for the true model, the candidate models are not restricted within a $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model. Thus we can use all candidate models which may result in parameter estimates with both high and low bias and variance.

Next, we discuss developing a set of weights that will help us to build a combined model average estimator. The weights are based on the idea of bias variance tradeoff. We develop the weights based on mean square error of the model average estimator, which takes into account both bias and variance of the estimator. When sample size is small the model average estimator may be based on biased candidate models. Since the weights are based on mean square error of the model average estimator, biased estimators may end up having lower mean square error than the true model. However, the weights chosen often display good optimality properties, for example, the parameter estimates converging to the true value as n becomes large. Thus it can be shown that in most of the cases weights that are chosen to combine the candidate models highlight the contribution of the true model.

Our approach to weight selection is based on the mean squared error (MSE) properties of the model average estimator similar to that discussed in Liang et al. [2011]. Specifically, a consistent estimate of the mean square error of the model average estimator is proposed, and the weights are chosen such that the MSE estimate is minimized. Using this weights, we show that model averaging performs better or no worse than several existing and commonly used model selection or model averaging methods.

Model averaging method was also discussed in a Bayesian framework; see, e.g.

Raftery and Hoeting [1998] and Hoeting et al. [1999]. A weighted average of the posterior distributions under every available candidate model was used for estimation and prediction purposes. The weights were determined by posterior model probabilities. Model averaging in a frequentist setup, as in Hjort and Claeskens [2003] and also ours, precludes the need to specify any prior distributions, thus removing any possible oversight due to faulty choice of priors. The question in a frequentist setting is how to obtain the weights by a data-driven approach.

In section 2.2, we propose a general framework that subsumes the framework of Hjort and Claeskens [2003] and study asymptotic properties of model average estimators. We also derive a consistent estimator for the mean square error of the model average estimator and use it to facility our choice of data-driven weights. In section 2.3.1, the model averaging methodology developed is illustrated in generalized linear models and particularly linear and logistic model setups. We also develop the choice of weights for the model average estimator in the linear and logistic setup. In section 2.4, a simulation study is carried out to examine the performance of the proposed estimator and to compare its performance with existing methods.

2.2 General Framework

2.2.1 Basic Notation and Set up.

Consider n independent data points $\boldsymbol{y} = (y_1, \cdots, y_n)$ sampled from a distribution having density of the form $f(y_i) \equiv f(y_i, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the unknown parameter of interest.

Here $\boldsymbol{\beta}$ can be written as $\boldsymbol{\beta} = (\boldsymbol{\theta}, \boldsymbol{\gamma})$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, $p \geq 0$, are the parameters that are always included in every candidate model and $\boldsymbol{\gamma} \in \mathbb{R}^q$, is the remaining set of parameters that may or may not be included in the candidate models. We assume that p and q are given. Following the paradigm of model averaging, instead

of choosing one particular candidate model as the "correct" model, we consider all possible combinations of the q parameters as different candidate models. In another word, each candidate model contains the common parameters $\boldsymbol{\theta}$ and a unique $\boldsymbol{\gamma}$ that includes m of q components of the parameter, $0 \leq m \leq q$.

We define $\mathcal{M}$ as the set of candidate models in our analysis. The choice of $\mathcal{M}$ can vary depending on the problem one is trying to solve. For example, $\mathcal{M}$ can contain all possible 2^q candidate models. Or, one can always choose a subset of the 2^q possible models as $\mathcal{M}$. In Hansen [2007], a set of nested models has been used as candidate models, with $|\mathcal{M}| = q + 1$. In Hjort and Claeskens [2003] $\mathcal{M}$ includes candidate models that are within a $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model. Our development encompasses both setups as there are no restrictions on $\mathcal{M}$, $\mathcal{M}$ can include any number of candidate models between 1 and 2^q. Similar setup was used in Liang et al. [2011], where the framework was based on a unrestricted $\mathcal{M}$ as well, but the development was done in a linear regression framework.

Let the parameter in the true model be given by $\boldsymbol{\beta}_{true} = (\boldsymbol{\theta}_{True}, \boldsymbol{\gamma}_{True})$. Let m^{true} be the components of $\boldsymbol{\gamma}$ that are present in the true model. Define $\mathcal{M}_{\in}$ as the collection of the candidate models that contain the true model, thus every model in $\mathcal{M}_{\in}$ contain each and every one of the m^{true} components of $\boldsymbol{\gamma}$. Define, $\mathcal{M}_{\overline{\in}} = \mathcal{M} - \mathcal{M}_{\in} \subset \mathcal{M}$. So, $\mathcal{M}_{\overline{\in}}$ will contain candidate models for which at least one of those m^{true} components are not present. Clearly $\mathcal{M} = \mathcal{M}_{\in} \cup \mathcal{M}_{\overline{\in}}$.

In Hjort and Claeskens [2003] the authors provided the first formal theoretical treatment of frequentist model averaging. The work was done in a general parametric setup. In their framework the presence of a common parameter in all the candidate models is similar to our framework, but the treatment of $\boldsymbol{\gamma}$ is different. In the earlier work the model containing just $\boldsymbol{\theta}$ is called a narrow model and the true model is chosen of the form $f(\boldsymbol{y}) = f(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma}_0 + \delta/\sqrt{n})$. Here, parameter δ determines how far a candidate model can vary from the narrow model and $\boldsymbol{\gamma}_0$ is the value of $\boldsymbol{\gamma}$ for

which any extended model reduces down to the narrow model. Thus, this choice of true model essentially requires that the all candidate models are within a $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model. Any model that is beyond $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model is excluded from the analysis. In this paper, we remove this rather restrictive constraint and develop the model average estimator under a true model that has not restrictions imposed on it. As mentioned above, the parameter for the true model be given by $\boldsymbol{\beta}_{true} = (\boldsymbol{\theta}_{True}, \boldsymbol{\gamma}_{True})$, where $\boldsymbol{\gamma}_{True}$ may or may not have any of the q components. Thus in our model setup there are no restrictions on the choice of true model or on the set of candidate models as in Hjort and Claeskens [2003]. Indeed, we can treat the setup considered in Hjort and Claeskens [2003] as a special case of ours by restricting the value of $\boldsymbol{\gamma}_{True}$, such that all the candidate models will have bias of order $\mathcal{O}(1/\sqrt{n})$ or less.

In model averaging, every candidate model includes a unique $\boldsymbol{\gamma}$ that may or may not include all q components. Thus parameters from different candidate models will have different lengths for the parameter $\boldsymbol{\beta}$. To bring all of them at the same length and for ease of presentation we introduce the idea of augmentation. We use a simple example to illustrate the idea. Let us consider the a linear regression setup, where $\boldsymbol{y}$ is the vector of responses and $\boldsymbol{X}$ is the design matrix with full column rank $p+q$. We consider only nested models as candidate models as done in Hansen [2007]. We also assume the first p columns of $\boldsymbol{X}$ are always included in the candidate models. Then the k^{th} candidate model includes the first $p+k$ columns of $\boldsymbol{X}$, $k = 0, \cdots, q$. Then the augmented estimator for the k^{th} candidate model will be given by

$$\widetilde{\boldsymbol{\beta}}_k = (\widehat{\boldsymbol{\beta}}_k, 0) = \begin{bmatrix} (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^T \boldsymbol{y} \\ 0 \end{bmatrix}$$

Similar augmentation technique has been used before. Most notably, in Hansen [2007] the author used this augmentation on a set of nested candidate models.

Next we move on to the general setup. Let $\boldsymbol{\beta}_k$ be the parameter for the k^{th} model in $\mathcal{M}$. Define the log-likelihood for the i^{th} observation in the k^{th} model as $\ell_{k,i} = \log f(y_i, \boldsymbol{\beta}_k)$. The maximum likelihood estimate (MLE) of $\boldsymbol{\beta}_k$ for the k^{th} model is thus defined as the maximizer of $\sum_{i=1}^{n} \ell_{k,i}$. We also define the score function of the k^{th} model as $S_k(\boldsymbol{\beta})$. Then $\boldsymbol{\beta}_{true}$ is a solution of the equation $\mathbb{E}(S_{true}(\boldsymbol{\beta})) = 0$. Following these notations, as in the example above, for the k^{th} model $\in \mathcal{M}$ the augmented maximum likelihood estimator is given by

$$\widetilde{\boldsymbol{\beta}}_k = (\widehat{\boldsymbol{\beta}}_k, \boldsymbol{c}), \quad \text{where } \widehat{\boldsymbol{\beta}}_k \text{ is the MLE for } k^{\text{th}} \text{ model},$$

here $\boldsymbol{c}$ is the fixed value that is used for augmentation. This fixed value augmentation does not affect the parameter, only appends the length of the parameter. Note that in general linear model this value is $\boldsymbol{c} = 0$. Further details on this can be found in the Appendix. The model average estimator is defined as

$$\sum_{k \in \mathcal{M}} w_k \mu(\widetilde{\boldsymbol{\beta}}_k), \tag{2.2.1}$$

where $0 \leq w_k \leq 1 \forall\ k$ and $\sum_{k \in \mathcal{M}} w_k = 1$. For the model $k \in \mathcal{M}$, let us also define $\boldsymbol{\beta}_k^* \in \mathbb{R}^{p+m}$ as the solution of the equation $\mathbb{E}S_k(\boldsymbol{\beta}) = 0$, where $S_k(\boldsymbol{\beta})$ is the score function of the k^{th} model having $p + m$ parameters. Define, as before, $\widetilde{\boldsymbol{\beta}}_k^* \in \mathbb{R}^{p+q}$ as the $\boldsymbol{c}-$augmented version of $\boldsymbol{\beta}_k^*$. Note that while $\widetilde{\boldsymbol{\beta}}_k^*$ may not be close to $\boldsymbol{\beta}_{true}$, $\widetilde{\boldsymbol{\beta}}_k \to \widetilde{\boldsymbol{\beta}}_k^*$, a.s., due to consistency of MLE under usual regularity conditions. In this section we will focus on deriving the asymptotic properties of the model average estimators. We now present our main result.

2.2.2 Main Results.

We now develop the framework under traditional conditions of regularity, which are sufficient to apply familiar maximum likelihood asymptotics arguments. These conditions are described in the Appendix. For details of such conditions, see Lehmann and Casella [1998], Lehmann [1999] and Van der Vaart [2000].

We also assume that the variance matrix of the score statistic is finite and positive definite. We establish a few more notations necessary for analysis in the various candidate models. Define $\mathbf{H}_k = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\ell_k''(\boldsymbol{\beta}_k^*)\right]$. Here we assume that the regularity conditions as described in Section 4 in Hoadley [1971] are satisfied. Here $\boldsymbol{\mu} \in \mathbb{R}^{p+q} \to \mathbb{R}^{\ell}$ is a general function and $\nabla\boldsymbol{\mu} \in \mathbb{R}^{\ell\times(p+q)}$. We assume that $\boldsymbol{\mu} : \mathbb{R}^{p+q} \to \mathbb{R}^{\ell}$ be a function that is 1^{st} order partially differentiable at $\boldsymbol{\beta}_{true}$. We also assume $\mathbf{H}_k$ is invertible.

$$(\mathbf{A}1)\ \lim_n \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\max_{k\in\mathcal{M}}\|\nabla\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k^*)\mathbf{H}_k^{-1}\ell_{k;i}'\|\ \mathbb{I}\ \left\{\max_{k\in\mathcal{M}}\|\nabla\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k^*)\mathbf{H}_k^{-1}\ell_{k;i}'\| > \sqrt{n}\epsilon\right\}\right] = 0. \tag{2.2.2}$$

for any $\epsilon > 0$. This condition is straightforward and is satisfied in a wide array of cases. We provide such examples in the cases of linear and generalized linear models in Section 2.3.1. Condition $(\mathbf{A}1)$ implies that the contribution of $\nabla\mu^{(dropped)}(\boldsymbol{\beta}_k^*)\mathbf{H}_k^{-1}\ell_{k;i}'$ to the total variance, for each model k in the set $\mathcal{M}$ and for each $1 \le i \le n$ is asymptotically negligible. We discuss this condition further in Section 2.3.1, in particular, we describe sufficent conditions under which it is satisfied. We state the following theorem about the asymptotic distribution of the model average estimator.

Theorem 2.1. *Let $\widetilde{\boldsymbol{\beta}}_k$ be the $\boldsymbol{c}$-augmented MLE as define in (2.2) for the k^{th} model in $\mathcal{M}$. Let $0 \le w_k \le 1$ for $k \in \mathcal{M}$ be model weights so that $\sum_k w_k = 1$. Under the assumption $(\boldsymbol{A}1)$ in (2.2.2) above, the asymptotic distribution of the model average*

estimator for $\boldsymbol{\mu}(\boldsymbol{\beta}_{true})$ *is given as,*

$$\sqrt{n}\sum_{k\in\mathcal{M}} w_k(\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k) - \boldsymbol{\mu}(\boldsymbol{\beta}_{true})) - \sqrt{n}w_k(\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_{true,k})^T\partial\mu(\boldsymbol{\beta}_{true})/\partial\boldsymbol{\beta} \xrightarrow{\text{D}} \mathcal{N}\left(0,\ \boldsymbol{\Sigma}_w\right), \tag{2.2.3}$$

where the variance $\boldsymbol{\Sigma}_w$ *is given by*

$$\boldsymbol{\Sigma}_w = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\sum_k w_k\nabla\mu(\widetilde{\boldsymbol{\beta}}_k^*)^T\boldsymbol{H}_k^{-1}\ell'_{k;i}\|_2^2\right]. \tag{2.2.4}$$

The proof of the theorem is given in the Appendix. The weights considered in this theorem so far are fixed. However in practice, we need to estimate the weights using data. In case the weights considered are being computed from the data, we assume $w_k^{(n)}(\boldsymbol{y})$, the weight assigned to the kth model converges to w_k as n goes to infinity. Using a simple application of Slutsky's lemma it can be shown that the earlier result in Theorem 2.1 holds when the weights are replaced by data dependent weights.

In our development of model average estimator we considered the estimation of $\boldsymbol{\beta} = (\boldsymbol{\theta}, \boldsymbol{\gamma})$ and considered the candidate models to be constructed via different subsets of only the $\boldsymbol{\gamma}$ parameter. Thus all the candidate models had $\boldsymbol{\theta}$ in common. We can use Theorem 2.1 to construct asymptotic convergence results for the common parameter $\boldsymbol{\theta}$. If we consider the function given by $(\boldsymbol{\theta}, \boldsymbol{\gamma}) \mapsto \boldsymbol{\theta}$; a function that extracts the $\boldsymbol{\theta}$ parameter, then by direct application of Theorem 2.1 we can derive the asymptotic distribution of $\boldsymbol{\theta}$ as given below in Corollary 2.1.

Corollary 2.1. *Let* $\boldsymbol{\theta}$ *be the common parameter for all candidate models in* $\mathcal{M}$*. Let* $\boldsymbol{\beta}_{true} = (\boldsymbol{\theta}_{true}, \boldsymbol{\gamma}_{true})$, $\boldsymbol{\beta}_k^* = (\boldsymbol{\theta}_k^*, \boldsymbol{\gamma}_k^*)$, $\widehat{\boldsymbol{\beta}}_k = (\widehat{\boldsymbol{\theta}}_k, \widehat{\boldsymbol{\gamma}}_k)$*. Then under the same setup as in Theorem 2.1*

$$\sqrt{n}\sum_{k\in\mathcal{M}} w_k(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{true}) - \sqrt{n}\sum_{k\in\mathcal{M}} w_k(\boldsymbol{\theta}_k^* - \boldsymbol{\theta}_{true}) \xrightarrow{\text{D}} \mathcal{N}\left(0,\ \boldsymbol{\Sigma}_w^{\boldsymbol{\theta}}\right), \tag{2.2.5}$$

where the variance is given by $\boldsymbol{\Sigma}_w^{\boldsymbol{\theta}} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\| \sum_k w_k [\boldsymbol{I}_p, \boldsymbol{0}] \boldsymbol{H}_k^{-1} \ell'_{k;i} \|_2^2\right]$.

In Hjort and Claeskens [2003] the development was done with a choice of true model that essentially required that the all candidate models are within a $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model. We broaden this framework by using $\boldsymbol{\beta}_{true}$ in our analysis. Next we show that the results described in Hjort and Claeskens [2003] can be proved to be a special case of our result. For that purpose, we now discuss our results in the setup studied in Hjort and Claeskens [2003].

First we describe the misspecified model setup that is used in the aforementioned paper. This setup is based on iid data $Y_1, ..., Y_n$ from density f. The parameter of interest is $\mu = \mu(f)$, where $\mu : \mathbb{R}^{p+q} \to \mathbb{R}$. The model that includes just $\boldsymbol{\theta}$ is defined as the narrow model, while any extended model $f(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ reduces to the narrow model for $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$; here $\boldsymbol{\gamma}_0$ is fixed and known. For the k^{th} model the maximum likelihood estimator is $\widehat{\mu}_k = \mu(\widehat{\boldsymbol{\theta}}_k, \widehat{\boldsymbol{\gamma}}_k, \boldsymbol{\gamma}_{0,k^c})$. Thus in this setup, if a parameter γ_j is not included in the candidate model, we set $\gamma_j = \gamma_{j,0}$. The data is assumed to be generated from a density

$$f_{true}(y) = f(y, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{n}), \tag{2.2.6}$$

where $\boldsymbol{\delta}$ signify the deviation of the model in directions $1, ..., q$. Thus in this case $\boldsymbol{\beta}_{true} = (\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{n})$. Let us write $\boldsymbol{\beta}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$. We will also write $\mu_{true} = \mu(\boldsymbol{\beta}_{true})$, which is the estimand under study. Under this misspecification model, Hjort and Claeskens [2003] derived asymptotic normality result for the model average estimator $\sum_k w_k \widehat{\mu}_k$. To describe their result, let us first define

$$S(y) = \begin{bmatrix} U(y) \\ V(y) \end{bmatrix} = \begin{bmatrix} \partial log f(y, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)/\partial \boldsymbol{\theta} \\ \partial log f(y, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)/\partial \boldsymbol{\gamma} \end{bmatrix} \text{ so that } var(S(Y)) = \begin{bmatrix} \mathbf{J}_{00} & \mathbf{J}_{01} \\ \mathbf{J}_{01} & \mathbf{J}_{11} \end{bmatrix} = \mathbf{J}_{full}.$$

Let $\overline{U}_n = n^{-1} \sum_i U(Y_i)$ (and similarly for $\overline{V}_n$). Let us denote by $V_k(Y), \overline{V}_{n;k}$ resp.,

the appropriately subsetted vectors obtained from $V(Y), \overline{V}_n$, with the subset indices corresponding to that of $\widehat{\boldsymbol{\gamma}}$ in model $k \in \mathcal{M}$. Also define $\mathbf{J}_k = var(U(Y), V_k(Y))^T$ for all $k \in \mathcal{M}$. Then Hjort and Claeskens [2003] shows that,

$$\sqrt{n}(\sum_k w_k \widehat{\mu}_k - \mu_{true}) \xrightarrow{\text{D}} \sum_k w_k \Lambda_k, \qquad \text{where,}$$
$$\Lambda_k = \begin{pmatrix} \partial\mu(\boldsymbol{\beta}_0)/\partial\boldsymbol{\theta} \\ \partial\mu(\boldsymbol{\beta}_0)/\partial\boldsymbol{\gamma}_k \end{pmatrix}^T \left\{ \mathbf{J}_k^{-1} \begin{pmatrix} \mathbf{J}_{01}\boldsymbol{\delta} \\ \pi_k \mathbf{J}_{11}\delta \end{pmatrix} + \mathbf{J}_k^{-1} \begin{pmatrix} \sqrt{n}(\overline{U}_n - \mathbb{E}U_k(Y_1)) \\ \sqrt{n}(\overline{V}_{n,k} - \mathbb{E}V_k(Y_1)) \end{pmatrix} \right\}$$
$$- \left(\frac{\partial\mu(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\gamma}} \right)^T \boldsymbol{\delta}. \tag{2.2.7}$$

In the above, $\pi_k \in \mathbb{R}^{|M_k| \times q}$ is the projection matrix that projects any vector $\boldsymbol{u} \in \mathbb{R}^q$ to $\boldsymbol{u}_k \in \mathbb{R}^{|M_k|}$ with indices as given by $M_k \in \mathcal{M}$. From our asymptotic convergence result in Theorem 2.1 it follows that,

Corollary 2.2. *under the misspecification model (2.2.6), the asymptotic bias and variance in (2.2.7) matches that in Theorem 2.1.*

Let us consider the a linear regression setup, where $\boldsymbol{y}$ is the vector of responses and $\boldsymbol{X}$ is the design matrix with full column rank $p + q$. We also assume the first p columns of $\boldsymbol{X}$ are always included in the candidate models. As discussed earlier in the development in Section 2.2, calculating model average estimator involves averaging over candidate models. To construct model average estimator $\widehat{\mu}^{ave}$ for some function μ in the linear regression set up, we need to estimate $\widehat{\boldsymbol{\beta}}_k$ for all $1 \leq k \leq \mathcal{M}$. This estimation procedure is computationally intensive especially when $|\mathcal{M}|$ is large. Sometimes one particular parameter, say, β_t can be of interest. Among the set of candidate models, some models may include that particular covariate $\boldsymbol{x}_t$, while others do not. So if we consider the problem of estimating the regression coefficient β_t for an explanatory variable $\boldsymbol{x}_t$, an alternate approach could be only including the candidate models in the analysis that contains $\boldsymbol{x}_t$. Then the idea is to find a model average

estimate of β_t over those models only. We find the estimate of β_t by only using the models that regress on $\boldsymbol{x}_t$ and thus estimates the parameter. Then we assign new weights to those estimates that are proportional to the original weights. A scaled version of the original weights can be used as the new weights, and the new weights sum up to 1. The model average estimator averages the estimates of β_t across all models which include it, using these new weights.

2.3 Selection of Weights in Frequentist Model Averaging.

The key idea behind model averaging is to acknowledge the uncertainty regarding which model is the truth. We weight all candidate models to incorporate this uncertainty. This is done by developing a set of weights that to some degree is a measure of evidence of each candidate model. In the following development we assume that the true model is included in the set of candidate models. For each of the candidate models, we assign a weight w_i to model M_i, for all i. We restrict $0 \leq w_i \leq 1$ for all i, and also impose the constraint that $\sum_i w_i = 1$. Under these restrictions, model weights may be thought of as probabilities associated with each model. If $w_i < w_j$, then in some sense model M_j is more likely, or more plausible than the competing model M_i.

In this section we propose a set of weights for model average estimator. We minimize an estimate of the mean squared error to obtain weights that would be used to combine the candidate models. Similar weights were discussed in Liang et al. [2011], where the authors minimized an unbiased estimator of mean squared error to obtain the weights. However, their work was done in linear model. In this section we propose a set of weights for model average estimator in general parametric models.

From Theorem 2.1 we calculate the asymptotic mean squared error (AMSE) of μ

as,

$$Q(\boldsymbol{w}) = \sum_{k \in \mathcal{M}} w_k \{\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k^*) - \boldsymbol{\mu}(\boldsymbol{\beta}_{true})\}^2 + \boldsymbol{\Sigma}_w.$$

Let the estimate of $Q(\boldsymbol{w})$ be $\widehat{Q}_n(\boldsymbol{w})$, where $\widehat{Q}_n(\boldsymbol{w})$ is consistent for $Q(\boldsymbol{w})$. We want to obtain $w_1, \cdots, w_N$ such that the trace of the estimate of the MSE proposed is minimized, denoted by $\widehat{Q}_n(\boldsymbol{w})$. Since these weights are based on the mean square error of the model average estimator, which takes into account both bias and variance of the estimator. When sample size is small the model average estimator may be based on biased candidate models. Since the weights are based on mean square error of the model average estimator, biased estimators may end up having lower mean square error than the true model. Next, we focus on the behavior of weights when the sample size is large. We want the chosen weights to have good optimality properties, for example, the parameter estimates converging to the true value as n becomes large. It can be shown that when properly chosen weights are used to combine the candidate models, resulting model average estimator is asymptotically equivalent to an estimator based on the true model. In this section we demonstrate such a choice of weights, while similar examples can be found in literature. As describes before we obtain weights $\boldsymbol{w}_n^*$ by minimizing $\widehat{Q}_n(\boldsymbol{w})$. We want to show, $\widehat{Q}_n(\boldsymbol{w}_n^*) \xrightarrow{\mathcal{P}} Q(\boldsymbol{w}^*)$, where $\boldsymbol{w}^* = \text{argmin}_{\boldsymbol{w}} Q(\boldsymbol{w})$. Then,

Theorem 2.2. *(a) Under some mild regularity conditions,* $\widehat{Q}_n(\boldsymbol{w}_n^*) \xrightarrow{\mathcal{P}} Q(\boldsymbol{w}^*)$.
(b) Suppose further that $Sup_{\boldsymbol{w}} \left|\widehat{Q}_n(\boldsymbol{w}) - Q(\boldsymbol{w})\right| \xrightarrow{\mathcal{P}} 0$, *for any* $\boldsymbol{w}$, *and* $\boldsymbol{w}^*$ *is a well separated point of minimum of* $Q(\boldsymbol{w})$, $\widehat{Q}_n(\boldsymbol{w}_n^*) \leq \widehat{Q}_n(\boldsymbol{w}^*) + o_{\mathbb{P}}(1)$. *Then,* $\boldsymbol{w}_n^* \xrightarrow{\mathcal{P}} \boldsymbol{w}^*$.

This can be proven by noting $\widehat{Q}_n(\boldsymbol{w}_n^*) = Q(\boldsymbol{w}_n^*) + o_{\mathbb{P}}(1) \geq Q(\boldsymbol{w}^*) + o_{\mathbb{P}}(1)$. Also, $\widehat{Q}_n(\boldsymbol{w}_n^*) - Q(\boldsymbol{w}^*) \leq \widehat{Q}_n(\boldsymbol{w}^*) - Q(\boldsymbol{w}^*) = o_{\mathbb{P}}(1)$.

Using these weights we can state the following theorem about the asymptotic distribution of the model average estimator.

Theorem 2.3. *Let* $0 \leq \boldsymbol{w}_{nk}^* \leq 1$ *for* $k \in \mathcal{M}$ *are model weights so that* $\sum_k w_{nk}^* = 1$ *and* $\boldsymbol{w}_n^* \xrightarrow{\mathcal{P}} \boldsymbol{w}^*$. *And* $\widetilde{\boldsymbol{\beta}}_k$ *be the* $\boldsymbol{c}$*-augmented MLE as define in (2.2) for the* k^{th} *model in* $\mathcal{M}$. *Then, under the assumption* $(\boldsymbol{A}1)$ *in (2.2.2) above, the asymptotic distribution of the model average estimator for* $\boldsymbol{\mu}(\boldsymbol{\beta}_{true})$ *is given as,*

$$\sqrt{n} \sum_{k \in \mathcal{M}} w_{nk}^* (\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k) - \boldsymbol{\mu}(\boldsymbol{\beta}_{true})) \xrightarrow{\text{D}} \mathcal{N}\left(0,\ \boldsymbol{\Sigma}_{w^*}\right), \tag{2.3.1}$$

where the variance $\boldsymbol{\Sigma}_{w^*}$ *is given by*

$$\boldsymbol{\Sigma}_{w^*} = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\| \sum_k w_{nk}^* \nabla \mu(\widetilde{\boldsymbol{\beta}}_k^*)^T \boldsymbol{H}_k^{-1} \ell'_{k;i} \|_2^2 \right]. \tag{2.3.2}$$

When sample size increases, it is observed that the behavior of the model average estimator and the true model estimator is similar, which will be illustrated in a simulation study later. In next section we will develop the asymptotic distribution of the model average estimator in a linear regression setup. The weights selection process will also be described. The estimator proposed in the section above is not unbiased in general. But under certain specific framework such as linear regression, general liner model or exponential family it can be simplified and the estimators can be developed so that they are optimal or near optimal. We can develop consistent or unbiased estimators in linear regression framework as detailed in Liang et al. [2011]. This estimator of the MSE of the model average estimator could also be used to derive the model weights.

2.3.1 Selection of Weights in General Linear Models

We now discuss the model average estimator described in Section 2.2 for generalized linear models (GLM). As discussed before, we consider $\mathbb{E}y_i = g(\boldsymbol{x}_i^T \boldsymbol{\beta})$, where $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ are independent observations from a response variable $\boldsymbol{y}$, $\boldsymbol{x}_i$ is a vector

of explanatory variables and $\boldsymbol{\beta} \in \mathbb{R}^{p+q}$ is the vector of unknown parameters. Consider a link function g that connects the mean and the linear predictor $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. We let the first p models appear in all possible candidate models and consider a set $\mathcal{M}$ of 2^q models. We want to estimate some function $\boldsymbol{\mu}(\boldsymbol{\beta})$ and the final model average estimator is given by,

$$\boldsymbol{\mu}(\boldsymbol{\beta}) = \sum_{k \in \mathcal{M}} w_k \boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k).$$

Since the set up for Theorem 2.1 is for a general parametric model, the same asymptotic convergence results hold for GLM models. In particular we discuss two special cases of linear and logistic regressions. Note that linear regression is the case when g is the identity map and $y_i \sim \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$. For logistic regression, the link function is the logit function $t \mapsto \log(t/(1-t))$ and $y_i \sim Bin(1, p_i)$. Our framework encompasses all general linear class of models and similar results can be derived for them as well.

Prediction in Linear Regression Framework.

We can use the results developed in the general framework to compute the model average estimator in a linear regression framework. Consider the linear regression model,

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is a vector of observations, $\mathbf{X} \in \mathbb{R}^{n \times p+1}$ is a non-random design matrix, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p) \in \mathbb{R}^{p+1}$ is the vector of unknown parameters, with $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Additionally, we assume that $\mathbf{X}$ has full column rank i.e. $\text{rank}(\mathbf{X}) = p + 1$.

Let $\mathcal{M} = \{M_k\}_{k=1}^{|\mathcal{M}|}$ be the set of candidate models Here M_k denotes a particular set of features having cardinality $|M_k|$. Define $\mathbf{X}_k \in \mathbb{R}^{n \times |M_k|}, 1 \leq k \leq |\mathcal{M}|$ as the

design matrix of the k^{th} candidate model with the features in M_k. We consider zero -augmentation of the parameter set $\boldsymbol{\beta}_k$ for all k. Let $\widetilde{\mathbf{X}}_k \in \mathbb{R}^{n\times p}$ be the augmented version of $\mathbf{X}_k$ with the missing columns replaced by the $\mathbf{0}$ vector. In our analysis, all the candidate models contain the intercept term corresponding to β_0. With the rest of the p components, we can construct 2^p candidate models all of which are included in our analysis.

Let us fix a $\boldsymbol{x}^* \in \mathbb{R}^{p+1}$. Define $\boldsymbol{x}_k^* \in \mathbb{R}^{|M_k|}$ so that $\boldsymbol{x}_k^*$ consists of those components of $\boldsymbol{x}^*$ indexed by $M_k \in \mathcal{M}$. Consider the particular choice of the function $\mu : \mathbb{R}^{p+1} \to \mathbb{R}$ so that for $\boldsymbol{b} \in \mathbb{R}^{p+1}$, $\mu(\boldsymbol{b}) = \boldsymbol{x}^{*T}\boldsymbol{b}$. Clearly the $\nabla\mu(\boldsymbol{\beta}) = \boldsymbol{x}^*$. For the following discussion, we are interested in the model average estimator of $\mu(\boldsymbol{\beta}_{true}) = \boldsymbol{x}^*\boldsymbol{\beta}_{true}$, which is given by $\widehat{\mu}^{ave} = \sum_k w_k \boldsymbol{x}_k^{*T}\widehat{\boldsymbol{\beta}}_k$ where $w_k \geq 0$ and $\sum_k w_k = 1$.

For the k^{th} candidate model with $\boldsymbol{\beta}_k \in \mathbb{R}^{|M_k|}$, the score function is given by $\boldsymbol{\ell}_k'(\boldsymbol{\beta}_k) = \mathbf{X}_k^T(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k)$ and the hessian matrix is given by $\mathbf{H}_k = -\mathbf{X}_k^T\mathbf{X}_k$. Thus our hessian matrix satisfies the condition as it does not depend on $\boldsymbol{y}$. Similarly we note that referring condition ($\mathbf{A}1$), in this example,

$$|\nabla\mu(\widetilde{\boldsymbol{\beta}}_k^*)\mathbf{H}_k^{-1}\ell_{k;i}'| = \left|(y_i - [\mathbf{X}_k]_{i,\bullet}^T\boldsymbol{\beta}_k^*)\ \boldsymbol{x}_k^{*T}(\mathbf{X}_k^T\mathbf{X}_k)^{-1}[\mathbf{X}_k]_{i,\bullet}\right| = |c_{ik}(\varepsilon_i + A_{ik})|,$$

where $c_{ik} = \boldsymbol{x}_k^{*T}(\mathbf{X}_k^T\mathbf{X}_k)^{-1}[\mathbf{X}_k]_{i,\bullet}$ and $A_{ik} = \boldsymbol{x}_i^T(\boldsymbol{\beta}_{true} - \widetilde{\boldsymbol{\beta}}_k^*))$ are fixed constants. Note that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The condition ($\mathbf{A}2$) is satisfied when for any arbitrary $\epsilon > 0$,

$$\lim_{n\to\infty}\frac{1}{n}\max_{1\leq i\leq n}\mathbb{E}\left\{\max_{k\in\mathcal{M}}|c_{ik}(\varepsilon_i + A_{ik})|\right\}^2\mathbb{I}\left\{\max_{k\in\mathcal{M}}|c_{ik}(\varepsilon_i + A_{ik})| > \sqrt{n}\epsilon\right\} = 0.$$

Moreover, if $|c_{ik}| \leq C$ for some fixed constant $C > 0$ then we can reduce the condition further to,

$$\lim_{n\to\infty}\frac{1}{n}\max_{1\leq i\leq n}\mathbb{E}\left\{\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}|\right\}^2\mathbb{I}\left\{\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}| > \sqrt{n}\epsilon\right\} = 0.$$

It is appropriate to to note that we can have a bound of c_{ik} as

$$\begin{aligned}\max_k |c_{ik}| &= \max_k |\boldsymbol{x}_k^{*T}(\mathbf{X}_k^T\mathbf{X}_k)^{-1}[\mathbf{X}_k]_{i,\bullet}| \\ &\leq \|\boldsymbol{x}^*\|\|\boldsymbol{x}_i\| \max_k \frac{1}{\lambda^2_{min}(\mathbf{X}_k)}.\end{aligned}$$

Here $\lambda_{min}(\mathbf{X}_k)$ denotes the smallest singular value of $\mathbf{X}$. Now by application of Cauchy-Schwarz inequality,

$$\begin{aligned}&\frac{1}{n}\mathbb{E}\left\{\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}|\right\}^2 \mathbb{I}\left\{\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}| > \sqrt{n}\epsilon\right\} \\ &\qquad\leq \frac{1}{n}\left\{\mathbb{E}\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}|^4\right\}^{1/2}\left\{\mathbb{P}(\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}| > \sqrt{n}\epsilon)\right\}^{1/2} \\ &\qquad\leq \frac{1}{n}\sum_{k\in M}\mathbb{E}(\varepsilon_i + A_{ik})^4\left\{\sum_{k\in\mathcal{M}}\mathbb{P}(|\varepsilon_i + A_{ik}| > \sqrt{n}\epsilon)\right\}^{1/2} \\ &\qquad\leq \left\{\frac{A_{ik}^4}{n^2} + 6\frac{A_{ik}^2\sigma^2}{n^2} + \frac{3\sigma^4}{n^2}\right\}^{1/2}\left\{\sum_{k\in\mathcal{M}}\mathbb{P}(|\varepsilon_i| > \sqrt{n}\epsilon - |A_{ik}|)\right\}^{1/2}.\end{aligned} \tag{2.3.3}$$

Thus it follows that for $|\mathcal{M}|$ finite, as n goes to infinity, the right hand side of the above equation (2.3.3) goes to zero and thus condition (**A**2) is satisfied.

The MLE for the k^{th} model is given by $\widehat{\boldsymbol{\beta}}_k = (\mathbf{X}_k^T\mathbf{X}_k)^{-1}\mathbf{X}_k^T\boldsymbol{y}$. Let $\boldsymbol{\beta}_k^*$ be such that $\mathbb{E}\boldsymbol{\ell}_k(\boldsymbol{\beta}_k^*) = \mathbf{0}$; $\mathbb{E}\boldsymbol{\ell}_k(\boldsymbol{\beta}_k)$ being the score function of the k^{th} model. Solving which we find that,

$$\boldsymbol{\beta}_k^* = (\mathbf{X}_k^T\mathbf{X}_k)^{-1}\mathbf{X}_k^T\mathbf{X}\boldsymbol{\beta}_{true}. \tag{2.3.4}$$

As discussed in Section 2.2.1, the entire set of candidate models can be divided in to two categories. The 1^{st} category contains the ones that are biased and is denoted by $\mathcal{M}_{\notin}$ and the second category contains ones that are not and is denoted by $\mathcal{M}_{\in}$. So, for $k \in \mathcal{M}_{\in}$ we have $\boldsymbol{\beta}_k^* = \boldsymbol{\beta}_{true}$, whereas for $k \in \mathcal{M}_{\notin}$ we have $\boldsymbol{\beta}_k^* \neq \boldsymbol{\beta}_{true}$. Therefore

the bias term of model average estimator $\widehat{\mu}^{ave}$ can be written as,

$$\sum_{k\in\mathcal{M}_{\notin}} w_k(\boldsymbol{x}_k^{*T}\boldsymbol{\beta}_k^* - \boldsymbol{x}^{*T}\boldsymbol{\beta}_{true}) = \boldsymbol{x}_k^{*T}(\mathbf{X}_k^T\mathbf{X}_k)^{-1}\mathbf{X}_k^T\mathbf{X}\boldsymbol{\beta}_{true} - \boldsymbol{x}^{*T}\boldsymbol{\beta}_{true}.$$

Since the weights assigned to the models are unknown, we propose an estimate of the mean squared error (MSE) and minimize the MSE to obtain weights that would be assigned to the candidate models. From Theorem 2.1, the mean squared error (MSE) of $\widehat{\mu}^{ave}$ is given by

$$\begin{aligned} Q(\boldsymbol{w}) &= \mathbb{E}(\boldsymbol{x}_k^{*T}\widehat{\boldsymbol{\beta}}_k - \boldsymbol{x}^{*T}\boldsymbol{\beta}_{true})^2 \\ &= \left\{ \left[\sum_{k\in\mathcal{M}_{\notin}} w_k(\boldsymbol{x}_k^{*T}\boldsymbol{\beta}_k^* - \boldsymbol{x}^{*T}\boldsymbol{\beta}_{true}) \right]^2 \right. \\ &\qquad \left. + \frac{1}{n^2}\sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}} w_k w_{k'} \boldsymbol{x}_k^{*T}\mathbf{H}_k^{-1}\, \mathbb{E}\boldsymbol{\ell}_k'(\boldsymbol{\beta}_{true})\boldsymbol{\ell}_{k'}'(\boldsymbol{\beta}_{true})^T\, \mathbf{H}_{k'}^{-1^T}\boldsymbol{x}_k^* \right\}. \end{aligned}$$

We want to propose an estimate for the MSE stated. Since $\mathbf{H}_k$ does not depend on $\boldsymbol{y}$, we focus on estimating $\mathbb{E}\boldsymbol{\ell}_k'(\boldsymbol{\beta}_{true})\boldsymbol{\ell}_{k'}'(\boldsymbol{\beta}_{true})^T$. Now,

$$\mathbb{E}\boldsymbol{\ell}_k'(\boldsymbol{\beta}_{true})\boldsymbol{\ell}_{k'}'(\boldsymbol{\beta}_{true})^T = \mathbf{X}_k^T\mathbb{E}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_{true})(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_{true})^T\mathbf{X}_{k'} = \sigma^2\mathbf{X}_k^T\mathbf{X}_{k'},$$

so that the MSE is given by,

$$\begin{aligned} Q^{(linear)}(\boldsymbol{w}) = \left\{ \sum_{k\in\mathcal{M}_{\notin}}\sum_{k'\in\mathcal{M}_{\notin}} w_k w_{k'}(\boldsymbol{x}_k^{*T}\boldsymbol{\beta}_k^* - \boldsymbol{x}^{*T}\boldsymbol{\beta}_{true})(\boldsymbol{x}_{k'}^{*T}\boldsymbol{\beta}_{k'}^* - \boldsymbol{x}^{*T}\boldsymbol{\beta}_{true}) \right. \\ \left. + \frac{\sigma^2}{n^2}\sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}} w_k w_{k'}\boldsymbol{x}_k^{*T}(\mathbf{X}_k^T\mathbf{X}_k)^{-1}\, \mathbf{X}_k^T\mathbf{X}_{k'}\, (\mathbf{X}_{k'}^T\mathbf{X}_{k'})^{-1}\boldsymbol{x}_k^* \right\}. \end{aligned}$$

Let us define the estimates,

$$\widehat{\boldsymbol{\beta}}_{full} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}, \qquad \widehat{\sigma}_{full}^2 = \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{full}\|^2/n. \tag{2.3.5}$$

Thus $(\widehat{\boldsymbol{\beta}}_{full}, \widehat{\sigma}_{full})$ are consistent estimates of $(\boldsymbol{\beta}_{true}, \sigma)$. We propose $Q^{(linear)}(\boldsymbol{w})$ as

$$\widehat{Q}^{(linear)}(\boldsymbol{w}) = \text{trace}\left\{ \sum_{k\in\mathcal{M}_{\notin}} \sum_{k'\in\mathcal{M}_{\notin}} w_k w_{k'} (\boldsymbol{x}_k^{*T}\widehat{\boldsymbol{\beta}}_k - \boldsymbol{x}^{*T}\widehat{\boldsymbol{\beta}}_{full})(\boldsymbol{x}_{k'}^{*T}\widehat{\boldsymbol{\beta}}_{k'} - \boldsymbol{x}^{*T}\widehat{\boldsymbol{\beta}}_{full}) + \frac{\widehat{\sigma}^2_{full}}{n^2} \sum_{k\in\mathcal{M}} \sum_{k'\in\mathcal{M}} w_k w_{k'} \boldsymbol{x}_k^{*T} (\mathbf{X}_k^T\mathbf{X}_k)^{-1}\, \mathbf{X}_k^T\mathbf{X}_{k'}\, (\mathbf{X}_{k'}^T\mathbf{X}_{k'})^{-1} \boldsymbol{x}_k^* \right\}. \tag{2.3.6}$$

We obtain the weights for model average estimator $\boldsymbol{w} = (w_1, \cdots, w_{|\mathcal{M}|})$ such that $\widehat{Q}^{(linear)}(\boldsymbol{w})$ in (2.3.6) is minimized.

Estimation in Logistic Regression Framework.

In this section we develop model average estimators for generalized linear models (GLM). We specifically focus on Logistic regression models which is widely employed type of GLM; it is used for modeling dichotomous responses based on a set of continuous or categorical features. See Hosmer and Lemeshow [2000] for details.

Let $\boldsymbol{y} \in \mathbb{R}^n$ ne n independent copies of a dichotomous response variable Y taking values 0/1. Let $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n\times(p+1)}$ be a set of features. The logit model is given by,

$$p_i = P(y_i = 1|\mathbf{X}) = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}, \quad \forall i = 1, \cdots, n,$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ are the set of unknown parameters of interest. Alternatively, the linear predictor in logistic regression has the interpretation as the conditional log odds, i.e.

$$\log\left[\frac{P(y_i = 1|\mathbf{X})}{P(y_i = 0|\mathbf{X})}\right] = \boldsymbol{x}_i^T\boldsymbol{\beta}.$$

Assuming y_i's are independent observations the log-likelihood for logistic regression

can be written as,

$$\ell_k(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X}) = \log \prod_{i=1}^{n} \frac{\exp(y_i \boldsymbol{x}_i^T \boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} = \sum_{i=1}^{n} y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \log(1+\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})).$$

As before, let $\mathcal{M} = \{M_k\}_{k=1}^{|\mathcal{M}|}$ be the set of candidate models Here M_k denotes a particular set of features having cardinality $|M_k|$. Define $\mathbf{X}_k = (\boldsymbol{x}_{(k)1}, \cdots, \boldsymbol{x}_{(k)n})^T \in \mathbb{R}^{n \times |M_k|}, 1 \leq k \leq |\mathcal{M}|$ as the design matrix of the k^{th} candidate model with the features in M_k. Thus $\boldsymbol{x}_{(k)i} \in Re^{|M_k|}$. Let $\boldsymbol{\beta}_k \in \mathbb{R}^{|M_k|}$ be the parameter vector with components corresponding to the index set M_k. We consider zero -augmentation of the parameter set $\boldsymbol{\beta}_k$ for all k as was done for linear regression models.

For this discussion, we consider estimation of a function of the form $\boldsymbol{p} : \mathbb{R}^{p+1} \to \mathbb{R}^n$ given by

$$\boldsymbol{p}(\boldsymbol{\beta}) = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}, \tag{2.3.7}$$

which are calculated component wise. Let the unknown true parameter in our model be $\boldsymbol{\beta}_{true} \in \mathbb{R}^{p+1}$. Then $\boldsymbol{p}_{true} = \boldsymbol{p}(\boldsymbol{\beta}_{true}) = \exp(\mathbf{X}\boldsymbol{\beta}_{true})/(1+\exp(\mathbf{X}\boldsymbol{\beta}_{true})) \in \mathbb{R}^n$ calculated component wise. To estimate the parameter $\boldsymbol{p}^{true}$, we consider the model average estimator given by

$$\boldsymbol{p}^{ave} = \sum_{k \in \mathcal{M}} w_k \boldsymbol{p}(\widetilde{\boldsymbol{\beta}}_k),$$

where $\widetilde{\boldsymbol{\beta}}_k$ is the 0-augmented version of the MLE for $\widehat{\boldsymbol{\beta}}_k$ of $\boldsymbol{\beta}_k$ for the k^{th} model. The score function for the k^{th} model is given by

$$\boldsymbol{\ell}_k^{'}(\boldsymbol{\beta}_k) = \sum_i y_i \boldsymbol{x}_{(k)i} - \sum_i \frac{\exp(\boldsymbol{x}_{(k)i}^T \boldsymbol{\beta}_k)}{1+\exp(\boldsymbol{x}_{(k)i}^T \boldsymbol{\beta}_k)} \boldsymbol{x}_{(k)i} = \mathbf{X}_k^T(\boldsymbol{y} - \boldsymbol{p}_k) \quad \forall\ 1 \leq k \leq |\mathcal{M}|.$$

Here we have used the notation $\boldsymbol{p}_k = (p_{(k)1}, \cdots, p_{(k)n})$ as the vector of probabilities as

computed for the k^{th} model defined by $p_{(k)i} = \exp(\boldsymbol{x}_{(k)i}^T\boldsymbol{\beta}_k)/(1+\exp(\boldsymbol{x}_{(k)i}^T\boldsymbol{\beta}_k))$. The hessian of the log-likelihood is given by

$$\boldsymbol{\ell}_k''(\boldsymbol{\beta}_k) = \sum_{i=1}^{n} \frac{\exp(\boldsymbol{x}_{(k)i}^T\boldsymbol{\beta}_k)}{(1+\exp(\boldsymbol{x}_{(k)i}^T\boldsymbol{\beta}_k))^2}\boldsymbol{x}_{(k)i}\boldsymbol{x}_{(k)i}^T = \mathbf{X}_k^T\mathbf{W}_k(\mathbf{I}_n - \mathbf{W}_k)\mathbf{X}_k \quad \forall\ 1 \le k \le |\mathcal{M}|.$$

To estimate the bias of the model average estimator we need to find a $\boldsymbol{\beta}_k^*$ which is the solution of the equation

$$\mathbb{E}[\boldsymbol{\ell}_k'(\boldsymbol{\beta}_k)] = \mathbb{E}(\mathbf{X}_k^T(\boldsymbol{y} - \boldsymbol{p}_k)) = \mathbf{0}. \tag{2.3.8}$$

Thus $\boldsymbol{\beta}_k^*$ is also a solution of $\mathbf{X}_k^T(\boldsymbol{p}_{true} - \boldsymbol{p}_k) = 0$. Let us denote by $\boldsymbol{p}_k^* = \exp(\mathbf{X}_k\boldsymbol{\beta}_k^*)/(1+\exp(\mathbf{X}_k\boldsymbol{\beta}_k^*)) \in \mathbb{R}^n$ calculated component wise. Define as before $\mathbf{W}_k^* = \text{diag}(\boldsymbol{p}_k^*) \in \mathbb{R}^{n\times n}$ and $\mathbf{W}^{true} = \text{diag}(\boldsymbol{p}_{true}) \in \mathbb{R}^n$. We can use iterative re-weighted least squares (IRLS) method to solve the equation. See Holland and Welsch [2007] for more details. Let $\boldsymbol{\beta}_k^{*(i)}$ be the solution of (2.3.8) at the i^{th} stage of the IRLS algorithm. The coefficients for the $(i+1)^{\text{th}}$ stage is then given by

$$\begin{aligned}
&\boldsymbol{\beta}_k^{*(i+1)} \\
&= \boldsymbol{\beta}_k^{*(i)} + [\mathbf{X}_k^T\mathbf{W}_k(\mathbf{I}_n - \mathbf{W}_k)\mathbf{X}_k]^{-1}\mathbf{X}_k^T\left[\frac{\exp(\mathbf{X}\boldsymbol{\beta}_{true})}{1+\exp(\mathbf{X}\boldsymbol{\beta}_{true})} - \frac{\exp(\mathbf{X}_k\boldsymbol{\beta}_k)}{1+\exp(\mathbf{X}_k\boldsymbol{\beta}_k)}\right]\Bigg|_{\boldsymbol{\beta}_k=\boldsymbol{\beta}_k^{*(i)}} \\
&= \boldsymbol{\beta}_k^{*(i)} + [\mathbf{X}_k^T\mathbf{W}_k(\mathbf{I}_n - \mathbf{W}_k)\mathbf{X}_k]^{-1}\mathbf{X}_k^T(\boldsymbol{p}_{true} - \boldsymbol{p}_k)\Big|_{\boldsymbol{\beta}_k=\boldsymbol{\beta}_k^{*(i)}}.
\end{aligned}$$

Since the weights assigned to the models are unknown, we follow the setup in Section 2.2 and propose an estimate of the mean squared error and use that to obtain weights that would be assigned to the candidate models. In order to calculate the MSE we need to calculate the gradient of $\boldsymbol{\mu}$ i.e. $\nabla\boldsymbol{p}^{(dropped)} \in \mathbb{R}^{n\times M_k}$ for $1 \le k \le |\mathcal{M}|$, which is given by

$$\nabla\boldsymbol{p}^{(dropped)}(\boldsymbol{\beta}_k^*) = (\mathbf{I}_n - \mathbf{W}_k^*)\mathbf{W}_k^*\mathbf{X}_k, \quad 1 \le k \le |\mathcal{M}|.$$

Now note that,

$$\begin{aligned}
\mathbb{E}\boldsymbol{\ell}'_k(\boldsymbol{\beta}^*_k)\boldsymbol{\ell}'_{k'}(\boldsymbol{\beta}^*_k)^T &= \mathbf{X}_k^T\mathbb{E}(\boldsymbol{y}-\boldsymbol{p}^*_k)(\boldsymbol{y}-\boldsymbol{p}^*_{k'})^T\mathbf{X}_{k'} \\
&= \mathbf{X}_k^T\mathbb{E}(\boldsymbol{y}-\boldsymbol{p}_{true}-(\boldsymbol{p}^*_k-\boldsymbol{p}_{true}))(\boldsymbol{y}-\boldsymbol{p}_{true}-(\boldsymbol{p}^*_{k'}-\boldsymbol{p}_{true}))^T\mathbf{X}_{k'} \\
&= \mathbf{X}_k^T\mathbb{E}(\boldsymbol{U}-(\boldsymbol{p}^*_k-\boldsymbol{p}_{true}))(\boldsymbol{U}-(\boldsymbol{p}^*_{k'}-\boldsymbol{p}_{true}))^T\mathbf{X}_{k'} \\
&= \mathbf{X}_k^T\left[\mathbb{E}\boldsymbol{U}\boldsymbol{U}^T+(\boldsymbol{p}^*_k-\boldsymbol{p}_{true})(\boldsymbol{p}^*_k-\boldsymbol{p}_{true})^T\right]\mathbf{X}_{k'} \\
&= \mathbf{X}_k^T\left[\mathbf{W}^{true}+(\boldsymbol{p}^*_k-\boldsymbol{p}_{true})(\boldsymbol{p}^*_{k'}-\boldsymbol{p}_{true})^T\right]\mathbf{X}_{k'},
\end{aligned}$$

where we have used the notation that $\boldsymbol{U}=\boldsymbol{y}-\boldsymbol{p}_{true}$ so that $\mathbb{E}\boldsymbol{U}=\mathbf{0}$ and $var(\boldsymbol{U})=\mathbb{E}\boldsymbol{U}\boldsymbol{U}^T=\mathbf{W}^{true}$. Using these results, the MSE estimate for $\widehat{\boldsymbol{p}}^{ave}$ is be given by,

$$\begin{aligned}
Q^{(logistic)}(\boldsymbol{w}) = \text{trace}\Bigg\{ & \sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}} w_k w_{k'}(\boldsymbol{p}^*_k-\boldsymbol{p}_{true})(\boldsymbol{p}^*_k-\boldsymbol{p}_{true})^T \\
& + \frac{\sigma^2}{n^2}\sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}} w_k w_{k'}(\mathbf{I}_n-\mathbf{W}^*_k)\mathbf{W}^*_k\mathbf{X}_k[\mathbf{X}_k^T\mathbf{W}^*_k(\mathbf{I}_n-\mathbf{W}^*_k)\mathbf{X}_k]^{-1} \\
& \qquad \times\ \mathbf{X}_k^T\left[\mathbf{W}^{true}+(\boldsymbol{p}^*_k-\boldsymbol{p}_{true})(\boldsymbol{p}^*_{k'}-\boldsymbol{p}_{true})^T\right] \\
& \qquad \times\ \mathbf{X}_{k'}[\mathbf{X}_{k'}^T\mathbf{W}^*_{k'}(\mathbf{I}_n-\mathbf{W}^*_{k'})\mathbf{X}_{k'}]^{-1}\mathbf{X}_k^T\mathbf{W}^*_k(\mathbf{I}_n-\mathbf{W}^*_k)\Bigg\}.
\end{aligned}$$

We can obtain $w_1,\cdots,w_N$ such that a consistent estimate of the MSE $Q^{(logistic)}(\boldsymbol{w})$ is minimized, similar to the development done in linear regression setup. These weights can be assigned to individual models for developing the model average estimator.

2.4 Simulation Study & Real Data Analysis

2.4.1 Simulation Study

Large sample behavior & bias variance tradeoff.

In this section we study the large sample behavior of the model average estimator and the weights chosen. This is done under a linear regression framework. We use the

following simple linear regression setup with three regressor variables (not including the intercept) with unknown parameters $\beta_1, \beta_2, \beta_3$. The possible candidate models can be any of the $2^3 = 8$ parameter combinations possible. Namely, $[M_1] : \boldsymbol{y} = \beta_0\mathbf{1} + \boldsymbol{\epsilon_1}$, $[M_2] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \boldsymbol{\epsilon_2}$, $[M_3] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_2\boldsymbol{x_2} + \boldsymbol{\epsilon_3}$, $[M_4] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_3\boldsymbol{x_3} + \boldsymbol{\epsilon_4}$, $[M_5] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \boldsymbol{\epsilon_5}$, $[M_6] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \beta_3\boldsymbol{x_3} + \boldsymbol{\epsilon_6}$, $[M_7] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_2\boldsymbol{x_2} + \beta_3\boldsymbol{x_3} + \boldsymbol{\epsilon_7}$, $[M_8] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \beta_3\boldsymbol{x_3} + \boldsymbol{\epsilon_8}$. Here $\boldsymbol{\epsilon_i} \sim N(0, \sigma^2)$ $\forall i = 1, \cdots, 8$ and $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}$ are the regressors. We compare the mean square error of the parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ using model average estimator and the oracle mean square error, which is the mean square error assuming the true model from which the data is generated is known. Two different scenarios have been considered when studying the behavior of model average estimator under different sample size.

In the first scenario we use the true model $[M_{true}] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \boldsymbol{\epsilon_3}$, with $\beta_0 = 2$, $\beta_1 = 4$ and $\beta_2 = 0.5$. Depending on the choice of true model, different candidate models will have different biases. For example, when M_{true} is the true model, estimators from M_5 and M_8 will have no bias, whereas estimators from the rest of the candidate models will be biased. In this case we have two candidate models M_2 and M_5 that are really close to the true model, with M_2 being a biased candidate model. We vary the sample size from 100 to 1000 and compare the performance of model average estimator with the true one.

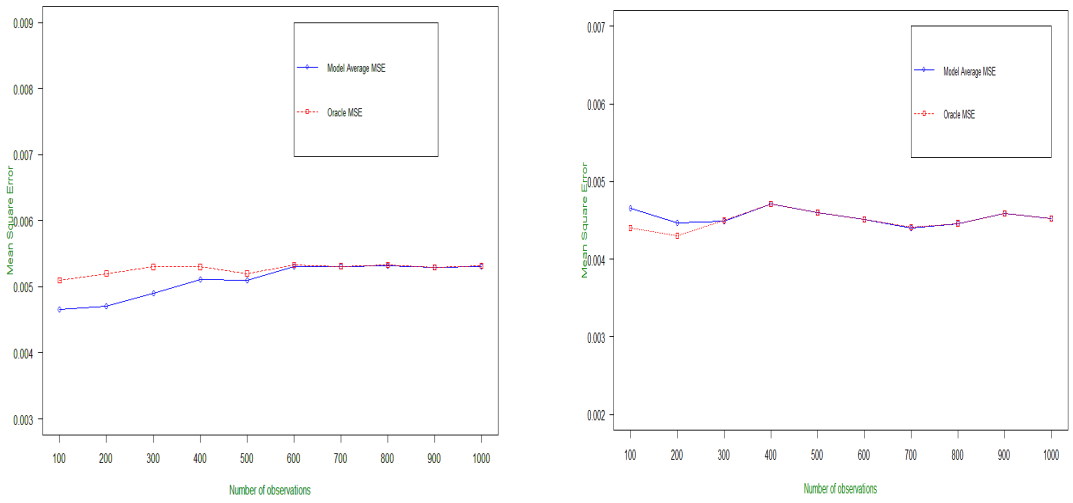

Figure 2.1: (a) Bias variance tradeoff of model average estimator and (b) large sample behavior of model average estimator

From Figure 2.1(a) we can see that, model average estimator performs better for small sample sizes. When the sample size is small, model average estimator selects smaller candidate models (models with fewer parameters than the true model) that have increased bias but low variance, and thus less mean square error than the true model. However, when sample size is increased, we see that the weights converge to the true weights and both model average mean square error and oracle mean square error are similar to each other.

Next we consider the true model $[M_{true}] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \boldsymbol{\epsilon_3}$, with $\beta_0 = 2$, $\beta_1 = 4$ and $\beta_2 = 5$. In this case there are no candidate models close to the true model. We vary the sample size again from 100 to 1000, and observe the performance of the model average estimator. This is done to examine the effectiveness of the weights chosen. We can see from Figure 2.1 (b) as sample size increases the mean square error of model average estimator approaches that of the true model. Therefore as sample size increases model average estimator performs as well as the true model when mean square error is considered.

Comparison with model selection.

In this section, we use the same linear regression setup to perform a simulation study that compares the performance of the frequentist model average estimator with proposed weights with model selection by comparing the coverage probability. The true model from which the data is generated is

$$M_{true} : \boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \boldsymbol{\epsilon}$$

, with $\beta_0 = 2$ and $\beta_1 = 1.5$. Also, we assume $\text{cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = 1/\sqrt{3}$. This simulation is motivated from a similar example in Berk et al. [2000]. We use best subset selection methods for model selection. Best subset model selection method is used to select between: $M_0 : \boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x_1} + \boldsymbol{\epsilon}$ and $M_1 : \boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \boldsymbol{\epsilon}$. Similar results were observed while using AIC/BIC. Our parameter of interest is β_1. In this framework we next compare model averaging with model selection. The comparison is done by varying the values of β_2 over a range and computing the coverage probability of the estimator of β_2. The results are presented in the following table. The results in the following table are based on $m = 100$ and $m = 10000$ simulations. The sample size considered is $n = 10000$.

Table 2.1: Coverage probability for the (a) model average estimator with proposed, weights (b) estimator selected using best subset selection, (c) oracle estimator

m	p	β_2	(a)*Proposed*	(b) Model Selection	(c) Oracle
1000	3	0.07	0.93	0.89	0.95
		0.12	0.92	0.89	0.95
		0.55	0.90	0.85	0.95
		1.1	0.90	0.83	0.95
10000	3	0.045	0.95	0.93	0.95
		0.18	0.94	0.91	0.95
		0.53	0.92	0.89	0.95
		1.21	0.91	0.87	0.95

Comparison with existing frequentist model averaging methods using a linear regression framework.

In this section we use a linear regression model and perform a simulation study that compares the performance of the frequentist model average estimator with proposed weights with existing methods in model averaging. In Hjort and Claeskens [2003] the authors proposed an averaging scheme, Frequentist Model Averaging (FMA) that combines estimators from different models assuming the data is coming from a local misspecification framework. This assumption in turn specifies a set of models that can contribute to the averaging process. In this method any candidate model used has to have a bias of $\mathcal{O}(1/\sqrt{n})$ or less, whereas in our proposed method the choice of candidate models is unrestricted. In Liang et al. [2011] authors proposed a selection of optimal weights to be used in a linear model framework. The idea was to propose an unbiased estimate of MSE of the model average estimator and then minimizing the trace of the MSE estimate the weights were obtained. Their proposed estimator

(OPT) uses this choice of weights and combines all plausible candidate models. This selection of weights has shown to exhibit optimality properties with respect to the mean square error of the estimator.

We use the same linear regression setup as in Section 2.4.1 with three regressor variables (not including the intercept) with unknown parameters $\beta_1, \beta_2, \beta_3$. The possible candidate models can be any of the $2^3 = 8$ parameter combinations possible. Namely, $[M_1] : \boldsymbol{y} = \beta_0\mathbf{1} + \boldsymbol{\epsilon_1}$, $[M_2] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \boldsymbol{\epsilon_2}$, $[M_3] : \boldsymbol{y} = \beta_0\mathbf{1}+\beta_2\boldsymbol{x_2}+\boldsymbol{\epsilon_3}$, $[M_4] : \boldsymbol{y} = \beta_0\mathbf{1}+\beta_3\boldsymbol{x_3}+\boldsymbol{\epsilon_4}$, $[M_5] : \boldsymbol{y} = \beta_0\mathbf{1}+\beta_1\boldsymbol{x_1}+\beta_2\boldsymbol{x_2}+\boldsymbol{\epsilon_5}$, $[M_6] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \beta_3\boldsymbol{x_3} + \boldsymbol{\epsilon_6}$, $[M_7] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_2\boldsymbol{x_2} + \beta_3\boldsymbol{x_3} + \boldsymbol{\epsilon_7}$, $[M_8] : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \beta_3\boldsymbol{x_3} + \boldsymbol{\epsilon_8}$. Here $\boldsymbol{\epsilon_i} \sim N(0, \sigma^2)$ $\forall i = 1, \cdots, 8$ and $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}$ are the regressors.

In order to compare three different methods we use three different simulation setups. In the first setup, the purpose is to evaluate the proposed estimator relative to the FMA estimator when both are considered in a setup where candidate models include the true model from which the data was generated. The second setup, which is based on the same setting, but the choice of true model is different, examines the performance of the FMA estimator when it combines a different, more restricted set of models than the proposed estimator. Finally, in the third setup we choose a much bigger misspcification framework and evaluate the performance of all three estimators.

The true model from which the data is generated is $M_{true} : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2}+\beta_3\boldsymbol{x_3}+\boldsymbol{\epsilon}$, with $\beta_0 = 2$, $\beta_1 = 3$ and $\beta_2 = 1$ In Table 2 we vary the value of β_3 and observe the performance of FMA, OPT and the proposed estimator. The comparison is done by computing the mean square error of the estimator $\hat{\boldsymbol{\mu}}$ as described in Section 3.1. The results in the following table are based on $m = 100$ simulations. The sample size considered is $n = 500$.

Table 2.2: Mean square error for the (a)model average estimator with proposed, weights (b) model average estimator with Liang's [2011] weights, (c)Hjort's [2003] model average estimator with AIC based weights, (d) oracle estimator

β_3	$(a)Proposed$	(b) Liang/MSE	(c) Hjort/AIC	(d) Oracle
0.001	0.00051	0.00051	0.00062	0.00044
0.005	0.00267	0.00267	0.00268	0.00213
0.01	0.00235	0.00234	0.00281	0.00207
0.05	0.00111	0.00111	0.00128	0.00109
0.1	0.00089	0.00089	0.00102	0.00086
0.5	0.00249	0.00249	0.00249	0.00248

We compare the performance of FMA, OPT and the proposed estimator. FMA estimator is based on AIC based weights. From Table 2.2 we can see that the proposed estimator outperforms the FMA estimator, while the OPT estimator shows similar performance as the proposed estimator. We note that the set of candidate models were the same for all three methods, thus this table focuses on comparing the weights that were used to combine the model average estimator in each case. We can see that the proposed weights perform better than the AIC based weights, and are also on par with the MSE based weights from Liang et al. [2011], which were shown to be optimal in the same publication. Also we note that, as the value of β_2 increases all three methods perform similarly. The reason being, when β_2 is large the evidence for $M_{true} : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \beta_3\boldsymbol{x_3} + \boldsymbol{\epsilon}$ gets stronger and all other methods choose weights that favor M_{true}.

Next we focus on comparing the three methods when the set of candidate model varies across them. The setting is similar to that of Table 2.2, but the choice of assumed true model is different. This assumed true model was used during the development of Hjort and Claeskens [2003], where candidate models are assumed to

be in a close neighborhood of the true model. But as in real life we have no idea what this true model should be, in our proposed method we use all possible candidate models that were available. Therefore, we use this setup to examine the performance of the FMA estimator when it combines candidate models under a particular true model assumption. The true model from which the data is generated is $M_{true} : \boldsymbol{y} = \beta_0\mathbf{1}+\beta_1\boldsymbol{x_1}+\beta_2\boldsymbol{x_2}+\beta_3\boldsymbol{x_3}+\boldsymbol{\epsilon}$, with $\beta_0 = 2$, $\beta_1 = 3$ and $\beta_3 = 0.01$. We vary the value of β_2 and compare the performance of the different estimators. But, the FMA estimator operates under the assumed true model, $M'_{true} : \boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \gamma + \boldsymbol{\epsilon}$, with $\gamma = 1$ The results in the following table are based on $m = 100$ simulations as before. From Table 2.3 we can see that the proposed and OPT estimator performs better than the FMA estimator which was based on a restricted set of candidate models.

Table 2.3: Mean square error for the (a)model average estimator with proposed, weights (b) model average estimator with Liang's [2011] weights, (c)Hjort's [2003] model average estimator with AIC based weights, (d) oracle estimator

β_2	(a) Proposed	(b) Liang/MSE	(c) Hjort/AIC	(d) Oracle
0.001	0.00317	0.00317	1.04127	0.00296
0.005	0.00356	0.00356	1.00310	0.00331
0.01	0.00277	0.00277	0.98291	0.00266
0.05	0.00398	0.00398	0.93442	0.00391
0.1	0.00479	0.00479	0.81071	0.00462
0.5	0.00512	0.00512	0.25021	0.00512

Table 2.3 compares the performance of all three methods where the set of candidate models is not same. The AIC based Hjort's estimator is based on a restricted set of candidate models whereas the remaining two methods are based all candidate models available. We can see from the results that the proposed and Liang et al.

[2011]'s model average estimator works well. However, since the true model assumption was wrong, the AIC based model average estimator performs poorly. This is again due to the fact that the estimator was developed using a restriction over the set of candidate models and the true model. Thus, in this section we compare the proposed model average estimator with the existing estimators. Table 2.2 compares the weight choices when set of candidate model is the same and Table 2.3 compares the estimators when the set of candidate models are restricted. From both the tables it is apparent that the proposed estimator works well.

Finally, in Table 2.4 the data are simulated from a model that is not included in the set of candidate models. We use six regressor variables in the true model. The rationale behind this was to evaluate the performance of all three estimators under a setup when the truth is different than the choices considered. The true model contains three additional regressors that are not included in the set of candidate models used in the analysis. We compare the prediction error of all three method with that of the true model. The results in the following table are based on $m = 100$ simulations.

Table 2.4: Prediction error for the (a)model average estimator with proposed, weights (b) model average estimator with Liang's [2011] weights, (c)Hjort's [2003] model average estimator with AIC based weights and (d) oracle estimator from the true model.

β_2	(a) Proposed	(b) Liang/MSE	(c) Hjort/AIC	(d) Oracle
0.001	18.73	18.89	20.17	0.00345
0.005	17.28	17.11	19.34	0.00564
0.01	18.71	18.17	20.13	0.00567
0.05	19.49	19.51	21.16	0.00752
0.1	19.44	18.54	19.33	0.00642
0.5	18.91	19.88	21.86	0.00768

From Table 2.4 we see that none of the estimators are performing well, which was expected as the choices of candidate models were all wrong. However, the proposed method choose the candidate model that is closest to the true model. Also, using the proposed method we can properly estimate the true model parameters that are present in the candidate models considered.

To conclude, from Table 2.2 and 2.3 above we observe that the proposed estimator and OPT performs better than the FMA estimator. In the setup considered in Table 2.3 FMA estimator is combined based on a restricted set of candidate models, whereas the remaining two estimators uses all 8 candidate models. This affects the performance of the FMA estimator as seen in Table 2.3. Also proposed and OPT weights seem to perform better than AIC weights.

Comparison with existing frequentist model averaging methods using a logistic regression framework.

In this section we study the large sample behavior of the model average estimator and the weights chosen using a logistic regression framework. The logit model is given by,

$$p_i = P(y_i = 1|\mathbf{X}) = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}, \quad \forall i = 1, \cdots, n,$$

We use the above logistic regression setup with three regressor variables (not including the intercept) with unknown parameters $\beta_1, \beta_2, \beta_3$ and $n = 500$. The possible candidate models can be any of the $2^3 = 8$ parameter combinations possible, similar to the linear regression setup described in Section 2.4.1. The true model from which the data is generated contains intercept and the regressor $\boldsymbol{x_1}$. Here, $\beta_0 = 2$, $\beta_1 = 3$ and we use different values of β_2 to compare the estimators. In this setup we compare the performance of the proposed estimator along with Hjort's estimator with AIC based weights. As similar to the linear regression case proposed weights outperform

the AIC based weights. The results are based on $m = 100$ simulations.

Table 2.5: Mean square error for the (a)model average estimator with proposed, weights (b)Hjort's [2003] model average estimator with AIC based weights and (c) oracle estimator from the true model.

β_2	(a) Proposed	(b) Hjort/AIC	(c) Oracle
0.001	0.00420	0.00434	0.00389
0.005	0.00214	0.00231	0.00109
0.01	0.00715	0.00819	0.00560
0.05	0.00811	0.00832	0.00655
0.1	0.00912	0.00955	0.00411
0.5	0.00412	0.00447	0.00412

2.4.2 Analysis of Prostate Cancer Data.

The data for this example come from a study by Stamey et al. [1989]. They examined the relationship between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. The variables are log cancer volume (lcavol), log prostateweight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp),Gleason score (gleason), and percent of Gleason scores 4 or 5. Here svi is a binary variable, and gleason is an ordered categorical variable. The model selection results are based on a best-subset selection using an all-subsets search. From this we obtain an estimated prediction error that we use for comparing model selection estimate with model averaging estimate. The data is divided randomly into a training set of size 67 and a test set of size 30. We repeat the test and training breakup 5 times and average over the results. Best-subset selection selected a model containing two predictors lcvol and lweight.

Using model averaging on the dataset the weights were computed as follows. The models assigned the most weights were with features lcavol, lweight, svi, pgg45, lcp, gleason and lbph and the model with lcavol, lweight, svi, pgg45, lcp, gleason, lbph and age. Whereas AIC dependent weights gives more weight to a smaller model containing lcavol and lweight. The 90% prediction inteval for antigen levels in given below. The prediction interval was computed for one set of test data.

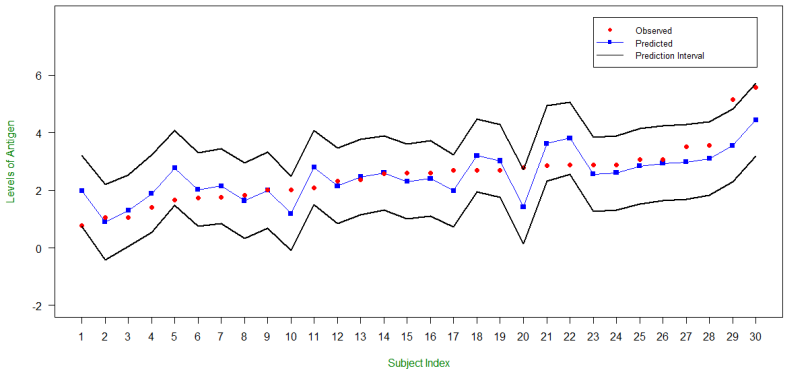

Figure 2.2: Observed value and predicted interval for antigen level

2.5 Discussion

A model averaging estimator incorporates model uncertainty into the analysis by combining a set of competing candidate models rather than choosing just one. It also provides an insurance against selecting a poor model thus improving the risk in estimation. In Hjort and Claeskens [2003] the authors proposed a formal framework for frequentist model averaging as detailed before. In Hjort and Claeskens [2006] and Claeskens and Hjort [2008] variable selection methods for the Cox proportional hazards regression model were discussed along with the choice of weights. In Hansen [2007] a new set of weights were derived using Mallow's criterion. In Liang et al. [2011], the authors proposed an unbiased estimator of the risk and a set of optimal weights were chosen by minimizing the trace of the unbiased estimator. Further

details about model selection and averaging can also be found in Lien and Shrestha [2005], Karagrigoriou et al. [2009] and Wei and McNicholas [2012]. Recently, with the development of theory of model averaging, it has been used in many areas of application. The application of Frequentist model selection and weighting schemes have been a focus of discussion in Bates and Granger [1969], where the authors used it for forecasting airline passenger data. Similar approach was used in Danilov and Magnus [2004b,a] for forecasting stock market data. Pesaran et al. [2009] discusses dealing with the risk of using false models in portfolio management.

In this paper, we propose a more general framework where the choice of true model is not fixed. The truth can be any one or a mixture of the candidate models. Models that have large biases are not excluded from our analysis. We also study the behavior of frequentist model average estimator with an optimal weighting scheme to combine all the individual candidate models. As an illustration, we derive the model average estimator in the linear regression framework. The asymptotic distribution for model average estimator is also given. A linear regression model setup is used to simulate different scenarios to compare the performance of the proposed model average estimator with existing methods. Mean square error of the estimator is used for the purpose of comparison. We also implement the weighting scheme proposed by Liang et al. [2011] and compare their performance to AIC based weights. The simulation results indicate that under certain model specifications, the proposed estimator works better than Hjort and Claeskens [2003]'s estimator. And in some cases, the proposed estimator works better even than the estimator which is based on the model from which the data is actually generated. If the model average estimator follows an asymptotic normal distribution, as discussed in the main results, then one can construct confidence intervals based on Theorem 2.1 for model average estimators. If we could specify an asymptotically correct estimator for the variance of the model average estimator, one could propose a theoretically correct construction for

such confidence intervals.

There are many ways a regression model can be misspecified. The functional form of the model may not be correctly specified or there may be dependencies among the predictor variables. Misspecification in most cases is often interpreted as a case of left out variables. In these instances, the normality assumption among random errors are violated. This results in the estimates being biased as discussed in Giles et al. [1992]. These estimates can harm the decision making process, so one should be very attentive while fitting and choosing models in the presence of misspecification. Many methods have been used to measure and limit misspecification in model fitting. Ramsey Regression Equation Specification Error Test (RESET), discussed in Thursby and Schmidt [1977] being a test that is useful in a linear regression setup.

In model averaging, if the true model is not included in the set of candidate models, we end up using an estimate that is biased. If all the models are misspecified, the weights derived by AIC or by using a consistent or unbiased estimator of mean square error are not optimal and should be used after careful consideration. When the true model is not included in the analysis thus all the candidate models are wrong, there have been developments in model selection that takes care of the bias resulting from selection. See Hurvich and Tsai [1989, 1991]. A penalized version of AIC and BIC have been derived that performs better than other selection criteria. One can follow a similar path and derive the model averaging weights based on a sightly modified criteria.

Another problem with model averaging is that the number of optional parameters in analysis could be very high. For example, if there are 30 parameters we could end up using as many as 2^{30}candidate models. This may be time consuming and not ideal in certain fields of study. However, as suggested in this paper, a statistician can choose to use all or very few candidate models as per the scope of the study. This could be explored in further development.

2.6 Appendix

2.6.1 Regularity Conditions and Assumptions.

In this section we state the regularity conditions that were used throughout the paper. We assume that the density function satisfies the following conditions.

(a) Θ is an open subset of $\mathbb{R}^p$, and the support of the density function is independent of $\boldsymbol{\beta}$.

(b) The true parameter value is an interior point of the parameter space.

(c) $\ell'_{k;i}$ and $\ell''_{k;i}(\boldsymbol{\beta}^*_k)$ exists and $\ell'_{k;i}$ is a continuous function of $\boldsymbol{\beta}$.

(d) $\mathbb{E}[\ell'_{k;i}] = 0$ and $\mathbb{E}[\ell'_{k;i}\ell'^T_{k;i}] = -\mathbb{E}[\ell''_{k;i}(\boldsymbol{\beta}^*_k)]$. These conditions are standard conditions for asymptotic normality of maximum likelihood estimators.

(e) $\lim_{n\to\infty} \frac{1}{n}\left[\ell''_k(\boldsymbol{\beta}^*_k)\right] \to \mathbf{H}_k$ and $\mathbf{H}_k$ is positive definite.

(f) For some $\epsilon > 0$, $\sum_i \mathbb{E}|\lambda'\ell'_{k;i}(\boldsymbol{\beta}_{true})|^{2+\epsilon}/n^{(2+\epsilon)/2} \to 0$ for all $\epsilon \in \mathbb{R}^p$.

(g) There exists $\epsilon > 0$ and random variables $B_i(y_i)$ $sup\left\{|\ell''_{k;i}(\boldsymbol{\beta}^*_k)| : ||t - \boldsymbol{\beta}_{true}|| \le \epsilon\right\} \le B_i(y_i)$ and $\mathbb{E}|B_i(y_i)|^{1+\delta} \le K$, where δ and K are positive constants.

Consider a functional $\mu : \mathbb{R}^{p+q} \to \mathbb{R}$. Define $\mu^{(dropped)} : \mathbb{R}^{p+m} \to \mathbb{R}$ as the same function as μ with only the $(q-m)$ corresponding arguments dropped. For any $\boldsymbol{b} = (b_1, \cdots, b_p, b_{p+1}, \cdots, b_{p+m})$ with $1 \le m \le q$ define the $\boldsymbol{c}$-*augmented* version of $\boldsymbol{b}$ as $\widetilde{\boldsymbol{b}} = \{\boldsymbol{b}, \boldsymbol{c}\} \in \mathbb{R}^{p+q}$ with some fixed $\boldsymbol{c} \in \bar{\mathbb{R}}^{q-m}$ inserted at the place of missing components. Let the indices of the missing components be $\{p + i_1, \cdots, p + i_{q-m}\}$. We define $\widetilde{\mu} : \mathbb{R}^{p+m} \to \mathbb{R}$ as the restriction of $\mu : \mathbb{R}^{p+q} \to \mathbb{R}$ subject to $b_{p+i_1} = c_1, \cdots, b_{p+i_{q-m}} = c_{q-m}$. Clearly then $\mu(\widetilde{\boldsymbol{b}}) = \widetilde{\mu}(\boldsymbol{b})$. Given a function μ, the fixed value $\boldsymbol{c}$ is chosen in such a way that $\mu(\widetilde{\boldsymbol{b}}) = \mu^{(dropped)}(\boldsymbol{b})$. We assume that $\boldsymbol{\mu} : \mathbb{R}^{p+q} \to \mathbb{R}^\ell$ be a function that is 1^{st} order partially differentiable at $\boldsymbol{\beta}_{true}$. Note that by definition of

$\boldsymbol{c}$-augmentation, $\mu(\widetilde{\boldsymbol{\beta}}_k) = \mu^{(dropped)}(\widehat{\boldsymbol{\beta}}_k)$. For ease of reading, in the subsequent proof, we omit the superscript '$(dropped)$'.

2.6.2 Proofs of Theorems.

Proof of Theorem 2.1. From usual regularity conditions on the log-likelihood, as related to M-estimation it can be shown that $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\right) = -\mathbf{H}_k^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \ell'_{k;i}(\boldsymbol{\beta}_k^*)\right] + o_{\mathbb{P}}(1)$. For more detail and exact conditions see [Van der Vaart, 2000, Chapter 5]. Now by application of Taylor expansion,

$$\mu(\widehat{\boldsymbol{\beta}}_k) - \mu(\boldsymbol{\beta}_k^*) = \nabla\mu(\boldsymbol{\beta}_k^*)^T(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*) + o_{\mathbb{P}}(\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|),$$

$$\sqrt{n}(\mu(\widehat{\boldsymbol{\beta}}_k)-\mu(\boldsymbol{\beta}_k^*)) = -\nabla\mu(\boldsymbol{\beta}_k^*)^T\left\{\mathbf{H}_k^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \ell'_{k;i}(\boldsymbol{\beta}_k^*)\right] + o_{\mathbb{P}}(1)\right\} + o_{\mathbb{P}}(\sqrt{n}\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|).$$

Thus it follows that for $0 \le w_k \le 1$ with $\sum_{k\in\mathcal{M}} w_k = 1$,

$$\begin{aligned}
&\sqrt{n}\sum_{k\in\mathcal{M}} w_k(\mu(\widehat{\boldsymbol{\beta}}_k) - \mu(\boldsymbol{\beta}_{true})) \\
&= \sqrt{n}\sum_{k\in\mathcal{M}} w_k(\mu(\boldsymbol{\beta}_k^*) - \mu(\boldsymbol{\beta}_{true})) + \sqrt{n}\sum_{k\in\mathcal{M}} w_k(\mu(\widehat{\boldsymbol{\beta}}_k) - \mu(\boldsymbol{\beta}_k^*)) \\
&= \sqrt{n}\sum_{k\in\mathcal{M}} w_k(\mu(\boldsymbol{\beta}_k^*) - \mu(\boldsymbol{\beta}_{true})) - \sum_{k\in\mathcal{M}} w_k\nabla\mu(\boldsymbol{\beta}_k^*)^T\mathbf{H}_k^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \ell'_{k;i}(\boldsymbol{\beta}_k^*)\right] \\
&+ o_{\mathbb{P}}\left(\sum_{k\in\mathcal{M}} \sqrt{n}\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|\right) \\
&= \sqrt{n}\sum_{k\in\mathcal{M}} w_k(\mu(\boldsymbol{\beta}_k^*) - \mu(\boldsymbol{\beta}_{true})) + \frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{-\sum_{k\in\mathcal{M}} w_k\nabla\mu(\boldsymbol{\beta}_k^*)^T\mathbf{H}_k^{-1}\ell'_{k;i}(\boldsymbol{\beta}_k^*)\right\} \\
&+ o_{\mathbb{P}}\left(\sum_{k\in\mathcal{M}} \sqrt{n}\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|\right) \\
&= \sqrt{n}\sum_{k\in\mathcal{M}} w_k(\mu(\boldsymbol{\beta}_k^*) - \mu(\boldsymbol{\beta}_{true})) + \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i + o_{\mathbb{P}}\left(\sum_{k\in\mathcal{M}} \sqrt{n}\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|\right),
\end{aligned}$$

where we have used the definition that $Z_i = -\sum_{k\in\mathcal{M}} w_k \nabla\mu(\boldsymbol{\beta}_k^*)^T \mathbf{H}_k^{-1} \ell'_{k;i}(\boldsymbol{\beta}_k^*)$. First note that $\sqrt{n}\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\| = o_{\mathbb{P}}(1)$ via consistency of MLE. Note that Z_i's are independent and $\mathbb{E}Z_i = 0$. Now fix $\epsilon > 0$. In order to prove the asymptotic normality of the quantity $(1/\sqrt{n})\sum_i Z_i$ we invoke the Lindeberg-Feller central limit theorem (see Billingsley [2008]). This requires verification of the so called Lindeberg condition, given by $(1/n)\sum_{i=1}^n \mathbb{E}Z_i^2 \mathbb{I}\{|Z_i| > \sqrt{n}\epsilon\}$. Let us denote $Y_{ki} = \nabla\mu(\boldsymbol{\beta}_k^*)\mathbf{H}_k^{-1}\ell'_{k;i}$. Now,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbb{E}Z_i^2 \mathbb{I}\left\{|Z_i| > \sqrt{n}\epsilon\right\} &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\underbrace{\left(\sum_{k\in\mathcal{M}} w_k Y_{ki}\right)^2}_{=A,\text{ say}} \underbrace{\mathbb{I}\left\{|\sum_{k\in\mathcal{M}} w_k Y_{ki}| > \sqrt{n}\epsilon\right\}}_{=B,\text{ say}} \\
&\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\sum_{k\in\mathcal{M}} w_k Y_{ki}^2 \, \mathbb{I}\left\{\max_{k\in\mathcal{M}} |Y_{ki}| > \sqrt{n}\epsilon\right\}\right] \\
&\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\max_{k\in\mathcal{M}} |Y_{ki}|^2 \mathbb{I}\left\{\max_{k\in\mathcal{M}} |Y_{ki}| > \sqrt{n}\epsilon\right\}\right].
\end{aligned}$$

Here the inequality in the second line is derived by first noting that if $A, B > 0$ and $A < C, B < D$, then $AB < CD$. Secondly, note that $A = (\sum_{k\in\mathcal{M}} w_k Y_{ki}) \leq \sum_{k\in\mathcal{M}} w_k Y_{ki}^2$ by Jensen's inequality. Also since

$$\sqrt{n}\epsilon < |\sum_{k\in\mathcal{M}} w_k Y_{ki}| \leq \max_{k\in\mathcal{M}} \sum_k |w_k| = 1,$$

it follows that

$$\mathbb{I}\left\{|\sum_{k\in\mathcal{M}} w_k Y_{ki}| > \sqrt{n}\epsilon\right\} \leq \mathbb{I}\left\{\max_{k\in\mathcal{M}} |Y_{ki}| > \sqrt{n}\epsilon\right\}.$$

Now take $C = \sum_{k\in\mathcal{M}} w_k Y_{ki}^2$ and $D = \mathbb{I}\{\max_{k\in\mathcal{M}} |Y_{ki}| > \sqrt{n}\epsilon\}$. Now by condition (2.2.2), the Lindeberg-Feller condition is satified for $(1/\sqrt{n})Z_i$'s whence it follows

that $(1/\sqrt{n})\sum_{i=1}^{n} Z_i \sim \mathcal{N}(0, \sigma_w^2)$, where σ_w^2 is given by

$$\sigma_w^2 = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\sum_k w_k \nabla\mu(\boldsymbol{\beta}_k^*)^T \mathbf{H}_k^{-1} \ell'_{k;i}\right]^2.$$

The theorem follows. ■

Proof of Corollary 2.2. As defined before, for the k^{th} candidate model, let $\boldsymbol{\beta}_k^* \in \mathbb{R}^{p+|M_k|}$ be the solution of the equation $\mathbb{E}S_k(\boldsymbol{\beta}) = 0$, where $S_k(\boldsymbol{\beta})$ is the score function for the k^{th} model. Let $\boldsymbol{\beta}_{0,k} = (\boldsymbol{\theta}_0, \pi_k\boldsymbol{\gamma}_0)^T \in \mathbb{R}^{p+|M_k|}$. Therefore, $\mathbb{E}(\ell'_k(\boldsymbol{\beta}_k^*)) = \mathbf{0}$. Then, by Taylor's theorem and appropriate regularity conditions on the density function, it follows that asymptotically, $\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_{0,k} \approx \mathbf{J}_k^{-1}\mathbb{E}(\ell'_k(\boldsymbol{\beta}_0))$. Now note that following [Hjort and Claeskens, 2003, Page 37],

$$\mathbb{E}(\ell'_k(\boldsymbol{\beta}_0)) = \begin{pmatrix} \mathbf{J}_{01}\boldsymbol{\delta}/\sqrt{n} + o(1/\sqrt{n}) \\ \pi_k\mathbf{J}_{11}\boldsymbol{\delta}/\sqrt{n} + o(1/\sqrt{n}) \end{pmatrix},$$

so that,

$$\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_{0,k} \approx \mathbf{J}_k^{-1}\begin{pmatrix} \mathbf{J}_{01}\boldsymbol{\delta}/\sqrt{n} \\ \pi_k\mathbf{J}_{11}\boldsymbol{\delta}/\sqrt{n} \end{pmatrix}. \tag{2.6.1}$$

In order to prove the corollary, we first match the bias terms. Note that in Theorem 2.1, the bias term is given by

$$\sqrt{n}\sum_{k\in\mathcal{M}} w_k(\mu(\boldsymbol{\beta}_k^*, \gamma_{0,k^c}) - \mu(\boldsymbol{\beta}_{true})).$$

Thus consider term buy term, the bias of the k^{th} component is given by

$$\begin{aligned}\sqrt{n}(\mu(\boldsymbol{\beta}_k^*, \boldsymbol{\gamma}_{0,k^c}) - \mu(\boldsymbol{\beta}_{true})) &= \sqrt{n}(\mu(\boldsymbol{\beta}_k^*, \boldsymbol{\gamma}_{0,k^c}) - \mu(\boldsymbol{\beta}_0)) - \sqrt{n}(\mu(\boldsymbol{\beta}_{true}) - \mu(\boldsymbol{\beta}_0)) \\ &\approx \sqrt{n}(\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_{0,k})^T \begin{pmatrix} \partial\mu(\boldsymbol{\beta}_0)/\partial\boldsymbol{\theta} \\ \partial\mu(\boldsymbol{\beta}_0)/\partial\boldsymbol{\gamma}_k \end{pmatrix} - \left(\frac{\partial\mu(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\gamma}}\right)^T \boldsymbol{\delta} \\ &= \begin{pmatrix} \partial\mu(\boldsymbol{\beta}_0)/\partial\boldsymbol{\theta} \\ \partial\mu(\boldsymbol{\beta}_0)/\partial\boldsymbol{\gamma}_k \end{pmatrix}^T \mathbf{J}_k^{-1} \begin{pmatrix} \mathbf{J}_{01}\boldsymbol{\delta} \\ \pi_k \mathbf{J}_{11}\boldsymbol{\delta} \end{pmatrix} - \left(\frac{\partial\mu(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\gamma}}\right)^T \boldsymbol{\delta},\end{aligned}$$

where the last term follows from (2.6.1). This matches the bias term in (2.2.7). Looking at the variance term, note that from (2.3.2), the variance of the k^{th} term is given by,

$$var(\nabla\mu(\boldsymbol{\beta}_k^*, \boldsymbol{\gamma}_{0,k^c})^T \mathbf{H}_k^{-1} (\sum_{i=1}^n \ell'_{k;i}(\boldsymbol{\beta}_k^*)/\sqrt{n}).$$

From (2.6.1), via Taylors theorem it follows that $\nabla\mu(\boldsymbol{\beta}_k^*, \boldsymbol{\gamma}_{0,k^c}) \approx \nabla\mu(\boldsymbol{\beta}_0)$. Also note that from standard theory of maximum likelihood estimation,

$$\begin{aligned}\mathbf{H}_k^{-1}(\boldsymbol{\beta}_k^*)(\sum_{i=1}^n \ell'_{k;i}(\boldsymbol{\beta}_k^*)/\sqrt{n}) &\approx \sqrt{n}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*) \\ &= \sqrt{n}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{0,k}) - \sqrt{n}(\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_{0,k}) \\ &= \mathbf{J}_k^{-1} \begin{pmatrix} \sqrt{n}\overline{U}_n \\ \sqrt{n}\overline{V}_{n,k} \end{pmatrix} - \mathbf{J}_k^{-1} \begin{pmatrix} \mathbf{J}_{01}\boldsymbol{\delta} \\ \pi_k \mathbf{J}_{11}\boldsymbol{\delta} \end{pmatrix} \\ &= \mathbf{J}_k^{-1} \begin{pmatrix} \sqrt{n}(\overline{U}_n - \mathbb{E}U_k(Y_1)) \\ \sqrt{n}(\overline{V}_{n,k} - \mathbb{E}V_k(Y_1)) \end{pmatrix}.\end{aligned}$$

Here the last inequality follows from Lemma 3.1 in Hjort and Claeskens [2003]. Hence it follows that asymptotically both the bias and variance terms are equal. ■

Chapter 3
Dose Finding In Combination Therapy

3.1 Introduction

With the development of new drugs in pharmaceutical industry, clinical trials involving treatments that combine multiple drugs are also increasing in number. Sometimes even if a drug fails to exhibit efficacy when used in isolation, it may demonstrate increased efficacy when used in combination with some other drug. Multiple drugs used together can interact and may enhance the effectiveness of the treatment. Drug synergy can occur because of biological activity or pharmacokinetics. One may also see negative effects of synergy that might amplify the side effects and result in a safety issue. The goal of combination therapy is to achieve better patient response, particularly for cancer patients who are non respondent to conventional single agent therapies. In oncology, for example, multiple drugs used together can induce a synergistic treatment effect by targeting different pathways.

In phase I clinical trials, we often focus on understanding the toxicity profile of the drugs studied. Multiple dose levels are used in the trial and the toxicity observed is studied. Often the purpose of a phase I trial is to find the maximum tolerated dose (MTD) of the drug under study. The MTD is the highest possible dose that does not exhibit an unacceptable amount of toxicity in the subjects. Thus often this dose (MTD) is the one with a probability of toxicity that is closest to the trial's pre specified target. In a single drug trial finding MTD could be rule based or model based. In a model based approach finding MTD involves modeling a

dose toxicity relationship. Here, one can either use a model based on dose levels or use pre specified dose toxicity probabilities. In continual reassessment method (CRM) O'Quigley et al. [1990] a parametric function is used to model the relationship between the true dose toxicity probabilities and the pre specified toxicity probabilities. This dose toxicity relationship is then modelled using a Bayesian adaptive design that updates the relationship based on the information received from the clinical trial.

Finding the MTD in a drug combination trial with two or more drugs is different and more complicated than a single drug trial. In a single drug trial often toxicity is assumed to be monotonically increasing with dose level, which helps to identify the unique MTD. But in a combination trial multiple dose combinations can have the same levels of toxicity. Also information about drug - drug interaction is not known, which may lead to unknown patterns of toxicity. Thus ordering the dose level combinations is more complicated when dealing with multiple drugs. Finding the combination dose levels with similar level of toxicity may involve investigating numerous dose combinations, which, in real life may be complicated because of the small number of subjects present in the trial. In theory, in a multiple drug setting, an infinite number of possible dose combinations may achieve the same target toxicity level, as we assume dose-toxicity surface is continuous. In practice, however, such choices are often restricted by pre defining a set of dose level combinations used in the trial. Different approaches for finding dose combinations have been discussed in many recent literature. Yin and Yuan [2009] proposed copula regression models for analyzing dose combination. Thall and Cook [2004] proposed a six-parameter model to define the probability of toxicity that had properties similar to that of logistic regression. Further details about dose finding designs can be found in Staw and Ross [1987], Thall et al. [1999], Thall and Cook [2004], Zhou [2010], Huo et al. [2012], Sweeting and Mander [2012] etc. Using Bayesian parametric models to describe the dose-toxicity surface in clinical trials has become increasingly popular. Often adaptive

designs are used for the purpose as one can update the model as soon as new data are collected. In this paper, we apply the same idea in a two dimensional dose toxicity analysis. A combination model that is able to capture different natures of synergy and antagonism has been used for analysis. We model the rates of toxicity of the combined drugs via a copula-type regression. We also use an adaptive design with a hierarchical model where the model parameters can be separated into those that relate to the marginal dose-toxicity response and those that relate to the interaction between the two drugs. We use a three-parameter copula-type regression model stated in Yin and Yuan [2009], that can be treated as an extension of the popular continuous reassessment method (CRM) used in single-drug dose-finding trials.

Different trial designs involving drug combinations can be used to find the optimal dose level based on both safety and efficacy of the drugs. Also different escalation strategies may result in different MTDs being identified and recommended for later phases, where by utilizing efficacy data an optimal dose level is selected. In this paper, we use a strategy for dose escalation and also explore the dose-toxicity space for proposing future dose levels. Rest of this paper is detailed as follows. The model framework used in the analysis is described in Section 2. We also look into an extension where three drugs are combined together. Section 3 details a simulation study. In Section 4 we present a case study that is developed by using the data from oncology clinical trials. This case study includes two individual drug trials as well as the combination trial with both the drugs used together.

3.2 Parametric Model for Dose Combination

3.2.1 Joint Toxicity Model

We follow a model setup similar to Yin and Yuan [2009] for combined toxicity modelling. Let A and B be two drugs used in a drug combination trial. In a drug

combination study, the probability of toxicity corresponding to doses can be specified in advance based on prior information available (See section 2.2 for details). Let p_i be the pre specified probability of toxicity corresponding to A_i, the ith dose for drug A, and q_j be that of B_J, the jth dose for drug B. To accommodate uncertainty in the marginals we introduce probabilities of toxicity for drug A and drug B respectively as p_i^{α} and q_j^{β}, where $\alpha \geq 0$ and $\beta \geq 0$ are unknown parameters characterizing the effect of the individual drugs. However, when two or more drugs are combined, each drug would not act independently and there would be a drug drug interactive effect. This interaction would affect the joint toxicity profile.

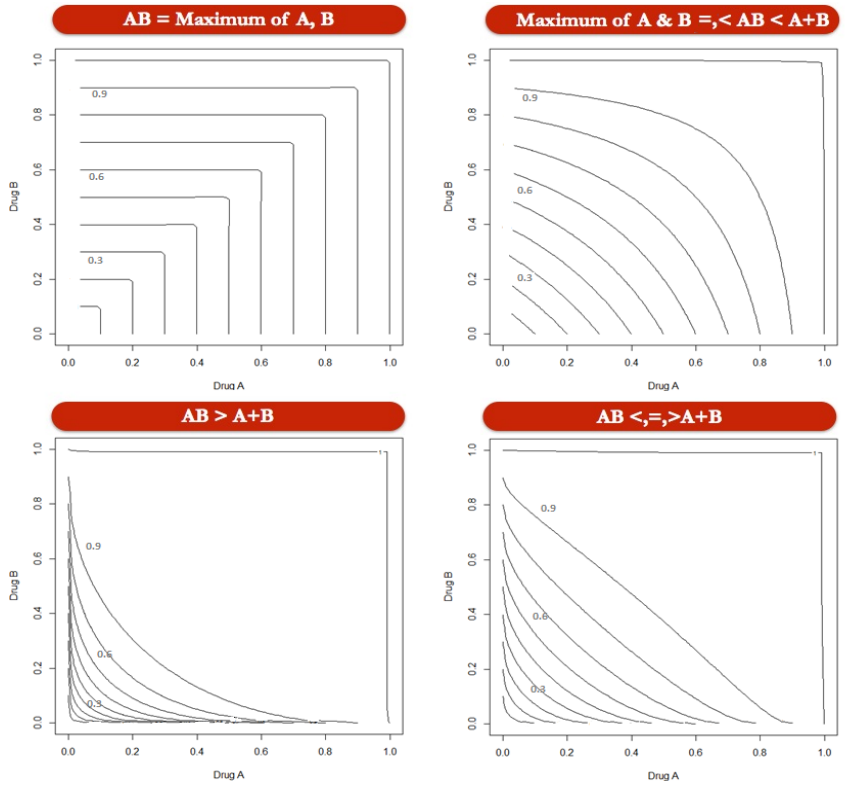

Figure 3.1: Different toxicity contours for combination of drug A and B

Figure 1 above shows different type of combined effects for drug A and B. For a combination model it is desired that the joint toxicity function would be able to capture both synergy and antagonism if present in the data. As seen in Figure 1 in the

third and fourth contour plots the combined toxicity is much worse than individual ones, which might be the scenario in real life.

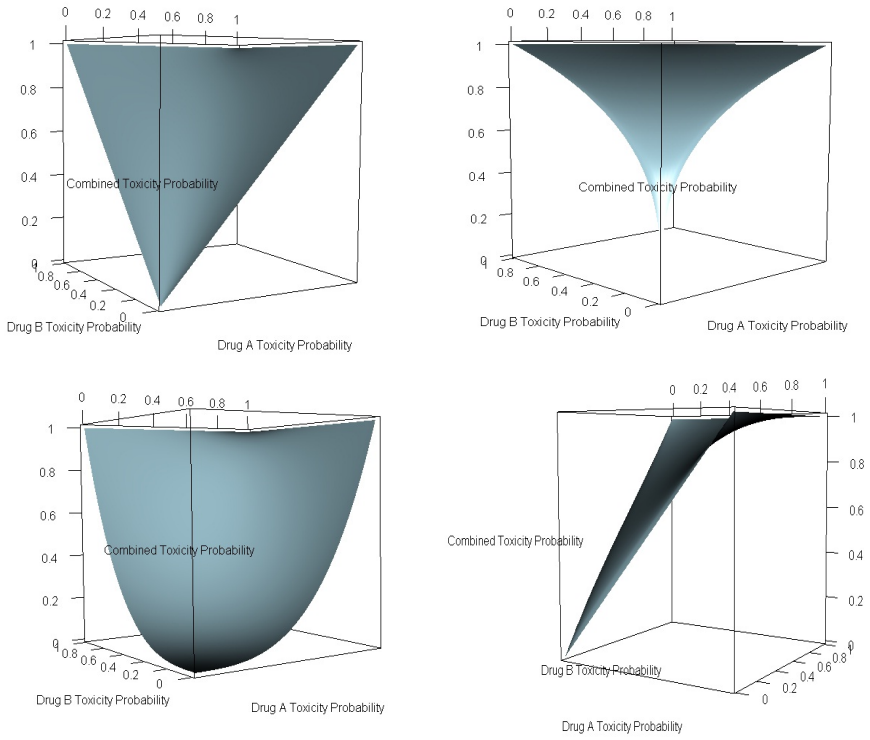

Figure 3.2: Different toxicity surface for the two-drug combination

In Figure 2, we illustrate the joint toxicity probability surface based on model in the two-dimensional probability space. Depending on the three parameters α, β, γ the toxicity probability surface may have various shapes. Therefore, it is critical to choose an appropriate model to link the joint toxicity probability π_{ij} with the p_i and q_j. In a combination trial, we often observe a binary toxicity outcome for the drug combination. Thus, for a subject treated at (A_i, B_j), we have information about toxicity indicator being present or not. Our model links the joint toxicity probability π_{ij} to the individual probabilities of toxicity p_i^{α} and q_j^{β} through a copula-type regression model. Copula models are a widely used class of models that define the joint probability distribution based on the marginal distributions and a dependence parameter. We use the following model as given in Yin and Yuan [2009] for drug

combinations given by the Gumbel Hougaard copula

$$\pi_{ij} = 1 - \exp(-[\{-log(1-{p_i}^{\alpha})\}^{1/\gamma} + \{-log(1-{q_j}^{\beta})\}^{1/\gamma}]^{\gamma})$$

where as mentioned before p_j is the pre-specified best guess toxicity probability for drug A, q_j is the pre-specified best guess toxicity probability for drug B and α and β characterizes the individual drug effects and γ characterizes drug-drug interactive effect respectively.

This copula model usually satisfies some basic model conditions. For example, if we have no drug present, the joint toxicity probability should be 0. Also if the probability of toxicity of one drug is 0 (which would be a individual drug trial) the joint toxicity probability would be the toxicity of the other drug. Or, if the toxicity probability of either drug is 1, the joint toxicity probability should be 1. This model can also capture different dose toxicity contours as shown in Figure 1. By only changing the drug drug interaction parameter γ one can model different synergy effects (with the other two parameters fixed). This makes the model easy to interpret, and computationally easier with only three parameters present. Moreover, if only one drug is tested, say $p_i > 0$ and $q_j = 0$ the above model reduces to the CRM, with $\pi_{ij} = p_i^{\alpha}$. We can use other copula models as well, depending on the focus of the study and computational convenience.

Next, one can construct the likelihood function on the basis of the binomial distribution with the probabilities π_{ij}. Suppose that, at the current stage of the trial, among n_{ij} subjects treated at the combined dose level (A_i, B_j), x_{ij} subjects have reported toxicity. The likelihood is then given by

$$l(\alpha, \beta, \gamma | data) \propto \prod_i \prod_j \pi_{ij}^{x_{ij}} (1-\pi_{ij})^{n_{ij}-x_{ij}}.$$

Assuming prior distributions of the model parameters are independent, The joint

posterior distribution is given by

$$\pi(\alpha, \beta, \gamma|data) \propto l(\alpha, \beta, \gamma|data)\pi(\alpha)\pi(\beta)\pi(\gamma).$$

From the joint posterior distribution we can obtain the full conditional distributions of the model parameters. After the outcomes of each cohort of subjects have been observed, one can use the Gibbs sampling algorithm to sample from the posterior distributions of the unknown parameters. Thus, the posterior estimates can be easily obtained, which could lead to the next stage of the trial design.

3.2.2 Marginal Probabilities for Dose Combination

We perform a dose response/toxicity study for each individual drug and obtain the marginal toxicity from this analysis. This will allow us to establish a relationship between dose and toxicity for each drug considered. Developing a model based on the data from individual study also makes the marginals more robust, as oppose to using the actual data itself. Here we assume that toxicity increases as dose increases for each drug.

There are many models that are used as popular choices for modeling the dose response relationship in clinical trials. Here the expected response is assumed monotonically related to dose, and there exists a lower and upper asymptote for the expected response. These properties are basic desirable features for modeling many clinical trial dose response curves. In practice, these models have been found to fit well to many data sets. One example of a model with similar properties is the Emax model.

We illustrate our modeling technique using the Emax model for dose response.

The model is given by,

$$\mu_{ij} = E_0 + \frac{E_{max(i)} \times dose_j}{Ed_{50} + dose_j} + \epsilon$$

where ϵ is $N(0, \sigma^2)$ and Y denotes a response of interest, E_0, E_{max}, ED_{50} are unknown parameters and ϵ is a random error assumed to be normally and independently distributed with constant variance.

Emax could be set different for study characteristics that are of interest (ie, population, etc). In a cancer trial, for example Emax could set to be different for each cancer type. Then for the ith cancer type and jth dose level, out of n_{ij} subjects, if y_{ij} have been identified with toxicity. We have, $y_{ij} \sim Bin(p_{ij}, n_{ij})$. We use Emax model to estimate the probabilities of the distribution given above. To achieve this goal we utilize the logit function as follows

$$logit(p_{ij}) = log(\frac{p_{ij}}{1 - p_{ij}}) = \mu_{ij}.$$

Here E_0 represents the placebo response, E_{max} represents the maximum possible response as dose approaches infinity, ED_{50} is the dose that produces half of the E_{max} effect.

Other than obtaining marginal toxicity probability for each individual drug this method also allows us to combine data across different type of cancer or different trials. Since in clinical trials data is scarce pulling information from multiple studies can help building the combined model. Combining data across trials with a different emax parameter takes account the variation among different type of cancer combined. At the same time it borrows information across different types and makes the model more robust.

3.2.3 Joint Toxicity Model for more than two drugs

In practice, drug combination trials could involve more than a pair of drugs, each with several pre specified doses. When the dimension of the drug combination space is higher than 2, dose finding becomes a much more complicated process, for which most of the currently available methods might not work well. This method, however, can be easily generalized to such a high dimensional dose combination problem. For example, if three drugs are combined in a trial, we denote p_i to be the physician-specified probability of toxicity for the ith dose of drug A, $i = 1, ..., I$ q_j to be that for the jth dose of drug B, $j = 1, ..., J$ and r_k to be that for the kth dose of drug C, $k = 1, ..., K$ i.e. the triplet p_i, q_j, r_k represents the pre specified probabilities of toxicity that are associated with the combined drug doses A_i, B_j, C_k. By incorporating a power parameter for each prior probability of toxicity, the true probabilities of toxicity are p_i^{α} , q_j^{β},r_k^{ψ}, where α, β, ψ characterize individual drug effects. In this three-dimensional toxicity probability space, we can still quantify the joint toxicity probabilities through a copula-type model. Then, we have

$$\pi_{ijk} = 1 - \exp(-[\{-log(1 - {p_i}^{\alpha})\}^{1/\gamma} + \{-log(1 - {q_j}^{\beta})\}^{1/\gamma} + \{-log(1 - {r_k}^{\psi})\}^{1/\gamma}]^{\gamma})$$

where γ is the parameter characterizing interaction effect.

However, sometimes in a clinical trial a third drug (drug C) is combined to an existing study of two drugs (drug A and B). In a case like that, we can utilize the original copula model and treat the first combination toxicity probabilities as marginal. Therefore, from the existing data, one can develop the pre specified marginal toxicity profile of A and B. Then this marginal could be used to delevop the three drug combination model, by using the copula function. Then, we have

$$\pi_{ijk} = 1 - \exp(-[\{-log(1 - {p_{ij}}^{\delta})\}^{1/\gamma} + \{-log(1 - {r_k}^{\psi})\}^{1/\gamma}]^{\gamma}).$$

Where p_{ij} is the pre-specified toxicity probabilities when drugs A and B are used in combination. r_k is the pre-specified probabilities for drug C. The corresponding likelihood based on the binomial distribution is given by

$$\prod_i \prod_j \prod_k \pi_{ijk}^{x_{ijk}} (1 - \pi_{ijk})^{n_{ijk} - x_{ijk}}$$

where, among n_{ijk} subjects treated at A_i, B_j, C_k, x_{ijk} subjects have experienced toxicity. Prior specifications and posterior derivations are similar to that of the two drug combination.

3.2.4 Combination for Ordinal Toxicity Measures

We can modify the parametric model chosen for combination of drugs when the toxicity variable is ordinal in nature. In Houede et al. [2010] authors discussed modelling toxicity in combination with ordinal variable. Here we use a similar idea and model the conditional distribution instead of the joint distribution by using the copula regression function. We first model the marginal distribution of each ordinal outcome as a function of dose d. For the marginal distributions, we will use a copula model as used in the section 2.1. For the effects of the dose i of drug A and dose j of drug B, we define the copula model as stated earlier by using the probability function

$$\omega_{ij}^y = 1 - \exp(-[\{-log(1 - p_i(y)^\alpha)\}^{1/\gamma} + \{-log(1 - q_j(y)^\beta)\}^{1/\gamma}]^\gamma)$$

denoted as $\omega^y(\theta, d)$, where θ is the set of parameters (α, β, γ) and d is the dose level combination for drug A and B (i, j). Y is the ordinal variable with the ordinal outcomes denoted by y. For example, let us consider a toxicity variable with four

levels,

$$\begin{aligned} Y = 0 \quad & \text{No SAE} \\ = 1 \quad & \text{Grade1} - 2 \\ = 2 \quad & \text{Grade3} - 4 \\ = 3 \quad & \text{Grade5} \end{aligned}$$

where SAE is serious adverse events occurring during the dose response study. Before we specify marginal probabilities for the ordinal outcomes $Y = y$, we first define the conditional probabilities,

$$Pr(Y_k \geq y | Y_k \geq y - 1, \theta, d) = \omega^y(\theta, d).$$

Therefore,

$$\begin{aligned} \omega^1 &= P(Y \geq 1 | Y \geq 0) = P(\text{Any SAE} | \text{Total}) \\ \omega^2 &= P(Y \geq 2 | Y \geq 1) = P(\text{Grade 3-5} | \text{Any SAE}) \\ \omega^3 &= P(Y \geq 3 | Y \geq 2) = P(\text{Grade 5} | \text{Grade 3-5}) \end{aligned}$$

Then, for all $y \geq 1$, the joint density of the ordinal variable Y is given by,

$$\pi^y(\theta, d) = (1 - \omega^{y+1}(\theta, d)) \prod_{m=0}^{y} \omega^m(\theta, d).$$

Thus, we can define the cdf of a ordinal variable with values $0, 1, 2, 3$ is

$$\begin{aligned} F(y|\theta, d) &= 0 \quad if \quad y \leq 0 \\ &= 1 - \pi^1(\theta, d) - \pi^2(\theta, d) - \pi^3(\theta, d) \quad \text{if} \quad y \geq 0 \quad and \quad < 1 \\ &= 1 - \pi^1(\theta, d) - \pi^2(\theta, d) \quad \text{if} \quad y \geq 1 \quad and \quad < 2 \\ &= 1 - \pi^2(\theta, d) \quad \text{if} \quad y \geq 2 and \quad < 2 \\ &= 1 \quad \text{if} \quad y \geq 3. \end{aligned}$$

Next, we can construct the likelihood function on the basis of the multinomial distribution with the probabilities $\pi^y(\theta, \quad d)$. Suppose that, at the current stage of the trial, among n_{ij} subjects treated at the combined dose level (A_i, B_j), x_{ij}^y subjects have reported $Y = y$ level toxicity. The likelihood is then given by

$$l(\alpha, \beta, \gamma|data) \propto \prod_y \binom{n_{ij}}{x_{ij}^y} \pi^y(\theta, d)^{x_{ij}^y}.$$

Assuming prior distributions of the model parameters are independent, The joint posterior distribution is given by $\pi(\alpha, \beta, \gamma|data) \propto l(\alpha, \beta, \gamma|data)\pi(\alpha)\pi(\beta)\pi(\gamma)$. From the joint posterior distribution we can obtain the full conditional distributions of the model parameters. After the outcomes of each cohort of subjects have been observed, one can use the Gibbs sampling algorithm to sample from the posterior distributions of the unknown parameters. Thus, the posterior estimates can be easily obtained, which could lead to the next stage of the trial design. Here we note that defining each outcome as an ordinal variable having three or more levels and recording two outcome variables provides a much more informative patient outcome than a binary variable. A simulation study is presented later in 3.3.2 that utilizes the ordinal toxicity framework.

3.3 Simulation

3.3.1 Binary Toxicity

We investigated the operating characteristics of our two-dimensional Bayesian copula dose finding method through simulation studies under different toxicity scenarios. For example, let us assume we have drug A with dose levels $(0.01, 0.03, 0.1, 0.3, 1)$ and drug B with dose levels $(0.01, 0.03, 0.1, 0.3, 1)$. Prior to the dose combination study, individual studies with drug A and drug B were performed. We use the individual toxicity probabilities from these studies in our joint toxicity copula model. We only use the dose levels that are less than or equal to the MTD that was selected from the individual trials on A and B.

Subjects are treated in cohorts, for example, a cohort size of 3. When a certain dose combination cohort reaches or exceeds the targeting toxicity limit, subjects are treated with a lower dose level. As phase I trial is frequently used to identify toxic dose levels of the drugs used, a conservative dose escalation approach is often preferred. To avoid overdosing, which is a concern in any clinical trial, often dose escalation is restricted to a one- step procedure. This can be done by restricting admissible dose combinations to the neighboring dose levels. A non diagonal escalation strategy escalates the dose level of drug A while keeping drug B dose fixed. A diagonal dose escalation increases both drug dose levels at the same time. A diagonal dose escalation approach could help one to obtain the MTD much faster than the non-diagonal approach, though the risk of overdosing is higher. However, in the copula combination method both the individual drug profiles are well known. This knowledge lowers the risk of overdosing even if a diagonal escalation is used. In this paper, we restrict dose escalation or de-escalation to one dose level at a time, while also allowing a move along the diagonal direction. So, at a cohort with dose level $(0.1, 0.1)$ we have three possibilities for dose escalation, namely $(0.1, 0.3), (0.3, 0.1)$ and $(0.3, 0.3)$.

For the combination trial with drug A and B, the dose finding algorithm works as follows.

1. At first, subjects are treated at a pre specified starting dose combination (eg. the lowest).

2. If the target toxicity limit (TTL) is not reached, at the current dose combination the dose is escalated to the neighboring dose combination which has the probability of toxicity closest to TTL. For example, if subjects are treated at $(0.1, 0.1)$ dose level, possible dose combinations for next stage would include $(0.1, 0.3), (0.3, 0.1)$ and $(0.3, 0.3)$. Next we compare the posterior mean probability of these three dose levels and choose the one closest to TTL.

3. At the current dose combination, if the TTL is exceeded, dose level is de-escalated to the neighboring dose combination which has the probability of toxicity lower than and closest to TTL.

4. We pre specify a maximum sample size. Once the sample size has been reached, the dose combination that has the probability of toxicity that is closest to the target toxicity limit is selected as the MTD combination. In our simulation, we define a specific targeting toxicity limit of 33%.

	Drug A				
Drug B	0.01	0.03	0.1	0.3	1
0.01	0.030	0.085	0.137	0.213	0.268
0.03	0.049	0.104	0.154	0.229	0.283
0.1	0.089	0.141	0.189	0.261	0.313
0.3	0.167	0.215	0.259	0.324	0.372
1	0.239	0.283	0.323	0.383	0.426

Table 3.1: True toxicity probabilities used for the simulation. Toxicity probabilities are in agreement with prior.

	Drug A				
Drug B	0.01	0.03	0.1	0.3	1
0.01	0.123	0.240	0.317	0.410	0.468
0.03	0.157	0.274	0.349	0.439	0.495
0.1	0.212	0.325	0.397	0.482	0.535
0.3	0.303	0.408	0.474	0.551	0.598
1	0.378	0.475	0.535	0.605	0.648

Table 3.2: True toxicity probabilities used for the simulation. Toxicity probabilities are toxic.

Table 1 above shows the true toxicity probabilities used for the simulation. We use gamma priors with mean 1 for all three parameters in the model. When using toxicity probabilities that are in agreement with the prior, we have 3 MTD dose combinations (doses within 2.5% of TTL 33%), $(0.1, 1), (0.3, 0.3), (1, 0.1)$. During the simulation, 87% of the trials selected one of the three true MTD's. For the second scenario true MTD's were $(0.1, 0.01), (0.1, 0.03), (0.03, 0.1)$ and the percentage

of MTD recommendations that coincides with the true MTD's were 62%. Similarly the six parameter logistic model managed to identify 57% and 84% MTD's intriduced. The simulation were done based on 1000 trials.

3.3.2 Ordinal Toxicity

Next we perform a simulation study that utilizes an ordering of the toxicity variable. Depending on a study one can categorise the toxicity variable in multiple categories according to severity of issues occuring. For example,

$$\begin{aligned} Y = 0 \quad & \text{No SAE} \\ = 1 \quad & \text{Grade1} - 2 \\ = 2 \quad & \text{Grade3} - 4 \\ = 3 \quad & \text{Grade5} \end{aligned}$$

Here each category $Y = y$ is composed of a discrete set of preferred terms, including those of greatest clinical relevance. Ideally, an ordinal toxicity variable with multiple categories will be more informative than a binary toxicity variable, which should improve the predicitive ability of a dose toxicity model. Aside from the overall prediction and dose selection as it was done in 3.3.1, we are also interested in finding the conditional toxicity prediction. At a given dose combination level, by using an ordinal toxicity variable we can predict higher or worse levels of toxicity. For example, we will gather all available information about Grade 2 toxicity incidents and we will use that to predict any Grdae 3 or 4 toxicity incidents. For example, let us assume we have drug A with dose levels $(0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$ and drug B with dose levels $(0.1, 0.2, 0.3, 0.4, 0.5)$. Prior to the dose combination study, individual studies with drug A and drug B were performed. We use the individual toxicity probabilities as predicted using the Emax model from these studies in our joint toxicity copula model.

We only use the dose levels that are less than or equal to the MTD that was selected from the individual trials on A and B.

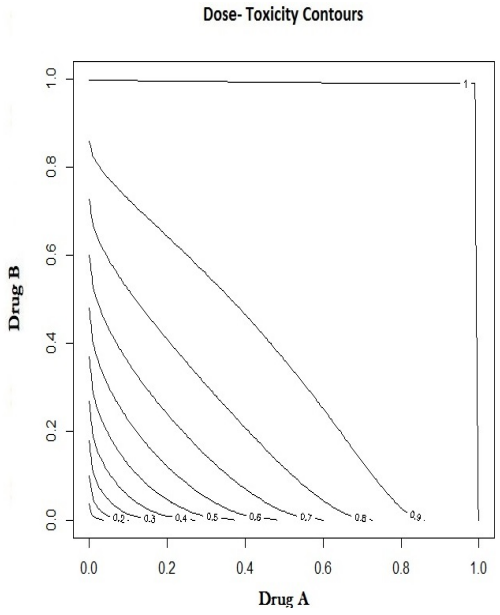

Figure 3.3: Estimated toxicity contour for combination of drug A and B

In our combination model subjects are treated in cohorts. Similar to the binary toxicity scenario, we restrict dose escalation or de-escalation to one dose level at a time, while also allowing a move along the diagonal direction. When a certain dose combination cohort reaches or exceeds the targeting toxicity limit, subjects are treated at a lower dose level. The target toxicity limit is fixed at 33%. The toxicity indicator is determined based on Grade 3 or higher toxicity events. Using this setup, we have the following 9 dose combination cohorts in our study, $(0.1, 0.1), (0.2, 0.2), (0.2, 0.3), (0.3, 0.3), (0.3, 0.4), (0.3, 0.5), (0.4, 0.5), (0.4, 0.6), (0.5, 0.6)$. The toxicty contour in 3.3 was obtained using the parameter estimates after all 9 cohorts.

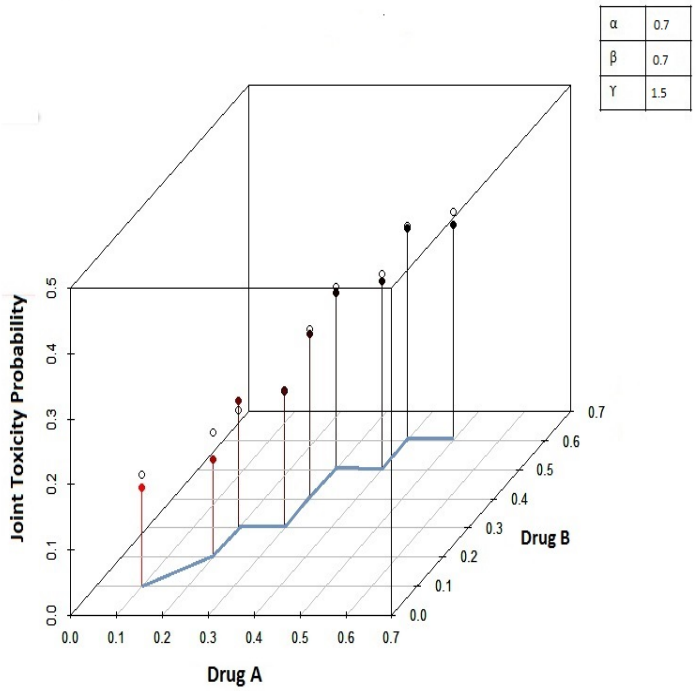

Figure 3.4: Observed and estimated ordinal toxicity probability for drug A and B

Next, we focus on the probability of observing a toxicity of Grade 5 given Grade 3 or 4 has been observed. The following graph details for each cohort the predicted and observed percentage of Grade 5 toxicity events. The predicted percentage was calculated at each dose combination level, using the hierarchical model and based on the information about Grade 3-4 and Grade 2 toxicity information available at that point in the study. From the graph above we can see that the proposed model works really well and the predicted percentage of Grade 5 incidents are becoming more and more accurate as the number of cohorts increse. The average cohort size for this study was between $21 - 24$ for all cohorts.

3.4 Case Study

The drug combination copula model was used to analyze the data obtained from clinical trials. A combination trial with two drugs, referred as drug A and B from now on was used, along with the individual trial information for drug A and B. Drug A had 5 dose levels $(0.01, 0.03, 0.1, 0.3, 1)$ and drug B had 3 dose levels $(0.03, 0.3, 1)$ in the individual study.

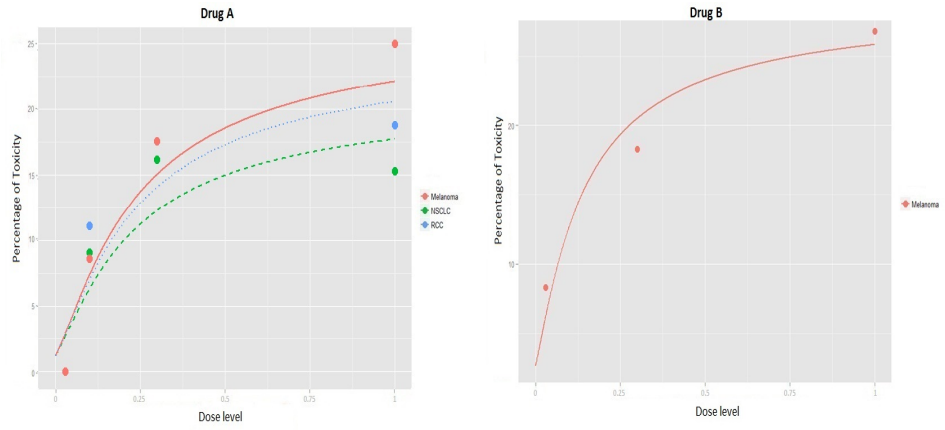

Figure 3.5: Observed and estimated individual toxicity probability for drug A and B

In the individual trial, drug A was used on multiple type of cancer subjects. These data was pooled together and analyzed using an Emax model for dose response. We obtained the marginal toxicity probabilities using this model. The graph above shows the dose toxicity relationship for drug A and B. For drug A we had 3 type of cancer subjects, and we used different Emax parameter for all 3 types, the remaining parameters were the same across all types. This allowed us to pool the data and combine the information together as mentioned in section 2.2. If sufficient information is available for each type one can always fit different dose response model to different type of cancer as well. For drug B only one type of cancer data was available and an emax model was used to obtain the dose toxicity relationship. A dose combination study with drug A and drug B was performed with a dose escalation scheme

similar to as mentioned in Section 3. The dose combination study had six cohorts $(0.03, 0.3), (0.1, 0.3), (0.3, 0.3), (0.3, 0.1), (0.1, 0.01), (0.3, 0.01)$.

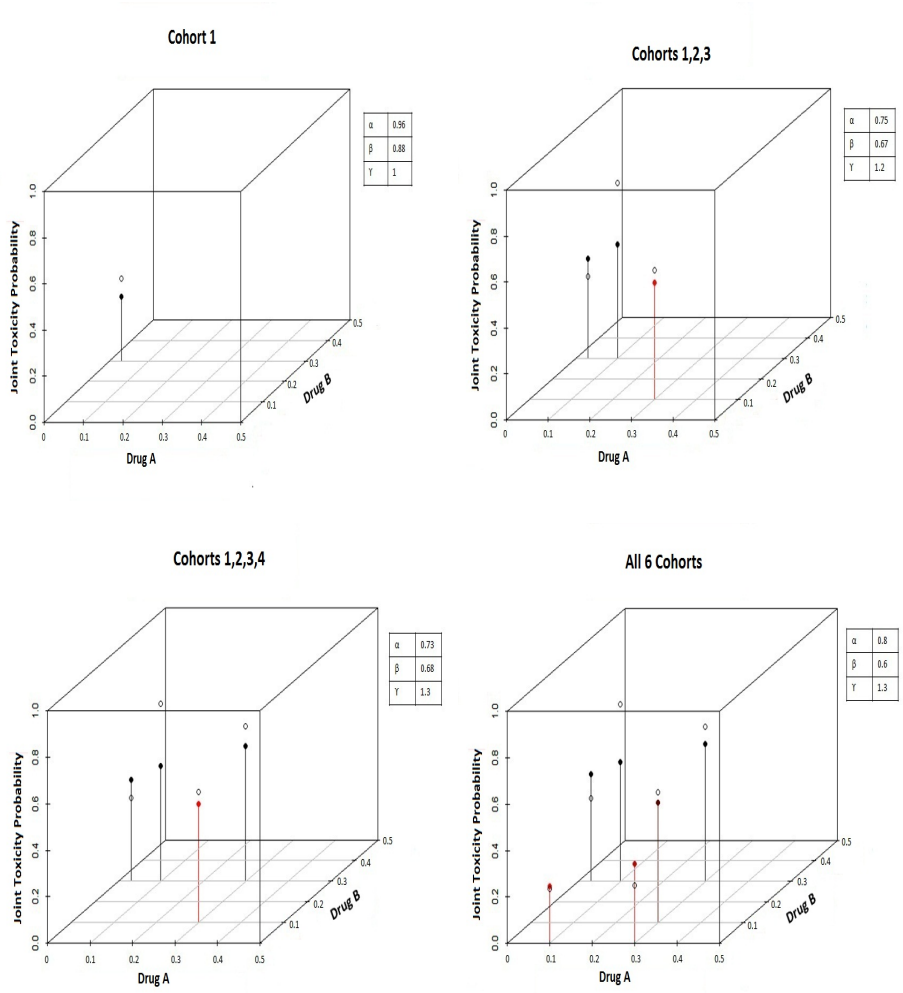

Figure 3.6: Observed and estimated joint toxicity probability drug A and B combination study

In figure 3.6 details the estimated and toxicity of the combination study are given. The snapshots of the model updated after the inclusion of every new cohort are also shown. Respective parameter estimates are given as well.

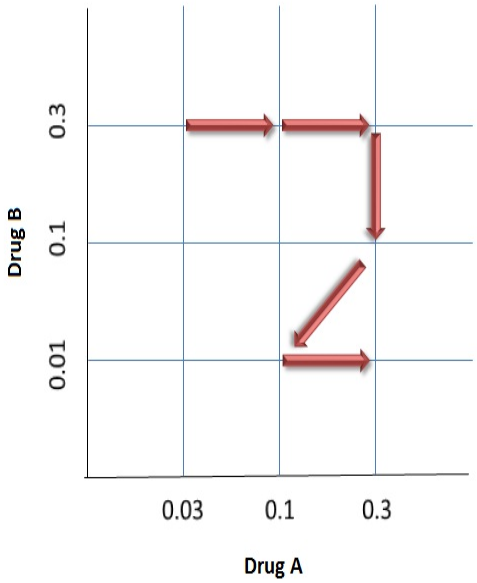

Figure 3.7: Dose Escalation in drug A and B combination study

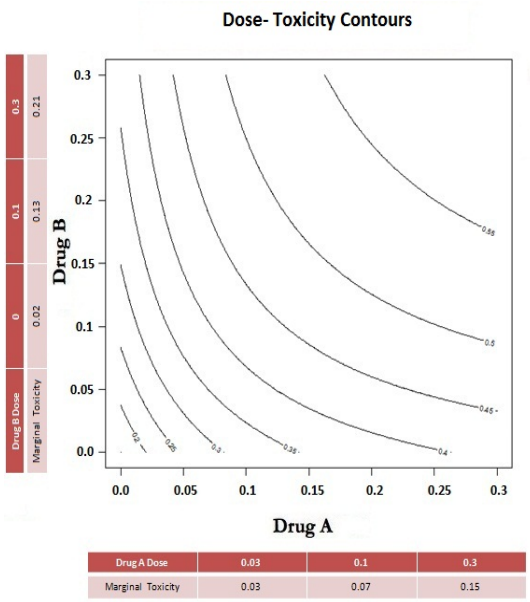

Figure 3.8: Estimated joint toxicity contour in drug A and B combination study

Next we study the contours estimated from the combination model. These contours can be extremely useful for future dose selection. In figure 3.8, different toxicity contours are shown. These contour lines, spread across different dose levels gives us

information about estimated toxicity. For example, the dose level $(0.1, 0.1)$ would have a toxicity between 40% to 45%. Also, if our target toxicity limit is 50% , then following the 0.5 contour we can identify which dose levels would likely result in lower toxicity. This knowledge could be really helpful for selecting the next cohort dose levels. Figure 3.8 also shows the dose escalation steps during the combination study.

The copula model retains the advantage of CRM, where the power parameter associated with the marginal probabilities can be updated during the analysis. This feature of the copula model allows us to update the toxicity profile of the individual drugs. In the Figure below we compare the marginal toxicity profile of drug A obtained from the individual study with the the marginal toxicity profile of drug A obtained from the combination study. We observe that, when used together with drug B, drug A exhibits an increased toxicity level. Also, as expected, as drug B dose level increases the marginal toxicity of drug A also increases.

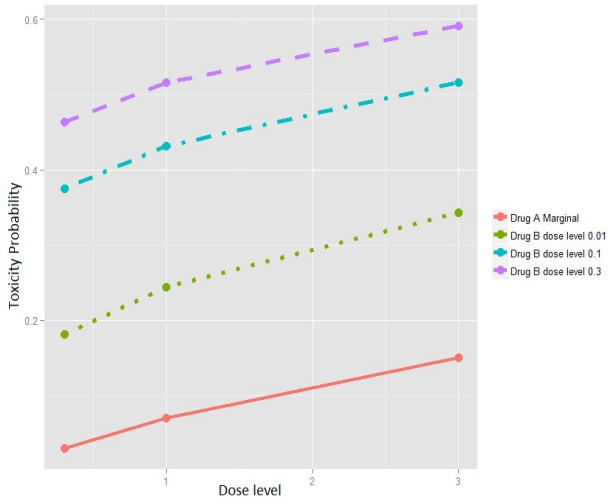

Figure 3.9: Drug A toxicity probability keeping drug B fixed

3.5 Discussion

There has been many methods that are developed for single-agent dose finding trials. Given the enormous advances in medicine and large numbers of new drugs to be tested, interest in finding combinations of drugs for patient treatment has grown. The goal of combination therapy is to achieve better patient response, particularly for cancer patients who are refractory to conventional therapies. In oncology, for example, combining agents can induce a synergistic treatment effect, allowing the clinician to target tumour cells with differing drug susceptibilities, and to achieve a higher intensity of dose with non-overlapping toxicities. Our research is motivated by many recent and more emerging dose finding drug combination clinical trials at the M. D. Anderson Cancer Center. One example is to find the MTD combination of a small molecule receptor (orally administered with four dose levels) and a mammalian target of rapamycin inhibitor (an intravenous drug with four dose levels) resulting in 16 combinations. The combined drugs are expected to induce a synergistic treatment effect by targeting different pathways. The trend of drug combination trials poses a great challenge to finding theMTDcombination with two or more drugs, particularly with small sample sizes in phase I studies. In a single-agent trial, we typically assume a monotonically increasing order of toxicity with respect to the dose. For any given dose, there are at most two adjacent doses and the order of toxicity is known. In contrast, for a two-drug combination dose space, there are up to eight adjacent doses, including diagonal and off-diagonal doses, as shown in Fig. 1. More importantly, complex drug–drug interactive effects often lead to unknown patterns of toxicity. Thus, the monotonic order of toxicity with respect to the dose level is lost, and it becomes unclear which dose combination should be assigned under a decision of dose escalation or de-escalation. Moreover, when two or more drugs are combined, the dimension of the dose space expands in a multiplicative fashion. This rapid increase in the dose dimension naturally requires a larger sample size, which can easily double

or triple that of a single-agent trial.

In this chapter, we discussed a phase I clinical trial adaptive design for combination therapy. We use a Bayesian hierarchical copula to model the joint toxicity probability of multiple drugs. This dose finding clinical trial is designed by borrowing strength from all the available data, individual and combination. The three parameter copula model links the joint toxicity probability of the drug combination to the individual drug toxicity probabilities. It captures different natures of interactive drug - drug effect in the combined dose toxicity surface. This model also allows us to incorporate the pre specified probabilities of toxicity of the dose combinations on the basis of the data obtained prior to the trial. This makes the estimation of the combination dose toxicity contour more efficient. We use a dose escalation strategy that allows both diagonal and non- diagonal increase in dose levels. Using this escalation strategy we estimate the MTD and obtain dose toxicity contours for dose combination. A simulation study demonstrates the performance of the dose escalation scheme which is further validated in the case study.

In this development the joint toxicity model was built based on a toxicity indicator. Depending on the purpose and the requirement of the study any binary toxicity variable can be used. Dose limiting toxicity is a widely used toxicity indicator in clinical trials. Similarly any adverse events (drug related or overall) can also be used to model the combined toxicity. Often adverse events are classified into multiple categories depending on the severity of their nature. Thus it is possible to use an ordinal toxicity variable to obtain a better idea about the dose toxicity surface. Grade 2, 3-4, 5 adverse events could be used for such a purpose. The method of combining such ordinal outcomes have been discussed in Section 2.4. Further simulation and case studies will be performed in future to assess the performance of this method. While the discussion in this paper is focused on dose-toxicity relationship, one can use similar methodologies to study efficacy data. Based on both dose-toxicity and

dose-efficacy contours, safe and effective dose combination levels could be reported for further analysis in Phase III trials.

Chapter 4

Conclusion

A model averaging estimator incorporates model uncertainty into the analysis by combining a set of competing candidate models rather than choosing just one. It also provides an insurance against selecting a poor model thus improving the risk in estimation. In this dissertation, we propose a more general framework where the choice of true model is not fixed. The truth can be any one or a mixture of the candidate models. Models that have large biases are not excluded from our analysis. We also study the behavior of frequentist model average estimator with an optimal weighting scheme to combine all the individual candidate models. As an illustration, we derive the model average estimator in the linear and logistic regression framework. The asymptotic distribution for model average estimator is also given. A linear regression model setup is used to simulate different scenarios to compare the performance of the proposed model average estimator with existing methods. Mean square error of the estimator is used for the purpose of comparison. We also implement the weighting scheme proposed by Liang et al. [2011] and compare their performance to the proposed and AIC based weights. The simulation results indicate that the proposed estimator works better than existing model averaging estimators.

In model averaging, if the true model is not included in the set of candidate models, we end up using an estimate that is biased. If all the models are misspecified, the weights derived by AIC or by using a consistent or unbiased estimator of mean square error are not optimal and should be used after careful consideration. When the true model is not included in the analysis thus all the candidate models are wrong, there

have been developments in model selection that takes care of the bias resulting from selection.A penalized version of AIC and BIC have been derived that performs better than other selection criteria. One can follow a similar path and derive the model averaging weights based on a sightly modified criteria.

Next, we discussed a phase I clinical trial adaptive design for combination therapy. We used a Bayesian hierarchical copula to model the joint toxicity probability of multiple drugs. This dose finding clinical trial is designed by borrowing strength from all the available data, individual and combination. The three parameter copula model links the joint toxicity probability of the drug combination to the individual drug toxicity probabilities. It captures different natures of interactive drug - drug effect in the combined dose toxicity surface. This model also allows us to incorporate the pre specified probabilities of toxicity of the dose combinations on the basis of the data obtained prior to the trial. This makes the estimation of the combination dose toxicity contour more efficient. We use a dose escalation strategy that allows both diagonal and non- diagonal increase in dose levels. Using this escalation strategy we estimate the MTD and obtain dose toxicity contours for dose combination. A simulation study demonstrates the performance of the dose escalation scheme which is further validated in the case study.

In this development the joint toxicity model was built based on a toxicity indicator. Depending on the purpose and the requirement of the study any binary toxicity variable can be used. Dose limiting toxicity is a widely used toxicity indicator in clinical trials. Similarly any adverse events (drug related or overall) can also be used to model the combined toxicity. Often adverse events are classified into multiple categories depending on the severity of their nature. Thus it is possible to use an ordinal toxicity variable to obtain a better idea about the dose toxicity surface. Grade 2, 3-4, 5 adverse events could be used for such a purpose. The method of combining such ordinal outcomes have been discussed as well. Further simulation was performed

to assess the performance of this method. While the discussion here is focused on dose-toxicity relationship, one can use similar methodologies to study efficacy data. Based on both dose-toxicity and dose-efficacy contours, safe and effective dose combination levels could be reported for further analysis in Phase III trials.

Bibliography

J. M. Bates and C. W. J. Granger. The Combination of Forecasts. *OR*, 20:451–468, 1969. ISSN 01605682. doi: 10.1057/jors.1969.103. 37

B. R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. VALID POST-SELECTION INFERENCE By Richard Berk, Lawrence Brown , Andreas Buja , Kai Zhang and Linda Zhao . 2000. 28

P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008. 41

S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model Selection: An Integral Part of Inference. *Biometrics*, 53:603–618, 1997. ISSN 0006341X. doi: 10.2307/2533961. 4

G. Claeskens and N. L. Hjort. *Model Selection and Model Averaging*, volume 44. 2008. ISBN 9780521852258. doi: 10.1006/jmps.1999.1278. 4, 36

D. Danilov and J. R. Magnus. On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122:27–46, 2004a. ISSN 03044076. doi: 10.1016/j.jeconom.2003.10.018. 37

D. Danilov and J. R. Magnus. Forecast Accuracy After Pretesting with an Application to the Stock Market. *Journal of Forecasting*, 23:251–274, 2004b. ISSN 02776693. doi: 10.1002/for.916. 37

D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:45–97, 1995. ISSN 00359246. 4

D. E. A. Giles, O. Lieberman, and J. A. Giles. The Optimal Size of a Preliminary Test of Linear Restrictions in a Misspecified Regression Model. *Journal of the American Statistical Association*, 87:1153–1157, 1992. 38

B. E. Hansen. Least Squares Model Averaging. *Econometrica*, 75:1175–1189, 2007. ISSN 00129682. doi: 10.1111/j.1468-0262.2007.00785.x. 8, 9, 36

T. Hastie and R. Tibshirani. The elements of statistical learning: data mining, inference and prediction. 173:693–694, 2005. doi: 10.1111/j.1467-985X.2010.00646_6.x. 4

N. L. Hjort and G. Claeskens. Frequentist Model Average Estimators. *Journal of the American Statistical Association*, 98(464):879–899, Dec. 2003. ISSN 0162-1459. doi: 10.1198/016214503000000828. URL http://www.tandfonline.com/doi/abs/10.1198/016214503000000828. 2, 5, 7, 8, 9, 13, 14, 29, 31, 36, 37, 42, 43

N. L. Hjort and G. Claeskens. Focussed information criteria and model averaging for Cox's hazard regression model. *Journal of American Statistical Association*, 101: 1449–1464, 2006. ISSN 01621459. doi: 10.1198/016214506000000069. 36

B. Hoadley. Asymptotic Properties of Maximum Likelihood Estimators for the Independent Not Identically Distributed Case, 1971. ISSN 0003-4851. 11

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian Model Averaging. *Statistical Science*, 14:121–149, 1999. ISSN 08834237. doi: 10.2307/2676803. 7

P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6:813–827, 2007. ISSN 0361-0926. doi: 10.1080/03610927708827533. 24

D. W. Hosmer and S. Lemeshow. *Applied logistic regression*. 2000. ISBN 0471356328. doi: 10.1198/tech.2002.s650. 22

N. Houede, P. F. Thall, H. Nguyen, X. Paoletti, and A. Kramar. Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase I/II trials. *Biometrics*, 66:532–540, 2010. ISSN 0006341X. doi: 10.1111/j.1541-0420.2009.01302.x. 53

L. Huo, Y. Yuan, and G. Yin. Bayesian dose finding for combined drugs with discrete and continuous doses. *Bayesian Analysis*, 7:1035–1052, 2012. ISSN 19360975. doi: 10.1214/12-BA735. 45

C. M. Hurvich and C. L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989. 38

C. M. Hurvich and C.-L. Tsai. Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika*, 78:499–509, 1991. ISSN 00063444. doi: 10.1093/biomet/78.3.499. 38

A. Karagrigoriou, S. Lee, and K. Mattheou. A model selection criterion based on the BHHJ measure of divergence, 2009. ISSN 03783758. 37

E. L. Lehmann. *Elements of large-sample theory*. 1999. ISBN 0387985956. doi: 10.2307/1271493. 11

E. L. Lehmann and G. Casella. *Theory of Point Estimation*, volume 41. 1998. ISBN 0387985026. doi: 10.2307/1270597. 11

H. Liang, G. Zou, A. T. K. Wan, and X. Zhang. Optimal Weight Choice for Frequentist Model Average Estimators. *Journal of the American Statistical Association*, 106(495):1053–1066, Sept. 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.tm09478.

URL http://www.tandfonline.com/doi/abs/10.1198/jasa.2011.tm09478. 6, 8, 15, 17, 29, 31, 32, 36, 37, 69

D. Lien and K. Shrestha. Estimating the optimal hedge ratio with focus information criterion. *Journal of Futures Markets*, 25:1011–1024, 2005. ISSN 02707314. doi: 10.1002/fut.20166. 37

D. Madigan, A. E. Raftery, J. C. York, J. M. Bradshaw, and R. G. Almond. Strategies for Graphical Model Selection. In *Selecting Models from Data*, volume 98195, pages 91–100. 1994. doi: 10.1007/978-1-4612-2660-4_10. 4

J. O'Quigley, M. Pepe, and L. Fisher. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, 46:33–48, 1990. ISSN 0006-341X. doi: 10.2307/2531628. 45

M. H. Pesaran, C. Schleicher, and P. Zaffaroni. Model averaging in risk management with an application to futures markets. *Journal of Empirical Finance*, 16:280–305, 2009. ISSN 09275398. doi: 10.1016/j.jempfin.2008.08.001. 37

A. E. Raftery and J. A. Hoeting. Bayesian Model Averaging for Linear Regression. 1998. 7

T. A. Stamey, J. N. Kabalin, M. Ferrari, and N. Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. IV. Anti-androgen treated patients. *The Journal of urology*, 141:1088–1090, 1989. 35

B. M. Staw and J. Ross. Behavior in Escalation Situations: Antecedents, Prototypes, and Solutions. *Research in Organizational Behavior*, 9:39, 1987. ISSN 01913085. 45

M. J. Sweeting and A. P. Mander. Escalation strategies for combination therapy

Phase i trials. *Pharmaceutical Statistics*, 11:258–266, 2012. ISSN 15391604. doi: 10.1002/pst.1497. 45

P. F. Thall and J. D. Cook. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics*, 60:684–693, 2004. ISSN 0006341X. doi: 10.1111/j.0006-341X.2004.00218.x. 45

P. F. Thall, E. H. Estey, and H. G. Sung. A new statistical method for dose-finding based on efficacy and toxicity in early phase clinical trials. *Invest New Drugs*, 17: 155–167, 1999. 45

J. G. Thursby and P. Schmidt. Some properties of tests for specification error in a linear regression model. *Journal of the American Statistical Association*, 72(359):pp. 635–641, 1977. ISSN 01621459. URL http://www.jstor.org/stable/2286231. 38

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. 11, 40

Y. Wei and P. D. McNicholas. Mixture Model Averaging for Clustering and Classification. *arXiv preprint arXiv12125760*, 2012. 37

G. Yin and Y. Yuan. Bayesian dose finding in oncology for drug combinations by copula regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(2):211–224, 2009. ISSN 1467-9876. doi: 10.1111/j.1467-9876.2009.00649.x. 45, 46, 48

Y. Zhou. Adaptive designs for phase i dose-finding studies, 2010. ISSN 07673981. 45