Fall 2013

Best-subset model selection based on multitudinal assessments of likelihood improvements

Knute Derek Carter
University of Iowa

BEST-SUBSET MODEL SELECTION BASED ON MULTITUDINAL ASSESSMENTS OF LIKELIHOOD IMPROVEMENTS

by

Knute Derek Carter

A thesis submitted in partial
fulfillment of the requirements for the
Doctor of Philosophy degree in Biostatistics
in the Graduate College of The University of Iowa

December 2013

Thesis Supervisor: Professor Joseph E. Cavanaugh

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Knute Derek Carter

has been approved by the Examining Committee for
the thesis requirement for the Doctor of Philosophy
degree in Biostatistics at the December 2013 graduation.

Thesis Committee: _____
Joseph Cavanaugh, Thesis Supervisor

_____
Kathryn Chaloner

_____
William Clarke

_____
Jane Pendergast

_____
Joseph Lang

_____
Eric Foster

In memory of
Dr. Jane Chalmers
1965–2008

If we set ourselves the problem, in its essence one of frequent occurrence, of finding the arbitrary elements in a function of known form, which best suit a set of actual observations, we are met at the outset by an arbitrariness which appears to invalidate any results we obtain. In the general problem of fitting a theoretical curve, either to an observed curve, or to an observed series of ordinates, it is, indeed, possible to specify a number of different standards of conformity between the observations and the theoretical curve, which definitely lead to different though mutually approximate results. This mutual approximation, though convenient in practice in that it allows a computer to make a choice of the method which is arithmetically simplest, is harmful from the theoretical standpoint as tending to obscure the practical discrepancies, and the theoretical indefiniteness which actually exist.

R. A. Fisher, On an absolute criterion
for fitting frequency curves.

ACKNOWLEDGMENTS

Firstly, I would like to acknowledge my late wife Dr. Jane Chalmers. If it were not for her I would never have made my way across the world to The University of Iowa. You told me that I was to continue on with life and finish my Ph.D. Here it is! I wish you could read it.

My knowledge and understanding of statistics and biostatistics has developed substantially thanks to the wonderful faculty I have had the privilege to learn from. Two faculty in particular I would like to recognize are Professor Mike Jones and Professor Ying Zhang from whom I learnt so much. I cannot express enough how grateful I am to the entire biostatistics department for their support through the most difficult sections of my journey. Thank you to Terry Kirk, the source of departmental knowledge, always there to help.

Professor Cavanaugh. Joe. Thank you for introducing me to the area of model selection. You have been my guiding hand, allowing me to pursue my ideas, travelling the path with me. As you know, this thesis is not where we began, the many hours spent in discussion with you have brought us here. I am proud of what we have achieved. You are an outstanding academic and mentor.

My sincere appreciation to my thesis committee for your thoughts and feedback. Professor Kathryn Chaloner, thank you for leading a terrific department and for your support over the years. Professor Bill Clarke, my initial academic advisor, ever-mindful of practical utility. Professor Jane Pendergast, my first departmental contact, and a source of sound advice. Professor Joe Lang, many wonderful ideas, always inspiring to talk with. Professor Eric Foster, for carefully checking my work. Every question helps me think about my work in a different way.

Thank you to the Public Policy Center for generously supporting me over the years as a graduate research assistant; and to Betsy Momany and all who gave me

their support when I needed it most.

To Mum and Dad, thank you for all you have done for me throughout my life; reading to me as child, encouraging me to attend Urrbrae, picking me up late at night from university, supporting our decision to relocate to Iowa. You have always been there for me, and I would not be the person I am without you. Gran, I love you. Your constant busyness and just get it done attitude is an inspiration. You can finally call me doctor.

Finally, a big thank you to my family Jean, Wade, Ryan, Isaac, and Ellis, for the love and happiness they bring to my life. My dear wife Jean Willard, you are a wonderful mother, friend, and companion 6). Thank you for reading through my thesis and providing feedback, despite the Greek nomenclature, and for listening to my occassional frustrations. You have been a constant source of support over these past few years as I have worked to complete this thesis—thank you.

Wade and Ryan, you have been there from the start, you have been a wonderful distraction, I am finished now.

ABSTRACT

Given a set of potential explanatory variables, one model selection approach is to select the best model, according to some criterion, from among the collection of models defined by all possible subsets of the explanatory variables. A popular procedure that has been used in this setting is to select the model that results in the smallest value of the Akaike information criterion (AIC). One drawback in using AIC is that it can lead to the frequent selection of overspecified models. This can be problematic if the researcher wishes to assert, with some level of certainty, the necessity of any given variable that has been selected.

This thesis develops a model selection procedure that allows the researcher to nominate, *a priori*, the probability at which overspecified models will be selected from among all possible subsets. The procedure seeks to determine if the inclusion of each candidate variable results in a *sufficiently improved fitting term*, and hence is referred to as the SIFT procedure. In order to determine whether there is sufficient evidence to retain a candidate variable or not, a set of threshold values are computed. Two procedures are proposed: a naive method based on a set of restrictive assumptions; and an empirical permutation-based method.

Graphical tools have also been developed to be used in conjunction with the SIFT procedure. The graphical representation of the SIFT procedure clarifies the process being undertaken. Using these tools can also assist researchers in developing a deeper understanding of the data they are analyzing.

The naive and empirical SIFT methods are investigated by way of simulation under a range of conditions within the standard linear model framework. The performance of the SIFT methodology is compared with model selection by minimum AIC; minimum Bayesian Information Criterion (BIC); and backward elimination based on p-values. The SIFT procedure is found to behave as designed—asymptotically

selecting those variables that characterize the underlying data generating mechanism, while limiting the selection of false or spurious variables to the desired level.

The SIFT methodology offers researchers a promising new approach to model selection, whereby they are now able to control the probability of selecting an overspecified model to a level that best suits their needs.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

> Causas rerum naturalium non plures admitti debere, quàm quæ & vera sint & earum Phænomenis explicandis sufficiunt.
>
> Isaac Newton, Philosophiæ Naturalis Principia Mathematica.

1.1 Principles and Philosophy of Model Selection

Suppose we have information available on n cases or subjects, indexed by i. For each case there is a response variable y_i of interest, and a vector $\boldsymbol{z}_i$, consisting of observations on p potential explanatory variables.

Statistical modeling arises when we wish to characterize the response variable y_i, in some sensible manner in terms of the $\boldsymbol{z}_i$. When y_i is stochastic, a common choice is to model the mean function, $\mathrm{E}\,[y_i|\boldsymbol{z}_i]$. We may propose a parametric model involving a set of unknown parameters, $\boldsymbol{\theta}$, and use a function $M(\boldsymbol{z}_i, \boldsymbol{\theta})$ to define our model.

If the true form of $M(\boldsymbol{z}_i, \boldsymbol{\theta})$ is known, then the problem is estimation of $\boldsymbol{\theta}$, and model selection is not an issue. The problem of model selection arises when we have a collection of models to consider, $M_1, M_2, \ldots, M_k$, and we wish to select one model as "best" from among them.

This raises the question, "in what manner do we define 'best'?" A seemingly natural way to proceed would be to select the model that, with respect to some measure, most closely approximates the observed data y_i. There is, however, a problem with this approach: through the construction of ever-more complicated models we can increase the fit of the model. Thus, in general, we will be guided towards selecting a complex model, one that is overly tailored to the data at hand. Such a model will likely incorporate spurious features peculiar to the data used in selecting the model. These features are not informative of the phenomenon that we

ultimately wish to understand.

This leads us to the concept of parsimony. The introductory quote to this chapter translates to, "No more causes of natural things should be admitted than such as are both true and sufficient to explain the phenomena." We do not just wish to find a model that fits the current data well, we wish to find a model that provides us with information about the truth of the phenomenon we are studying. We should strive to identify and eliminate from our model anything that is not needed to describe the data we are analyzing. The notion that we should use the smallest model that suitably encapsulates the system being studied is the principle of parsimony.

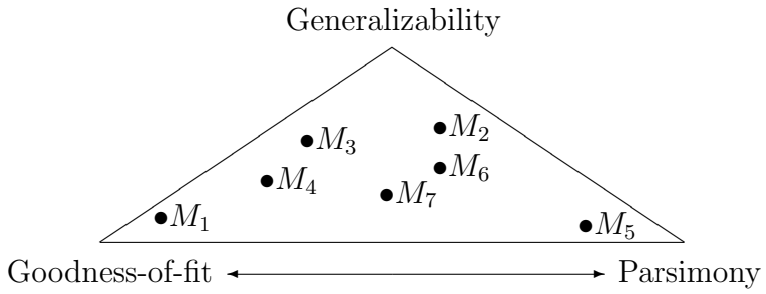

Figure 1.1: Competing principles of model selection.

In statistical modeling there is a tension between how well a model fits the observed data, its goodness-of-fit, and how parsimonious the model is. A schematic depiction of competing principles of model selection is presented in Figure 1.1. In the statistical setting, parsimony is often conceptualized in terms of the number of variables contained in the model; that is, the model size. A smaller model corresponds to greater parsimony. It should be noted that using this definition, it is possible for two equally parsimonious models to be very different in their goodness-of-fit. A correctly specified model should have a much better goodness-of-fit when compared

with a model having the same number of variables, none of which are related to the outcome.

Generalizability is a third principle within the model selection philosophy. Typically, generalizability is considered to be how well a particular model will apply to some future or alternate data set. A correctly specified model should be more generalizable than a model that is either missing key variables or includes spurious variables. Thus, two models of equal parsimony may result in different levels of generalizability.

The preceeding suggests, there is no definitive "best" model: optimality is a balance between goodness-of-fit, parsimony, and generalizability, while attempting to exclude spurious information.

1.2 Commonly Used Model Selection Methods

Numerous model selection procedures and variants have been proposed across a range of statistical frameworks. Many model selection procedures can be characterized as being synonymous with variable selection. That is, given a particular model structure, which subset of potential variables should we select to represent our final model? Hotelling (1940) provides an early exposition of the selection of variables for use in prediction.

Possibly the most well known model selection criterion is the Akaike information criterion (AIC), introduced by Akaike (1973, 1974). The AIC linked the information theoretic criterion of Kullback and Leibler (1951) with the principle of maximum likelihood. The AIC was developed as an asymptotically unbiased estimator of the expected Kullback–Leibler discrepancy between the true data generating mechanism and the fitted model.

For small sample sizes, AIC is biased in the estimation of the Kullback–Leibler discrepancy. Within the standard linear model regression framework, Sugiura (1978)

introduced a bias-corrected version of AIC, which is denoted as AICc. Hurvich and Tsai (1989) extended AICc to other regression and time series frameworks, and later examined the performance of AICc for underspecified models (Hurvich and Tsai, 1991). The derivations and justifications of AIC and AICc were developed within differing frameworks; a unified derivation of the two criteria was provided by Cavanaugh (1997).

Another popular selection criterion in the linear regression framework is the conceptual predictive statistic, C_P, of Mallows (1973, 1995). C_P was originally introduced by Gorman and Toman (1966) for the selection of variables when considering all possible subsets. Within the linear model framework, Fujikoshi and Satoh (1997) introduced modified versions of both AIC and C_P. These modified versions, MAIC and MC_P, were developed to achieve smaller bias for underspecified models. Asymptotically, AIC, AICc, MAIC, C_P and MC_P all yield equivalent model selections.

The Bayesian information criterion, developed by Schwarz (1978), is the result of a large-sample Bayesian approach to the model selection problem. BIC has the same general form as AIC, however, its penalty term is not a constant multiple of the model dimension, but rather a multiple that depends on the sample size. For samples greater than seven, the penalty term for BIC is larger than the penalty for AIC. Kass and Raftery (1995) demonstrated the link between BIC and Bayes factors. Sometimes BIC is referred to as the Schwarz criterion or SIC.

Much of the work related to AIC, BIC, and their variants has been conducted under the assumption that there exists an underlying true model of finite dimension. Shibata (1980, 1981) examined properties of selection criteria under the assumption that the regression function has infinitely many nonzero parameters, or an increasing number of variables as the sample size increases. Under this assumption, it was demonstrated that AIC is *asymptotically efficient*, meaning that asymptotically AIC

will select the model that minimizes the mean squared error of prediction. BIC is not asymptotically efficient for infinite dimensional models. In the finite setting, BIC will asymptotically select the true model with probability one if the model is within the candidate set, and is therefore referred to as *consistent*. AIC is not a consistent criterion.

The texts of Konishi and Kitagawa (2010) and McQuarrie and Tsai (1998) both provide detailed expositions of the aforementioned methods. These texts also cover a number of other model selection approaches that have not been introduced here—comparing and contrasting the array of methods.

Another branch of model selection involves model averaging, whereby a single model is not selected but several models may be combined with the hope to retain the strengths of each. Burnham and Anderson (2004) are proponents of such an approach, which they term *multimodel inference*. Yang (2005) considers model averaging in an attempt to find a criterion that can share the strengths of both AIC and BIC. Yang ultimately concludes that a criterion cannot hold both the property of consistency and asymptotic efficiency. A comprehensive account of model averaging can be found in Claeskens and Hjort (2008).

1.3 Objective and Scope of Thesis

The objective of this thesis is to develop a general model selection procedure, applicable to best-subset selection from among all possible subsets, where the probability of including spurious variables can be limited to some *a priori* nominated level. The focus will be limited to the standard linear model framework. Simulations are conducted under the assumption that there exists a true underlying data generating mechanism which can be identified from the set of models under consideration. Variables that are not part of the designated data generating mechanism will be referred to as false or spurious predictors.

The procedure developed herein has been designed primarily for analyses involving a relatively small number of potential predictor variables, perhaps fifteen to twenty at most, mainly due to computational considerations. As a general guide, to apply the procedure, the minimum sample size should exceed the greater of forty, or eight times the number of predictor variables being considered.

1.4 Overview of Thesis

In Chapter 2, concepts and results that will be referred to and used within this thesis will be introduced. Stepwise selection methods will be briefly discussed. The Kullback–Leibler information or divergence will be introduced, followed by a short overview of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The final section will cover, in some detail, results from a paper by Vuong (1989) that examined likelihood ratio tests for misspecified models.

Chapter 3 presents the details of the determination of evidence thresholds that will be later used in the *sufficiently improved fitting term* (SIFT) procedure presented in Chapter 5. Two versions of threshold determination are presented: a naive version based on a set of simplistic assumptions; and an empirically-based version to account for greater complexity within the data.

Graphical analysis tools that were developed are described in Chapter 4. The proposed tools are for use in conjunction with the SIFT procedure for best-subset model selection. Three novel plots will be presented: the deviance plot; the likelihood ratio plot; and the likelihood ratio model identification plot.

In Chapter 5, the SIFT procedure is introduced. The general concept is first described, and a general notation is defined to accommodate either the naive or empirical thresholds. The details of the SIFT procedure is then described in a detailed, step-by-step fashion.

Chapter 6 will study the behavior of the SIFT procedure in comparison to

the methods introduced in Chapter 2. The performance measures that will be used for comparisons will be outlined, and some computational issues discussed. A range of simulation settings will be investigated: independent predictors; correlated predictors; differing false admission rates; and noisy data or small effects.

An application of the SIFT procedure to a publicly-available data set will be presented in Chapter 7, demonstrating the procedure in conjunction with the plots introduced in Chapter 4. An example that includes an interaction term will also be presented to illustrate how to handle such terms within the SIFT methodology.

Finally, in Chapter 8, an overall summary will be provided, some limitations will be discussed, and a few ideas for future directions will be suggested.

CHAPTER 2
PRELIMINARIES AND BACKGROUND MATERIAL

The content of this thesis will focus exclusively on the fitting of statistical models using standard linear models with Gaussian errors. It will be assumed that we have information available on n independent cases or subjects; these shall be indexed by i. For each case there is a response variable y_i of interest, and a vector of p potential explanatory variables $\boldsymbol{z}_i = (x_{i1}, \ldots, x_{ip})'$. The vector of responses will be denoted by $\boldsymbol{y} = (y_1, \ldots, y_n)'$, and $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})'$ will denote the data for the jth explanatory variable. The $n \times (p+1)$ matrix, $\mathbf{X} = [\mathbf{1} \;\; \boldsymbol{x}_1 \;\; \boldsymbol{x}_2 \;\; \cdots \;\; \boldsymbol{x}_p]$, will be referred to as the design matrix, where $\mathbf{1}$ is a conformable vector of ones used to estimate the intercept term. An intercept term will be present in all models considered in this thesis. Appropriate modifications to the techniques developed herein can be made by the reader if models without an intercept term are to be considered.

2.1 Stepwise Selection Methods

There are many variants of stepwise model selection, which may alternatively be referred to as stepwise regression. These methods follow algorithms that sequentially add or remove variables from the candidate model according to some specified criterion. A forward selection procedure will start with the null (or mean) model and sequentially add variables; while a backward selection procedure will start with the largest candidate model and sequentially delete variables. Bidirectional procedures are also sometimes used and allow for both the addition and removal of variables as the algorithm progresses. Common criteria that are used for such procedures include p-values obtained either from analysis of variance F-tests or Wald-based t-tests from regression analysis; Mallows' C_p statistic (Mallows, 1973); and either the Akaike or Bayesian information criterion.

One of the major objections to the use of stepwise model selection techniques

is that they are well known to be susceptible to overfitting, and the results are often presented as though they were based on an *a priori* analysis with valid p-values. Another objection is that any selected model is path-dependent, and so there may exist a better model that is not attainable via any of the algorithmic paths available. Despite these concerns, most if not all major statistical packages offer some form of stepwise selection procedure as an option.

For comparative purposes, a backward selection procedure has been included in the simulations presented in Chapter 6. The procedure used was based on the p-values obtained from a regression analysis. Initially, the largest candidate model was fit. The variable with the largest p-value greater than 0.05 was removed, and another regression analysis was performed excluding the removed variable. This process was repeated until all remaining variables in the model had a p-value less than 0.05.

2.2 Kullback–Leibler Divergence

Kullback and Leibler (1951) introduced what has come to be known as the Kullback–Leibler information or directed divergence. If we have two densities $\psi(y)$ and $f(y)$, with corresponding distribution functions F and Ψ, then the Kullback–Leibler (directed) divergence of distribution F from Ψ is defined as

$$\begin{aligned} D_{KL}(\Psi||F) &= \int \ln\left(\frac{\psi(y)}{f(y)}\right) d\Psi \\ &= \int \ln \psi(y)\, d\Psi - \int \ln f(y)\, d\Psi \\ &= -H(\Psi) + H(\Psi, F) \\ &\geq 0. \end{aligned}$$

The term $H(\Psi) = -\int \ln \psi(y)\, d\Psi$ is known as the entropy of Ψ; and $H(\Psi, F)$ as the cross-entropy of F with Ψ. For distributions Ψ and F, the cross-entropy of F with Ψ can never be less than the entropy of Ψ, and hence the Kullback–Leibler divergence

is always non-negative. Equality of the entropy and cross-entropy occurs if and only if ψ and f are almost everywhere equal, which will result in a Kullback–Leibler divergence of zero. Kullback and Leibler characterize the quantity $D_{KL}(\Psi||F)$ as the mean information per observation for discriminating between the distributions Ψ and F under Ψ. Hence, if F had a very similar distribution to Ψ, then the mean information available for discrimination will be small when compared to the mean information that would be available if F was radically different from Ψ. Note that the Kullback–Leibler divergence is asymmetric since it is dependent on the assumed reference distribution Ψ, and if the roles of Ψ and F were to be interchanged, then so would the reference distribution under which the expectations are defined.

2.3 Akaike Information Criterion (AIC)

In what became a landmark paper, Hirotugu Akaike proposed a model selection procedure that linked the information theoretic criterion of Kullback and Leibler with the classical maximum likelihood principle (Akaike, 1973). Akaike described his work as an extension of the maximum likelihood principle and developed what is now known as the Akaike information criterion or AIC. For a parametric model with density $f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k$ is a vector of k unknown parameters, AIC is defined as minus twice the log-likelihood plus twice the number of parameters that are being estimated. Thus, if $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator for $\boldsymbol{\theta}_k$, and $\mathcal{L}(\hat{\boldsymbol{\theta}}_k|\boldsymbol{y};\mathbf{X})$ is the log-likelihood of the data evaluated at $\hat{\boldsymbol{\theta}}_k$, then

$$\begin{aligned} \text{AIC} &= -2\sum_{i=1}^{n} \ln f(y_i|\boldsymbol{z}_i;\hat{\boldsymbol{\theta}}_k) + 2k \\ &= -2\mathcal{L}(\hat{\boldsymbol{\theta}}_k|\boldsymbol{y};\mathbf{X}) + 2k. \end{aligned} \tag{2.1}$$

The $-2\mathcal{L}(\hat{\boldsymbol{\theta}}_k|\boldsymbol{y};\mathbf{X})$ term is a goodness-of-fit term and gets smaller for better fitting models. The problem in using only the goodness-of-fit to select a model is that the addition of any additional complexity to a model will result in an improvement in

this term. Thus the need for the additional $2k$ term in (2.1), which is often referred to as the penalty term. The penalty term is an asymptotic bias correction term, so that in large-sample settings, AIC is an approximately unbiased estimator of the Kullback–Leibler discrepancy between the fitted model and the true data generating model.

The penalty adjustment of AIC is obtained via asymptotic arguments and thus its propriety depends on a large sample size. For small sample sizes, the penalty of $2k$ is smaller than the true bias adjustment necessary. The corrected AIC of Hurvich and Tsai (1989) provides a penalty term that is unbiased regardless of sample size for the Gaussian linear model framework.

An inherent feature of AIC is that it favors overfitting to underfitting. Thus, even under an ideal setting (a finite number of explanatory variables that form the data generating mechanism; a finite number of superfluous variables; and a nested set of models to select from), there will always be a non-zero probability of selecting superfluous variables, no matter what the sample size. For the most optimistic case with only one superfluous variable, the probability of its inclusion is approximately 0.14. Shibata (1980, 1981) showed that under certain conditions with an infinite dimensional system, that AIC is asymptotically efficient, meaning that it will aymptotically select the model with the smallest prediction error.

One further cautionary issue with AIC is that its development assumes that the models under consideration are either correctly specified or overspecified. Thus, the validity of AIC may be difficult to gauge under a scenario where none of the models under consideration are able to capture the true underlying data generating mechanism. For model selection purposes, it is commonly argued that for under-specified models, the contribution from the goodness-of-fit term will overwhelm the penalty term for practical purposes.

The bias correction implemented with AIC when comparing *two* correctly

or overspecified nested models is asymptotically correct. Now suppose we have a collection of overspecified models that are all based on the same number of variables. The AIC for any one of these models will be an asymptotically unbiased estimator of the Kullback-Leibler divergence between the fitted model and the correctly specified model. However, if we select the model with the smallest AIC from among this collection, then the minimum AIC value will no longer provide an unbiased estimator. Thus, the use of the minimum AIC as a selection criterion, when there are multiple models of the same size, cannot be justified in terms of its relationship with the Kullback–Leibler divergence. Despite this known shortcoming, the use of unadjusted AIC values for best-subset selection of models continues to be used.

2.4 Bayesian Information Criterion (BIC)

The derivation of the Bayesian information criterion by Schwarz (1978) implements a Bayesian approach under which the model to be selected is the model that is *a posteriori* most probable. The asymptotic result obtained is not dependent on what prior was applied to the models or to the model parameters under consideration. Despite a very different derivation to that of AIC, the BIC shares the same basic form—minus twice the log-likelihood plus a penalty that is a multiple of the number of parameters that are being estimated. The BIC is defined as

$$\text{BIC} = -2\mathcal{L}(\hat{\boldsymbol{\theta}}_k|\boldsymbol{y}; \mathbf{X}) + k \ln n.$$

For a sample size greater than seven, BIC will have a larger penalty term than AIC. Within a finite dimensional setting, if a correctly specified model is under consideration, then asymptotically, BIC will choose the correct model with probability one. Hence, BIC is referred to as a consistent model selection procedure. Under the infinite dimensional conditions of Shibata (1980, 1981), BIC does not have the asymptotic efficiency property.

2.5 Likelihood Ratio Tests for Misspecified Models

In this section, key results from a paper by Vuong (1989) are introduced. This work builds upon the results of earlier works that examined maximum likelihood estimation of misspecified models (Wald, 1943; White, 1982; Lien and Vuong, 1987). Vuong extended the theoretical framework of the likelihood ratio test in two important ways. Most importantly, he relaxed the assumption that one of the models under consideration correctly encapsulates the data generating mechanism. Under the relaxed condition, the likelihood ratio test then becomes a test of whether the two competing models are equally close to the (unknown) data generating mechanism, against a hypothesis that one model is closer. Properties of these tests can be characterized in terms of the Kullback–Leibler divergence. The second extension naturally follows the first, and is the generalization of the likelihood ratio test to the case of non-nested models, either strictly non-nested models or overlapping models. Asymptotic results are derived for each of these new classes of likelihood ratio tests. The procedure developed in this thesis rests only upon the theoretical results pertaining to the case of strictly nested models. The necessary regularity conditions and relevant results are summarized here.

Suppose we have information on n cases. For each case, $i = 1, \ldots, n$, we have a response Y_i, and a vector of covariates $\boldsymbol{Z}_i$. The following five assumptions (regularity conditions) are taken to be satisfied.

Assumption 1

(a) The data are i.i.d. with common true distribution Ψ, and

(b) $\Psi_{Y|\boldsymbol{Z}}(y_i|\boldsymbol{z}_i)$ has density $\psi(y_i|\boldsymbol{z}_i)$ relative to some measure ν_Y.

We shall consider that we have two competing conditional parametric family

models:

$$\begin{aligned} \mathbf{F}_\theta &\equiv \{F_{Y|\boldsymbol{Z}}(y_i|\boldsymbol{z}_i;\boldsymbol{\theta}); \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset R^p\} \quad \text{and} \\ \mathbf{G}_\gamma &\equiv \{G_{Y|\boldsymbol{Z}}(y_i|\boldsymbol{z}_i;\boldsymbol{\gamma}); \boldsymbol{\gamma} \in \boldsymbol{\Gamma} \subset R^q\}. \end{aligned}$$

Later in this section it will be assumed that $\mathbf{G}_\gamma \subset \mathbf{F}_\theta$, implying $\boldsymbol{\Gamma} \subset \boldsymbol{\Theta}$; however the results hold more broadly, and so this more general notation will be retained for the interim exposition. Furthermore, unless otherwise noted, any notation or conditions ascribed to $\mathbf{F}_\theta$ are assumed to also extend to $\mathbf{G}_\gamma$.

Assumption 2

(a) $F_{Y|\boldsymbol{Z}}(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})$ has density $f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})$ relative to ν_Y, and

(b) $f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$.

Assumption 3

(a) $|\ln f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})|$ is dominated by an integrable function, independent of $\boldsymbol{\theta}$, and

(b) $\int \ln f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})\, d\Psi$ has a unique maximum on $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}_*$.

The value $\boldsymbol{\theta}_*$ will be referred to as the pseudo-true value of $\boldsymbol{\theta}$ for the conditional model family $\mathbf{F}_\theta$. Furthermore, we shall define a shorthand notation such that $F_* \equiv F_{Y|Z}(y_i|\boldsymbol{z}_i;\boldsymbol{\theta}_*)$.

Assumption 4

(a) $\ln f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})$ is twice continuously differentiable on $\boldsymbol{\Theta}$, and

(b) there exist integrable functions, independent of $\boldsymbol{\theta}$, that dominate

$$\left|\frac{\partial^2 \ln f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right| \quad \text{and} \quad \left|\frac{\partial \ln f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \ln f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right|.$$

Assumption 4 ensures the existence of the information matrices

$$\begin{aligned} A_f(\boldsymbol{\theta}) &\equiv \mathrm{E}_\psi\left[\frac{\partial^2 \ln f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \quad \text{and} \\ B_f(\boldsymbol{\theta}) &\equiv \mathrm{E}_\psi\left[\frac{\partial \ln f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \ln f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right], \end{aligned}$$

where E_ψ is the expectation taken under Ψ.

Assumption 5

(a) $\boldsymbol{\theta}_*$ is an interior point of $\boldsymbol{\Theta}$, and

(b) $A_f(\boldsymbol{\theta}_*)$ is non-singular.

Under Assumptions 1–5, the (quasi) maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ exists, and furthermore it can be shown that

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) \xrightarrow{D} \mathrm{N}(\mathbf{0}, \mathrm{A}_\mathrm{f}^{-1}(\boldsymbol{\theta}_*)\mathrm{B}_\mathrm{f}(\boldsymbol{\theta}_*)\mathrm{A}_\mathrm{f}^{-1}(\boldsymbol{\theta}_*)).$$

Thus the (quasi) maximum likelihood estimator is consistent for the pseudo-true value of $\boldsymbol{\theta}$.

The likelihood ratio statistic for the model $\mathbf{F}_\theta$ against the model $\mathbf{G}_\gamma$ is defined as

$$\Lambda_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\gamma}}_n) \equiv -2\sum_{i=1}^{n} \ln \frac{g(y_i|\boldsymbol{z}_i; \hat{\boldsymbol{\gamma}}_n)}{f(y_i|\boldsymbol{z}_i; \hat{\boldsymbol{\theta}}_n)}.$$

It can be shown that

$$\begin{aligned}
\frac{\Lambda_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\gamma}}_n)}{2n} &\xrightarrow{a.s.} E_\psi\left[\ln \frac{f(y_i|\boldsymbol{z}_i; \boldsymbol{\theta}_*)}{g(y_i|\boldsymbol{z}_i; \boldsymbol{\gamma}_*)}\right] \\
&= E_\psi\left[\ln \frac{\psi(y_i|\boldsymbol{z}_i)}{g(y_i|\boldsymbol{z}_i; \boldsymbol{\gamma}_*)}\right] - E_\psi\left[\ln \frac{\psi(y_i|\boldsymbol{z}_i)}{f(y_i|\boldsymbol{z}_i; \boldsymbol{\theta}_*)}\right] \\
&= D_{KL}(\Psi||G_*) - D_{KL}(\Psi||F_*) \\
&= H(\Psi, G_*) - H(\Psi, F_*). \qquad (2.2)
\end{aligned}$$

For the remainder of the chapter it will be assumed that $\mathbf{G}_\gamma$ is nested within $\mathbf{F}_\theta$. Note that under this condition, $\mathbf{G}_\gamma$ must be a member of the same distributional family defined by $\mathbf{F}_\theta$. Under most practical implementations this would also result in a parameterization such that $\boldsymbol{\Gamma} \subset \boldsymbol{\Theta}$.

Under the null hypothesis that the two competing model parameterizations are in fact equivalent at their pseudo-true values, that is, $H_0 : g(y_i|\boldsymbol{z}_i; \boldsymbol{\gamma}_*) \overset{a.e.}{=} f(y_i|\boldsymbol{z}_i; \boldsymbol{\theta}_*)$,

then

$$\Lambda_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\gamma}}_n) \xrightarrow{D} Q_p(\cdot),$$

where $Q_p(\cdot)$ is the distribution function of a weighted sum of p independent χ^2_1 variables, with weights defined by F and G. The exact form of $Q_p(\cdot)$ can be found in Vuong (1989). The specific details are not presented here, since for our purposes we only require the result that the likelihood ratio converges to some well-behaved distributional form, as has been established by Vuong.

If $f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta}_*) \overset{a.e.}{=} \psi(y_i|\boldsymbol{z}_i)$, that is the larger model appropriately captures the underlying data generating mechanism, then under H_0 we get the familiar result of Wilks (1938),

$$\Lambda_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\gamma}}_n) \xrightarrow{D} \chi^2_{p-q}.$$

Under the alternative hypothesis that the competing models do not have pseudo-true equivalence, $H_A : g(y_i|\boldsymbol{z}_i;\boldsymbol{\gamma}_*) \neq f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta}_*)$, then

$$\Lambda_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\gamma}}_n) \xrightarrow{a.s.} +\infty.$$

For the remainder of this thesis we shall be restricting our attention to the standard linear model framework. Hence, the models we will be examining are defined in terms of a linear combination of some subset of the potential explanatory variables. Candidate models can therefore be distinguished by the set of variables they include. The following notation shall be used from this point forward to provide an implicit referential identification of models. Let $\mathcal{M}_{k,j}$ be a model that includes variable x_j and let $\mathcal{M}_k$ be the corresponding model without x_j, and let their respective p.d.f.s be $f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta})$ and $g(y_i|\boldsymbol{z}_i;\boldsymbol{\gamma})$. We shall then define

$$\begin{aligned} \delta_{k,j} &\equiv \widehat{\Lambda}_n(\mathcal{M}_{k,j}, \mathcal{M}_k) \\ &\equiv \Lambda_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\gamma}}_n). \end{aligned}$$

If $g(y_i|\boldsymbol{z}_i;\boldsymbol{\gamma}_*) \overset{a.e.}{=} f(y_i|\boldsymbol{z}_i;\boldsymbol{\theta}_*)$, then we will say the models have pseudo-true equivalence, which we shall denote by $\mathcal{M}_k^* = \mathcal{M}_{k,j}^*$. Similarly, we will define $D_{KL}(\Psi||\mathcal{M}_{k,j}^*) \equiv D_{KL}(\Psi||F_*)$ and $D_{KL}(\Psi||\mathcal{M}_k^*) \equiv D_{KL}(\Psi||G_*)$. That is, we will allow the interchange of notation that describes the model under consideration with the pseudo-true distribution that is induced by that model. If the data generating mechanism has p.d.f. $\psi(\cdot)$, we will assign it the model notation $\mathcal{M}_\psi$.

Table 2.1: Model relationship possibilities for nested models, and properties of the associated Kullback–Leibler divergences and likelihood ratio statistics.

Model Relationship	$D_{KL}(\Psi\|\|\mathcal{M}_k^*)$	$D_{KL}(\Psi\|\|\mathcal{M}_{k,j}^*)$	$\delta_{k,j}/2n$	$\delta_{k,j}$
$\mathcal{M}_k^* = \mathcal{M}_{k,j}^* = \mathcal{M}_\psi$	0	0	$\to 0$	$\to \chi_1^2$
$\mathcal{M}_k^* \neq \mathcal{M}_{k,j}^* = \mathcal{M}_\psi$	$+$	0	$\to k_1$	$\to \infty$
$\mathcal{M}_k^* = \mathcal{M}_{k,j}^* \neq \mathcal{M}_\psi$	$+$	$+$	$\to 0$	$\to Q_p(\cdot)$
$\mathcal{M}_k^* \neq \mathcal{M}_{k,j}^* \neq \mathcal{M}_\psi$	$++$	$+$	$\to k_2$	$\to \infty$

where $k_1 = H(\Psi, \mathcal{M}_k^*) - H(\Psi) > 0$ and $k_2 = H(\Psi, \mathcal{M}_k^*) - H(\Psi, \mathcal{M}_{k,j}^*) > 0$, and $+$ and $++$ are both positive numbers such that $+ < ++$.

Table 2.1 summarizes the different underlying relationships between nested models that are possible. The first two scenarios presented correspond to the situation in which the larger of the two models is correctly specified in terms of its pseudo-true parameters. A model that is correctly specified in terms of its pseudo-true parameters does not imply that all variables in the model are required; unnecessary variables will have a pseudo-true parameter value of zero, and thus we may in general terms describe such a model as overspecified.

The first case further assumes that the nested model is also correctly specified, which implies that the larger model is in fact overspecified. Under this first scenario,

the Kullback–Leibler divergence between each of these models and the true distribution is zero. The likelihood ratio statistic, $\delta_{k,j}$ converges to a chi-square distribution, and $\delta_{k,j}/2n$ converges to zero, indicating that there is no information asymptotically to distinguish between the two competing models relative to the true distribution.

The second case assumes that the pseudo-true distributions of the nested models differ. The larger of the two models again has a Kullback–Leibler divergence of zero. In this case, if the larger of the two models is correctly specified, then the nested model must be underspecified; while if the larger model is overspecified, then the nested model must be omitting a required variate. In either case, the Kullback–Leibler divergence for the nested model will be positive, and the expected average information per observation available to discriminate between these two models will be $k_1 = H(\Psi, \mathcal{M}_k^*) - H(\Psi)$.

The final two scenarios correspond to the situation in which the larger of the two models does not have pseudo-true equivalence with the true distribution. This will be the case if the true distribution is not attainable by any of the models under consideration, in which event neither of the first two scenarios are even possible. It will also be the case if the larger of the two models is either strictly underspecified, or is missing a necessary variable but includes extraneous variables.

If the nested models have the same pseudo-true distribution, then the Kullback–Leibler divergence of each from the true distribution will be positive but equal, and thus there is asymptotically no information to distinguish them. If the nested models differ in their pseudo-true distributions, then the smaller of the two will have a larger Kullback–Leibler divergence from the true distribution, and the expected average information per observation available to discriminate between these two models will be $k_2 = H(\Psi, \mathcal{M}_k^*) - H(\Psi, \mathcal{M}_{k,j}^*)$.

CHAPTER 3
EVIDENCE THRESHOLD

We shall continue to use the notation introduced in Section 2.5. That is, $\mathcal{M}_{k,j}$ will denote a model that includes variable x_j, and $\mathcal{M}_k$ will be the corresponding model for which x_j is not included. The likelihood ratio statistic comparing model $\mathcal{M}_{k,j}$ to the nested model $\mathcal{M}_k$ is defined as $\delta_{k,j}$. If there are p potential explanatory variables, then when considering all possible subsets, there will be 2^{p-1} models that do not contain variable x_j. Thus, the number of $\delta_{k,j}$ values that will be realized for variable x_j is 2^{p-1}.

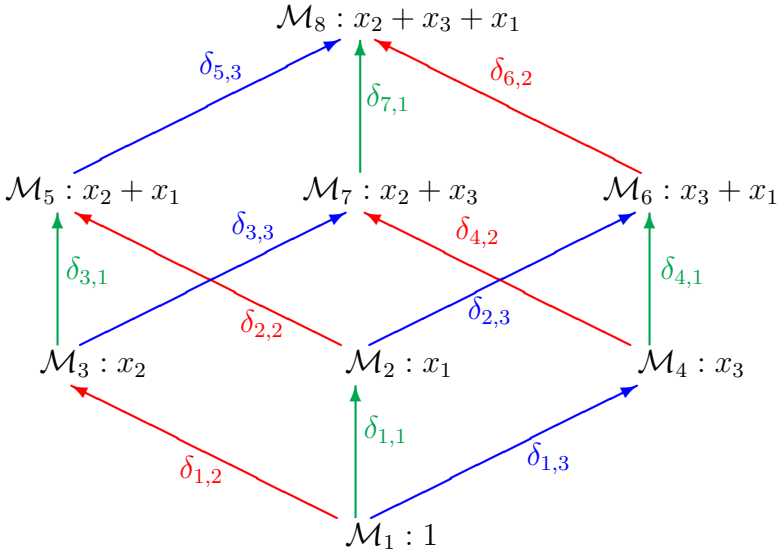

Figure 3.1: Schematic of all possible models and likelihood ratios for three explanatory variables.

Figure 3.1 illustrates the general set up for the case when there are three explanatory variables, and we are considering likelihood ratio values between models that have a difference in parameter space dimension of one. In this case, there are a total of eight potential models, resulting in four likelihood ratio statistics for each of

the three predictors. Note that the indexing used for k is arbitrary.

3.1 Null Behavior of Likelihood Ratio Statistics

The SIFT procedure developed in Chapter 5 relies on the assessment of all $p \times 2^{p-1}$ values of $\delta_{k,j}$. Given this large number of model assessments that will be made, how then should we decide what magnitude of improvement constitutes fair evidence to justify the inclusion of any given variable? This may be rephrased as the classical question: "Is there evidence that the data at hand are inconsistent with what we might observe by chance alone?" In order to answer this question there are two primary issues that need to be considered: the multiplicity of within variable assessments; and the assessment of multiple variables.

In order to address these issues, we shall consider the null model setting under which none of the predictors are intrinsically related to the response data, and the pseudo-true distribution of the null model corresponds to that of the data generating mechanism. As stated in Section 2.5, under these conditions, the asymptotic distribution of any $\delta_{k,j}$ is χ^2_1. If we further assume that the explanatory variables are uncorrelated with each other, then it can be argued that $\delta_{k,j} - \delta_{k',j} \approx 0$ for any $\mathcal{M}_k \neq \mathcal{M}_{k'}$. Thus, under these null conditions, for any given variable x_j, we can regard the set of $\delta_{k,j}$ as essentially a single observation from a χ^2_1 distribution.

Figure 3.2 exemplifies, via a numeric example, the previous assertion. Three independent variables were generated under the described null conditions, and the resulting twelve likelihood ratio statistics are presented. The likelihood ratio statistics associated with the inclusion of x_1 were the smallest, and ranged from 0.004 to 0.015. The range of values for $\delta_{k,2}$ was 0.046 to 0.051; and for $\delta_{k,3}$, was 1.660 to 1.674. So for any given explanatory variable, the variation in the likelihood ratio statistics appears to be small, as asserted. In order to reduce the complexity of the problem, it seems that treating the set of $\delta_{k,j}$ for any given j as essentially a single observation

from a χ^2_1 distribution may provide a reasonable approximation. This assumption may, however, result in some small bias to be incorporated into the ζ_p threshold values derived later in this section.

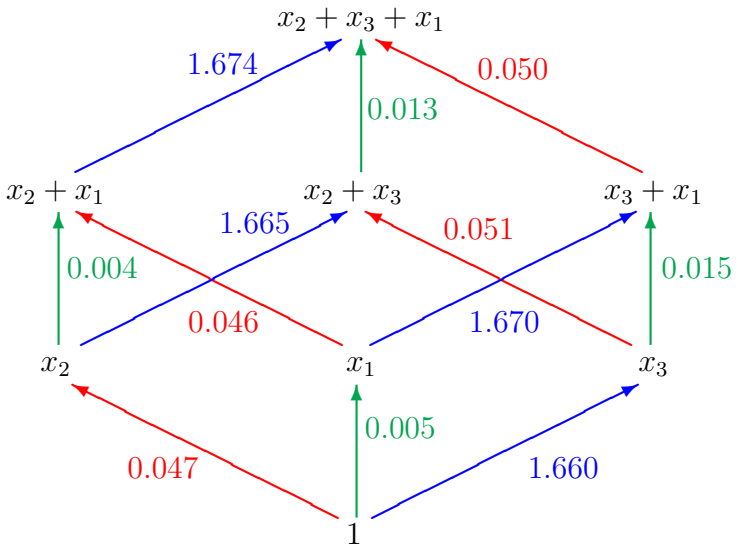

Figure 3.2: Numeric example of likelihood ratios for three explanatory variables under a null model.

The issue of multiple variable assessments is well known, and many solutions to this problem have been developed, of which the Bonferroni correction is possibly the best known. In general, our aim is to control the probability of including spurious variables in the selected model—variables that are not associated with the underlying data generating mechanism. Using methods such as backward selection, this probability will increase as the number of spurious variables being considered increases.

Under the null conditions and assumptions put forward in this section, we now consider p independent χ^2_1 variables. If we want to restrict the probability of including one or more spurious variables to be less than α, we could then use as a threshold value the $100(1-\alpha)$th percentile of the maximum of p independent χ^2_1

variables. This value can be derived as follows.

Let $X_1, \ldots, X_p$ be p independent χ^2_1 variables. Then for $j = 1, \ldots, p$, the variate X_j has density

$$f(x_j) = \frac{1}{\sqrt{2\pi}} x_j^{-1/2} e^{-x_j/2}$$

and cumulative distribution function

$$\begin{aligned} F_{X_j}(x_j) &= \frac{2}{\sqrt{\pi}} \int_0^{\sqrt{x_j/2}} e^{-t^2/2} dt \\ &= 2\Phi\left(x_j^{1/2}\right) - 1, \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Therefore, if $X_{(p)}$ is the pth order statistic of these values, then it will have density

$$f_{X_{(p)}}(x) = \frac{p}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} \left[2\Phi\left(x^{1/2}\right) - 1\right]^{p-1}$$

and cumulative distribution function

$$F_{X_{(p)}}(x) = \left[2\Phi\left(x^{1/2}\right) - 1\right]^p.$$

Now, we wish to find the threshold

$$\zeta_{p,\alpha} : F_{X_{(p)}}\left(\zeta_{p,\alpha}\right) = 1 - \alpha,$$

so

$$\Phi\left(\zeta_{p,\alpha}^{1/2}\right) = \frac{1 + \sqrt[p]{1-\alpha}}{2},$$

and therefore

$$\zeta_{p,\alpha} = \left[\Phi^{-1}\left(\frac{1 + \sqrt[p]{1-\alpha}}{2}\right)\right]^2. \tag{3.1}$$

Table 3.1 provides the threshold values, $\zeta_{p,\alpha}$, from (3.1) for α levels of 0.1, 0.05, and 0.01, and for the number of explanatory variables ranging from one through fifteen. The threshold for $\alpha = 0.05$ when $p = 1$ is the familiar value of 3.84. The

rate at which $\zeta_{p,\alpha}$ increases with respect to p decreases as p increases, the change in $\zeta_{p,\alpha}$ being relatively small for large values of p. The value α will be called the false admission probability, and $100\alpha\%$ the false admission rate.

Table 3.1: Naive $\zeta_{p,\alpha}$ threshold values for 1 to 15 predictors.

p	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	2.71	3.84	6.63
2	3.80	5.00	7.87
3	4.47	5.70	8.61
4	4.96	6.20	9.13
5	5.34	6.60	9.54
6	5.65	6.92	9.88
7	5.92	7.20	10.16
8	6.16	7.44	10.41
9	6.37	7.65	10.62
10	6.55	7.84	10.82
11	6.72	8.01	11.00
12	6.87	8.17	11.16
13	7.02	8.31	11.31
14	7.15	8.45	11.44
15	7.27	8.57	11.57

3.2 Empirically Determined Thresholds

Due to the potential lack of independence of the likelihood ratio statistics being assessed, establishing exact results for the threshold values is a non-trivial exercise.

In order to better capture the complexities of the data which were not accounted for by the threshold values proposed in the previous section, an empirically-based approach is developed here.

Assume we have data $\boldsymbol{y}$ and $\mathbf{X}$ available to us, from which we calculate the likelihood ratio statistics $\delta_{k,j}$. Let $\boldsymbol{y}^{(b)}$ be a permuted vector of the actual response data $\boldsymbol{y}$. Now let $\delta_{k,j}^{(b)}$ be the value corresponding to $\delta_{k,j}$, using $\boldsymbol{y}^{(b)}$ as the response data, in place of $\boldsymbol{y}$. Repeat this for B permutations of the response data, so that $b = 1, \ldots, B$. Permuting the data in this way will retain and account for any correlation structure embedded within the design matrix. It will also take into account any within-variable variation in the $\delta_{k,j}$ values that was assumed away in the previous section.

In the development of the SIFT procedure in Chapter 5, a variable will be determined to have sufficient evidence for inclusion if its smallest $\delta_{k,j}$ is found to be greater than some threshold value. It is that threshold value we are attempting to determine here. So for a given permutation $\boldsymbol{y}^{(b)}$, we will first need to identify the smallest likelihood ratio statistic for each variable, and then select the maximum of these. That is, let $\zeta^{(b)} = \max_j \left(\min_k \delta_{k,j}^{(b)} \right)$. To obtain the required threshold, $\hat{\zeta}_\alpha | \mathbf{X}$, we then select the $100(1-\alpha)$th percentile of the set $\{\zeta^{(1)}, \ldots, \zeta^{(B)}\}$.

In order to attain a relatively stable value of $\hat{\zeta}_\alpha | \mathbf{X}$, the number of permuted data sets, B, to be processed before evaluating the desired percentile will need to be large. For the simulations presented in Chapter 6, B was taken to be 50,000.

CHAPTER 4
GRAPHICAL ANALYSIS TOOLS

During the development of the analytical procedure presented in Chapter 5, a number of graphical analysis tools were developed to assist and inform the researcher. These plots will be introduced and explained in this chapter by way of a contrived example.

For our example, we will consider a situation in which we have four potential explanatory variables, x_1, x_2, x_3, and x_4. Suppose we have $n = 40$ cases, and let x_{ij} denote the value observed for the ith case for variable x_j. Let $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})'$ be the vector of observations for variable x_j, where $j = 1, \ldots, 4$. We shall randomly generate the x_{ij} to be independently and identically distributed (i.i.d.) from a Uniform$(0, 1)$ distribution. In order to induce a correlation between $\boldsymbol{x}_1$ and $\boldsymbol{x}_4$, we will redefine $\boldsymbol{x}_4$ to be $0.4\boldsymbol{x}_1 + 0.6\boldsymbol{x}_4$. The response variable for the ith case will be defined as $y_i = 16 + 6x_{i1} + 3x_{i2} + \varepsilon_i$, where the ε_i are randomly generated i.i.d. N$(0, 1)$ values. Thus, the underlying true data generating mechanism is dependent only on x_1 and x_2, while x_3 and x_4 are spurious variables that are not intrinsically part of the data generating mechanism.

Table 4.1: Correlation matrix of predictor and response variables for the contrived example.

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{y}$
$\boldsymbol{x}_1$	1.000	-0.102	-0.106	0.642	0.845
$\boldsymbol{x}_2$		1.000	0.331	-0.273	0.274
$\boldsymbol{x}_3$			1.000	-0.084	0.013
$\boldsymbol{x}_4$				1.000	0.445
$\boldsymbol{y}$					1.000

Table 4.1 presents the observed pairwise correlations from a data set simulated as previously described. The largest correlation observed was between x_1 and the response variable; this is unsurprising since x_1 has the largest coefficient within the data generating mechanism. A substantial correlation between x_1 and x_4 is evident, as was deliberately induced; this is further reflected in the pairwise correlation between x_4 and the response variable. Given the form of the data generating mechanism, we would expect the positive correlation observed between x_2 and the response variable. The expectations of all other correlations are zero, thus the moderately-sized correlations observed between x_2 with x_3 or x_4 do not reflect any structural relationship, and are due entirely to sampling variability.

4.1 Deviance Plot

The deviance plot is a way in which to graphically portray information about the goodness-of-fit of each model under consideration. For each model, we can compute a goodness-of-fit statistic. For our development, this is negative twice the log-likelihood evaluated at the maximum likelihood estimate. When considering all possible subsets, the largest candidate model will provide the best fit to the data, and thus will have the smallest goodness-of-fit value. If we subtract that minimum value from the value obtained under each other model we will obtain what we shall call the deviance, akin to that described by Nelder and Wedderburn (1972).

If we have p variables under consideration, then the 2^p goodness-of-fit values will be displayed. For the example we are examining, there are four variables under consideration, and therefore there are a total of sixteen models. The goodness-of-fit and deviance statistics for each of these sixteen models are presented in Table 4.2.

Figure 4.1 plots the deviance values of Table 4.2. The set of vertical tick marks at any observed deviance value specifies the model that produced that deviance. For each variable, a blue tick mark above the dotted guideline indicates the inclusion of

Table 4.2: Goodness-of-fit and deviance statistics for the contrived example.

Model	Variables				$-2\mathcal{L}(\mathcal{M}_k)$	Deviance
$\mathcal{M}_1$	.	.	.	.	183.035	74.944
$\mathcal{M}_2$	.	.	x_3	.	183.029	74.938
$\mathcal{M}_5$	.	x_2	.	.	179.921	71.830
$\mathcal{M}_6$	.	x_2	x_3	.	179.624	71.533
$\mathcal{M}_3$	.	.	.	x_4	174.187	66.096
$\mathcal{M}_4$	.	.	x_3	x_4	174.062	65.971
$\mathcal{M}_7$	.	x_2	.	x_4	164.726	56.635
$\mathcal{M}_8$	.	x_2	x_3	x_4	164.249	56.158
$\mathcal{M}_9$	x_1	.	.	.	132.931	24.840
$\mathcal{M}_{10}$	x_1	.	x_3	.	131.420	23.329
$\mathcal{M}_{11}$	x_1	.	.	x_4	130.615	22.524
$\mathcal{M}_{12}$	x_1	.	x_3	x_4	129.092	21.001
$\mathcal{M}_{13}$	x_1	x_2	.	.	108.360	0.269
$\mathcal{M}_{14}$	x_1	x_2	x_3	.	108.301	0.210
$\mathcal{M}_{15}$	x_1	x_2	.	x_4	108.135	0.044
$\mathcal{M}_{16}$	x_1	x_2	x_3	x_4	108.091	0.000

Note that the model numbering is entirely arbitrary.

that variable in the model; while a gray tick mark below indicates the exclusion of that variable from the model. So from Figure 4.1, we can read that the model with a deviance of just over 21 is the model that only includes variables x_1, x_3, and x_4. The left-most model with a deviance of zero is the largest candidate model, while the right-most model corresponds to the null or mean only model.

From Figure 4.1, we can immediately see that all models that did not include x_1

fitted more poorly than did models that included x_1. If we now restrict our attention to only the subset of models that included x_1, we then make the same observation about x_2. We are unable to make any such clear inference about variables x_3 and x_4. These observations are consistent with what we know to be true about the underlying data generating mechanism.

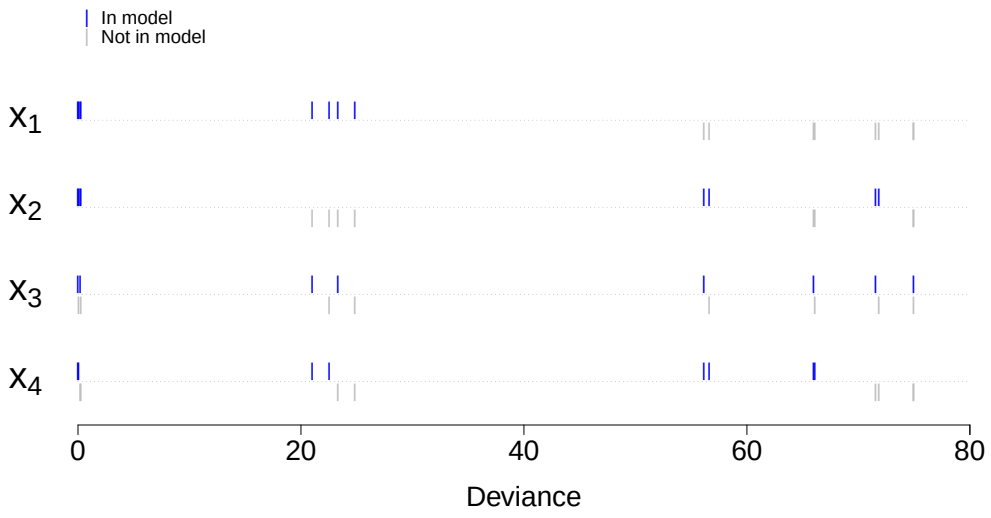

Figure 4.1: Deviance plot for the contrived example.

4.2 Likelihood Ratio Plot

The likelihood ratio plot provides a way to visualize the $\delta_{k,j}$ values, which are the likelihood ratio statistics obtained when comparing model $\mathcal{M}_k$ with $\mathcal{M}_{k,j}$. The thirty-two $\delta_{k,j}$ values from our example can be found in Table 4.3. Depending on the base model, the improvement in the goodness-of-fit statistic by the addition of variable x_1 ranged from 43.57 to 71.56. Thus, the addition of x_1 to any base model improves the fit of the model substantially. If we turn our attention to x_4, we can see that adding x_4 to models that did not include x_1 resulted in some improvement to the fit. However, when added to models that included x_1, the improvement was

small. Since x_3 is neither part of the data generating mechanism, nor related to any component of the data generating mechanism, the improvement of fit due to the inclusion of x_3 is small no matter what model is initially being considered.

Table 4.3: Likelihood ratio statistics ($\delta_{k,j}$), and estimated Kullback–Leibler divergence differences ($\delta_{k,j}/2n$) for the contrived example.

Base model				$\delta_{k,j}$				$\delta_{k,j}/2n$			
				$+x_1$	$+x_2$	$+x_3$	$+x_4$	$+x_1$	$+x_2$	$+x_3$	$+x_4$
.	.	.	.	50.10	3.11	0.01	8.85	0.626	0.039	0.000	0.111
.	.	x_3	.	51.61	3.40	.	8.97	0.645	0.043	.	0.112
.	.	.	x_4	43.57	9.46	0.13	.	0.545	0.118	0.002	.
.	.	x_3	x_4	44.97	9.81	.	.	0.562	0.123	.	.
.	x_2	.	.	71.56	.	0.30	15.20	0.895	.	0.004	0.190
.	x_2	x_3	.	71.32	.	.	15.37	0.892	.	.	0.192
.	x_2	.	x_4	56.59	.	0.48	.	0.707	.	0.006	.
.	x_2	x_3	x_4	56.16	.	.	.	0.702	.	.	.
x_1	.	.	.	.	24.57	1.51	2.32	.	0.307	0.019	0.029
x_1	.	x_3	.	.	23.12	.	2.33	.	0.289	.	0.029
x_1	.	.	x_4	.	22.48	1.52	.	.	0.281	0.019	.
x_1	.	x_3	x_4	.	21.00	.	.	.	0.263	.	.
x_1	x_2	.	.	.	.	0.06	0.23	.	.	0.001	0.003
x_1	x_2	x_3	.	.	.	.	0.21	.	.	.	0.003
x_1	x_2	.	x_4	.	.	0.04	.	.	.	0.001	.

The second part of Table 4.3, which lists the values of $\delta_{k,j}/2n$, has been included to illustrate numerically the concepts discussed earlier in relation to Table 2.1 and (2.2).

Models that did not include x_1 have a larger cross-entropy with the underlying data generating mechanism than models that included x_1. The estimated difference in the per observation information available to discriminate between models that did or did not include x_1 ranged from 0.545 to 0.895. This is contrasted with variable x_3, for which there is approximately zero information difference per observation through the omission of x_3. This indicates that models which included x_3 are as close to the 'truth' as models that excluded x_3. Therefore, x_3 is probably not required.

Figure 4.2 plots the $\delta_{k,j}$ values of Table 4.3. This plot provides the researcher with an easier-to-digest version of the data than the tabular form. A tool such as this would be even more critical if a greater number of variables were being considered. This graphical summary of the data, does not, however, allow for the identification of the underlying base reference model from which each likelihood ratio statistic was calculated. Thus, for example, we are no longer able tell that the larger values observed on the x_4 row correspond to the models in which x_1 was absent. In order to be able to recover such information graphically, the likelihood ratio model identification plot described in the next section was developed.

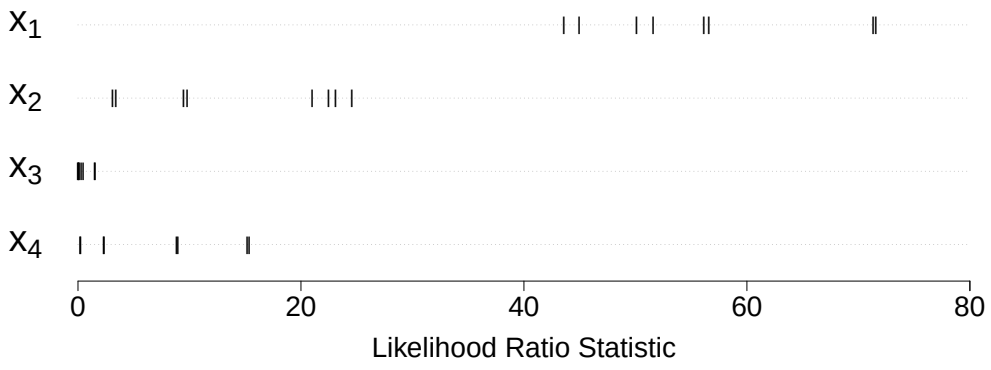

Figure 4.2: Likelihood ratio plot for the contrived example.

4.3 Likelihood Ratio Model Identification Plot

The likelihood ratio model identification plot allows the researcher to select one of the variables of interest and graphically assess the likelihood ratio statistics for that variable against the corresponding base models. As an illustration in the context of our example, we may wish to further examine the models associated with the likelihood ratio statistics for x_4. The model identification plot for x_4 is presented in Figure 4.3. The variable of interest, x_4 in our case, is now located at the top of the graph, and its likelihood ratio statistics are plotted as tick marks in red that cross the dotted guideline. The specification of the corresponding base models is achieved in a similar manner to that used for the deviance plot introduced earlier. A blue tick mark above the dotted line indicates inclusion of a variable in the model; and a gray tick mark below the line indicates exclusion from the model. Thus, for the right-most model with the largest likelihood ratio statistic, we can see that this value arises when adding x_4 to the base model of x_2 and x_3. Another feature we are able to readily discern is the observation made earlier, that the largest likelihood ratio statistics for models that include x_4 occur when x_1 is not in the base model.

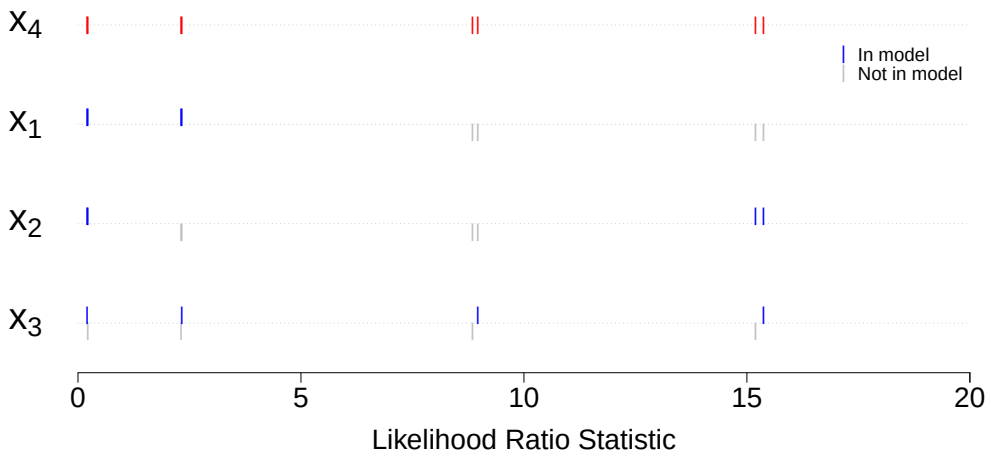

Figure 4.3: Likelihood ratio model identification plot of $\delta_{k,4}$ values for the contrived example.

The use of these graphical tools, in conjunction with subsetting options that allow for forced inclusion or exclusion of variables to the model set, provides the researcher with the ability to undertake a model selection search across all possible subsets from an entirely visual perspective.

CHAPTER 5
SUFFICIENTLY IMPROVED FITTING TERM (SIFT) PROCEDURE

5.1 Overview

In this chapter, the proposed SIFT procedure is described in detail. A brief overview summarizing the general execution of the procedure is first presented.

Initially, any variable that improves the model fit to a degree greater than a specified threshold, when added to any other model under consideration, will be admitted. After admitting this first set of variables, the number of models under consideration is reduced. The threshold is consequently adjusted, and a check is again made for variables that always improve the model fit by more than the revised threshold. This is repeated until no further variables are admitted.

For the set of models still under consideration, all variables that never improve the model fit to a degree greater than the threshold are removed, further reducing the set of models under consideration. A check is then made among the remaining models, to identify any variables that always improve the fit by more than the threshold. If so, they are admitted, and the threshold is further adjusted. The process of identifying variables for removal then admittance is repeated until no further variables are removed or admitted.

If all of the variables under consideration have been either removed or admitted, then we have identified our final model. If this is not the case, then we have a set of undecided variables. Each undecided variable is characterized as follows: (1) there exists at least one model, still under consideration, that is improved by more than the threshold by the addition of the undecided variable; and (2) there exists at least one model that is not improved by more than the threshold by the addition of the undecided variable. In order to identify a unique final model, these undecided variables will be added singly, then pairwise, and so on, until a combination is found

that results in the evidence for all remaining undecided variables to be entirely above or below the threshold. The smallest such model found will be selected, and in the case of two or more equal-sized models being identified, the model with the best fit will be selected.

5.2 Threshold Notation

In Chapter 2, two procedures to determine the evidence threshold were proposed: a naive method based on a set of restrictive assumptions; and an empirical method involving data permutation. It is hoped that the naive method will capture and reflect the essence of the distributional characteristics of the likelihood ratio statistics; and that departures from the restrictive assumptions made will be minor in their consequence for any naturally occurring data. The empirical approach retains the exact structure contained within the design matrix, and thus it should account better for departures from the assumptions of the naive method. It may, however, be too dependent on the data at hand, and be overly subject to peculiarities that may be contained within.

The SIFT procedure described in this chapter relies on using the appropriate evidence threshold. The appropriate threshold will vary as the algorithm progresses, and variables are admitted to the final model set. At any given point in the algorithm, the threshold is determined by the variables that have not been admitted into the model.

Thus, for example, if we had eight potential explanatory variables, the first naive threshold value would be $\zeta_{8,\alpha}$. If two variables were found to exceed this threshold, across the set of models under consideration, then the new revised threshold value would be $\zeta_{6,\alpha}$. If further variables were consequently added to the final variable set, the threshold would continue to be adjusted accordingly.

For the empirical approach, the threshold to be used is $\hat{\zeta_\alpha}|\mathbf{X}$, where $\mathbf{X}$ is the

design matrix comprised of all variables that have not been admitted into the final model set. Thus, for the example described in the paragraph above, the initial matrix would be $\mathbf{X}^{(0)} = [\mathbf{1} \ \boldsymbol{x}_1 \ \boldsymbol{x}_2 \ \cdots \ \boldsymbol{x}_8]$, and the threshold would be $\hat{\zeta_\alpha}|\mathbf{X}^{(0)}$. If variables $\boldsymbol{x}_2$ and $\boldsymbol{x}_5$ were the two variables initially admitted into the final model set, then the next revised design matrix to be used would be $\mathbf{X}^{(1)} = [\mathbf{1} \ \boldsymbol{x}_1 \ \boldsymbol{x}_3 \ \boldsymbol{x}_4 \ \boldsymbol{x}_6 \ \boldsymbol{x}_7 \ \boldsymbol{x}_8]$, and $\hat{\zeta_\alpha}|\mathbf{X}^{(1)}$ would be the revised threshold. This threshold would be adjusted again if further variables were added into the final variable set, as the algorithm progressed.

In order to describe the SIFT procedure, the threshold has been denoted as ξ_p in Section 5.3. If using the naive threshold, then $\xi_p = \zeta_{p,\alpha}$. If using the empirical threshold, then $\xi_p = \hat{\zeta_\alpha}|\mathbf{X}$, where the matrix $\mathbf{X}$ is defined as described previously, and $\text{rank}(\mathbf{X}) = p + 1$.

5.3 SIFT Procedure

Step 1: Initial variable inclusion

The first step identifies all of the variables that always improve the model fit by more than the currently determined threshold. Let $\Delta_j = \{\delta_{k,j}\}$ be the set of all likelihood ratio statistics for variable x_j. Let t be an iteration counter, initialized at $t = 0$. Let R initially be an empty set; R will ultimately be used to collect the set of all variables to be removed from the candidate model set. The initial set of selected variables, $G^{(0)}$, consists of those variables with a minimum likelihood ratio statistic greater than the current threshold. That is,

$$G^{(0)} = \{x_j : \min \Delta_j > \xi_p\}.$$

There are $p^{(0)} = \#G^{(0)}$ such variables, where $\#A$ denotes the cardinality of set A.

Step 2: Threshold adjustment; additional variable inclusion

The threshold ξ_p is obtained under the assumption that none of the variables under consideration are related to the response variable. For the naive method, this

threshold increases with the number of variables under consideration. In Step 1, we have identified $p^{(0)}$ variables as having sufficient evidence to include in any model we are building. If these $p^{(0)}$ variables are indeed related to the response variable, then the naive threshold will be too conservative (large) for subsequent purposes, and it would therefore be more prudent to continue under the assumption that the $p^{(0)}$ variables are required. Hence, only $p - p^{(0)}$ variables should be viewed as under consideration, and we should now use $\xi_{p-p^{(0)}}$ as a more appropriate threshold value. For the empirical method, if we have admitted $p^{(0)}$ variables, and we are now considering the remaining $p - p^{(0)}$ variables, these remaining variables should be used to determine $\xi_{p-p^{(0)}}$. Thus, the new empirical threshold would be based on the reduced design matrix as described in Section 5.2.

The set of variables that exceed the adjusted threshold, and the number of such variables, are then respectively given by

$$G^{(t+1)} = \{x_j \notin G^{(t)} : \min \Delta_j > \xi_{p-p^{(t)}}\} \cup G^{(t)}, \quad \text{and} \;\; p^{(t+1)} = \#G^{(t+1)}. \tag{5.1}$$

Step 3: Check for no further change

If $p^{(t+1)} = p^{(t)}$, then no additional variables have been found that exceed the revised threshold, so continue to Step 4. Otherwise, if $p^{(t+1)} > p^{(t)}$, increment t by 1, revise the threshold, and further update $G^{(t+1)}$ and $p^{(t+1)}$ according to (5.1) in Step 2.

Step 4: Current set definitions

The set of selected variables and the number of such variables will be respectively notated as

$$G^* = G^{(t)}, \quad \text{and} \;\; p^* = \#G^*. \tag{5.2}$$

The set of variables that are still undecided (that is, they are neither in the

selected set nor in the removed set), is notated as

$$U = \{x_j : x_j \notin G^* \cup R\}. \tag{5.3}$$

Let $\mathcal{M}^*$ denote the set of all models to be considered further. This set consists of all models that contain all of the currently selected variables; none of the removed variables; and all combinations of variables from the undecided set. That is,

$$\mathcal{M}^* = \{\mathcal{M}_k : x_j \in \mathcal{M}_k \ \ \forall x_j \in G^* \ \ \& \ \ x_j \notin \mathcal{M}_k \ \ \forall x_j \in R\}. \tag{5.4}$$

The set of likelihood ratios for each of the undecided variables still under consideration is therefore given by

$$\Delta_j^* = \{\delta_{k,j} : \mathcal{M}_k \in \mathcal{M}^*, x_j \in U\}. \tag{5.5}$$

Step 5: Identify variables for removal followed by inclusion

Having identified a set of variables for inclusion, we will now identify variables that do not have sufficient evidence to warrant further consideration. Such variables have a maximum likelihood ratio statistic across all models still under consideration that is less than the currently specified threshold. The set R is the set of variables that have been removed due to insufficient evidence, and will be updated as

$$R \leftarrow R \cup \{x_j : x_j \in U, \ \max \Delta_j^* < \xi_{p-p*}\}.$$

Having now possibly removed some variables from consideration, we will update $U, \mathcal{M}^*$, and Δ_j^* using (5.3), (5.4), and (5.5) of Step 4.

The removal of candidate variables may result in some variables of the undecided set to have a minimum likelihood ratio statistic across the remaining models under consideration that exceeds the current threshold value. Any such variables will be identified by updating G^*, according to

$$G^* \leftarrow G^* \cup \{x_j : x_j \in U, \ \min \Delta_j^* > \xi_{p-p^*}\}.$$

Having possibly modified our sets we update $U, \mathcal{M}^*, \Delta_j^*$, and p^* using (5.3), (5.4), (5.5), and (5.2) of Step 4 respectively.

Step 6: Check to see if we have identified a final model

If $\#\mathcal{M}^* = 1$, then there is only one model that is now currently under consideration, and thus $\mathcal{M}^*$ is the final selected model.

Step 7: Check to see if any further variables for clear inclusion or removal can be found

Having selected or removed variables in Step 5, it is possible that there may be further variables that should now be clearly included or removed. If the likelihood ratio statistics for a given variable from the undecided set straddle the current threshold, then we do not have clear evidence for its inclusion or exclusion. Thus, if for all $x_j \in U$, $\min \Delta_j^* < \xi_{p-p*} < \max \Delta_j^*$, then continue to Step 8; otherwise repeat Steps 5 through 7.

Step 8: Organize undecided variables into sets of sets

At this stage, the algorithm has identified all variables that have clear evidence to warrant inclusion. Given that these variables are included in the model, we have also identified those variables for which no reasonable evidence is provided by the data at hand, and have indicated them for removal. There is no clear evidence for any of the variables that remain in the undecided set as to whether they should be kept or removed. For each of these undecided variables, there exists from among the models still under consideration, at least one model for which the variable provides insufficient improvement, and at least one model with an improvement above the currently designated threshold. Hence, the algorithm as described thus far has reached a stopping point, and we must now take some modified approach to find a satisfactory model.

From this point, we shall consider in a sequential manner all possible subsets that can be formed by variables in the undecided set. Hence, we shall define sets of

sets, by letting

$$U_b = \{\text{all possible sets of } b \text{ variables from } U\} \quad \text{for } \ b = 1, \ldots, \#U.$$

For example, if the undecided set, $U = \{x_1, x_2, x_3\}$, had three variables, then we would form three sets of sets: the set of singleton variables, $U_1 = \{\{x_1\}, \{x_2\}, \{x_3\}\}$; the set of pairwise combinations, $U_2 = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_1, x_3\}\}$; and the triplet combination, $U_3 = \{\{x_1, x_2, x_3\}\}$.

We shall also re-initialize the iteration counter to be $t = 1$.

Step 9: Find the smallest irreducible models

Currently, G^* is the set of variables for which we have sufficient evidence for inclusion. To this set we shall augment, in sequence, the elements from the sets established in Step 8. For each $\upsilon \in U_t$, let

$$G^*_\upsilon = G^* \cup \upsilon.$$

Now for each of these augmented sets, we shall identify the set of models that need to be further considered,

$$\mathcal{M}^*_\upsilon = \{\mathcal{M}_k : x_j \in \mathcal{M}_k \ \ \forall x_j \in G^*_\upsilon \ \ \& \ \ x_j \notin \mathcal{M}_k \ \ \forall x_j \in R\},$$

and also the relevant likelihood ratio statistics for each variable per model to be accounted for,

$$\Delta^*_{j\upsilon} = \{\delta_{k,j} : \mathcal{M}_k \in \mathcal{M}^*_\upsilon, \ x_j \in U \setminus \upsilon\}.$$

Finally, for each of the models we are considering, we shall identify which of them, if any, result in the likelihood ratio statistics for all remaining undecided variables to be entirely above or below the threshold,

$$E = \{\upsilon : (\max \Delta^*_{j\upsilon} < \xi_{p-p^*-\#\upsilon} \text{ or } \ \min \Delta^*_{j\upsilon} > \xi_{p-p^*-\#\upsilon}) \ \ \forall x_j \in U \setminus \upsilon\}.$$

Step 10: Check to see if a final model has been identified

If $\#E = 0$, then the addition of any element from U_t does not result in a clear separation of the remaining likelihood ratio statistics at the current threshold. This indicates the need for a more complex model; hence, we shall increment t by 1, and repeat Step 9.

If $\#E = 1$, then there exists one and only one element from U_t, which when augmented with G^*, results in a clear separation of the remaining likelihood ratio statistics. Let $\eta \in E$, then the final model that will be selected is given by

$$\mathcal{M}_\eta^* = G^* \cup \eta \cup \{x_j : x_j \in U \setminus \eta,\ \min \Delta_{j\eta}^* > \xi_{p-p^*-\#\eta}\}. \tag{5.6}$$

If $\#E > 1$, then we have $\#E$ essentially equivalent models to choose from. Let $e = \#E$, and let the e elements of E be $\eta_1, \ldots, \eta_e$, with corresponding models $\mathcal{M}_{\eta_1}^*, \ldots, \mathcal{M}_{\eta_e}^*$, as defined by (5.6).

If there exists an $r \in \{1, 2, \ldots, e\}$ such that $\#\mathcal{M}_{\eta_r} < \#\mathcal{M}_{\eta_s}$ for all $r \neq s \in \{1, 2, \ldots, e\}$, then $\mathcal{M}_{\eta_r}$ will be selected as the final model, since it is the smallest among those models still under consideration.

If no such r exists, then let $q = \min\{\#\mathcal{M}_{\eta_1}^*, \ldots, \#\mathcal{M}_{\eta_e}^*\}$. Let $\mathcal{M}_\varnothing$ denote the null or mean only model, and $\mathcal{M}_\bullet$ denote the largest candidate model. Then select as the final model

$$\mathcal{M}_\eta^* : \arg\max_\eta \Lambda(\mathcal{M}_\eta^*, \mathcal{M}_\varnothing) \quad \text{for} \ \ \eta \in E \ \ \& \ \ \#\mathcal{M}_\eta^* = q.$$

This is equivalent to selecting

$$\mathcal{M}_\eta^* : \arg\min_\eta \Lambda(\mathcal{M}_\eta^*, \mathcal{M}_\bullet) \quad \text{for} \ \ \eta \in E \ \ \& \ \ \#\mathcal{M}_\eta^* = q.$$

Asymptotically, for a finite dimension data generating mechanism, the probability that Steps 8 through 10 will be required to establish a final model will approach zero.

CHAPTER 6
LINEAR MODEL SIMULATIONS

In this chapter, we shall examine the comparative performance of a number of model selection methods under a variety of simulated settings. Five methods will be studied: (1) backward elimination; (2) minimum AIC; (3) minimum BIC; (4) naive SIFT; and (5) empirical SIFT. As noted earlier, we will be working within the framework where we have p potential explanatory variables and we wish to select a model based on some linear combination of these p variables. Thus, we will need to consider a total of 2^p models.

All simulations were conducted using version 2.15.3 of the R statistical software.

6.1 Performance Measures

In order to compare the model selection methods, a number of performance measures will be examined. For each simulation setting, these performance measures will be presented in two sets of graphs. The first set will present: (a) the mean total model error; (b) the mean number of predictors selected; (c) the mean number of of true predictors selected; and (d) the mean number of false predictors selected. The second set will present the percentage of selected models that were (a) correctly specified; (b) underspecified; (c) both under and overspecified; and (d) overspecified.

The mean total model error is the sum of the squared differences between the true mean and the fitted value from the selected model. Suppose we have design matrix $\mathbf{X}$, and a corresponding parameter vector $\boldsymbol{\beta}_0$ that defines the true data generating mechanism, such that $\mathrm{E}[\boldsymbol{y}] = \mathbf{X}\boldsymbol{\beta}_0$. Here, not all of the elements of $\boldsymbol{\beta}_0$ need be non-zero. If we select a model characterized by $\hat{\boldsymbol{\beta}}$, where the elements that correspond to non-selected variables have a value of zero, then the total model error is defined as $(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{X}\hat{\boldsymbol{\beta}})$. For every sample size, this value is calculated for each simulated data set. These values are then averaged to produce

the mean total model error.

For each simulation of a given sample size, the total number of variables selected, not including the intercept term, is retained and then averaged over the simulated data sets to establish the mean number of predictors selected. Variables that are part of the data generating mechanism are referred to as true predictors, while variables not part of the data generating mechanism are referred to as false (or spurious) predictors. The final two graphs of the first set decompose the mean number of predictors selected into the mean number of true predictors selected and the mean number of false predictors selected.

Models will be classified into four mutually exclusive categories: correctly specified; underspecified; overspecified; and both under and overspecified. A correctly specified model contains *only and all of* the true predictors. An underspecified model contains *only but not all of* the true predictors. An overspecified model contains *all of* the true predictors, *and some of* the false predictors. A model that is both under and overspecified contains *some but not all of* the true predictors, *and some of* the false predictors. The percentage of models at each sample size within each of these categories is presented in the second set of graphs.

6.2 Computational Issues

For any given model, the computation of the log-likelihood in the linear model setting requires computation of the three quantities: $\boldsymbol{y}'\boldsymbol{y}$; $\mathbf{X}'\boldsymbol{y}$; and $\mathbf{X}'\mathbf{X}$. For a single model, and with p not too large, it does not take very long to compute these quantities even for large sample sizes. However, once we are considering all possible subsets, we have many more models to be evaluated. For example, if we had 10 predictor variables then we would now need to calculate these quantities across all 1,024 potential models.

It is observed, however, that the quantity $\boldsymbol{y}'\boldsymbol{y}$ will remain constant for all

of the models being evaluated, and thus need only be calculated once, rather than recomputing the exact same quantity 2^p times. A similar but slightly more complicated observation can be made with respect to the $\mathbf{X}'\boldsymbol{y}$ vector and the $\mathbf{X}'\mathbf{X}$ matrix. Let $\mathbf{X}_p$ be the design matrix for the largest candidate model being considered. Once we have computed $\mathbf{X}_p'\boldsymbol{y}$, we can easily obtain the required vector for any of the subset models, simply by selecting the relevant set of elements from $\mathbf{X}_p'\boldsymbol{y}$. Similarly, after computing $\mathbf{X}_p'\mathbf{X}_p$, we can obtain the sum of squares and cross-products matrix for all other models we are considering, simply by selecting the appropriate set of rows and columns from $\mathbf{X}_p'\mathbf{X}_p$. Thus, after we have calculated these three quantities once for the largest candidate model, we can quickly extract the information we need from these summary statistics, without the need for recomputation for all other models under consideration. Substantial improvements in efficiency can be achieved by implementing the preceeding technique.

Prior to the widespread availability of high performance computing, the need to discover computationally efficient methods where the results of one analysis could be reused for the next analysis was more critical. Authors such as Schatzoff et al. (1968) and Furnival (1971) developed techniques specific to the all possible regression setting. Rediscovering and implementing techniques such as these may yield significant practical improvements to computational efficiency.

6.3 Independent Predictors: 5% False Admission Rate

The first simulation will consider the setting of independent predictor variables. For a given sample size, eight explanatory variables were randomly generated as follows:

$$x_{ij} \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1) \quad \text{for} \;\; i = 1, \ldots, n \quad \text{and} \;\; j = 1, \ldots, 8.$$

The data generating mechanism is a function of three of the eight variables,

$$y_i = 16 + 1.0x_{i1} + 0.8x_{i2} + 0.6x_{i3} + \varepsilon_i \quad \text{where} \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \text{N}(0, 1).$$

Simulations were conducted at sample sizes of 25, 50, 75, 100, 125, 150, 175, 200, 300, 400, 500, 750, 1,000, 1,500, 2,000, and 2,500. A total of 5,000 simulated data sets were compiled at each sample size. For all simulations, the design matrix remained fixed at each sample size. The component that varied from simulation to simulation was ε_i, and through this, the y_i. The false admission rate was set at 5% for the naive and empirical SIFT methods.

Figure 6.1 presents the first set of results. Across all of the figures, the results from the naive and empirical SIFT methods were extremely close. The two SIFT methods differed most for smaller n, where the total model error for the naive method was slightly larger than for the empirical method. A feature common to all methods was the initial increase in total model error followed by a decline which approached a plateau as n increased. For sample sizes above 400, the mean total model error for the SIFT methods dropped below that for the minimum AIC. For larger sample sizes, the results for the SIFT methods were closely aligned with the results for the minimum BIC. The p-value backward elimination method was more liberal than either SIFT or minimum BIC, but was more conservative than the minimum AIC method. Thus, it had an intermediate profile.

The minimum AIC consistently selected larger models than the other methods; selecting around 3.8 variables on average for large n, compared with around 3.05 for SIFT and 3.26 for backward elimination. Once the sample size reached about 750, all methods were consistently selecting all of the variables within the data generating mechanism. Asymptotically, the minimum BIC method will not admit any of the false predictors. For the largest sample size presented, the SIFT methods admitted an average of 0.05 false predictors; minimum AIC admitted 0.78 predictors; and

backward selection 0.23 predictors.

Figure 6.2 presents the second set of results. Minimum AIC selected the correctly specified model in 42.56% of the simulated samples for a sample size of 2,500; the balance of 57.44% of models were overspecified. Even for small n, the minimum AIC was more likely to choose a model that was both under and overspecified, as opposed to a strictly underspecified model. For large n, none of the methods selected models that were either underspecified, or both under and overspecified. For the largest sample size presented, 21.00% of models selected by backward elimination were overspecified. The rate of overspecification for the naive and empirical SIFT methods were 5.08% and 5.34% respectively—very close to the nominated false admission rate of 5%. Overspecification by minimum BIC for the largest sample size was 2.60%—this would continue to approach zero as the sample size is increased.

6.4 Independent Predictors: 30% False Admission Rate

In the previous section, it was observed that the performance of the SIFT method was rather similar to that attained by using the minimum BIC. One advantage of the SIFT procedure over the other methods presented is that the false admission rate can be nominated in advance of the data analysis. For each of the other methods presented, the false admission rate will be a function of the number of false predictors being considered—the greater the number, the higher the rate of false admission. Furthermore, for any particular application, the rate of false admission will remain unknown for these other methods. By using SIFT, researchers can control the false admission rate to a known level, which can be determined *a priori* according to their specific needs.

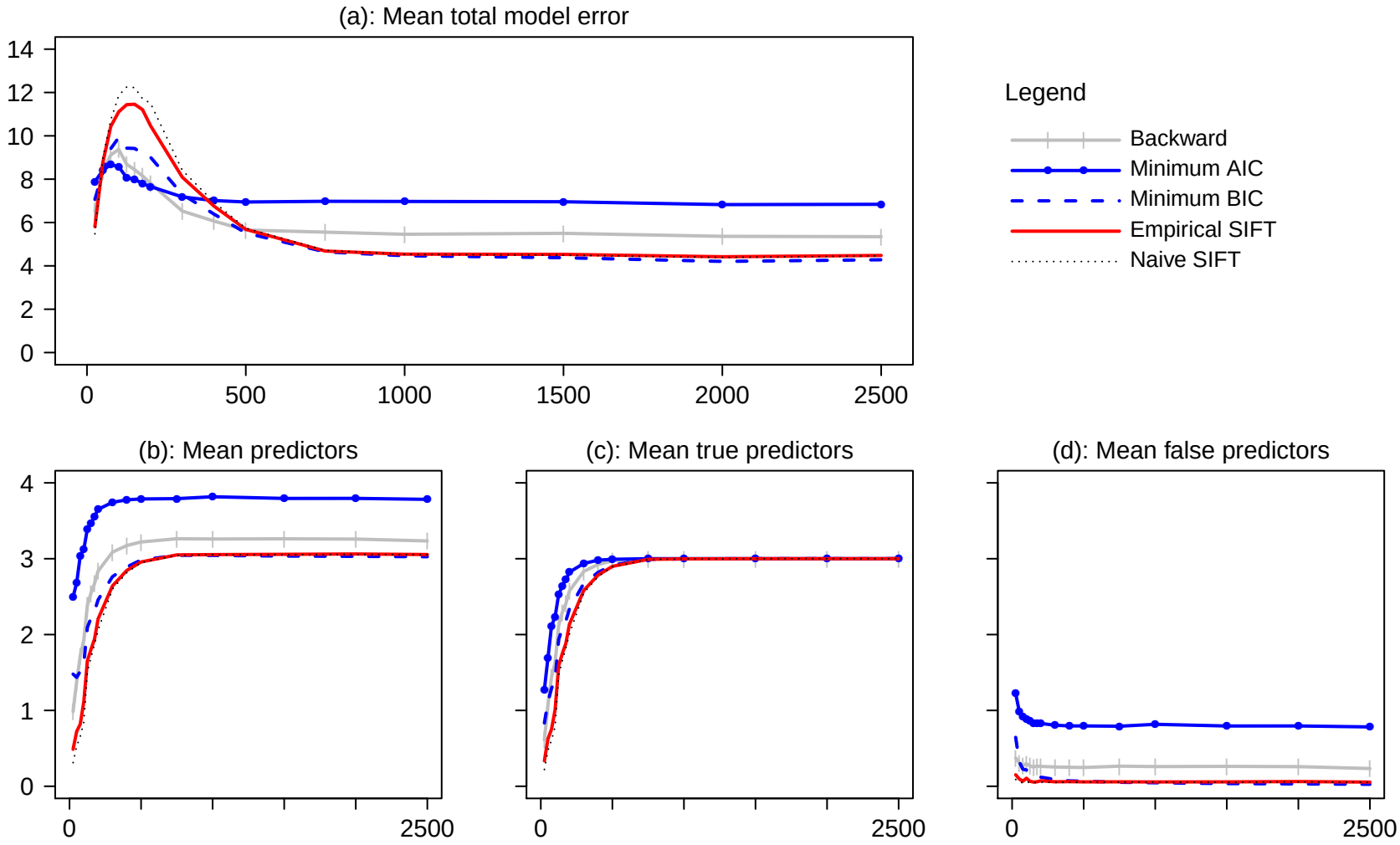

Figure 6.1: Comparison of model selection methods for independent explanatory variables with a 5% false admission rate—mean total model error and mean number of predictors selected.

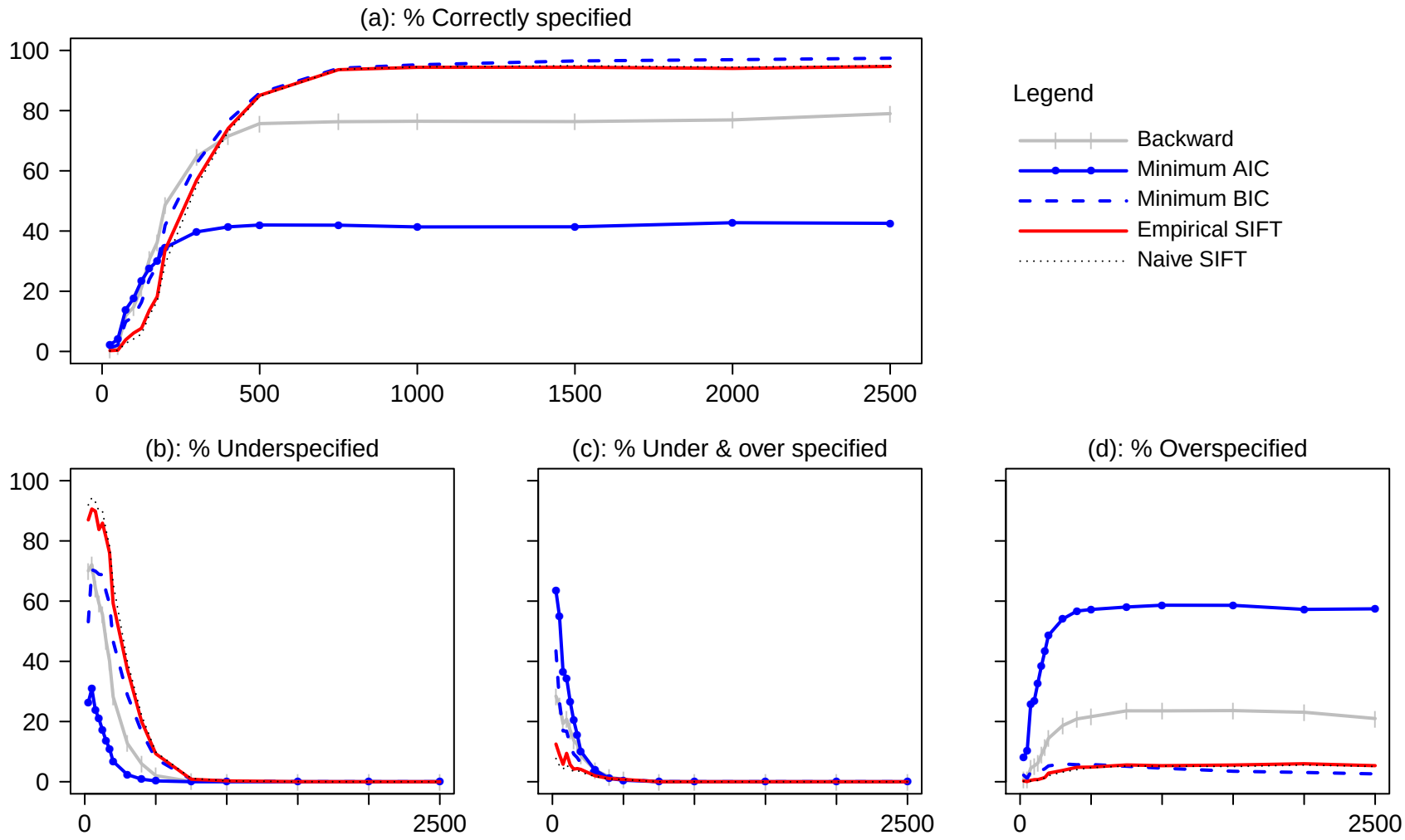

Figure 6.2: Comparison of model selection methods for independent explanatory variables with a 5% false admission rate—type of model selected.

In order to exemplify this property of the SIFT procedure, we present in this section a set of simulations where the false admission rate for the SIFT method has been set to 30%. The simulations of this section are identical to the conditions described in Section 6.3, with the sole exception of the differing false admission rate for the SIFT procedure. Thus, in the figures that follow, the results for backward elimination, minimum AIC, and minimum BIC, are identical to those of Section 6.3—only the results for the SIFT methods change.

Figure 6.3 shows that for smaller sample sizes, the behavior of the SIFT method, with respect to total model error, now more closely mirrors that of minimum AIC or backward elimination. For large sample sizes, the total model error for SIFT was a little greater than that for backward elimination. Consistent with relaxing the false admission rate, the mean number of predictors selected by SIFT was higher than that observed in Section 6.3. The average number for large n increased to around 3.38 predictors in total, with an average of 0.38 false predictors.

The type of models selected are shown in Figure 6.4. As expected, the percentage of models correctly specified by SIFT decreased. For the largest sample size, the naive and empirical methods respectively selected the correctly specified model 71.10% and 69.58% of the time. Again, both SIFT methods achieved results that were consistent with the nominated false admission rate. The percentage of models correctly specified was slightly higher than the nominated level for the naive SIFT. This may indicate that the threshold for the naive SIFT is possibly slightly more conservative than need be. For large sample sizes, all of the incorrectly specified models were in the overspecified category.

6.5 Correlated Predictors: 5% False Admission Rate

Sections 6.3 and 6.4 both examined simulations under the assumption that the set of predictor variables were mutually independent. Thus, the conclusion that the

naive and empirical SIFT methods produced such similar results may be somewhat unsurprising, given that the simulation conditions reflected those under which the naive method was proposed. In this section, we will examine a scenario under which correlation is present within the design matrix. We will initially randomly generate eight explanatory variables as was done in Section 6.3, so that

$$x_{ij} \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1) \quad \text{for} \;\; i = 1, \ldots, n \quad \text{and} \;\; j = 1, \ldots, 8.$$

However, in order to introduce correlation among the variables, we shall redefine the explanatory variables in the following way:

$$x_{ij} \;\; \leftarrow \;\; 0.2x_{i,j-1} + 0.8x_{ij} \quad \text{for} \;\; j = 2, \ldots, 8.$$

We will continue to use the same data generating mechanism for the response variable,

$$y_i = 16 + 1.0x_{i1} + 0.8x_{i2} + 0.6x_{i3} + \varepsilon_i \quad \text{where} \;\; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \text{N}(0,1).$$

Simulations were carried out 5,000 times at each of the sample sizes specified in Section 6.3. The false admission rate for the SIFT methods was set at 5%.

The results under the correlated setting were quite similar to those obtained under the condition of independence. In the correlated setting depicted in Figure 6.5, the total model error required a larger sample size before reaching the asymptotic plateau. The limiting values for each method were nearly identical to those attained under independence, with the exception of the empirical SIFT, which had an average total model error of 4.62 in the correlated setting, compared with 4.48 under the independence setting. The rate at which variables were admitted was slower for the correlated setting, and hence the sample size required before each of the methods consistently selected all of the true predictors is larger.

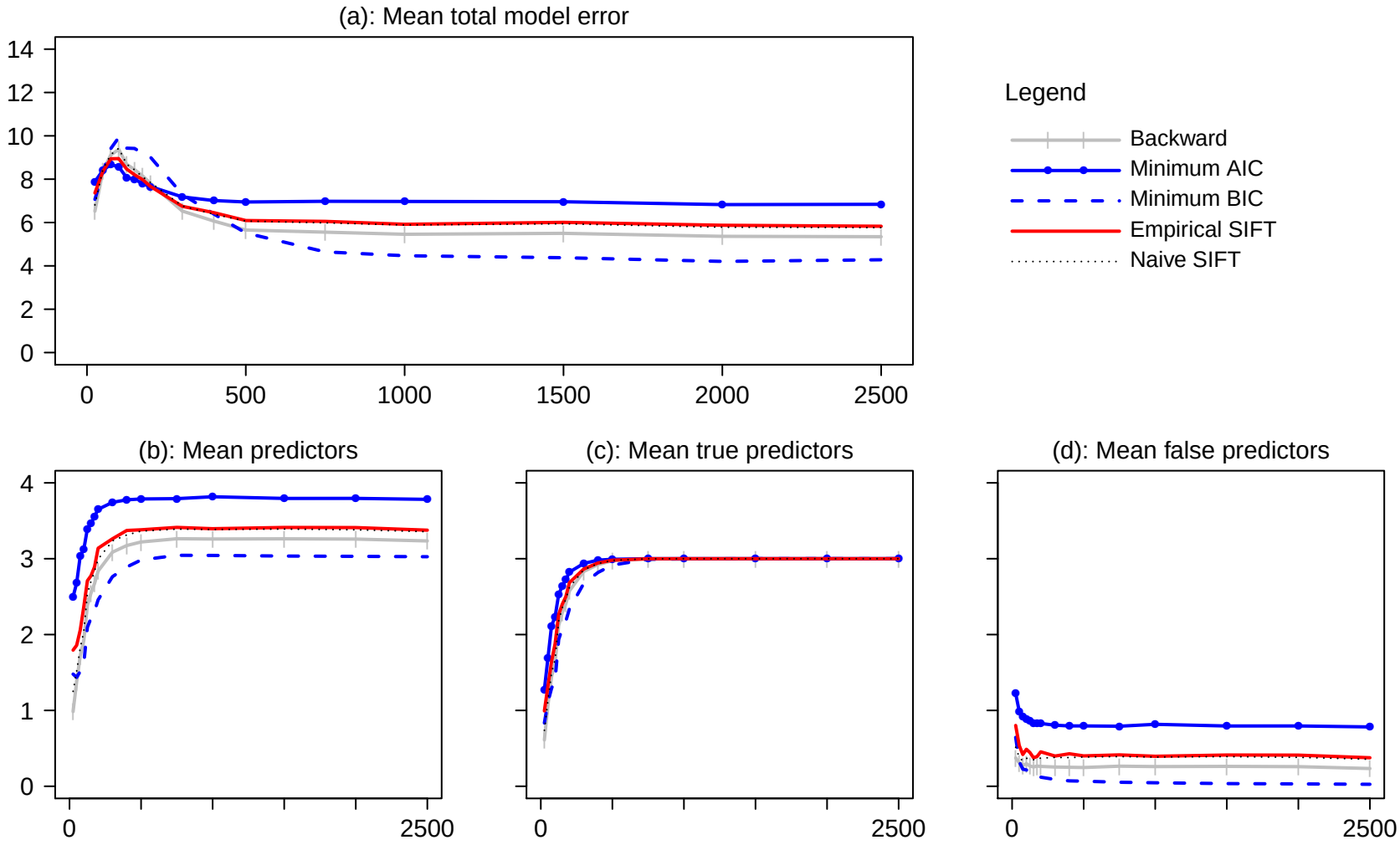

Figure 6.3: Comparison of model selection methods for independent explanatory variables with a 30% false admission rate—mean total model error and mean number of predictors selected.

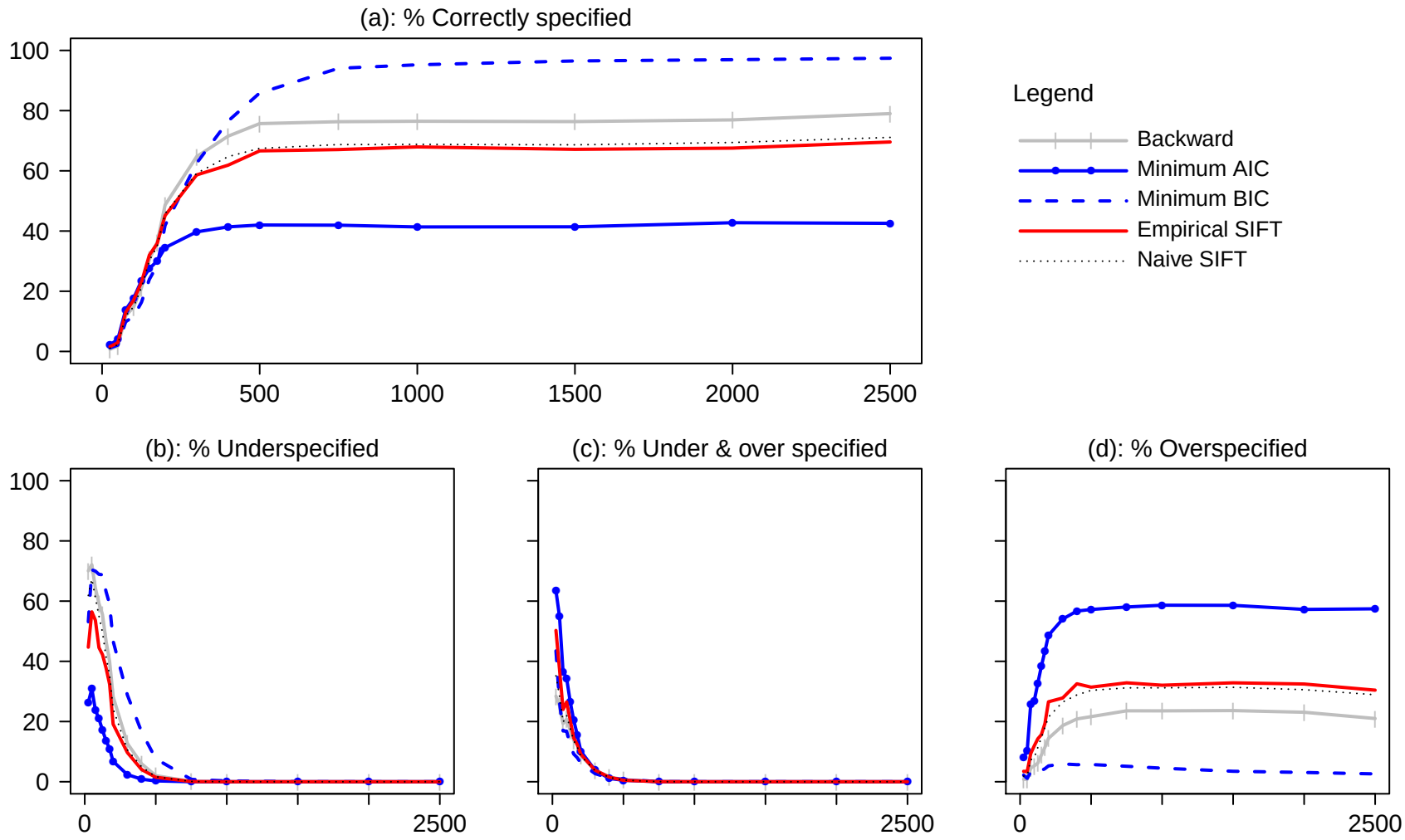

Figure 6.4: Comparison of model selection methods for independent explanatory variables with a 30% false admission rate—type of model selected.

Figure 6.6 presents the breakdown of the type of models selected. As was the case in Section 6.3, once the sample size had reached 2,500, none of the methods selected models that omitted any of the variables within the data generating mechanism. For the largest sample size, the percentage of models selected that were overspecified was 56.72% for the minimum AIC, 21.22% for backward elimination, and 2.76% for minimum BIC. The rate of overspecification for the naive and empirical SIFT methods were 5.10% and 7.16% respectively. Based on these results, it may seem that the empirical SIFT is more liberal than it perhaps should be, and the naive SIFT may in fact produce results closer to the nominated false admission rate than empirical SIFT.

6.6 Correlated Predictors with Noisy Data or Small Effects: 5% False Admission Rate

The final setting that will be investigated is one in which we have noisy data, or equivalently small effects, that will require large sample sizes in order to detect. Some correlation between the set of predictor variables will also be introduced. As was done before, we will initially randomly generate eight explanatory variables, such that

$$x_{ij} \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1) \quad \text{for} \;\; i = 1, \ldots, n \quad \text{and} \;\; j = 1, \ldots, 8.$$

In order to introduce some correlation within the variable set, we will redefine the explanatory variables x_3 and x_8 as

$$\begin{aligned} x_{i3} &\leftarrow 0.2x_{i2} + 0.8x_{i3}, \quad \text{and} \\ x_{i8} &\leftarrow 0.2x_{i1} + 0.8x_{i8}. \end{aligned}$$

We will modify the data generating mechanism for the response variable by reducing the magnitude of the coefficients,

$$y_i = 16 + 0.14x_{i1} + 0.12x_{i2} + 0.10x_{i3} + \varepsilon_i \quad \text{where} \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \text{N}(0, 1).$$

Simulations were conducted at sample sizes of 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 8,000, 9,000, 10,000, 15,000, 20,000, 30,000, 40,000, 50,000, 75,000, and 100,000. For each sample size, a total of 5,000 simulated data sets were again generated. The false admission rate for the SIFT methods was set at 5%.

Unlike AIC, for which the penalty term does not vary with the sample size, the penalty term for BIC is a function of the sample size, and increases with increasing sample size. It is this property that makes model selection by BIC more and more conservative with larger sample sizes, thus enabling it to asymptotically select the correctly specified model with probability one. It is also this property that will distinguish the SIFT procedure from minimum BIC in the noisy data setting.

Under the previous settings, it was observed that as a function of sample size, the SIFT method initially resulted in a larger total model error. In the noisy data setting, it is observed in Figure 6.7 that the minimum BIC has the largest initial total model error. The total model error for the SIFT methods dropped below that for the minimum AIC once the sample size reached about 20,000. This did not occur for the minimum BIC until a sample size of more than 30,000 had been reached. The pattern of results for the number of predictors selected was similar to the previous simulation sets, with the exception that the minimum BIC is now the method that is slowest in admitting the true predictors.

The effect of the large penalty term for BIC is evident in Figure 6.8, especially in the percentage underspecified graph. For the largest sample sizes, none of the methods chose models that were both under and overspecified. None of the methods selected an underspecified model for the largest sample size presented.

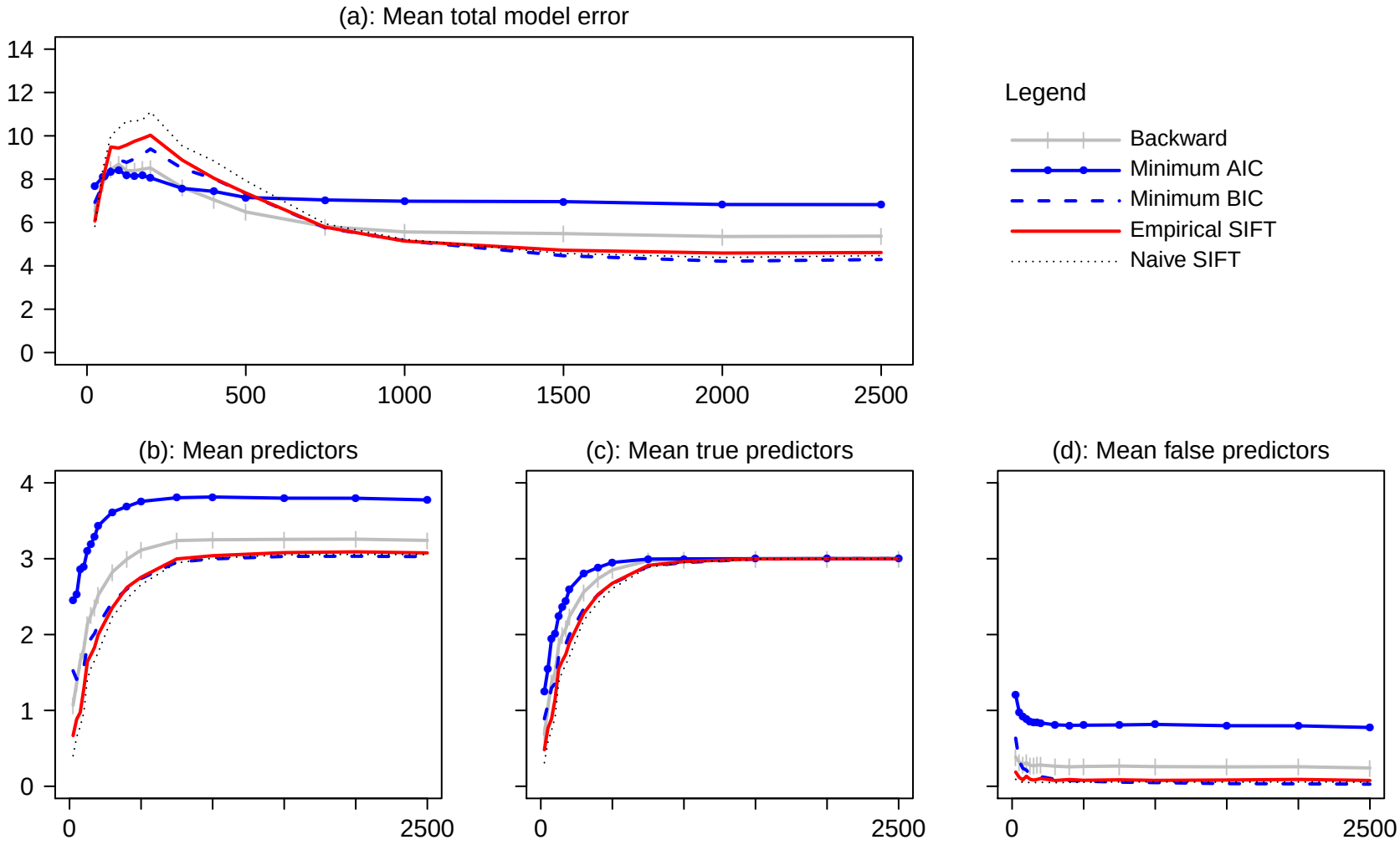

Figure 6.5: Comparison of model selection methods for correlated explanatory variables with a 5% false admission rate—mean total model error and mean number of predictors selected.

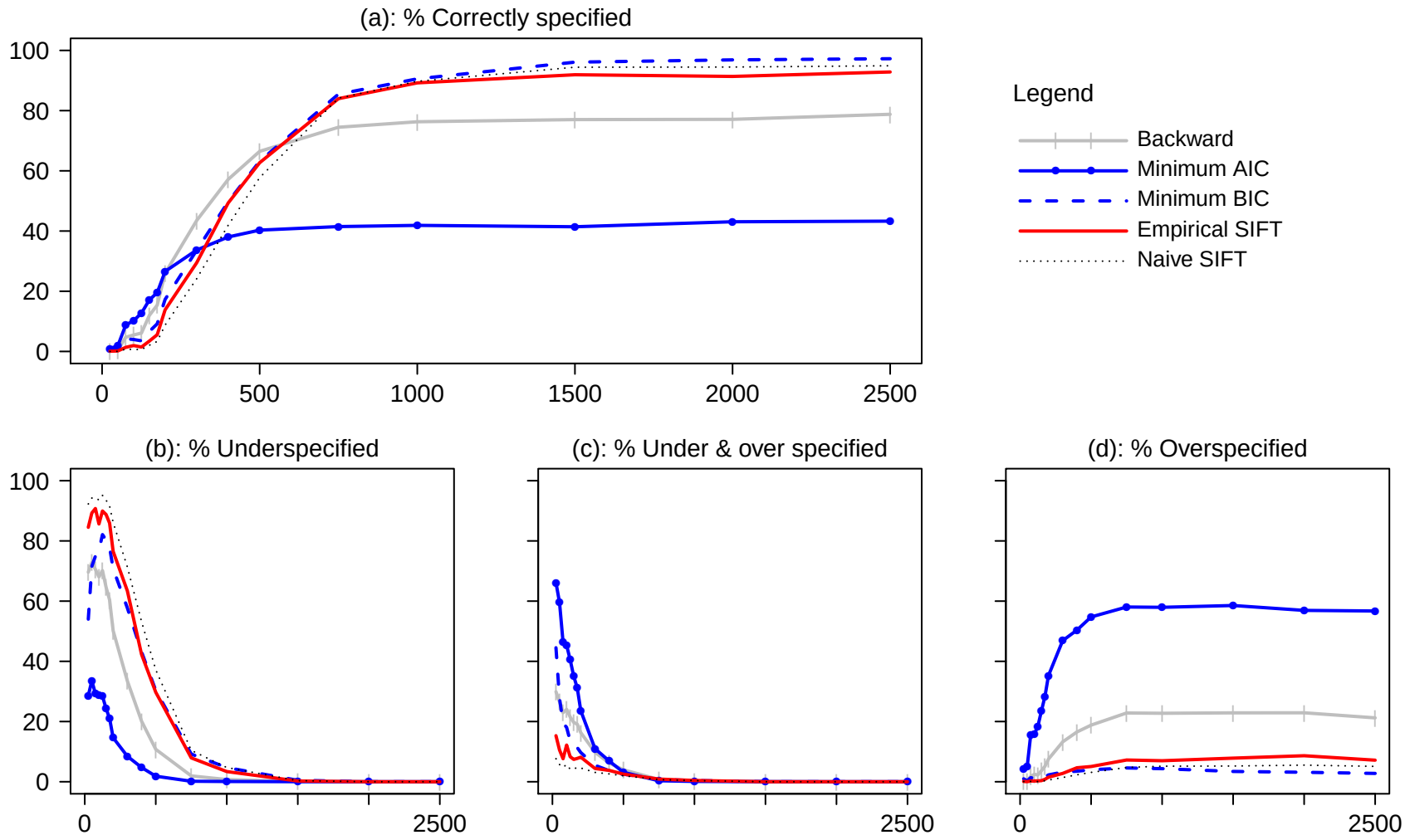

Figure 6.6: Comparison of model selection methods for correlated explanatory variables with a 5% false admission rate—type of model selected.

For the largest sample size, 58.62% of the models selected by minimum AIC were overspecified, and 23.56% for backward elimination. The selection of overspecified models by both SIFT methods was 5.44% for a sample size of 100,000, and was 0.34% for the minimum BIC.

6.7 Discussion of Results

In this chapter, we have investigated under a variety of settings the performance of the proposed SIFT methodology. The SIFT procedure was designed to allow for the control of the rate at which selected models include false or spurious variables. To this end, the SIFT procedure was observed to generally perform as designed.

The empirical SIFT was designed to attempt to better account for more general conditions than the naive SIFT. However, in the correlated setting of Section 6.5, the percentage of models overspecified by the empirical SIFT was 7.16% compared with 5.10% for the naive SIFT. This suggests that the empirical SIFT may in fact be more liberal than desired, or perhaps requires a larger sample size to achieve the nominated level.

For settings under which effects could be detected with sample sizes on the order of hundreds or thousands, and the false admission rate was 5%, the performance of the SIFT methods were similar to that observed by the use of minimum BIC. Under settings in which tens of thousands of observations were required to detect effects, the penalty term for BIC is such that the criterion will initially be more conservative than SIFT. Thus, BIC will require a larger sample size before selecting correctly specified models at the same rate as SIFT.

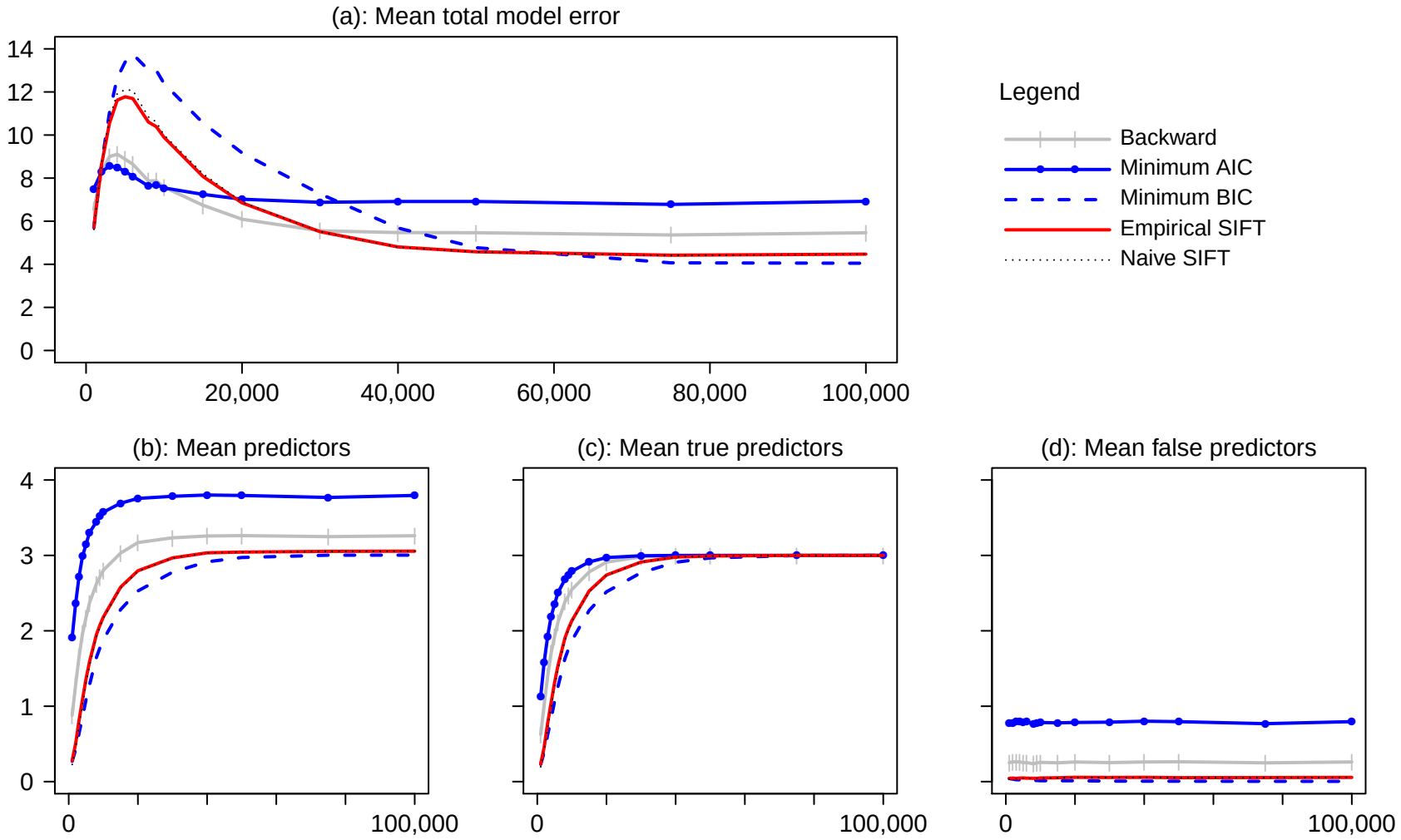

Figure 6.7: Comparison of model selection methods for mildly correlated explanatory variables with small effect sizes and a 5% false admission rate—mean total model error and mean number of predictors selected.

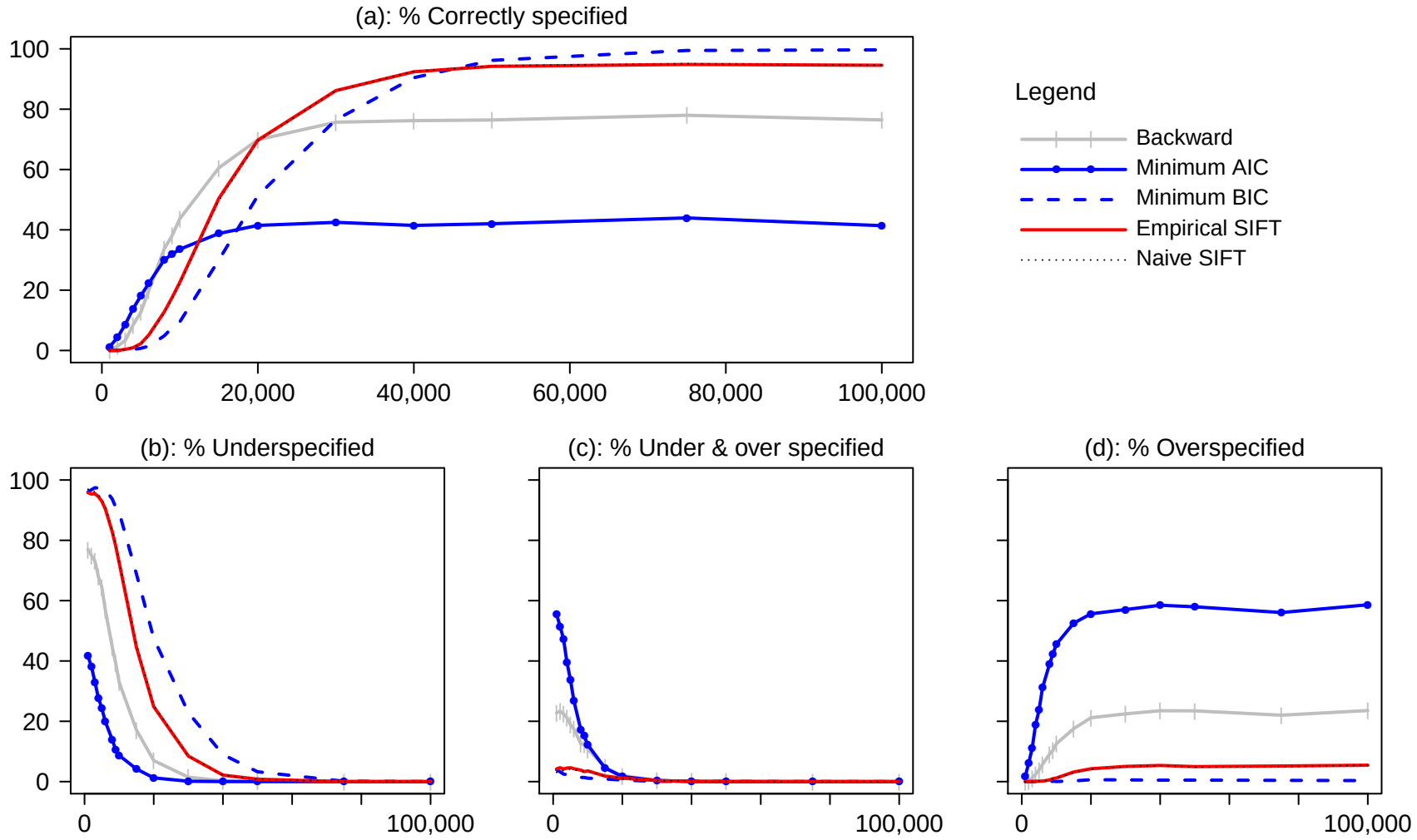

Figure 6.8: Comparison of model selection methods for mildly correlated explanatory variables with small effect sizes and a 5% false admission rate—type of model selected.

While for smaller sample sizes, the total model error attained using the minimum AIC will be smaller than for SIFT, eventually SIFT will result in a smaller total model error. For any given application, depending on the size of the effects relative to the noise of the system, it will be unknown as to which side of this crossing point the data are on. Thus, it is unknown as to whether a minimum AIC model or a SIFT model will produce the lower total model error.

As expected, model selection by backward elimination based on p-values produced results that were more conservative than minimum AIC, but more liberal than minimum BIC.

6.8 Concordance of Model Selections

In this section, we examine the concordance of the type of model selected between different selection procedures. Simulations were conducted for a moderate sample size of 300, and for a large sample size of 2,500. For each sample size, a total of 5,000 simulationed data sets were compiled under the conditions specified in Section 6.3. Selected models were classified into one of the four categories described in Section 6.1. Comparisons are made with respect to the model selections of the empirical SIFT, the minimum AIC and the minimum BIC.

Table 6.1 presents the results for the sample size of 300. There were no data sets for which the empirical SIFT selected a model from a larger class than the minimum AIC, as indicated by the upper triangular nature of the first sub-table. The same conclusion is made when comparing the empirical SIFT to minimum BIC, or when comparing minimum BIC to minimum AIC. The empirical SIFT and minimum BIC resulted in the same classification for 90.8% of the simulated data sets. Among models correctly specified by either empirical SIFT or minimum BIC, the minimum AIC was more likely to choose an overspecified model than a correctly specified model. Both the empirical SIFT and minimum BIC were more likely to

select an underspecified model among models that were classified as both under and overspecified by AIC.

Table 6.1: Concordance of model selections, $n = 300$.

	Under-specified	Correctly specified	Under and over-specified	Over-specified	Total
Empirical SIFT		Minimum AIC			
Underspecified	105	600	175	784	1,664
Correctly specified	–	1,362	–	1,664	3,026
Under & overspecified	–	–	9	93	102
Overspecified	–	–	–	208	208
Total	105	1,962	184	2,749	5,000
Empirical SIFT		Minimum BIC			
Underspecified	1,287	315	45	17	1,664
Correctly specified	–	2,953	–	73	3,026
Under & overspecified	–	–	92	10	102
Overspecified	–	–	–	208	208
Total	1,287	3,268	137	308	5,000
Minimum BIC		Minimum AIC			
Underspecified	105	445	165	572	1,287
Correctly specified	–	1,517	–	1,751	3,268
Under & overspecified	–	–	19	118	137
Overspecified	–	–	–	308	308
Total	105	1,962	184	2,749	5,000

The results for the large sample size of 2,500 are presented in Table 6.2. At this sample size, none of the methods selected models that were underspecified or

models that were both under and overspecified. All models correctly specified by minimum AIC were also correctly specified by both empirical SIFT and mimimum BIC. All models that were classified as overspecified by empirical SIFT were classified correctly or as overspecified by minimum BIC.

Table 6.2: Concordance of model selections, $n = 2{,}500$.

	Under-specified	Correctly specified	Under and over-specified	Over-specified	Total
Empirical SIFT		Minimum AIC			
Underspecified	–	–	–	–	–
Correctly specified	–	2,095	–	2,616	4,711
Under & overspecified	–	–	–	–	–
Overspecified	–	–	–	289	289
Total	–	2,095	–	2,905	5,000
Empirical SIFT		Minimum BIC			
Underspecified	–	–	–	–	–
Correctly specified	–	4,711	–	–	4,711
Under & overspecified	–	–	–	–	–
Overspecified	–	147	–	142	289
Total	–	4,858	–	142	5,000
Minimum BIC		Minimum AIC			
Underspecified	–	–	–	–	–
Correctly specified	–	2,095	–	2,763	4,858
Under & overspecified	–	–	–	–	–
Overspecified	–	–	–	142	142
Total	–	2,095	–	2,905	5,000

In contrast to the smaller sample size, the empirical SIFT was now more likely to select an overspecified model than minimum BIC. All models classified as overspecified by BIC were classified the same by the empirical SIFT. Asymptotically the SIFT procedure should select on overspecified model at the false admission rate, which in this case is 5%, while the probability that BIC will select an overspecified model will approach zero.

CHAPTER 7
APPLICATION

7.1 Description of Example

In order to provide an example of an application of the SIFT procedure, we will use the diabetes data set considered by Efron et al. (2004) in their Least Angle Regression article, referred to as the LARS paper. The data set consists of ten potential explanatory variables and an outcome variable for 442 cases. The originating source of the data was not stated or referenced. For the analysis in the LARS paper, the explanatory variables were all centered at zero and scaled to have unit length. For the application presented here, we will be using the untransformed data that has been made publicly available by one of the authors (Hastie, 2003). The LARS paper describes the explanatory variables as "age, sex, body mass index, average blood pressure and six blood serum measurements". The blood serum measurements are not identified, and are simply labeled S1 though S6 in the untransformed data set. The transformed data used in the LARS paper is included as part of the `lars` package in R (Hastie and Efron, 2013), where the blood serum measurements are labeled *tc*, *ldl*, *hdl*, *tsh*, *ltg*, and *glu*. The outcome variable is simply described as, "a measure of disease progression one year after baseline" (Efron et al., 2004).

In a personal communication, the authors of the LARS paper were unable to provide any further information about the outcome variable or the blood serum measurements, and only identified the original source of the data as arising from a consulting project by the first author (Efron et al., 2013). Table 7.1 provides a description of the variables in the data set. The descriptions and units listed are based on a considered interpretation of the variable labels given previously, and an examination of the data distributions for each variable as compared with a

chart outlining reference ranges for blood tests (Häggström, 2009). As such, these descriptions should only be taken as indicative and not definitive.

Table 7.1: Description of variables in the diabetes data set.

Name	Description
y	Measure of disease progression one year after baseline
Age	Age (years)
Sex	Male/Female
BMI	Body Mass Index (kg/m^2)
MAP	Mean Arterial Pressure (mmHg)
TC	Total Cholesterol (mg/dL)
LDL	Low-Density Lipoprotein (mg/dL)
HDL	High-Density Lipoprotein (mg/dL)
TSH	Thyroid-Stimulating Hormone (mmol/L)
LTG	Triglycerides (mmol/L)
GLU	Glucose (mg/dL)

Table 7.2 presents the observed pairwise correlations between each of the ten explanatory variables, and also their correlations with the outcome variable. The outcome variable exhibited the strongest correlation with BMI and LTG; and with the exception of Sex, it had a non-negligible correlation with all other explanatory variables. With the exception of HDL, a positive correlation was observed between all other variables. HDL was negatively correlated with every variable except for total cholesterol (TC), for which the correlation was weak. The strongest observed correlation was 0.90, between total cholesterol and LDL. There were a number of other pairs with correlations greater than 0.5, and the majority were greater than 0.25 in magnitude, thus indicating a high level of inter-relatedness within this data

set.

Table 7.2: Pairwise correlation of variables in the diabetes data.

	Age	Sex	BMI	MAP	TC	LDL	HDL	TSH	LTG	GLU
Age	1.00	0.17	0.19	0.34	0.26	0.22	-0.08	0.20	0.27	0.30
Sex		1.00	0.09	0.24	0.04	0.14	-0.38	0.33	0.15	0.21
BMI			1.00	0.40	0.25	0.26	-0.37	0.41	0.45	0.39
MAP				1.00	0.24	0.19	-0.18	0.26	0.39	0.39
TC					1.00	0.90	0.05	0.54	0.52	0.33
LDL						1.00	-0.20	0.66	0.32	0.29
HDL							1.00	-0.74	-0.40	-0.27
TSH								1.00	0.62	0.42
LTG									1.00	0.46
y	*0.19*	*0.04*	*0.59*	*0.44*	*0.21*	*0.17*	*-0.39*	*0.43*	*0.57*	*0.38*

7.2 Results and Discussion

Figure 7.1 plots the formula-based naive threshold values and the empirically-derived thresholds for the diabetes data, calculated using a false admission rate of 5%. The naive threshold is the same for all models that consider the same number of variables. The empirical thresholds are based on the set of variables that have not been admitted. Thus, depending on the structure of the explanatory design matrix and its correlation structure, the thresholds will differ among variable sets of the same size. For the diabetes data, only one set of size ten can be formed, and thus there is only one empirical threshold that involves ten variables. For models based on nine variables there are ten thresholds; and $\binom{10}{2} = 45$ thresholds for models of eight variables; etc. Due to the high degree of correlation between variables in

the diabetes data set, the empirically-derived thresholds are often lower than the corresponding naive threshold, and sometimes considerably lower.

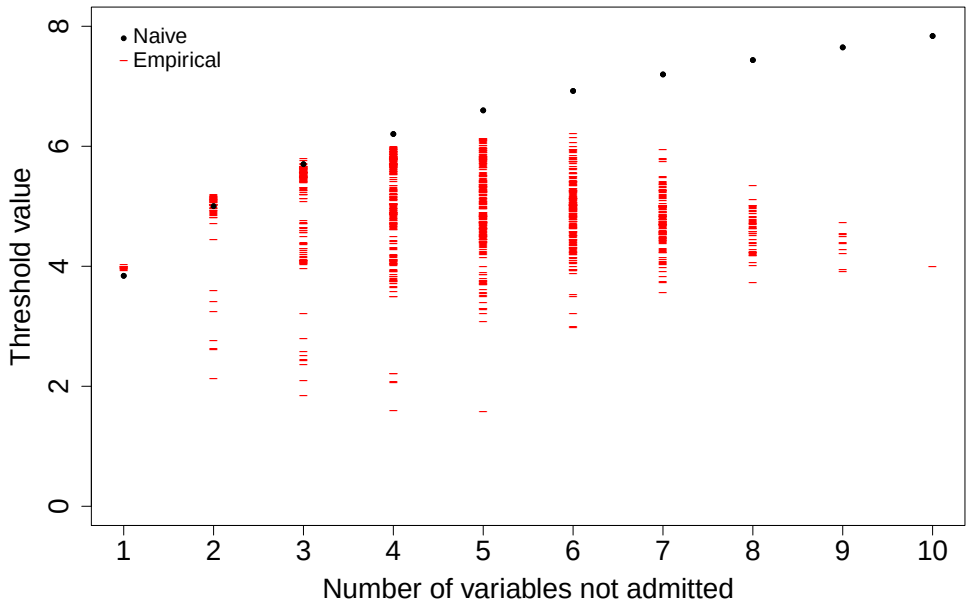

Figure 7.1: Naive and empirical threshold values for the diabetes data, with a 5% false admission rate.

The deviance plot introduced in Chapter 4 is presented in Figure 7.2 for the diabetes data. Some key features are readily observed. All models that resulted in a small deviance included the variables Sex, BMI, MAP and LTG. Conversely, models that did not include either BMI or LTG had a large deviance. The model with the largest deviance, among those that had at least one explanatory variable, was the model that included only Sex.

Figure 7.3 presents the likelihood ratio plot for the diabetes data. For each of the ten explanatory variables, all 512 associated likelihood ratio statistics are displayed. Consistent with the earlier observations from the deviance plot, the addition of BMI, MAP, or LTG to any base model under consideration resulted in an improvement to the likelihood ratio that was well above the empirical threshold

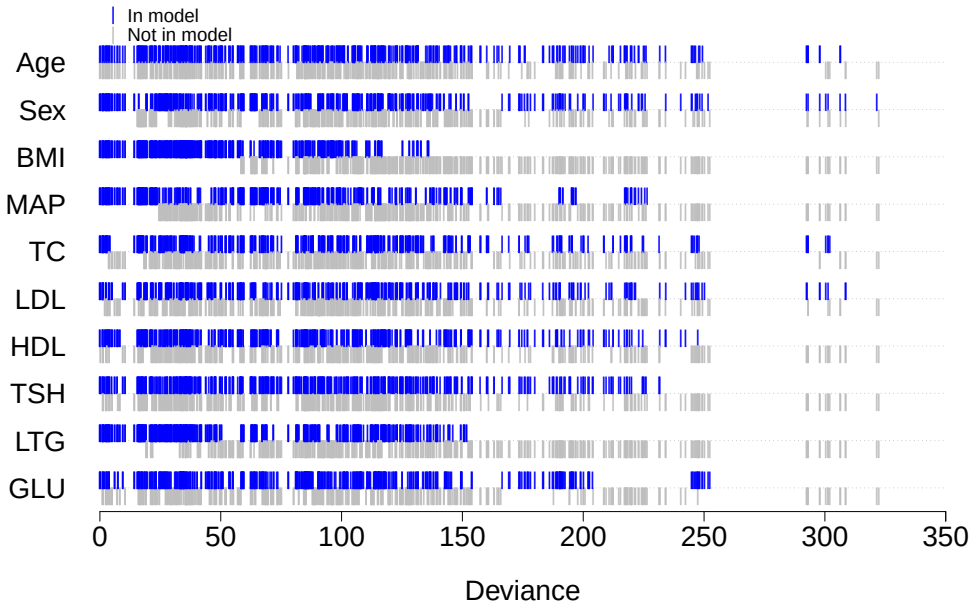

Figure 7.2: Deviance plot for the diabetes data.

of 3.97. This represents Step 1 of the SIFT procedure, and so $G^{(0)} = \{\text{BMI, MAP, LTG}\}$ and $p^{(0)} = 3$.

Having identified three variables that exceeded the nominated evidence threshold, we now adjust the threshold value to be based on the remaining seven undecided variables. The revised empirical threshold value was 4.69. The likelihood ratio plot for the undecided variables, given that BMI, MAP, and LTG are in the base model, is presented in Figure 7.4, with the updated threshold value. To more clearly distinguish the roles of each variable, the labels for variables in the base model have been prepended with the addition symbol and colored green. Across the remaining models under consideration, the minimum likelihood ratio value for Sex was 4.81, exceeding the threshold value. Therefore Sex is also admitted into the base model. This is Step 2, and now $G^{(1)} = \{\text{BMI, MAP, LTG, Sex}\}$ and $p^{(1)} = 4$. In Step 3, we determine that $p^{(1)} > p^{(0)}$; that is, we just added a variable, and thus we need to repeat Step 2.

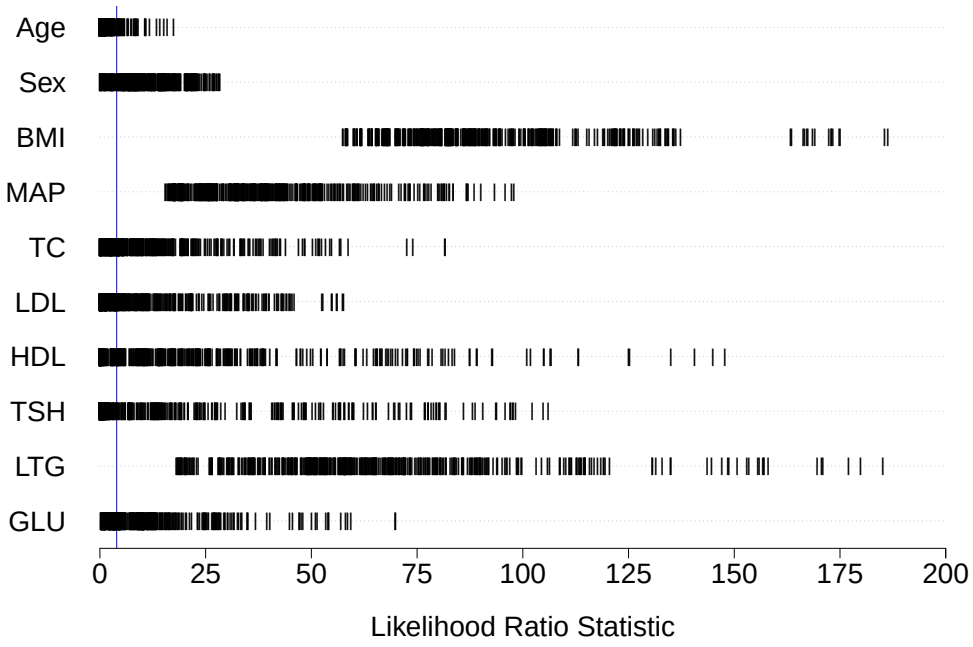

Figure 7.3: Likelihood ratio plot for the diabetes data, with an empirical threshold of 3.97.

Repeating Step 2 for $G^{(1)}$, we obtain a new threshold value of 4.46 based on the six undecided variables. The updated likelihood ratio plot is presented in Figure 7.5. From this plot, it is observed that none of the remaining variables have a minimum likelihood ratio change greater than the threshold. Thus, no further variables are added, and so $p^{(2)} = p^{(1)}$, and we progress to Steps 4 and 5. The maximum likelihood ratio change for both Age and GLU is below the threshold. Therefore they will be removed from the set of potential candidate variables, and $R = \{\text{Age, GLU}\}$.

The likelihood ratio plot for the remaining pool of candidate models, after the exclusion of Age and GLU, is presented in Figure 7.6. The labels for variables in the removed set are distinguished by being parenthesized and colored red. None of the remaining four undecided variables had all of their respective likelihood ratio statistics entirely above or below the threshold value. Thus, there are no further variables at this stage that can be identified for clear inclusion or exclusion, since

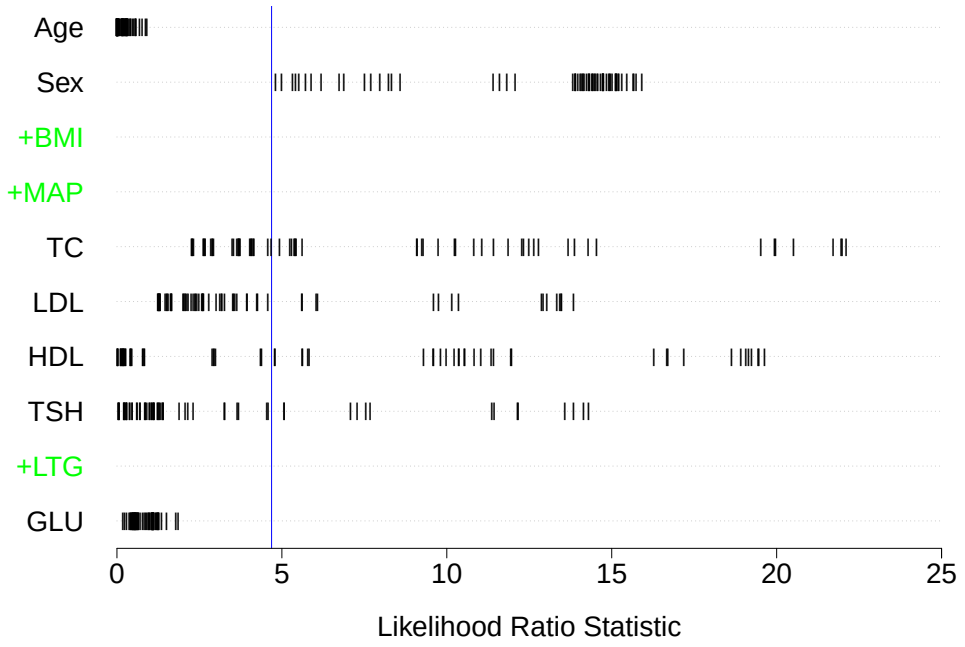

Figure 7.4: Likelihood ratio plot when BMI, MAP, LTG $\in \mathcal{M}_k$, with an empirical threshold of 4.69.

they all have values that straddle the threshold. The number of models currently under consideration, $\#\mathcal{M}^*$, is sixteen. This exhausts Steps 6 and 7, and we must therefore establish a final model using Steps 8 through 10.

The current undecided set is $U = \{\text{TC, LDL, HDL, TSH}\}$, and therefore $U_1 = \{\{\text{TC}\}, \{\text{LDL}\}, \{\text{HDL}\}, \{\text{TSH}\}\}$. When TC was added to the base model and the remaining likelihood ratio statistic values were examined as described in Step 9, the set of values for each of the other three variables continued to straddle the threshold value. This same situation held when either LDL or TSH was added to the base model. However, when HDL was added to the base model, the maximum of the likelihood ratio values for each of the other three variables was below the threshold, as illustrated in Figure 7.7. Thus $E = \{\text{HDL}\}$, and the variable set identfied as the final model was {Sex, BMI, MAP, HDL, LTG}.

The analysis of the diabetes data in the LARS paper by Efron et al. (2004)

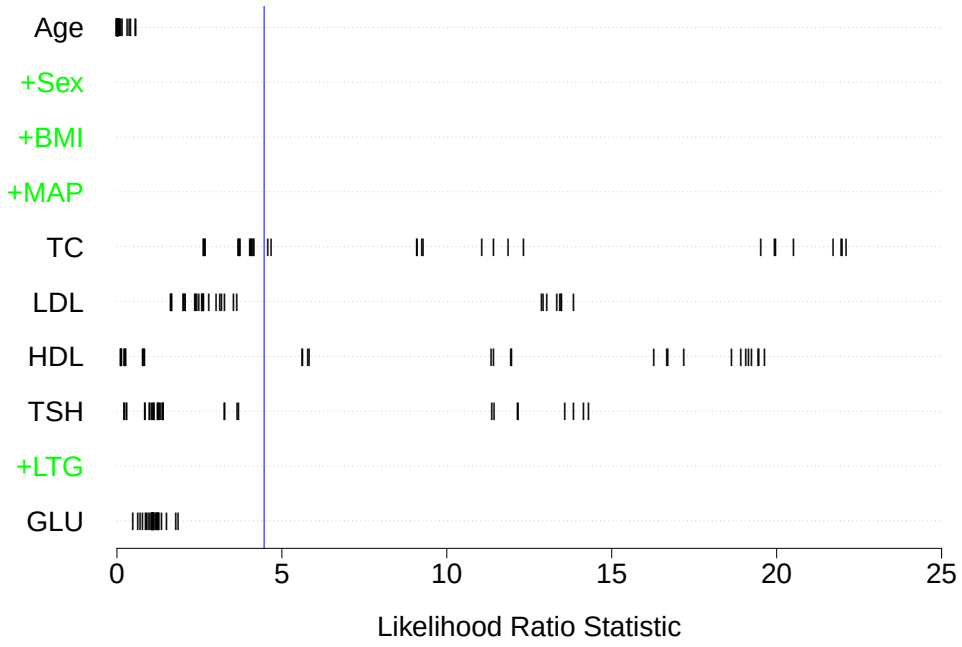

Figure 7.5: Likelihood ratio plot when Sex, BMI, MAP, LTG $\in \mathcal{M}_k$, with an empirical threshold of 4.46.

does not directly identify a final model since it requires specification of a constraint value. However, the first five variables that would be admitted as the constraint value is increased are (in order), BMI, LTG, MAP, HDL, and Sex. Thus the model selected by SIFT maintains a consistency with that obtained by Efron et al. (2004).

For a single application, the majority of the processing time is in establishing the empirical threshold values. To calculate the empirical threshold values based on 10,000 permuted data sets for the diabetes data took approximately fifty minutes. The time to compute the naive threshold values, or to run the SIFT procedure once the threshold values had been determined, took only a fraction of a second.

7.3 Interaction Variables

The ability of a model selection procedure to adequately handle desired features such as interactions is an important consideration. In this section, an example of

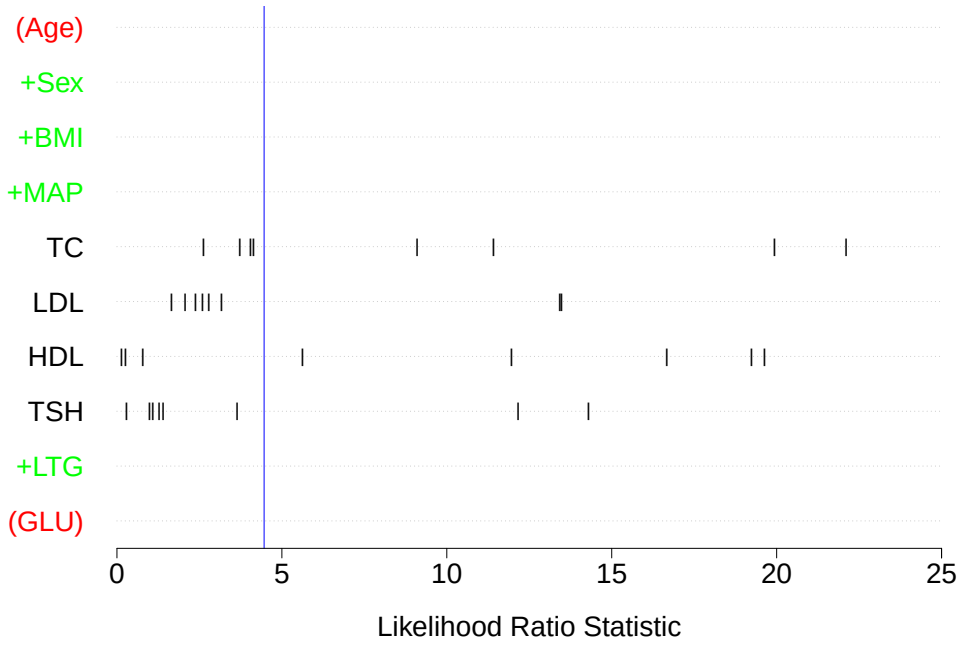

Figure 7.6: Diabetes likelihood ratio plot when Sex, BMI, MAP, LTG $\in \mathcal{M}_k$, and Age, GLU $\in R$, with an empirical threshold of 4.46.

how to use the SIFT procedure with an interaction is presented, again using a false admission rate of 5%. A complication that arises for automated selection procedures when an interaction is introduced is the need to maintain the lower-order terms in the model whilst the interaction term remains in the potential candidate variable set. To exemplify how to achieve this with the use of SIFT, we shall again use the diabetes data, but will also include an interaction term between Age and Sex.

Since an interaction term is present, it will assumed that the main effects of Age and Sex are required, and thus only models that contain Age and Sex in the base shall be considered initially. The minimum and maximum likelihood ratio values for each of the other eight variables and the interaction term are given in Table 7.3. The empirical threshold obtained was 4.59. The minimum values observed for BMI, MAP, and LTG all exceeded this value, and therefore sufficient evidence to warrant their inclusion was observed.

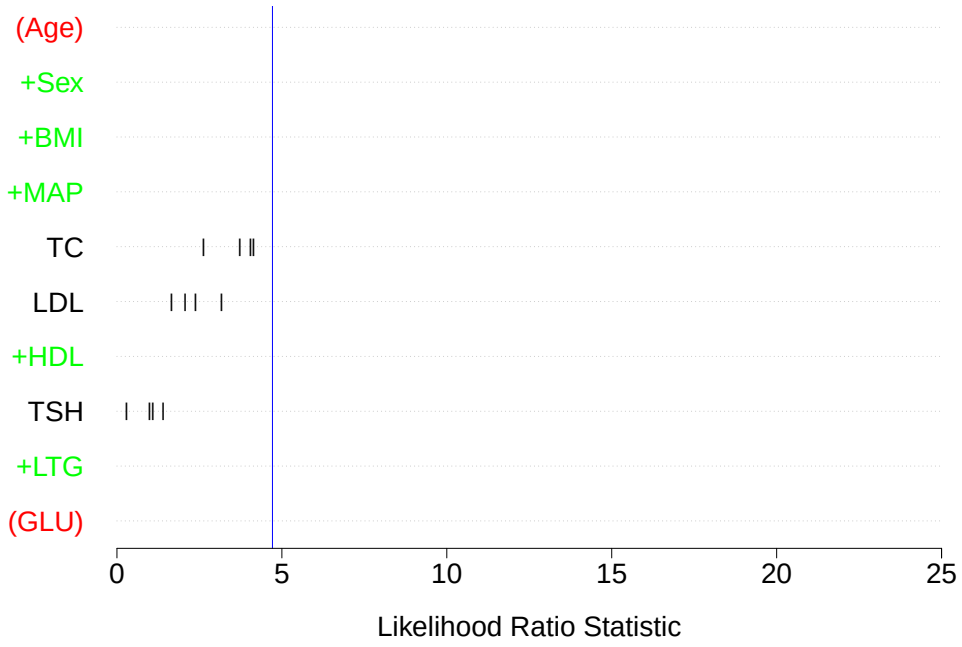

Figure 7.7: Diabetes likelihood ratio plot when Sex, BMI, MAP, HDL, LTG $\in \mathcal{M}_k$, and Age, GLU $\in R$, with an empirical threshold of 4.82.

Table 7.3: Minimum and maximum likelihood ratio values when Age, Sex $\in \mathcal{M}_k$.

Δ_j	Age×Sex	BMI	MAP	TC	LDL	HDL	TSH	LTG	GLU
min	1.3	57.5	18.9	<0.1	<0.1	<0.1	<0.1	18.4	0.6
max	15.0	179.5	86.5	76.0	54.9	146.7	106.5	183.5	60.7

Empirical threshold of 4.59.

As per the previous example, these three variables were added to the base set and the restricted likelihood ratio values were assessed against the revised threshold. This resulted in the data presented in Table 7.4. The minimum likelihood ratio value of 12.0 for the interaction term is greater than the threshold of 4.78, and thus the interaction term must also be included in the final model set.

The process is repeated with the interaction term added, leaving only five

Table 7.4: Minimum and maximum likelihood ratio values when Age, Sex, BMI, MAP, LTG $\in \mathcal{M}_k$.

Δ_j	Age×Sex	BMI	MAP	TC	LDL	HDL	TSH	LTG	GLU
min	12.0	—	—	2.6	1.6	<0.1	0.2	—	0.6
max	15.0	—	—	22.0	13.5	19.5	14.1	—	2.3

Empirical threshold of 4.78.

variables undecided as shown in Table 7.5. The revised threshold was 3.71. None of the remaining variables had a minimum likelihood ratio value greater than this value. Thus, we now turn our attention to the maximum values. We observe that the maximum value for GLU is lower than the threshold value, and therefore should be excluded from further consideration.

Table 7.5: Minimum and maximum likelihood ratio values when Age, Sex, Age×Sex, BMI, MAP, LTG $\in \mathcal{M}_k$.

Δ_j	Age×Sex	BMI	MAP	TC	LDL	HDL	TSH	LTG	GLU
min	—	—	—	2.8	1.8	<0.1	0.3	—	0.9
max	—	—	—	20.1	12.6	17.4	13.2	—	2.3

Empirical threshold of 3.71.

Removing GLU from the set of models under consideration resulted in the likelihood ratio value ranges presented in Table 7.6. The threshold value remained at 3.71. The likelihood ratio statistics for each of the four remaining variables being considered straddled this threshold, so no further clear evidence for inclusion or exclusion was available. In order to establish a final model, the subset possibilities of these undecided variables must now be examined, just as was required in the previous section.

Table 7.6: Minimum and maximum likelihood ratio values when Age, Sex, Age×Sex, BMI, MAP, LTG $\in \mathcal{M}_k$, and GLU $\notin \mathcal{M}_k$.

Δ_j	Age×Sex	BMI	MAP	TC	LDL	HDL	TSH	LTG	GLU
min	—	—	—	2.8	1.8	0.1	0.4	—	—
max	—	—	—	20.1	12.6	17.4	13.2	—	—

Empirical threshold of 3.71.

Table 7.7 shows the results when it was assumed that HDL is in the model. When HDL was assumed to be part of the model, then both LDL and TSH fail to exceed the threshold value of 4.05, indicating their removal. The TC variable still straddled the threshold. However, upon the exclusion of models that included LDL or TSH, only one likelihood ratio value remained for TC—it was equal to 3.52, and thus, TC would consequently be removed. As was the case in the previous section, including any of the alternate three variables, TC, LDL, or TSH, in the base, always resulted in a situation where further variables would need to be added. Thus the set of variables in the final model, when considering an Age by Sex interaction, will be {Age, Sex, Age×Sex, BMI, MAP, HDL, LTG}.

Table 7.7: Minimum and maximum likelihood ratio values when Age, Sex, Age×Sex, BMI, MAP, HDL, LTG $\in \mathcal{M}_k$, and GLU $\notin \mathcal{M}_k$.

Δ_j	Age×Sex	BMI	MAP	TC	LDL	HDL	TSH	LTG	GLU
min	—	—	—	2.8	1.8	—	0.4	—	—
max	—	—	—	4.1	2.6	—	1.7	—	—

Empirical threshold of 4.05.

For this example, the interaction term was retained. If at some point in the analysis, it had been determined that there was insufficient evidence for the inclusion

of the interaction term, it would then be removed from the candidate model set. Upon the removal of the interaction term, the two main effects would be placed into the undecided set and the analysis continued until a final model was determined.

It should be noted that even when considering all possible subsets and choosing the best model from among them, there still may be a better model from a larger model class, as has been observed for these data. For this application, if we used the minimum BIC instead of SIFT, then the same model would have been chosen. If we had used the minimum AIC then the selection of HDL would have been replaced with the selection of both TC and LDL, resulting in a slightly larger model.

CHAPTER 8
CONCLUSION

8.1 Summary

The SIFT procedure was designed to be a model selection method that would allow for the specification of a nominated probability that would limit the chance of returning a model that was overspecified. To that end, the content of this thesis supports the successful attainment of that goal. Under conditions in which the true data generating mechanism is contained within the set of models being evaluated, the SIFT method will asymptotically select all of the terms of the true model while limiting the selection of false or spurious variables to the nominated level.

Using the minimum AIC as a criteria for model selection can lead to a high percentage of overspecified models, which may be undesirable. The use of minimum BIC limits overspecification but may be more conservative than needed. The SIFT procedure allows the researcher to specify the rate at which overspecified models will be identified. This rate may vary depending on the nature of the research being undertaken.

In addition to the development of the SIFT procedure, a number of graphical tools have also been proposed and described, which allow researchers to gain a deeper insight into the relationships contained within the data.

8.2 Limitations

A limitation of the SIFT procedure as it has been developed is that it only assesses the inclusion or exclusion of individual variables; it does not jointly assess the contribution of a group of variables. For instance, consider a factor level variable with q levels, which would be represented in the design matrix as $q - 1$ indicator variables. The SIFT procedure as currently developed requires that the likelihood

ratio values used are based on models that have a difference in parameter space dimensions of one. Thus, the method will separately determine the inclusion or exclusion of each of the $q-1$ indicator variables—some may be retained while others removed. If the researcher is comfortable with this behavior, then the SIFT procedure can be used. If this behavior is undesirable, then the SIFT as currently developed is not suitable. An ad hoc rule could be introduced to remove the set only if all $q-1$ indicators failed to reach the threshold, and to retain the entire set if any of the $q-1$ indicators surpassed the threshold.

Computational issues arise when attempting to analyze all possible subsets. Even for a moderate number of predictor variables, the number of models that needs to be assessed gets large very quickly, doubling for each additional variable in the potential variable set. To use a procedure such as SIFT for larger explanatory variable sets will require the discovery of additional computational efficiencies, and the need for greater computational power, perhaps via parallel processing. The burden could be reduced if some reliable method of initial screening could be quickly implemented to reduce the search set, by identifying either variables that definitely should be in the final model set and/or variables that definitely should not be in the final model set. If this could be achieved without initially fitting all possible models, then the computational burden could be reduced.

8.3 Future Directions

The SIFT procedure is based on the log-likelihood statistic and a set of resulting likelihood ratios; as such, the SIFT procedure should work in a range of likelihood-based modeling frameworks. In particular, it should be immediately applicable to the framework of generalized linear models. Further work investigating the performance of SIFT for various types of generalized linear models could further solidify the benefits of this method.

The relative merits of the naive method compared with the empirical method in determining the evidence threshold could be further explored. Calculation of the empirical threshold is very computationally intensive. The discovery of methods to determine the empirical thresholds more quickly would increase the utility of using these thresholds. One line of research would be to develop a crude empirical threshold that could be computed efficiently from the correlation matrix of the design matrix rather than using a permutation-based approach.

An extension of the procedure that could handle factor variables in a seamless and appropriate manner would also be a valuable methodological addition.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *2nd International Symposium on Information Theory*, Budapest, pp. 267–281. Akedemia Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC-19*(6), 716–723.

Burnham, K. P. and D. R. Anderson (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research 33*(2), 261–304.

Cavanaugh, J. E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters 33*, 201–208.

Claeskens, G. and N. L. Hjort (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–451.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2013, September). Personal Communication.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics 41*, 155–160.

Fujikoshi, Y. and K. Satoh (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika 84*(3), 707–716.

Furnival, G. M. (1971). All possible regressions with less computation. *Technometrics 13*(2), 403–408.

Gorman, J. W. and R. J. Toman (1966). Selection of variables for fitting equations to data. *Technometrics 8*(1), 27–51.

Häggström, M. (2009, September). Blood values sorted by mass and molar concentration. http://commons.wikimedia.org/wiki/File:Blood_values_sorted_by_mass_and_molar_concentration.png.

Hastie, T. (2003, January). Diabetes data. http://www.stanford.edu/∼hastie/Papers/LARS/diabetes.data.

Hastie, T. and B. Efron (2013). *lars: Least Angle Regression, Lasso and Forward Stagewise*. R package version 1.2.

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *The Annals of Mathematical Statistics 11*(3), 271–283.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76*(2), 297–307.

Hurvich, C. M. and C.-L. Tsai (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika 78*(3), 499–509.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kennard, R. W. (1971). A note on the C_p statistic. *Technometrics 13*(4), 899–900.

Konishi, S. and G. Kitagawa (2010). *Information Criteria and Statistical Modeling*. Springer.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Lien, D. and Q. H. Vuong (1987). Selecting the best linear regression model: A classical approach. *Journal of Econometrics 35*, 3–23.

Mallows, C. L. (1973). Some comments on C_P. *Technometrics 15*(4), 661–675.

Mallows, C. L. (1995). More comments on C_P. *Technometrics 37*(4), 362–372.

McQuarrie, A. D. R. and C.-L. Tsai (1998). *Regression and Time Series Model Selection*. World Scientific.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Newton, I. (1686). *Philosophiæ Naturalis Principia Mathematica*.

Schatzoff, M., R. Tsao, and S. Fienberg (1968). Efficient calculation of all possible regressions. *Technometrics 10*(4), 769–779.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics 8*(1), 147–164.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika 68*(1), 45–54.

Sin, C.-Y. and H. White (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics 71*, 207–225.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics A7*, 13–26.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*(2), 307–333.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society 54*(3), 426–482.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics 9*, 60–62.

Yanagihara, H. and C. Ohmoto (2005). On distribution of AIC in linear regression models. *Journal of Statistical Planning and Inference 133*, 417–433.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? a conflict between model identification and regression estimation. *Biometrika 92*(4), 937–950.