Use Of Vapnik-Chervonenkis Dimension in Model Selection

by

Merlin T. Mpoudeu

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Statistics

Under the Supervision of Professor Bertrand Clarke

Lincoln, Nebraska

August, 2017

Use Of Vapnik-Chervonenkis Dimension in Model Selection

Merlin T. Mpoudeu, Ph.D.

University of Nebraska, 2017

Advisor: Professor Bertrand Clarke

In this dissertation, I derive a new method to estimate the Vapnik-Chervonenkis Dimension (VCD) for the class of linear functions. This method is inspired by the technique developed by Vapnik et al. Vapnik et al. (1994). My contribution rests on the approximation of the expected maximum difference between two empirical Losses (EMDBTEL). In fact, I use a cross-validated form of the error to compute the EMDBTEL, and I make the bound on the EMDBTEL tighter by minimizing a constant in of its right upper bound. I also derive two bounds for the true unknown risk using the additive $(\widehat{ERM}_1)$ and the multiplicative $(\widehat{ERM}_2)$ Chernoff bounds. These bounds depend on the estimated VCD and the empirical risk. These bounds can be used to perform model selection and to declare with high probability, the chosen model will perform better without making strong assumptions about the data generating process (DG).

I measure the accuracy of my technique on simulated datasets and also on three real datasets. The model selection provided by VCD was always as good as if not better than the other methods under reasonable conditions. To understand the behaviour of my method, I introduced the concept of 'consistency at the true model'. I compare my technique to established model selection techniques such as two forms of empirical risks minimization (ERM's), the Bayes Information Criterion (BIC), Smoothly Clipped Absolute Deviation (SCAD, see Fan and Li (2001)), and Adaptive LASSO (ALASSO, see Zou (2006a)). On simulations, we conclude that when design points are well chosen and the sample size is sufficient, our method performs as well as BIC and better than ERM's. On real datasets with dependence structure, our method performs better than BIC. As a generality, sparsity methods such as SCAD and ALASSO, gave models that were so oversimplified as to be invalid.

Thus under reasonable conditions on design points and sample size, our method never gave models that were obviously deficient compared to those of other methods.

Acknowledgements

I would especially like to thank my advisor Bertrand Clarke. Without his insight, encouragement, probing questions, excitement, and suggestions, this thesis would have ended up much differently. I am very proud of this work, and I owe him a tremendous amount for helping it to become something valuable. I would also like to thank Professor Kent M. Eskridge for the help that he provided through his Friday consulting group.

I would like to thank Dr. Jennifer Clarke, Dr. Yumou Qiu and Dr Stephen Scott for being my graduate committee members and reviewing my dissertation. Their help and suggestions have been invaluable on both academic and personal levels.

I am grateful to my friends (Cyrile Nzouda, Salfo Bikienda, and Kismiantini); you have been with me since the beginning, and your support through the last four years has been invaluable.

I would like to express my gratitude to my parents, Mbialeu Charles and Siewe Philomene, and my brother and sisters for their encouragement throughout this graduate school process. I would also like to thank two and most import people in my life: My fiancee Anne Mbenya and my 20 month old son Charles Mpoudeu for their support and patience.

Table of Contents

List of Figures

List of Tables

Nomenclature

ALASSO	Adaptive Least Angle Shrinkage Selector Operator
BIC	Bayesian Information Criterion
DG	Data Generator
EMDBTEL	Expected Maximum Difference Between Two Empirical Losses
ERM	Empirical Risk Minimization
LHS	Left-Hand Side
RHS	Right-Hand Side
SCAD	Smoothly Clipped Absolute Deviation
VCD	Vapnik-Chervonenkis Dimension

INTRODUCTION

Given a dataset of size n, there are 2^n different dichotomies of the observations. If we have a collection of functions $\mathcal{F}$ so that given any dichotomy of the dataset, there is a function $f \in \mathcal{F}$ that is consistent with the dichotomy then we say that $\mathcal{F}$ generates the 2^n datasets. The Vapnik-Chervonenkis Dimension (VCD) is the maximum value of data points all whose dichotomies can be generated by $\mathcal{F}$. This measures the richness of $\mathcal{F}$.

The earliest usage of what we now call the VCD seems to be in Vapnik and Chervonenkis (1968). A translation into English was published as Vapnik and Chervonenkis (1971). The VCD was initially called an index of a collection of sets with respect to a sample and was developed to provide sufficient conditions for the uniform convergence of empirical distributions. It was extended to provide a sense of dimension for function spaces, particularly for functions that represented classifiers. After two decades of development this formed the foundation for the field of Statistical Learning Theory, summarized in Vapnik (1999) and extensively treated in Vapnik (1998). An intuitive treatment of VCD from the classification perspective can be found in Moore (2001). Aside from the development in statistical learning, VCD plays an important role in various real analysis settings. For instance, VCD defines VC classes of functions that have relatively small covering numbers, see Pollard (1984), Devroye (1996), Van der Waart (1998), and van de Geer (2000) amongst others. At the risk of oversimplification, the VCD of a space of functions is currently seen mostly as a parameter characterizing bounds of cumulative risks in term of the size of the function class.

One of our goals is to estimate the true value h of the VCD (noted h_T) of a data generator (DG) by, say, $\hat{h}$ and use $\hat{h}$ as the estimated complexity of the model class; e.g, linear models. Given $\hat{h}$, the next step is to perform model selection. In fact, we use the notion of 'consistency at the true model' to do so. We define 'consistency at the true model' to be

1. consistency at the true model in the usual sense of the term, basically we expect that $\hat{h} \to h_T$ in probability as the sample size increases, coupled with

2. some kind of identifiable bad behaviour away from the true model, getting worse as the wrong model moves further from the true model.

If we have a list of non-nested models with nested VCD's denoted by h, and we would like to perform

model selection, we can estimate h by setting

$$\hat{h} = \arg\min_k \left| VCD\left(P_k(\cdot \mid \beta)\right) - \hat{h}_k \right| \tag{1}$$

where $\{P_k\left(\cdot \mid \beta\right) | k = 1, 2, \cdots, K\}$ is some set of models and $\hat{h}_k$ is calculated using model k. Writing $VCD\left(P_k(\cdot \mid \beta)\right) = h_k$, we see that $\hat{h}_k$ is a function of the VCD of the conjectured model and $\arg\min_k$ in (1) removes this dependence. In the special case that $P_k(\cdot \mid \beta)$'s are linear i.e. $P_k(\cdot \mid \beta) = \sum_{j=0}^k \beta_j x_j$, we will see that $h_T = k + 1$ (here, $x_0 \equiv 1$). In practice we an approximate form of (1) by thresholding the difference within the $\arg\min$; see (3.6) in Subsec. 3.4.2.

We can also order the inclusion of the covariates using a shrinkage method such as SCAD (see Fan and Li (2001)) or the correlation (see Fan and Lv (2008)) and the use $\hat{h}$ to pick a model. We also upper bounded the true but unknown risk by an expression that depends on $\hat{h}$. We used that expression for model selection. In fact, for each model on the list, we estimate a VCD for it and use the estimate in an upper bound of the unknown risk. Then we pick the model with the smallest upper bound on its risk. Asymptotically, the chosen model should perform well with higher probability. This technique is called Empirical Risk Minimization and is suitable for non-nested model lists.

While VCD is not in general equivalent to the real dimension, in the special case of linear models VCD and the real dimension are the same. Thus, at a minimum, with linear models, we can cut down the plausible models to those that have at least $\hat{h}$ parameters – and hopefully not too many more.

The estimation of h_T is done in three steps: first we find the Expected Maximum Difference Between Two Empirical Losses (EMDBTEL). This value depends on the model class (the class of linear models for this case), the design points, the number of bootstrap samples we draw from our data, and the number of intervals used to discretrize the loss function. In fact, if the number of the design points is 10, we will have 10 different values of the EMDBTEL. Since, the bound on the EMDBTEL is not tight, the next step is to find the best factor on the RHS of (2.21) i.e. the value of c that will minimize the upper bound. The final step is the estimation of h_T. This is done by minimizing the squared distance between the EMDBTEL (the LHS of (2.21)) and the upper bound (the RHS of (2.21)). The intuition is that this minimum corresponds to the decision problem, or more precisely the model class, that has the fastest convergence to zero of its expected supremal difference of cumulative empirical losses and hence would be the 'right' problem for us to solve if only we knew it. Therefore, even though the model class is only implicitly defined, we can take its h_T or empirically, $\hat{h}$, as a lower bound on the VCD of model classes. in the examples presented here our objective function has a unique minimum that can be taken as an estimate of $\hat{h}$. In practice, for finite sample sizes this is best regarded as a lower bound on h_T. Indeed, in practice, we can estimate several values of $\hat{h}$, take their average and standard deviation as a way to get what is effectively a bootstrap confidence interval for h_T.

Our contribution rests on two pioneering earlier contributions. First using zero-one loss functions, Vapnik and collaborators Vapnik et al. (1994) developed the asymptotic expected supremal difference of the cumulative empirical losses in classification problems. Thus, it focused more on probabilities than expectations. The asymptotic expression they obtained had three terms combined in an ad hoc way. We did not use the asymptotic form that they obtained to estimate $\hat{h}$. Nevertheless, our derivation of an asymptotic upper bound for the expected supremal difference of empirical cumulative losses is based on their work. However, our derivation is more complex e.g., to allow for continuous loss functions, and we use the cross-validation form of the error. McDonald McDonald (2012) Chaps. 2 and 3 provides a lucid discussion on this. Secondly, McDonald et al. McDonald et al. (2011) examine the technique from Vapnik et al.Vapnik et al. (1994) and tried to establish consistency of $\hat{h}$ for h_T, again in the classification context. The key result in their proof rested on a result from van de Geer (2000). However, this result does not correctly fill the gap in their proof. Our simulations with the objective function in Vapnik et al. (1994) leads us to believe that consistency will be very hard to prove – if in fact it is true.

Our main contributions in this dissertation are:

- We re-derive the Vapnik et al. (1994) bounds for the regression case (not classification).

- We correct errors in the earlier derivation so we obtain a single objective function to estimate h.

- We improve the objective function by using a cross validation form of the errors.

- We improve the objective function by optimizing over a constant that appears in it.

- We correct the original Vapnik et al. (1994) algorithm for our case.

- We verify that our $\hat{h}$ works well in simulation and for three datasets.

The motivation of this dissertation comes from the fact that in statistics, there are many different model selection techniques that have been developed. These include best subset selection, forward and backward model selection, Bayesian Information Criterion (BIC), Smoothly Clipped Absolute Deviation, Adaptive Least Angle Shrinkage Selector Operator, etc. However, none of these techniques use the complexity of the data generator (DG).

The remainder of this dissertation is structured as follows. In chapter 1, we discuss the notion of complexity in the VCD sense in general. We give two definitions of the notion of VCD. We also prove that for the case of linear regression, VCD is just the number of parameters in the saturated model. We also present summaries of Vapnik et al. (1994) and McDonald et al. (2011). In Chapter 2 we present the main theoretical work we have accomplished to date. It rests on discretizing bounded loss functions so that upper bounds for the distinct regions of the expected supremal difference of empirical losses can be derived in Sec. 2.1. In Chapter 3 we implement the theory developed in Chapter 2 on synthetic datasets.

In Chapters 4 and 5 we implement the theory developed in Chapter 2 on three real datasets namely the Tour De France, the Abalone and the Wheat data. In Chapter 6, we conclude our work, and we present ideas for future work, in particular what we can do to improve our estimator for h_T.

Chapter 1

VAPNIK-CHERVONENKIS DIMENSION, COVERING NUMBER AND COM-PLEXITY

There are a variety of intuitive concepts in Statistics that have been formalized in ways that make good sense to those who are familiar with them but are often impenetrable to those who are not. As example, we have information. A series of incisive articles including Soofi (1994), Ebrahimi et al. (1999) and Ebrahimi et al. (2010) develop this concept admitting its 'intangibility'. Another example is model uncertainty. This topic has been addressed in many settings but remains relatively hard to quantify see Draper (1995) (and discussion) for one of the earliest contributions, Clyde and George (2004) for a more recent contribution from the Bayes perspective, and Berk et al. (2013) for a treatment of the effect of model uncertainty on inference. A third example, and the subject of this review, is complexity.

The aim of this chapter is to give some background on the notion of complexity. In fact, there are many ways to measure the complexity or the richness of a class of functions. In Statistical Learning Theory, the indexes to measure the richness of a collection of function are VCD, ND Covering Number, in Information Theory the notion of code length is used and this is an extension of Shannon Entropy and mutual information. The remainder of this chapter will be developed as follows: In Sec. 1.1, we will give two definitions of the notion of VCD. In Sec. 1.2, we will show that VCD for the class of linear models is just the number of parameters. In Sec. 1.3, we will talk about Covering Number, Entropy, Growth Function and how they are related.

1.1 Definition of VCD

The aim of the VCD is to measure the richness or capacity, the complexity of a collection of functions. There are at least three distinct ways to approach the definition of VCD: geometric, combinatorial, and via covering numbers. We will begin with the geometric definition and then proceed to the combinatorial definition. For the sake of simplicity, we only present here the first two definitions of VCD.

1.1.1 Geometric definition of VC dimension

Let $\mathcal{F}$ be a collection of classifiers on some domain $\mathcal{X}$. The domain of $\mathcal{X}$ is usually a finite d dimensional space. Let $k \geq 1$ be an integer, and let $\mathcal{X}_k = \{x_1, \cdots, x_k\} \subseteq \mathcal{X}$. A dichotomy of $\mathcal{X}_k$ is a partition of $\mathcal{X}_k$ into two disjoint sets. We say that $\mathcal{X}_k$ is identified by $\mathcal{F}$ if and only if for any dichotomy of $\mathcal{X}_k$, there exists a function $f \in \mathcal{F}$ consistent with the dichotomy. That is let U be any dichotomy of $\mathcal{X}_k$, there exists $f \in \mathcal{F}$ such that $\forall u \in U\ f(u) = 1$ and $\forall u \in U^c \mathcal{X}_k,\ f(u) = 0$. Thus when we have a class of subsets of $\mathcal{X}$, we will say that the class of subset is generated by $\mathcal{F}$ if each element of the collection can be identified by $\mathcal{F}$. Succinctly, the VCD of $\mathcal{F}$ is the largest k for which every single element of the power set of $\mathcal{X}_k$ $(P(\mathcal{X}_k))$ can be identified by $\mathcal{F}$ in other words, $P(\mathcal{X}_k)$ is generated by $\mathcal{F}$. An infinite VC dimension means that $\mathcal{F}$ maintains full richness for all sample sizes.

Before giving some examples on how to calculate the VCD using the geometric definition, we will first emphasize some aspects of VCD. Note that, in the definition of VCD we did not say that for all sets of size k, their power set should be generated by $\mathcal{F}$. It's enough to show that there exist a set of size k such that its power set can be generated by $\mathcal{F}$, then we say that the VCD is at least k. So a simple way to prove that the VCD of a collection of function is k can be done in two steps. First prove that the collection of functions can generate any set of size less or equal to k. Second prove that for any set of size $k + 1$, we cannot identify a dichotomy for that set. The following examples are toy examples that will serve for illustrative purposes.

Let $\mathcal{X} = \mathcal{R}$ and let $\mathcal{F}$ be a collection of indicator functions of the intervals in the real line of the form $[a, \infty)$ for some $a \in \mathcal{R}$. That's

$$f(x) = \chi_{[a,\infty)}(x) = \begin{cases} 1, & \text{if } x \in [a, \infty); \\ 0, & \text{otherwise.} \end{cases}$$

The VCD is at least 1. Let's select a sample of one point $(k = 1)$ and position it on the real line. We want to show that there are a's which generate $\{\{\emptyset\}, \{x\}\}$. By taking $a = x - 1$, $f(x) = \chi_{[x-1,\infty)}(x)$ identifies $\{x\}$ and $f(x) = \chi_{[x+1,\infty)}(x)$ identifies $\emptyset$. Now the question becomes: is there any set of size 2 that can be generated by $\mathcal{F}$? Choose two points $\mathcal{X}_2 = \{x_1, x_2\}$, we have $P(\mathcal{X}_2) = \{\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_1, x_2\}\}$. Let's assume without loss of generality that $x_1 < x_2$. In fact, we cannot find any a that identifies x_1 alone thus; the VCD of this collection of functions is 1.

Let $\mathcal{X} = \mathcal{R}$ and let $\mathcal{F}$ be a collection of indicator functions of the intervals in the real line of the form $(a, b]$ for some $a, b \in \mathcal{R}^2$. That's

$$f(x) = \chi_{[a,b)}(x) = \begin{cases} 1, & \text{if } x \in (a, b]; \\ 0, & \text{otherwise.} \end{cases}$$

From the previous case, we can say that the VCD is at least 2. Let's choose two points $\mathcal{X}_2 = \{x_1, x_2\}$, we have $P(\mathcal{X}_2) = \{\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_1, x_2\}\}$. Let's assume without loss of generality that $x_1 < x_2$. We need to show that there are values of a, b which identify each element of $P(\mathcal{X}_2)$. The following functions $f(x) = \chi_{(x_2, x_2+1]}(x)$, $f(x) = \chi_{(x_1-1, \frac{x_1+x_2}{2}]}(x)$, $f(x) = \chi_{(\frac{x_1+x_2}{2}, x_2+1]}(x)$, $f(x) = \chi_{(x_1-1, x_2+1]}(x)$ respectively identify $\{\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_1, x_2\}\}$. Now to show that the VCD is at most two, we need to show that any set of three points cannot be generated by $\mathcal{F}$. It is sufficient to show that one of the dichotomies is not identifiable. Let $\mathcal{X}_3 = \{x_1, x_2, x_3\}$ and assume without loss of generality that $x_1 < x_2 < x_3$. There is no $\chi_{(a,b]}$ that identifies $\{x_1, x_3\}$. Since the choice of points was arbitrary, we say that the VCD is 2.

Another interesting example that is more related to Statistics is the case of half space in the plane. Let $\mathcal{X} = \mathcal{R}^2$ and $\mathcal{F}$ is the collection of functions (half plane) of the form $f(x) = ax + b$ where $a, b \in \mathcal{R}^2$. For this case, the VCD is at least 3 because any 3 non co-linear points in $\mathcal{R}^2$ can be generated by $\mathcal{F}$. In fact, the power set will have 8 different sets (dichotomies) where four of them are the mirror of the other four. By placing one point on one side of the line (half plane) and the other two on the other side of the line we can identify 3 dichotomies and by placing all three points on one side of the plane we identify the fourth dichotomy. The other four are just a mirror image of what we just did. So we can identify them just by symmetric. Now we just have to prove that there is not any set of four points that can be generated by our collection of functions. Here, there are two cases to be considered. Firstly, let's assume without loss of generality that that the four points lie in the convex hull generated by them. In this case, the labelling which assigns a positive sign to the pair on one diagonal and a negative sign on the other diagonal cannot be identified by a straight line. Secondly, let's assume that three of the points form a convex hull and the fourth point is an interior point. The dichotomy where the points on the convex hull are all positives and the interior point is negative cannot be identified by a half space. So the VCD is 3.

VCD can be extended to $\mathcal{F}$'s containing functions that assume values in $\mathbb{R}$ by looking at level sets. In fact, we will define level sets as the collection of points in the domain of f for which $f(x) \geq \beta$ where $\beta \in \mathcal{R}$. Let

$$\mathcal{F}^* = \{\chi_{\{z | f(z) \geq \beta\}}(\cdot) \mid f \in \mathcal{F}, \beta \in \mathcal{R}\} \tag{1.1}$$

be the collection of indicator functions generated by $\mathcal{F}$. Now, one can set $\text{VCdim}(\mathcal{F}) = \text{VCdim}(\mathcal{F}^*)$.

As an example of this consider the densities $\mathcal{F} = \{\phi_\sigma(\cdot) \mid \sigma \in \mathbb{R}\}$ where $\phi_\sigma(\cdot)$ is the density of a $N(0, \sigma^2)$ distribution. The collection of points for which $\phi_\sigma(x) \geq \beta$ i.e., the level set is symmetric intervals around zero, i.e., $\mathcal{F}^* = \{\chi_{[-a,a]} \mid a \in \mathbb{R}^+\}$. By reasoning similar to that above, $\mathcal{F}^*$ has VC dimension one. (There is no set of two distinct points on the real line that has a $\mathcal{P}(\mathcal{X}_2)$ that can be generated by the sets in $\mathcal{F}^*$ but any $\mathcal{P}(\mathcal{X}_1)$ can be generated.) Extending this, let $\mathcal{F} = \{\phi_{\mu,\sigma}(\cdot) \mid \mu \in \mathbb{R}, \sigma \in \mathbb{R}\}$

where $\phi_{\mu,\sigma}(\cdot)$ is the density of a $N(\mu,\sigma^2)$ distribution. Now, $\mathcal{F}^*$ consists of all closed intervals of the form $[a,b]$ in $\mathbb{R}$. In this case, $\text{VCD}(\mathcal{F}) = 2$.

Now consider $\mathcal{F} = \{\phi_{\mu,\Sigma}(\cdot) \mid \mu \in \mathbb{R}^2, \Sigma = \sigma^2 I_2, \sigma \in \mathbb{R}^+\}$, densities of two independent normal with different means. Then $\mathcal{F}^*$ consists of all circles in $\mathbb{R}^2$. In this case, the circle will classify everything inside the circle as positive and everything outside aas negative. It is easy to see that any set of three points in the plane that forms a non-degenerate triangle gives a $\mathcal{P}(\mathcal{X}_3)$ that can be generated by $\mathcal{F}^*$. However, no set of four points gives a $\mathcal{P}(\mathcal{X}_4)$ that can be generated by $\mathcal{F}^*$. Hence, $\text{VCD}(\mathcal{F}) = 3$.

Analogously, we can see that if the variances are not equal, i.e., $\mathcal{F} = \{\phi_{\mu,\Sigma}(\cdot) \mid \mu \in \mathbb{R}^2, \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2), \sigma_1, \sigma_2 \in \mathbb{R}^2\}$ consists of all ellipses, then it is possible to produce five distinct points whose $\mathcal{P}(\mathcal{X}_5)$ can be generated by $\mathcal{F}$. However, there is no set of six points in the plane giving a $\mathcal{P}(\mathcal{X}_6)$ that can be generated by $\mathcal{F}$. So, $\text{VCdim}(\mathcal{F}) = 5$.

It is left to the reader to work out the VCD of the full set of normals in the plane, i.e., $\mathcal{F} = \{\phi_{\mu,\Sigma}(\cdot) \mid$ where $\mu \in \mathbb{R}^2, \text{diag}(\Sigma) = (\sigma_1^2, \sigma_2^2), \sigma_1, \sigma_2 \in \mathbb{R}^2$ and $\sigma_{12} = \sigma_{21} \in \mathbb{R}\}$. (It is not hard to produce six points in the plane whose $\mathcal{P}(\mathcal{X}_6)$ can be generated from $\mathcal{F}$. Arguing that no set of seven points leads to a $\mathcal{P}(\mathcal{X}_7)$ that can be generated by $\mathcal{F}$ is harder.)

Looking back to the examples studied so far, one might think that there is a one-to-one relationship between VCD of a collection of functions and the number of parameters in the model require to represent the class of functions. In fact, this observation is not always true. A one dimensional family can have infinite VCD. Here is an example.

Let $\mathcal{X} = \mathcal{N}$, $\mathcal{F} = \{f_\alpha(x) : \quad$ where $\alpha \in \mathbb{R}\}$

$$f_\alpha(x) = \begin{cases} 1, & \text{if the } x^{th} \text{ bit in the binary representation of } \alpha \text{ is 1;} \\ 0, & \text{if the } x^{th} \text{ bit in the binary representation of } \alpha \text{ is 0.} \end{cases}$$

For any number x that you give, I will always find a number (α) such that the xth bit in the binary representation is 1, for instance $x = 5$, in this case, $\alpha = 125$ and the binary representation of $\alpha = 125 = 1111101$. Thus every $\mathcal{S} \subseteq \mathcal{X}$ can be generated by $\mathcal{F}$, therefore $\text{VCdim}(\text{F}) = \infty$

1.1.2 Combinatorial definition of VCD

In Subsec. 1.1.1 the VCD was associated with a collection of functions $\mathcal{F}$. In this section, we will look at the VCD of a collection of measurable sets.

Let $\mathcal{C}$ be a collection of measurable subsets of a space and let $\mathcal{X} = \{x_1, ..., x_n\}$ be a set of n points in the same space. Consider the collection of sets $C \cap \{x_1, ..., x_n\} \ \forall C \in \mathcal{C}$. As before, we say a subset A of $\mathcal{X}$ can be identified by $\mathcal{C}$ if there exists $C \in \mathcal{C}$ such that $A = C \cap \{x_1, ..., x_n\}$, and $\mathcal{P}(\mathcal{X})$ is generated by $\mathcal{C}$ if and only if $\forall A \in \mathcal{P}(\mathcal{X})$, there exists $C \in \mathcal{C}$ such that $A = C \cap \{x_1, \ldots, x_n\}$. In statistical learning

jargon, they use the terminology 'picked out' for identified and shattered for generated. As before, the VC dimension k of the class $\mathcal{C}$ of sets is the smallest n for which no set of size $n+1$ can be generated by $\mathcal{C}$.

Formally, Van Der Vaart and Wellner (1996) defines

$$\Delta(\mathcal{C}, x_1, ..., x_n) = \#\left\{ \{x_1, ..., x_n\} \cap C : C \in \mathcal{C} \right\}, \tag{1.2}$$

so that the VCD is

$$k = \inf\left\{ n : \max_{\{x_1, ..., x_n\}} \Delta(\mathcal{C}, x_1, ..., x_n) < 2^n \right\}. \tag{1.3}$$

The infimum over an empty set can be defined to be zero, so the VCD of a set is only infinity when $\mathcal{C}$ can generate $\mathcal{P}(\mathcal{X}_k)$ for arbitrarily large k. Devroye et al. Devroye et al. (2013) Chap. 12.4 uses a slight generalization of this by defining

$$s(\mathcal{C}, n) = \max_{x_1, ..., x_n} \Delta(\mathcal{C}, x_1, ..., x_n)$$

so that $s(\mathcal{C}, n) \leq 2^n$ noting that if there is an n for which $s(\mathcal{C}, n) < 2^n$ then for all $k \geq n$, $s(\mathcal{C}, k) < 2^k$. The VC dimension k is then the last time equality holds, that is, the largest k for which $s(\mathcal{C}, k) = 2^k$.

Checking whether a subset of $\{x_1, ..., x_n\}$ can be identified (or 'picked out') by some $C \in \mathcal{C}$ is much like checking whether the indicator function for a set gives the right pattern of zeros and ones – ones for the x_i's in the set identified and zero on the other x_i's. Consider the equivalence relation on sets defined by saying that two sets in a class of sets are equivalent if and only if they identify the same subsets of $\{x_1, ..., x_n\}$. Then, the number of equivalence classes in $\mathcal{C}$ depends on $\{x_1, ..., x_n\}$. For fixed n, the equivalence classes heuristically are like a data driven sub-σ-field from the σ field generated by $\mathcal{C}$. Thus the combinatorial interpretation of VC dimension is an effort to formalize the idea that finitely many data points only permits finitely many events to be distinguished.

Let $\mathcal{F}$ be a class of measurable functions on a sample space and let $f \in \mathcal{F}$. The subgraph of $f : \mathcal{X} \to \mathbb{R}$ is

$$SG_f = \{(x, t) : t < f(x)\}. \tag{1.4}$$

Now, $\mathcal{F}$ is called a VC class if and only if the collection of sets $\{SG_f \mid f \in \mathcal{F}\}$ has finite VC dimension as defined by 1.3. When $\mathcal{F}$ is a collection of indicator functions, it is straightforward to see that $\mathcal{F}$ is a VC class with $\text{VCdim}(\mathcal{F}) = k$ in the sense of 1.3 if and only if $\mathcal{F}$ has VC dimension k in the geometric sense in Subsec. 1.1.1. Simply note that the subgraphs from a collection of indicator functions is a class of sets

that can generate a particular $\mathcal{P}(\mathcal{X}_k)$ in the geometric sense if and only if it satisfies 1.3, i.e., generates all the sets in the class $\mathcal{P}(\mathcal{X}_k)$ by using Δ to pick them out. More generally, results in Van Der Vaart and Wellner (1996) Sec. 2.6.5 suggest that in general the definition of a VC class of functions (in terms of 1.3) is equivalent to the geometric definition by using the indicator functions in 1.1.

Part of the reason 1.3 is regarded as combinatorial is from the following intuition and result. By definition, a VC class of functions $\mathcal{F}$ with $\mathsf{VCdim}(\mathcal{F}) = k$ identifies fewer than 2^h subsets from any $\mathcal{P}(\mathcal{X}_h)$ when $h > k$. It is intuitive that as h increases the smaller the fraction of the possible sets that $\mathcal{F}$ can identify. That is, if $h = k + \ell$ then even though $\mathcal{F}$ could identify $2^{h-\ell}$ sets the actual number is polynomial as h increases not exponential with a smaller exponent, i.e., the dropoff in how many sets can be identified as h increases is fast. This surprising fact is called Sauer's Lemma, see Sauer (1972). Let $\mathcal{C}$ be a class of functions and define the 'shatter' function

$$s(n) = \max_{\{x_1, ..., x_n\}} \Delta_n(\mathcal{C}, x_1, ..., x_n).$$

The shatter function is the largest number of subsets that $\mathcal{C}$ can identify from an n-tuple of points. It is naturally compared with 2^n, the cardinality of the power set from n points. If a set of size n is generated by $\mathcal{C}$ then $\mathsf{VCdim}(\mathcal{C}) \geq n$ and $s(n) = 2^n$. If there is no such set then $\mathsf{VCdim}(\mathcal{C}) < n$ and $s(n) < 2^n$. Sauer's Lemma gives a polynomial bound on $s(n)$. In particular, Sauer's Lemma gives that if a collection of sets has VC dimension bounded by k then any set of n elements can only be split n^k ways. The proof requires recognizing $s(n)$ can be represented as a sum of binomial coefficients to which an inductive argument can be applied

Theorem 1.1.1. *If* $\mathsf{VCdim}(\mathcal{C}) = k$, *then* $s(n) \leq \left(\frac{en}{k}\right)^k$.

Proof. See Van Der Vaart and Wellner (1996). $\qquad\square$

This sort of sudden qualitative change in s will be seen in other quantities such as the growth function, see Sec. 1.3.

1.2   VC dimension in statistics

In section 1.1 we gave two definitions of VC dimension and we went through some examples of VCD where we derive the VCD of various sets of functions. In this section, we will formally prove the VCD of some collection of functions. In fact, we will derive the VC dimension of the class of linear functions.

1.2.1 VCD for Regression

The VCD is a quantity that is defined for a collection of functions. It is the largest number of points the collection of functions can separate in all possible ways or equivalently the largest n for which the growth function is 2^n. Note that we really mean the largest number of points the collection of functions can separate and all possible ways; this does not imply that every set of n points can be separate in possible ways by our collection of functions. Our goal here is to prove the following theorem:

Theorem 1.2.1. *Let $\mathcal{R}^d = \mathcal{X}$, $\mathcal{S} = \{x_1, x_2, \cdots, x_n\} \subseteq \mathcal{R}^d$ and let*

$$\mathcal{F} = \{(w, \beta) : f(x) = sign(wx - \beta)\}$$

where $w \in \mathcal{R}^d$ and $\beta \in \mathcal{R}$. Then a set $\mathcal{S} = \{x_1, x_2, \cdots, x_n\} \subseteq \mathcal{R}^d$ is generated by $\mathcal{F}$ if and only if the set $\left\{(x_1' - 1), (x_2' - 1), \cdots, (x_n' - 1)\right\}$ is linearly independent in $\mathcal{R}^{d+1}$. It follows that VCdim($\mathcal{F}$)$=d+1$

Proof. The proof of this Theorem will follow the same stream of reasoning as the proof of Theorem : 3.4 in Anthony and Bartlett (2009). We will prove this theorem in two steps. Let's assume that if $\mathcal{S}$ is generated by $\mathcal{F}$, then $\left\{(x_1' - 1), (x_2' - 1), \cdots, (x_n' - 1)\right\}$ is linearly independent in $\mathcal{R}^{d+1}$. We will prove this by contradiction, that's we assume that $\mathcal{S}$ is generated by $\mathcal{F}$ but $\left\{(x_1' - 1), (x_2' - 1), \cdots, (x_n' - 1)\right\}$ is linearly dependent in $\mathcal{R}^{d+1}$. Let $V = (v_1, v_2, \cdots, v_{2^n}) = \begin{pmatrix} w_1 & w_2 & \cdots & w_{2^n} \\ \beta_1 & \beta_2 & \cdots & \beta_{2^n} \end{pmatrix}$ be the total number of partitions. $\mathcal{S}$ is generated by $\mathcal{F}$ means that $\forall\ v_i \in V, \exists\ (w_i, \beta_i) \in \mathcal{F}$ such that

$$\begin{pmatrix} x_1' & -1 \\ x_2' & -1 \\ \vdots & \vdots \\ x_n' & -1 \end{pmatrix} \begin{pmatrix} w_i \\ \beta_i \end{pmatrix} = v_i.$$

Since $\left\{(x_1' - 1), (x_2' - 1), \cdots, (x_n' - 1)\right\}$ are linearly dependent, one can expresses one element of this vector as the linear combination of the others, that is

$$\begin{pmatrix} x_n' & -1 \end{pmatrix} = \sum_{k=1}^{n-1} \alpha_i \begin{pmatrix} x_k' & -1 \end{pmatrix}, \quad \text{where at least one } \alpha_k \neq 0.$$

Using this, the nth coordinate of any dichotomy v_i can be expressed as follows

$$
\begin{aligned}
y_{ni} &= \begin{pmatrix} x_n^{'} & -1 \end{pmatrix} \begin{pmatrix} w \\ \beta \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n-1} \alpha_i x_i^{'} & -\sum_{i=1}^{n-1} \alpha_i \end{pmatrix} \begin{pmatrix} w \\ \beta \end{pmatrix} \\
&= w \sum_{i=1}^{n-1} \alpha_i x_i^{'} - \beta \sum_{i=1}^{n-1} \alpha_i = \sum_{i=1}^{n-1} \left(\alpha_i x_i^{'} w - \beta \alpha_i \right) \\
&= \sum_{i=1}^{n-1} \alpha_i \left(w x_i^{'} - \beta \right) = \sum_{j=1}^{n-1} \alpha_j v_{ji}, \quad \text{where } v_{ji} = w_i x_j^{'} - \beta_i
\end{aligned}
$$

Let's assume that x_j with $\alpha_j \neq 0$ and let $y_{ji} = sign(\alpha_j)$ for all $1 \leq j \leq n-1$. Let x_j and let $y_{ji} = -1$, we will prove that this dichotomy is not identifiable. $y_{ji} = sign(w_i x_j - \beta_i) = sign(\alpha_j) \Rightarrow \alpha_j (w_i x_j - \beta_i) \geq 0 \Rightarrow \sum_{j=1}^{n-1} \alpha_j (w_i x_j - \beta_i) \geq 0 \Rightarrow y_{ji} \geq 0$, which is a contradiction since $y_{ji} = -1$. Then it follows that $VCD(\mathcal{F}) \leq d+1$.

For the second part, we assume that $\left\{ (x_1^{'} - 1), (x_2^{'} - 1), \cdots, (x_n^{'} - 1) \right\}$ is linearly independent and prove that $\mathcal{S}$ can be generated by $\mathcal{F}$. Because of the linear independence assumption, the matrix

$$
\begin{pmatrix}
x_1^{'} & -1 \\
x_2^{'} & -1 \\
\vdots & \vdots \\
x_n^{'} & -1
\end{pmatrix}
$$

has row-rank n, therefore, for any $v \in V$, there is a unique solution (w, β) to the following equation

$$
\begin{pmatrix}
x_1^{'} & -1 \\
x_2^{'} & -1 \\
\vdots & \vdots \\
x_n^{'} & -1
\end{pmatrix}
\begin{pmatrix} w \\ \beta \end{pmatrix} = v.
$$

Thus it follows that $\mathcal{S}$ can be generated by $\mathcal{F}$, this implies that $\text{VCD}(\mathcal{F}) \geq d+1$. These two results together imply that $\text{VCD}(\mathcal{F}) = d+1$. $\qquad \square$

We comment that the VCD for the regression function in a recursive partitioning model, i.e., a tree, can be readily worked out as a function of the number of splits in the tree. If there are, say, K splits, it is easy to verify that the VCD is $k+1$. Analogous results for neural networks – even single hidden layer neural networks – are not so easy to derive. Nevertheless, in the 1990's there were many efforts, chiefly by computer scientists, to evaluate the VCD of classes of neural networks particularly in a classification context. While many of these results were helpful in general terms, it is not clear how they can be used in our present context of regression because often only bounds or asymptotic expressions for the VCD in

terms of the number of hidden nodes were obtained.

1.3 Covering Numbers, Entropy, Growth Function and VC-dimension

The goal is to assign a class of functions $\mathcal{F}$ a measure that generalizes the concept of real dimension. This will necessitate defining covering numbers, their finite sample analogs called random entropy numbers, and then generalizations of entropy numbers called the annealed entropy and the growth function. The VCD will emerge from a key shape property of the growth function. This section is largely a revamping of parts of Vapnik (Vapnik (1998), Chap.3).

To start, recall the simplest measure of the size of a set is its cardinality, but since $\mathcal{F}$ is typically uncountable that's no help. What is a help is the concept of a covering number since it is also a measure of size. Since $\mathcal{F}$ is a subset of a normed space with norm $\| \cdot \|$, given an ϵ one can measure $\mathcal{F}$'s size by the minimal number of balls of radius ϵ required to cover $\mathcal{F}$. This is called the covering number and is written $N(\epsilon, \mathcal{F}, \| \cdot \|)$. The ϵ-entropy of $\mathcal{F}$, or the Boltzmann entropy, is $\ln N(\epsilon, \mathcal{F}, \| \cdot \|)$, its natural log. The covering number of a class of indicator functions $\mathcal{F}$ is related to the number of subsets that can be picked out from $\{x_1, ..., x_n\}$ by the functions in $\mathcal{F}$. Indeed, write $\mathcal{F} = \{1_C : C \in \mathcal{C}\}$ so that $\mathcal{C}$ is the class of sets whose indicator functions are in $\mathcal{F}$. Now, a complicated argument based on Sauer's Lemma gives a bound on the covering number for $\mathcal{F}$. The result is that there is a universal constant K so that for $1 > \epsilon > 0$ and $r \geq 1$, if h is the (geometric) VCD of $\mathcal{C}$ then

$$N(\epsilon, \mathcal{F}, L_r(\mu)) = N(\epsilon, \mathcal{C}, L_r(\mu)) \leq \frac{Kh(4e)^h}{\epsilon^{r(h-1)}}$$

where L_r is the Lebesgue space with r-th power norm with respect to μ, see Van Der Vaart and Wellner (1996) (Chap. 2.6).

The more typical definition of the entropy, also called the Shannon entropy, is minus the expected log of the density; this is the expectation of a sort of random entropy. The log density plays a role similar to the Boltzmann entropy because both represent a codelength.

While the relationships among the ϵ-entropy of $\mathcal{F}$, the VCD of $\mathcal{F}$, and other quantities are deep, all of these are population quantities. So, given a finite sample, it is important to have an analog of covering number. This is provided by counting the number of sets a sample of size n can distinguish. Define two events as distinguishable using a sample $z_1, ..., z_n$, if and only if there is at least one z_i that belongs to one event and not the other. It is immediate that not all sets in a large class $\mathcal{C}$ are distinguishable given a sample; different samples can have different collections of distinguishable sets; and the number of distinguishable sets can depend on the sample as well.

The next task is to identify which sets it is important for a sample to be able to distinguish. For a set

of indicator functions $Q(z, \alpha)$ with support sets C_α and $\alpha \in \Lambda$, consider the vector of length n

$$q(\alpha) = (Q(z_1, \alpha), ..., Q(z_n, \alpha)).$$

For fixed α, $q(\alpha)$ is a vertex of the unit cube in $\mathbb{R}^n$. As α ranges over Λ, $q(\alpha)$ hops from vertex to vertex. Let $N(\Lambda, z_1, ..., z_n)$ be the number of vertices $q(\alpha)$ lands on. Then, $N(\Lambda, z_1, ..., z_n)$ is the number of sets the sample $z_1, ..., z_n$ can distinguish.

Let the Z_i be IID. Then, $N(\Lambda, Z_1, ..., Z_n)$ is a random variable bounded by 2^n for given n. The random entropy of the set $\mathcal{F}$ of indicator functions $Q(\cdot, \alpha)$ for $\alpha \in \Lambda$ is $\ln N(\Lambda, z_1, ..., z_n)$ and the entropy of $\mathcal{F}$ is

$$H(\Lambda, n) = \int \ln N(\Lambda, z_1, ..., z_n) dF(z_1, ..., z_n).$$

The entropy of real valued functions in general is similar to the indicator function case, but uses the concept of a covering number more delicately. First consider a set of bounded real valued functions $\mathcal{F}_C$ defined by

$$\forall \alpha \in \Lambda \quad |Q(z, \alpha)| \le C.$$

Again $q(\alpha, z_1, ..., z_n)$ can be used to generate a region $\mathcal{R}$ in the n dimensional cube of side length $2C$. The random covering number of $\mathcal{R}$ for a given ϵ is $N(\Lambda, \epsilon, z_1, ..., z_n)$ giving $\ln N(\Lambda, \epsilon, z_1, ..., z_n)$ as the random ϵ-entropy. (To define the covering number, the norm must be specified as well as the ϵ and region to be covered. In this case, the natural choice is the metric $\rho(q(\alpha), q(\alpha')) = \max_{i=1,...,n} |Q(z_i, \alpha) - Q(z_i, \alpha')|$.) The ϵ-entropy (not random) for $\mathcal{F}_C$ is

$$H(\Lambda, \epsilon, n) = \int H(\Lambda, \epsilon, z_1, ..., z_n) dF(z_1, ..., z_n).$$

Equipped with the entropy, it is easy to specify two closely related quantities that will arise in the bounds to be established that involve the VCD. First, the annealed entropy is

$$H_{ann}(\Lambda, n) = \ln E N(\Lambda, Z_1, ..., Z_n)$$

and, second, the growth function is

$$G(\Lambda, n) = \ln \sup_{z_1, ..., z_n} N(\Lambda, z_1, ..., z_n).$$

Clearly, $H(\Lambda, n) \le H_{ann}(\Lambda, n) \le G(\Lambda, n)$, the result of Jensen's inequality on the entropy $H(\Lambda, n)$ and taking the supremum of the integrand.

The growth function does not depend on the probability measure; it is purely data driven.

The key theorem linking VCD to a collection of indicator functions is the following. It provides an interpretation for h separate from its definition in terms of shattering. Although h arises in the argument by using the geometric definition of shattering, in principle, there is no reason not to use the result of this Theorem as a third definition for h. Let $C(n, i)$ be the number of combinations of i items from n items.

Theorem 1.3.1. *Vapnik (1998)The growth function $G(\Lambda, n)$ for a set of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$ is linear-logarithmic in the sense that*

$$
G(\Lambda, n) = \begin{cases} = n \ln 2 & \text{if } n \leq h; \\ \leq \ln(\sum_{i=0}^{h} C(n, i) \leq h \ln(en/h) = h(1 + \ln(n/h)) & \text{if } n > h. \end{cases}
$$

Comment: *If $h = \infty$ then the second case in the bound never applies so $G(\Lambda, n) = n \ln 2$. The essence of the result is that if a set of functions has finite VCD, then its growth function is initially linear, meaning that the amount of learning per datum accumulated is proportional to the sample size, up to the VCD, after which the learning per datum drops off to a log rate.*

1.4 Review of Vapnik et al. (1994) and McDonald et al. (2011)

In this section we will present the work of Vapnik et al Vapnik et al. (1994) that leads to the empirical estimation of VCD. We will also present the work of McDonald et al. (2011) who tried to argue that using the procedure of Vapnik et al. (1994), the estimated $\hat{h}$ will concentrate around the true value.

Let's first set the stage and notation that Vapnik used in his 1994 paper. $f(x, \alpha), \alpha \in \Lambda$ is the set of binary classifiers. The n sample pairs $Z = (x, y)$ where $X \in \mathcal{R}^p$, $Y \in \{0, 1\}$ is $(Z_1, Z_1, \cdots, Z_n)$. The probability of error between y and $f(x, \alpha)$ is

$$
p(\alpha) = E |Y - f(X, \alpha)|.
$$

The expectation is taken with respect to the joint distribution between x, and y $(f(x, y))$. Since $f(x, y)$ is unknown, we use the empirical error

$$
\nu(\alpha) = \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i, \alpha)|.
$$

In Vapnik Vapnik and Chervonenkis (1991), it was shown that a function that minimizes the empirical risk will be consistent if and only if the following one sided uniform convergence holds

$$
\lim_{n \to \infty} P \left(\sup_{\alpha \in \Lambda} |p(\alpha) - \nu(\alpha)| \geq \epsilon \right) = 0.
$$

In the 1930's Kolmogorov and Smirnov found the law of distribution of the maximum deviation between a distribution function and the empirical function for any random variable. This can be stated as follows: for the collection of functions $f(x, \alpha)$, where $\alpha \in \Lambda$ the equality

$$P\left(\sup_{\alpha \in \Lambda}(p(\alpha) - \nu(\alpha)) \geq \epsilon\right) = \exp\left(-2\epsilon^2 n - 2\sum_{j=2}^{\infty}(-1)^j \exp\left(-2\epsilon^2 j^2 n\right)\right)$$

holds for large n where $p(\alpha) = Ef(x, \alpha)$ and $\nu(\alpha) = \frac{1}{n}\sum_{i=1}^{n} f(x_i, \alpha)$.

In Vapnik Vapnik (1998), it was shown that

$$P\left(\sup_{\alpha \in \Lambda}(p(\alpha) - \nu(\alpha)) \geq \epsilon\right) \leq \min\left(1, \exp\left\{\left(c_1 \frac{\ln\left(\frac{2n}{h}\right) + 1}{\frac{n}{h}} - c_2 \epsilon^2\right) n\right\}\right)$$

As before, this inequality does not depend on the probability distribution. Using these results, Vapnik et al. in Vapnik et al. (1994) showed that the expected maximum difference between two empirical losses can be bounded by a function which depends only on the VCD of the classifier. For comparison purposes we present this argument next.

Let $Z^1 = Z_1, Z_2, \cdots, Z_n$ and $Z^2 = (Z_{n+1}, Z_{n+2}, \cdots, Z_{2n})$ be two independent identically distributed copies of Z. Let $\nu_1(Z^1, \alpha)$ and $\nu_2(Z^2, \alpha)$ be the empirical risk using Z^1 and Z^2. Vapnik et al Vapnik et al. (1994). claims the following bound on the expectation of the maximum difference between $\nu_1(z^1, \alpha)$ and $\nu_2(z^2, \alpha)$,

$$\xi(n) = E\left(\sup_{\alpha \in \Lambda}\left(\nu_1(z^1, \alpha) - \nu_2(z^2, \alpha)\right)\right) \leq \begin{cases} 1, & \text{if } \frac{n}{2} \leq 0.5; \\ c_1 \frac{\ln\left(\frac{2n}{h}\right) + 1}{\frac{n}{h}}, & \text{if } \frac{n}{h} \text{ is small;} \\ c_2 \sqrt{\frac{\ln\left(\frac{2n}{h}\right) + 1}{\frac{n}{h}}}, & \text{if } \frac{n}{h} \text{ is large.} \end{cases} \tag{1.5}$$

Equation (1.5) can be bounded by

$$\Phi_h(n) = \begin{cases} 1, & \text{if } \frac{n}{h} \leq 0.5; \\ a\frac{\ln\left(\frac{2n}{h}\right) + 1}{\frac{n}{h} - k}\left(\sqrt{1 + \frac{b\left(\frac{n}{h} - k\right)}{\ln\left(\frac{2n}{h}\right) + 1}} + 1\right), & \text{otherwise.} \end{cases} \tag{1.6}$$

The constants $a = 0.16$ and $b = 1.2$ are claimed to be universal and given in Vapnik et al. (1994). These constants can be used for estimation of VCD for other classes of functions; they represent the trade off error between small and large n/h. The constant $k = 0.14927$ was chosen so that $\Phi_h(0.5) = 1$, although it is unclear what this means. If the LHS of (1.5) were known, we would just solve for h and get an estimate $\hat{h}$. Because we do not have that information, we estimate $\hat{\xi}(n)$ numerically. The algorithm proposed for estimation h in Vapnik et al. (1994) is as follows.

Result: Obtain $\widehat{\xi}(n_j)$'s
Given:

- A collection of classifiers $\mathcal{F}$;

- A set of design points $N_L = \{n_1, n_2, \cdots, n_l\}$;

- An integer W for the number of bootstrap samples;

for $j = 1, 2, \cdots, l$

1. Generate a random sample of size $2n_j$ $Z^{2n} = (Z_1, Z_2, \cdots, Z_{2n_j})$;

2. Divide the random sample into two groups; Z^1 and Z_2;

3. Flip the class labels for the second set: Z^2;

4. Merge the two sets to train the binary classifier;

5. Separate the sets and flip the labels on the second set back again;

6. Calculate the training error of the estimated classifier $\widehat{f}$ respectively on Z^1 and Z^2 with the correct labels;

7. Calculate

$$\widehat{\xi}_i(n_j) = \left| \nu_1(Z^1, \alpha) - \nu_2(Z^2, \alpha) \right|$$

8. Repeat steps 1-7 W times

9. Calculate

$$\widehat{\xi}(n_j) = \frac{1}{W} \sum_{i=1}^{W} \widehat{\xi}_i(n_j)$$

End for
Algorithm 1: Generation of the $\widehat{\xi}(n_j)$'s using the Vapnik et al. Vapnik et al. (1994) algorithm.

Having enough values of $\widehat{\xi}(n_j)$ we can use nonlinear regression to estimate h via

$$\widehat{\xi}(n) = \Phi_h(n) + \epsilon(n),$$

where $\epsilon(n)$ has mean zero and an unknown distribution. Now we estimate $\hat{h}$ by

$$\hat{h} = \min_h \sum_{l=1}^{|N_L|} \left(\hat{\xi}(n_l) - \Phi_h(n) \right)^2.$$

One of the problems with this method is that the derivation of $\Phi_h(n)$ uses a conditioning argument in which the conditioning set has probability going to zero so the derivation of $\Phi_h(n)$ is in question. Moreover, in our work with algorithm 1 (not shown here) we have found it to be unstable.

McDonald et al. (2011) tried to prove the algorithm in Vapnik et al. (1994) for estimating h was consistent. Their main result was the following

Theorem 1.4.1. *Let $\delta \geq \frac{4}{\sqrt{2Wk}} \max(24c_1, 29)$ and suppose that $h \leq M$. Then $\exists c_1, c_2, c_3$ and W, k so*

that

$$P\left(\left|\hat{h} - h\right| \geq \delta\right) \leq 13 \exp\left(-\frac{Wkc_2\delta^2}{16c_3}\right).$$

This result as stated is not true although the parts of the proof that are correct are a tour-de-force. What makes the proof incomplete is a gap in a sequence of upper bounds. One of the bounds that look intuitively reasonable does not in fact follow from the covering number argument the authors advance. So we believe that a version of Theorem 1.4.1 is true but will required a more careful treatment. Appendix A provides some details.

Finally for this chapter we credit these results for motivating us and providing us with a framework for our contributions. In particular, we have borrowed literally from the technique of proof of these results. However, we also remind readers that our results are for regression not classification; we have change the mathematical form of the error and adding an extract optimization, amongst other improvements. In short, we think we have taken a technique whose rudiments had been identified, have refined them, have assembled and have shown how to use the overall result for analysis.

Chapter 2

BOUNDS ON EXPECTED SUPREMAL DIFFERENCE OF EMPIRICAL LOSSES

We are going to bound the Expected Maximum Difference Between Two Empirical Losses (EMDBTEL) for the case of regression. In fact, we convert the regression problem into m classification problems by discretizing the empirical loss. We will use this bound to derive an estimator of the VCD for the class of linear functions. In Sec: 2.1, we will present the extension of Vapnik bounds and in Sec: 2.2, we will present adjustment in the computation of the EMDBTEL.

2.1 Extension of Vapnik-Chervonenkis Bounds

Let $Z = (X, Y)$ be a pair of observations and write $Z^{2n} = (Z_1, \ldots, Z_{2n})$ be a vector of $2n$ independent and identically distributed (IID) copies of Z. Let

$$Q(z, \alpha) = L\left(y, f\left(x, \alpha\right)\right)$$

be a bounded real valued loss function, where $\alpha \in \Lambda$, an index set, and assume $\forall\, \alpha,\ 0 \leq Q(Z, \alpha) \leq B$ for some $B \in \mathbb{R}$. Consider the discretization of Q using m disjoint intervals (with union $[0, B)$) given by

$$Q_j^*\left(z, \alpha, m\right) = \begin{cases} \frac{(2j+1)B}{2m}, & \text{if } Q\left(z, \alpha\right) \in I_j = \left[\frac{jB}{m}, \frac{(j+1)B}{m}\right); \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

and j takes values from 0 to $m-1$. For $j \neq j'$, $support Q_j^*(Z, \alpha, m) \cap support Q_{j'}^*(z, \alpha, m) = \emptyset$. Now, consider indicator functions for Q being in an interval of the same form. That is, let

$$\chi_{I_j}(Q\left(z, \alpha\right)) = \begin{cases} 1, & \text{if } Q\left(z, \alpha\right) \in I_j = \left[\frac{jB}{m}, \frac{(j+1)B}{m}\right); \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

and write

$$n_{1j}^* = \sum_{i=1}^{n} \chi_{I_j}\left(Q(z_i^1, \alpha, m)\right) \quad \text{and} \quad n_{2j}^* = \sum_{i=n+1}^{2n} \chi_{I_j}\left(Q(z_i^2, \alpha, m)\right)$$

for the number of data points whose losses land inside the interval I_j in the first and second half of the sample of size n, respectively.

Proposition 2.1.1. *The sequence of measurable functions $\langle Q^*(\cdot, \alpha, m)\rangle\big|_{m \geq 1} = \sum_{j=0}^{m-1} Q_j^*(z, \alpha, m)$ converges to $Q(\cdot, \alpha) \in [0, B]$ a.e. in the underlying probability of the measure space.*

Proof. we want to prove that $\forall \epsilon \geq 0$, there exists $m_0 > 0$ such that whenever $m > m_0$, $|Q^*(z, \alpha, m) - Q(z, \alpha)| \leq \epsilon$. Let $\epsilon = \frac{B}{2m}$. $\forall z$ there exists j such that

$$
\begin{aligned}
\frac{jB}{m} &\leq Q(z, \alpha) \leq \frac{(j+1)B}{m} \\
\frac{jB}{m} - \frac{(2j+1)B}{2m} &\leq Q(z, \alpha) - Q^*(z, \alpha, m) \leq \frac{(j+1)B}{m} - \frac{(2j+1)B}{2m} \\
-\frac{B}{2m} &\leq Q^*(z, \alpha, m) - Q(z, \alpha) \leq \frac{B}{2m} \\
|Q^*(z, \alpha, m) - Q(z, \alpha)| &\leq \frac{B}{2m} = \epsilon
\end{aligned}
$$

$\square$

Let

$$
Z^1 = (Z_1, \ldots, Z_n) \quad \text{and} \quad Z^2 = (Z_{n+1}, \ldots, Z_{2n})
$$

be the first and second half of the sample of size $2n$, respectively, and write

$$
\nu_{1j}^* \left(z^1, \alpha, m\right) = \frac{n_{1j}^* Q_j^* \left(z^1, \alpha, m\right)}{n} \quad \text{and} \quad \nu_{2j}^* \left(z^2, \alpha, m\right) = \frac{n_{2j}^* Q_j^* \left(z^2, \alpha, m\right)}{n},
$$

for the empirical risk using $Q^* (z, \alpha, m)$ on the first and second half of the sample, respectively for the jth interval. Also, let

$$
\nu_1 \left(z^1, \alpha\right) = \frac{1}{n} \sum_{i=1}^{n} Q \left(z_i^1, \alpha\right) \quad \text{and} \quad \nu_2 \left(z^2, \alpha\right) = \frac{1}{n} \sum_{i=n+1}^{2n} Q \left(z_i^2, \alpha\right)
$$

be the empirical risk function using $Q (z, \alpha)$ on the first and second half of the sample, respectively.

To begin to control the expected supremal difference between bounded loss functions, let $\epsilon > 0$ and define the events

$$
A_{\epsilon,m} = \left\{ z^{2n} : \sup_{\alpha \in \Lambda} \left(\nu_1^* \left(z^1, \alpha\right) - \nu_2^* \left(z^2, \alpha\right) \right) \geq \epsilon \right\} \tag{2.3}
$$

$$
\text{where} \quad \nu_k^*(z^k, \alpha) = \sum_{j=0}^{m-1} \nu_{kj}^*(z^k, \alpha, m) \quad \text{and} \quad k = 1, 2.
$$

Since $A_{\epsilon,m}$ is defined on the entire range of the loss function and we want to partition the range into m intervals, let

$$
A_{\epsilon,m} = \left\{ z^{2n} \mid \sup_{\alpha \in \Lambda} \left\{ \sum_{j=0}^{m-1} \nu_{1j}^*\left(z^1, \alpha\right) - \sum_{j=0}^{m-1} \nu_{2j}^*\left(z^2, \alpha\right) \right\} \geq \epsilon \right\}
$$

$$
= \left\{ z^{2n} \mid \sup_{\alpha \in \Lambda} \sum_{j=0}^{m-1} \left(\nu_{1j}^*(z^1, \alpha, m) - \nu_{2j}^*(z^2, \alpha, m) \right) \geq \epsilon \right\}
$$

$$
\subseteq \left\{ z^{2n} \mid \exists\, j \;\; \sup_{\alpha \in \Lambda} \left(\nu_{1j}^*(z^1, \alpha, m) - \nu_{2j}^*(z^2, \alpha, m) \right) \geq \frac{\epsilon}{m} \right\}
$$

$$
\subseteq \bigcup_{j=0}^{m-1} \left\{ z^{2n} \mid \sup_{\alpha \in \Lambda} \left(\nu_{1j}^*(z^1, \alpha, m) - \nu_{2j}^*(z^2, \alpha, m) \right) \geq \frac{\epsilon}{m} \right\}
$$

$$
\text{where} \quad A_{\epsilon,m,j} = \left\{ z^{2n} : \sup_{\alpha \in \Lambda} \left(\nu_{1j}^*\left(z^1, \alpha, m\right) - \nu_{2j}^*\left(z^2, \alpha, m\right) \right) \geq \frac{\epsilon}{m} \right\}
$$

so that

$$
A_{\epsilon,m} \subseteq \bigcup_{j=0}^{m-1} A_{\epsilon,m,j}.
$$

Let suprema over Λ within $A_{\epsilon,m,j}$ be achieved at

$$
\alpha_j^* = \alpha_j^*(z^{2n}) = \arg \sup_{\alpha \in \Lambda} \left(\nu_{1j}^*\left(z_1, \alpha, m\right) - \nu_{2j}^*\left(z_2, \alpha, m\right) \right).
$$

Next, for any fixed z^{2n}, and any given α_j, form the vector

$$
(Q^*(z_1, \alpha_j, m), Q^*(z_2, \alpha_j, m), \ldots, Q^*(z_{2n}, \alpha_j, m))
$$

of the middle values of the intervals I_j, for $j = 0, 1, 2, \ldots, m-1$. For any α_j and $\alpha_{j'}$, write

$$
\alpha_j \sim \alpha_{j'} \iff
$$

$$
(Q^*(z_1, \alpha_j, m), \ldots, Q^*(z_{2n}, \alpha_j, m)) = (Q^*(z_1, \alpha_{j'}, m), \ldots, Q^*(z_{2n}, \alpha_{j'}, m)).
$$

So, for any fixed $Z^{2n} = z^{2n}$ it is seen that $\sim$ is an equivalence relation on Λ and therefore partitions Λ into disjoint equivalence classes. Denote the number of these classes by N_j^Λ and write

$$
N_j^\Lambda = N_j^\Lambda\left(Z^{2n}\right) = N_j^\Lambda(z_1, z_2, \ldots, z_{2n}).
$$

We define values α_{jk}^* as the canonical representatives of the equivalence classes where $k \in K_j$ is the kth equivalence class. Clearly, $\#(K_j) = N_j^\Lambda\left(Z^{2n}\right)$ and K_j is treated simply as an index set.

To make use of the above partitioning of A_ϵ, consider mapping the space $\mathcal{Z}^{2n}$ onto itself using $(2n)!$

distinct permutations $T = T_i$. Then, if f is integrable with respect to the distribution function of Z_i, its Riemann-Stieltjes integral satisfies

$$\int_{Z^{2n}} f\left(Z^{2n}\right) dF\left(Z^{2n}\right) = \int_{Z^{2n}} f\left(T_i Z^{2n}\right) dF\left(Z^{2n}\right),$$

and this gives

$$\int_{Z^{2n}} f\left(Z^{2n}\right) dF\left(Z^{2n}\right) = \int_{Z^{2n}} \frac{\sum_{i=1}^{(2n)!} f\left(T_i Z^{2n}\right)}{(2n)!} dF\left(Z^{2n}\right). \tag{2.4}$$

To make use of 2.4 we need the following results from Vapnik (1998).

Theorem 2.1.1. *In this theorem, we put together some background results from Vapnik Vapnik (1998).*

1. *Let $H_{\text{ann}}^{\Lambda}(n) = E \ln N^{\Lambda}(Z^n)$ denote the annealed entropy and $G^{\Lambda}(n) = \sup_{z^n} \ln N^{\Lambda}(Z^n)$ be the growth function. Then*

$$H_{ann}^{\Lambda}(n) \leq G^{\Lambda}(n) \leq h\left(1 + \ln\left(\frac{n}{h}\right)\right).$$

2. *Let $\Gamma = \sum_k \frac{\binom{m}{k}\binom{2n-m}{n-k}}{\binom{2n}{n}}$ where the summation is over k so that*

$$\left|\frac{k}{n} - \frac{m-k}{n}\right| > \epsilon, \quad \max(0, \ m - n) \leq k \leq \min(m, \ n)$$

where n and $m < 2n$ are arbitrary positive integers. Note that

$$\left|\frac{k}{n} - \frac{m-k}{n}\right| > \epsilon$$
$$\Leftrightarrow \left|\frac{2k}{n} - \frac{m}{n}\right| > \epsilon$$
$$\Leftrightarrow \frac{2}{n}\left|k - \frac{m}{2}\right| > \epsilon$$
$$\Leftrightarrow \left|k - \frac{m}{2}\right| > \frac{\epsilon n}{2}, \quad \max(0, \ m - n) \leq k \leq \min(m, \ n).$$

Then, $\exists\, C > 0$ so that $\Gamma < Ce^{-\epsilon n^2}$.

3. *Let $Q(\cdot, \alpha)$ for $\alpha \in \Lambda$ be a set of real-valued(not necessary bounded) non-negative functions and let $H_{ann}^{\Lambda}(n)$ be the annealed entropy of the level sets for $\{Q(\cdot, \alpha) \mid \alpha \in \Lambda\}$. Then, for any $1 < p \leq 2$, and a constant $D_p(\alpha)$ to be defined later, we have that*

$$P\left\{\sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i, \alpha)}{D_p(\alpha)} \geq \epsilon\right\} <$$
$$4\exp\left\{\left(\frac{H_{ann}^{\Lambda, \beta}(n)}{n^{2-\frac{2}{p}}} - \frac{\epsilon^2}{2^{1+\frac{2}{p}}}\right) n^{2-\frac{2}{p}}\right\}.$$

Proof. Clause 1 is proved in Theorem 4.3 in Vapnik (1998), p. 145; Clause 2 is proved in Sec. 4.13 of Vapnik (1998), p. 163, and Clause 3 is proved in Sec. 5.4 Vapnik (1998) p. 195. $\qquad \square$

To present our first important result let

$$\Delta_j \left(T_i Z^{2n}, \alpha_j^*, m \right) = \nu_{1j}^* \left(T_i Z_1, \alpha_j^*, m \right) - \nu_{2j}^* \left(T_i Z_2, \alpha_j^*, m \right).$$

We have the following.

Proposition 2.1.2. *Let* $\epsilon > 0$, $m \in \mathcal{N}$, *and* $h = VCdim\{Q(\cdot, \alpha) : \alpha \in \Lambda\}$, *and* n *be the sample size. If* h *is finite, then*

$$P\left(A_{\epsilon, m} \right) \leq 2m \left(\frac{2ne}{h} \right)^h \exp\left\{ -\frac{n\epsilon^2}{m^2} \right\}.$$

Proof. First, by elementary manipulations we have

$$
\begin{aligned}
P\left(A_{\epsilon, m} \right) &\leq P\left(\bigcup_{j=0}^{m-1} A_{\epsilon, m, j} \right) \leq \sum_{j=0}^{m-1} P\left(A_{\epsilon, m, j} \right) \\
&= \sum_{j=0}^{m-1} P\left(\left\{ Z^{2n} : \sup_{\alpha \in \Lambda} \left(\nu_{1j}^*(Z_1, \alpha, m) - \nu_{2j}^*(Z_2, \alpha, m) \right) \geq \frac{\epsilon}{m} \right\} \right).
\end{aligned}
$$

Continuing the equality gives

$$
\begin{aligned}
&= \sum_{j=0}^{m-1} P\left(\left\{ Z^{2n} : \sup_{\alpha \in \Lambda} \left(\left(\nu_{1j}^*(T_i Z_1, \alpha, m) - \nu_{2j}^*(T_i Z_2, \alpha, m) \right) \right) \geq \frac{\epsilon}{m} \right\} \right) \\
&= \sum_{j=0}^{m-1} P\left(\left\{ Z^{2n} : \left(\nu_{1j}^*(T_i Z_1, \alpha_j^*, m) - \nu_{2j}^*(T_i Z_2, \alpha_j^*, m) \right) \geq \frac{\epsilon}{m} \right\} \right) \\
&= \sum_{j=0}^{m-1} P\left(\left\{ Z^{2n} : \Delta_j \left(T_i Z, \alpha_j^*, m \right) \geq \frac{\epsilon}{m} \right\} \right) \\
&= \frac{1}{(2n)!} \sum_{j=0}^{m-1} \sum_{i=1}^{(2n)!} P\left(\left\{ Z^{2n} : \Delta_j \left(T_i Z, \alpha_j^*, m \right) \geq \frac{\epsilon}{m} \right\} \right) \\
&= \frac{1}{(2n)!} \sum_{j=0}^{m-1} \sum_{i=1}^{(2n)!} \int I_{\left\{ Z^{2n} : \Delta_j \left(T_i Z, \alpha_j^*, m \right) \geq \frac{\epsilon}{m} \right\}} \left(z^{2n} \right) dP(Z^{2n}). \quad (2.5)
\end{aligned}
$$

Let the equivalence classes in Λ under $\sim$ be denoted Λ_k. Then, the equivalence classes $\left[\alpha_{jk}^* \right]$ for the α_{jk}^*'s provide a partition for Λ. That is, $\Lambda = \bigcup_{k=1}^{N_j^\Lambda(z^{2n})} [\alpha_{jk}^*]$ because $\alpha_{jk}^* \in \Lambda_k$ and hence $[\alpha_{jk}^*] = \Lambda_k$. In

addition, α_{jk}^* is the maximum value of α_j in the kth equivalence class. So,

$$
\begin{aligned}
I_{\left\{Z^{2n}:\Delta_j\left(T_iZ,\alpha_j^*,m\right)\geq\epsilon\right\}}\left(z^{2n}\right) \;\leq\; & I_{\left\{Z^{2n}:\Delta_j\left(T_iZ,\alpha_{1j}^*,m\right)\geq\frac{\epsilon}{m}\right\}}\left(z^{2n}\right) \\
& +\;\cdots+I_{\left\{Z^{2n}:\Delta_j\left(T_iZ,\alpha_{N_j^\Lambda(z^{2n})j}^*,m\right)\geq\frac{\epsilon}{m}\right\}}\left(z^{2n}\right) \\
=\; & \sum_{k=1}^{N_j^\Lambda(z^{2n})} I_{\left\{Z^{2n}:\Delta_j\left(T_iZ,\alpha_{kj}^*,m\right)\geq\frac{\epsilon}{m}\right\}}\left(z^{2n}\right) \qquad (2.6)
\end{aligned}
$$

where

$$
\begin{aligned}
A_{\epsilon,m,j,k} \;=\; & \left\{Z^{2n}:\Delta_j\left(T_iZ,\alpha_{kj}^*,m\right)\geq\frac{\epsilon}{m}\right\} \\
=\; & \left\{Z^{2n}:\sup_{\alpha_j\in\Lambda_k}\left(\nu_{1j}^*(T_iZ_1,\alpha_j,m)-\nu_{2j}^*(T_iZ_2,\alpha_j,m)\right)\geq\frac{\epsilon}{m}\right\}
\end{aligned}
$$

Now, using 2.6, 2.5 is bounded by

$$
\begin{aligned}
P\left(A_{\epsilon,m}\right) \;\leq\; & \frac{1}{(2n)!}\sum_{j=0}^{m-1}\sum_{i=1}^{(2n)!}\int\sum_{k=1}^{N_j^\Lambda(z^{2n})} I_{\left\{Z^{2n}:\Delta_j(T_iZ,\alpha_{kj}^*,m)\geq\frac{\epsilon}{m}\right\}}\left(z^{2n}\right)dP\left(z^{2n}\right) \\
=\; & \int\sum_{j=0}^{m-1}\sum_{k=1}^{N_j^\Lambda(z^{2n})}\left[\frac{1}{(2n)!}\sum_{i=1}^{(2n)!} I_{\left\{Z^{2n}:\Delta_j(T_iZ,\alpha_{kj}^*,m)\geq\frac{\epsilon}{m}\right\}}\left(z^{2n}\right)\right]dP\left(z^{2n}\right). \qquad (2.7)
\end{aligned}
$$

The expression in square brackets in 2.7 is the fraction of the number of the $(2n)!$ permutations T_i of Z^{2n} for which $A_{\epsilon,m,j,k}$ is closed under T_i for any fixed equivalence class Λ_k. It is equal to

$$
\Gamma_j = \sum_k \frac{\binom{m_j^*}{k}\binom{2n-m_j^*}{m_j^*-k}}{\binom{2n}{n}}
$$

where

$$
\begin{cases}
\left\{k:\left|\frac{k}{n}-\frac{m_j^*-k}{n}\right|\geq\frac{\epsilon}{m}\right\} & ; \\
m_j^*=n_{1j}^*+n_{2j}^* & .
\end{cases}
$$

Here, Γ_j is the probability of choosing exactly k sample data points whose losses fall in interval I_j respectively in the first and second half of the sample such that $A_{\epsilon,m,j,k}$ holds. m_j^* is the number of data points from the first and the second half of the sample whose losses landed inside interval j. Using Theorem 2.1.1, Clause II, we have $\Gamma_j\leq 2\exp\left(-\frac{n\epsilon^2}{m^2}\right)$. So, using this in equation 2.7 gives $P\left(A_\epsilon\right)$ is

upper bounded by

$$\int \sum_{j=0}^{m-1} \sum_{k=1}^{N_j^\Lambda(z^{2n})} 2 \exp\left(-\frac{n\epsilon^2}{m^2}\right) dP(z^{2n}) = 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \int \sum_{j=0}^{m-1} \sum_{k=1}^{N_j^\Lambda(z^{2n})} dP(z^{2n})$$

$$= 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \sum_{j=0}^{m-1} \int \sum_{k=1}^{N_j^\Lambda(z^{2n})} dP(z^{2n}) = 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \sum_{j=0}^{m-1} \int N_j^\Lambda(z^{2n}) dP(z^{2n})$$

$$= 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \sum_{j=0}^{m-1} E\left(N_j^\Lambda(z^{2n})\right). \tag{2.8}$$

Since Theorem 2.1.1, Clause I gives

$$H_{ann}(Z^{2n}) = \ln\left(E\left(N_j^\Lambda(Z^{2n})\right)\right) \leq G(2n) \leq h\ln\left(\frac{2ne}{h}\right) \Rightarrow E\left(N_j^\Lambda(z^{2n})\right) \leq \left(\frac{2ne}{h}\right)^h.$$

Using this m times in 2.8 gives the Proposition. $\qquad\square$

Proposition 2.1.1 gives us an upper bound on the expected supremal difference. We will use this upper bound to bound the unknown true risk in the following proposition. Let $R(\alpha_k)$ be the true unknown risk at α_k and $R_{emp}(\alpha_k)$ be the empirical risk at α_k.

Proposition 2.1.3. *With probability $1 - \eta$, the inequality*

$$R(\alpha_k) \leq R_{emp}(\alpha_k) + m\sqrt{\frac{1}{n}\log\left(\left(\frac{2m}{\eta}\right)\left(\frac{2ne}{h}\right)^h\right)} \tag{2.9}$$

holds simultaneously for all functions $Q(z, \alpha_k)$, $k = 1, 2, \cdots, K$.

This inequality suggests that, the best model will be the one that minimizes the RHS of inequality 2.9. The use of inequality 2.9 in model selection is called Empirical Risk Minimization (ERM).

Proof. To obtain inequality (2.9), we have to equate the RHS of Proposition 2.1.2 to some positive number $0 \leq \eta \leq 1$. That's

$$\eta = 2m\left(\frac{2ne}{h}\right)^h \exp\left(-\frac{n\epsilon^2}{m^2}\right)$$

and solve it for ϵ, one will get

$$\epsilon = m\sqrt{\frac{1}{n}\log\left(\left(\frac{2m}{\eta}\right)\left(\frac{2ne}{h}\right)^h\right)}. \tag{2.10}$$

Proposition 2.1.3 can be obtained from the additive Chernoff bounds, expression 4.4 in Vapnik (1998) as

follows

$$R\left(\alpha_k\right) \quad \leq \quad R_{emp}\left(\alpha_k\right) + \epsilon. \tag{2.11}$$

Using (2.10) in inequality (2.11), completes the proof. $\qquad\square$

Proposition 2.1.4. *With probability* $1 - \eta$*, the inequality*

$$R\left(\alpha_k\right) \leq R_{emp}\left(\alpha_k\right) + \frac{m^2}{2n}\log\left(\frac{2m}{\eta}\left(\frac{2ne}{h}\right)^h\right)\left(1 + \sqrt{1 + \frac{4nR_{emp}\left(\alpha_k\right)}{m^2\log\left(\frac{2m}{\eta}\left(\frac{2ne}{h}\right)^h\right)}}\right) \tag{2.12}$$

holds simultaneously for all K *functions in the set* $Q\left(z, \alpha_k\right)$*,* $k = 1, 2, \ldots, K$*.*

As we said before, this inequality suggests that, the best model will be the one that minimizes the RHS of 2.12. The use of inequalities (2.1.3) and (2.12) in model selection are called Empirical Risk Minimization (ERM).

Proof. Let $\epsilon, \eta > 0$. Then, inequality (4.18) in Vapnik (1998) gives, with probability at least $1 - \eta$, that

$$\frac{R\left(\alpha_k\right) - R_{emp}\left(\alpha_k\right)}{\sqrt{R\left(\alpha_k\right)}} \leq \epsilon.$$

Routine algebraic manipulations and completing the square give

$$\left(R\left(\alpha_k\right) - 0.5\left(\epsilon^2 + 2R_{emp}\left(\alpha_k\right)\right)\right)^2 - 0.25\left(\epsilon^2 + 2R_{emp}\left(\alpha_k\right)\right)^2 \leq -R_{emp}^2\left(\alpha_k\right).$$

Taking the square root on both sides and re-arranging gives

$$R\left(\alpha_k\right) \leq R_{emp}\left(\alpha_k\right) + 0.5\epsilon^2\left(1 + \sqrt{1 + \frac{4R_{emp}\left(\alpha_k\right)}{\epsilon^2}}\right)$$

Using (2.10) in the last inequality completes the proof of the Proposition. $\qquad\square$

The next Theorem will bound the EMDBTEL using proposition 2.1.1

Theorem 2.1.2. *Let* $h = VCDim\left\{Q\left(\cdot, \alpha\right) : \alpha \in \Lambda\right\}$*,* $m \in \mathcal{N}$ *be the number of discretized intervals,* $n \in \mathcal{N}$ *be the sample size. If* $h < \infty$*, we have*

$$E\left(\sup_{\alpha \in \Lambda}\left|\nu_1^*(z^1, \alpha) - \nu_2^*(z^2, \alpha)\right|\right) \leq m\sqrt{\frac{1}{n}\ln\left(2m^3\left(\frac{2ne}{h}\right)^h\right)} + \frac{1}{m\sqrt{n\ln\left(2m^3\left(\frac{2ne}{h}\right)^h\right)}}$$

Proof. The LHS of the statement of the theorem equals

$$\int_0^\infty P\left(A_{\epsilon,m}\right) d\epsilon \leq \int_0^u d\epsilon + \int_u^\infty P\left(A_\epsilon\right) d\epsilon \leq u + 2m \left(\frac{2ne}{h}\right)^h \int_u^\infty \exp\left(-\frac{n\epsilon^2}{m^2}\right) d\epsilon. \qquad (2.13)$$

Observing that

$$\epsilon > u \Rightarrow \frac{n\epsilon^2}{m^2} > \frac{nu\epsilon}{m^2} \Rightarrow \int_u^\infty \exp\left(-\frac{n\epsilon^2}{m^2}\right) d\epsilon$$

$$\leq \int_u^\infty \exp\left(-\frac{nu\epsilon}{m^2}\right) d\epsilon = \frac{m^2}{nu} \exp\left(-\frac{nu^2}{m^2}\right),$$

the bound in (2.13) gives

$$E\left(\sup_{\alpha \in \Lambda} \left(\nu_1^*(z^1,\alpha) - \nu_2^*(z^2,\alpha)\right)\right) \leq u + 2m^3 \left(\frac{2ne}{h}\right)^h \frac{1}{nu} \exp\left(-\frac{nu^2}{m^2}\right). \qquad (2.14)$$

substituting

$$u = m\sqrt{\frac{1}{n}\ln\left(2m^3\left(\frac{2ne}{h}\right)^h\right)}$$

in (2.14) gives the statement of the Theorem. $\qquad \square$

Now, we turn our attention to developing a second upper bound on the expected supremal difference. We begin with two elementary lemmas.

Lemma 2.1.1. *Let $\epsilon > 0$. The probability of the supremal difference between two empirical losses is bounded by*

$$P\left(\sup_{\alpha \in \Lambda}\left(|\nu_1\left(Z^1,\alpha\right) - \nu_2\left(Z^2,\alpha\right)| \geq \epsilon\right)\right) \leq$$

$$2P\left(\sup_{\alpha \in \Lambda}\left(|\nu_1(Z^1,\alpha) - E(Q(Z^1,\alpha))| \geq \frac{\epsilon}{2}\right)\right).$$

Proof. The left hand side is bounded by

$$P\left(\sup_{\alpha \in \Lambda}\left|\nu_1\left(Z^1,\alpha\right) - E\left(Q(Z^1,\alpha)\right) + E\left(Q(Z^1,\alpha) - \nu_2\left(Z^2,\alpha\right)\right)\right| \geq \epsilon\right)$$

$$\leq P\left(\sup_{\alpha \in \Lambda}\left|\nu_1\left(Z^1,\alpha\right) - E\left(Q(Z^1,\alpha)\right)\right| \geq \frac{\epsilon}{2}\right) +$$

$$P\left(\sup_{\alpha \in \Lambda}\left|E\left(Q(Z^1,\alpha)\right) - \nu_2\left(Z^2,\alpha\right)\right| \geq \frac{\epsilon}{2}\right)$$

$$= 2P\left(\sup_{\alpha \in \Lambda}\left|\nu_1\left(Z^1,\alpha\right) - E\left(Q(Z^1,\alpha)\right)\right| \geq \frac{\epsilon}{2}\right).$$

$\qquad \square$

To state Lemma 2.1.2, let

$$\overline{Q(Z,\alpha)}^- = \frac{1}{n}\sum_{i=1}^{n}Q(z_i,\alpha)^- \quad \text{and} \quad \overline{Q(Z,\alpha)}^+ = \frac{1}{n}\sum_{i=1}^{n}Q(z_i,\alpha)^+$$

and assume

$$D_p(\alpha) = \int_0^\infty \sqrt[p]{P\{Q(z,\alpha) \ge c\}}dc < \infty.$$

Lemma 2.1.2. *Let $Q(Z,\alpha)$ be a nonnegative function. Then,*

$$P\left(\sup_{\alpha\in\Lambda}\frac{\left(E\left(Q(Z,\alpha)\right)-\overline{Q(Z,\alpha)}\right)}{D_p(\alpha)} \ge \epsilon\right) \le$$

$$P\left(\sup_{\alpha\in\Lambda}\frac{\left(E\left(Q(Z,\alpha)^-\right)-\overline{Q(Z,\alpha)}^-\right)}{D_p(\alpha)} \ge \frac{\epsilon}{2}\right) +$$

$$P\left(\sup_{\alpha\in\Lambda}\frac{\left(E\left(Q(Z,\alpha)^+\right)-\overline{Q(Z,\alpha)}^+\right)}{D_p(\alpha)} \ge \frac{\epsilon}{2}\right).$$

Proof. The LHS is

$$P\left(\sup_{\alpha\in\Lambda}\frac{1}{D_p(\alpha)}\left(\left(E\left(\overline{Q(Z,\alpha)}^+\right)-\overline{Q(Z,\alpha)}^+\right)+\left(E\left(Q(Z,\alpha)^-\right)-\overline{Q(Z,\alpha)}^-\right)\right) \ge \epsilon\right)$$

$$\le P\left(\left\{\left(\sup_{\alpha\in\Lambda}\frac{1}{D_p(\alpha)}\left(E\left(Q(Z,\alpha)^+\right)-\overline{Q(Z,\alpha)}^+\right)+\sup_{\alpha\in\Lambda}\frac{1}{D_p(\alpha)}\left(E\left(Q(Z,\alpha)^-\right)-\overline{Q(Z,\alpha)}^-\right)\right) \ge \epsilon\right\}\right.$$

$$\left.\cap\left\{\sup_{\alpha\in\Lambda}\frac{1}{D_p(\alpha)}\left(E\left(Q(Z,\alpha)^+\right)-Q(Z,\alpha)^+|\right) \ge \frac{\epsilon}{2}\right\}\right)$$

$$+P\left(\left\{\left(\sup_{\alpha\in\Lambda}\frac{1}{D_p(\alpha)}\left(E\left(Q(Z,\alpha)^+\right)-\overline{Q(Z,\alpha)}^+\right)+\sup_{\alpha\in\Lambda}\frac{1}{D_p(\alpha)}\left(E\left(Q(Z,\alpha)^-\right)-\overline{Q(Z,\alpha)}^-\right)\right) \ge \epsilon\right\}\right.$$

$$\left.\cap\left\{\sup_{\alpha\in\Lambda}\frac{1}{D_p(\alpha)}\left(E\left(Q(Z,\alpha)^+\right)-\overline{Q(Z,\alpha)}^+\right) \ge \frac{\epsilon}{2}\right\}\right)$$

$$\le P\left(\sup_{\alpha\in\Lambda}\frac{1}{D_p(\alpha)}\left(E\left(Q(Z,\alpha)^-\right)-\overline{Q(Z,\alpha)}^-\right) \ge \frac{\epsilon}{2}\right)$$

$$+ P\left(\sup_{\alpha\in\Lambda}\frac{1}{D_p(\alpha)}\left(E\left(Q(Z,\alpha)^+\right)-\overline{Q(Z,\alpha)}^+\right) \ge \frac{\epsilon}{2}\right).$$

$\square$

Proposition 2.1.5. *Let $h = VCDim\{Q(\cdot,\alpha) : \alpha \in \Lambda\}$, where $Q(\cdot,\alpha)$ is unbounded, $n \in \mathcal{N}$ be the sample size. If $h < \infty$, and*

$$D_p(\alpha) = \int_0^\infty \sqrt[p]{P\{Q(z,\alpha) \ge c\}}dc \le \infty$$

where $1 < p \le 2$ *is some fixed parameter, we have*

$$P\left(\sup_{\alpha \in \Lambda} \left|\nu_1\left(Z^1, \alpha\right) - \nu_2\left(Z^2, \alpha\right)\right| \ge \epsilon\right) \le \quad 16\left(\frac{ne}{h}\right)^h \exp\left\{-\left(\frac{\epsilon n^{1-\frac{1}{p}}}{D_p(\alpha^*)2^{2.5+\frac{1}{p}}}\right)^2\right\}, \qquad (2.15)$$

where $\alpha^* = \sup_{\alpha \in \Lambda} D_p(\alpha)$.

Proof. By Lemma 2.1.1, we have

$$P\left(\sup_{\alpha \in \Lambda} \left|\nu_1\left(Z^1, \alpha\right) - \nu_2\left(Z^2, \alpha\right)\right| \ge \epsilon\right) \le \quad 2P\left(\sup_{\alpha \in \Lambda} \left|\nu_1\left(z^1, \alpha\right) - E\left(Q\left(z^1, \alpha\right)\right)\right| \ge \frac{\epsilon}{2}\right) \qquad (2.16)$$

To bound the RHS of 2.16, use Lemma 2.1.2 to observe that

$$P\left(\sup_{\alpha \in \Lambda} \left|\nu_n\left(z, \alpha\right) - E\left(Q(z, \alpha)\right)\right| \ge \frac{\epsilon}{2}\right) \le P\left(\sup_{\alpha \in \Lambda} \left|E\left(Q(z, \alpha)^+\right) - \frac{1}{n}\sum_{i=1}^n Q(z_i, \alpha)^+\right| \ge \frac{\epsilon}{4}\right)$$

$$+ P\left(\sup_{\alpha \in \Lambda} \left|E\left(Q(z, \alpha)^-\right) - \frac{1}{n}\sum_{i=1}^n Q(z, \alpha)^-\right| \ge \frac{\epsilon}{4}\right). \qquad (2.17)$$

Each probability on the right hand side of (2.17) can be bounded. Since α^* is a maximum, we have $1/D_p(\alpha^*) \le 1/D_p(\alpha)$. Thus, for either the positive or negative parts in (2.17) we have

$$\sup_{\alpha \in \Lambda} \frac{E\left(Q(z, \alpha)\right) - \nu_n(z, \alpha)}{D_p(\alpha^*)} \le \sup_{\alpha \in \Lambda} \frac{E\left(Q(z, \alpha)\right) - \nu_n(z, \alpha)}{D_p(\alpha)}$$

$$P\left(\sup_{\alpha \in \Lambda} \frac{E\left(Q(z, \alpha)\right) - \nu_n(z, \alpha)}{D_p(\alpha^*)} \ge \frac{\epsilon}{4D_p(\alpha^*)}\right) \le$$

$$P\left(\sup_{\alpha \in \Lambda} \frac{E\left(Q(z, \alpha)\right) - \nu_n(z, \alpha)}{D_p(\alpha)} \ge \frac{\epsilon}{4D_p(\alpha^*)}\right) \Rightarrow$$

$$P\left(\sup_{\alpha \in \Lambda} \left(E\left(Q(z, \alpha)\right) - \nu_n(z, \alpha)\right) \ge \frac{\epsilon}{4}\right) \le$$

$$P\left(\sup_{\alpha \in \Lambda} \frac{E\left(Q(z, \alpha)\right) - \nu_n(z, \alpha)}{D_p(\alpha)} \ge \frac{\epsilon}{4D_p(\alpha^*)}\right)$$

Letting $\delta = \frac{\epsilon}{4D_p(\alpha^*)}$ and using Theorem 2.1.1, Clause 3, the last inequality gives

$$P\left(\sup_{\alpha \in \Lambda} \left(E\left(Q(z, \alpha)\right) - \nu_n(z, \alpha)\right) \ge \frac{\epsilon}{4}\right) \le 4\exp\left\{\left(\frac{H_{ann}^{\Lambda, \beta}(n)}{n^{2-\frac{2}{p}}} - \frac{\delta^2}{2^{1+\frac{2}{p}}}\right) n^{2-\frac{2}{p}}\right\}. \qquad (2.18)$$

Using (2.18), both terms on the right in (2.17) can be bounded. This gives

$$P\left(\sup_{\alpha \in \Lambda} \left|\nu_n(z, \alpha) - E\left(Q(z, \alpha)\right)\right| \ge \frac{\epsilon}{2}\right) \le 8\exp\left\{\left(\frac{H_{ann}^{\Lambda, \beta}(n)}{n^{2-\frac{2}{p}}} - \frac{\delta^2}{2^{1+\frac{2}{p}}}\right) n^{2-\frac{2}{p}}\right\}. \qquad (2.19)$$

The presence of β in the the exponent of the annealed entropy will not change the validity of Clause I of Theorem 2.1.1, so we have, $H_{ann}^{\Lambda, \beta}(n) \le G^\Lambda(n) \le \ln\left(\frac{en}{h}\right)^h$ therefore, $\exp\left(H_{ann}^{\Lambda, \delta}(n)\right) \le \exp\left(G(n)\right) \le$

$\left(\frac{en}{h}\right)^h$. Using this in (2.19) gives

$$P\left(\sup_{\alpha\in\Lambda}\left(\nu_n(z,\alpha)-E\left(Q(z,\alpha)\right)\right)\geq\frac{\epsilon}{2}\right)\leq 8\left(\frac{ne}{h}\right)^h\exp\left\{-\left(\frac{\delta n^{1-\frac{1}{p}}}{2^{0.5+\frac{1}{p}}}\right)^2\right\}$$

$$=8\left(\frac{ne}{h}\right)^h\exp\left\{-\left(\frac{\epsilon n^{1-\frac{1}{p}}}{D_p(\alpha^*)2^{2.5+\frac{1}{p}}}\right)^2\right\}.$$

Recalling the extra factor of 2 in Inequation (2.16) gives the statement of the Proposition. $\qquad\square$

Theorem 2.1.3. *Let $h=VCDim\left\{Q\left(\cdot,\alpha\right):\alpha\in\Lambda\right\}$ where $Q\left(\cdot,\alpha\right)$ is unbounded, $n\in\mathcal{N}$ is the sample size. If $h\leq\infty$, and*

$$D_p\left(\alpha\right)=\int_0^\infty\sqrt[p]{P\left\{Q\left(z,\alpha\right)\geq c\right\}}dc\leq\infty$$

where $1<p\leq 2$ is some fixed parameter, we have

$$E\left(\sup_{\alpha\in\Lambda}|\nu_1(Z_1,\alpha)-\nu_2(Z_2,\alpha)|\right)\leq\frac{D_p(\alpha^*)2^{2.5+\frac{1}{p}}\sqrt{h\ln\left(\frac{ne}{h}\right)}}{n^{1-\frac{1}{p}}}+\frac{16D_p(\alpha^*)2^{2.5+\frac{1}{p}}}{n^{1-\frac{1}{p}}\sqrt{h\ln\left(\frac{ne}{h}\right)}}.\qquad(2.20)$$

Proof. Using the integral of probabilities identity, the left-hand side of (2.20) equals

$$\int_0^\infty P\left(\sup_{\alpha\in\Lambda}|\nu_1(Z_1,\alpha)-\nu_2(Z_2,\alpha)|\geq\epsilon\right)d\epsilon$$

$$\leq\int_0^\infty 16\left(\frac{ne}{h}\right)^h\exp\left\{-\left(\frac{\epsilon n^{1-\frac{1}{p}}}{D_p(\alpha^*)2^{2.5+\frac{1}{p}}}\right)^2\right\}d\epsilon$$

$$\equiv\int_0^u d\epsilon+16\left(\frac{ne}{h}\right)^h\int_u^\infty\exp\left\{-\left(\frac{\epsilon n^{1-\frac{1}{p}}}{D_p(\alpha^*)2^{2.5+\frac{1}{p}}}\right)^2\epsilon^2\right\}d\epsilon$$

$$\leq u+16\left(\frac{ne}{h}\right)^h\int_u^\infty\exp\left\{-\left(\frac{n^{1-\frac{1}{p}}}{D_p(\alpha^*)2^{2.5+\frac{1}{p}}}\right)^2 u\epsilon\right\}d\epsilon$$

$$=u+\left(\frac{D_p(\alpha^*)2^{2.5+\frac{1}{p}}}{n^{1-\frac{1}{p}}}\right)^2\frac{16\left(\frac{ne}{h}\right)^h}{u}\exp\left\{-\left(\frac{un^{1-\frac{1}{p}}}{D_p(\alpha^*)2^{2.5+\frac{1}{p}}}\right)^2\right\}.$$

Choosing

$$u=D_p(\alpha^*)2^{2.5+\frac{1}{p}}\frac{\sqrt{h\ln\left(\frac{ne}{h}\right)}}{n^{1-\frac{1}{p}}}$$

gives the statement of the Theorem. $\qquad\square$

Finally, we summarize the results of Theorem 2.1.2, and 2.1.3. It identifies the conditions under which the tightest upper bound we are able to derive for the expected supremal difference holds.

Theorem 2.1.4. *Assume that $h \to \infty$, $\frac{n}{h} \to \infty$, $m \to \infty$, $\ln(m) = o(n)$,*

$$D_p(\alpha) = \int_0^\infty \sqrt[p]{P\{Q(z,\alpha) \geq c\}} dc \leq \infty,$$

here $p = 2$, we have that

$$E\left(\sup_{\alpha \in \Lambda} |\nu_1(Z_1, \alpha) - \nu_2(Z_2, \alpha)|\right) \leq \min(1, 8D_p(\alpha^*))\sqrt{\frac{h}{n}\ln\left(\frac{2ne}{h}\right)}. \tag{2.21}$$

Proof. Theorems 2.1.2 and 2.1.3 give

$$
\begin{aligned}
E\left(\sup_{\alpha \in \Lambda} |\nu_1^*(Z_1, \alpha, m) - \nu_2^*(Z_2, \alpha, m)|\right) &\leq \sqrt{\frac{\ln(2m)}{n} + \frac{h}{n}\ln\left(\frac{2ne}{h}\right)} \\
&\overset{\infty}{=} \sqrt{\frac{h}{n}\ln\left(\frac{2ne}{h}\right)},
\end{aligned} \tag{2.22}
$$

where $\overset{\infty}{=}$ indicates a limit as $n \to \infty$ has been taken. Similarly, Theorem 2.1.3 gives

$$E\left(\sup_{\alpha \in \Lambda} |\nu_1(Z_1, \alpha) - \nu_2(Z_2, \alpha))|\right) \leq 8D_p(\alpha^*)\sqrt{\frac{h}{n} + \frac{h}{n}\ln\left(\frac{2ne}{h}\right)} + E(n, h) \tag{2.23}$$

where $E(n, h)$ is of smaller order than the first term on the right in 2.23. Thus,

$$E\left(\sup_{\alpha \in \Lambda} |\nu_1(Z_1, \alpha) - \nu_2(Z_2, \alpha))|\right) \leq 8D_p(\alpha^*)\sqrt{\frac{h}{n}\ln\left(\frac{2ne}{h}\right)}. \tag{2.24}$$

Taking the minimum over the RHS of equations (2.22) and (2.24) gives the Theorem. $\square$

In Chap. 3, we will see that this more or less direct extension of Vapnik's bounds performs badly because, typically the bound is far too loose. This happens because the error criterion is too "small". So, in the next sub-section we change the error criterion to make it 'bigger' while preserving the structure of the derivation of the upper bounds in Proposition 2.1.2, and in Theorem 2.1.2, 2.1.3 and 2.1.4. This makes the upper bound tighter and as we will see in Chap. 3, we can optimize over the constant factor in the upper bound of Theorem 2.1.4 to get a better estimate for h.

2.2 Change in the Expected Maximum Deviation Quantity

The numerical effect of the change in the calculation of EMDBTEL will be observed in Sec. . In this section, we will restate all results using a cross-validation form of error. Here, we will present only the main changes. In fact, the computation of the maximum difference is done with the idea of cross-validation in mind. Thus, the losses whose difference we take are found using data that was not used to build the

model. Moreover, instead of using one model, we use two, one for each half sample. The two models are identical in that they have the same covariates.

The stream of reasoning used to develop the results in the previous section will remain the same here. However, there are some changes in detail that will be noted.

Here we identify main changes so that the remainder of the proof will follow. Let $Z = (x, y)$ be a pair of observations and write $Z^1 = (z_1, z_2, \cdots, z_n)$ and $Z^2 = (z_{n+1}, z_{n+2}, \cdots, z_{2n})$ two vectors of n independent and identically distributed (IID) copies of Z. Let

$$Q_1\left(z_1, \alpha_1\right) = L\left(y, f\left(x, \alpha_1\right)\right) \quad \text{and} \quad Q_2\left(z_2, \alpha_2\right) = L\left(y, f\left(x, \alpha_2\right)\right)$$

be two bounded real valued loss functions where $\alpha_i \in \Lambda$, $i = 1, 2$, an index set, and assume $\forall\ \alpha_i$, $0 \leq Q_i\left(z_i, \alpha_i\right) \leq B_i$ for some $B_i \in \mathbb{R}$, $i = 1, 2$. Consider the discretization of Q_i using m disjoint intervals (with union $[0; B_i)$) given by

$$Q^*_{ij}\left(z_i, \alpha_i, m\right) = \begin{cases} \frac{(2j+1)B_i}{2m}, & \text{if } Q_i\left(z, \alpha\right) \in\ I_j = \left[\frac{jB_i}{m}, \frac{(j+1)B_i}{m}\right),\ i = 1, 2; \\ \\ 0 & \text{otherwise.} \end{cases} \tag{2.25}$$

Where $j = 0, 1, \cdots, m - 1$. Now, consider indicator functions for Q_i being in an interval of the same form. That is, let

$$\chi_{I_{ij}}(Q_i\left(z, \alpha, m\right)) = \begin{cases} 1, & \text{if } Q_i\left(z, \alpha\right) \in I_j = \left[\frac{jB_i}{m}, \frac{(j+1)B_i}{m}\right),\ i = 1, 2; \\ \\ 0, & \text{otherwise.} \end{cases} \tag{2.26}$$

and write
$$n^*_{1j} = \sum_{i=1}^{n} \chi_{I_{1j}}\left(Q_1(z_i^2, \alpha_1, m)\right) \quad \text{and} \quad n^*_{2j} = \sum_{i=n+1}^{2n} \chi_{I_{2j}}\left(Q_2(z_i^1, \alpha_2, m)\right)$$

for the number of data points whose losses land inside the interval I_j in the first and second half of the sample of size n, respectively. The empirical loss for each model will now be written as follows:

$$\nu^*_{1j}\left(z^2, \alpha_1, m\right) = \frac{n^*_{2j}Q^*_{2j}\left(z^2, \alpha_1, m\right)}{n}, \ \nu^*_{2j}\left(z^1, \alpha_2, m\right) = \frac{n^*_{1j}Q^*_{1j}\left(z^1, \alpha_2, m\right)}{n}.$$

We see that in the computation of the empirical loss, we use datasets that were not used to fit the model. The event $A_{\epsilon,m}$ will now be defined as

$$A_{\epsilon,m} = \left\{ z^{2n} : \sup_{\alpha_1, \alpha_2}\ \left(\nu^*_1\left(z^2, \alpha_1\right) - \nu^*_2\left(z^1, \alpha_2\right)\right) \geq \epsilon \right\} \tag{2.27}$$

$$\text{where } \nu_1^* \left(z^2, \alpha_1 \right) = \sum_{j=0}^{m-1} \nu_{1j}^* \left(z^2, \alpha_1, m \right), \ \nu_2^* \left(z^1, \alpha_2 \right) = \sum_{j=0}^{m-1} \nu_{2j}^* \left(z^1, \alpha_2, m \right).$$

Since A_ϵ is defined on the entire range of our loss function, and we want to partition the range into m disjoint intervals, let

$$
\begin{aligned}
A_{\epsilon,m} &= \left\{ z^{2n} \mid \sup_{\alpha_1, \ \alpha_2 \in \Lambda} \left\{ \sum_{j=0}^{m-1} \nu_{1j}^* \left(z^2, \alpha_1 \right) - \sum_{j=0}^{m-1} \nu_{2j}^* \left(z^1, \alpha_2 \right) \right\} \geq \epsilon \right\} \\
&\subseteq \left\{ z^{2n} \mid \exists \ j : \sup_{\alpha_1, \ \alpha_2 \in \ \Lambda} \left(\nu_{1j}^* \left(z^2, \alpha_2, m \right) - \nu_{2j}^* \left(z^1, \alpha_2, m \right) \right) \geq \frac{\epsilon}{m} \right\} \\
&\subseteq \bigcup_{j=0}^{m-1} \left\{ z^{2n} \mid \sup_{\alpha_1, \ \alpha_2 \ \in \Lambda} \left(\nu_{1j}^* \left(z^2, \alpha_1, m \right) - \nu_{2j}^* \left(z^1, \alpha_2, m \right) \right) \geq \frac{\epsilon}{m} \right\} \\
&\subseteq \bigcup_{j=0}^{m-1} A_{\epsilon,m,j}
\end{aligned}
$$

where $A_{\epsilon,m,j} = \left\{ z^{2n} \mid \sup_{\alpha_1, \ \alpha_2 \ \in \Lambda} \left(\nu_{1j}^* \left(z^2, \alpha_1, m \right) - \nu_{2j}^* \left(z^1, \alpha_2, m \right) \right) \geq \frac{\epsilon}{m} \right\}$. The suprema over Λ within $A_{\epsilon,m,j}$ will be achieved at

$$\alpha_j^* = \alpha_j^* \left(z^{2n} \right) = \arg \sup_{\alpha_1, \ \alpha_2} \left(\nu_{1j}^* \left(z^2, \alpha_1, m \right) - \nu_{2j}^* \left(z^1, \alpha_2, m \right) \right).$$

Every development done in page 23 will stay unchanged. Using the cross-validation form of the error and using two different models will not affect Theorem 2.1.1. However, since the use of these changes are intense in the proof of Proposition 2.1.2, we will restate it and redo the proof taking into account these changes.

Proposition 2.2.1. *Let* $\epsilon \geq 0, \quad m \in \mathcal{N}, and, \quad h = VCD \left\{ Q \left(\cdot, \alpha \right) : \alpha \in \Lambda \right\}$. *If* h *is finite, then*

$$P \left(A_{\epsilon,m} \right) \leq 2m \left(\frac{2ne}{h} \right)^h \exp \left\{ -\frac{n\epsilon^2}{m^2} \right\}. \tag{2.28}$$

This proposition is very similar in spirit to Proposition 2.1.2. The difference is mainly in the setting. This being so, in our proof of Proposition 2.2.1, we will emphasize the steps where changes to the argument are important.

Proof. Let $\Delta_j^* \left(T_i Z^{2n}, \alpha_j^*, m \right) = \nu_{1j}^* \left(T_i Z_2, \alpha_{1j}^*, m \right) - \nu_{2j}^* \left(T_i Z_1, \alpha_{2j}^*, m \right)$, where $\alpha_j = \left(\alpha_{1j}, \ \alpha_{2j} \right)$ and

$\alpha_j^* = \{\alpha_{1j}, \; \alpha_{2j} : \arg\max \Delta_j\left(T_i Z, \alpha_j, m\right)\}$. Using some manipulations, we have

$$
\begin{aligned}
P\left(A_{\epsilon,m}\right) \quad &\leq \quad P\left(\bigcup_{j=0}^{m-1} A_{\epsilon,m,j}\right) \\
&\leq \quad \sum_{j=0}^{m-1} P\left(A_{\epsilon,m,j}\right) \\
&= \quad \sum_{j=0}^{m-1} P\left(\left\{Z^{2n} : \sup_{\alpha_1, \; \alpha_2 \in \Lambda} \left(\nu_{1j}^*(Z_2, \alpha_1, m) - \nu_{2j}^*(Z_1, \alpha_2, m)\right) \geq \frac{\epsilon}{m}\right\}\right).
\end{aligned}
$$

Continuing the equality gives that the RHS equals

$$
\sum_{j=0}^{m-1} P\left(\left\{Z^{2n} : \sup_{\alpha_1, \; \alpha_2 \in \Lambda} \left(\left(\nu_{1j}^*(T_i Z_2, \alpha_1, m) - \nu_{2j}^*(T_i Z_1, \alpha_2, m)\right)\right) \geq \frac{\epsilon}{m}\right\}\right) =
$$

$$
\sum_{j=0}^{m-1} P\left(\left\{Z^{2n} : \left(\nu_{1j}^*(T_i Z_2, \alpha_{1j}^*, m) - \nu_{2j}^*(T_i Z_1, \alpha_{2j}^*, m)\right) \geq \frac{\epsilon}{m}\right\}\right)
$$

$$
\begin{aligned}
&= \quad \sum_{j=0}^{m-1} P\left(\left\{Z^{2n} : \Delta_j^*\left(T_i Z, \alpha_j^*, m\right) \geq \frac{\epsilon}{m}\right\}\right) \\
&= \quad \frac{1}{(2n)!} \sum_{j=0}^{m-1} \sum_{i=1}^{(2n)!} P\left(\left\{Z^{2n} : \Delta_j^*\left(T_i Z, \alpha_j^*, m\right) \geq \frac{\epsilon}{m}\right\}\right) \\
&= \quad \frac{1}{(2n)!} \sum_{j=0}^{m-1} \sum_{i=1}^{(2n)!} \int I_{\left\{Z^{2n} : \Delta_j^*\left(T_i Z, \alpha_j^*, m\right) \geq \frac{\epsilon}{m}\right\}}\left(z^{2n}\right) dP(Z^{2n}). \quad (2.29)
\end{aligned}
$$

Let the equivalence classes in Λ under $\sim$ be denoted Λ_k. Then, the equivalence classes $[\alpha_{jk}^*]$ for the α_{jk}^*'s provide a partition for Λ. That is, $\Lambda = \bigcup_{k=1}^{N_j^\Lambda(z^{2n})} [\alpha_{jk}^*]$ because $\alpha_{jk}^* \in \Lambda_k$ and hence $[\alpha_{jk}^*] = \Lambda_k$. In addition, α_{jk}^* is the maximal value of α_{1j} and α_{2j} in the kth equivalence class. So,

$$
\begin{aligned}
I_{\left\{Z^{2n} : \Delta\left(T_i Z, \alpha_j^*, m\right) \geq \epsilon\right\}}\left(z^{2n}\right) \quad &\leq \quad I_{\left\{Z^{2n} : \Delta_j^*\left(T_i Z, \alpha_{1j}^*, m\right) \geq \frac{\epsilon}{m}\right\}}\left(z^{2n}\right) \\
&\quad + \quad \cdots + I_{\left\{Z^{2n} : \Delta_j^*\left(T_i Z, \alpha_{N_j^\Lambda(z^{2n})j}^*, m\right) \geq \frac{\epsilon}{m}\right\}}\left(z^{2n}\right) \\
&= \quad \sum_{k=1}^{N_j^\Lambda(z^{2n})} I_{\left\{Z^{2n} : \Delta_j^*\left(T_i Z, \alpha_{kj}^*, m\right) \geq \frac{\epsilon}{m}\right\}}\left(z^{2n}\right) \quad (2.30)
\end{aligned}
$$

where

$$
\begin{aligned}
A_{\epsilon,m,j,k} \quad &= \quad \left\{Z^{2n} : \Delta_j^*\left(T_i Z, \alpha_{kj}^*, m\right) \geq \frac{\epsilon}{m}\right\} \\
&= \quad \left\{Z^{2n} : \sup_{\alpha_{1j}, \alpha_{2j} \in \Lambda_k} \left(\nu_{1j}^*(T_i Z_2, \alpha_{1j}, m) - \nu_{2j}^*(T_i Z_1, \alpha_{2j}, m)\right) \geq \frac{\epsilon}{m}\right\}
\end{aligned}
$$

Now, using (2.30), (2.29) is bounded by

$$P\left(A_{\epsilon,m}\right) \leq \frac{1}{(2n)!} \sum_{j=0}^{m-1} \sum_{i=1}^{(2n)!} \int \sum_{k=1}^{N_j^\Lambda(z^{2n})} I_{\left\{Z^{2n}:\Delta_j^*(T_iZ,\alpha_{kj}^*,m)\geq\frac{\epsilon}{m}\right\}}\left(z^{2n}\right) dP\left(z^{2n}\right)$$

$$= \int \sum_{j=0}^{m-1} \sum_{k=1}^{N_j^\Lambda(z^{2n})} \left[\frac{1}{(2n)!} \sum_{i=1}^{(2n)!} I_{\left\{Z^{2n}:\Delta_j^*(T_iZ,\alpha_{kj}^*,m)\geq\frac{\epsilon}{m}\right\}}\left(z^{2n}\right)\right] dP\left(z^{2n}\right). \quad (2.31)$$

The expression in square brackets in (2.31) is the fraction of the number of the $(2n)!$ permutations T_i of Z^{2n} for which $A_{\epsilon,m,j,k}$ is closed under T_i for any fixed equivalence class Λ_k. It is equal to

$$\Gamma_j = \sum_k \frac{\binom{m_j^*}{k}\binom{2n-m_j^*}{m_j^*-k}}{\binom{2n}{n}}$$

where

$$\begin{cases} \left\{k : \left|\frac{k}{n} - \frac{m_j^*-k}{n}\right| \geq \frac{\epsilon}{m}\right\} & ; \\ m_j^* = n_{1j}^* + n_{2j}^* & . \end{cases}$$

Here, Γ_j is the probability of choosing exactly k sample data points whose losses fall in interval I_j respectively in the first and second half of the sample such that $A_{\epsilon,m,j,k}$ holds. m_j^* is the number of data points from the first and the second half of the sample whose loss land inside interval j. Using Theorem 2.1.1 Clause II, we have $\Gamma_j \leq 2\exp\left(-\frac{n\epsilon^2}{m^2}\right)$. So, using this in 2.31 gives $P\left(A_{\epsilon,m}\right)$ is upper bounded by

$$\int \sum_{j=0}^{m-1} \sum_{k=1}^{N_j^\Lambda(z^{2n})} 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) dP(z^{2n}) = 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \int \sum_{j=0}^{m-1} \sum_{k=1}^{N_j^\Lambda(z^{2n})} dP(z^{2n})$$

$$= 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \sum_{j=0}^{m-1} \int \sum_{k=1}^{N_j^\Lambda(z^{2n})} dP(z^{2n}) = 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \sum_{j=0}^{m-1} \int N_j^\Lambda(z^{2n}) dP(z^{2n})$$

$$= 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \sum_{j=0}^{m-1} E\left(N_j^\Lambda(z^{2n})\right). \quad (2.32)$$

Since Theorem 2.1.1 Clause I gives

$$H_{ann}(Z^{2n}) = \ln\left(E\left(N_j^\Lambda(Z^{2n})\right)\right) \leq G(2n) \leq h\ln\left(\frac{2ne}{h}\right) \Rightarrow E\left(N_j^\Lambda(z^{2n})\right) \leq \left(\frac{2ne}{h}\right)^h.$$

Using this m times in (2.32) gives the Proposition. $\qquad\square$

Because changes mentioned in the computation of the EMDBTEL do not affect Propositions 2.1.3, 2.1.4 and Theorems 2.1.2, 2.1.3, and 2.1.4, we restate them here for convenience.

Proposition 2.2.2.   *1. With probability $1 - \eta$, the inequality*

$$R\left(\alpha_k\right) \leq R_{emp}\left(\alpha_k\right) + m\sqrt{\frac{1}{n}\log\left(\left(\frac{2m}{\eta}\right)\left(\frac{2ne}{h}\right)^h\right)} \qquad (2.33)$$

holds simultaneously for all K functions in the set $\{Q\left(z, \alpha_k\right), k = 1, 2, \cdots, K\}$.

2. With probability $1 - \eta$, the inequality

$$R\left(\alpha_k\right) \leq R_{emp}\left(\alpha_k\right) + \frac{m^2}{2n}\log\left(\frac{2m}{\eta}\left(\frac{2ne}{h}\right)^h\right)\left(1 + \sqrt{1 + \frac{4nR_{emp}\left(\alpha_k\right)}{m^2\log\left(\frac{2m}{\eta}\left(\frac{2ne}{h}\right)^h\right)}}\right) \qquad (2.34)$$

holds simultaneously for all K functions in the set $\{Q\left(z, \alpha_k\right), k = 1, 2, \cdots, K\}$.

Theorem 2.2.1.   *1. If $h < \infty$, we have*

$$E\left(\sup_{\alpha_1, \alpha_2 \in \Lambda}\left|\nu_1^*(z^2, \alpha_1) - \nu_2^*(z^1, \alpha_2)\right|\right) \leq m\sqrt{\frac{1}{n}\ln\left(2m^3\left(\frac{2ne}{h}\right)^h\right)} + \frac{1}{m\sqrt{n\ln\left(2m^3\left(\frac{2ne}{h}\right)^h\right)}}$$
$$(2.35)$$

2. If $h \leq \infty$, and

$$D_p\left(\alpha\right) = \int_0^\infty \sqrt[p]{P\left\{Q\left(z, \alpha\right) \geq c\right\}}dc \leq \infty$$

where $1 < p \leq 2$ is some fixed parameter, we have

$$E\left(\sup_{\alpha_1, \alpha_2 \in \Lambda}\left|\nu_1(Z_2, \alpha_1) - \nu_2(Z_1, \alpha_2)\right|\right) \leq \frac{D_p(\alpha^*)2^{2.5 + \frac{1}{p}}\sqrt{h\ln\left(\frac{ne}{h}\right)}}{n^{1 - \frac{1}{p}}} + \frac{16D_p(\alpha^*)2^{2.5 + \frac{1}{p}}}{n^{1 - \frac{1}{p}}\sqrt{h\ln\left(\frac{ne}{h}\right)}}. \quad (2.36)$$

3. Assume that $h \to \infty$, $\frac{n}{h} \to \infty$, $m \to \infty$, $\ln\left(m\right) = o(n)$, and

$$D_p\left(\alpha\right) = \int_0^\infty \sqrt[p]{P\left\{Q(z, \alpha) \geq c\right\}}dc \leq \infty$$

where $p = 2$. Then we have that

$$E\left(\sup_{\alpha_1, \alpha_2 \in \Lambda}\left|\nu_1(Z_2, \alpha_1) - \nu_2(Z_1, \alpha_2)\right|\right) \leq \min\left(1, 8D_p(\alpha^*)\right)\sqrt{\frac{h}{n}\ln\left(\frac{2ne}{h}\right)}. \qquad (2.37)$$

The expression in the RHS of inequality (2.37) is a decreasing function of h for fixed n and an increasing function of n for fixed h. This observation is confirmed in figure (2.1) Now that we have all the results that we want, we can estimate the VCD and use it in model selection. We show how to do this for simulated and real dataset in Chap. 3.
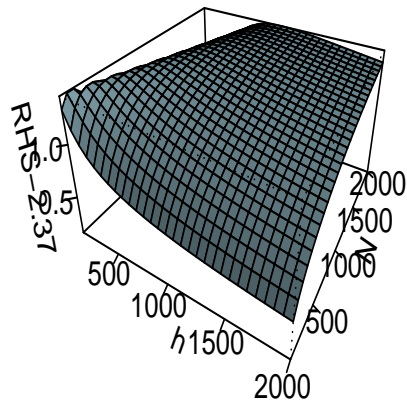
Figure 2.1: Perspective Plot of the RHS of equation (2.37)

Chapter 3

SIMULATION STUDIES

Our goal here is to implement the theory developed in Chapter 2. We start by giving a definition of what we call 'consistency at the true model' in Sec. 3.1. In Sec. 3.2, we implement Theorem 2.1.4 computationally. We develop this implementation in the simulation setting of Subsec. 3.2.1. The model class we use for simulated data is the class of linear models. In Subsec. 3.2.2, we begin by using the original algorithm of Vapnik et al., Vapnik et al. (1994) to estimate the LHS of Theorem 2.1.4 and see that often it does not give satisfactory results. In particular we will see that the estimates are not accurate. We also see a sense in which Vapnik et al., Vapnik et al. (1994) original algorithm –Algorithm 1 in Chap. 1– might not be too bad.

We can do better.

Specifically, we suggest that the inaccuracy in the estimate of VCD using Vapnik's original algorithm 1 arises from insufficient variation in the computation of the Maximum Difference Between Two Empirical Losses and from the fact that k, a, b in equation (1.6) were chosen in sub-optimal ways. To correct for these, we use a data-driven estimate for a constant in the RHS of (2.21) from Theorem 2.1.4. We also use a cross-validated form of the error, and instead of taking the maximum difference, we take the mean difference between two empirical losses. This is done in Secs. 3.2 and 3.3. In Sec. 3.4, we present cases where our method from Sec. 3.3 works properly. In Subsec. 3.4.3, we look at cases where our method does not work and show how tuning some parameters dramatically improves our estimates.

3.1 Definition of 'consistency at the true model'

Definition 3.1.1. We define 'consistency at the true model' to be

1. consistency at the true model in the usual sense of the term i.e $\hat{h} \to h_T$ in probability coupled with

2. some kind of identifiable bad behaviour away from the true model, usually getting worse as the wrong model moves further from the true model.

We argue that the 2nd clause is important because if we are in the unfortunate case of having a wrong model, clause 2 can help us move toward a better model. Moreover, since in practice no models are

correct to infinite precision, clause 2 is essential for good modeling. Our proposed methodology for using estimate $\hat{h}$ of h_T is based on this definition. We begin by assuming that there is a true model. The model need not be true to infinite precision, but hopefully it is plausible enough that when we try to discredit it, it will not in fact be discredited.

For any model, we can evaluate the LHS of equation (2.21) from Theorem 2.1.4 by Algorithm 2 below. Then, we can use nonlinear regression in (3.5) to get $\hat{h}$. So, it is seen that $\hat{h}$ is a function of the conjectured model. In principle, for any given model class that we might take as true, the VCD can be found. This is particularly easy for linear models. − Theorem 1.2.1 shows that the VCD for linear models is just the number of parameters for a saturated model.

Since our goal is to estimate the true VCD, when a conjectured model $P\left(\cdot \mid \beta\right)$ is linear, we expect $VCD\left(P\left(\cdot \mid \beta\right)\right) \cong \hat{h}$. By the same logic, if $P\left(\cdot \mid \beta\right)$ is far from the true model, we expect $VCD\left(P\left(\cdot \mid \beta\right)\right) \gg \hat{h}$ or $VCD\left(P\left(\cdot \mid \beta\right)\right) \ll \hat{h}$. That is, there are conditions that would satisfy clause 2 of our definition of 'consistency at the true model'. Taken together, clause 1 and two of 3.1.1 suggest that we estimate h_T by seeking

$$\hat{h} = \arg \min_{k} \left| VCD\left(P_k\left(\cdot \mid \beta\right)\right) - \hat{h}_k \right| \tag{3.1}$$

where $\{P_k\left(\cdot \mid \beta\right) | k = 1, 2, \cdots, K\}$ is some set of models and $\hat{h}_k$ is calculated using model k. In the case of linear model, with $q = 1, 2, \cdots, Q$ explanatory variables, we get

$$\hat{h} = \arg \min_{q} \left| q - \hat{h}_q \right| \tag{3.2}$$

where $\hat{h}_q$ is the estimated VCD for model of size q. Note that (3.1) can identify a good model even when consistency at the true model fails. The reason is that (3.1) only requires a minimum at the VCD not convergence to the true VCD. That is, (3.1) is a slightly weaker condition than consistency at the true model in that clause 1 need not be satisfied. Indeed, clause 1 can be regarded as what we what, while (3.1) is what we can find. Both criteria narrow the search for the true model.

It would be seen in the simulation that, often, clause 2 is not perfectly satisfied. For instance there may be two values of h which achieve (3.1) at least approximately. For instance if the true value is $p = 30$, there may be a larger value p (see the four panels in Fig. 3.11). In these cases, usually the smallest value of $\hat{h}$ that approximately achieves the $\arg \min$ is usually the best choice. This means that from a strictly pragmatic standpoint, imposing some level of parsimony on the model selection is helpful − whence the introduction of a threshold in (3.1) or (3.2) as used in (3.6). For ease of exposition, we usually drop the explicit use of the threshold t. However, in practice, we must remember that taking a precise $\arg \min$ is sub-optimal compared to allowing a little flexibility via t.

3.2 Implementation of Theorem 2.1.4

Before presenting our simulations, let's first re-examine Theorem 2.1.4. In fact, Theorem 2.1.4 has two parts. The Left-Hand Side (LHS) that represents the Expected Maximum Difference Between Two Empirical Losses (EMDBTEL) and the Right-hand side (RHS) that represents the upper bound of the EMDBTEL. The connection between these two quantities is the VCD that we would like to estimate. The VCD appears explicitly in the expression of the RHS of (2.21) from Theorem 2.1.4, whereas in the LHS of (2.21) from Theorem 2.1.4, the VCD is implicit in the computation. More details on the calculation of the LHS of (2.21) from Theorem 2.1.4 are given in Sect 3.3.

Our objective is to find the value of VCD that will minimize the squared distance between the RHS and the LHS of (2.21) of Theorem 2.1.4.

After many simulations, we observed that the RHS bound of (2.21) from Theorem 2.1.4 was neither tight nor stable. This finding was the same as McDonald et al. (2011). So, to make the bound tighter, instead of taking $\min\left(1, D_p(\alpha^*)\right)$ (the constant factor at the RHS of (2.21) from Theorem 2.1.4), we replace it by c, and use our data to find a value of $\hat{c}$ that makes the upper bound as tight as possible. This approach is different from Vapnik et al., Vapnik et al. (1994) who asserted (without justification) that the constants in the objective function were universal.

Our estimation of $\hat{h}$ has three steps. The first step is to estimate the Expected Maximum Difference Between Two Empirical Losses (EMDBTEL) i.e the RHS of (2.21) in Theorem 2.1.4. This step requires that we define the model class and hence its complexity in VCD. We also need to choose the number of design points ($N_L = n_1, n_2, \cdots, n_l$). Design points represent different sample sizes that will be used to compute each value of the EMDBTEL use in the non-linear regression see (3.5). They are chosen so they will cover $[0, n]$ roughly uniformly, where n is the sample size. Thus, we have used a uniform grid, and this has worked well in our examples. However, we note that Shao et al. Shao et al. (2000) proposed a design procedure for specifying the N_L's and the number of repeated experiments at a given sample size. We have not assessed how well their recommendations perform; we leave this as future work. Also, in this first step, we use the number m of disjoint intervals as defined in Chap. 2 to convert our regression problem into m classification problems.

The second step is the calculation of c. This step is very important because a good value of c will make the bound tighter and the probability of getting a good estimate $\hat{h}$ will be higher. To find c, we define a relatively fine grid of values, e.g. let c range from 0.1 to 100, in steps of size 0.01. For each value of c, we compute the squared distance between the RHS and the LHS of Theorem 2.1.4, and pick the $\hat{c}$ that gives the minimum squared distance. This is possible because we have a predefined model complexity.

The third step is to find $\hat{h}$. Once we have calculated the LHS of Theorem 2.1.4 and we have the value

of $\hat{c}$, the only unknown value in Theorem 2.1.4 is h. To estimate the assumed true value h_T of $\hat{h}$, we find the value h that minimizes the squared distance between the RHS and the LHS of Theorem 2.1.4 by non-linear regression.

The point of the rest of this section is to argue that the algorithm proposed by Vapnik et al. Vapnik et al. (1994), i.e Algorithm 1, to estimate the LHS of (2.21) from Theorem 2.1.4 and thereby estimate the VCD can behave poorly even on simulated data. In fact, as we said before, to estimate the VCD we first need to find a good way to estimate the LHS of inequality (2.21) in Theorem 2.1.4. For our first set of simulations, using the computational procedure of Vapnik et al. Vapnik et al. (1994) i.e Algorithm 1 EMDBTEL in the regression setting gives minimally adequate results that we can improve. In particular Algorithm 1 does not show variability in the estimate of VCD across a wide range of models.

3.2.1 Simulation Settings

Here, our simulations are based on linear models, since for these, we know the VCD equals the number of parameters in the model. To establish notation, we write the regression function as a linear combination of the covariates X_j, $j = 0, 1, \cdots, p$,

$$y = f(x, \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \sum_{j=0}^{p} \beta_j x_j.$$

Given a dataset, $\{(x_i, Y_i, \ i = 1, 2, \cdots, n)\}$, the matrix representation is

$$Y = X'_{nxp}\beta_{px1} + \epsilon_{nx1} \quad \text{where} \quad X'_{nxp} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

and $\epsilon_{n \times 1}$ is an n-dimension column vector of mean zero. Now, the least squared estimator $\widehat{\beta}$ is given by

$$\widehat{\beta} = \left(X'X \right)^{-1} X'Y.$$

Our simulated data is analogous. We write

$$Y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

where $\epsilon \sim N(0, \sigma_\epsilon = 0.4)$ $\quad x_0 = 1,$ $\quad \beta_j \sim N(\mu = 5, \sigma_\beta = 3),$ for $j = 1, 2, \cdots p$ and

$$x_j \sim N(\mu = 5, \sigma_x = 2).$$

and are all independent We center and scale all our variables including the response. Initially, we use a nested sequence of model lists. If our covariates are highly correlated, what we can do before applying our method is to de-correlated our variables by studentized or sphere the data. If our interest in only on prediction, we will not be interested in VCD, however, we can use VCD to narrow down the class of predictors we have to consider.

3.2.2 Direct Extension of the Algorithm in Vapnik et al. Vapnik et al. (1994)

The objective here is to argue that direct implementation of Theorem 2.1.4 using Algorithm 1 to estimate the LHS of (2.21) in Theorem 2.1.4 often gives poor results on simulated datasets. Later, in Sec. 4.2, we also show this for the real datasets. Indeed, we were unable to replicate the simulation results given in Vapnik et al. Vapnik et al. (1994) using their computational technique. To implement Theorem 2.1.4, we took the constant $\min\left(1, D_p\left(\alpha^*\right)\right)$ to be 1, since it does not affect the minimization over h.

We implemented simulations for two different model sizes $p = 15$ and 60. For both model sizes, we implemented two sets of models. The first set of models used only a subset of variables that was used to generate our response. For instance, out of 15 explanatory variables used to generate our response, our modeling started by using only 9 of them. We added covariates one at a time with their corresponding parameters. Once we get to 15 explanatory variables, we were at the true model. So the explanatory variables added after this were decoys and have corresponding true coefficients zero. More formally, we started with models of the form

$$y = \beta_0 + \sum_{j=1}^{l} \beta_j x_j,$$

where $l \leq p$ and p is the number of parameters used to simulate the data. When we proceeded to include decoys explanatory variables, we used models of the form

$$y = \beta_0 + \sum_{j=1}^{m} \beta_j x_j, \text{ where, } m > p, \text{ and } \forall m > p, \beta_j = 0.$$

Taken together, this gives a simple nested sequence of models in which the correct terms are added first. In practice, we suggest that, any collection of models built using the same explanatory variables can be ordered by any shrinkage method such as ALASSO Efron et al. (2004), SCAD Fan and Li (2001), elastics net Zou and Hastie (2005), etc. that satisfies the oracle property.

Thus for $p = 15$, we fitted 14 different models centered at the true model and each time we estimated the VCD. In Table 3.1a, the first model used 9 out of the 15 covariates, and we added covariates one at a time up to 15. For the second group of 8 models, we used all 15 covariates and adding decoys one at a time. In Table 3.1a, the first model in the second group has the 15 correct covariates and one decoy. Later models used more decoys. The left column represents the model size used to estimate the VCD.

The right column shows the estimate of VCD.

Model size	$\hat{h}$
9	50
10	50
11	50
12	50
13	50
14	50
15	2
16	2
17	2
18	2
19	2
20	2
21	2
22	2
23	2

(a) Estimates of $\hat{h}$, $\sigma_\epsilon = 0.4$, $N_L = 50, 100, 150, 200, 250, 300, 350$, $N = 400$, $m = 15$, $p = 15$. Using Vapnik's algorithm, here Algorithm 1

Model Size	$\hat{h}$
53	75
54	75
55	75
56	75
57	75
58	75
59	75
60	2
61	2
62	2
63	2
64	2
65	2
66	2
67	2
68	2

(b) Estimation of $\hat{h}$, $\sigma = 0.4$, $N_L = 75, 150, 225, 300, 375, 450, 525, 600$, $N = 600$, $m = 10$, $p = 60$

Table 3.1: Direct implementation of Using Vapnik's algorithm, here Algorithm 1 for $p = 15$ and 60.

From Table 3.1a, we observe that, there is no variability in the estimate of $\hat{h}$, except for the large sudden decrease at the true model $p = 15$. For the cases where the conjectured model is a subset of the true model, $\hat{h}$ is just equal to the minimum value of the design points and is far from the true value ($h_T = 15$). When adding decoy covariates in the model, the estimate that we get is the same ($\hat{h} = 2$) for all model sizes. We performed the analogous simulation for $p = 60$. Results of these simulations are contained in Table 3.1b. The observations are qualitatively the same as for $p = 15$.

From Tables 3.1a and 3.1b, we can also observe that the estimated $\hat{h}$ equals $\min N_L$ for subsets of the true model and suddenly drops to a much lower value that is different from $\min N_L$ when the conjectured model is exactly equal to the true model and is constant thereafter. Moreover, neither the large nor the small values of $\hat{h}$ are true. Thus, even though we can identify the location of the drop as the point where the true model occurs, $\hat{h}$ is not a direct estimate of the VCD, and so is not consistent at the true model. Otherwise put, $\hat{h}$ merely permits us to locate the true model on a nested list of models, contrary to the claims made by Vapnik et al.,Vapnik et al. (1994).

Overall, the implementation of Theorem 2.1.4 on synthetic data faces problems such as lack of variability and inaccuracy. The lack of variability merely makes it hard to tell in a complex setting whether the true model is more or less complicated than the conjectured model. That is, clause 2 fails. The inaccuracy is obvious in that estimates of $\hat{h}$ were far from the true number of parameters. On the positive side, we do retain the ability to identify a true model and the true value of h is within the range of values of $\hat{h}$ found by Vapnik et al.'s method. In an effort to improve the Vapnik et al. Vapnik et al. (1994) technique, we consider an alternative formulation in the next section.

3.3   An Estimator of the VCD

Recall that from Theorem 2.1.4, the upper bound is

$$\Phi_h(n) = \min\left(1, D_p(\alpha^*)\right) \sqrt{\frac{h}{n} \log\left(\frac{2ne}{h}\right)}. \tag{3.3}$$

This is slightly different from the form derived in Vapnik et al. (1994) and studied in McDonald et al. (2011). Moreover, although $\min\left(1, D_p(\alpha^*)\right)$ does not affect the optimization, it might not be the best constant for the inequality in (2.21). So, we replace it with an arbitrary constant c and optimize over it to make our upper bound as tight as possible. In contrast to Vapnik et al. Vapnik et al. (1994), our computation of c is data driven not 'universal'. We let c vary from 0.01 to 100 in steps of size 0.01. However, we have observed in practice, the best value of $\hat{c}$ is usually between 1 and 8. The technique that we use to estimate $\hat{h}$ is also different from that in Vapnik et al. Vapnik et al. (1994). Indeed, our Algorithm 2 below accurately encapsulates the way the LHS of (2.24) is formed unlike Algorithm 1. In particular, we use two large bootstrapping procedures, one as a proxy for calculating expectations and the second as a proxy for calculating a maximum. In Vapnik et al.'s algorithm (Algorithm 1), to account for the maximum variability they have to change the label of the second data. In our case we use cross-validation and bootstrapping. In fact, we split the data into two groups, using the first group of data, we fitted model I and with the second group of data we fitted model II. To estimate the error from the first model, we compared the predictions from the first model and the second group of data and

the other way around for the second model. We comment that Jun Shao et al. Shao (1993) suggested in a cross-validation context that as much as data as possible should be reserved for testing rather than fitting.

To explain how we find our estimate of the RHS of (2.21) from Theorem 2.1.4, we start by replacing the sample size n in (3.3) with a specified value of design point, so that the only unknown is h. Thus, formally, we replace (3.3) by

$$\Phi_h^*(n_l) = \widehat{c}\sqrt{\frac{h}{n_l}\log\left(\frac{2n_l e}{h}\right)},$$

where $\hat{c}$ is the optimal data driven constant. If we knew the LHS of (2.24) even computationally, we could use it to estimate h. However, in general we don't know the LHS of (2.24). Instead, we generate one observation of

$$\widehat{\xi}(n_l) = E\left(\sup_{\alpha_1,\alpha_2 \in \Lambda}(\nu_1^*(z_2,\alpha_1) - \nu_2^*(z_1,\alpha_2))\right) = \Phi_h^*(n_l) + \epsilon(n_l) \tag{3.4}$$

for each design point n_l by bootstrapping. In (3.4), we assume $\epsilon(n_l)$ has a mean zero but an otherwise unknown distribution. We can therefore obtain a list of values of $\widehat{\xi}(n_l)$ for the elements of N_L. To generate $\widehat{\xi}(n_l)$ for each n_l our algorithm uses a cross-validation error defined in step 5. We use our computation procedure for finding the LHS of (2.24) as in algorithm 2.

Note that this algorithm is parallelizable because different n_l can be sent to different nodes to speed up the process of estimating $\hat{\xi}()$ for all n_l. After obtaining $\hat{\xi}(n_l)$ for each value of n_l, we estimate h_T by minimizing the squared distance between $\hat{\xi}(n_l)$ and $\Phi(n_l)$. Our objective function is

$$f_{n_l}(h) = \sum_{l=1}^{|N_L|}\left(\hat{\xi}(n_l) - \hat{c}\sqrt{\frac{h}{n_l}\log\left(\frac{2n_l e}{h}\right)}\right)^2. \tag{3.5}$$

We comment that optimizing (3.5) does not in general lead to closed form solutions. For instance, differentiating (3.5) with respect to h yields

$$\frac{\partial f_{n_l}(h)}{\partial h} = \sum_{l=1}^{|NL|}\frac{\hat{c}\left(\log\left(\frac{2n_l e}{h}\right) - 1\right)\left(\hat{c}\sqrt{\frac{h}{n_l}\log\left(\frac{2n_l e}{h}\right)} - \hat{\xi}(n_l)\right)}{n_l\sqrt{\frac{h}{n_l}\log\left(\frac{2n_l e}{h}\right)}}$$

Setting this equal to zero only leads to an implicit expression for $\widehat{h}$ that can only be solved numerically. Doing this would give our estimate of the VCD. However, for ease of computation, we use grid search on (3.5) to estimate h, even though a quadratic optimization might be more efficient. To evaluate the performance of our $\hat{h}$, we first present some simulation studies using linear models.

Result: An estimate of a vector of values containing $\hat{\xi}(n_l)$.
Inputs:

- A collection of regression models $\mathcal{G} = \{g_\beta : \beta \in \beta\}$;

- A dataset;

- Two integers b_1 and b_2 for the number of bootstrap samples;

- Integer m for the number of disjoint intervals use to discretize the losses;

- A set of design points $N_L = \{n_1, n_2, \cdots, n_l\}$.

1. For each $g = 1, 2, \cdots, l$ do;

2. Take a bootstrap sample of size $2n_g$ (with replacement) from our dataset;

3. Randomly divide the bootstrap data into two groups G_1 and G_2 of size n_g each;

4. Fit two models one for G_1 and one for G_2;

5. The mean square error of each model is calculated using the covariate andthe response from the other model, thus: For instance

$$MSE_1 = \left(predict(Model_1, X_2) - Y_2\right)^2, \text{ and}$$

$$MSE_2 = \left(predict(Model_2, X_1) - Y_1\right)^2;$$

6. Discretize the loss function ie discretize MSE_1 and MSE_2 into m disjoint intervals;

7. Estimate $\nu_{1j}^*(Z_2, \alpha_1, m)$ and $\nu_{2j}^*(Z_1, \alpha_2, m)$ using MSE_1 and MSE_2 respectively for each interval;

8. Compute the difference $\left|\nu_{1j}^*(z_2, \alpha_1, j) - \nu_{2j}^*(z_1, \alpha_2, j)\right|$

9. Repeat steps $1 - 7$ b_1 times, take the mean intervalwise and sum it across all intervals so we have:

$$\hat{\xi}_i(n_l) = \sum_{j=0}^{m-1} mean \left|\nu_{1j}^*(z_2, \alpha_1, j) - \nu_{2j}^*(z_1, \alpha_2, j)\right|;$$

10. Repeat step $1 - 8$ b_2 times and calcuate

$$\hat{\xi}(n_l) = \frac{1}{b_2} \sum_{i=1}^{b_2} \hat{\xi}_i(n_l).$$

11. End for.

Algorithm 2: Generate $\hat{\xi}(n)$'s.

3.4 Analysis of Synthetic data

The simulations in this section are based on the setting in Sec. 3.2.1. As required for any valid estimation, the technique that we use to estimate the VCD does not require knowledge of the true number of parameters in the model. In Subsec. 3.4.1, we present simulation results to show that our estimator for VCD is consistent for the VCD of the true model. Of course, since our results are only simulations, we do not get clause 1 of consistency at the true model perfectly on all occasions: sometimes our $\hat{h}$ is off by one. (As noted later, this can often be corrected if the sample size is larger.) So we are led to believe that, with enough data our $\hat{h}$ is consistent in the sense of Def. 3.1.1. On the other hand, for the sake of good prediction, it may sometimes be good to add one or two explanatory variables beyond what a model selection or principle identifies as optimal. In Subsec. 3.4.3, we will look at simulations where results do not initially appear to be consistent with the theory. Roughly, we show that for larger values of p, larger values of n are needed. Also, as p increases, we must choose n_l's that are properly spread out over $[0, n]$. These studies help us calibrate the estimating procedure for h_T given n and a reasonable range for the number of parameters. We suggest this is necessary because (3.3) is only an upper bound (see Theorem 2.1.4) that tightens as n increases.

To recap: Our estimation procedure has three steps. First, we obtain $\hat{\xi}(n_l)$'s; The algorithm for its estimation was given in Sec. 3.3. Second step we find the value of $\hat{c}$ by minimizing (3.5). This is possible because h is known in the conjectured model. It is the complexity of the model used to estimate $\hat{\xi}(n_l)$. Third, we find the value of $\hat{h}$ that minimizes (3.5).

3.4.1 Simulation cases where our proposed method works

Simulations are implemented for model sizes $p = 15, 30, 40,$ and 50 and we present the results for all cases. For $p = 15, 30$, the choice of parameters in our simulation are all the same. When the sample size is $N = 400$; the design points are $N_L = \{50, 100, 150, 200, 250, 300, 400\}$; $m = 10$; and the number of bootstrap samples is $b_1 = b_2 = 50$. For $p = 40, 50$ and 60, the choice of the parameters in our simulation is all the same. When the sample size is $N = 600$; the design points are $N_L = (75, 150, 225, 300, 375, 450, 525, 600)$; $m = 10$; and the number of bootstrap samples is $b_1 = b_2 = 50$. For these cases, we fitted two sets of models, the first set uses a subset of our covariates to estimate the VCD, and in the second set, we added some decoys (their corresponding β in the generation of the response are zeros). Outputs of simulations are given in Figures 3.1-3.5. We emphasize that we use Def. 3.1.1 for our $\hat{h}$. $\widehat{ERM}_1$ and $\widehat{ERM}_2$ use the point where the sharpest decreases occurs, and BIC is simply minimized to identify a good model.

By examining Figures 3.1 to 3.5, we see that, for each given true model of pre-specified size, we fitted
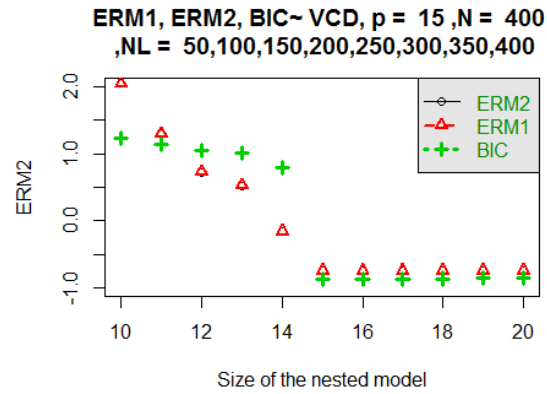
(a) Values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC
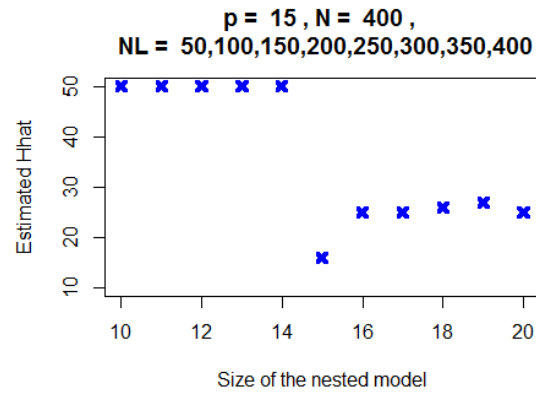
(b) Estimates of $\hat{h}$

Figure 3.1: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 15$, $\sigma_\epsilon = 0.4$, $\sigma_\beta = 3$, $\sigma_x = 2$
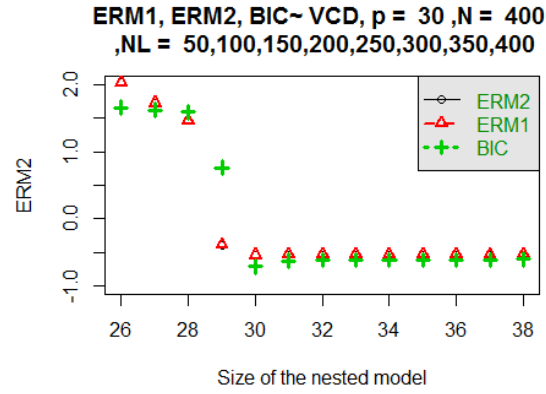
(a) Values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC
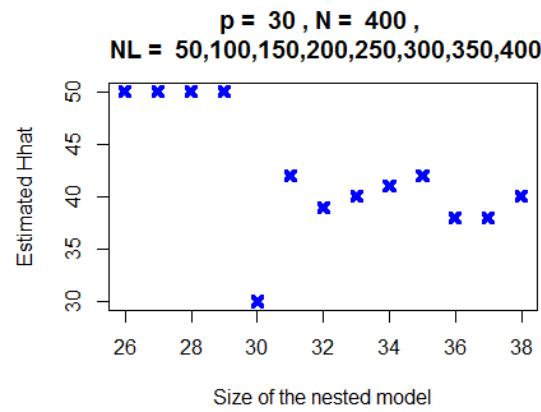
(b) Estimates of $\hat{h}$

Figure 3.2: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 30$, $\sigma_\epsilon = 0.4$, $\sigma_\beta = 3$, $\sigma_x = 2$
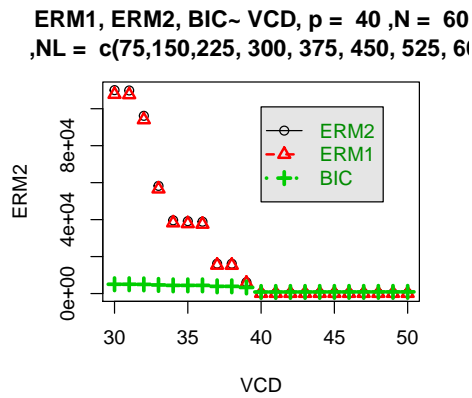
(a) Values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC
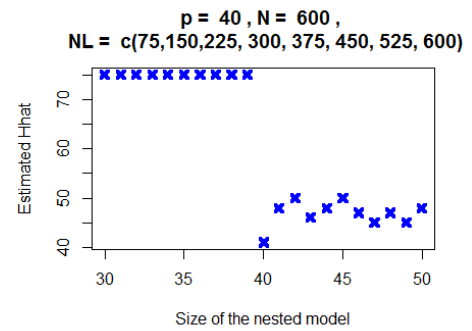
(b) Estimates of $\hat{h}$

Figure 3.3: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 40$
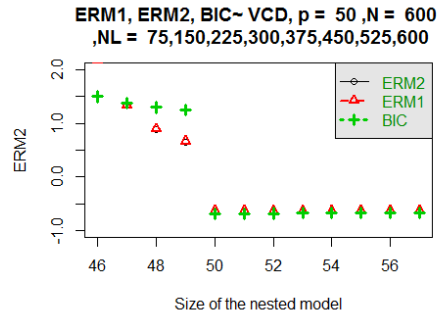
(a) Values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC
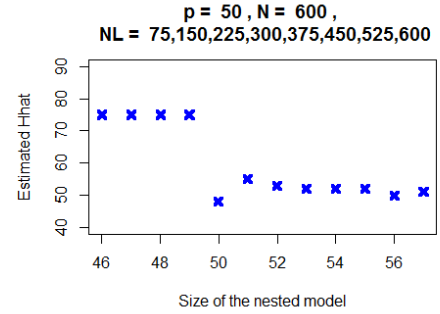
(b) Estimates of $\hat{h}$

Figure 3.4: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 50$
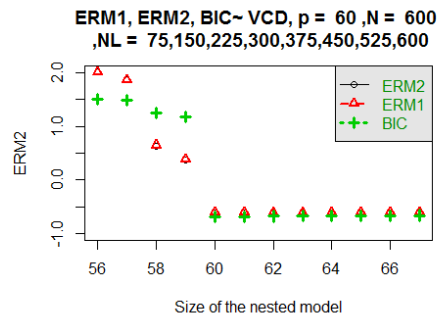
(a) Values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC
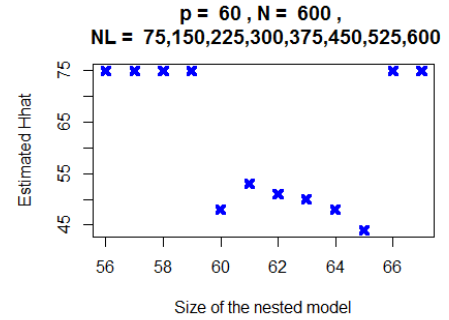
(b) Estimates of $\hat{h}$

Figure 3.5: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 60$

a list of nested models. We start with models that used a subset of covariates to generate the response, we added one variable at a time up to the correct size of the model, and after that, we added decoys. These tables also contain information on the upper bound of the true unknown risk implemented using Proposition 2.1.3 ($\widehat{ERM}_1$) and Proposition 2.1.4 ($\widehat{ERM}_2$), and also the Bayesian Information Criterion (BIC) of the corresponding model. We comment that $\widehat{ERM}_1$ and $\widehat{ERM}_2$ are the standard style of probabilistic upper bounds for the empirical risk under two senses of distance, one is additive (additive Chernoff bound Vapnik (1998)) and the other is multiplicative (multiplicative Chernoff bounds Vapnik (1998)). Note that $\widehat{ERM}_1$ and $\widehat{ERM}_2$ from Propositions 2.1.3 and 2.1.4 are of exactly the same form as would be derived using our cross-validated form of the errors namely the clause 1 and 2 from Proposition 2.2.2.

To ensure $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC will fit on the same graph, we have scaled their values. Their exact values are recorded in the tables of Appendix B.

We see that when the size of the conjectured model is strictly less than that of the true model, the estimated VCD is equal to the minimum value of the design points, and the values of the $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC are extremely high. Furthermore, these latter values typically decrease as the conjectured models become similar to the true model. For this range of model sizes, when the conjectured model exactly matches the true model, the estimated VCD ($\hat{h}$) is closest to the true value. The biggest discrepancy (of size 2) occurs for $p = 50$; by contrast, for every other case the difference between the true value and the estimated VCD is at most one.

From Figures 3.1-3.4, the behaviour of $\widehat{ERM}_1$, $\widehat{ERM}_2$ or BIC is the same. In fact, when the conjectured model is a subset of the true model, we see a consistent and substantial decrease of $\widehat{ERM}_1$, $\widehat{ERM}_2$ or BIC; and a sudden drop of these statistics when the conjectured model perfectly matches the true model. This sudden drop can be considered as an indicator of the true model. After this point, $\widehat{ERM}_1$ usually flatlines; in fact, $\widehat{ERM}_1$ does not have the ability to discriminate well over models in the sense of clause 2 of Def. 3.1.1. However, $\widehat{ERM}_2$ can still identify the true model since its smallest value occurs at the sudden drop. Also, $\widehat{ERM}_2$ often increases (albeit slowly) from its minimum as p increases, thereby often satisfying clause 2 of Def. 3.1.1 at least in a minimal sense. We also see that BIC has good power of discrimination since its smallest values occur at the true model thereafter, when we add decoys the values of BIC are bigger than those of the true model although sometimes not by much. Finally, the smallest discrepancy between the size of the model (h) and $\hat{h}$ usually occurs at the true model. This indicates that $\hat{h}$ satisfies clause 1 of Def. 3.1.1, i.e appears to be consistent for the true model. In addition, $\hat{h}$ generally increases as the size of the model becomes bigger although pass a certain point it may flatline as well. The problem with flatlining or even decreasing past a certain value of h is worse when N is not large enough relative to p.

In Figure 3.5, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC behave as before. In fact, we observe a decrease as the conjectured model becomes similar to the true model and there is a big drop as the conjectured model exactly matches the true model. However, at the true model, there is a big discrepancy between $\hat{h}$ and the size of the true model. We suggest that this discrepancy occurs because the sample size is too small comparatively to p and the choice of the design points is poor. In Sec. 3.4.3 we will observe the effect of the sample size and make tentative recommendations for how to choose design points well.

For the present, we note that Figure 3.5 gives us the estimates of $\widehat{ERM}_1$ and $\widehat{ERM}_2$ when the sample size is $N = 600$, and N_L takes on values from 75 to 600 in steps of size 75. We observe that $\widehat{ERM}_1$ and $\widehat{ERM}_2$ have a very low power of discrimination between models in the sense of clause 2 of Def. 3.1.1. However, from Figure 3.17 where the sample size is $N = 2000$ and $N_L = 500, 700, 1000, 1500, 2000$, we see that after the sudden drop in the estimate of $\widehat{ERM}_1$ and $\widehat{ERM}_2$, (that is an indicator of the true model), that $\widehat{ERM}_2$ tends to discriminate better than $\widehat{ERM}_1$.

Table 3.2: Relative increase of the sample size given the size of the model

	Fig. 3.1b	Fig. 3.2b	Fig. 3.3b	Fig. 3.4b	Fig. 3.5b
$\frac{N}{p}$	27	13	15	12	10

Table 3.2 gives the ratio of the sample size to the size of the model. In fact, from Fig 3.1b to Fig 3.5b, we see that as $\frac{N}{p}$ increases, we have a good estimate of $\hat{h}$ and clause 2 of Def. 3.1.1 seems to be satisfied better. That is, the higher $\frac{N}{p}$ is, the better clause 2 of Def. 3.1.1 is satisfied. From Fig. 3.1b to Fig. 3.3b, we see that clause 2 of Def. 3.1 is satisfied, in Fig. 3.4b, clause 2 is minimally satisfied, and in Fig. 3.5b, clause 2 is not satisfied.

We argue that estimating VCD directly is better than using $\widehat{ERM}_1$ or $\widehat{ERM}_2$. There are several reasons. First, the computation of $\widehat{ERM}_1$ and $\widehat{ERM}_2$ requires $\hat{h}$. It also requires a threshold η be chosen (see Proposition 2.2.2) and is more dependent on m than $\hat{h}$ is. Being more complicated than $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ will break down faster than $\hat{h}$. This is seen, for instance in Table B.6, where $\hat{h}$ identifies the correct model (even if it does not estimate h_T accurately due to small sample sizes) while $\widehat{ERM}_1$ fails clause 2 of Def. 3.1.1. $\widehat{ERM}_2$ behaves better than $\widehat{ERM}_1$ because the former is multiplicative (a stronger criterion) but can still give smaller values for larger model sizes, see Table B.9. In these cases when $\widehat{ERM}_1$ or $\widehat{ERM}_2$ performs poorly, $\hat{h}$ successfully identifies the true model and usually gives a value a value relatively close to the correct VCD. As a further point, Tables B.7 and B.9 show that $\widehat{ERM}_1$ often performs notably worse than $\widehat{ERM}_2$ or $\hat{h}$, and this may be attributed to the design points. More generally, we propose the following based on efficiency.

Specifically, we argue that $\widehat{ERM}_1$, and $\widehat{ERM}_2$ break down faster than $\hat{h}$ (under Def. 3.1.1) with increasing p, if the sample size is held constant. Otherwise put, $\widehat{ERM}_1$ and $\widehat{ERM}_2$ are inefficient compared to $\hat{h}$, in that to do as well as $\hat{h}$ or (BIC) for fixed p, the sample size for $\widehat{ERM}_1$ and $\widehat{ERM}_2$
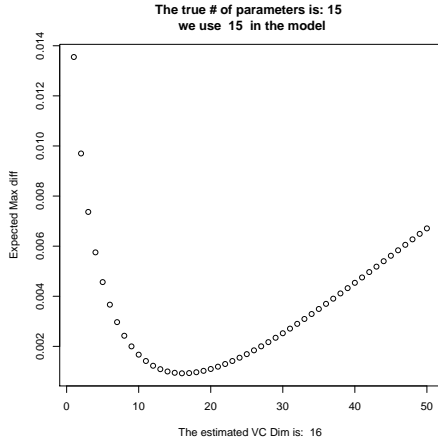
(a) The true Number of Parameters is $p = 15$, $N = 400$, $N_L = 50, 100, 150, 200, 250, 300, 350, 400$
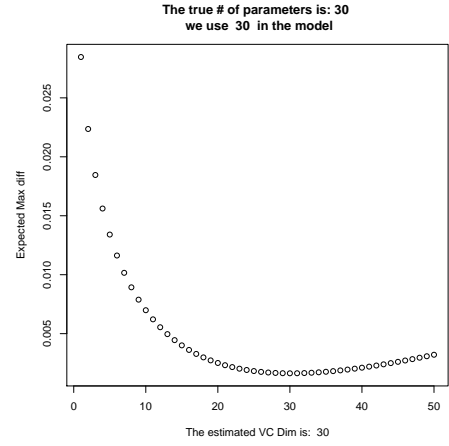
(b) The true Number of Parameters is $p = 30$, $N = 400$, $N_L = 50, 100, 150, 200, 250, 300, 350, 400$

Figure 3.6: Estimates of VCD for $p = 15, 30$

must be larger. We see this phenomenon in Tables B.1-B.4 where we let $p = 15, 30, 40$, and $p = 50$ while n ranges from 400 to 600. In these cases we see that the inefficiency makes $\widehat{ERM}_1$ and $\widehat{ERM}_2$ less variable, though they still oscillate somewhat when decoys are included. So $\widehat{ERM}_1$, and $\widehat{ERM}_2$ do not satisfy clause 2 of consistency at the true model very well in contrast to $\hat{h}$ using (3.2).

Graphs 3.8a and 3.8b from Figure 3.6-3.8 are the graphs of the objective function (2.21). On the x-axis, we have the VCD and on the y-axis we have the value of the objective function. The minimum of these graphs is the estimate $\hat{h}$.

3.4.2 Changes in simulation settings

Our goal in this subsection is to see how our estimates will behave when we change some of our simulation settings. In fact, we change the values of σ_ϵ from 0.4 to 0.8. We also change σ_x and σ_β and we estimate $\hat{h}$ when the size of the conjectured model exactly matches the true model. We repeated this 67 times and each time we only changed the seed used to simulate the response and the covariates. (We want to do this 100 times, but time did not allow this.) By doing this, we can see how our estimates will vary given different seeds.

Figs 3.9 to 3.10 give the estimate of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$, and BIC when we increase σ_ϵ form 0.4 to 0.8, the true number of parameters is $p = 30$ and $p = 15$ respectively. Figs 3.10a and 3.9a are the graphs of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC against the sizes of the conjectured model. Again, the behaviour of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC is pretty much the same of before; in fact, we observe a big drop in their values when the size of the conjecture model exactly matches the size of the true and do not discriminate well when the size of the conjectured model passes the size of true model.
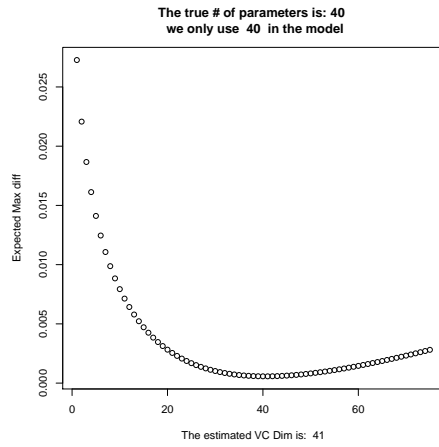
(a) The true Number of Parameters is $p = 40$, $N = 600$, $N_L = 75, 150, 225, 300, 375, 450, 525, 600$
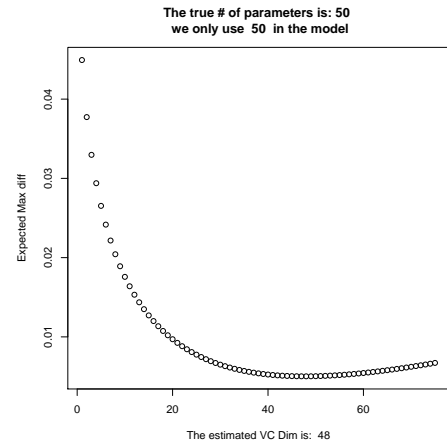
(b) The true Number of Parameters is $p = 50$, $N = 600$, $N_L = 75, 150, 225, 300, 375, 450, 525, 600$

Figure 3.7: Estimates of VCD for $p = 40, 50$
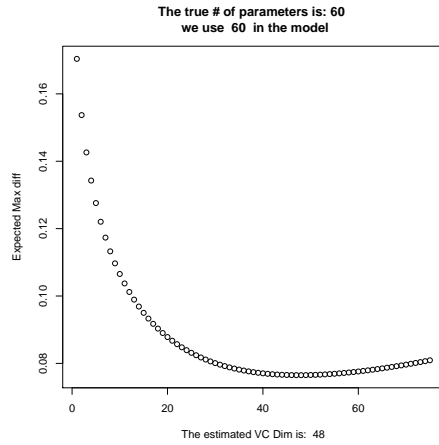
(a) The true Number of Parameters is $p = 60$, $N = 600$, $N_L = 5, 150, 225, 300, 375, 450, 525, 600$
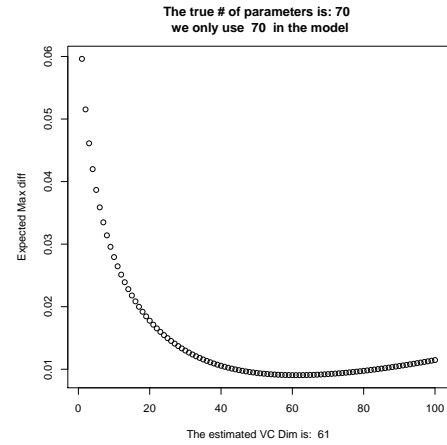
(b) The true Number of Parameters is $p = 70$, $N = 700$, $N_L = 100, 200, 300, 400, 500, 600, 700$

Figure 3.8: Estimates of VCD for $p = 60, 70$
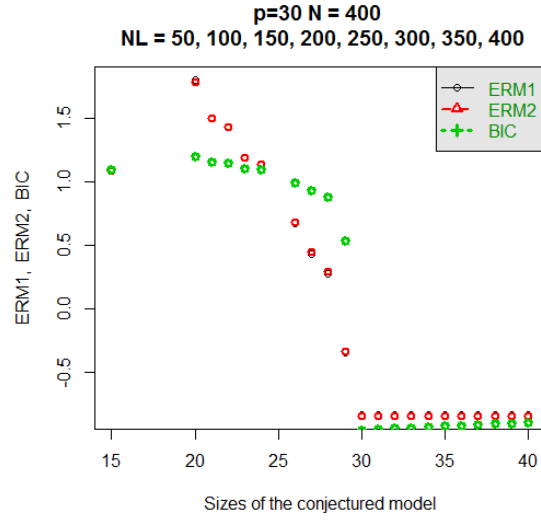
(a) The true Number of Parameters is $p = 30$, $N = 400$, $N_L = 50, 100, 150, 200, 250, 300, 350, 400$
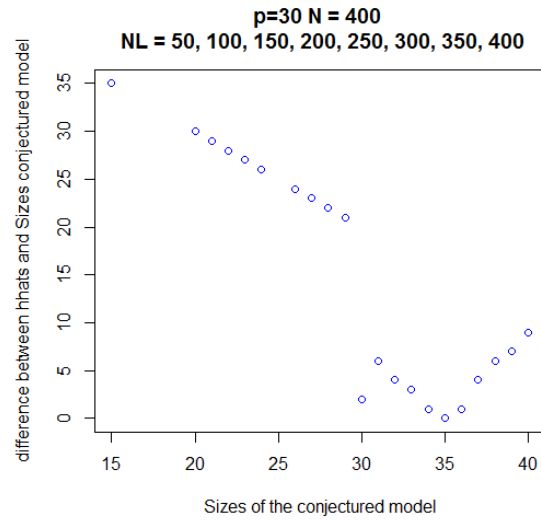
(b) The true Number of Parameters is $p = 30$, $N = 400$, $N_L = 50, 100, 150, 200, 250, 300, 350, 400$

Figure 3.9: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$, BIC for $p = 30$, $\sigma_\epsilon = 0.8$
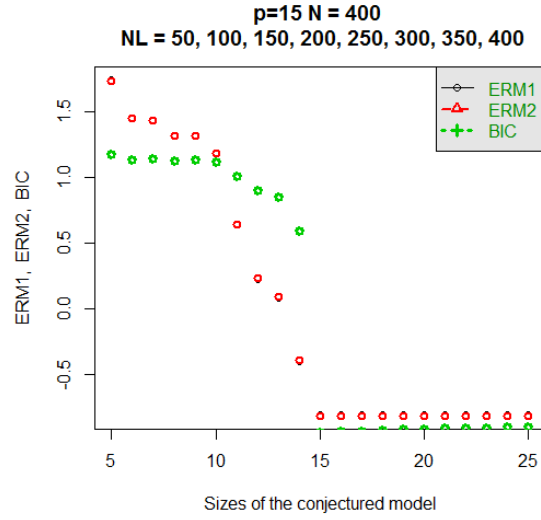
(a) The true Number of Parameters is $p = 15$, $N = 400$, $N_L = 50, 100, 150, 200, 250, 300, 350, 400$
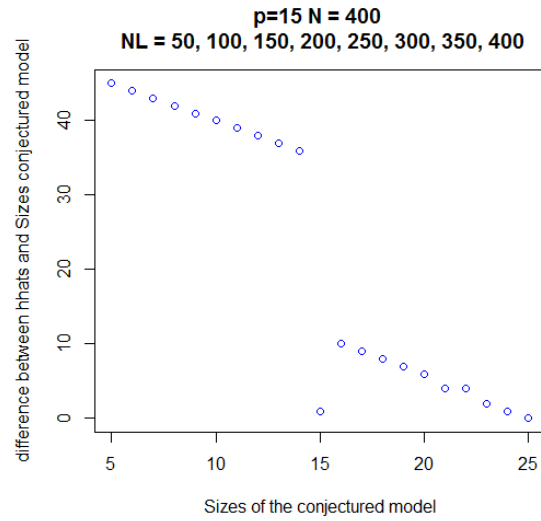
(b) The true Number of Parameters is $p = 15$, $N = 400$, $N_L = 50, 100, 150, 200, 250, 300, 350, 400$

Figure 3.10: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$, BIC for $p = 15$, $\sigma_\epsilon = 0.8$
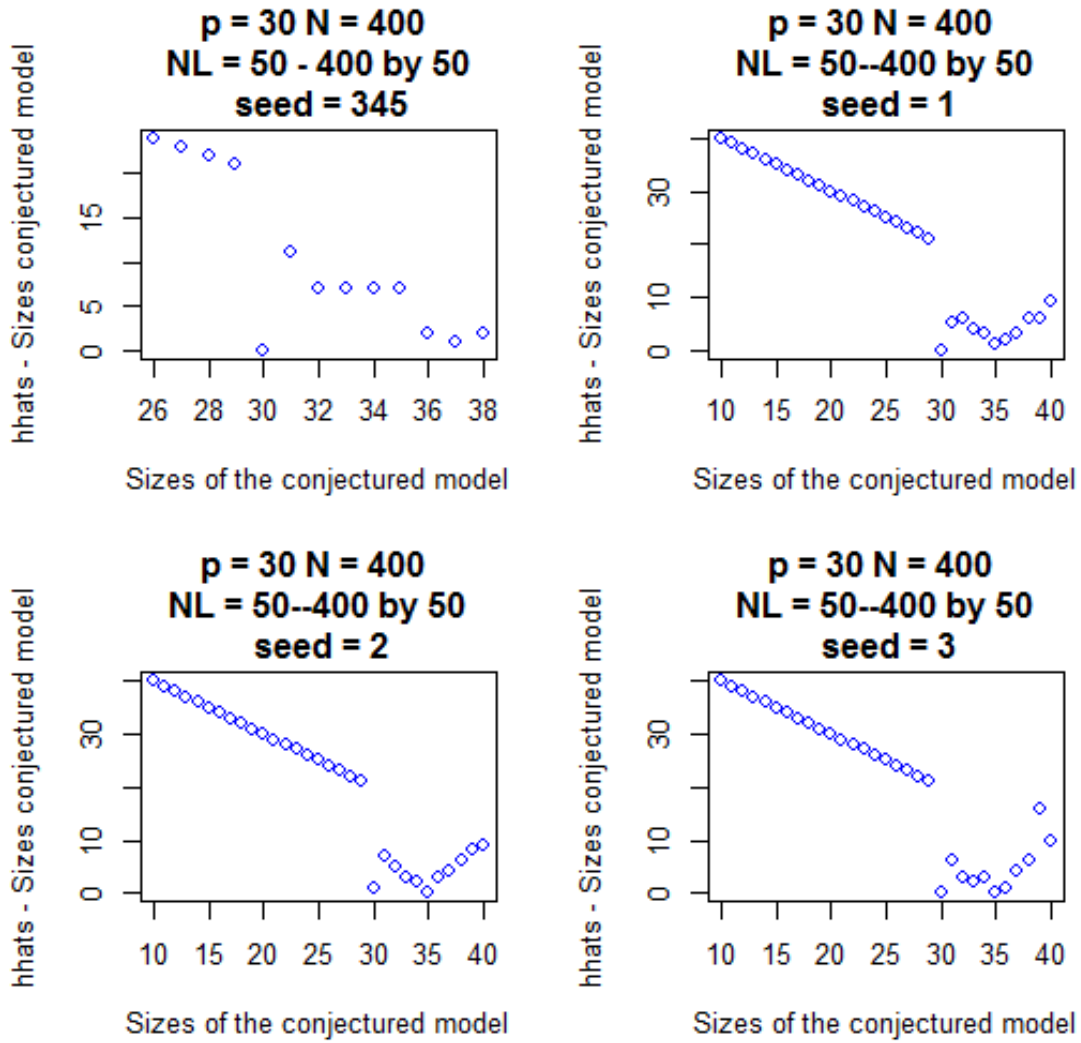
Figure 3.11: Estimates of $\hat{h}$ for $p = 30$ for different seeds

Figs 3.10b and 3.9b are graphs of the difference between $\hat{h}$ and the size of the conjectured model for $p = 15$ and $p = 30$ respectively with $\sigma_\epsilon = 0.8$. We observed that the minimum distance between the size of the conjectured model and $\hat{h}$ occurs at models of size 15 and 25 when the true $p = 15$ and a model of size 35 when the true $p = 30$. However, at the true model $\hat{h} = 16$ and $\hat{h} = 28$ respectively for $p = 15$ and $p = 30$. We comment that a parsimony argument possibly using a threshold as in (3.6) applied to Fig. 3.9b could lead to the choice $\hat{h} = 30$.

Fig 3.11 is the graph of the estimated $\hat{h}$ for $p = 30$ using different seeds in the simulate of the X's. In these case, we retain $\sigma_\epsilon = 0.4$, $\sigma_\beta = 3$ and $\sigma_X = 2$. We observe that when we change the seed, the model that we pick using $\hat{h}$ is not always the true model. For instance, when the seed is 2, the minimum discrepancy between $\hat{h}$ and the size of the conjectured model occurs when the size of the conjectured model is 35. When the seed is 3, the minimum discrepancy between $\hat{h}$ and the size of the conjectured model occurs at either 30 or 35. Again, we observe that parsimony arguments may improve our selection
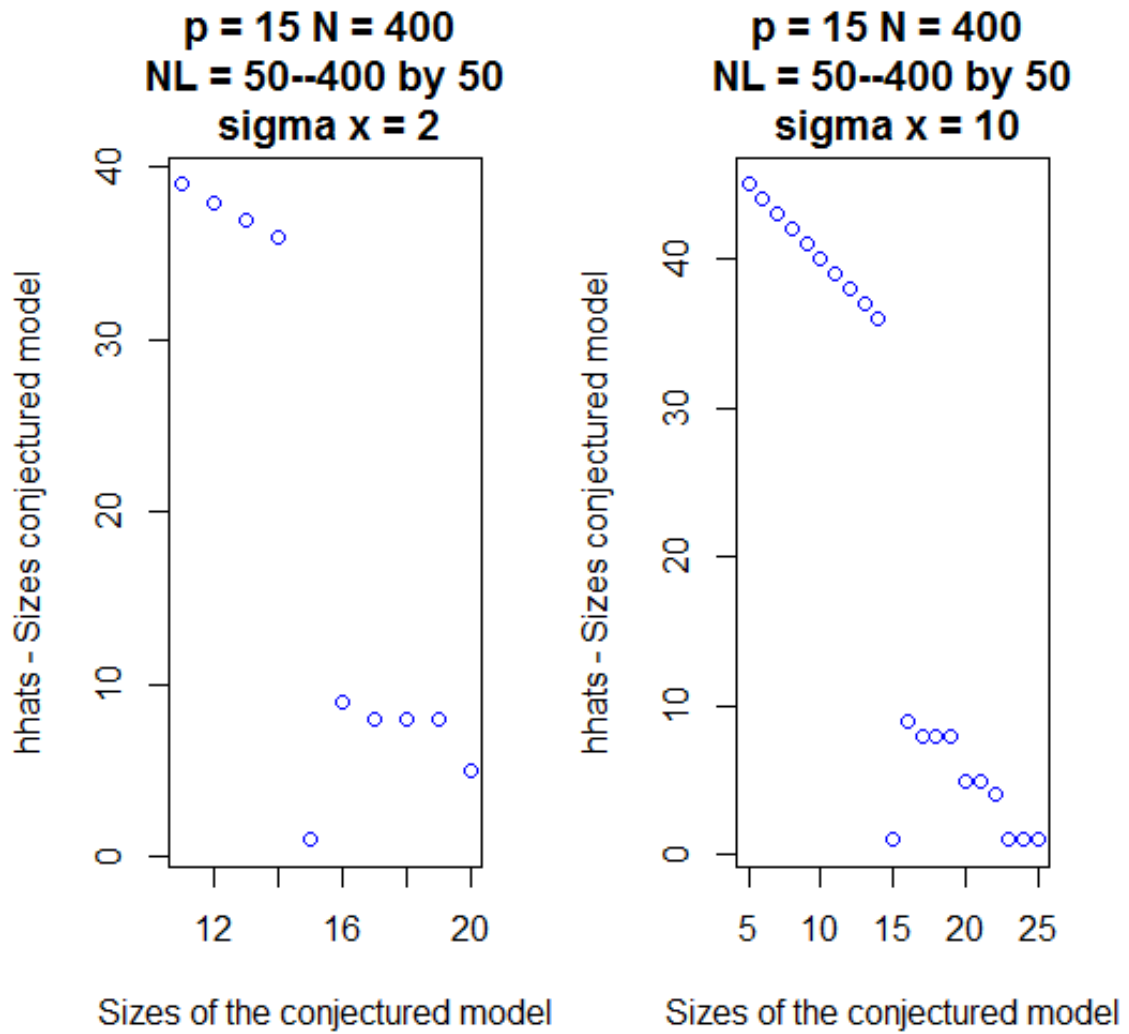
Figure 3.12: Estimates of $\hat{h}$ for $p = 15$ with different σ_x's use to simulate the covariates

of $\hat{h}$ i.e., make $\hat{h}$ reliably closer to h_T

Fig. 3.12 estimates of $\hat{h}$ when the standard deviation used to simulate the covariates increases from 2 to 10. We observe that with a small variability in our covariates our method still picks the true h. (the small discrepancy between $\hat{h}$ and the size of the conjectured model occurs at the true model.) However, we note that as σ_X increases, we may have to invoke parsimony to make $\hat{h}$ closer to h_T. Indeed, the right panel of Fig. 3.12 shows that when there is a high variability in the covariates the method does not discriminate well – the method did not discriminate amongst models of size 15, 23, 24, and 25.

Fig. 3.13 gives the estimate of $\hat{h}$ at the true model when the seed used to simulate the response and covariates varies from 1 to 67. Fig. 3.13a shows a sampling distribution for $\hat{h}$. Here, there is no change in the simulation settings, we merely computed different values of $\hat{h}$ with the seed changing from 1 to 67. We observe that most of the estimates of $\hat{h}$ are around 30. The average value of $\hat{h}$ is 29.07 with a standard deviation of 0.94. Fig. 3.13b gives the same observation, albeit with a little more variability. In
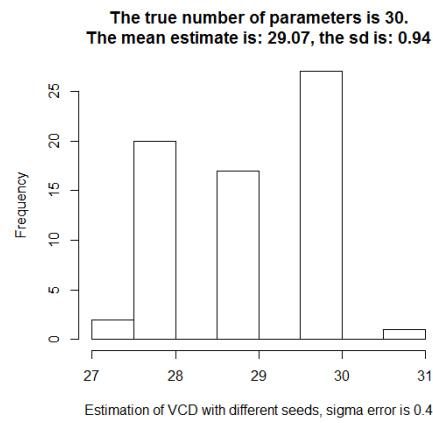
(a) Sampling distribution of $\hat{h}$, $\sigma_\epsilon = 0.4$, $\sigma_\beta = 3$, $\sigma_X = 2$.
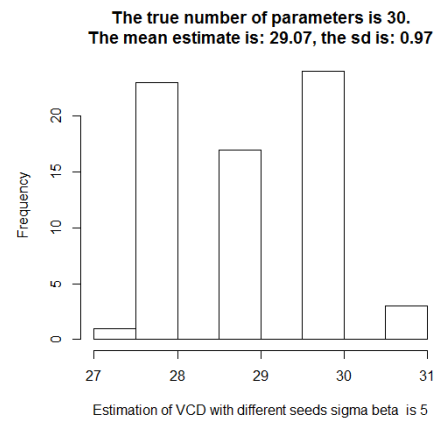
(b) Sampling distribution of $\hat{h}$, $\sigma_\epsilon = 0.4$, $\sigma_\beta = 5$, $\sigma_X = 2$.
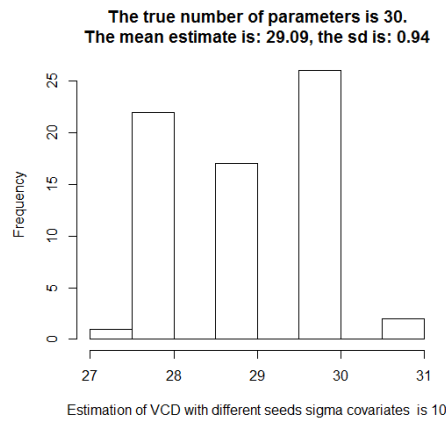
(c) Sampling distribution of $\hat{h}$, $\sigma_\epsilon = 0.4$, $\sigma_\beta = 3$, $\sigma_X = 10$.

Figure 3.13: Effect of changing the seed on $\hat{h}$ at the true model. The true number of parameters is 30
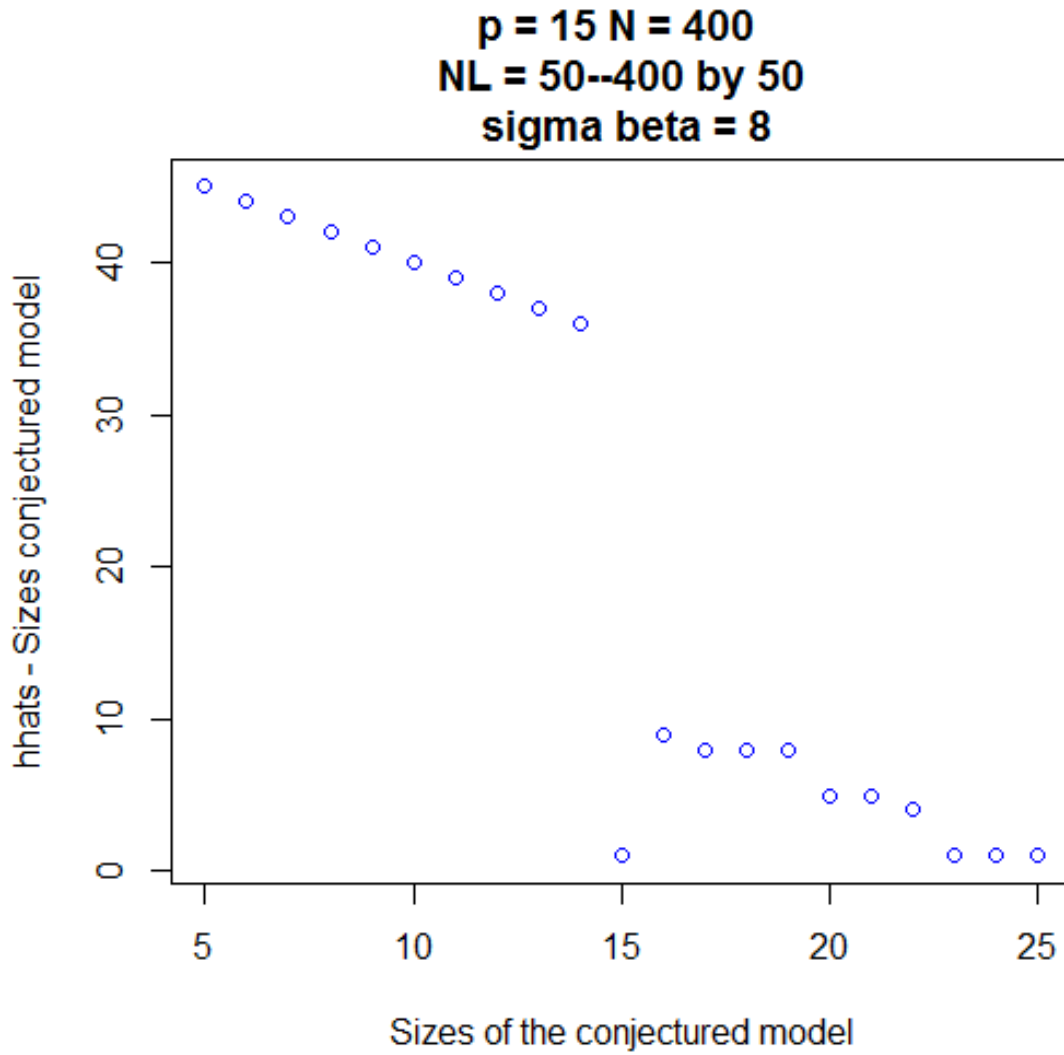
Figure 3.14: Estimates of $\hat{h}$ for $p = 15$, $\sigma_\beta = 8$

fact, the average value of $\hat{h}$ is 29.07 with a standard deviation of 0.97. Again, Fig. 3.13c is a sampling distribution for $\hat{h}$ for $\sigma_X = 10$. The analysis is as before. The average value of $\hat{h}$ is 29.09 with a standard deviation of 0.94. For the case of varying σ_X, we expect at most only slight change we have studentized the covariates.

In Fig. 3.14, we set $\sigma_\beta = 8$. We see that $\hat{h}$ does not discriminate well. In fact, $\hat{h}$ does not detect a difference between models of size 15, 23, 24, and 25. As in other settings, we suggest parsimony arguments and thresholding can improve the performance of $\hat{h}$. In the present case, simply choosing the smallest value of $\hat{h}$ (parsimony) identifies the true model.

Based on our simulation results, we suggest that in practice, $\hat{h}$ be chosen by a criterion of the form

$$\hat{h} = \arg\min_k \left\{ \left| \hat{h}_k - VCD\left(p_k(\cdot, \beta)\right) \right| \leq t \right\} \tag{3.6}$$

for some reasonable threshold t. Often can be taken to be 0, but not always. In practice, we recommend $t = 2$.

3.4.3 Dependency on The Sample Size and Design Points

Our goal here is to show how we can improve the quality of our estimate by increasing N or tuning the design points. At the beginning of this section, we started observing the effect of sample size on the quality of our estimates. Here, we emphasize both the sample size and the effect of design points. We perform simulations for model sizes $p = 60$, and 70. For $p = 70$, the sample size is set to be $N = 700$, design points vary from 100, to 700, by 100, everything else is still the same as in the previous subsection. To observe the effect of sample size and design points on $\hat{h}$, $\widehat{EMR}_1$ and $\widehat{ERM}_2$ we set the sample size to be $N = 2000$ and $N_L = 500, 700, 1000, 1500, 2000$ for either $p = 60$ or $p = 70$ and keep everything else unchanged.

Figures 3.5 and 3.15 give estimates of $\hat{h}$, the upper bound of the true unknown risk using Propositions 2.1.3 ($\widehat{ERM}_1$) and 2.1.4 ($\widehat{ERM}_2$), and the BIC as before, for small sample sizes $N = 600, 700$ and their corresponding design points. Given that the size of the conjectured model is strictly less than the size of the true model, $\hat{h}$ is equal to the smallest design point. However, when the conjectured model exactly matches the true model, $\hat{h} \approx 50, 61$ underestimates h_T. When the conjectured model is more complex than the true model, we see that $\hat{h}$ still underestimates h_T in most cases. Our observations about $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC are still the same as before.

Figure 3.16 gives estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC when $N = 700$, $N_L = 100, 200$, $300, 400, 500, 600$, and 700 and the model size is $p = 60$. We see that $\hat{h} = 57$; this estimate is closer to the true value than that from Figure 3.5. We do not observe any change in the qualitative behaviour of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC. This shows that small changes in sample size or design points may have large numerical effects on the values of $\hat{h}$, $\widehat{ERM}_1$ and $\widehat{ERM}_2$ and BIC.

Figure 3.17 is qualitatively the same as Figure 3.5. The difference is the sample size and design points. In fact, in Figure 3.17, the sample size is $N = 2000$, design points are $N_L = 500, 700, 1000, 1500, 2000$, whereas in Figure 3.5, $N = 600$, design points vary from 75 to 600 in steps of 75. The behaviour of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC are still the same as previously described. We infer from this that as the sample size increases, $\hat{h}$ moves closer to its true value. In these figures, the design points have also shifted. This leads us to suggest that to get the optimal estimate $\hat{h}$, not only must N increase, the value design points must also increase so that the range of values in the set of design points covers $[0, N]$.

When comparing Figure 3.18 to Figure 3.5, the difference is that in Figure 3.18 $N = 2000$, whereas (in Figure 3.5, $N = 600$). We observe that when there is a big enough increase in N relative to the design points (N_L being constant, but small compare to sample size) $\hat{h}$ converges to the true value, but

at a lower rate even though $\hat{h}$ only moved from 48 to 50, in both cases, the model identified is very close to the true model. Indeed, in Figure 3.5, $\hat{h}$ identifies $p = 61$ or 62 and in Figure 3.18, $\hat{h}$ identifies $p = 61$.

Figure 3.16 and Figure 3.19 are almost the same despite the substantial increase in N. In Figure 3.19, the sample size is $N = 2000$, and in Figure 3.16 the sample size is $N = 700$, design points $N_L = 100, 200, 300, 400, 500, 600, 700$ for the two figures. We observe that the qualitative behaviour of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC is unchanged and $\hat{h}$ at the true model is nearly the same for both simulations and close to the true VCD. In both cases, the model identified by $\hat{h}$ is qualitatively the same and close to the true model. $\hat{h}$ tends to increase as the size of the wrong model increases i.e clause 2 of Def. 3.1.1 is approximately satisfied. This comparison suggests that if the design points are large relative to p then $\hat{h}$ can be found accurately without necessitating large sample sizes, i.e., for well chosen design points, $\frac{N}{p} \approx 15$ will be sufficient. (Usually, the rule of thumb in parameter estimate is to set 10 datas per parameter. Here, we recommend a higher sample size because we are doing model selection as well as parameter estimation. Often, this method will not work with smaller sample sizes. However, few model selection techniques work with small sample sizes outside very restricted settings)

These observations show that, with enough data points, and by choosing design points well (usually half of the size of the data), we can obtain good estimates of $\hat{h}$. Therefore, we can do model selection using clause 1 of Def. 3.1.1 and a form of clause 2. Usually, with this adjustment, $\hat{h}$ will perform better than either $\widehat{ERM}_1$ or $\widehat{ERM}_1$. In contrast to BIC, for a given sample size, our method performs roughly as well as BIC for model selection when the design points are well chosen, but also provides an estimate of the VCD.

We leave the question of optimally choosing the design points as future work even though we have conjectured that design points should broadly cover $[0, N]$ with more design points in the upper half of the interval as N increases, i.e the size of the design points should track N. We also recommend that a set of design points be chosen so that the shape of the objective function (2.21) vs VCD is convex and has a well defined minimum.

Summary

In Subsect. 3.2.2 we looked at Vapnik's method for Theorem 2.1.4 on synthetic datasets. We observed that this did not give good results reliably. We assume that this inaccuracy on the estimate of $\hat{h}$ is due first to the lack of variability in the error criterion, second to the fact that the bound is not tight enough. To fix these problems, we used the estimator of h in Sec. 3.3 that accounts for these. In Sec. 3.4, we estimated h using our new estimator $\hat{h}$, and our estimates were closer to the true value. To understand the behaviour of our method, we introduced the concept of 'consistency at the true model'. We also observed how $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC decrease as the conjectured model moves closer to the true
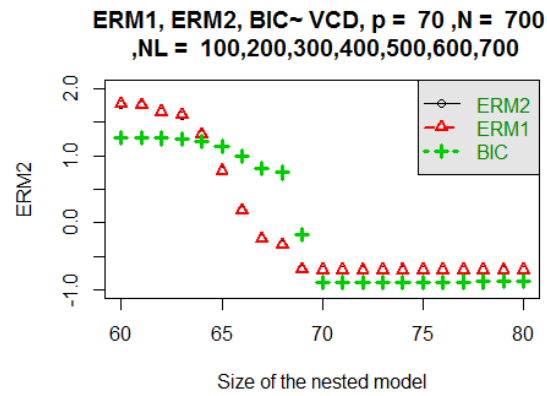
(a) Values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 70$
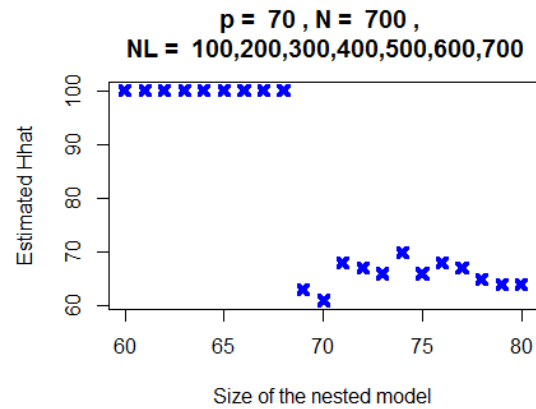
(b) Estimate of $\hat{h}$ for $p = 70$

Figure 3.15: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 70$
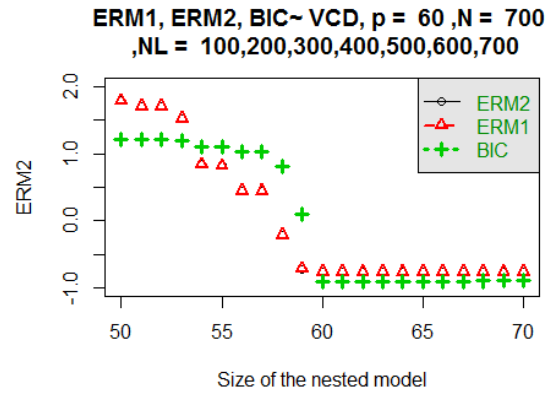
(a) Values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 60$
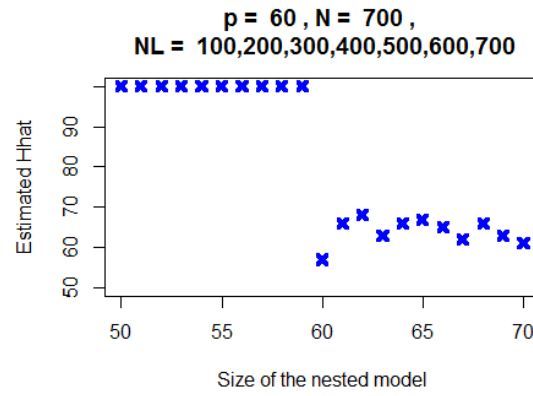
(b) Estimate of $\hat{h}$ for $p = 60$

Figure 3.16: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 60$, $\hat{h} = 57$
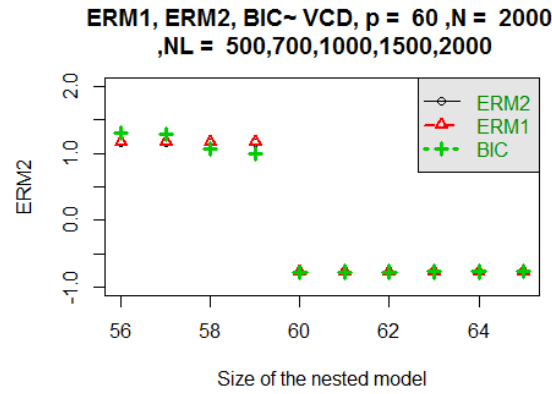
(a) Values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 60$
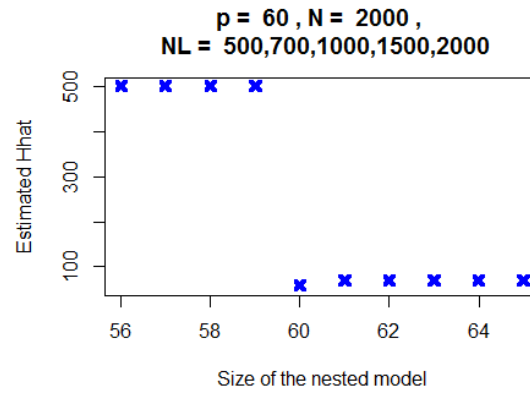
(b) Estimate of $\hat{h}$ for $p = 60$

Figure 3.17: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 60$, $\hat{h} = 59$
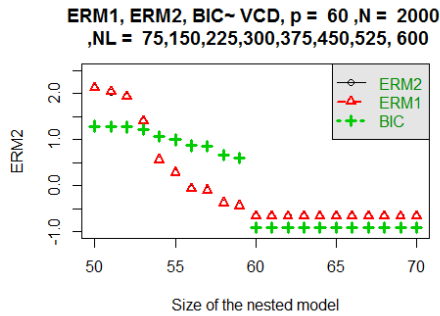
(a) Values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 60$
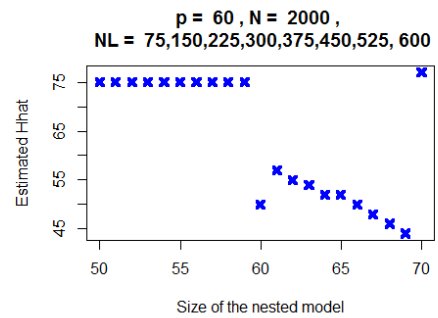
(b) Estimate of $\hat{h}$ for $p = 60$

Figure 3.18: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 60$
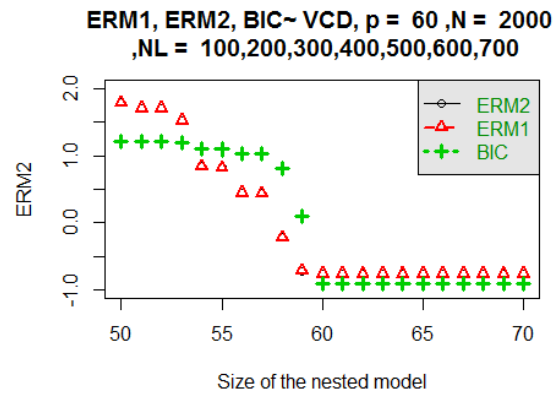
(a) Values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 60$
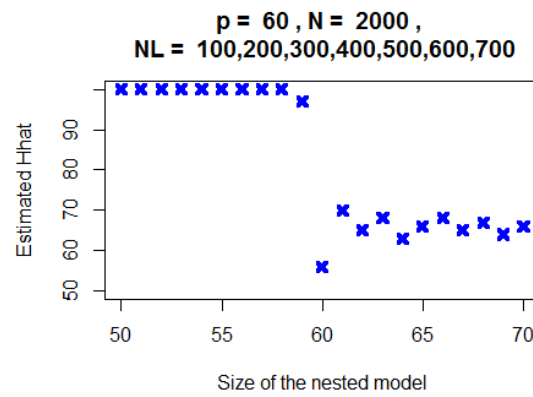
(b) Estimate of $\hat{h}$ for $p = 60$

Figure 3.19: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 60$
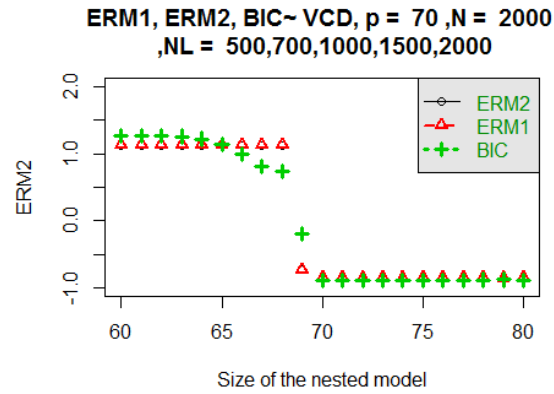
(a) Values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC
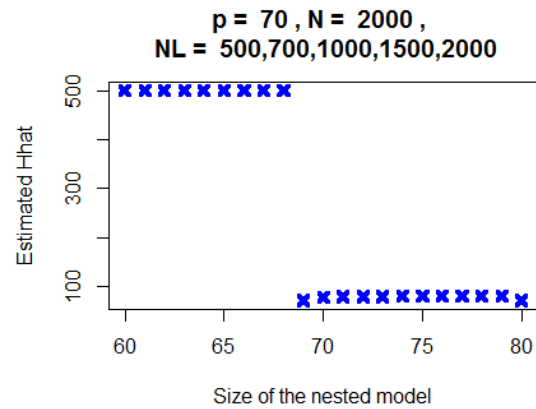
(b) Estimate of $\hat{h}$ for $p = 70$

Figure 3.20: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for $p = 70$

model and how they drop suddenly when the two models are equal. In Subsect. 3.4.3, we investigated how we can tune the design points and the sample size to improve the estimate of $\hat{h}$. We observed that, given a sample size, if we choose a good set of design points we can improve the estimation of h by $\hat{h}$. We also saw that as the sample size increases, $\hat{h}$ is consistent for the true VCD, and this may be independent of the selection of the design points. So, we suggest that the Theorem in McDonald et al. (2011) is true, but at a very slow rate and only when $c = \hat{c}$ and for a cross-validated form of the error is used. We also note that the design points must also be uniformly chosen such that its range cover the size of the dataset, and more design points are better than fewer design points.

Chapter 4

ANALYSIS OF TOUR DE FRANCE DATASET

After estimating $\hat{h}$ on synthetic datasets, the natural question is: How do we estimate $\hat{h}$ with real data and use it for model selection? We will see that the estimate of $\hat{h}$ and the upper bounds of Propositions 2.1.3 and 2.1.4 will guide our way to the choice of the best model. Since the goal of this chapter is to evaluate our method on a real dataset, we have chosen Tour De France.[1]

We start this chapter by giving some information about our dataset in Sec. 4.1. Then, in Sec. 4.2 we analyze our dataset using two classes of models with a model list based on $Year$ and $Distance$; the first class is a sequence of nested models and the second class consists of non-nested models. We evaluate our method by comparing $\hat{h}$ to BIC, $\widehat{ERM}_1$ and $\widehat{ERM}_2$. In the last Section, we look at the effect of outliers in the estimates $\hat{h}$, $\widehat{ERM}_1$ and $\widehat{ERM}_2$. In Sec. 4.3 we re-analyze the data using a model list made up of variables $Year$, $Distance$, Age of the winner and $Stages$ won. A natural question is why not combine the two model lists into one. The answer is that the resulting model would have 9 variables, but our dataset would have size 103 (including outliers) or 99 (with outliers removed). Thus, we would have enough data for parameters estimation, but not for model selection, which is the point of using h.

We admit that in the present context, the Tour De France dataset is a toy dataset, because it is small in terms of the number of observations compared to the number of quantities that we will be computing from it. However, the results seem reasonable and this chapter is important mostly to demonstrate how $\hat{h}$ can be used and compared to other methods. In Chap. 5 we will use much larger datasets and avoid problems with over-analysis.

4.1 Descriptive Analysis of Tour De France Data

The full data set has $n = 103$ dependent data points. The data points in the set are dependent because many cyclists competed in the Tour for more than one year. Indeed, one may argue that it is better analyzed as an auto-regressive time series. Here we ignore the dependence structure for the sake of comparing our method with independent data. Each data point has a value of the response variable, the average speed in kilometer per hour (km/h) of the winner (Speed) of the Tour from 1903 to 2016.

---

[1]Tour De France Data was collected by Bertrand Clarke. More information can be found at http://www.letour.fr/

(a) Speed VS Year

(b) Speed VS Distance
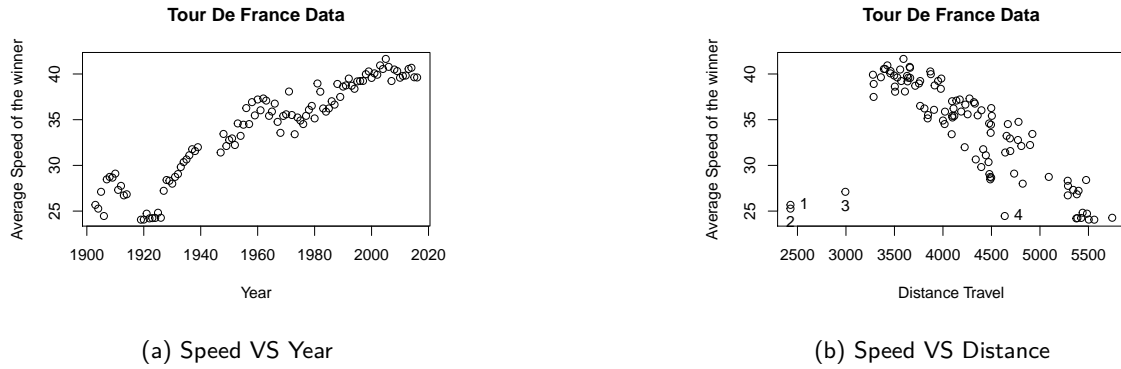
Figure 4.1: Tour De France from 1903 to 2016

However during World War 1 and 2 there was no Tour De France, so we do not have data points for those periods of time. We also see the effect of World War 1 on the speed of the winner of the tour. The lowest speeds were after World War 1. This is probably due to the death of the young men. These data points are potential outliers. After World War 2, there was also a decrease in average winning speed, but the decrease was less than that after World War 1. From Fig. 4.1a, we see that there is a curvilinear relationship between Speed and Year (Y). In Fig. 4.1b, we see that there is a linear relationship between Speed and Distance and that the variability of speed increases with the Distance (D). The four observations labeled 1-4 that are away from the bulk of the data are from the early years of the Tour and are also considered as potential outliers. The Tour De France dataset also has information on the Age of the winner (A), the number of stages won by the winner (S), and the distance (Km) (D) of the Tour De France. We use these data in a separate analysis with a different model list in Sec. 4.3.

Table 4.1: Correlation between covariates

	Speed	Year	Distance	Stages	Age
Speed	1	0.94	-0.69	-0.18	0.07
Year	0.94	1	-0.63	-0.22	0.21
Distance	-0.69	-0.63	1	0.11	-0.08
Stages	-0.18	-0.22	0.11	1	-0.20
Age	0.07	0.21	0.08	-0.20	1

Table 4.2: Correlation between Speed and D^2, Y^2 and $Y:D$

	$Distance^2$	$Year^2$	Year:Distance
Speed	-0.53	-0.24	0.02

Tables 4.1 and 4.2 give us Pearson correlation between our variables. We see that there is a strong correlation between Speed and Year (0.94). There is a negative but strong linear relationship between speed and distance (-0.69) and weak linear relationship between speed and stages won (-0.18), whereas, there is a positive and very weak relationship between speed and Age of the winner (0.07). We also notice that the correlation between speed and the squared distance is not strong (-0.53), quite weak

(a) Number of Stages won by the winner

(b) Estimates of $\hat{h}$
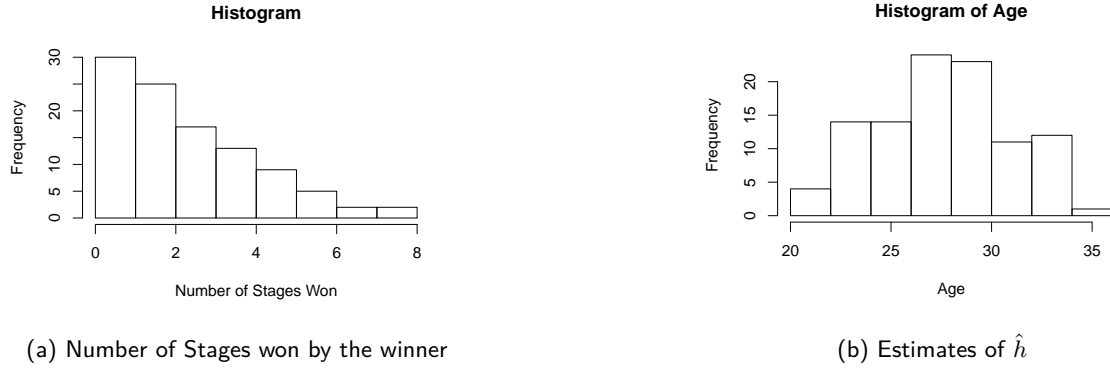
Figure 4.2: Histogram of Stages won and Age of the winners

between speed and $Year^2$ (-0.24) and almost absent between speed and the interaction between Year and distance (0.02). Tables similar to 4.1 and 4.2 can be computed using Kendall's τ for instance; the results are qualitatively the same.

From Figure 4.2, we see that most of the winners of the tour had age ranging from 25 to 30 years and the number of stages that they won ranged between 0 and 1. This suggest that the winners were the ones who went fast enough to do well but not so as to over exert themselves, becoming tired or injured as a result.

4.2 Analysis of a nested and non-nested collection of model lists of Tour De France

4.2.1 A nested model list

We identify a nested model list using Y, D, Y^2, D^2 and $Y:D$ as covariates. Because the size of the dataset is small, we can only use a small model list. We order the variables using the Smoothly Clipped Absolute Deviation (SCAD) Fan and Li (2001) shrinkage method because it perturbs parameter estimates the least and satisfies an oracle property.

Under SCADFan and Li (2001), the order of inclusion of variables is Y, D, D^2, Y^2, and $Y:D$. We therefore fit five different models.

Table 4.3: Direct Implementation of Vapnik's method for the nested models

Model	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
Y	20	316.74	495.22	419.14
$Y,\ D$	20	218.11	451.06	411.13
$Y,\ D,\ D^2$	20	176.09	317.27	365.68
$Y, D,\ D^2,\ Y^2$	20	172.75	312.89	368.24
$Y, D,\ D^2,\ Y^2,\ Y:D$	20	172.05	311.97	372.43

Table 4.4: Estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC of the nested model using our method

Model	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
Y	4	16.42	44.95	79.67
$Y,\ D$	4	15.10	42.83	71.66
$Y,\ D,\ D^2$	4	11.21	36.37	**26.21**
$Y, D,\ D^2,\ Y^2$	**4**	11.09	36.16	28.77
$Y, D,\ D^2,\ Y^2,\ Y:D$	4	**11.06**	**36.11**	32.96

The estimate of VCD requires that we choose some values to use in the analysis. Since the size of our dataset is 103, we set $m = 10$, we choose to vary N_L from 20 to 100 by 10. And, we set the number of bootstrap samples to be $b_1 = b_2 = 50$.

Tables 4.3 and 4.4 give us estimates of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC of the nested models using Algorithm 1 and Algorithm 2 respectively. It is seen that Vapnik's original method is helpful only if it is reasonable to surmise that there are exactly 15 missing variables. Our method uniquely identifies one of the models on the list. Even though there is likely no model for Tour De France dataset that is accurate to infinite precision, our method is giving a useful result.

From Table 4.3, we observe that $\hat{h} = 20$, the smallest design point. This indicates that the objective function used to estimate $\hat{h}$ did not give a useful answer. This is the problem that we face most of the time we directly implement Algorithm 1 to estimate the LHS of 2.21 in Theorem 2.1.4.

In Table 4.4, we see $\hat{h} = 4$ no matter which model is chosen. This indicates that we need 4 explanatory variables to explain our response. Thus, the best model is the one with Y, D, D^2, and Y^2. However, if we use BIC for model selection we select the model with Y, D, and D^2. We prefer the model chosen by $\hat{h}$ because there does not appear to be as strong a curvilinear relationship between speed and distance as there appears to be between speed and Year e.g via Y^2. We attribute the lesser performance of BIC to the fact that its derivation rests heavily on the assumption that the data are independent. Minimizing $\widehat{ERM}_1$ and $\widehat{ERM}_2$ leads to the model with five parameters. There is nothing a priori wrong with this, but a smaller model (of size 4 using $\hat{h}$) is preferred when justifiable. Alternatively, we may regard the difference among the 3th, 4th and 5th models as trivial for $\widehat{ERM}_1$, $\widehat{ERM}_2$, so they effectively lead to the model with Y, D and D^2 as terms, since $\widehat{ERM}_1$ and $\widehat{ERM}_2$ both have a large decrease from the 2 term to the 3 term model. That is, they give the same result as BIC which we think is inferior to the model chose by $\hat{h}$.

4.2.2 The non-nested cases

Now, we replicate what we did above, but without using SCAD to nest the models. Since VCD and real dimension of the parameters in the regression function coincide for linear models, we considered all first and second order terms for this case. We fitted 31 different models, 10 models with only two variables,

10 with 3 variables, 5 with 4 variables, 5 with 1 variable and 1 with 5 variables. For each model in the set, we estimate $\hat{h}$, the upper bound of Propositions 2.1.3 and 2.1.4 and BIC.

Table 4.5: Models of size one using Year and Distance as covariates

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y	4	16.42	44.95	79.67
D	4	59.05	105.39	239.79
D^2	4	78.08	130.13	270.86
Y^2	4	**14.56**	**41.94**	**65.96**
$Y:D$	4	59.05	105.39	239.79

Table 4.6: Models of size two using Year and Distance as covariates

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y, D	4	15.10	42.83	71.66
Y, Y^2	4	16.42	44.95	79.67
Y, D^2	4	15.10	42.83	71.66
$Y, Y:D$	4	**14.56**	**41.94**	**65.96**
D, Y^2	4	46.46	88.51	217.12
D, D^2	4	35.86	73.83	186.69
$D, Y:D$	4	46.14	88.07	216.30
Y^2, D^2	4	100.97	159.03	298.92
$Y^2, Y:D$	4	100.91	158.96	303.50
$D^2, Y:D$	4	78.03	130.06	275.42

Table 4.7: Models of size three using Year and Distance as covariates

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
$Y, Y:D, D^2$	4	12.91	37.90	40.16
$D, Y:D, D^2$	4	24.82	57.86	145.74
Y, D^2, Y^2	4	13.78	40.67	61.87
$Y, Y^2, Y:D$	4	14.26	41.46	67.33
Y, D, Y^2	4	13.48	40.17	58.25
Y, D, D^2	4	**11.21**	**36.37**	**26.21**
$Y, D, Y:D$	4	14.51	41.86	70.05
D, Y^2, D^2	4	34.91	72.49	188.12
$D, Y^2, Y:D$	4	23.79	56.31	140.20
$Y^2, D^2, Y:D$	4	77.59	129.50	279.43

Next, we compare $\hat{h}$, $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC, for the case of non-nested model lists of size 31. We emphasize here that for the cases of non-nested model lists, common practice is to make $\widehat{ERM_1}$ and $\widehat{ERM_2}$ as small as possible. For all these cases, Tables 4.5-4.8 show how $\hat{h}$, $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC behave. From Tables 4.5-4.8, we observe that, for any model in this set of models, the estimated VCD is 4. This indicates the robustness in the estimated VCD for this dataset. We can also see that the behavior of the estimated upper bounds from using either Propositions 2.1.3 or 2.1.4 ($\widehat{ERM_1}$ or $\widehat{ERM_2}$) are the same within each model size. More specifically within each size class, the minimum of $\widehat{ERM_1}$ and $\widehat{ERM_2}$ occurs at the same model. We note that the overall minima of $\widehat{ERM_1}$ and $\widehat{ERM_2}$ occurs for Y,

Table 4.8: Models of size four using Year and Distance as covariates

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y, D, Y^2, D^2	4	**11.09**	**36.16**	**28.77**
$Y, D, D^2, Y:D$	4	11.20	36.35	30.67
$Y, Y^2, D^2, Y:D$	4	11.99	37.70	43.09
$D, Y^2, D^2, Y:D$	4	19.03	49.06	114.79
$Y, D, Y^2, Y:D$	4	11.93	39.12	45.79

$D, Y^2, D^2, Y:D$ (See the last line of Table 4.4, 11.06 and 36.11 respectively.), but the minimum of the models of size four (see Table 4.8 is 11.09 and 36.16 respectively) are nearly the same and nearly lead to the same model as in the nested case (See Table 4.4 where the values are 11.09 and 36.16 respectively.).

The behavior of the BIC is similar to that of the estimated upper bounds $\widehat{ERM_1}$ and $\widehat{ERM_2}$. The smallest BIC model always matches the model with the smallest estimated upper bound, for each model size. However, the smallest BIC over all model sizes occurs for a model of size three, namely Y, D, and D^2. This is not the case when we look at $\widehat{ERM_1}$ or $\widehat{ERM_2}$. As observed before, if we look at graph (a) on Figure 4.1, we see that there is a curvature when we regress Speed on Year; this indicates that the best model should be non-linear in Year. So the best model appears to be the model given by $\hat{h}$, $\widehat{ERM_1}$ or $\widehat{ERM_2}$ (since they give nearly identical values for the model of size 5 and the best model of size 4) not BIC. We can improve the result from $\widehat{ERM_1}$ and $\widehat{ERM_2}$ if we eliminate the extra term $Y:D$ that they include. Looking at Table 4.2, the correlation between $Y:D$ and the average speed is nearly zero. So adding this information to $\widehat{ERM_1}$ or $\widehat{ERM_2}$ improves them to $\hat{h}$ by allowing us to eliminate the extra term $Y:D$. In simulations, we argued that $\hat{h}$ was better than $\widehat{ERM_1}$, $\widehat{ERM_2}$. For efficiency reason; here we argue that $\hat{h}$ is more robust against deviation from independence than BIC.

4.2.3 Analysis of The Tour De France dataset with outliers removed for nested and non-nested cases.

The observations just after World War One may be outliers. Our goal is to see how our estimate will behave after we remove these observations.

As before, we analyze this reduced dataset in two steps. In the first step, we identify the nested model lists by SCAD Fan and Li (2001). Then, for each model in the class, we will estimate $\hat{h}$, and obtain $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC. In the second step, without defining any structure, we fit all models of size 1, 2, 3, 4, and 5.

We first investigate the correlation between the average speed and our covariates before and after removing observations after world war one. Table 4.9 gives us the same information as Table 4.1, but with outliers removed. When comparing the two tables, we see that the correlation between Speed and Year, Speed and Stage won, and between speed and Age increases when outliers are removed. However,

Table 4.9: Correlation between covariates with outliers removed

	Speed	Year	Distance	Stages won	Age
Speed	1	0.96	-0.56	-0.24	0.37
Year	0.96	1	-0.57	-0.25	0.40
Distance	-0.56	-0.57	1	0.15	-0.15
Stages won	-0.24	-0.25	0.15	1	-0.18
Age	0.37	0.40	-0.15	-0.18	1

Table 4.10: Correlation between Speed and D^2, Y^2 and $Y : D$ with outliers removed

	$Distance^2$	$Year^2$	Year:Distance
Speed	-0.5	-0.38	-0.08

it decreases between speed and Distance. Moreover, when we look at correlation between Speed and D^2 (-0.53), we see a slight decrease when outliers are removed, and a slight increase on the correlation between speed and Y^2 (-0.38, Table 4.10). We still did not have much correlation between speed and the interaction between Year and Distance(-0.08).

Table 4.11: Nested models using $Year$ and $Distance$ as covariates with outliers removed

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y	4	12.87	40.72	44.55
Y, D^2	4	12.01	39.26	37.84
Y, D^2, D	4	11.66	38.36	37.41
Y, D^2, D, Y^2	**4**	11.48	38.34	39.20
Y, D^2, D, Y^2, $Y : D$	4	11.35	38.13	41.83

Under SCADFan and Li (2001), the order of inclusion of our covariates is: Y, D^2, D, Y^2 and $Y : D$. This order is different from when we use all data points. Recall that, when we used all data points, D was included before D^2 and D^2 was included after y^2. With this new ordering we fit 5 different models.

From Table 4.11, if we choose a model using $\hat{h}$, we get the same answer as in Sec. 4.2.1, the model with four variables: Y, D^2, D, Y^2. The interaction between Year and distance (Y:D) is not included because of the low correlation between Speed and $Y : D$ (-0.08). Also as before, BIC indicates a model of size 3 having Y, D^2, D as covariates and $\widehat{ERM_1}$ and $\widehat{ERM_2}$ chose a model of size 5. The reasoning in Subsec. 4.2.1 for why we think that the $\hat{h}$ chosen model is best continues to hold.

Next, we turn to the non-nested cases. Tables 4.12-4.15 give analogous results to those in Tables 4.5–4.8 and our observations about $\hat{h}$, $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC are unchanged. In fact, we see that removing outliers did not affect $\hat{h}$, however, there is a change in the values of $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC; this suggests that $\hat{h}$ may be relatively robust against outliers; Alternatively it may only reflect that $\hat{h}$ must be an integer.

For the non-nested case, if we use either $\widehat{ERM_1}$ or $\widehat{ERM_2}$ for model selection, we will choose the model with the smallest value. Now, from Tables 4.12-4.15, it is seen that both $\widehat{ERM_1}$ and $\widehat{ERM_2}$ pick the most complex model with all five parameters. BIC continues to choose Y, D^2 D. Again we invoke

Table 4.12: Models of size one using Year and Distance as covariates with outliers removed

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y	4	**12.87**	**40.72**	**44.55**
D	4	69.60	121.45	246.36
Y^2	4	85.40	141.85	267.20
$Y:D$	4	98.61	158.60	281.68
D^2	4	75.14	128.66	254.21

Table 4.13: Models of size two using Year and Distance as covariates with outliers removed

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
$Y,\,D$	4	12.83	40.64	48.57
Y,Y^2	4	12.39	39.91	43.04
$Y,\,Y:D$	4	12.76	40.52	47.69
$Y,\,D^2$	4	**12.01**	**39.26**	**37.84**
D,Y^2	4	49.11	94.14	214.56
$D,Y:D$	4	52.92	99.31	222.46
D,D^2	4	44.22	87.43	203.35
$Y^2,Y:D$	4	84.23	140.35	270.36
Y^2,D^2	4	74.07	127.27	257.29
$Y:D,D^2$	4	75.14	128.65	258.75

the argument of Subsec. 4.2.1 to argue that the model chosen using $\hat{h}$ is the best.

4.3 Analysis of a nested and non-nested collection of model lists of Tour De France using different covariates

4.3.1 Outliers retained

In this Subsection, we use (Y, D, S, A) as covariates, estimate $\hat{h}$, obtain $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC using for two cases: When covariates are nested and when they are not.

We start by identifying the nested structure of the covariates. Under SCAD Fan and Li (2001), the order of inclusion is $Year$, Age, $Distance$ and $Stageswon$. With this order, we fit 4 different models. For the non-nested case, we fit 15 models. In this list, we fit 6 with 2 variables, 4 with 3 variables, 4 with 1 variable and 1 with all four variables.

In Table 4.16, we have the nested models. We see that $\hat{h} = 4$ for all models. This indicates that a model with all four variables should give good predictions of the average speed of the winner of the Tour. However, if we use $\widehat{ERM_1}$, $\widehat{ERM_2}$ or BIC, the best model will have $Year$, Age and $Distance$ as covariates. However, on panel (a) of Figure 4.2, we see that most of the winners of the Tour won between zero and one stage, so stages won is nearly constant. This suggests that the model chosen by $\widehat{ERM_1}$, $\widehat{ERM_2}$ or BIC is better than the one chose by $\hat{h}$.

For the non-nested case, we see from Tables 4.17 − 4.20 that, for any conjectured model, the estimated

Table 4.14: Models of size three using Year and Distance as covariates with outliers removed

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y, D, Y^2	4	12.09	39.40	43.54
$Y, D, Y:D$	4	12.76	40.52	52.22
Y, D, D^2	4	**11.66**	**38.66**	**37.41**
$D, Y^2, Y:D$	4	23.93	58.24	138.23
D, Y^2, D^2	4	40.50	82.25	198.38
$Y^2, Y:D, D^2$	4	73.98	127.15	261.72
$Y, Y:D, D^2$	4	11.83	38.94	39.80
$D, Y:D, D^2$	4	32.26	71.26	174.88
Y, Y^2, D^2	4	11.95	39.16	41.57
$Y, Y^2, Y:D$	4	12.34	39.81	46.79

Table 4.15: Models of size four and five using Year and Distance as covariates with outliers removed

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y, D, Y^2, D^2	4	**11.48**	**38.34**	**39.20**
$Y, Y^2, D, Y:D$	4	11.93	39.12	45.79
$Y^2, D, D^2, Y:D$	4	21.99	55.28	132.40
$Y, D, D^2, Y:D$	4	11.66	38.66	41.97
$Y, Y^2, D^2, Y:D$	4	11.81	38.90	44.02
$Y, D, Y^2, D^2, Y:D$	4	**11.35**	**38.13**	**41.83**

VCD for the Tour de France data is 4. Thus, if we choose a model using $\hat{h}$, the model that we will choose will again contain Year, Distance, Stages won and Age as covariates. $\widehat{ERM_2}$ behaves as before, and suggests that the best model can either have three parameters Y, D and A, or four parameters Y, D, S and A, since both models have same value for $\widehat{ERM_2}$. BIC and $\widehat{ERM_1}$ confirm the smallest model, i.e with Y, D and A.

We comment that the qualitative appearance of graphs of expected maximum difference VS VCD for Y, D, S, A model list is the same in Figs. $3.6 - 3.8$.

4.3.2 Analysis of Tour De France dataset with outliers removed using Y, D, S, and A as covariates

Recall that the outliers are the observations just after World War One. As we did earlier, we find the best model using this set of covariates in two settings. First, we use SCAD Fan and Li (2001) to order the inclusion of variables into the model. Second, we fit all models of size 1, 2, 3, and 4. Under SCAD, the order of inclusion is $Year$, $Distance$, Age, and the number of $Stage\ won.$ This order is different from that with all observations where we found Y, A, D, S. With this order, we fit 4 models.

From Table 4.21, as before, we observed $\hat{h} = 4$. So the best model under $\hat{h}$ has Y, D, A, and the number of Stages won (S). $\widehat{ERM_1}$ and $\widehat{ERM_2}$ do not discriminate between model of size three and four since they have the same value. In fact, on physical grounds, we can choose the model with Y, D, A since most of the winners of the Tour won between zero and one stage. The BIC picks the model with

Table 4.16: Nested models using all covariates

Model Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
Y	4	16.42	44.95	79.67
Y, A	4	14.79	42.32	68.45
Y, A, D	4	**14.11**	**41.22**	**65.71**
Y, A, D, S	4	14.31	41.22	70.33

Table 4.17: Models of size one using all covariates

Model Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
Y	4	**16.42**	**44.95**	**79.67**
D	4	59.05	105.39	239.79
S	4	103.70	162.43	301.81
A	4	106.42	165.82	304.62

Y and D, omitting A and S. In fact, under $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$, the model with Y, D, A is probably the best since S cannot be very important and the value for it for $\hat{h}$, $\widehat{ERM}_1$, and $\widehat{ERM}_2$ are very similar. The BIC model with Y and D as variables means in particular that A and S have little effect. This is likely true for S (see Fig 4.2 (a)) but less so for A (see Fig 4.2 (b)). Thus in these cases, $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ lead with some extra reasoning to the same model Y, D, A. Since $\widehat{ERM}_1$, $\widehat{ERM}_2$ are derived using $\hat{h}$, this is not particularly surprising.

Thus, with and without outliers, analysis from Table 4.22 to 4.24 is qualitatively the same as the one we have done before.

Summary

In this chapter – Chapter 4 we implemented our theory on Tour de France data. We fitted two set of models. The first set is based on the following covariates Y, D, D^2, Y^2, and $Y : D$, the second contains Y, D, S, and A. Whichever of the two model lists is chosen, $\hat{h} = 4$. The best model to predict the average speed of the winner of the Tour De France dataset should have 4 variables.

Table 4.25 puts together the best model for each size across all sizes when using Year, distance and their combination as covariates. Table 4.26 gives us the same information but with outliers removed. A closer look at these tables informs us that outliers did not have any effect on $\hat{h}$, whereas it did have an effect on $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC since models of size one and two differ from both tables. Over those two tables, BIC picks the model with 3 covariates, whereas $\widehat{ERM}_1$, and $\widehat{ERM}_2$ pick the biggest model. Note also that with $\widehat{ERM}_1$, and $\widehat{ERM}_2$ there is a graduate decrease not a sudden drop to indicate the location of the true model.

Table 4.27 and Table 4.28 respectively put together the best model for each size across all sizes when using Y, D, A and S as covariates with and without outliers removed. We see that outliers did not have any effect on $\hat{h}$, however, it did have an effect on $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC since models of size two differ

Table 4.18: Models of size two using all covariates

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y, D	4	15.10	42.82	71.66
Y, S	4	16.36	44.85	83.79
Y, A	4	**14.79**	**42.32**	**68.45**
D, S	4	57.85	103.80	242.12
D, A	4	57.32	103.09	241.08
S, A	4	103.54	162.23	306.28

Table 4.19: Models of size three using all covariates

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y, D, S	4	15.06	42.77	75.91
Y, D, A	4	**14.11**	**41.22**	**65.71**
Y, S, A	4	14.79	42.32	73.07
D, S, A	4	56.64	102.19	244.35

from both tables. Over those two tables, BIC picks the model with only Y as covariate. When outliers are not removed, $\widehat{ERM_1}$ picks the model of size 3, whereas when outliers are removed, $\widehat{ERM_1}$ did not discriminate between models of size three and four. Across both tables, $\widehat{ERM_2}$ did not discriminate between models of size three and four, its values are the same. We do note, however, that the nesting in Tables 4.27 and 4.28 may indicate a sensitivity to outliers of SCAD.

If we have to choose between the best models from both set of models, with or without outliers, and base our choice either on $\widehat{ERM_1}$ or $\widehat{ERM_2}$, we will have to choose the best model from the first set, since $\widehat{ERM_1}$ or $\widehat{ERM_2}$ is smaller than that of the second set; this suggests that we will need 5 explanatory variables to predict the Average speed of the winner of the Tour, and those variables are: Y, D, Y^2, D^2. and $Y : D$. Moreover, the correlation between the average speed of the winner of the tour and $Y : D$ is almost not existent, also because there is not a big difference between models of size four and five (with and without outliers) in $\widehat{ERM_1}$ and $\widehat{ERM_2}$. We will say that the best model is the model with Y, D, Y^2, D^2 – The model chosen by using $\hat{h}$ in the nested case.

Concerning the second model class (Y, D, A, S), with outliers (see Table 4.16), $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC pick the model containing Y, A and D. However, when outliers are removed (see Table 4.21), $\widehat{ERM_1}$ and $\widehat{ERM_2}$ did not discriminate between models containing Y, D, A and the model containing Y, D, A and S. Since, on physical ground, stages won is just like a decoy, we will say that the model containing Y, D, and A is the one that we should choose and we can use $\hat{h}$ to get this model by invoking the extra information that speed is independent of stages won.

We conclude that the best model is the model chosen by $\hat{h}$ with the following variables: Y, D, Y^2, D^2.

Table 4.20: Models of size four using all covariates

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y, D, S, A	4	14.31	41.22	70.33

Table 4.21: Nested models using all covariates with outliers removed

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y	4	16.42	44.95	79.67
Y, D	4	12.83	40.64	**48.57**
Y, D, A	4	**12.81**	**40.61**	52.89
Y, D, A, S	4	**12.81**	**40.61**	57.41

Table 4.22: Models of size one using all covariates with no outliers

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
A	4	86.32	143.10	268.35
S	4	93.64	152.33	276.50

Table 4.23: Models of size two using all covariates with outliers removed

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y, D	4	**12.82**	**40.64**	**48.57**
Y, A	4	12.84	40.67	48.79
Y, S	4	12.87	40.71	49.09
D, S	4	67.16	118.26	247.25
D, A	4	61.88	111.30	238.80
S, A	4	83.38	139.26	269.33

Table 4.24: Models of size three using all covariates with outliers removed

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y, D, A	4	**12.81**	**40.61**	**52.89**
Y, D, S	4	12.83	40.64	53.11
D, A, S	4	60.62	109.62	241.21
Y, A, S	4	12.85	40.67	53.31
Y, D, A, S	4	**12.81**	**40.61**	57.41

Table 4.25: Best models across all model sizes using Y, D based on $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y^2	4	14.56	41.94	65.96
$Y, Y:D$	4	14.56	41.94	65.96
Y, D, D^2	4	11.21	36.37	26.21
Y, D, Y^2, D^2	4	11.09	36.16	28.77
$Y, D, Y^2, D^2, Y:D$	4	11.06	36.11	32.96

Table 4.26: Best models across all model sizes using Y, D based on $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC with outliers removed

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
Y	4	12.87	40.72	44.55
Y, D^2	4	12.01	39.26	37.84
Y, D, D^2	4	11.66	38.66	37.41
Y, D, Y^2, D^2	4	11.48	38.34	39.20
$Y, D, Y^2, D^2, Y:D$	4	11.35	38.13	41.83

Table 4.27: Best models across all model sizes using Y, D, A, S based on $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC

Model Size	$\widehat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
Y	4	16.42	44.95	79.67
Y, A	4	14.79	42.32	68.45
Y, A, D	4	14.11	41.22	65.71
Y, D, A, S	4	14.31	41.22	70.33

Table 4.28: Best models across all model sizes using Y, D, A, S based on $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC with outliers removed

Model Size	$\widehat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
Y	4	12.87	40.72	44.55
Y, D	4	12.82	40.64	48.57
Y, A, D	4	12.81	40.61	52.89
Y, D, A, S	4	12.81	40.61	57.41

Chapter 5

ANALYSIS OF MORE COMPLEX DATASETS

In Chapter 3, we used synthetic datasets to look at the implementation of theories developed in Chapter 2. In that context, we estimated VCD for different sizes of linear models and we observed some variability in the estimate of VCD. The biggest discrepancy occurred when the size of the true model was $p = 60$. However, this discrepancy was reduced when N was increased. In Chapter 4, we followed essentially the same steps for the Tour De France data. We argue that the Tour de France data set was a toy data because the size (103 observations) was small compared to the number of models that we fitted. In this chapter, we will use larger and more complex datasets to demonstrate the effectiveness of our new methodology.

We start by analyzing the Abalone dataset of Primary Industry and Fisheries (1996) in Sec. 5.1. We compare the model chosen by our method to that of SCAD Fan and Li (2001), ALASSO Zou (2006b) and BIC. We will also replicate what we have done on Tour de France dataset in Chapter 4 on the Wheat dataset in Sec. 5.2. In Sec.5.2.1, we describe the wheat dataset. In Sec. 5.2.2 we estimate the VCD using phenotype covariates by location. In Sec. 5.3, we pool together all location and perform the analysis. In Sec. 5.4, we add variables representing the design structure of the dataset to estimate $\hat{h}$.

5.1 Analysis of the Abalone dataset of Primary Industry and Fisheries (1996)

The Abalone dataset has been widely used in statistics and in machine learning as a benchmark dataset. It is known to be very difficult to analyze as either a classification or as a regression problem. Our goal in this section is to see how our method will perform on the Abalone data set and compare the result to other model selection techniques such as SCAD Fan and Li (2001), Adaptive Lasso (ALASSO) Zou (2006b), and BIC.

5.1.1 Descriptive Analysis of Abalone dataset

The Abalone dataset of Primary Industry and Fisheries (1996) has 4177 observation 8 covariates. Sex is a nominal variable with 3 categories: Male, Female and Infant. Length (mm) is the longest shell
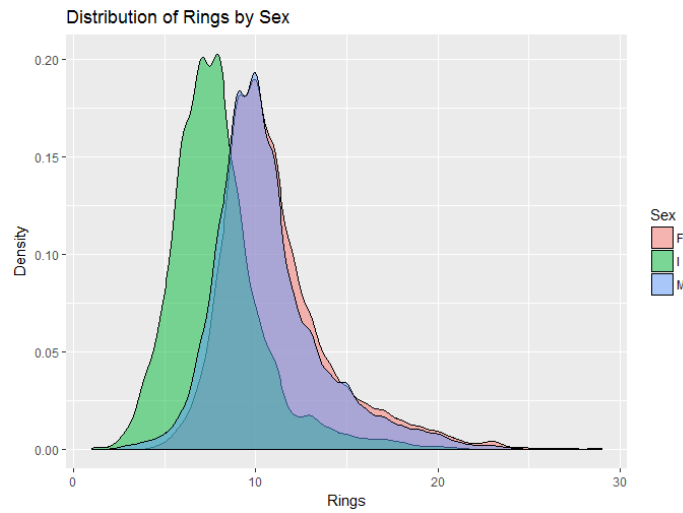
(a) Distribution of $Rings$ by Sex
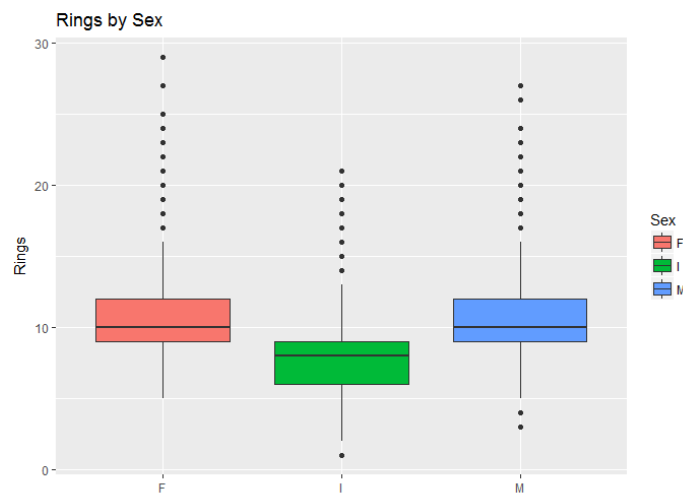
(b) Boxplot of $Rings$ by Sex

Figure 5.1: Density and Boxplot of Abalone

measurement, Diameter (mm), Height (mm) is the height measures with the meat, Whole weight (grams) is the whole weight of the abalone, Shucked weight (grams) is the weight of the meat, Viscera weight (grams) is the gut weight after bleeding, Shell weight is the shell weight after being dried. The response variable, i.e., the Y, is rings. The number of rings is roughly the age of an abalone, since abalone typically grow one ring per year. This dataset is known to be very hard to analyze because many of these covariates are function of each other.

Fig. 5.1a gives the distribution of the rings by Sex. We see that Female and Male abalone have essentially the same distribution. The distribution of Infant abalone is shifted left from the common distribution of Female and Male. All three distributions are skewed to the right but are close to a normal distribution over much of their support. From this graph, we see that if $Rings \leq 5$, the abalone is likely to be an Infant. However if your $Rings \geq 15$, it is more probable that the abalone is a Female or a Male.

	Length	Diameter	Height	WholeWT
Length	1	0.99	0.83	0.93
Diameter	0.99	1	0.83	0.93
Height	0.83	0.83	1	0.82
WholeWT	0.93	0.93	0.82	1
ShuckedWT	0.90	0.89	0.77	0.97
VisceraWT	0.90	0.90	0.80	0.97
ShellWT	0.90	0.91	0.82	0.96
Rings	0.56	0.57	0.56	0.54

Table 5.1: First set of correlation

	ShuckedWT	VisceraWT	ShellWT	Rings
Length	0.90	0.90	0.90	0.56
Diameter	0.89	0.90	0.91	0.57
Height	0.77	0.80	0.82	0.56
WholeWT	0.97	0.97	0.96	0.54
ShuckedWT	1	0.93	0.88	0.42
VisceraWT	0.93	1	0.91	0.50
ShellWT	0.88	0.91	1	0.63
Rings	0.42	0.5	0.63	1

Table 5.2: Second set of correlation

Fig. 5.1b is the box plot of Rings by Sex. We still observe that Female and Male Abalone have the same distribution. We observe some data points outside of the whisker plot; these observations are potential outliers but they are less than 1% of the sample..

Tables 5.1 and 5.2 give all pairwise correlations of the variables. None of the correlation is obviously low. In fact, the smallest correlation (0.42) is between $Rings$ and $Shucked\ weight$. Moreover, we see relatively high correlation between our explanatory variables. In linear models this would leads us to expect problems from multicollinearity. Overall, it is not clear which, if any, of the explanatory can be omitted when explaining the response ($rings$). We also observe that $Length$ is highly correlated with other explanatory variables but not so much with the response ($Rings$).

Figs. 5.2 and 5.3 show all pairwise plots of the form $Rings$ versus covariates (color coded by Sex). We see that no matter which covariates are chosen, the variability in the Rings increases as the size of the covariates increases. We also observe that there is likely to be a curvilinear relationship between Rings and the covariates. However it is so weak for $Rings$ vs $Diameter$ and $Rings$ vs $Length$ that linear terms in these covariates may be adequate.

5.1.2 Statistical Analysis of the Abalone data

Our goal here is to evaluate our method on a more complex dataset. The model that we use to estimate the complexity of the response 'Rings' is a linear combination of all the variables. To accomplish this, we first order the inclusion of variables in the model using correlation Fan and Lv (2008). Under correlation
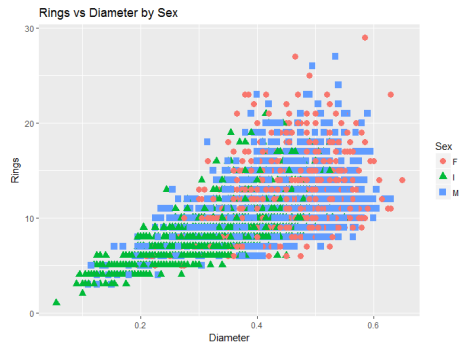
(a) Scatter plot of *Rings* VS *Diameter*
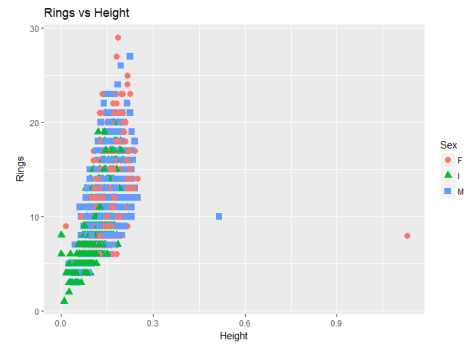
(b) Scatter plot of Rings VS Height
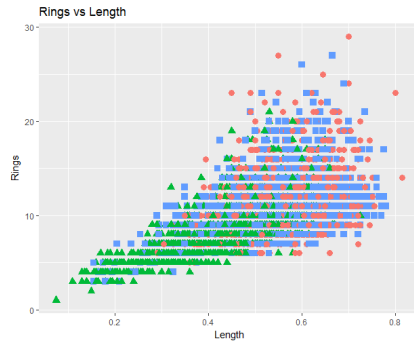
(c) Scatter plot of Rings VS Length
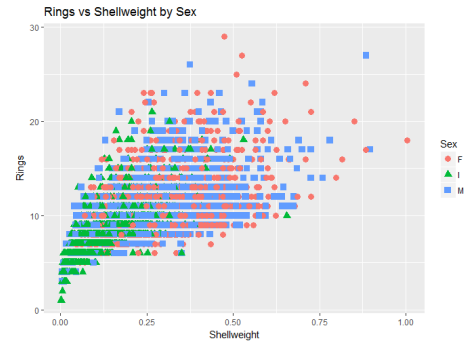
(d) Scatter plot of *Rings* VS *Shell weight*

Figure 5.2: Scatter plot of $Rings$ VS $Diameter, Height, Length$ by Sex
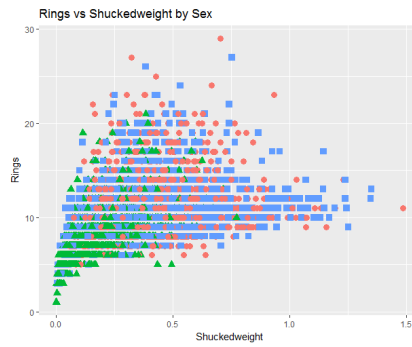
(a) Scatter plot of Rings VS Shucked weight
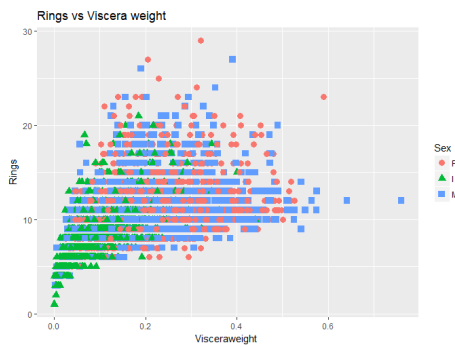
(b) Scatter plot of Rings VS Viscera weight
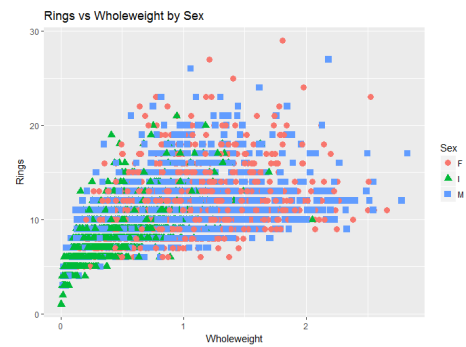
(c) Scatter plot of Rings VS Whole weight

Figure 5.3: Scatter plot of $Rings$ VS $Shell\ weight, Shucked\ weight$ and $Whole\ weight$ by Sex

between $Rings$ and each of the explanatory variables, the order of inclusion of variables is as follows: $Shell\ weight,\ Diameter,\ Height,\ Length,\ Whole\ weight,\ Viscera\ weight$ and $Shucked\ weight.$ Using this ordering, we fit seven different models, estimated $\hat{h}$, and found values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC. These values are in Table 5.3. We also compare our method to other model selection techniques based on sparsity such as SCAD Fan and Li (2001) and Adaptive Lasso Zou (2006b).

From Table 5.3, we observe that $\hat{h} = 8$ except for model of size 5. We regard $\hat{h} = 9$ for a model of size 5 as a random fluctuation since it is close to 8 and our method while stable is not perfectly so. The model chosen using $\hat{h}$ is the biggest model because this model has the smallest distance between its size and $\hat{h}$. However, the fact that model is first order of size 7 and $\hat{h} = 8$ suggest there may be a missing variable in the dataset i.e., unavoidable bias. In fact, observational data is prone to bias than design of experiments, but bias is always present in practice. Note that a missing variable in a model may correspond to many actual variables that would have to be measured. We see some variability in the estimate of $\widehat{ERM}_1$ and $\widehat{ERM}_2$ as we include variables in the model. We see a drop when $Diameter$ is included, and the values go back up when $height$ and $Length$ are included. There is a big decrease at the 6th model and a slight increases at the 7th model. This observation is similar for BIC. So $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC pick the model of size 6 while our method picks the model of size seven and suggest there is a bias from at least a missing variable. We regard the results from using $\hat{h}$ as more plausible physically in the present case.

Table 5.3: Nested models using covariates for Abalone

Model Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
$Shellweight$	8	26315	26524	19567
$Shellweight,\ Diameter$	8	22839	23033	18983
$Shellweight,\ Diameter,\ Height$	8	26045	26253	19540
$Shellweight,\ Diameter,\ Height,\ Length$	8	25745	25952	19500
5	9	23477	23684	19124
6	8	20389	20573	18551
7	8	20507	20696	18575

Next we turn to the results of a sparsity driven analysis. We comment that since there are seven explanatory variables and $N = 4177$, sparsity per se is not necessarily an important property for a model to have. We present these results (for comparative purpose only).

Our first comparison uses SCAD Fan and Li (2001) as a model selection technique. Fig 5.4a gives the value of the cross-validation error for each value of $\log(\lambda)$. The optimal value of λ is seen to be $\hat{\lambda} = 0.0027$. With this value of $\hat{\lambda}$, the best model must have 6 variables. Fig 5.4b is the trace of parameters for different value of λ. Using the optimal value of $\hat{\lambda}$, the variables that get into the model in order are $Shell\ weight,\ Shucked\ weight,\ Height,\ Diameter,\ Viscera\ weight$, and $Whole\ weight.$

(a) Optimal value of λ (b) Order of Inclusion of variables in model
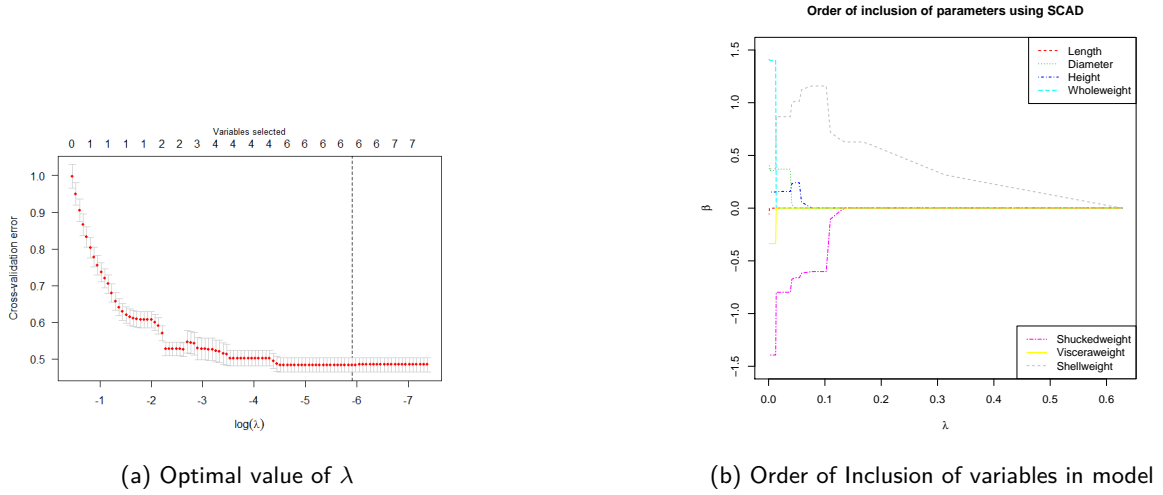
Figure 5.4: Analysis of Abalone dataset using SCAD

Thus, under SCAD, we are led to the model

$$\widehat{Rings} = 0.36 \cdot Diameter + 0.15 \cdot Height + 1.40 \cdot wholeweight$$
$$- 1.39S \cdot shuckedweight - 0.34 \cdot Visceraweight \qquad (5.1)$$
$$+ 0.37 \cdot Shellweight.$$

Analogous analysis under Adaptive LASSO Zou (2006b) leads us to the same six terms and the model is:

$$\widehat{Rings} = 3 + 11.62 \cdot Diameter + 11.69 \cdot Height + 9.21 \cdot wholeweight$$
$$- 20.24 \cdot Shuckedweight - 9.79 \cdot Visceraweight \qquad (5.2)$$
$$+ 8.63 \cdot Shellweight.$$

Even though (5.1) and (5.2) have the same terms, the coefficients are very different. This may occur because there is a high correlation between covariates as observed in Tables 5.1 and 5.2; a prediction based analysis might resolve this question however that is not our point here. Note that the models in (5.1) and (5.2) both included $Shucked\ weight$ but neither included $Length$, whereas the models chosen by BIC, $\widehat{ERM}_1$ and $\widehat{ERM}_2$ include $Length$ but not $Shucked\ weight$. That is, the sparsity models use the same variables, the $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC used the same variables (albeit a different set) and $\hat{h}$ includes all the variables, suggesting that some are missing. As before, we regard the model chosen by $\hat{h}$ as the most reasonable.

5.2 Analysis of the Wheat dataset

The Wheat dataset has 2912 observation, 104 varieties. The experimental study was conducted in seven locations; Lincoln, NE, in 1999 to 2001, and Mead and Sidney, NE, in 2000 and 2001. The design used in Lincoln, NE, in 1999 was a Randomized Complete Block Design (RCBD) with four replicates. In others years, the design was an Incomplete Block design with four replicates where each replication consisted of eight incomplete blocks of thirteen entries. The environments are diverse and representative of wheat producing areas of Nebraska. More information concerning the dataset and the design structure can be found in Campbell et al. (2003). The response variable is $Yield$ (MG/ha), the covariates that we used are 1000 kernel weight (TKWT), kernels per spike (KPS), Spikes per square meter (SPSM), Height of the plant (HT), Test weight (TSTWT(KG/hl)), and Kernels per squared meter (KPSM) in wheat. $Yield$ was measured by harvesting all four rows with a small plot combine.

The analysis of the wheat dataset will be done using phenotype data location by location first. Then we will provide an analysis combining the data from all locations primarily for comparison purposes.

5.2.1 Description of the Wheat dataset

From Fig. 5.5a, the variable of interest is $Yield$. We see that lines connecting locations by varieties are crossing; this indicates there is interaction between varieties and Locations. We can also observed that there is a lot of variations in Lincoln 1999. On graph 5.5b, we see how the yield varies from one location to another. The smallest value of the yield occurs in Lincoln in 1999, and the highest yield occurs at Lincoln in 2001. The most variable location is Sidney in 2000. All data points outside of the whisker plot can be considered as potential outliers.

From Fig. 5.6a, we see that there is a reasonably linear relationship between YIELD and KPSM –although the variance appears to increase slowly with $KPSM$. On Fig. 5.6b, there is fairly good linear relationship between YIELD and SPSM although again the variance increases with $SPSM$. However, on Fig. 5.6c, there is a curvilinear relationship between YIELD and TSTWT, and the variance starts small (as a function of $TSTWT$), increases rapidly and then appears to decrease somewhat. Figs. 5.6d, 5.6e, and 5.6f suggest that the data do not reflect a strong relationship between $YIELD$ and any of HT, KPS, and $TKWT$.

From Table 5.4, we see a strong linear relationship between Yield and $TSTWT$ (0.80), between $Yield$ and $SPSM$ (0.74) and between Yield and $KPSM$ (0.93). There is also a very weak linear relationship between $YIELD$ and HT (0.04), $YIELD$ and KPS (0.022). We also notice some strong correlations between the phenotypic variables. For instance the correlation between $KPSM$ and $TSTWT$ is 0.64, between $KPSM$ and $SPSM$ is 0.83, between $TSTWT$ and $TKWT$ is 0.53. There is also weak
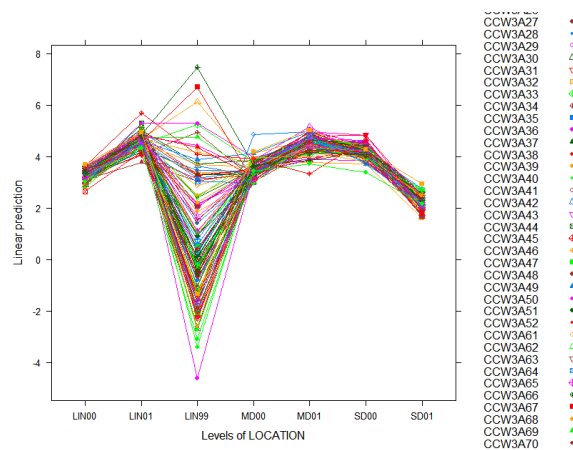
(a) Interaction between Locations and Varieties
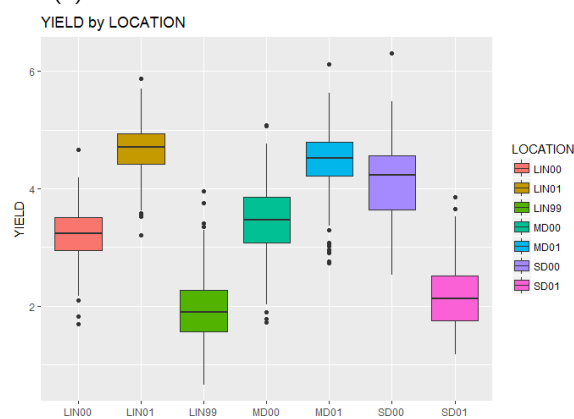
(b) Boxplot of Yield by Location

Figure 5.5: Interaction plot and boxplot of the Yield. Panel 5.5a shows symbols for each variety over the 7 locations. The symbols are connected by lines to shows interaction. Panel 5.5b pools over varieties at each location.
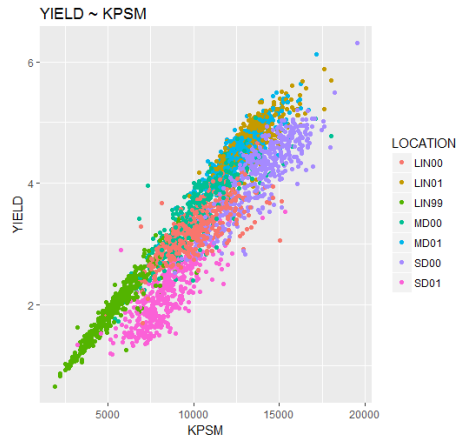
(a) Scatter plot of YIELD vs KPSM
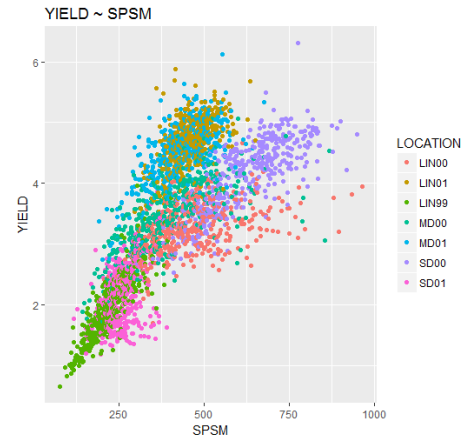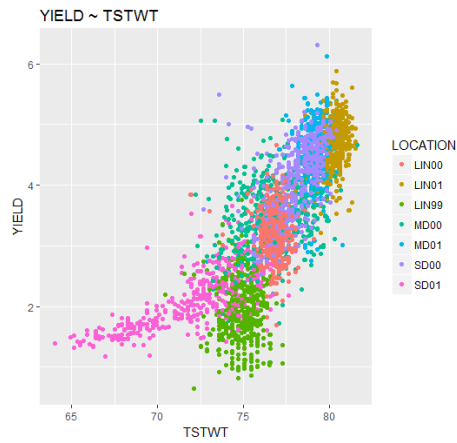
(b) Scatter plot of YIELD vs SPSM

(c) Scatter plot of YIELD vs TSTWT
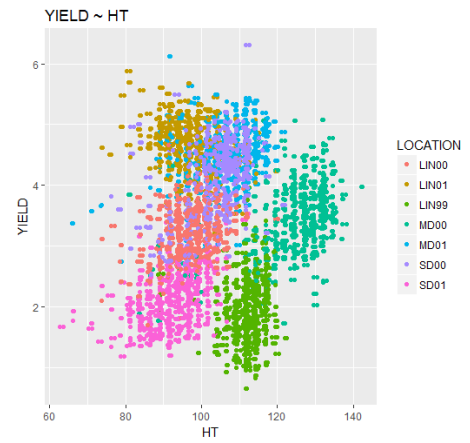
(d) Scatter plot of YIELD vs HT
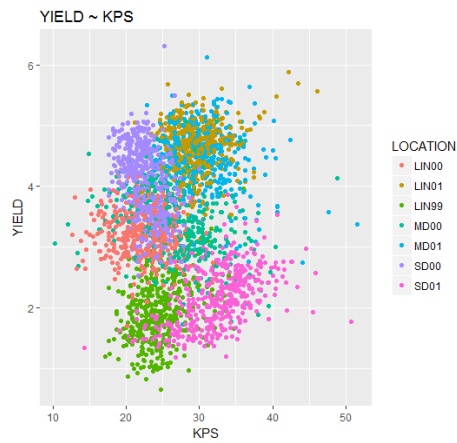
(e) Scatter plot of YIELD vs KPS
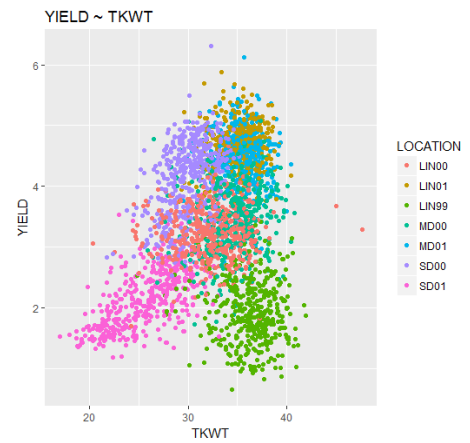
(f) Scatter plot of YIELD vs TKWT

Figure 5.6: Scatter plot of Yield versus phenotype covariates

Table 5.4: Correlation between phenotype covariates over all locations

	$YIELD$	HT	$TSTWT$	$TKWT$	$SPSM$	KPS	$KPSM$
$YIELD$	1.00	0.04	0.80	0.27	0.74	0.022	0.93
HT	0.04	1.00	0.09	0.38	0.02	-0.22	-0.072
$TSTWT$	0.80	0.09	1.00	0.53	0.53	-0.06	0.64
$TKWT$	0.27	0.38	0.53	1.00	-0.11	-0.09	-0.09
$SPSM$	0.74	0.02	0.53	-0.11	1.00	-0.5	0.83
KPS	0.022	-0.22	-0.06	-0.09	-0.5	1.00	0.014
$KPSM$	0.93	-0.07	0.64	-0.09	0.83	0.01	1.00

correlation between $TSTWT$ and HT (0.09), and between $TSTWT$ and KPS (-0.06).

Intuitively, we expect

$$YIELD = \beta_0 + \beta_1 \cdot TKWT \cdot KPSM + \epsilon$$

to be a good model because yield is essentially the product of the number of kernels and their average weight. Similarly,

$$YIELD = \beta_0 + \beta_1 \cdot TKWT \cdot KPS \cdot SPSM + \epsilon$$

should be a good model as well because $KPSM \cong KPS \cdot SPSM$. Thus using our phenotypic variables there is no unique good model. Moreover, these two models probably only capture the major effect of the explanatory variables. So both are over simplifications.

5.2.2 Estimation of VCD using phenotype covariates by Location

We analyze the data first for each location, so that we can see how our methodology performs. Here, we use the six phenotypic variables to estimate $\hat{h}$ and obtain $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC. In fact, the analysis is done in two settings. First, we find the order of inclusion of these variables in the model using correlation Fan and Lv (2008) with $Yield$. Second, we estimate $\hat{h}$ and obtain $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC. We will present the analysis of Lincoln 1999 and the across locations, the rest can be found in Appendix C

Analysis of Wheat data in Lincoln 1999

From Fig. 5.7a, we see that there is a strong linear relationship between $YIELD$ and $KPSM$. The variance is increasing very slowly with $KPSM$ so for practical purposes we regard it as relatively stable i.e essentially constant. There are also some data points not close to the majority of data but the overall trend is linear. From Fig. 5.7b, there is a weak linear relationship between $YIELD$ and $SPSM$. The variance increases as a function of $SPSM$. From Figs. 5.7c to 5.7f, there does not appear to be any non-trivial relationship between between $YIELD$ and any of $TSTWT, HT, KPS$, and $TKWT$. The observations are just spread around and do not appear to show any particular pattern.
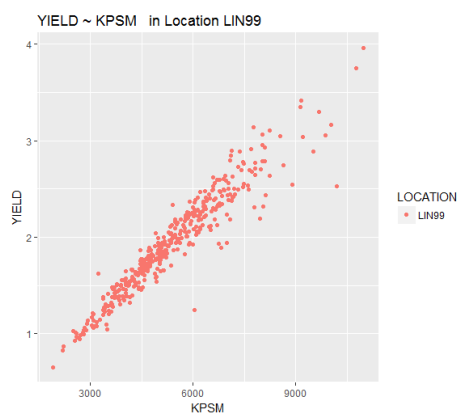
(a) Scatter plot of YIELD vs KPSM
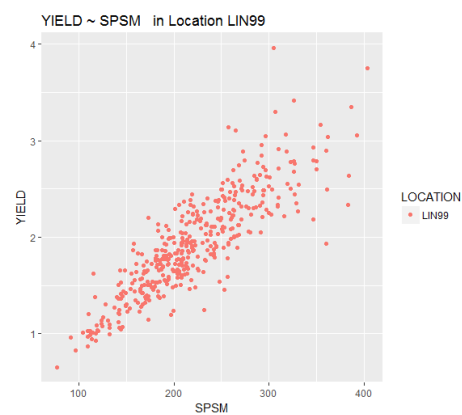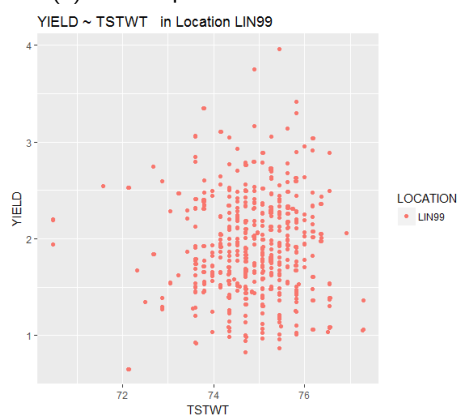
(b) Scatter plot of YIELD vs SPSM

(c) Scatter plot of YIELD vs TSTWT
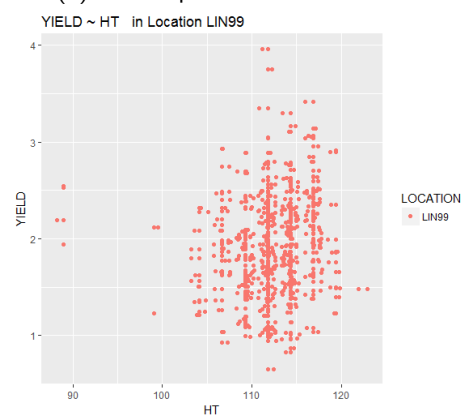
(d) Scatter plot of YIELD vs HT
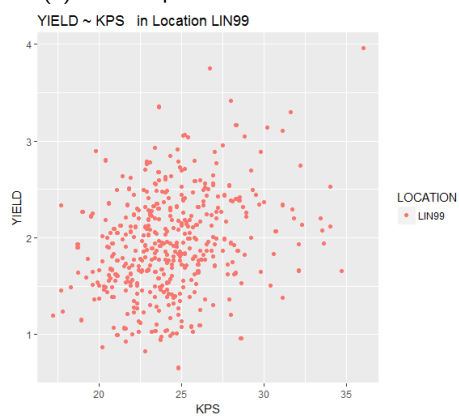
(e) Scatter plot of YIELD vs KPS
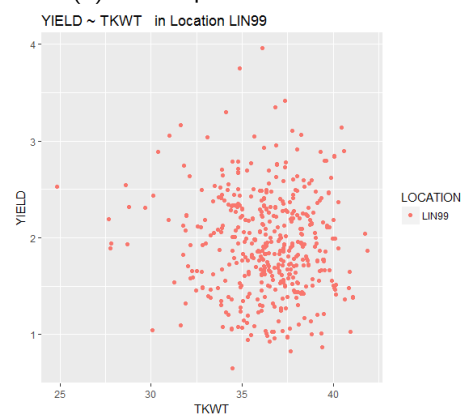
(f) Scatter plot of YIELD vs TKWT

Figure 5.7: Scatter plot of Yield versus phenotype data in Lincoln 1999

We compare our method to BIC and also to sparse model selection technique as SCAD Fan and Li (2001) and ALASSO Zou (2006b). We use correlation Fan and Lv (2008) between $YIELD$ and the phenotypic covariates to order the inclusion of phenotype variables and their products in the model. In fact, there are six linear terms, six squared terms, and fifteen cross products, this leads to a total of 27 variables and therefore, 27 nested models to fit. Using correlation, the order of inclusion of terms is: $TKWT \cdot KPSM, KPSM \cdot HT, KPSM \cdot TSTWT, SPSM \cdot KPS, KPSM, KPSM^2, KPSM \cdot SPSM, SPSM \cdot TKWT, KPSM \cdot KPS, SPSM, SPSM \cdot TSTWT, SPSM \cdot HT, SPSM^2, KPS \cdot HT, KPS \cdot TSTWT, KPS^2, KPS, KPS \cdot TKWT, HT^2, HT, HT \cdot TSTWT, TSTWT, TSTWT^2, TKWT \cdot HT, TKWT \cdot TSTWT, TKWT^2, TKWT.$

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
1	18	3.33	6.64	-1221.52
2	17	3.28	6.43	-1215.85
3	16	3.23	6.22	-1210.61
4	17	3.28	6.42	-1205.50
5	19	3.52	6.80	-1206.86
6	19	3.35	6.80	-1201.31
7	19	3.35	6.80	-1195.29
8	20	3.40	7.00	-1189.421
9	19	3.35	6.80	-1184.31

(a) Model sizes 1 to 9

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
10	19	3.35	6.80	-1178.39
11	20	3.39	6.99	-1172.38
12	20	3.90	6.99	-1167.28
13	20	3.99	6.99	-1161.4
14	19	3.33	6.76	-1159.72
15	20	3.38	6.96	-1154.10
16	20	3.76	6.96	-1148.37
17	21	3.42	7.16	-1142.42
18	21	3.42	7.16	-1136.79

(b) Model sizes 10 to 18

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
19	21	3.42	7.16	-1130.82
20	21	3.42	7.15	-1125.49
21	22	3.46	7.35	-1119.55
22	22	3.45	7.34	-1115.6
23	12	2.96	5.26	-1109.60
24	22	3.45	7.33	-1104.06
25	22	3.45	7.34	-1098.50
26	12	2.95	5.25	-1092.47
27	12	2.95	5.25	-1087.2

(c) Model sizes 19 to 27

Table 5.5: Estimation of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_1$ and BIC in LIN99. Size indicates the dimension of the parameter space of the linear model.

Table. 5.5, shows some variability in the estimate of $\hat{h}$, also in the values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC. The smallest difference between the size of the conjectured model and $\hat{h}$ occurs when the size of the conjectured model is 22; this indicates that the model chosen by $\hat{h}$ has 22 variables. $\widehat{ERM}_1$ and $\widehat{ERM}_2$ choose the most complex model with all parameters. BIC picks the smallest model with $TKWT \cdot KPSM$ as variable.

Now, we turn to analyzing the Lincoln 1999 dataset using shrinkage methods. From Fig. 5.8a, the optimal value of λ using SCAD is $\hat{\lambda} = 0.1281894$; with this value, the optimal model under SCAD should have one variable. From Fig. 5.8b, and using the optimal $\hat{h}$, the variable chosen is $KPSM \cdot TKWT$,
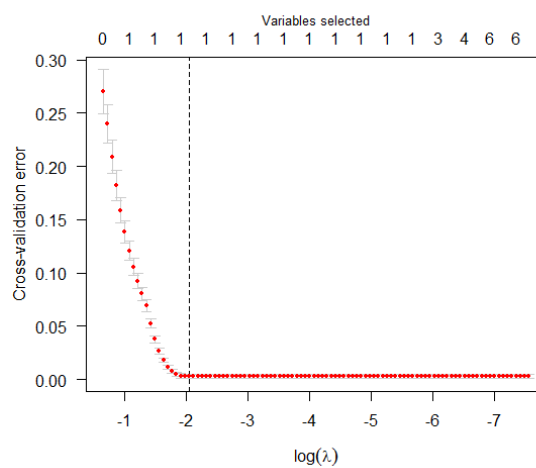
(a) Optimal value of λ for Lincoln 1999
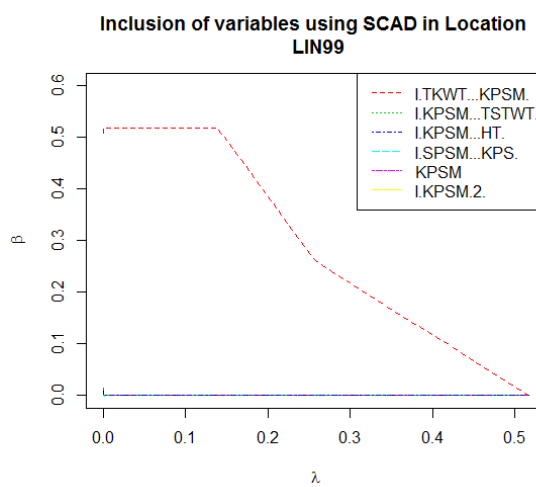
(b) Trace of variables for Lincoln 1999 using SCAD Fan and Li (2001)

Figure 5.8: Analysis of Wheat data set in Lincoln 1999 using SCAD Fan and Li (2001)

that is, the model is

$$\widehat{YIELD} = 1.93 + 0.52 \cdot KPSM \cdot TKWT. \tag{5.3}$$

Under Adaptive LASSO, a similar analysis leads to

$$\widehat{YIELD} = 1.93 + 0.48 \cdot KPSM \cdot TKWT + 0.038 \cdot SPSM \cdot KPS + 0.0043 \cdot TKWT. \tag{5.4}$$

Note that for the Lincoln 1999, SCAD and BIC give the same smallest model that is also intuitively reasonable. However, it is known there are effects on $Yield$ from explanatory variables other than $TKWT \cdot KPSM$. Since, we ignore incomplete blocks for the analysis, the estimated VCD could be mostly the incomplete block effects.

5.3 Multilocation analysis

For this analysis, we pool over locations, and we did not include location as a class variable inside the model. We start by ordering the inclusion of terms in the model using correlation Fan and Lv (2008). Under correlation, the nesting of our variables is as follows: $TKWT \cdot KPSM, KPSM \cdot TSTWT, KPSM,$ $SPSM \cdot KPS, KPSM^2, KPSM \cdot HT, SPSM \cdot TKWT, KPSM \cdot SPSM\ TSTWT^2, TSTWT, KPSM \cdot$ $KPS, SPSM \cdot TSTWT, SPSM, SPSM \cdot HT, SPSM^2, TKWT \cdot TSTWT, TKWT, HT \cdot TSTWT,$ $TKWT^2, KPS \cdot TKWT, KPS \cdot TSTWT, TKWT \cdot HT, KPS \cdot HT\ HT, HT^2, KPS, KPS^2$. With this order, we fit 27 models and each time we estimate $\hat{h}$ and find the the values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC.

From Table 5.6, we see that no matter which model you pick, $\hat{h} = 6$ and $\widehat{ERM}_1 = 7$ and $\widehat{ERM}_2 = 10$. We observe variability in the values of BIC. The smallest value of BIC occurs with the first model. Also, $\widehat{ERM}_1$ and $\widehat{ERM}_2$ pick a model of size 1. The fact that $\widehat{ERM}_1$ and $\widehat{ERM}_2$ is constant might indicate that variables included after the first one did not improve the prediction power of the model. However, $\hat{h}$ picked the model of size 6, a value reflecting a sort of middle point among $\hat{h}$'s found for the seven locations separately. Turning our attention to the sparsity methods, we observe from Fig. 5.9a that the optimal value of λ is $\hat{\lambda} = 0.29736$; this indicates that the optimal number of parameters is 1. From Fig. 5.9b, under SCAD, the variable chosen is $TKWT \cdot KPSM$, and the model can be written as:

$$\widehat{YIELD} = 3.43 + 1.12 \cdot KPSM \cdot TKWT. \tag{5.5}$$

Similarly, using ALASSO we have

$$\widehat{YIELD} = 3.43 + 1.12 \cdot KPSM \cdot TKWT \tag{5.6}$$

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
1	6	7	10	-9750
2	6	7	10	-9743
3	6	7	10	-9737
4	6	7	10	-9731
5	6	7	10	-9725
6	6	7	10	-9718
7	6	7	10	-9711
8	6	7	10	-9703
9	6	7	10	-9699

(a) Model sizes from 1 to 9

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
10	6	7	10	-9691
11	6	7	10	-9683
12	6	7	10	-9675
13	6	7	10	-9668
14	6	7	10	-9660
15	6	7	10	-9652
16	6	7	10	-9645
17	6	7	10	-9638
18	6	7	10	-9632

(b) Model sizes from 10 to 18

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
19	6	7	10	-9625
20	6	7	10	-9617
21	6	7	10	-9610
22	6	7	10	-9603
23	6	7	10	-9596
24	6	7	10	-9590
25	6	7	10	-9582
26	6	7	10	-9583
27	6	7	10	-9575

(c) Model sizes from 19 to 27

Table 5.6: Estimation of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_1$ and BIC for wheat data

We note here that ALASSO and SCAD are identical. BIC, $\widehat{ERM}_1$ and $\widehat{ERM}_2$ also pick $TKWT \cdot KPSM$ as the unique variable inside their model. the model chosen by $\hat{h}$ is the most complex one with 6 variables. Thus, all techniques except $\hat{h}$ simply choose the single most important term across locations while $\hat{h}$ tried to choose a model that was overall representative of the data.

Table 5.7: Nested models using all phenotype covariates

Model Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
$KPSM$	6	1226.38	1273.28	5774.36
$KPSM, TKWT$	6	56.88	66.56	-3234.779
$KPSM, TKWT, KPS$	6	50.75	59.86	-3567.844
$KPSM, TKWT, KPS, HT$	6	49.03	57.98	-3662.68
$KPSM, TKWT, KPS, HT, TSTWT$	6	48.41	57.30	-3692.71
$KPSM, TKWT, KPS, HT, TSTWT, SPSM$	6	48.02	56.86	-3709.31

We redid our analysis using first order linear models for the sake of comparison. From Table 5.7, we see that $\hat{h} = 6$ for all models in the list. This indicates that we should use all phenotypic variables to predict the yield. This model is not accurate because we see from Table 5.4 that there is a strong correlation between some of our covariates and we suspect the terms $SPAM \cdot TKWT$ is important. We also see a sudden drop in the values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC when we move from model of size 1 to model of size 2, and a gradual decrease thereafter.

from Tables 5.8 − 5.11, $\hat{h} = 6$, we see variability in the values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC. The

(a) Optimal value of λ across all locations

(b) Trace of variables across all location using SCAD Fan and Li (2001)
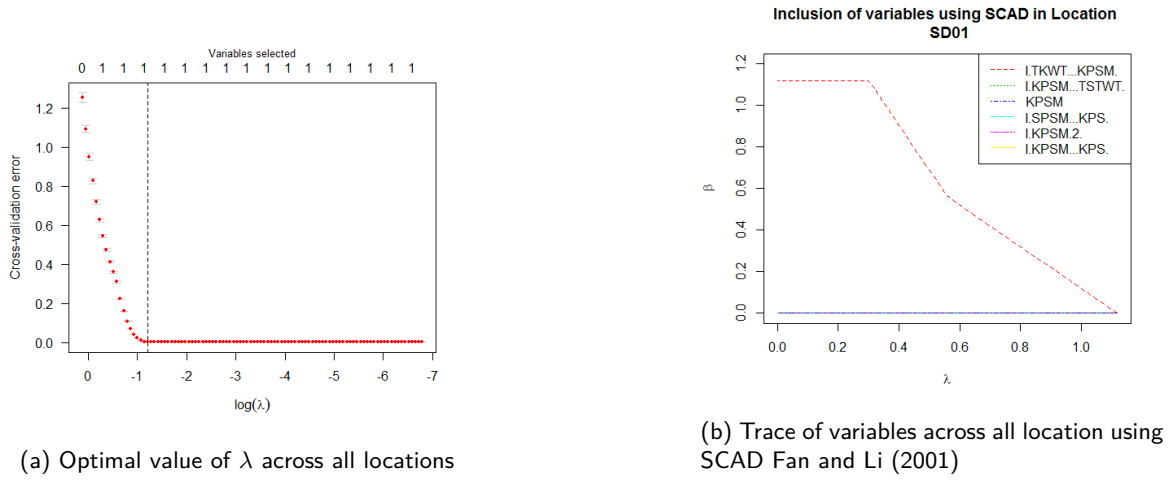
Figure 5.9: Analysis of Wheat data set with all locations combined using SCAD Fan and Li (2001)

Table 5.8: Models of size two using all phenotypes covariates

Model Size	$\widehat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
$HT, TSTWT$	6	1329.03	1377.88	6008.69
$HT, TKWT$	6	3364.70	3442.70	8715.42
$HT, SPSM$	6	1676.28	1731.20	6685.26
$HT, KPSM$	6	482.73	511.98	3054.36
HT, KPS	6	3647.63	3728.86	8950.62
$TKWT, KPS$	6	3374.08	3452.19	8723.53
$TKWT, KPSM$	6	57	67	-3235
$TKWT, SPSM$	6	1218.70	1265.45	5756.05
$TSTWT, KPS$	6	1314.87	1363.45	5977.47
$TSTWT, KPSM$	6	265.95	287.53	1311.74
$TSTWT, SPSM$	6	825.04	863.42	4618.51
$TSTWT, TKWT$	6	1226.38	1273.28	5774.36
$KPS, KPSM$	6	524.57	555.08	3297.08
$SPSM, KPS$	6	941.17	982.20	5002.59
$SPSM, KPSM$	6	510.22	540.31	3216.12

smallest values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC occur with the model of size 6. in agreement with $\hat{h}$. This model is not the best we can have because $KPSM \cdot TKWT$ is important to include. Note that here we believe the model list excludes many good models but $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC give the same model and $\hat{h}$ is a constant equal to its value on the earlier model list of size 27 when all locations are combined.

5.4 Estimate of VCD using the design structure of the model

Our objective in section is to see the impact of taking into account the design structure in our estimate. We implement our method on Lincoln 2000 and 2001 datasets. We are aware that the bootstrap technique that we used in our method will destroy the design used to collect this dataset, however, we took account of the blocking factor for completeness. Since the blocking variable is a factor variable with 32 different categories, we cannot use it to compute the correlation; so to include variable in the model, we use

Table 5.9: Models of size three using all phenotypes covariates

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
$HT, KPS, KPSM$	6	478.89	508.02	3039.01
$HT, SPSM, KPS$	6	877.51	917.11	4806.32
$HT, SPSM, KPSM$	6	459.07	487.58	2915.55
$HT, TKWT, KPS$	6	3360.45	3438.40	8719.71
$HT, TKWT, KPSM$	6	53.75	63.15	-3395.81
$HT, TKWT, SPSM$	6	1163.39	1209.06	5628.62
$HT, TSTWT, KPS$	6	1314.02	1362.59	5983.56
$HT, TSTWT, KPSM$	6	251.95	272.94	1161.43
$HT, TSTWT, SPSM$	6	823.61	861.97	4621.44
$HT, TSTWT, TKWT$	6	1220.27	1267.05	5767.78
$TSTWT, SPSM, KPS$	6	430.42	458.01	2727.35
$TSTWT, SPSM, KPSM$	6	253.69	274.75	1181.48
$TSTWT, TKWT, KPS$	6	1214.51	1261.19	5754.00
$TSTWT, TKWT, KPSM$	6	56.53	66.18	-3245.16
$TSTWT, TKWT, SPSM$	6	821.48	859.78	4613.88
$TKWT, KPS, KPSM$	6	50.75	59.86	-3567.844
$TKWT, SPSM, KPS$	6	247.12	267.90	1104.71
$TKWT, SPSM, KPSM$	6	50.36	59.44	-3590.39
$SPSM, KPS, KPSM$	6	449.08	477.27	2851.33

Table 5.10: Models of size four using all phenotypes covariates

Model Size	$\widehat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
$HT, SPSM, KPS, KPSM$	6	416.29	443.41	2637.80
$TSTWT, TKWT, SPSM, KPS$	6	244.28	264.94	1078.83
$TSTWT, TKWT, SPSM, KPSM$	6	50.12	59.17	-3596.89
$HT, TKWT, SPSM, KPS$	6	244.49	265.16	1081.35
$HT, TKWT, SPSM, KPSM$	6	48.64	57.55	-3686.69
$HT, TKWT, KPS, KPSM$	6	49.03	57.98	-3662.68
$HT, TSTWT, KPS, KPSM$	6	244.28	264.94	1078.83
$HT, TSTWT, SPSM, KPS$	6	411.54	438.51	2604.33
$HT, TSTWT, SPSM, KPSM$	6	234.78	255.02	962.65
$HT, TSTWT, TKWT, KPS$	6	1204.10	1250.58	5736.89
$HT, TSTWT, TKWT, KPSM$	6	52.87	62.19	-3436.98
$HT, TSTWT, TKWT, SPSM$	6	816.28	854.46	4603.33

the order of inclusion from the analysis of the Lincoln 2000 and 2001 datasets. For each model, we add the blocking variable. For instance, in Lincoln 1999, the first term that gets into the model is $TKWT \cdot KPSM$, the model that we fit has $TKWT \cdot KPSM$ plus $IBLK_j, j = 1, 2, \cdots, 32$ ($IBLK_j$ is the jth incomplete block). So that is the size of each model increases by 32. As another example, the 3rd model indicated in Table C.1 becomes

$$\widehat{YIELD} = \beta_0 + \beta_1 TKWT \cdot KPSM + \beta_2 KPSM \cdot TSTWT + \beta_3 KPSM \cdot HT + \sum_{k=1}^{32} \theta_k IBLK_k$$

Table 5.11: Models of size five and six using all phenotypes covariates

Model Size	$\widehat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
$HT, TKWT, SPSM, KPS, KPSM$	6	48.61	57.51	-3680.81
$HT, TSTWT, TKWT, KPS, KPSM$	6	48.41	57.30	-3692.71
$HT, TSTWT, TKWT, SPSM, KPS$	6	233.90	254.11	959.71
$HT, TSTWT, TKWT, SPSM, KPSM$	6	48.05	56.90	-3715.12
$HT, TSTWT, SPSM, KPS, KPSM$	6	226.56	246.45	866.46
$TSTWT, TKWT, SPSM, KPS, KPSM$	6	50.05	59.09	-3593.21
$KPSM, TKWT, KPS, HT, TSTWT, SPSM$	6	48.02	56.86	-3709.31

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
1	13	50	63	490
2	13	50	63	493
3	13	25	34	191
4	13	25	33	195
5	15	18	26	49
6	15	17	25	33
7	16	11	17	-173
8	16	9	15	-267
9	16	9	14	-274

(a) Model sizes from 1 to 9

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
10	16	9	14	-271
11	16	9	14	-267
12	16	9	14	-264
13	16	9	14	-259
14	16	8	14	-276
15	16	8	14	-270
16	16	8	14	-271
17	16	5	10	-519
18	15	5	9	-518

(b) Model sizes from 10 to 18

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
19	15	5	9	-518
20	15	5	9	-513
21	15	5	9	-507
22	15	5	9	-501
23	15	4	7	-778
24	15	2	4	-773
25	15	2	4	-3325
26	15	2	4	-3319
27	15	2	4	-3313

(c) Model sizes from 19 to 27

Table 5.12: Estimation of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_1$ and BIC in LIN00; note that the model is 32 plus the size of the model due to including $IBLK$

Analysis of Lincoln 2000 data with the design structure included in the model

Table 5.12 gives the estimate of $\hat{h}$, the values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC in Lincoln 2000. In these cases, we included the design structure represented by the incomplete block.

We observe small variability in the estimate of $\hat{h}$: it takes on values 13, 15, or 16. The smallest discrepancy between the size of the conjectured model and $\hat{h}$ occurs when the size of the conjectured model is 16. In fact, at that point $\hat{h} = 16$. We also see that there is variability in the values of $\widehat{ERM}_1$ and $\widehat{ERM}_2$ as the size of the conjectured model increases. The smallest value of $\widehat{ERM}_1$ namely two, occurs first when the size of the conjectured model is 24; this value stays constant up to the most complex model. So, so we would pragmatically say that $\widehat{ERM}_1$ picks a model of size 24. The smallest value of $\widehat{ERM}_2$ first occurs at the conjectured model of size 24 and stays constant up to the model of size 27.
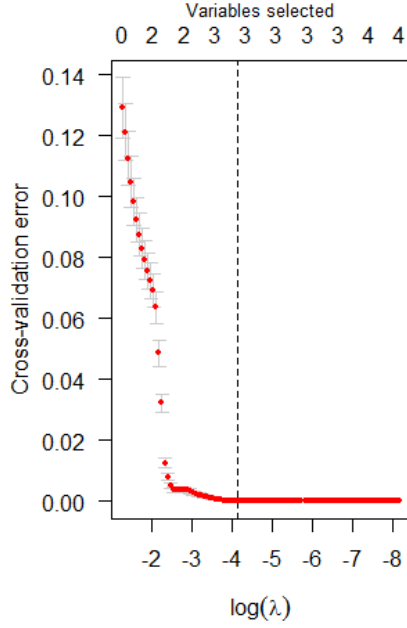
(a) Optimal value of λ In Lincoln 2000 with design structure included
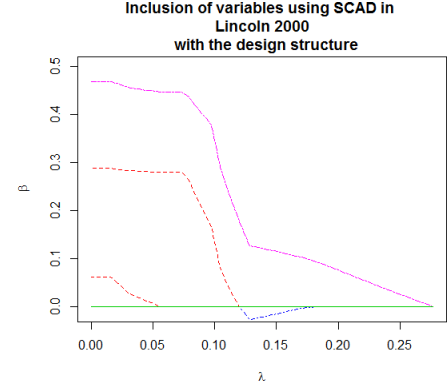
(b) Trace of variables in Lincoln 2000 with design structure included using SCAD Fan and Li (2001)

Figure 5.10: Analysis of Wheat data set in Lincoln 2000 with the design structure included in the model using SCAD Fan and Li (2001)

Being parsimonious, we would say that $\widehat{ERM}_2$ also picks a model of size 24. On the other hand, BIC picks the most complex model i.e.the model with the most variables.

Turning our attention to the sparsity methods, we observe from Fig. 5.10a that the optimal value of λ is $\hat{\lambda} = 0.0158356$; this indicates that the optimal number of parameters is 3. From Fig. 5.10b, under SCAD, the variables chosen are $TKWT \cdot KPSM$, $KPSM$ and $TKWT$. The model can be written as:

$$\widehat{YIELD} = -0.18 + 0.06 \cdot KPSM \cdot TKWT + 0.47 \cdot KPSM + 0.29 \cdot TKWT. \qquad (5.7)$$

The coefficients of $IBLK$ under SCAD were set to zero, but including $IBLK$ in our SCAD analysis give us two extra terms compare to (C.1).

We note here that all model selection techniques pick a different model list. BIC picks the most complex model list, $\widehat{ERM}_1$ and $\widehat{ERM}_2$ pick a model list of size 27, $\hat{h}$ pick a model of size 16 and SCAD pick a model of size 3. We also note that in (5.7) we have the interaction between $TKWT$ and $KPSM$ and their main effect. This is the first time that this occurs using shrinkage methods. The model given by SCAD is the best smallest model that we can have, however, the question is: Is this model the overall best model? I would say that this is not the overall best model because this model does not take into account the different type of varieties of wheat. (We comment that, here we have not used ALASSO

because the R code we used did not support categorical variables.)

The effect of including the design variables is substantial. In Table C.1, $\hat{h} = 1$; here $\hat{h} = 16$. Likewise, there are large difference in the other methods; for instance, compare (C.1) to (5.7), or $\widehat{ERM}_1$ and $\widehat{ERM}_2$ in Table C.1 to the values in Table 5.12.

Analysis of Lincoln 2001 data with the design structure included in the model

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
1	16	31	42	284
2	12	4	7	-783
3	10	3	5	-943
4	16	10	16	-239
5	16	10	16	-234
6	15	10	16	-234
7	14	7	11	-441
8	11	3	5	-922
9	12	3	6	-924

(a) Model sizes from 1 to 9

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
10	11	3	6	-918
11	12	3	6	-913
12	9	3	4	-1000
13	9	3	4	-1000
14	9	3	4	-992
15	9	3	4	-986
16	9	3	4	-980
17	8	3	4	-981
18	42	3	19	-1452

(b) Model sizes from 10 to 18

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
19	32	3	7	-1853
20	32	3	7	-1831
21	32	2	3	-1826
22	28	3	7	-1829
23	28	3	7	-1823
24	28	3	7	-1817
25	27	3	6	-1819
26	27	3	6	-1819
27	27	3	6	-1819

(c) Model sizes from 19 to 27

Table 5.13: Estimation of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_1$ and BIC in LINO1 with the design structure included

Table 5.13 gives the estimate of $\hat{h}$, the values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC in Lincoln 2001. In these cases, we included the design structure represented by the incomplete block.

We observe some a lot of variability in the estimate of $\hat{h}$: it takes on values from 8 to 42. The smallest discrepancy between the size of the conjectured model and $\hat{h}$ occurs when the size of the conjectured model is 27. In fact, at that point $\hat{h} = 27$. We also see that there is variability in the values of $\widehat{ERM}_1$ and $\widehat{ERM}_2$ as the size of the conjectured model increases. The smallest value of $\widehat{ERM}_1$ namely two, occurs first when the size of the conjectured model is 21. The smallest value of $\widehat{ERM}_2$ also occurs at the conjectured model of size 21. The smallest value of BIC occurs with model of size 19. Turning our attention to the sparsity methods, we observe from Fig. 5.11a that the optimal value of λ is $\hat{\lambda} = 0.00651666$; this indicates that the optimal number of parameters is 3. From Fig. 5.10b, under
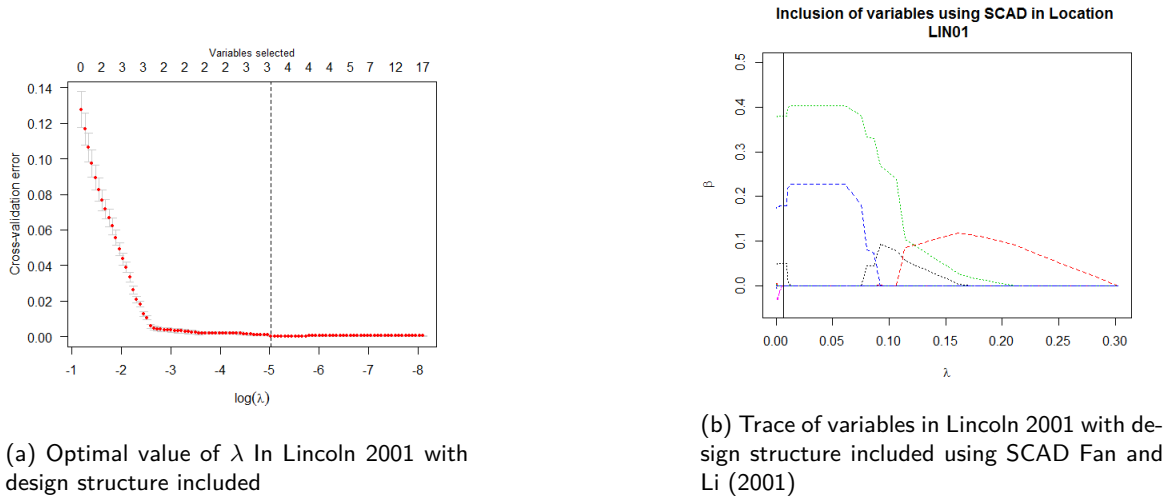
(a) Optimal value of λ In Lincoln 2001 with design structure included

(b) Trace of variables in Lincoln 2001 with design structure included using SCAD Fan and Li (2001)

Figure 5.11: Analysis of Wheat data set in Lincoln 2001 with the design structure included in the model using SCAD Fan and Li (2001)

SCAD, the variable chosen is $TKWT \cdot KPSM$, $KPSM$ and $TKWT$ and the model can be written as:

$$\widehat{YIELD} = 1.10 + 0.05 \cdot KPSM \cdot TKWT + 0.38 \cdot KPSM + 0.18 \cdot TKWT. \tag{5.8}$$

This has the same explanatory variables as (5.7).

We note again that all model selection techniques pick a different model. BIC picks a model with 19 variables, $\widehat{ERM}_1$ and $\widehat{ERM}_2$ pick a model of size 21, $\hat{h}$ pick a model of size 27 and SCAD pick a model of size 3. We also note that in (5.8) we have the interaction between $TKWT$ and $KPSM$ and their main effects.

Overall, the model sizes range from 3 (SCAD), 19 (BIC), 21 ($\widehat{ERM}_1$ and $\widehat{ERM}_2$) to 27 ($\hat{h}$). Since SCAD and BIC are sparsity or sparsity-like procedures, it is not surprising that they give the smallest models and unless a coarse model is desired, SCAD and possibly BIC can be ignored. In this example, we are left with models of size 21 and 27 and we recall that, compared to Bayes methods such as SCAD and BIC, our method is more sensitive to bias or, more precisely to exactly which variables are being included. We suspect that as in other cases, $\hat{h}$ is even more sensitive than $\widehat{ERM}_1$ and $\widehat{ERM}_2$ and has detected that some variables are relevant when blocking in included. Indeed, the elevated size of $\hat{h}$ shows how important the blocking is.

5.5 Analysis of Wheat data using SNP information

Our goal in this section is to see the effect of how taking the SNP (Single Nucleotide Polymorphism) information into account will affect the estimate of VCD. Since the variables representing the SNP's have so much missing data, we imply dropped all rows with missing SNP values and only use SNP's that are

complete (No missing values i.e., we dropped columns as well.) Thus we retained only 6 SNP's and our sample size was reduced to 2631. As before, we ordered the inclusion of covariates in our model using correlation. Under correlation is inclusion of terms in our model is as follows: $KPSM, TSTWT, SPSM,$ $KPS \cdot TKWT, TKWT, KPS \cdot TSTWT, TKWT \cdot KPSM, KPS \cdot HT, barc67, HT, cmwg680bcd366,$ $bcd141, barc86, gwm155, barc12, KPS, SPSM \cdot HT, KPSM \cdot HT, SPSM \cdot TKWT, SPSM^2, KPS^2,$ $HT^2, SPSM \cdot KPS, KPSM \cdot TSTWT, TKWT \cdot TSTWT, KPSM \cdot SPSM, TSTWT^2, KPSM^2,$ $SPSM \cdot TSTWT, HT \cdot TSTWT, TKWT^2, \ TKWT \cdot HT.$ Here the SNP's are symply denoted by their labels in the data. With this order, we fit 32 different models, each time, we estimate $\hat{h}$, find values of $\widehat{ERM_1}$, and $\widehat{ERM_2}$ and models for BIC, SCAD and ALASSO. Table 5.14 gives the estimate of $\hat{h}$,

Size	$\hat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
1	37	380	439	2381
2	37	193	235	587
3	37	185	226	477
4	37	181	222	430
5	36	37	55	-3930
6	37	37	56	-3924
7	37	8	17	-9100
8	37	8	17	-9093
9	36	8	16	-9089
10	36	8	16	-9081

(a) Model sizes from 1 to 10

Size	$\hat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
11	36	8	16	-9074
12	36	8	16	-9066
13	36	8	16	-9058
14	36	8	16	-9052
15	36	8	16	-9045
16	36	8	16	-9040
17	36	8	16	-9033
18	36	8	16	-9026
19	36	8	16	-9019
20	36	8	16	-9014

(b) Model sizes from 11 to 20

| Size	$\hat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
21	36	8	16	-9006
22	36	8	16	-8999
23	41	8	17	-8993
24	41	8	17	-8989
25	41	8	17	-8982
26	41	8	17	-8978
27	41	8	17	-8970
28	41	8	17	-8962
29	41	8	17	-8961
30	41	8	17	-8956
31	41	8	17	-8948
32	41	8	17	-8940

(c) Model sizes from 19 to 27

Table 5.14: Estimation of $\hat{h}$, $\widehat{ERM_1}$, $\widehat{ERM_1}$ and BIC with the SNPS included

the values of $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC. In these cases, we included the SNP'S variables.

We observe some variability in the estimate of $\hat{h}$: it takes values 36, 37, and 41. The smallest discrepancy between the size of the conjectured model and $\hat{h}$ occurs when the size of the conjectured model is 32. In fact, at that point $\hat{h} = 41$, suggests that the sample size may be low for our purpose here (We recall that model selection requires more data than parameter estimation). We also observe some variability in the values of $\widehat{ERM_1}$ and $\widehat{ERM_2}$ as the size of the conjectured model increases. There is a big drop in their value when the size of the conjectured model is 7, and thereafter, they become almost
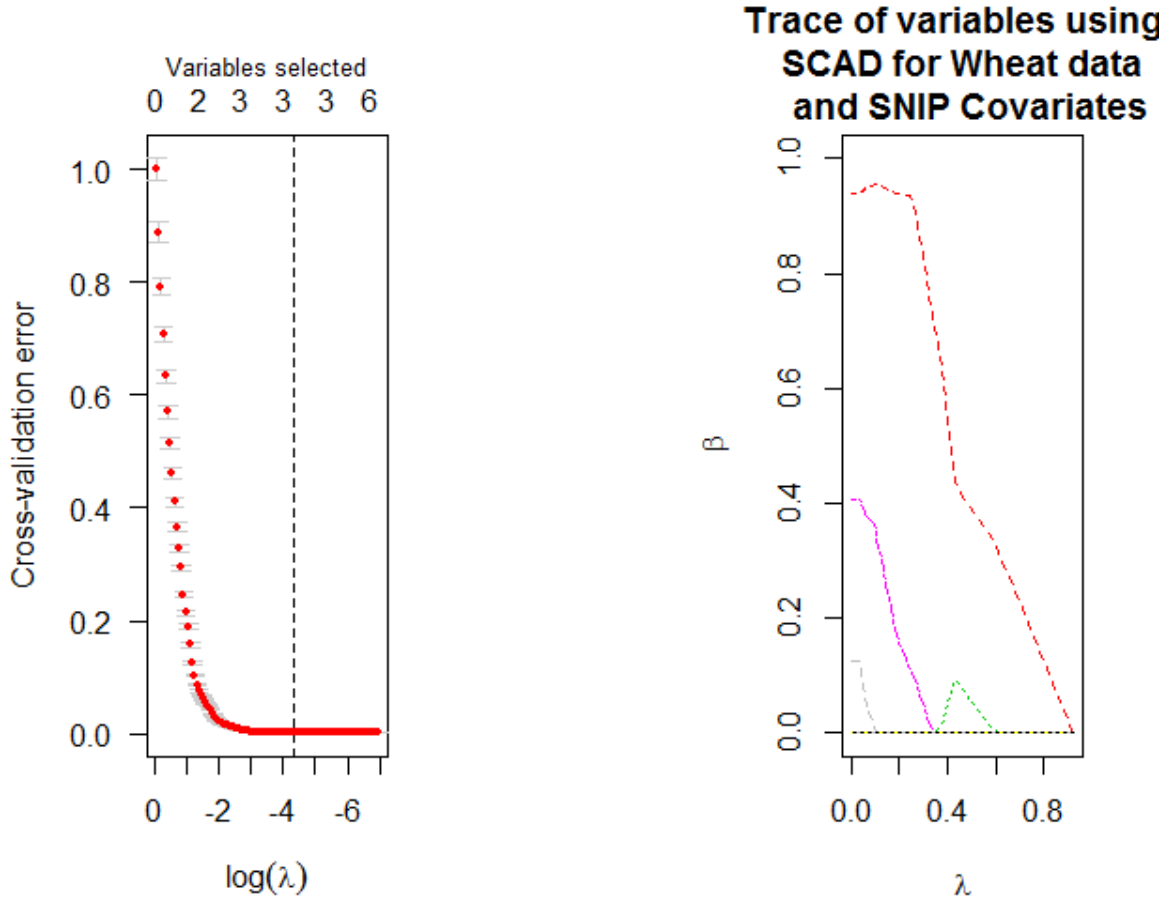
(a) Optimal value of λ for wheat data with SNP variables included

(b) Trace of variables for wheat dataset with SNP variables included using SCAD

Figure 5.12: Analysis of Wheat data with SNP covariates included in the model using SCAD

flat. The smallest value of $\widehat{ERM}_1$, namely eight, and $\widehat{ERM}_2$, namely 16, occur for models of size 7 and 9 respectively. The smallest value of BIC occurs for model of size 8.

Turning our attention to the sparsity methods, we observe from Fig. 5.12a that the optimal value of λ is $\hat{\lambda} = 0.01311611$; this indicates that the optimal number of parameters is 3. From Fig. 5.10b, under SCAD, the variable chosen is $TKWT \cdot KPSM$, $KPSM$ and $TKWT$ and the model can be written as:

$$\widehat{YIELD} = 0.13 \cdot KPSM \cdot TKWT + 0.94 \cdot KPSM + 0.41 \cdot TKWT. \tag{5.9}$$

Similarly, using ALASSO we have,

$$\widehat{YIELD} = 0.13 \cdot KPSM \cdot TKWT + 0.94 \cdot KPSM + 0.41 \cdot TKWT. \tag{5.10}$$

So, (5.10) and (5.9) are identical. We also note that BIC and $\widehat{ERM}_1$ pick the same variables. $\widehat{ERM}_2$ picks a model of size 9 (Large by one term), $\hat{h}$ picks the most complex model, and SCAD and ALASSO

pick a the smallest models, of size 3. We also note that in (5.9) and (5.10) we have the interaction between $TKWT$ and $KPSM$ and their main effects. As noted before, $\hat{h}$ tends to be most sensitive to bias, so it is not surprising that it is the largest and includes all SNP's. On the other hand, the first SNP to appear on the list is $barc67$, 9th on the list. This means SCAD, ALASSO, $\widehat{ERM}_1$ and BIC choose no SNP's. However, $\widehat{ERM}_2$ includes $barc67$, and $\hat{h}$ includes all the SNP's that had no missing values. Moreover, $\hat{h}$ being larger than 32 – the largest model size– suggests that other SNP's that may have been excluded due to missing values, or other missing variables such as variety, may be critical to explain $YIELD$.

1

Chapter 6

General Conclusions and Future Work

In this Dissertation, we have developed a general method to estimate the VCD h for classes of regression models.

Our method rests on minimizing an upper bound on an error. The upper bound depends on h and generates an estimator $\hat{h}$. We have applied our method to a variety of linear regression problems. Our method is highly data driven and at least for linear regression models with large enough sample sizes, seems to give better results on simulated and real datasets in contrast with other established methods. Otherwise put, at a minimum, the models selected by our method are, on balance, no worse and often better than models selected by other methods such as BIC, SCAD, ALASSO, and two forms of empirical risk minimization ($\widehat{ERM}$). The comparison between $\hat{h}$ and the $\widehat{ERM}$'s is not a surprise since the $\widehat{ERM}$'s are function of $\hat{h}$.

In Sec. 6.1, we review what the Dissertation has accomplished, and in Sec. 6.2, we identify gaps that remain to be filled and future work

6.1 General Conclusions

In Chapter 1, we reviewed the Vapnik-Chervonenkis concept of complexity often called a dimension here abbreviated VCD. Since it is relatively unfamiliar to statisticians, we recalled the geometric, combinatorial and covering number definitions of VCD. Then we summarized the work of Vapnik et al. Vapnik et al. (1994) and McDonald et al. McDonald et al. (2011) including the main theoretical results and the computational procedure.

All of this previous work was done for classification problems whereas our work generalizes it to regression problems. This led us to modify substantially not just the theory, but also the computing. Moreover we have had at various junctures to correct simplifying assumptions made by earlier authors so as to obtain good performance of the VCD as a model selection principle.

Chapter 2 presents the main mathematical results undergirding our methodology. Specifically, we convert a generic regression problem into m classifications problems and apply corrected form of Vapnik

et al to each of the m problems. This gives us a bound on the Expected Maximum Difference between Two Empirical Losses (EMDBTEL) in term of h, the correct VCD.

In Chapter 3, computational exploration showed this bound was not tight enough for good model selection. So we optimized the constant factor in the upper bound and changed the error criterion so it ressembled a cross-validation error rather than simply the difference between two models. We verified that the mathematics of Chapter 2 continued to hold with this change of the form of error.

The second major contribution of Chapter 3 was to correct the Vapnik et al Vapnik et al. (1994) computational procedure so it would accurately encapsulate the mathematical quantities the upper bound used. We denoted this estimate of h by $\hat{h}$. Our improved procedure is more computationally demanding. But not terribly so in term of running time. Thus, in our extensive simulations with linear models, we found the new method gave good results compared to Vapnik's original algorithm as well as compared to BIC and two forms of Empirical Risk Minimization propose by Vapnik Vapnik (1998).

In Chapter 3 we also introduced the concept of 'consistency at the true model' to help identify how exactly our method should be used. Indeed, all our method give is an estimate $\hat{h}$ of the VCD of the data generator (DG). To obtain this estimate in practice, we must conjecture the data generator and compare the various $\hat{h}$'s from the DG's to the known true h's from the DG's. Thus using $\hat{h}$ can always cut down a model list to a sub-list although in general $\hat{h}$ does not by itself identify a DG. To get around this problem we used nested model lists where the VCD of the models increased with their size. For these cases, $\hat{h}$ is able to give a unique model, apart from random variability – which can be considerable.

Our used of $\hat{h}$ required the selection of design points to use in a non-linear regression. Different design points lead to different values of $\hat{h}$ as seen in Chapter 3, however, we note a sample size effect: As n increases it seems that

1. the design points matter less and less and

2. the accuracy of $\hat{h}$ for selecting among nested models (with variables ordered for inclusion only by their absolute correlation with the response or by the order of inclusion using SCAD) increases.

In chapter 4, we estimated $\hat{h}$ using the Tour De France Dataset. We fitted two set of models. The first set is based on the following covariates Y, D, D^2, Y^2, and $Y : D$, the second contains Y, D, S, and A (Definitions and notations as in Chapter 4). Whichever of the two model lists is chosen, $\hat{h} = 4$. That is, the best model to predict the average speed of the winner of the Tour De France dataset should have 4 variables. We conclude that the better model list is the one containing Y, D, Y^2, D^2.

Our analysis in this example is exhaustive so we can get an indication of how our method performs relative to BIC and two forms of ERM. We emphasize that although this chapter uses real datasets it is really a sort of 'toy' problem since the sample size n is not large compared to the number of inferences we want to make.

In chapter 5, we implemented our method on more complex data sets. We also performed sparsity driven analyses, using SCAD and ALASSO in addition to estimating $\hat{h}$, and finding $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC. We examined 2 datasets that are commonly regarded as 'difficult' or complex. We started with the Abalone dataset and found $\hat{h} = 9$ no matter the conjectured linear model. However, there were only 7 variables since we only used first order models. We observed that SCAD and ALASSO picked the same variables, but the estimated coefficient for these variables were extremely different from the two models, this may be due to the high correlation between covariates. These models included $Shucked\ weight$ but not $Length$, whereas the models chosen by BIC, $\widehat{ERM}_1$ and $\widehat{ERM}_2$ included $Length$ but not $Shucked\ weight$. That is, the sparsity methods led to the same variables, and $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC used the same variables (albeit a different set) while $\hat{h}$ included all the variables, suggesting that some variables are missing i.e there is unavoidable bias. We regard the model chosen by $\hat{h}$ as the most reasonable because of the non-linear relationship between the response and the explanatory variables.

The second dataset that we used was the wheat dataset. We performed the estimation of $\hat{h}$ using our method by location first and second we combined all locations to perform a multilocation analysis. Overall, we see that complexity varies from one location to another; this can be observed by looking at the estimated $\hat{h}$, and the ordering of the variables in the models. The least complex location is Lincoln 2000 since the $\hat{h} = 1$ for all model lists. We also observe that $TKWT \cdot KPSM$ was the most important variable no matter which location you choose and also had the highest correlation with yield.

Thus, we provide a complete analysis for all 7 locations using the phenotypic variables. Then we turned to two other analyses. The first used the phenotypic variables and the design structure variables, the second used the phenotypic variables and the SNP variables. That is we have shown how our method extends to models permitting two data types. In both cases, we found that the results were more complex that with uni-type datasets and that some SNP's and some design variable were important to include. We could have done multi-type data analysis including phenotypic variables, design variables and SNP variables, but lacked of time to complete this. We note that the Wheat dataset has other variables in it e.g., variety, that we did not use. However, using them might have made the effective sample size too small relative to the model classes we wanted to use.

Over all, we found that our estimated VCD $\hat{h}$ had higher values for real datasets than our comparably sized synthetic datasets indicating $\hat{h}$ was in fact reflecting the complexity of the DG. Moreover, in all cases, $\hat{h}$ gave results that were reasonable even when other methods gave models that were obviously too parsimonious or otherwise implausible. We attribute this to the sensitivity of $\hat{h}$ to bias. By contrast, often other methods have some level of built-in-sparsity so they give smaller models than are reasonable.

As a pragmatic recommendation, we therefore suggest that for complex datasets having a large enough n that the conjecture model to be nested and the model selected by $\hat{h}$ be compared with the model

selected by other methods. That is, we think that no one method for model selection is *the* method to use. We think that judicious comparison of models selected unless different principles (sparsity, small bias, complexity, etc) is the appropriate way to do model selection. In this context, our method is simply another well performing method whose output should be taken into consideration for optimal model building.

6.2 Future Work

No Dissertation can solve all the problems related to a new methodology and this Dissertation is no exception.. There are (at least) five problems that we would like to address in future developing the VCD for model selection.

1. As it stands, $\hat{h}$ come from optimizing an upper bound. It is not clear how to re-do the analysis leading to the upper bound so as to derive a variance for $\hat{h}$. At the present, we can only suggest using a bootstrap approach: form $\hat{h}_1, \hat{h}_2, \cdots, \hat{h}_B$ from B bootstrap samples and then form $\hat{\sigma}_{\hat{h}}$ from them.

2. We have limited our work to linear models although our general theory applies to any class of regression models for which VCD's can be calculated. This includes Trees, Kernel methods, and neural networks, amongst others. These classes are more complicated but being non-linear may give better results in real problems.

3. A theoretical gap in our derivation in Chapter 2 is that we only have an upper bound on the EMDBTEL. We have tried to make it tight computationally, but having either a provably tight upper bound on the EMDBTEL or a provably tight lower bound on the EMDBTEL would make our inferences more convincing.

4. Another theoretical gap is to prove $\hat{h} \to h_{true}$ in probability. McDonald et al McDonald et al. (2011) tried to do this in the classification case but there are some gaps in their proof: For instance, they have a bound for the sum of absolute errors but they want a bound for the absolute error of the sum. We run into the exact same problem when we adapt their proof to regression, see Appendix A.4. However we think we have overcome it.

5. Although, we have shown in an example (see 5.4, 5.5) how our method can be used for two data types with linear models, the general question of how well our method performs in comparison to other methods for multi-type data remains to be explored. The issue is that as more data types become available the modeling becomes more complex and may outstrip $\hat{h}$'s ability to scale up with the number of data types and overall sample sizes.

In Sec. 5.2.2, and 5.3 we compared a location-by-location analysis with pooled data analysis. This is not the same as a multi-type data, but is a facet of the multi-type data problem and shows some of the complexities that may occur.

Obviously, this list is not complete, but it may serve as a warning that however well our method seems to perform in preliminary heats, reasonable questions about it remain to be answered before it is ready for general use.

Nomenclature

ALASSO Adaptive Least Angle Shrinkage Selector Operator

BIC Bayesian Information Criterion

DG Data Generator

EMDBTEL Expected Maximum Difference Between Two Empirical Losses

ERM Empirical Risk Minimization

LHS Left-Hand Side

RHS Right-Hand Side

SCAD Smoothly Clipped Absolute Deviation

VCD Vapnik-Chervonenkis Dimension

Appendix A

CHAPTER 2 APPENDIX

A.1   Proof of Theorem 2.2.1 clause 1

Proof. The LHS of the statement of the theorem equals

$$\int_0^\infty P\left(A_{\epsilon,m}\right) d\epsilon \leq \int_0^u d\epsilon + \int_u^\infty P\left(A_\epsilon\right) d\epsilon \leq u + 2m\left(\frac{2ne}{h}\right)^h \int_u^\infty \exp\left(-\frac{n\epsilon^2}{m^2}\right) d\epsilon. \qquad \text{(A.1)}$$

Observing that

$$\epsilon > u \Rightarrow \frac{n\epsilon^2}{m^2} > \frac{nu\epsilon}{m^2} \Rightarrow \int_u^\infty \exp\left(-\frac{n\epsilon^2}{m^2}\right) d\epsilon$$

$$\leq \int_u^\infty \exp\left(-\frac{nu\epsilon}{m^2}\right) d\epsilon = \frac{m^2}{nu} \exp\left(-\frac{nu^2}{m^2}\right),$$

we have

$$E\left(\sup_{\alpha_1,\alpha_2 \in \Lambda} |\nu_1\left(Z_2, \alpha_1\right) - \nu_2\left(Z_1, \alpha_2\right)|\right) \leq u + 2m^3\left(\frac{2ne}{h}\right)^h \frac{1}{nu} \exp\left(-\frac{nu^2}{m^2}\right). \qquad \text{(A.2)}$$

Let

$$f(u) = u + 2m^3\left(\frac{2ne}{h}\right)^h \frac{1}{nu} \exp\left(-\frac{nu^2}{m^2}\right),$$

$$f'(u) = 1 - 2m^3\left(\frac{2ne}{h}\right)^h \left[\frac{1}{nu^2} + \frac{2}{m^2}\right] \exp\left(-\frac{nu^2}{m^2}\right)$$

$$f'(u) = 0 \Rightarrow \left[\frac{1}{nu^2} + \frac{2}{m^2}\right] \exp\left(-\frac{nu^2}{m^2}\right) = \frac{1}{2m^3}\left(\frac{h}{2ne}\right)^h$$

$$\exp\left(-\frac{nu^2}{m^2}\right) = \frac{1}{2m^3}\left(\frac{h}{2ne}\right)^h \tag{A.3}$$

$$-\frac{nu^2}{m^2} = \ln\left(\frac{1}{2m^3}\left(\frac{h}{2ne}\right)^h\right) \tag{A.4}$$

$$u^2 = \frac{m^2}{n}\ln\left(2m^3\left(\frac{2ne}{h}\right)^h\right) \tag{A.5}$$

$$u = m\sqrt{\frac{1}{n}\ln\left(2m^3\left(\frac{2ne}{h}\right)^h\right)} \tag{A.6}$$

Substituting A.6 in A.2 gives the statement of the proof. $\qquad\square$

A.2 Proof of Theorem 2.2.1 clause 2

We will first state and proof the following Lemma

Lemma A.2.1. *Let $\epsilon \geq 0$. The probability of the supremal difference between two empirical losses using cross-validation form of the error is bounded above by*

$$P\left(\sup_{\alpha_1,\alpha_2\in\Lambda}|\nu_1\left(Z_2,\alpha_1\right)-\nu_2\left(Z_1,\alpha_2\right)|\geq\epsilon\right)$$
$$\leq 2P\left(\sup_{\alpha_1\in\Lambda}|\nu_1\left(Z_1,\alpha_1\right)-E\left(Z_2,\alpha_1\right)|\geq\frac{\epsilon}{2}\right) \tag{A.7}$$

Proof. The LHS is bounded by

$$P\left(\sup_{\alpha_1,\alpha_2\in\Lambda}|\nu_1\left(Z^2,\alpha_1\right)-E\left(Q(Z^2,\alpha_1)\right)+E\left(Q(Z^2,\alpha_1)-\nu_2\left(Z^1,\alpha_2\right)\right)|\geq\epsilon\right)$$
$$\leq\ P\left(\sup_{\alpha_1\in\Lambda}|\nu_1\left(Z^2,\alpha_1\right)-E\left(Q(Z^2,\alpha_1)\right)|\geq\frac{\epsilon}{2}\right)+$$
$$P\left(\sup_{\alpha_2\in\Lambda}|E\left(Q(Z^1,\alpha_2)\right)-\nu_2\left(Z^1,\alpha_2\right)|\geq\frac{\epsilon}{2}\right)$$
$$=\ 2P\left(\sup_{\alpha_1\in\Lambda}|\nu_1\left(Z^2,\alpha\right)-E\left(Q(Z^2,\alpha_1)\right)|\geq\frac{\epsilon}{2}\right).$$

if

$$P\left(\sup_{\alpha_1\in\Lambda}|\nu_1\left(Z^2,\alpha\right)-E\left(Q(Z^2,\alpha_1)\right)|\geq\frac{\epsilon}{2}\right)\geq$$
$$P\left(\sup_{\alpha_2\in\Lambda}|E\left(Q(Z^1,\alpha_2)\right)-\nu_2\left(Z^1,\alpha_2\right)|\geq\frac{\epsilon}{2}\right)$$

$\qquad\square$

Proposition A.2.1. *Let $h = VCDim\{Q\left(\cdot,\alpha\right):\alpha\in\Lambda\}$, where $Q\left(\cdot,\alpha\right)$ is unbounded, $n\in\mathcal{N}$ be the*

sample size. If $h < \infty$, and

$$D_p(\alpha_1) = \int_0^\infty \sqrt[p]{P\{Q(Z_2, \alpha_1) \geq c\}} dc \leq \infty$$

where $1 < p \leq 2$ is some fixed parameter, we have

$$P\left(\sup_{\alpha_1, \alpha_2 \in \Lambda} \left|\nu_1\left(Z^2, \alpha_1\right) - \nu_2\left(Z^1, \alpha_1\right)\right| \geq \epsilon\right) \leq \quad 16\left(\frac{ne}{h}\right)^h \exp\left\{-\left(\frac{\epsilon n^{1-\frac{1}{p}}}{D_p(\alpha_1^*)2^{2.5+\frac{1}{p}}}\right)^2\right\}, \quad \text{(A.8)}$$

where $\alpha_1^* = \sup_{\alpha_1 \in \Lambda} D_p(\alpha_1)$.

Proof. By Lemma A.2.1, we have

$$P\left(\sup_{\alpha_1, \alpha_2 \in \Lambda} \left|\nu_1\left(Z^2, \alpha_1\right) - \nu_2\left(Z^1, \alpha_2\right)\right| \geq \epsilon\right) \leq$$

$$2P\left(\sup_{\alpha_1 \in \Lambda} \left|\nu_1\left(z^2, \alpha_1\right) - E\left(Q\left(z^2, \alpha_1\right)\right)\right| \geq \frac{\epsilon}{2}\right)$$

$$\text{(A.9)}$$

To bound the RHS of A.9, use Lemma 2.1.2 to observe that

$$P\left(\sup_{\alpha \in \Lambda} \left|\nu_n(z, \alpha) - E(Q(z, \alpha))\right| \geq \frac{\epsilon}{2}\right) \leq P\left(\sup_{\alpha \in \Lambda} \left|E\left(Q(z, \alpha)^+\right) - \frac{1}{n}\sum_{i=1}^n Q(z_i, \alpha)^+\right| \geq \frac{\epsilon}{4}\right)$$

$$+ P\left(\sup_{\alpha \in \Lambda} \left|E\left(Q(z, \alpha)^-\right) - \frac{1}{n}\sum_{i=1}^n Q(z, \alpha)^-\right| \geq \frac{\epsilon}{4}\right). \quad \text{(A.10)}$$

Each probability on the right hand side of (A.10) can be bounded. Since α_1^* is a maximum, we have $1/D_p(\alpha_1^*) \leq 1/D_p(\alpha_1)$. Thus, for either the positive or negative parts in (A.10) we have

$$\sup_{\alpha_1 \in \Lambda} \frac{E\left(Q(Z_2, \alpha_1)\right) - \nu_1(Z_2, \alpha_1)}{D_p(\alpha_1^*)}$$

$$\leq \sup_{\alpha_1 \in \Lambda} \frac{E\left(Q(Z_2, \alpha_1)\right) - \nu_1(Z_2, \alpha_1)}{D_p(\alpha_1)}$$

$$P\left(\sup_{\alpha_1 \in \Lambda} \frac{E\left(Q(Z_2, \alpha_1)\right) - \nu_1(Z_2, \alpha_1)}{D_p(\alpha_1^*)} \geq \frac{\epsilon}{4D_p(\alpha_1^*)}\right)$$

$$\leq P\left(\sup_{\alpha_1 \in \Lambda} \frac{E\left(Q(Z_2, \alpha_1)\right) - \nu_1(Z_2, \alpha_1)}{D_p(\alpha_1)} \geq \frac{\epsilon}{4D_p(\alpha_1^*)}\right) \Rightarrow$$

$$P\left(\sup_{\alpha_1 \in \Lambda} \left(E\left(Q(Z_2, \alpha_1)\right) - \nu_1(Z_2, \alpha_1)\right) \geq \frac{\epsilon}{4}\right)$$

$$\leq P\left(\sup_{\alpha_1 \in \Lambda} \frac{E\left(Q(Z_2, \alpha_1)\right) - \nu_1(Z_2, \alpha_1)}{D_p(\alpha_1)} \geq \frac{\epsilon}{4D_p(\alpha_1^*)}\right).$$

Letting $\delta = \frac{\epsilon}{4D_p(\alpha^*)}$ and using Theorem 2.1.1, Clause 3, the last inequality gives

$$P\left(\sup_{\alpha_1 \in \Lambda} \left(E\left(Q(Z_2, \alpha_1)\right) - \nu_1(Z_2, \alpha_1)\right) \geq \frac{\epsilon}{4}\right) \leq 4\exp\left\{\left(\frac{H_{ann}^{\Lambda,\beta}(n)}{n^{2-\frac{2}{p}}} - \frac{\delta^2}{2^{1+\frac{2}{p}}}\right)n^{2-\frac{2}{p}}\right\}. \qquad (A.11)$$

Using (A.11), both terms on the right in (A.10) can be bounded. This gives

$$P\left(\sup_{\alpha_1 \in \Lambda} |\nu_1(Z_2, \alpha_1) - E\left(Q(Z_2, \alpha_1)\right)| \geq \frac{\epsilon}{2}\right) \leq 8\exp\left\{\left(\frac{H_{ann}^{\Lambda,\beta}(n)}{n^{2-\frac{2}{p}}} - \frac{\delta^2}{2^{1+\frac{2}{p}}}\right)n^{2-\frac{2}{p}}\right\}. \qquad (A.12)$$

The presence of β in the the exponent of the hannealed entropy will not change the validity of Clause I of Theorem 2.1.1, so we have, $H_{ann}^{\Lambda,\beta}(n) \leq G^{\Lambda}(n) \leq \ln\left(\frac{en}{h}\right)^h$ therefore, $\exp\left(H_{ann}^{\Lambda,\delta}(n)\right) \leq \exp\left(G(n)\right) \leq \left(\frac{en}{h}\right)^h$. Using this in (A.12) gives

$$P\left(\sup_{\alpha_1 \in \Lambda} \left(\nu_1(Z_2, \alpha_1) - E\left(Q(Z_2, \alpha_1)\right)\right) \geq \frac{\epsilon}{2}\right) \leq 8\left(\frac{ne}{h}\right)^h \exp\left\{-\left(\frac{\delta n^{1-\frac{1}{p}}}{2^{0.5+\frac{1}{p}}}\right)^2\right\}$$

$$= 8\left(\frac{ne}{h}\right)^h \exp\left\{-\left(\frac{\epsilon n^{1-\frac{1}{p}}}{D_p(\alpha_1^*)2^{2.5+\frac{1}{p}}}\right)^2\right\}.$$

Recalling the extra factor of 2 in inequality (A.9) gives the statement of the Proposition. $\qquad \square$

Now let prove Theorem 2.2.1 clause 2.

Proof. Using the integral of probabilities identity and Proposition A.2.1, the left-hand side of (2.36) equal

$$\int_0^\infty P\left(\sup_{\alpha_1,\alpha_2 \in \Lambda} |\nu_1(Z_2, \alpha_1) - \nu_2(Z_1, \alpha_2)| \geq \epsilon\right) d\epsilon$$

$$\leq \int_0^\infty 16\left(\frac{ne}{h}\right)^h \exp\left\{-\left(\frac{\epsilon n^{1-\frac{1}{p}}}{D_p(\alpha_1^*)2^{2.5+\frac{1}{p}}}\right)^2\right\} d\epsilon$$

$$\equiv \int_0^u d\epsilon + 16\left(\frac{ne}{h}\right)^h \int_u^\infty \exp\left\{-\left(\frac{\epsilon n^{1-\frac{1}{p}}}{D_p(\alpha_1^*)2^{2.5+\frac{1}{p}}}\right)^2 \epsilon^2\right\} d\epsilon$$

$$\leq u + 16\left(\frac{ne}{h}\right)^h \int_u^\infty \exp\left\{-\left(\frac{n^{1-\frac{1}{p}}}{D_p(\alpha_1^*)2^{2.5+\frac{1}{p}}}\right)^2 u\epsilon\right\} d\epsilon$$

$$= u + \left(\frac{D_p(\alpha_1^*)2^{2.5+\frac{1}{p}}}{n^{1-\frac{1}{p}}}\right)^2 \frac{16\left(\frac{ne}{h}\right)^h}{u} \exp\left\{-\left(\frac{un^{1-\frac{1}{p}}}{D_p(\alpha_1^*)2^{2.5+\frac{1}{p}}}\right)^2\right\}.$$

Choosing

$$u = D_p(\alpha_1^*)2^{2.5+\frac{1}{p}} \frac{\sqrt{h\ln\left(\frac{ne}{h}\right)}}{n^{1-\frac{1}{p}}}$$

gives the statement of the Theorem. $\qquad \square$

A.3 Proof of Theorem 2.2.1 clause 3

Proof. Theorem 2.2.1 clause 1 gives

$$
E\left(\sup_{\alpha_1,\alpha_2 \in \Lambda} |\nu_1^*(Z_2,\alpha_1,m) - \nu_2^*(Z_1,\alpha_2,m)|\right) \leq \sqrt{\frac{\ln(2m)}{n} + \frac{h}{n}\ln\left(\frac{2ne}{h}\right)}
$$

$$
\stackrel{\infty}{\cong} \sqrt{\frac{h}{n}\ln\left(\frac{2ne}{h}\right)}, \tag{A.13}
$$

where $\stackrel{\infty}{\cong}$ indicates a limit as $n \to \infty$ has been taken. Similarly, Theorem 2.2.1 clause 2 gives

$$
E\left(\sup_{\alpha_1,\alpha_2 \in \Lambda} |\nu_1(Z_2,\alpha_1) - \nu_2(Z_1,\alpha_2))|\right) \leq 8D_p(\alpha_1^*)\sqrt{\frac{h}{n} + \frac{h}{n}\ln\left(\frac{2ne}{h}\right)} + E(n,h) \tag{A.14}
$$

where $E(n,h)$ is of smaller order than the first term on the right in 2.23. Thus,

$$
E\left(\sup_{\alpha_1,\alpha_2 \in \Lambda} |\nu_1(Z_2,\alpha_1) - \nu_2(Z_1,\alpha_2))|\right) \leq 8D_p(\alpha_1^*)\sqrt{\frac{h}{n}\ln\left(\frac{2ne}{h}\right)}. \tag{A.15}
$$

Taking the minimum over the RHS of equations (A.13) and (A.15) gives the Theorem. $\qquad\square$

A.4 Proof of Consistency

Here, we offer a proof of consistency of $\hat{h}$ to h. In almost all respects this tentative of proof should be credited to McDonald et al. (2011). Our contribution is the examination of the gaps in their proof. Our two ugly hypotheses and the discussion about how to satisfy them fill the gaps in their proof. We have put it in as appendix rather than the main part of the thesis because we have not finished verifying all the details satisfactorily. Instead, we have not convince ourselves it is error free. Nevertheless, we have the following. ϵ_i's represent the ξ_i's in Algorithm 2

Lemma A.4.1. *Let $t \geq 0$ and let $\epsilon(n_l) = \frac{1}{W}\sum_{i=1}^{W}\epsilon_i(n_l)$, where $\epsilon_i(n_l)$'s are independent for any given n_l's. Then at any design point n_l, we have*

$$
E\left(\exp\left(t\epsilon_i(n_l)\right)\right) \leq \exp\left(\frac{t^2 B^2}{8Wm^2}\right) \tag{A.16}
$$

Proof. Fix $j = 0,1,2,\cdots,m-1$. Now, for any $t \geq 0$ and $i \geq 0$, and $\epsilon_i(n_l) \in \left(\frac{jB}{m}, \frac{(j+1)B}{m}\right]$, using Hoeffding's lemma we have

$$
E\left(\exp\left(t\epsilon_i(n_l)\right)\right) \leq \exp\left(\frac{t^2 B^2}{8m^2}\right).
$$

Thus, we have

$$
\begin{aligned}
E\left(\exp\left(t\epsilon_i\left(n_l\right)\right)\right) &= E\left(\exp\left(\frac{t}{W}\sum_{i=1}^{W}\epsilon_i\left(n_l\right)\right)\right) \\
&= E\left(\Pi_{i=1}^{W}\exp\left(\frac{t}{W}\epsilon_i\left(n_l\right)\right)\right) \\
&= \Pi_{i=1}^{W}E\left(\exp\left(\frac{t}{W}\epsilon_i\left(n_l\right)\right)\right) \\
&\leq \Pi_{i=1}^{W}\exp\left(\frac{t^2 b^2}{8Wm^2}\right) \\
&= \exp\left(\frac{t^2 B^2}{8Wm^2}\right)
\end{aligned}
\tag{A.17}
$$

$\square$

Proposition A.4.1. *Suppose that* $\{\epsilon_l \equiv \epsilon\left(n_l\right), l = 1, 2 \cdots k\}$ *is a set of random variables satisfying Lemma A.4.1. Then for any* $\gamma \in \mathbb{R}^k$ *and* $\rho \geq 0$, *we have*

$$
P\left(|\sum_{l=1}^{k}\epsilon_l\gamma_l| \geq \rho\right) \leq 2\exp\left(-\frac{2Wm^2\rho^2}{B^2\sum_{l=1}^{k}\gamma_l^2}\right)
\tag{A.18}
$$

Proof.

$$
\begin{aligned}
P\left(|\sum_{l=1}^{k}\epsilon_l\gamma_l| \geq \rho\right) &\leq 2P\left(\sum_{l=1}^{k}\epsilon_l\gamma_l \geq \rho\right) \\
&= 2P\left(\exp\left(t\sum_{l=1}^{k}\epsilon_l\rho_l\right) \geq \exp(t\rho)\right) \\
&\leq \frac{2E\left(\exp\left(t\sum_{l=1}^{k}\epsilon_l\rho_l\right)\right)}{\exp(t\rho)} \\
&= \frac{2E\left(\Pi_{l=1}^{k}\exp\left(t\epsilon_l\gamma_l\right)\right)}{\exp(t\rho)} \\
&= \frac{2\Pi_{l=1}^{k}E\left(\exp\left(t\epsilon_l\gamma_l\right)\right)}{\exp(t\rho)} \\
&\leq \frac{2\Pi_{l=1}^{k}\exp\left(\frac{t^2 B^2\gamma_l^2}{8Wm^2}\right)}{\exp(t\rho)} \\
&= 2\exp\left(\frac{t^2 B^2}{8Wm^2}\sum_{l=1}^{k}\gamma_l^2 - t\rho\right)
\end{aligned}
\tag{A.19}
$$

Since A.19 is true for all t, the RHS of A.19 attains its minimum for

$$
t = \frac{8Wm^2\rho}{2B^2\sum_{l=1}^{k}\gamma_l^2}
\tag{A.20}
$$

Replacing t by its value in A.19 gives the statement of the proposition. $\square$

Definition A.4.1. $\Phi = \{\phi_h : h \in (1, M]\}$. For $s = 0, 1, 2, \cdots$, let $\{\phi_j^s\}_{j=1}^{N_s}$ be the minimal $2^{-s}\tau$-covering set of $(\Phi, \| \cdot \|_Q)$. So $N_s = N(2^{-s}\tau, \Phi)$ and for each h, there exists a $\phi_h^s \in \{\phi_1^s, \phi_2^s, \cdots, \phi_{N_s}^s\}$ such that $\| \phi_h - \phi_h^s \|_Q \leq 2^{-s}\tau$. We choose $\phi_h^0 \equiv 0$, since $\| \phi_h \|_Q \leq \tau$. The key argument is based on the entropy of the restricted class $\Phi(\tau) = \{\phi_h \in \Phi : \| \phi_h - \phi_{h_T} \|_Q \leq \tau\}$. $\phi_h(n)$ is defined as follows:

$$\phi_h(n) = \hat{c}\sqrt{\frac{h}{n}\log\left(\frac{2ne}{h}\right)} \tag{A.21}$$

We assume without any proof that the following theorem is true.

Theorem A.4.1. *Let assume that we have a collection of functions inside $Im(\Phi)$ so that $\| \phi_h - \phi_h^s \|_2 \leq 2^{-s}\tau \implies \| \phi_h - \phi_h^s \|_1 \leq 2^{-s}c' f(\tau)$, where $f(\tau)$ is an increasing function of τ.*

Corollary A.4.1. *If $\sup_{\phi_h \in \Phi(\tau)} \| \phi_h \|_Q \leq \tau$ and the conclusion of Lemma A.4.1 holds for all design points n_l, then for*

$$\delta \geq \frac{2B\tau}{m\sqrt{kW}}(288\log(2))^{0.5}, \tag{A.22}$$

we have

$$P\left(\sup_{\phi_h \in \Phi}\left|\frac{1}{k}\sum_{l=1}^k \epsilon_l \phi_h(n_l)\right| \geq \delta\right) \leq 4\exp\left(-\frac{kWm^2\delta^2}{1152B^2\tau^2}\right). \tag{A.23}$$

Proof.

$$\left|\frac{1}{k}\sum_{l=1}^k \epsilon_l\left(\phi_h^s - \phi_s^{s-1}\right)\right| = \frac{1}{k}\sum_{l=1}^k sign(\epsilon_l)|\epsilon_l|sign\left(\phi_h^s - \phi_h^{s-1}\right)\left|\left(\phi_h^s - \phi_h^{s-1}\right)\right| \tag{A.24}$$

$$\left|\frac{1}{k}\sum_{l=1}^k \epsilon_l\left(\phi_h^s - \phi_s^{s-1}\right)\right| \leq \frac{1}{k}\sum_{l=1}^k \epsilon_l sign\left(\phi_h^s - \phi_h^{s-1}\right)\left(\phi_h^s - \phi_h^{s-1}\right) \tag{A.25}$$

So since the ϕ_h and ϕ_h^s are continuous, and if N_s is large enough, using Def. A.4.1 and Theorem A.4.1 we have

$$\left|\phi_h^s - \phi_h^{s-1}\right| \leq 3\tau 2^{-s}. \tag{A.26}$$

Using A.26, we have

$$-3\tau 2^{-s} \leq \phi_h^s - \phi_h^{s-1} \leq 3\tau 2^{-s} \implies$$

$$-3\tau 2^{-s}\epsilon_l \leq \left(\phi_h^s - \phi_h^{s-1}\right)\epsilon_l \leq 3\tau 2^{-s}\epsilon_l \implies$$

$$\left| \frac{1}{k} \sum_{l=1}^{k} \epsilon_l \left(\phi_h^s - \phi_h^{s-1} \right) \right| \leq \frac{3 \cdot 2^{-s} \tau}{k} \left| \sum_{l=1}^{k} \epsilon_l \right| \tag{A.27}$$

Let η_s be a positive number satisfying $\sum_{s=1}^{S} \eta_s \leq 1$, and using A.27, we have

$$P \left(\sup_{h \in [1,M]} \left| \frac{1}{k} \sum_{s=1}^{S} \sum_{l=1}^{k} \epsilon_l \left(\phi_h \left(n_l \right) - \phi_h \left(n_l \right) \right) \right| \geq \delta \right)$$

$$\leq \sum_{s=1}^{S} P \left(\sup_{h \in [1,M]} \frac{3 \cdot 2^{-s} \tau}{k} \left| \sum_{l=1}^{k} \epsilon_l \right| \geq \frac{\delta \eta_s}{2} \right)$$

$$= \sum_{s=1}^{S} P \left(\left| \sum_{l=1}^{k} \epsilon_l \right| \geq \frac{k \delta \eta_s}{3 \cdot 2^{-s+1} \tau} \right)$$

$$\leq 2 \sum_{s=1}^{S} \exp \left(-\frac{2 W m^2 k^2 \delta^2 \eta s^2}{9 \cdot B^2 k \cdot 2^{-2s+2} \tau^2} \right)$$

$$= 2 \sum_{s=1}^{S} \exp \left(-\frac{2 m^2 W k \delta^2 \eta^2}{9 B^2 \cdot 2^{-2s+2} \tau^2} \right)$$

Next, $\eta_s \geq \frac{2^{-s} \sqrt{s}}{8} \implies$

$$-\frac{2 m^2 W k \delta^2 \eta_s^2}{9 B^2 2^{-2s+2} \tau^2} \leq -\frac{W k \delta^2 m^2 s}{1152 B^2 \tau^2} \tag{A.28}$$

Using A.28 in the previous inequality, we have

$$2 \sum_{s=1}^{S} \exp \left(-\frac{2 m^2 W k \delta^2 \eta^2}{9 B^2 \cdot 2^{-2s+2} \tau^2} \right)$$

$$\leq 2 \sum_{s=1}^{S} \exp \left(-\frac{W k \delta^2 m^2 s}{1152 B^2 \tau^2} \right) \tag{A.29}$$

$$= \frac{2 \exp \left(-\frac{W k \delta^2 m^2}{1152 B^2 \tau^2} \right)}{1 - \exp \left(-\frac{W k \delta^2 m^2}{1152 B^2 \tau^2} \right)}$$

Using A.22, we have $\frac{m^2 \delta^2 k W}{1152 B^2 \tau^2} \geq \log(2) \implies \exp \left(-\frac{m^2 \delta^2 k W}{1152 B^2 \tau^2} \right) \leq \frac{1}{2} \implies$

$$P \left(\sup_{\phi_h \in \Phi} \left| \frac{1}{k} \sum_{l=1}^{k} \epsilon_l \phi_h \left(n_l \right) \right| \geq \delta \right) \leq 4 \exp \left(-\frac{k W m^2 \delta^2}{1152 B^2 \tau^2} \right). \tag{A.30}$$

$\square$

Theorem A.4.2. *Let*

$$\delta \geq \frac{2B}{m \sqrt{kW}} \left(288 \log(2) \right)^{0.5} \tag{A.31}$$

and suppose that $h_T \in (0, M]$. Then,

$$P \left(\| \phi_{\hat{h}} - \phi_{h_T} \|_Q \geq \delta \right) \leq \frac{28}{3} \exp \left(-\frac{kWm^2\delta^2}{4608B^2} \right). \tag{A.32}$$

Proof. Note that $\| \phi_{\hat{h}} - \phi_{h_T} \|_Q \leq \frac{2}{k} \sum_{l=1}^{k} \epsilon_l \left(\phi_{\hat{h}}(n_l) - \phi_{h_T}(n_l) \right)$. using this last inequality, we have

$$\begin{aligned} &P \left(\| \phi_{\hat{h}} - \phi_{h_T} \|_Q \geq \delta \right) \\ &\leq P \left(2 \left| \frac{1}{k} \sum_{l=1}^{k} \epsilon_l \left(\phi_{\hat{h}} - \phi_{h_T} \right) \right| \geq \delta \right) \\ &\leq P \left(\sup_{\phi_h \in \Phi(2^{s+1}\delta)} \left| \frac{1}{k} \sum_{s=1}^{S} \sum_{l=1}^{k} \epsilon_l \left(\phi_{\hat{h}}^s(n_l) - \phi_{h_T}^{s-1}(n_l) \right) \right| \geq 2^{2s-1}\delta^2 \right) \\ &\leq \sum_{s=1}^{S} P \left(\sup_{\phi_h \in \Phi(2^{s+1}\delta)} \left| \frac{1}{k} \sum_{l=1}^{k} \epsilon_l \left(\phi_{\hat{h}}(n_l) - \phi_{h_T}(n_l) \right) \right| \geq 2^{2s-1}\delta^2 \right) \end{aligned} \tag{A.33}$$

using corollary A.4.1, we have

$$\begin{aligned} &\sum_{s=1}^{S} P \left(\sup_{\phi_h \in \Phi(2^{s+1}\delta)} \left| \frac{1}{k} \sum_{l=1}^{k} \epsilon_l \left(\phi_{\hat{h}}(n_l) - \phi_{h_T}(n_l) \right) \right| \geq 2^{2s-1}\delta^2 \right) \\ &\leq \sum_{s=1}^{S} 4 \exp \left(-\frac{kWm^2 2^{4s-2}\delta^4}{1152B^2 2^{2s+2}\delta^2} \right) \\ &= 4 \sum_{s=1}^{S} \exp \left(-\frac{kWm^2\delta^2 2^{2s-4}}{1152B^2} \right) \\ &= 4 \exp \left(-\frac{kWm^2\delta^2}{4 \cdot 1152B^2} \right) + 4 \left(1 - \exp \left(-\frac{2kwm^2\delta^2}{1152B^2} \right) \right)^{-1} \exp \left(-\frac{2kWm^2\delta^2}{1152B^2} \right). \end{aligned} \tag{A.34}$$

Using A.31, we have

$$\exp \left(-\frac{2kWm^2\delta^2}{1152} \right) \leq \frac{1}{4}. \tag{A.35}$$

Substituting A.35 in A.34 gives

$$\begin{aligned} &4 \exp \left(-\frac{kWm^2\delta^2}{4 \cdot 1152B^2} \right) + 4 \left(1 - \exp \left(-\frac{2kWm^2\delta^2}{1152B^2} \right) \right)^{-1} \exp \left(-\frac{2kWm^2\delta^2}{1152B^2} \right) \\ &\leq 4 \exp \left(-\frac{kWm^2\delta^2}{4 \cdot 1152B^2} \right) + \frac{16}{3} \exp \left(-\frac{2 \cdot kWm^2\delta^2}{1152B^2} \right). \end{aligned} \tag{A.36}$$

Since the first exponential is the largest, this completes the proof. $\qquad \square$

Appendix B

CHAPTER 3 APPENDIX

c

Figures 3.17 and 3.20 give the estimates of $\hat{h}$ for larger sample size and N_L. Observations on these figures are qualitatively the same as in Figures 3.5, and 3.15b when the size of the conjectured model is strictly less than that of the true model. However, if the conjectured model exactly matches the true model, $\hat{h}$ is close to h_T. We also observe that we can discriminate between models using $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_1$ or BIC.

Table B.1: Estimation of $\hat{h}, \sigma_\epsilon = 0.4$, $m = 10$, $p = 15$, $N = 400$, $N_L = 50, 100, 150, 200,$ 250, 300, 350, 400

Model Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
11	50	74945.41	76872.40	3306.22
12	50	53641.17	55274.12	3178.42
13	50	46353.32	47872.53	3126
14	50	21047.24	22076.68	2816.11
15	16	66.87	110.64	493.02
16	25	68.61	123.44	504.88
17	25	68.30	123.03	508.88
18	26	68.35	124.15	514.66
19	27	68.37	125.21	520.29
20	25	67.97	122.59	524.77

Table B.2: Estimation of $\hat{h}, \sigma_\epsilon = 0.4$, $m = 10$, $p = 30$, $N = 400$, $N_L = 50, 100, 150,$
$200, 250, 300, 350, 400$

Model Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
26	50	116360.87	118757.68	3572.08
27	50	102187.23	104434.42	3526.1150
28	50	90273.38	92386.57	3482.52
29	50	6583.36	7166.91	2440.7975
30	30	66.30	125.44	571.34
31	42	82.38	150.10	684.47
32	39	82.21	147.48	690.63
33	40	82.09	148.14	695.71
34	41	81.84	148.5	699.91
35	42	81.58	149.03	704.01
36	38	80.61	144.55	705.24
37	38	79.60	143.22	704.67
38	40	79.61	144.85	710.20

Table B.3: Estimation of $\hat{h}, \sigma_\epsilon = 0.4$, $m = 10$, $p = 40$, $N = 600$, $N_L = 75, 150, 225, 300,$
$375, 450, 525, 600$

Model Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
36	75	37499.03	38860.06	4426.78
37	75	15313.82	16189.85	3895.68
38	75	15270.06	16144.87	3900.36
39	75	5583.17	6119.06	3302.60
40	41	97.90	163.4084	848.55
41	48	97.8	167.75	852.33
42	50	97.88	169.21	858.44
43	46	97.69	166.14	864.80
44	48	97.73	167.61	870.81
45	50	97.80	169.10	877.80
46	47	97.65	166.80	883.42
47	45	97.55	165.22	889.75

Table B.4: Estimation of $\hat{h}, \sigma_\epsilon = 0.4$, $m = 10$, $p = 50$, $N = 600$, $N_L = 75, 150, 225, 300,$
$375, 450, 525, 600$

Model Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
46	75	63643.52	65411.39	4808.19
47	75	44226.92	45703.52	4596.18
48	75	34270.58	35572.47	4449.52
49	75	29017.58	30216.93	4356.06
50	**48**	97.26	**167.71**	**906.15**
51	55	97.28	172.60	911.16
52	53	97.40	171.16	917.57
53	52	97.36	170.43	923.96
54	52	96.94	170.16	929.03
55	52	96.86	169.99	934.57
56	50	**96.65**	168.49	940.81
57	51	**96.65**	169.22	947.18

Table B.5: Estimation of $\hat{h}$, $\sigma_\epsilon = 0.4$, $m = 10$, $p = 60$, $N = 600$, $N_L = 75, 150, 225,$ 300, 375, 450, 525, 600

Model Size	$\hat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
56	75	55967.97	57626.89	4795.03
57	75	52809.85	54421.78	4766.58
58	75	26574.66	27723.16	4360.86
59	75	20887.67	21907.86	4222.73
60	48	96.05	166.14	962.10
61	53	95.72	169.13	964.82
62	51	95.62	167.65	971.13
63	50	95.57	166.90	977.47
64	48	95.48	165.39	983.85
65	44	95.29	162.29	990.25
66	75	96.41	183.47	995.92
67	75	96.41	183.47	1002.32

Table B.6: Estimation of $\hat{h}$, $\sigma_\epsilon = 0.4$, $m = 10$, $p = 60$, $N = 700$, $N_L = 100, 200, 300,$ 400, 500, 600, 700

Model Size	$\hat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
51	100	287528.97	291446.08	6546.30
52	100	287492.75	291409.61	6552.77
53	100	265973.71	269741.87	6504.86
54	100	186316.18	189473.13	6262.23
55	100	184015.42	187152.93	6260.08
56	100	139781.55	142518.55	6074.17
57	100	139329.81	142062.42	6078.45
58	100	62718.39	64558.12	5526.23
59	100	4337.61	4835.85	3661.76
60	57	109.96	184.46	1058.24
61	66	109.51	189.27	1059.55
62	68	109.56	190.48	1065.97
63	63	108.96	186.81	1069.64
64	66	109.06	188.69	1076.16
65	67	108.83	188.96	1080.88
66	65	108.75	187.70	1087.34
67	62	108.64	185.80	1093.87
68	66	108.76	188.30	1100.30

Table B.7: Estimation of $\hat{h}$, $\sigma_\epsilon = 0.4$, $m = 10$, $p = 60$, $N = 2000$, $N_L = 500, 700, 1000, 1500, 2000$

Model Size	$\hat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
56	500	INF	INF	15177.13
57	500	INF	INF	15068.43
58	500	INF	INF	13705.88
59	500	INF	INF	13234.45
60	59	307.79	381.07	2377.64
61	69	307.65	386.20	2382.83
62	69	307.30	385.80	2388.11
63	69	307.17	385.66	2394.87
64	70	307.12	386.14	2402.16
65	70	307.12	386.13	2409.73

Table B.8: Estimation of $\hat{h}, \sigma_\epsilon = 0.4$, $m = 10$, $p = 60$, $N = 2000$, $N_L = 100, 200, 300,$
$400, 500, 600, 700$ by 100

Model Size	$\hat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
52	100	823161.98	827589.97	18126.20
53	100	762207.66	766468.82	17979.94
54	100	532856.53	536420.52	17271.60
55	100	526203.77	529745.48	17254.07
56	100	400283.75	403373.66	16714.64
57	100	399003.71	402088.69	16715.83
58	100	180012.14	182086.56	15131.51
59	97	12722.63	13272.58	9839.12
60	56	307.63	379.31	2377.64
61	70	307.63	386.72	2382.83
62	65	307.15	383.65	2388.11
63	68	307.10	385.13	2394.87
64	63	306.93	382.35	2402.16
65	66	307.00	383.99	2409.73
66	68	306.94	384.95	2416.62
67	65	306.84	383.30	2424.04
68	67	306.81	384.30	2431.13

Table B.9: Estimation of $\hat{h}, \sigma_\epsilon = 0.4$, $m = 10$, $p = 60$, $N = 2000$, $N_L = 75, 150, 225, 300,$
$375, 450, 525, 600$

Model Size	$\hat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
52	75	804788.23	808699.78	18081.06
53	75	639641.54	643129.29	17629.31
54	75	378379.55	381063.24	16586.88
55	75	293689.27	296054.24	16087.73
56	75	185570.32	187451.28	15177.13
57	75	175086.88	176914.08	15068.43
58	75	88255.29	89554.04	13705.88
59	75	69458.37	70611.13	13234.45
60	50	307.46	375.63	2377.64
61	57	307.2945	379.4994	2382.83
62	55	306.8884	377.9203	2388.11
63	54	306.7329	377.1755	2394.87
64	52	306.6282	375.8976	2402.16
65	52	306.6232	375.8919	2409.73
66	50	306.4577	374.5224	2416.62
67	48	306.3705	373.2136	2424.04
68	46	306.2312	371.8198	2431.13

Table B.10: Estimation of $\hat{h}, \sigma_\epsilon = 0.4$, $m = 10$, $p = 70$, $N = 700$,$N_L = 100$ 200, 300 400, 500, 600, 700

Model Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
64	100	202138.73	205426.18	6378.25
65	100	176781.61	179857.21	6290.97
66	100	77787.73	79834.40	5729.38
67	100	41460.77	42960.20	5295.41
68	100	32595.39	33927.07	5133.53
69	63	1255.65	1486.44	2857.24
70	61	108.26	184.72	1111.20
71	68	107.47	187.76	1110.59
72	67	107.43	187.14	1117.12
73	66	107.27	186.36	1122.78
74	70	107.35	188.72	1128.94
75	66	107.12	186.17	1134.86
76	68	107.18	187.39	1141.37
77	67	107.12	186.74	1147.70
78	65	106.93	185.34	1153.41
79	64	106.82	184.63	1159.50
80	64	106.63	184.38	1164.70

Table B.11: Estimation of $\hat{h}, \sigma_\epsilon = 0.4$, $m = 10$, $p = 70$, $N = 2000$, $N_L = 500, 1000, 1500, 2000$

Model Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
66	500	INF	INF	15606.13
67	500	INF	INF	14349.52
68	500	INF	INF	13875.20
69	71	3483.22	3739.56	7322.59
70	71	306.38	385.8239	2442.84
71	78	307.03	389.93	2453.66
72	79	307.05	390.42	2461.26
73	79	307.02	390.38	2468.67
74	79	307.02	390.38	2476.26
75	80	307.02	390.86	2483.763
76	80	306.90	390.71	2490.527
77	80	306.88	390.69	2498.011

Table B.12: Estimation of $\hat{h}, \sigma_\epsilon = 0.8$, $m = 10$, $p = 30$, $N = 400$, $N_L = 50$, 100, 150, 200, 250, 300, 350, 400

| Model Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
20	50	37390.03	38756.26	3082
21	50	33214.76	34503.46	3040.59
22	50	32224.59	33494.20	3034.468
23	50	28807.88	30009.25	2995.62
24	50	28072.67	29258.84	2991.26
15	50	27402.22	28574.35	2987.58
26	50	21563.20	22604.96	2897.69
27	50	18187.27	19145.45	2835.56
28	50	15987.28	16886.74	2789.95
29	50	7172.88	7781.21	2475.14
30	28	247.65	347.37	1125.86
31	37	249.00	361.80	1133.08
32	36	248.50	359.86	1138.34
33	36	247.53	358.70	1142.73
34	35	246.90	356.61	1147.79
35	35	246.90	356.61	1153.78
36	35	246.44	356.06	1159.01
37	33	246.13	352.92	1164.70
38	32	245.56	350.83	1169.86
39	32	243.69	348.60	1172.71
40	31	243.43	346.85	1178.37

Appendix C

# CHAPTER 5 APPENDIX

Analysis of wheat data in Lincoln 2000

From Figs. C.1a and C.1b, we see a linear relationship between $YIELD$ and $KPSM$ and $YIELD$ and $SPSM$ respectively; there are some observations that are little bit a way from the bulk, but they seem few enough and don't appear to be strongly outlying. Hence we do not exclude them. From Figs. C.1c to C.1f, there does not appear to be any non-trivial relationship between the response and the phenotypic covariates. The observations are scattered and do not appear to have any useful pattern.

We implement our technique on the Lincoln 2000 data and compare the result to other model selection methods such as SCAD Fan and Li (2001), ALASSO Zou (2006b) and BIC. We use correlation Fan and Lv (2008) between the response ($YIELD$) and the phenotype data to order the inclusion of the phenotype variables and their products in the model. There are six linear terms, six squared terms 15 rectangular terms, a total of 27 explanatory variables. Under correlation, the order of inclusion is: $TKWT \cdot KPSM, KPSM \cdot TSTWT, KPSM \cdot HT, SPSM \cdot KPS, KPSM, KPSM^2, SPSM \cdot TKWT, KPSM \cdot SPSM, KPSM \cdot KPS, SPSM \cdot HT, SPSM, SPSM \cdot TSTWT, SPSM^2, HT, HT^2, HT \cdot TSTWT, TKWT \cdot HT, KPS^2, KPS \cdot TSTWT, TSTWT, TSTWT^2, KPS, KPS \cdot TKWT, KPS \cdot HT, TKWT, TKWT^2, TKWT \cdot TSTWT$

Table C.1 gives the estimates $\hat{h}$, and the values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC for the nested models following the ordering above. A model of size k has the 1st to the kth variable in the model, and variables are included following the order of inclusion stated above. Note that $\hat{h} = 1$ for all model lists, $\widehat{ERM}_1 = 0.74$, and $\widehat{ERM}_2 = 0.55$ for all model lists. BIC increases slowly as the size of the conjectured model increases leading to the one term model. The model chosen by $\hat{h}$ is also the simplest model using only $KPSM \cdot TKWT$ as the covariate. Note that $\hat{h} = 1$ for all model lists. The fact that either $\widehat{ERM}_1$ or $\widehat{ERM}_2$ does not change as the size of the conjectured model increases indicates that the variables that were included after $TKWT \cdot KPSM$ do not provide a lot of new information to the model. So $\widehat{ERM}_1$ and $\widehat{ERM}_2$ do not detect anything and end up picking the smallest model with $TKWT \cdot KPSM$ as variable. BIC gives the same result but decreases slowly indicating that the
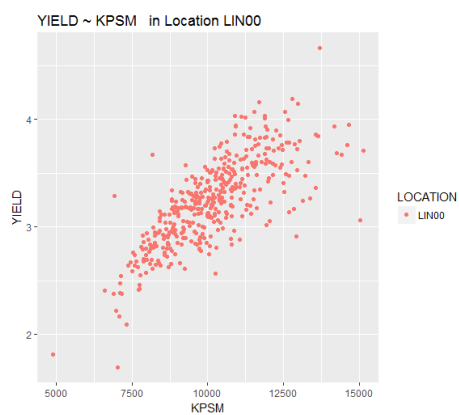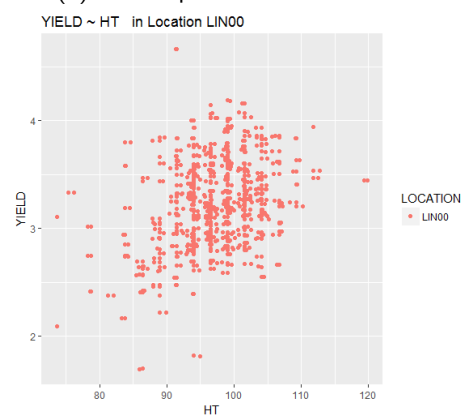
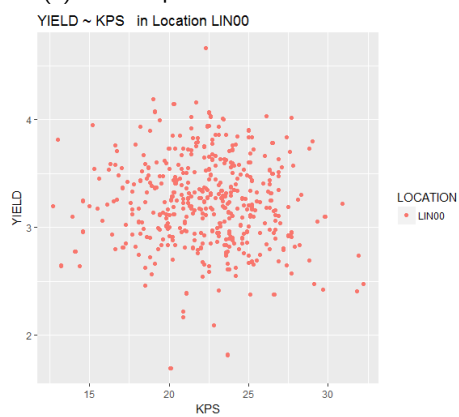(a) Scatter plot of YIELD vs KPSM
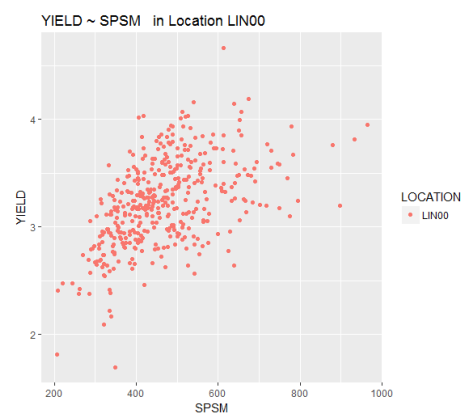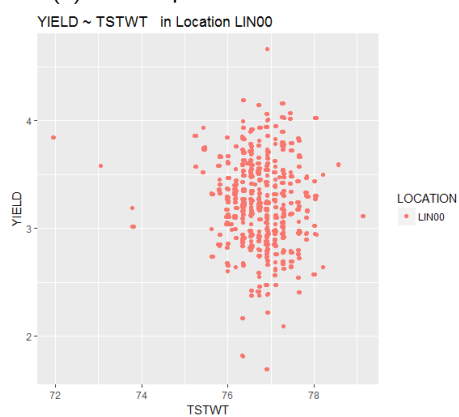
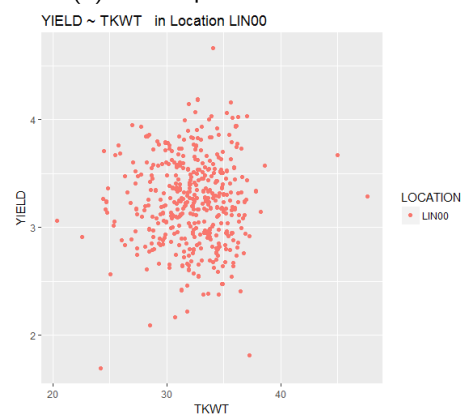(b) Scatter plot of YIELD vs SPSM

(c) Scatter plot of YIELD vs TSTWT

(d) Scatter plot of YIELD vs HT

(e) Scatter plot of YIELD vs KPS

(f) Scatter plot of YIELD vs TKWT

Figure C.1: Scatter plot of Yield versus phenotype data in Lincoln 2000

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
1	1	0.74	0.55	-3512.80
2	1	0.74	0.55	-3509.93
3	1	0.74	0.55	-3503.9
4	1	0.74	0.55	-3497.99
5	1	0.74	0.55	-3491.96
6	1	0.74	0.55	-3486.32
7	1	0.74	0.55	-3480.30
8	1	0.74	0.55	-3474.7
9	1	0.74	0.55	-3468.8

(a) Model sizes from 1 to 9

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
10	1	0.74	0.55	-3462.81
11	1	0.74	0.55	-3457.22
12	1	0.74	0.55	-3451.37
13	1	0.74	0.55	-3445.8
14	1	0.74	0.55	-3440.24
15	1	0.74	0.55	-3434.47
16	1	0.74	0.55	-3428.65
17	1	0.74	0.55	-3427.02
18	1	0.74	0.55	-3421.03

(b) Model sizes from 10 to 18

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
19	1	0.74	0.55	-3415.5
20	1	0.74	0.55	-3409.48
21	1	0.74	0.55	-3403.46
22	1	0.74	0.55	-3397.47
23	1	0.74	0.55	-3391.56
24	1	0.74	0.55	-3385.72
25	1	0.74	0.55	-3381.73
26	1	0.74	0.55	-3375.77
27	1	0.74	0.55	-3369.87

(c) Model sizes from 19 to 27

Table C.1: Estimation of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_1$ and BIC in LIN00

posterior spreads the probability of the best model over a relatively large class of models.

Turning our attention to sparsity analyses, we observe from Fig. C.2a that the optimal value of λ under SCAD is $\hat{\lambda} = 0.0999$; with this value of $\hat{\lambda}$, the best model should have one variable. From Fig. C.2b, using the optimal value of $\hat{\lambda}$, the variable chosen by SCAD is $KPSM \cdot TKWT$. That is, the model is:

$$\widehat{YIELD} = 3.23 + 0.40 \cdot KPSM \cdot TKWT. \tag{C.1}$$

Similarly, under Adaptive LASSO, we are led to the model

$$\widehat{YIELD} = 3.23 + 0.40 \cdot TKWT \cdot KPSM + 0.00046 \cdot KPSM. \tag{C.2}$$

Note that $\hat{h}$, SCAD and BIC choose the same variable (the interaction between $TKWT$ and $KPSM$), while ALASSO chooses $TKWT \cdot KPSM$ and $KPSM$. However, the coefficient of $KPSM$ in ALASSO model is close to zero; this indicates that under ALASSO, the variable that has the most effect on $YIELD$ is $TKWT \cdot KPSM$ and $KPSM$ by itself can probably be neglected. Overall, this indicates that in Lincoln 2000, $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$, BIC, SCAD, and ALASSO more or less agree. Note that some would argue that (C.2) is better than (C.1) because it has the main effect $KPSM$ present in the model.
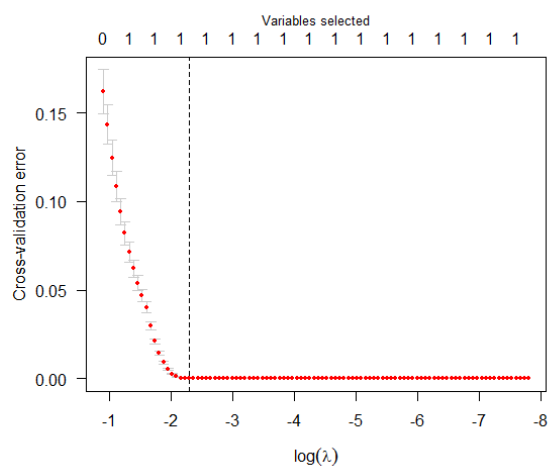
(a) Optimal value of λ for Lincoln 2000
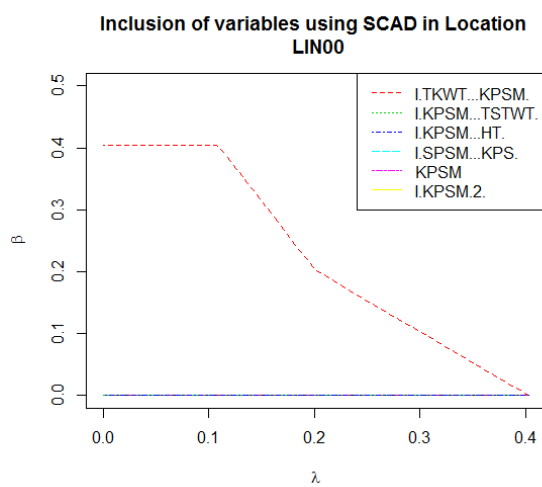
(b) Trace of variables for Lincoln 2000 using SCAD Fan and Li (2001)

Figure C.2: Analysis of Wheat data set in Lincoln 2000 using SCAD Fan and Li (2001)

Analysis of Wheat data in Lincoln 2001

From Fig. C.3a we see that there is a reasonably linear relationship between $YIELD$ and $KPSM$. The variability does not particularly increase as $KPSM$ increases. There are some observations that are away from the bulk of the data, but the overall trend is linear. In Fig. C.3b there is a weak linear relationship between $YIELD$ and $SPSM$ and the variance increases as a function of $SPSM$. From Fig. C.3c to C.3f there does not appear to be any non-trivial relationship between $YIELD$ and the phenotype covariates. Observations are scattered and some of them are far from the bulk of data.

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
1	18	17	26	-182
2	17	16	24	-197
3	17	16	24	-205
4	17	16	24	-203
5	18	16	24	-197
6	17	16	23	-198
7	13	9	14	-476
8	8	3	5	-1037
9	8	3	5	-1041

(a) Model sizes from 1 to 9

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
10	10	3	6	-1037
11	11	3	6	-1032
12	43	4	12	-11081
13	43	4	12	-1106
14	10	3	5	-1102
15	11	3	5	-1096
16	12	3	6	-1091
17	9	2	3	-1597
18	20	3	6	-1591

(b) Model sizes from 10 to 18

| Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
19	11	2	3	-1980
20	11	2	3	-1974
21	11	2	3	-1968
22	7	2	2	-1976
23	7	2	2	-1981
24	9	2	3	-1978
25	9	2	3	-1982
26	10	2	3	-1976
27	10	2	3	-1971

(c) Model sizes from 19 to 27

Table C.2: Estimation of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_1$ and BIC in LINO1

We use the same model list i.e., all terms of order ≤ 2 in the phenotypic variables. Under correlation, the order of inclusion is as follows: $KPSM \cdot TSTWT, KPSM, KPSM^2, SPSM \cdot TSTWT, SPSM, KPSM \cdot SPSM, SPSM \cdot TKWT, TKWT \cdot KPSM, KPSM \cdot KPS, SPSM^2, SPSM \cdot KPS, TSTWT, TSTWT^2, KPS^2, HT^2, KPS \cdot TSTWT, TKWT \cdot TSTWT, KPS, TKWT, TKWT^2, KPS \cdot TKWT, TKWT \cdot HT, HT, KPS \cdot HT, HT \cdot TSTWT, SPSM \cdot HT, KPSM \cdot HT$

Table C.2 gives the estimate of $\hat{h}$, the values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC in Lincoln 2001.

We see that the estimate $\hat{h}$ varies with the conjectured model. The smallest difference between the size of the conjectured model and $\hat{h}$ occurs either with the conjectured model of size 8 or 10. So we would choose $\hat{h} = 8$. We also observed some variability in the values of $\widehat{ERM}_1$ and $\widehat{ERM}_2$. Both values are pretty much constant from conjectured models of size 17 to 27, their values are either 2 or 3. In parsimonious ground, both choose models of size 17. The smallest value of BIC occurs for the model of
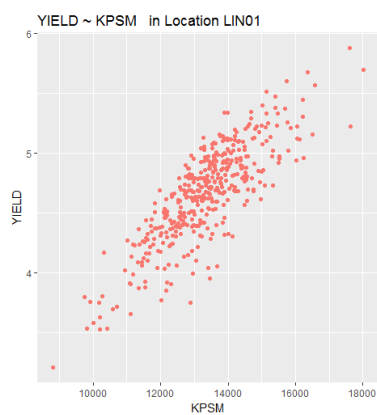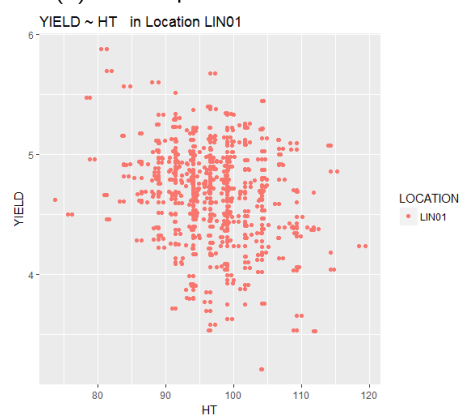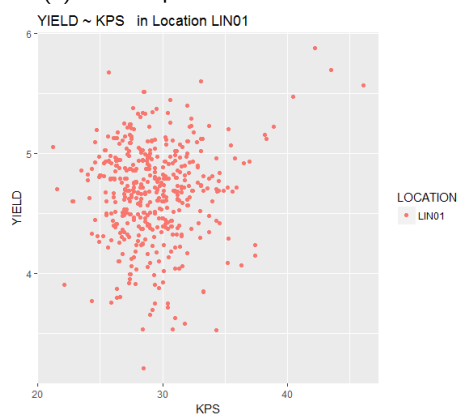
(a) Scatter plot of YIELD vs KPSM
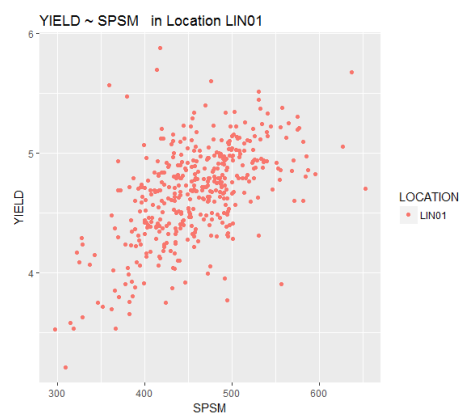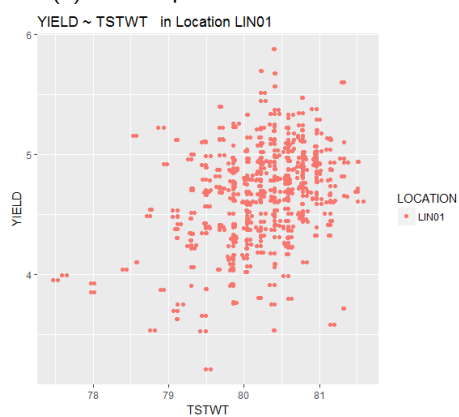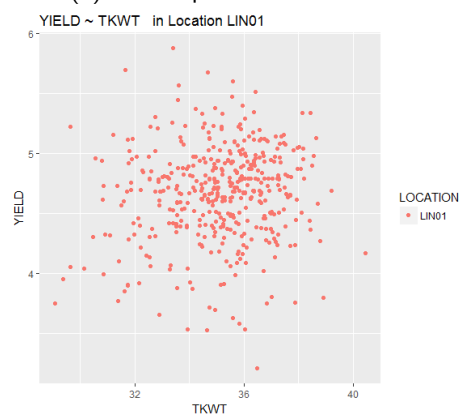
(b) Scatter plot of YIELD vs SPSM

(c) Scatter plot of YIELD vs TSTWT

(d) Scatter plot of YIELD vs HT

(e) Scatter plot of YIELD vs KPS

(f) Scatter plot of YIELD vs TKWT

Figure C.3: Scatter plot of Yield versus phenotype data in Lincoln 2001

Variables selected

(a) Optimal value of λ for Lincoln 2001

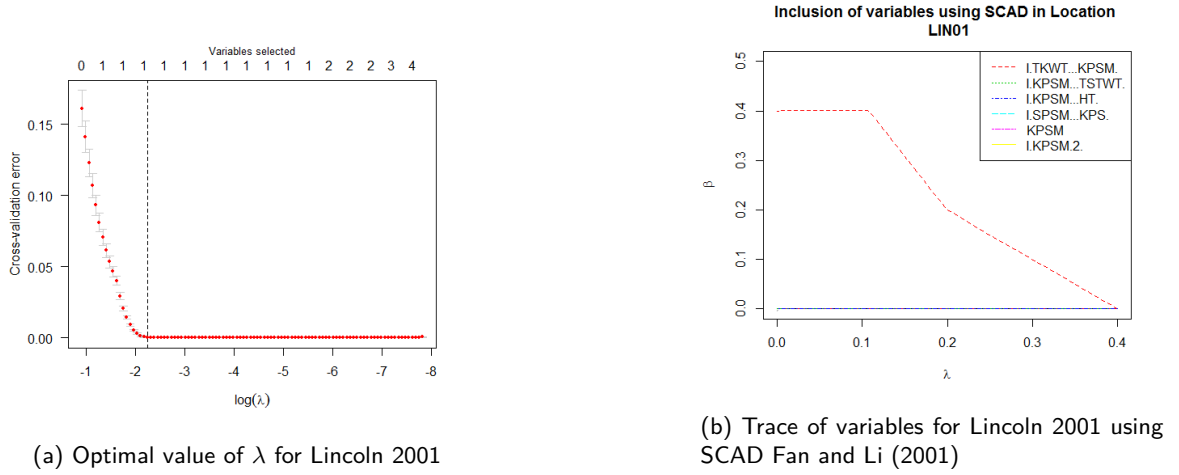(b) Trace of variables for Lincoln 2001 using SCAD Fan and Li (2001)

Figure C.4: Analysis of Wheat data set in Lincoln 2001 using SCAD Fan and Li (2001)

size 21. We observe that there is a big difference between models chosen by $\hat{h}$, BIC, and $\widehat{ERM}_1$ and $\widehat{ERM}_2$.

Turning to sparsity methods, from Fig. C.4a, the optimal value of λ is $\hat{\lambda} = 0.106$; with this optimal value of $\hat{\lambda}$, the best model has one variable. From Fig. C.4b, using the optimal value of $\hat{\lambda}$, the variable chosen by SCAD is $KPSM \cdot TKWT$. That is, the model is written as follows:

$$\widehat{YIELD} = 4.67 + 0.40 \cdot KPSM \cdot TKWT. \tag{C.3}$$

Under Adaptive LASSO, similarly, we are led to

$$\widehat{YIELD} = 4.67 + 0.40 \cdot KPSM \cdot TKWT + 0.0017 \cdot KPSM. \tag{C.4}$$

Note that $\hat{h}$ chooses a model that is very different from that of $\widehat{ERM}_1$, $\widehat{ERM}_2$, BIC, SCAD and ALASSO. $\widehat{ERM}_1$, $\widehat{ERM}_2$, BIC, SCAD pick the same model with $TKWT \cdot KPSM$ as variable. ALASSO picks $TKWT \cdot KPSM$ and $KPSM$, however, the coefficient of $KPSM$ in ALASSO is still pretty small. This means that under ALASSO, the variable that has the most effect on $YIELD$ is $TKWT \cdot KPSM$. Here, we would argue that while $TKWT \cdot KPSM$ is a reasonably good simple model; however it oversimplifies yield and the model chosen by $\hat{h}$ does not.

C.0.1 Analysis of Wheat data in Mead 2000

From Fig. C.5a we see that there is a linear relationship between $YIELD$ and $KPSM$, however, the variance slowly increases as the value of $KPSM$ increases, some data points are far from most of them, but the overall trend is linear. In Fig. C.5b, there is a weak linear relationship between $YIELD$ and $SPSM$, the variance increases faster as the value of $SPSM$ increases. From Figs. C.5c to C.5f, there

is no any obvious relationship between $YIELD$ and $TSTWT$, HT, KPS and $TKWT$; data are just spread all over the place, some are far away from the bulk of the data.

Size	$\hat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
1	9	1.78	2.89	**−1991.09**
2	9	1.78	2.89	-1985.2
3	8	1.71	2.67	-1979.41
4	8	1.71	2.67	-1973.38
5	8	1.71	2.66	-1973.73
6	8	1.74	2.66	-1967.93
7	8	1.7	2.66	-1961.9
8	**8**	1.7	2.66	-1955.87
9	8	1.7	2.66	-1950.09

(a) Model sizes from 1 to 9

Size	$\hat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
10	8	1.7	2.66	-1945.38
11	8	1.7	2.66	-1939.45
12	8	1.7	2.66	-1933.51
13	8	1.7	2.66	-1933.51
14	7	1.63	2.43	-1930.14
15	7	1.62	2.42	-1935.24
16	6	1.53	2.17	-1940.29
17	6	1.53	2.17	-1934.26
18	6	1.53	2.17	-1933.93

(b) Model sizes from 10 to 18

Size	$\hat{h}$	$\widehat{ERM_1}$	$\widehat{ERM_2}$	BIC
19	6	1.53	2.17	-1928.33
20	6	1.53	2.17	-1922.31
21	6	1.53	2.17	-1916.28
22	6	1.53	2.17	-1910.82
23	6	1.53	2.17	-1905.44
24	6	1.53	2.17	-1899.9
25	6	1.53	2.17	-1893.87
26	6	1.53	2.17	-1887.89
27	7	1.61	2.4	-1882.01

(c) Model sizes from 19 to 27

Table C.3: Estimation of $\hat{h}$, $\widehat{ERM_1}$, $\widehat{ERM_1}$ and BIC in Mead 2000

Using the same model lists as before, and using correlation to order the inclusion of terms in the model is as follows: $TKWT \cdot KPSM, KPSM \cdot TSTWT, SPSM \cdot KPS, KPSM, KPSM \cdot HT,$ $KPSM^2, KPSM \cdot SPSM, SPSM \cdot TKWT, SPSM \cdot TSTWT, SPSM \cdot HT\ SPSM, SPSM^2, KPSM \cdot$ $KPS, HT \cdot TSTWT, HT^2, HT, TKWT \cdot HT, TSTWT, TSTWT^2, TKWT \cdot TSTWT, TKWT^2, TKWT,$ $KPS \cdot HT,\ KPS \cdot TKWT,\ KPS \cdot TSTWT, KPS, KPS^2$.

Table C.3 gives the estimate of $\hat{h}$, the values of $\widehat{ERM_1}$, $\widehat{ERM_2}$ and BIC in Mead 2000.

We observe some variability in the estimate of $\hat{h}$, it ranges between 6 to 9. The smallest discrepancy between the size of the conjectured model and $\hat{h}$ occurs when the size of the conjectured model is 8. In fact, at that point $\hat{h} = 8$. We also see that $\widehat{ERM_1}$ and $\widehat{ERM_2}$ vary as the size of the conjectured model increases. The smallest value of either $\widehat{ERM_1}$ or $\widehat{ERM_2}$ occurs at the most complex model. BIC spreads the bulk of the posterior weight over several models even though the mode occurs at the smallest model.
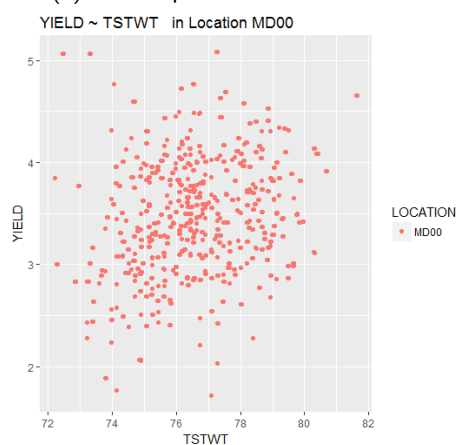
Now, let's move to shrinkage methods. Fig. C.6a is little bit different from the previous ones. At the optimal value of λ, we obtain seven because the CVE decreases unexpectedly from 7 to 6. Assuming that this is not a software problem, we argue that the optimal value of λ under SCAD is $\hat{\lambda} = 0.000915$. With
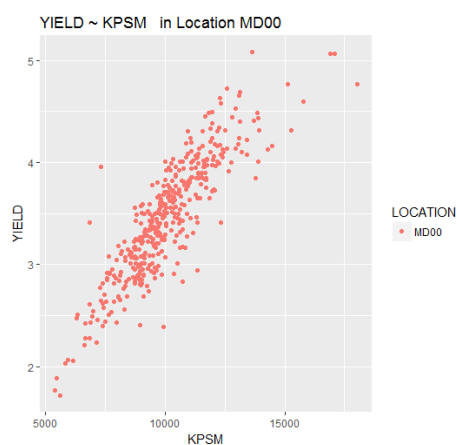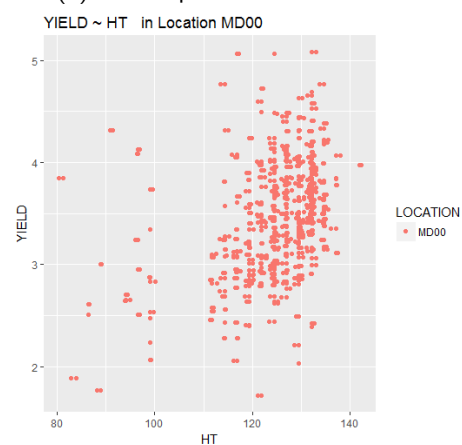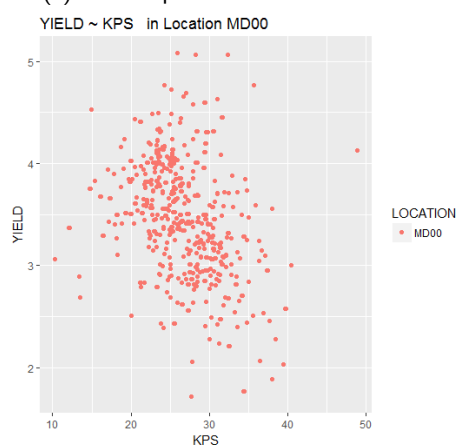
(a) Scatter plot of YIELD vs KPSM
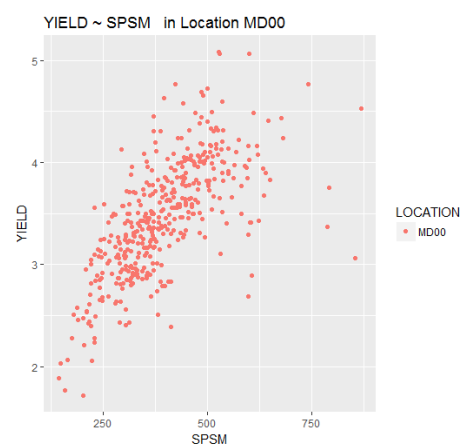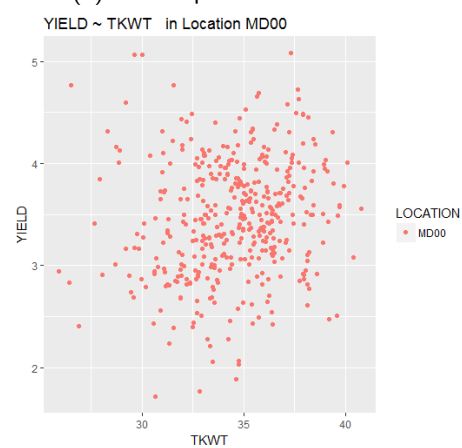

(b) Scatter plot of YIELD vs SPSM

(c) Scatter plot of YIELD vs TSTWT

(d) Scatter plot of YIELD vs HT

(e) Scatter plot of YIELD vs KPS

(f) Scatter plot of YIELD vs TKWT

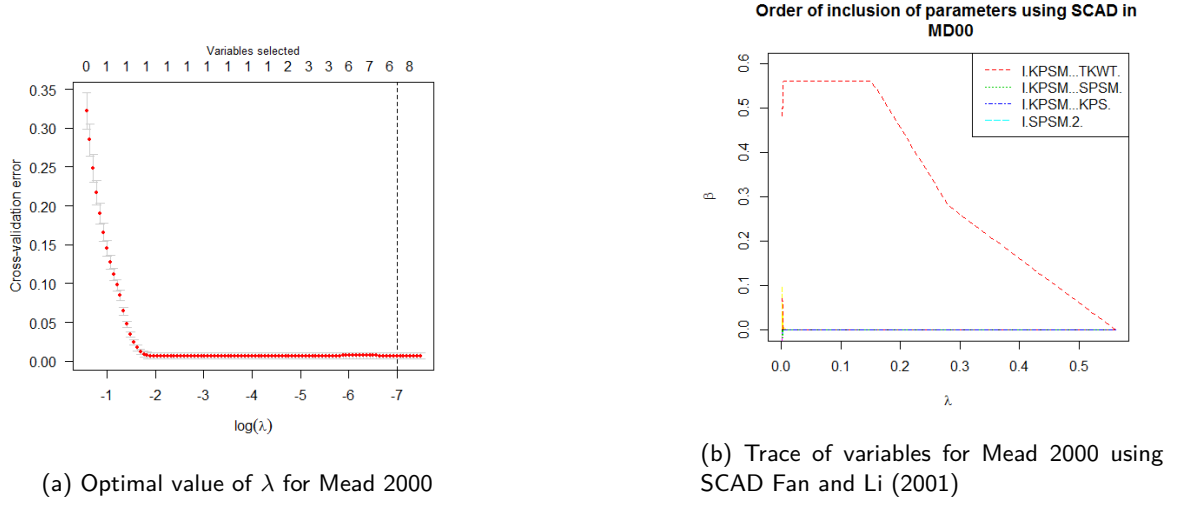Figure C.5: Scatter plot of Yield versus phenotype data in Mead 2000

(a) Optimal value of λ for Mead 2000

(b) Trace of variables for Mead 2000 using SCAD Fan and Li (2001)

Figure C.6: Analysis of Wheat data set in Mead 2000 using SCAD Fan and Li (2001)

this value, the optimal model is written as follows:

$$
\begin{aligned}
\widehat{YIELD} &= 3.46 + 0.50 \cdot KPSM \cdot TKWT - 0.01 \cdot SPSM^2 \\
&+ 0.035 \cdot TKWT^2 - 0.0009 \cdot KPS \cdot TKWT \\
&+ 0.0078 \cdot TKWT \cdot HT + 0.07 \cdot KPSM^2 - 0.015 \cdot KPS^2.
\end{aligned}
\tag{C.5}
$$

Similarly, under Adaptive LASSO, we have

$$
\begin{aligned}
\widehat{YIELD} &= 3.46 + 0.47 \cdot KPSM \cdot TKWT + 0.051 \cdot TKWT \\
&+ 0.096 \cdot SPSM \cdot KPS - 0.0053 \cdot KPS^2.
\end{aligned}
\tag{C.6}
$$

We observe that each model selection technique picks a different model list. $\hat{h}$ picks the model of size 8, BIC picks the simplest model, whereas $\widehat{ERM}_1$ and $\widehat{ERM}_2$ pick the most complex model. $TKWT \cdot KPSM$ is the variable that appears with all model lists. $KPSM^2$ is present in the model chosen by $\hat{h}$ and SCAD, similarly $SPSM \cdot KPS$ occurs in the ALASSO and $\hat{h}$ models. We argue here that the model chosen by $\hat{h}$ is probably better than other model selection technique, because it is the most plausible model. It is difficult to believe small models for this sort of data can be true to extremely high precision.

Analysis of Wheat dataset in Mead 2001

From Fig. C.7a, we observe that there is a linear relationship between $YIELD$ and $KPSM$. The variability slowly increases as a function of $KPSM$. But the overall trend is linear. On Fig. C.7b, there is a weak linear relationship between $YIELD$ and $SPSM$, the variability increases as $SPSM$ increases. There are data points that are away from the bulk of data , but not so much that as to suggest problems
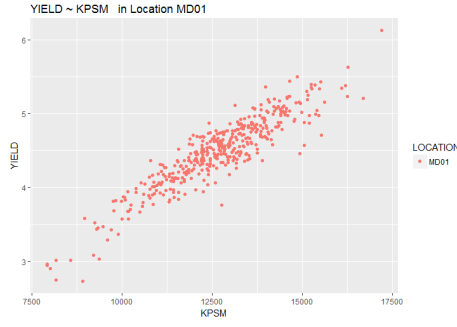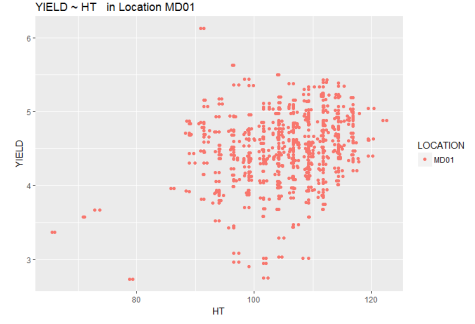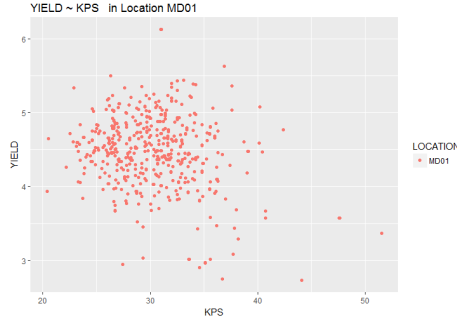
(a) Scatter plot of YIELD vs KPSM
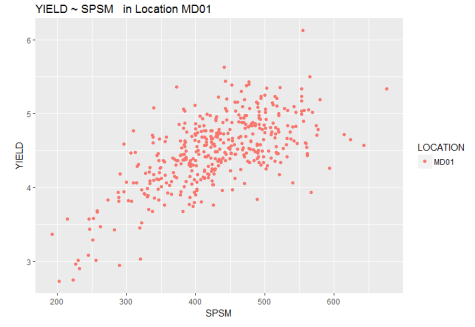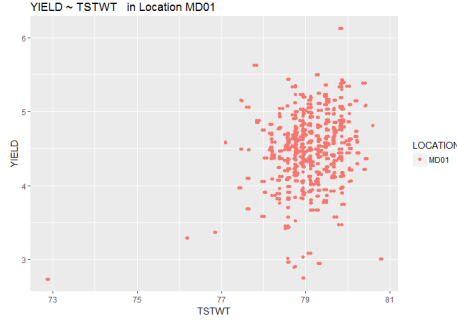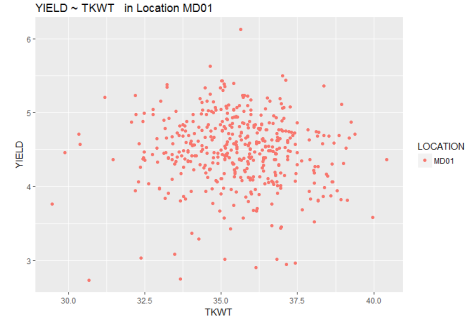
(b) Scatter plot of YIELD vs SPSM

(c) Scatter plot of YIELD vs TSTWT

(d) Scatter plot of YIELD vs HT

(e) Scatter plot of YIELD vs KPS

(f) Scatter plot of YIELD vs TKWT

Figure C.7: Scatter plot of Yield versus phenotype data in Mead 2001

with outliers. From Figs. C.7c to C.7f, there is no useful trend between $YIELD$ and $TSTWT$, HT, KPS and $TKWT$. While there are possibly outliers, they do not obviously suggest any pattern.

As we have done before, using correlation Fan and Lv (2008), the order of inclusion of variables is as follows: $TKWT{\cdot}KPSM, KPSM{\cdot}TSTWT, SPSM{\cdot}KPS, KPSM, KPSM^2, KPSM{\cdot}SPSM, KPSM{\cdot}HT, SPSM{\cdot}TKWT, SPSM{\cdot}TSTWT, SPSM, SPSM{\cdot}HT, SPSM^2, KPSM{\cdot}KPS, HT{\cdot}TSTWT, TSTWT, HT, TSTWT^2, HT^2, TKWT{\cdot}HT, TKWT{\cdot}TSTWT, KPS{\cdot}HT, TKWT, TKWT^2, KPS{\cdot}TSTWT, KPS, KPS \cdot TKWT, KPS^2$.

We compare our method in this location to shrinkage method such as SCAD Fan and Li (2001) and Adaptive LASSO Zou (2006b).

Table C.4 gives the estimate of $\hat{h}$, the values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC in Mead 2000. A model of size k has the 1st to the kth variable in the model, and variables are included following the order of inclusion stated above.

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
1	5	1.4	1.86	-2120
2	5	1.4	1.86	-2114
3	5	1.4	1.86	-2108
4	4	1.3	1.62	-2102
5	4	1.3	1.62	-2097
6	4	1.3	1.62	-2091
7	4	1.3	1.62	-2085
8	4	1.3	1.62	-2079
9	4	1.3	1.62	-2067

(a) Model sizes from 1 to 9

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
10	4	1.3	1.62	-2067
11	4	1.3	1.62	-2061
12	4	1.3	1.62	-2055
13	3	1.19	1.36	-2049
14	3	1.3	1.62	-2043
15	3	1.19	1.36	-2037
16	3	1.3	1.62	-2031
17	4	1.3	1.62	-2026
18	4	1.3	1.62	-2020

(b) Model sizes from 10 to 18

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
19	4	1.3	1.62	-2014
20	4	1.3	1.62	-2008
21	4	1.3	1.62	-2002
22	4	1.3	1.61	-1996
23	4	1.3	1.61	-1990
24	4	1.3	1.61	-1984
25	4	1.3	1.61	-1978
26	4	1.4	1.86	-1972
27	5	1.4	1.86	-1966

(c) Model sizes from 19 to 27

Table C.4: Estimation of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_1$ and BIC in Mead 2001

We observe some variability in the estimate of $\hat{h}$, it ranges between 3 to 5. The smallest discrepancy between the size of the conjectured model and $\hat{h}$ occurs when the size of the conjectured model is 4. In fact, at that point $\hat{h} = 4$. We also see that $\widehat{ERM}_1$ and $\widehat{ERM}_2$ vary as the size of the conjectured model increases. The smallest value of either $\widehat{ERM}_1$ or $\widehat{ERM}_2$ occurs at when the size of the conjectured model is either 13 or 15. BIC spreads the bulk of the posterior weight over several models even though the mode occurs at the smallest model.

Turning to shrinkage methods, from Fig. C.8a, the optimal value of λ is $\hat{\lambda} = 0.1200391$. With this value the optimal number of parameters is one. From Fig. C.8b, under SCAD, the variable chosen is $TKWT \cdot KPSM$. SO the model will be written as follows:

$$\widehat{YIELD} = 4.49 + 0.49 \cdot KPSM \cdot TKWT \tag{C.7}$$

Similarly, under Adaptive LASSO, we have

$$\widehat{YIELD} = 4 + 0.48 \cdot KPSM \cdot TKWT. \tag{C.8}$$

It is important to note that, for this dataset, SCAD, ALASSO, and BIC pick the same variables for the model and (C.7) and (C.8) are nearly the same. However, $\widehat{ERM}_1$ and $\widehat{ERM}_2$ pick a model of size 13 or 15 and $\hat{h}$ picks a model of size 4. We argue that the model chosen by $\hat{h}$ is probably the best because at
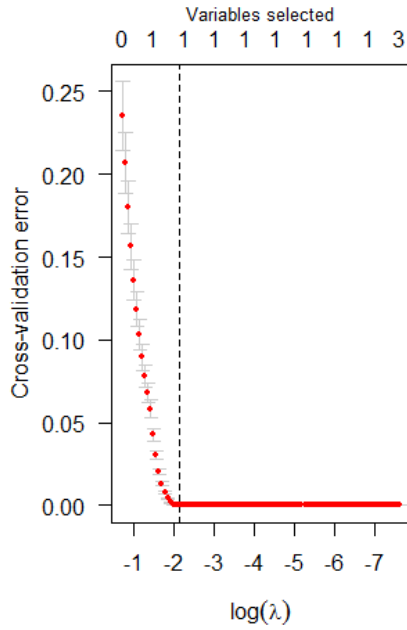
(a) Optimal value of λ for Mead 2001
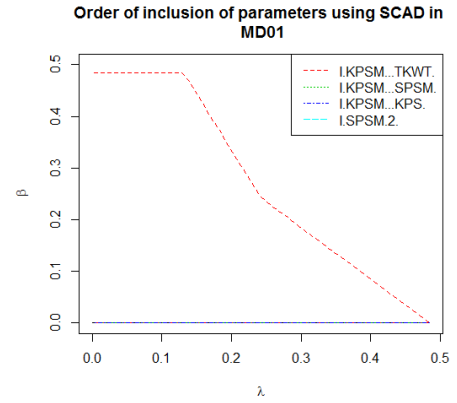
(b) Trace of variables for Mead 2001 using SCAD Fan and Li (2001)

Figure C.8: Analysis of Wheat data set in Mead 2001 using SCAD Fan and Li (2001)

least it includes the main effect $KPSM$ and the trems chosen by the shrinkage methods. we think that it is very odd that $\widehat{ERM}_1$ and $\widehat{ERM}_2$ that are function of $\hat{h}$ give up too much information and thereby let in too many terms.

Analysis of Wheat dataset in Sidney 2000

From Fig. C.9a, there is a linear relationship between $YIELD$ and $KPSM$. The variance is pretty much constant as $KPSM$ increases. There are some data points that are little bit away from the bulk of the data, but the overall data follows a linear trend. From Fig C.9b, as before, there is also a linear relationship between $YIELD$ and $SPSM$, we see some data points that are not close to the majority of data, we might consider them as potential outliers, however they don't appear to be very serious problem. From Figs. C.9c to C.9f, it is tempting to see weak linear relationship between $YIELD$ and $TSTWT$, HT, KPS and $TKWT$. However, they seem to be very weak.

As before, we order the inclusion of variables in the model using correlation Fan and Lv (2008). The nesting of our models follows this order of terms: $TKWT \cdot KPSM, KPSM \cdot TSTWT, SPSM \cdot TKWT, KPSM, SPSM \cdot KPS, KPSM^2, KPSM \cdot SPSM, KPSM \cdot HT, SPSM \cdot TSTWT, SPSM \cdot HT, SPSM, SPSM^2, TKWT \cdot HT, TSTWT^2, TSTWT, KPSM \cdot KPS, HT \cdot TSTWT, TKWT \cdot TSTWT, HT^2, HT, TKWT, TKWT^2, KPS \cdot TKWT, KPS \cdot HT, KPS \cdot TSTWT, KPS, KPS^2$. So, we fit 27 different models and each time we estimate $\hat{h}$, find the value of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC. We
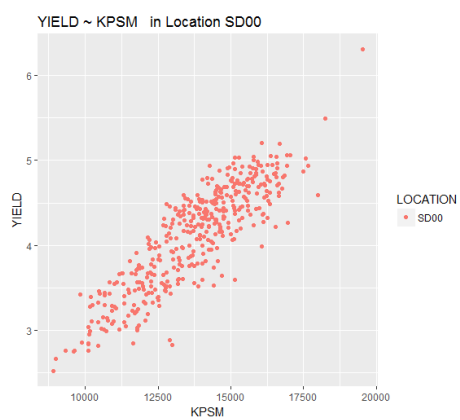
(a) Scatter plot of YIELD vs KPSM
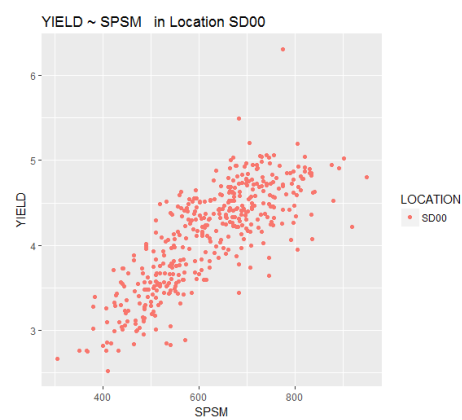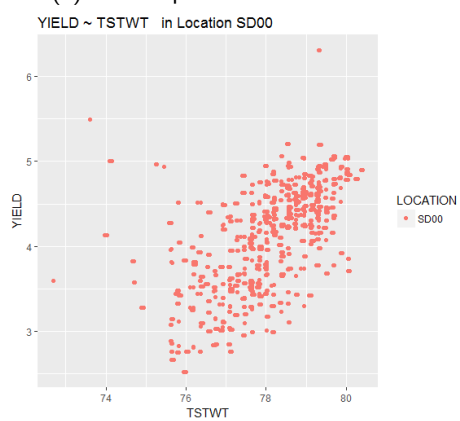
(b) Scatter plot of YIELD vs SPSM

(c) Scatter plot of YIELD vs TSTWT

(d) Scatter plot of YIELD vs HT

(e) Scatter plot of YIELD vs KPS

(f) Scatter plot of YIELD vs TKWT

Figure C.9: Scatter plot of Yield versus phenotype data in Sidney 2000

will also fit a SCAD and ALASSO models for comparison purposes.

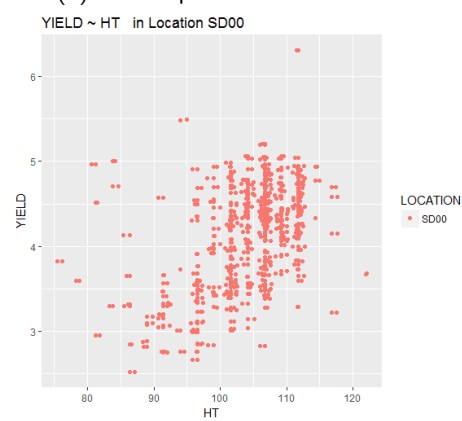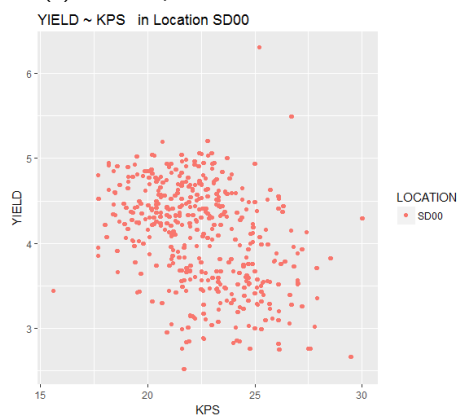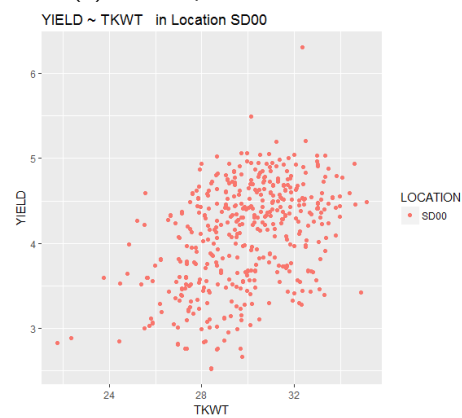| Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
1	2	1	1	-2298
2	2	1	1	-2292
3	1	0.83	0.71	-2288
4	1	0.83	0.71	-2282
5	1	0.83	0.71	-2276
6	1	0.83	0.71	-2271
7	1	0.83	0.71	-2265
8	1	0.83	0.71	-2259
9	1	0.83	0.71	-2254

(a) Model sizes from 1 to 9

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
10	1	0.83	0.71	-2248
11	1	0.83	0.71	-2248
12	1	0.83	0.71	-2237
13	1	0.83	0.71	-2231
14	1	0.83	0.71	-2225
15	1	0.83	0.71	-2219
16	1	0.83	0.71	-2214
17	1	0.83	0.71	-2208
18	1	0.83	0.71	-2202

(b) Model sizes from 10 to 18

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
19	1	0.82	0.71	-2196
20	1	0.82	0.71	-2190
21	1	0.82	0.71	-2184
22	1	0.82	0.71	-2179
23	1	0.82	0.71	-2173
24	1	0.82	0.71	-2167
25	1	0.82	0.71	-2161
26	1	0.82	0.71	-2155
27	1	0.82	0.71	-2149

(c) Model sizes from 19 to 27

Table C.5: Estimation of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_1$ and BIC in Sideney 2000

Table C.5 gives the estimate of $\hat{h}$, the values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC in Sidney 2000.

We observe small variability in the estimate of $\hat{h}$: it is either 1 or 2. The smallest discrepancy between the size of the conjectured model and $\hat{h}$ occurs when the size of the conjectured model is 2. In fact, at that point $\hat{h} = 2$. We also see that there is a small variability in the values of $\widehat{ERM}_1$ and $\widehat{ERM}_2$ as the size of the conjectured model increases. The smallest value of $\widehat{ERM}_1$ (0.82) occurs first when the size of the conjectured model is 19, however, when the size of the conjectured model is 3, $\widehat{ERM}_1 = 0.83$, so we would pragmatically say that $\widehat{ERM}_1$ picks a model of size 3. There is not much difference between $\widehat{ERM}_1$ for conjectured models of size 3 through 15. The smallest value of $\widehat{ERM}_2$ first occurs at the conjectured model of size 3. BIC picks the smallest i.e. one term model and, as usual, BIC spreads the bulk of the posterior weight over several models even though the mode occurs at the smallest model.

Fig. C.10a is the graph of the cross-validation error against the different values of λ. The optimal value of λ seen to be $\hat{\lambda} = 0.1498493$. From Fig. C.10b, and using the optimal value of $\hat{\lambda}$, we see that using SCAD, the only variable to enter the model is $TKWT \cdot KPSM$ and the model is

$$\widehat{YIELD} = 4.1 + 0.61 \cdot KPSM \cdot TKWT \tag{C.9}$$

(a) Optimal value of λ for Sidney 2000

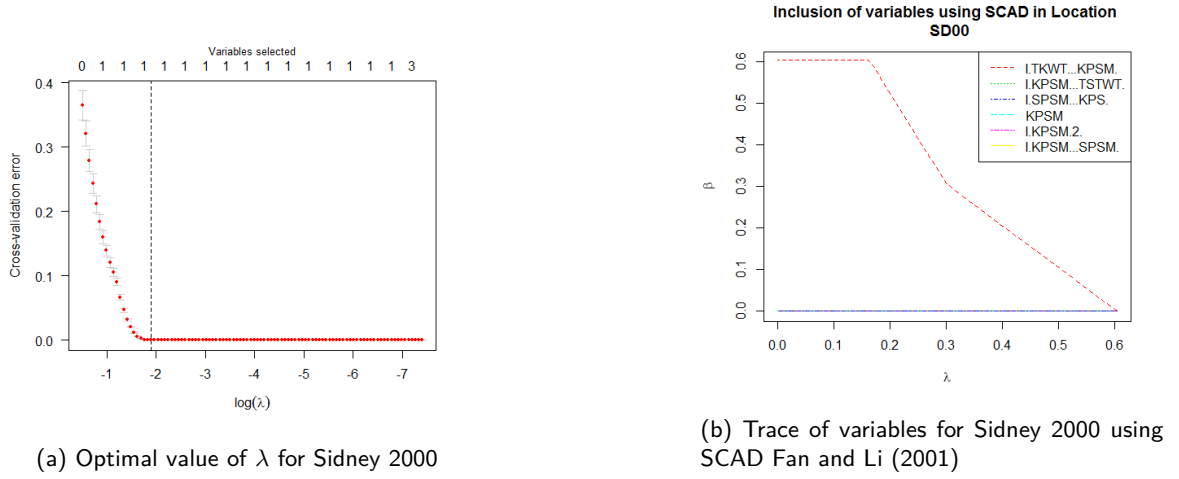(b) Trace of variables for Sidney 2000 using SCAD Fan and Li (2001)

Figure C.10: Analysis of Wheat data set in Sidney 2000 using SCAD Fan and Li (2001)

Similarly with ALASSO, we have

$$\widehat{YIELD} = 4.1 + 0.60 \cdot KPSM \cdot TKWT + 0.00118 \cdot SPSM \cdot TSTWT. \tag{C.10}$$

Overall, we can say that $TKWT \cdot TSTWT$ is the variable that is present in all model lists. SCAD and ALASSO are slightly different. ALASSO has an extra term ($SPSM \cdot TSTWT$), however the coefficient is almost zero (0.00118). This indicates that for ALASSO model, the most important variable to predict the yield is $TKWT \cdot KPSM$; this means that there is not any significant difference between SCAD and ALSSO. It is also important to note that for this dataset, the complexity of the model to predict the $YIELD$ is at most 3.

We comment that here and elsewhere, SCAD typically gives a model no larger than ALASSO because the SCAD penalty is effectively as small a penalty as possible subject to giving the oracle property.

Analysis of Wheat data in Sidney 2001

From Fig. C.11a, we see a linear trend between $YIELD$ and $KPSM$, the variance is constant as $KPSM$ increases. There are some observations that are away from the bulk of the data. From Fig. 5.6b we see that the variance increases as $SPSM$ increases, some observations are far from the bulk of the data. There is a weak linear trend between $YIELD$ and $SPSM$. In Fig. C.11c, we see a curvilinear relationship between $YIELD$ and $TSTWT$ and we see that the variance increases as a function of $TSTWT$. From Figs. C.11d to C.11f, it is tempting to see a linear retatiouship between $YIELD$ and the covariates. However, any such relation is weak and the variance increases as function of the covariates.
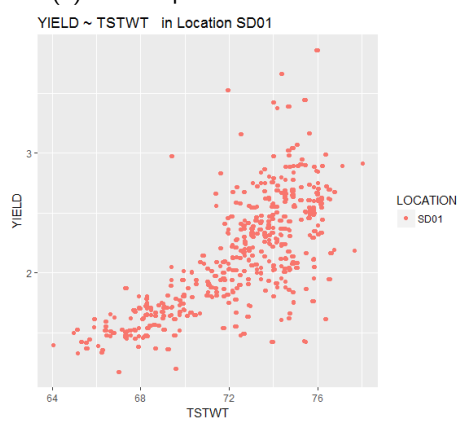
As before, we use the same model list i.e. all terms ≤ 2 in the phenotypic variables. We use correlation to order their inclusion in the model. Under correlation, the nesting order is as follows: $TKWT \cdot KPSM, KPSM \cdot TSTWT, KPSM, \ SPSM \cdot KPS, KPSM^2, KPSM \cdot KPS, SPSM \cdot$
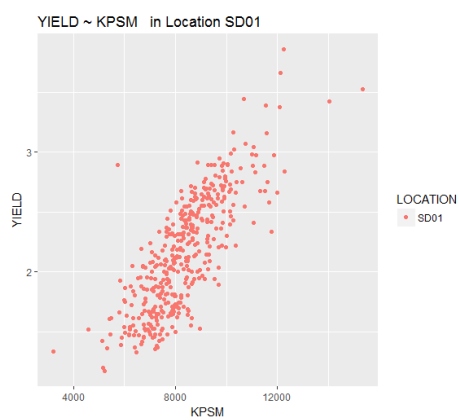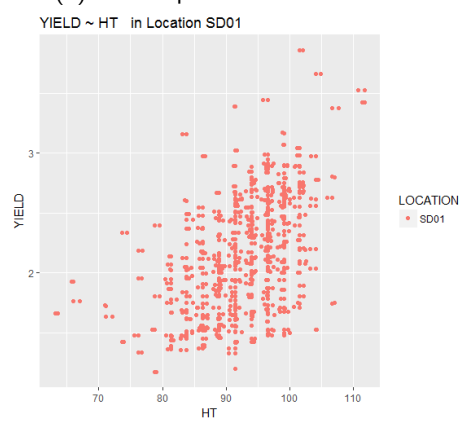
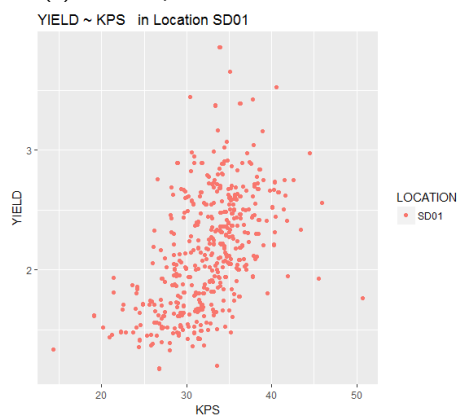(a) Scatter plot of YIELD vs KPSM
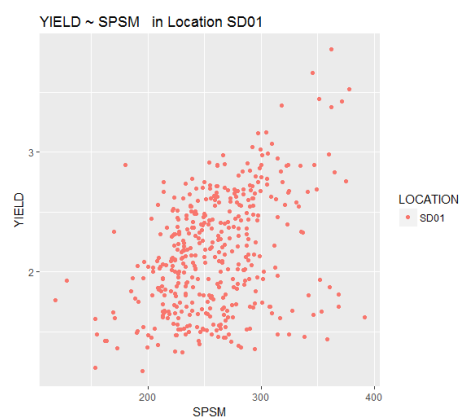
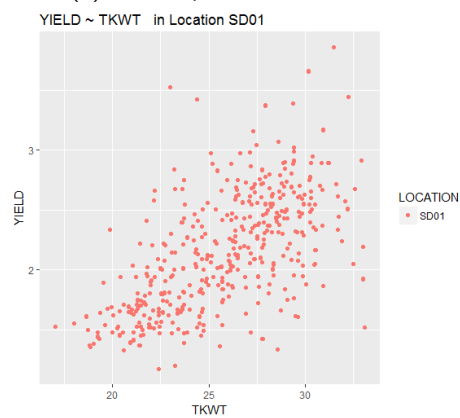(b) Scatter plot of YIELD vs SPSM

(c) Scatter plot of YIELD vs TSTWT

(d) Scatter plot of YIELD vs HT

(e) Scatter plot of YIELD vs KPS

(f) Scatter plot of YIELD vs TKWT

Figure C.11: Scatter plot of Yield versus phenotype data in Sidney 2001

$TKWT, KPSM \cdot HT, TKWT \cdot HT, TSTWT, TSTWT^2, HT \cdot TSTWT, KPS \cdot TKWT, KPSM \cdot SPSM, TKWT \cdot TSTWT, KPS \cdot HT, TKWT, KPS \cdot TSTWT, TKWT^2, SPSM \cdot TSTWT, SPSM \cdot HT, KPS, KPS^2, HT^2, HT, SPSM, SPSM^2.$

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
1	19	3.53	7.09	-1175
2	16	3.36	6.43	-1175
3	16	3.36	6.43	-1170
4	17	3.41	6.64	-1164
5	15	3.25	6.12	-1176
6	15	3.25	6.12	-1170
7	15	3.24	6.1	-1167
8	14	3.17	5.86	-1167
9	21	3.42	7.16	-1190

(a) Model sizes from 1 to 9

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
10	21	3.4	7.13	-1191
11	21	3.4	7.13	-1186
12	21	3.4	7.12	-1180
13	21	3.4	7.12	-1175
14	22	3.43	7.29	-1174
15	23	3.46	7.48	-1169
16	23	3.46	7.47	-1165
17	12	2.92	5.19	-1160
18	13	2.98	5.41	-1154

(b) Model sizes from 10 to 18

Size	$\hat{h}$	$\widehat{ERM}_1$	$\widehat{ERM}_2$	BIC
19	13	2.98	5.41	-1148
20	13	2.97	5.41	-1144
21	13	2.95	5.38	-1144
22	13	2.95	5.37	-1140
23	15	3.05	5.79	-1137
24	20	3.22	6.69	-1159
25	21	3.26	6.89	-1153
26	22	3.3	7.07	-1150
27	20	3.21	6.68	-1145

(c) Model sizes from 19 to 27

Table C.6: Estimation of $\hat{h}$, $\widehat{ERM}_1$, $\widehat{ERM}_1$ and BIC in Sideney 2001

Table C.6 gives the estimate of $\hat{h}$, the values of $\widehat{ERM}_1$, $\widehat{ERM}_2$ and BIC in Sidney 2001.

We observe variability in the estimate of $\hat{h}$, as it ranges between 12 and 23. The smallest discrepancy between the size of the conjectured model and $\hat{h}$ occurs when the size of the conjectured model is 24, 25, and 26, but we pick a model of size 24 in the ground of parsimony. In fact, at that point $\hat{h} = 20$ suggests that a slightly smaller model might also be reasonable. We also see that there is some variability in the values of $\widehat{ERM}_1$ and $\widehat{ERM}_2$ as the size of the conjectured model increases. The smallest value of $\widehat{ERM}_1$ (2.92) occurs when the size of the conjectured model is 17. The smallest value of $\widehat{ERM}_2$ occurs at conjectured model of size 17. We observe variability in the values of BIC. The smallest value of BIC occurs for model of size 10. We still observe that BIC spreads the bulk of the posterior weight over several models. Note that in this case $\hat{h}$, $\widehat{ERM}_1$ and $\widehat{ERM}_2$ increase and decrease, maximum loosely suggesting the size of the best model.

Turning to the sparsity methods, we observe from Fig. C.12a that the optimal value of λ is $\hat{\lambda} = 0.009559901$; this indicates that the optimal number of parameters is 1. From Fig. C.12b, under SCAD,

(a) Optimal value of λ for Sidney 2001

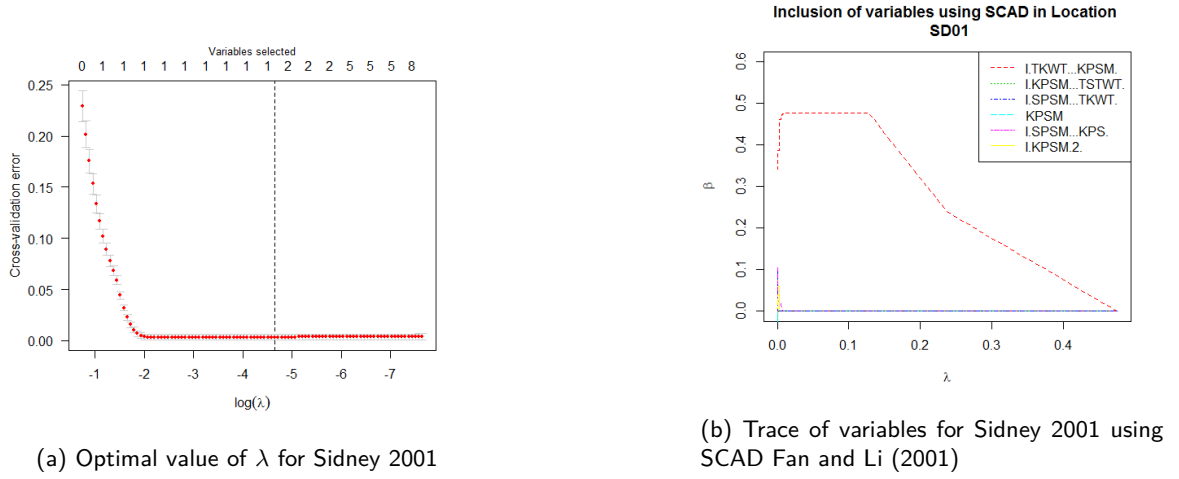(b) Trace of variables for Sidney 2001 using SCAD Fan and Li (2001)

Figure C.12: Analysis of Wheat data set in Sidney 2001 using SCAD Fan and Li (2001)

one variable is chosen, $TKWT \cdot KPSM$, and the model is:

$$\widehat{YIELD} = 2.15 + 0.48 \cdot KPSM \cdot TKWT. \tag{C.11}$$

Similarly, using ALASSO, we have the 15 terms model

$$
\begin{aligned}
\widehat{YIELD} = {} & 2.15 + 0.34 \cdot KPSM \cdot TKWT + 0.03 \cdot KPSM \cdot TSTWT \\
& - 0.04 \cdot SPSM \cdot KPS - 0.26 \cdot KPSM^2 + 0.37 \cdot KPSM \cdot SPSM \\
& + 0.08 \cdot SPSM \cdot TSTWT - 0.36 \cdot SPSM \cdot HT + 0.06 \cdot TKWT \cdot HT \\
& - 0.11 \cdot TSTWT^2 + 0.27 \cdot KPSM \cdot KPS + 0.05 \cdot TKWT \cdot TSTWT \\
& + 0.20 \cdot HT^2 - 0.21 \cdot KPS \cdot HT + 0.31 \cdot KPS \cdot TSTWT - 0.26 \cdot KPS.
\end{aligned}
\tag{C.12}
$$

Note here that the model chosen by ALASSO (C.12) is completely different from that of SCAD (C.11). Using $\hat{h}$ picks the most complex model followed by $\widehat{ERM}_1$, $\widehat{ERM}_2$, ALASSO, BIC and SCAD.

For this dataset, it is important to observe that all model selection techniques disagree. The natural question is: Which one should we choose? We also note that ALASSO, SCAD, BIC, $\widehat{ERM}_1$ and $\widehat{ERM}_2$ are submodels of the model chosen by $\hat{h}$, so we argue that the most reasonable model is the one given by $\hat{h}$.

Overall, the complexity of the model varies from one location to another one; this can be observed by looking at the estimate of $\hat{h}$, and the ordering of terms in the model. the least complex location is Lincoln 2001 since $\hat{h} = 1$ for all model lists. We observe that the most complex location is Sidney 2001, because most of the model selections pick a model with at least 10 parameters except SCAD that pick only $TKWT \cdot KPSM$. In Sidney 2001, the model chosen by $\hat{h}$ is as the aggregate of the other model lists. We also note that, usually, SCAD, ALASSO, and BIC pick the most simplified model and that

$\widehat{ERM}_1$ and $\widehat{ERM}_2$ are erratic. By contrast, $\hat{h}$ is consistently reasonable. Thus, in general, we argue that $\hat{h}$ gives good results –typically better than the other techniques we have implemented here. Bluntly, our method never give gives crazy or stupid models; the other techniques often do.

Bibliography

M. Anthony and P. L. Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

R. Berk, L. Brown, A. Buja, K. Zhang, L. Zhao, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

B. Campbell, P. S. Baenziger, K. Gill, K. M. Eskridge, H. Budak, M. Erayman, I. Dweikat, and Y. Yen. Identification of qtls and environmental interactions associated with agronomic traits on chromosome 3a of wheat. *Crop Science*, 43(4):1493–1505, 2003.

M. Clyde and E. I. George. Model uncertainty. *Statistical science*, pages 81–94, 2004.

L. Devroye. Gy orfi, l., lugosi, g., 1996. a probabilistic theory of pattern recognition. *Applications of mathematics*, 31, 1996.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

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

N. Ebrahimi, E. Maasoumi, and E. S. Soofi. Ordering univariate distributions by entropy and variance. *Journal of Econometrics*, 90(2):317–336, 1999.

N. Ebrahimi, E. S. Soofi, and R. Soyer. On the sample information about parameter and prediction. *Statistical Science*, pages 348–367, 2010.

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. doi: 10.1198/016214501753382273. URL `http://dx.doi.org/10.1198/016214501753382273`.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

D. McDonald. Generalization error bounds for time series. phd thesis, department of statistics, carnegie-mellon university, 2012.

D. J. McDonald, C. R. Shalizi, and M. Schervish. Estimated vc dimension for risk bounds. *arXiv preprint arXiv:1111.3404*, 2011.

A. W. Moore. Vc-dimension for characterizing classifiers. *Tutorial at http://www-2. cs. cmu. edu/awm/tutorials/vcdim08. pdf*, 2001.

M. R. D. M. R. L. T. D. of Primary Industry and H. T. . A. c. W. N. . . . w. Fisheries, Tasmania GPO Box 619F. Abalone data set. `https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/?C=D;O=D`, 1996.

D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

J. Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88 (422):486–494, 1993.

X. Shao, V. Cherkassky, and W. Li. Measuring the vc-dimension using optimized experimental design. *Neural computation*, 12(8):1969–1986, 2000.

E. S. Soofi. Capturing the intangible concept of information. *Journal of the American Statistical Association*, 89(428):1243–1254, 1994.

S. van de Geer. Empirical processes in m-estimation. cambridge series in statistical and probabilistic mathematics, 2000.

A. W. Van Der Vaart and J. A. Wellner. *Weak Convergence*. Springer, 1996.

A. Van der Waart. Asymptotic statistics, volume 27 of cambridge series in statistical and probabilistic mathematics 03, 1998.

V. Vapnik. Statistical learning theory. 1998, 1998.

V. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies to their probabilities. In *Soviet Math. Dokl*, volume 9, pages 915–918, 1971.

V. Vapnik, E. Levin, and Y. Le Cun. Measuring the vc-dimension of a learning machine. *Neural Computation*, 6(5):851–876, 1994.

V. N. Vapnik. An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5): 988–999, 1999.

V. N. Vapnik and A. J. Chervonenkis. The necessary and sufficient conditions for consistency of the method of empirical risk. *Pattern Recognition and Image Analysis*, 1(3):284–305, 1991.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Soviet Math. Dokl*, volume 9, pages 915–918, 1968.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101 (476):1418–1429, 2006a.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101 (476):1418–1429, 2006b.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.